GraphPad Statistics Guide

GraphPad Software Inc.
www.graphpad.com

This is one of three companion guides to GraphPad Prism 8.
All are available as web pages on graphpad.com.

This guide was published on April 1, 2020

Table of Contents

1   Welcome to Prism 8 Statistics Guide

This guide examines general principles of statistical analysis, looks at how to conduct those analyses in Prism, and how to interpret results of these analyses.

New to Statistical Analysis?

The [Principles of Statistics][12] section of this guide will help you learn about a number of fundamental topics in statistics!

Statistics in Prism

Check out our [Statistical analyses with Prism][198] section of this guide to start analyzing your data and interpreting your results!

More Guides!

You're currently browsing the Prism Statistics Guide. We also have our User Guide, focused on getting to know Prism, and our Curve Fitting Guide! Both of these guides do more than just explain how to use Prism. They also explain important concepts about data analysis!

- [Switch to the Prism 8 User Guide]

- [Switch to the Prism 8 Curve Fitting Guide]

2   Learn about analyses with Prism

The Prism user's guide explains the [general principles of analyzing data with Prism] including how to [normalize or transform data], and how to [simulate data].

3   How to cite these pages

When citing one of these pages, reference the name of the page, the date you accessed it and the URL. All the pages in the statistics and curve fitting guides were written by GraphPad's founder, Harvey Motulsky.

Use this example as a guide to cite a single page:

H. J. Motulsky, "Advice: When to plot SD vs. SEM", GraphPad Statistics Guide. Accessed 5 March 2016.

http://www.graphpad.com/guides/prism/7/statistics/index.htm?statwhentoplotsdvssem.htm

Note that the URL for each page in these guides is shown at the bottom of the page. Don't use the URL you see in the browser toolbar, as that may be generic and not apply just to the page you are looking at.

Use this example as a guide to citing the entire statistics guide:

H. J. Motulsky, GraphPad Statistics Guide. Accessed 5 March 2016. http://www.graphpad.com/guides/prism/7/statistics/index.htm

4   PRINCIPLES OF STATISTICS

The first half of this Guide reviews general principles of statistics, and is not at all specific to GraphPad Prism. It includes discussions of some important issues that many statistical text books barely mention, including:

- The problem of multiple comparisons[117] and the many ways you can get trapped by multiple comparisons[20] .

- Testing for equivalence[143]

- The danger of using outlier tests with lognormal distributions[157] and the problem of masking[163] which can make it harder to find two outliers than to find one.

- Why it doesn't make sense to automate the decision[148] to use a nonparametric test or not.

- The distinction between SD and SEM[51], and when to display each[52] .

- The advantages of reporting confidence intervals[64] .

- The most common misunderstanding about P values[71], and other misunderstandings[71] .

- Why you can't peek at the results and add more subjects if the results are not quite significant yet.[93]

- A simple [analogy to understand statistical power](#)[98].

- A set of [analysis checklists](#)[169]. Each checklist lists questions you should ask yourself before accepting the results of a statistical analysis.

The [second half of the guide](#)[196] explains how to analyze data with Prism. Even so, much of the content explains the alternative analyses and helps you interpret the results. These sections will prove useful no matter which statistical program you use.

4.1   The big picture

4.1.1   When do you need statistical calculations?

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H. G. Wells

When analyzing data, your goal is simple: You wish to make the strongest possible conclusion from limited amounts of data. To do this, you need to overcome two problems:

- Important findings can be obscured by biological variability and experimental imprecision. This makes it difficult to distinguish real differences from random variation.

- The human brain excels at finding patterns, even in random data. Our natural inclination (especially with our own data) is to conclude that differences are real and to minimize the contribution of random variability. Statistical rigor prevents you from making this mistake.

Statistical analyses are necessary when observed differences are small compared to experimental imprecision and biological variability.

Some scientists ask fundamental questions using clean experimental systems with no biological variability and little experimental error. If this describes your work, you can heed these aphorisms:

- If you need statistics to analyze your experiment, then you've done the wrong experiment.

- If your results speak for themselves, don't interrupt!

Other scientists work in fields where they look for relatively small differences in the face of large amounts of variability. In these fields, statistical methods are essential.

4.1.2   The essential concepts of statistics

If you know twelve concepts about a given topic you will look like an expert to people who only know two or three.

Scott Adams, creator of Dilbert

When learning statistics, it is easy to get bogged down in the details, and lose track of the big picture. Here are the twelve most important concepts in statistical inference.

Statistics lets you make general conclusions from limited data.

The whole point of inferential statistics is to extrapolate from limited data to make a general conclusion. "Descriptive statistics" simply describes data without reaching any general conclusions. But the challenging and difficult aspects of statistics are all about reaching general conclusions from limited data.

Statistics is not intuitive.

The word 'intuitive' has two meanings. One meaning is "easy to use and understand." That was my goal when I wrote *Intuitive Biostatistics*. The other meaning of 'intuitive' is "instinctive, or acting on what one feels to be true even without reason." Using this definition, statistical reasoning is far from intuitive. When thinking about data, intuition often leads us astray. People frequently see patterns in random data and often jump to unwarranted conclusions. Statistical rigor is needed to make valid conclusions from data.

Statistical conclusions are always presented in terms of probability.

"Statistics means never having to say you are certain."   If a statistical conclusion ever seems certain, you probably are misunderstanding something. The whole point of statistics is to quantify uncertainty.

All statistical tests are based on assumptions.

Every statistical inference is based on a list of assumptions. Don't try to interpret any statistical results until after you have reviewed that list. An assumption behind every statistical calculation is that the data were randomly sampled, or at least representative of, a larger population of values that could have been collected. If your data are not representative of a larger set of data you could have collected (but didn't), then statistical inference makes no sense.

Decisions about how to analyze data should be made in advance.

Analyzing data requires many decisions. Parametric or nonparametric test? Eliminate outliers or not? Transform the data first? Normalize to external control values? Adjust for covariates? Use weighting factors in regression? All these decisions (and more) should be part of experimental design. When decisions about statistical analysis are made after inspecting the data, it is too easy for statistical analysis to become a high-tech Ouja board -- a method to produce preordained results, rather an objective method of analyzing data. The new name for this is p-hacking.

A confidence interval quantifies precision, and is easy to interpret.

Say you've computed the mean of a set of values you've collected,or the proportion of subjects where some event happened. Those values describe the sample you've analyzed. But what about the overall population you sampled from? The true population mean (or proportion) might be higher, or it might be lower. The calculation of a 95% confidence interval takes into account sample size and scatter. Given a set of assumptions, you can be 95% sure that the confidence interval includes the true population value (which you could only know for sure by collecting an infinite amount of data). Of course, there is nothing special about 95% except tradition. Confidence intervals can be computed for any degree of desired confidence. Almost all results -- proportions, relative risks, odds ratios, means, differences between means, slopes, rate constants... -- should be accompanied with a confidence interval.

A P value tests a null hypothesis, and is hard to understand at first.

The logic of a P value seems strange at first. When testing whether two groups differ (different mean, different proportion, etc.), first hypothesize that the two populations are, in fact, identical. This is called the null

hypothesis. Then ask: If the null hypothesis were true, how unlikely would it be to randomly obtain samples where the difference is as large (or even larger) than actually observed? If the P value is large, your data are consistent with the null hypothesis. If the P value is small, there is only a small chance that random chance would have created as large a difference as actually observed. This makes you question whether the null hypothesis is true. If you can't identify the null hypothesis, you cannot interpret the P value.

"Statistically significant" does not mean the effect is large or scientifically important.

If the P value is less than 0.05 (an arbitrary, but well accepted threshold), the results are deemed to be statistically significant. That phrase sounds so definitive. But all it means is that, by chance alone, the difference (or association or correlation..) you observed (or one even larger) would happen less than 5% of the time. That's it. A tiny effect that is scientifically or clinically trivial can be statistically significant (especially with large samples). That conclusion can also be wrong, as you'll reach a conclusion that results are statistically significant 5% of the time just by chance.

"Not significantly different" does not mean the effect is absent, small or scientifically irrelevant.

If a difference is not statistically significant, you can conclude that the observed results are not inconsistent with the null hypothesis. Note the double negative. You cannot conclude that the null hypothesis is true. It is quite possible that the null hypothesis is false, and that there really is a difference between the populations. This is especially a problem with small sample sizes. It makes sense to define a result as being statistically significant or not statistically significant when you need to make a decision based on this one result. Otherwise, the concept of statistical significance adds little to data analysis.

Multiple comparisons make it hard to interpret statistical results.

When many hypotheses are tested at once, the problem of multiple comparisons makes it very easy to be fooled. If 5% of tests will be "statistically significant" by chance, you expect lots of statistically significant results if you test many hypotheses. Special methods can be used to reduce the problem of finding false, but statistically significant, results, but these methods also make it harder to find true effects. Multiple comparisons can be insidious. It is only possible to correctly

interpret statistical analyses when all analyses are planned, and all planned analyses are conducted and reported. However, these simple rules are widely broken.

Correlation does not mean causation.

A statistically significant correlation or association between two variables may indicate that one variable causes the other. But it may just mean that both are influenced by a third variable. Or it may be a coincidence.

Published statistics tend to be optimistic.

By the time you read a paper, a great deal of selection has occurred. When experiments are successful, scientists continue the project. Lots of other projects get abandoned.When the project is done, scientists are more likely to write up projects that lead to remarkable results, or to keep analyzing the data in various ways to extract a "statistically significant" conclusion. Finally, journals are more likely to publish "positive" studies. If the null hypothesis were true, you would expect a statistically significant result in 5% of experiments. But those 5% are more likely to get published than the other 95%.

4.1.3 Extrapolating from 'sample' to 'population'

The basic idea of statistics is simple:

  You want to use limited amounts of data to make general conclusions.

To do this, statisticians have developed methods based on a simple model: Assume that an infinitely large population of values exists and that your data (your 'sample') was randomly selected from this population. Analyze your sample and use the rules of probability to make inferences about the overall population.

This model is an accurate description of some situations. For example, quality control samples really are randomly selected from a large population. Clinical trials do not enroll a randomly selected sample of patients, but it is usually reasonable to extrapolate from the sample you studied to the larger population of similar patients.

In a typical experiment, you don't really sample from a population, but you do want to extrapolate from your data to a more general conclusion. The concepts of sample and population can still be used if you define the sample to be the data you collected and the population to be the data you would have collected if you had repeated the experiment an infinite number of times.

The problem is that the statistical inferences can only apply to the population from which your samples were obtained, but you often want to make conclusions that extrapolate even beyond that large population. For example, you perform an experiment in the lab three times. All the experiments used the same cell preparation, the same buffers, and the same equipment. Statistical inferences let you make conclusions about what would probably happen if you repeated the experiment many more times with that same cell preparation, those same buffers, and the same equipment.

You probably want to extrapolate further to what would happen if someone else repeated the experiment with a different source of cells, freshly made buffer, and different instruments. Unfortunately, statistical calculations can't help with this further extrapolation. You must use scientific judgment and common sense to make inferences that go beyond the limitations of statistics.

4.1.4   Why statistics can be hard to learn

Three factors make statistics hard to learn for some.

Probability vs. statistics

The whole idea of statistics is to start with a limited amount of data and make a general conclusion (stated in terms of probabilities). In other words, you use the data in your sample to make general conclusions about the population from which the data were drawn.

Probability theory goes the other way. You start with knowledge about the general situation, and then compute the probability of various outcomes. The details are messy, but the logic is pretty simple.

Statistical calculations rest on probability theory, but the logic of probability is opposite to the logic of statistics. Probability goes from general to specific, while statistics goes from specific to general. Applying the mathematics of probability to statistical analyses requires reasoning that can sometimes seem convoluted.

Statistics uses ordinary words in unusual ways

All fields have technical terms with specific meanings. In many cases, statistics uses words that you already know, but give them specific meaning. "Significance", "hypothesis", "confidence", "error", "normal" are all common words that statistics uses in very specialized ways. Until you learn the statistical meaning of these terms, you can be very confused when reading statistics books or talking to statisticians. The problem isn't that you don't understand a technical term. The problem is that you think you know what the term means, but are wrong. As you read these help screens be sure to pay attention to familiar terms that have special meanings in statistics.

> *When I use a word, it means just what I choose it to* mean — *neither more nor less.*
>
> Humpty Dumpty (amateur statistician) in Through the Looking Glass

Statistics is on the interface of math and science

Statistics is a branch of math, so to truly understand the basis of statistics you need to delve into the mathematical details. However, you don't need to know much math to use statistics effectively and to correctly interpret the results. Many statistics books tell you more about the mathematical basis of statistics than you need to know to use statistical methods effectively. The focus here is on selecting statistical methods and making sense of the results, so this presentation uses very little math. If you are a math whiz who thinks in terms of equations, you'll want to learn statistics from a mathematical book.

Parts of this page are excerpted from Chapter 2 of Motulsky, H.J. (2010). Intuitive Biostatistics, 2nd edition. Oxford University Press. ISBN=978-0-19-973006-3.

4.1.5   Don't be a P-hacker

Overview

Vickers told this story (1):

> Statistician: "Oh, so you have already calculated the P value?"
>
> Surgeon: "Yes, I used multinomial logistic regression."
>
> Statistician: "Really? How did you come up with that?"
>
> Surgeon: "Well, I tried each analysis on the SPSS drop-down menus, and that was the one that gave the smallest P value".

Basic rules of statistics

For statistical analyses to be interpretable at face value, it is essential that these three statements be true:

- All analyses were planned.

- All planned analyses were conducted exactly as planned and then reported.

- You take into account all the analyses when interpreting the results.

These simple and sensible rules are commonly violated in many ways as explained below.

Multiple ways to preprocess the data

Before the data are analyzed, some decisions get made. Which values should be deleted because they are so high or so low that they are considered to be mistakes? Whether and how to normalize? Whether and how to transform the data?

Sequential Analyses (ad hoc sample size determination)

To properly interpret a P value, the experimental protocol has to be set in advance. Usually this means choosing a sample size, collecting data, and then analyzing it.

But what if the results aren't quite statistically significant? It is tempting to run the experiment a few more times (or add a few more subjects), and then analyze the data again, with the larger sample size. If the results still aren't "significant", then do the experiment a few more times (or add more subjects) and reanalyze once again.

When data are analyzed in this way, it is impossible to interpret the results. This informal sequential approach should not be used.

If the null hypothesis of no difference is in fact true, the chance of obtaining a "statistically significant" result using that informal sequential approach is far higher than 5%. In fact, if you carry on that approach long enough, then every single experiment will eventually reach a "significant" conclusion, even if the null hypothesis is true. Of course, "long enough" might be very long indeed and exceed your budget or even your lifespan.

The problem is that the experiment continues when the result is not "significant", but stops when the result is "significant". If the experiment was continued after reaching "significance", adding more data might then result in a "not significant" conclusion. But you'd never know this, because the experiment would have been terminated once "significance" was reached. If you keep running the experiment when you don't like the results, but stop the experiment when you like the results, the results are impossible to interpret.

Statisticians have developed rigorous ways to handle sequential data analysis. These methods use much more stringent criteria to define "significance" to account for the sequential analyses. Without these special methods, you can't interpret the results unless the sample size is set in advance

Multiple Subgroups

Analyzing multiple subgroups of data is a form of multiple comparisons. When a treatment works in some subgroups but not others, analyses of subgroups becomes a form of multiple comparisons and it is easy to be fooled.

A simulated study by Lee and coworkers points out the problem. They pretended to compare survival following two "treatments" for coronary artery disease. They studied a group of real patients with coronary artery disease who they randomly divided into two groups. In a real study, they would give the two groups different treatments, and compare survival. In

this simulated study, they treated the subjects identically but analyzed the data as if the two random groups actually represented two distinct treatments. As expected, the survival of the two groups was indistinguishable (2).

They then divided the patients into six groups depending on whether they had disease in one, two, or three coronary arteries, and depending on whether the heart ventricle contracted normally or not. Since these are variables that are expected to affect survival of the patients, it made sense to evaluate the response to "treatment" separately in each of the six subgroups. Whereas they found no substantial difference in five of the subgroups, they found a striking result among the sickest patients. The patients with three-vessel disease who also had impaired ventricular contraction had much better survival under treatment B than treatment A. The difference between the two survival curves was statistically significant with a P value less than 0.025.

If this were an actual study, it would be tempting to conclude that treatment B is superior for the sickest patients, and to recommend treatment B to those patients in the future. But this was not a real study, and the two "treatments" reflected only random assignment of patients. The two treatments were identical, so the observed difference was absolutely positively due to chance.

It is not surprising that the authors found one low P value out of six comparisons. There is a 26% chance that one of six independent comparisons will have a P value less than 0.05, even if all null hypotheses are true.

If all the subgroup comparisons are defined in advance, it is possible to correct for many comparisons – either as part of the analysis or informally while interpreting the results. But when this kind of subgroup analysis is not defined in advance, it becomes a form of "data torture".

Multiple Predictions

In 2000, the Intergovernmental Panel on Climate Change made predictions about future climate. Pielke asked what seemed like a straightforward question: How accurate were those predictions over the next seven years? That's not long enough to seriously assess predictions of global warming, but it is a necessary first step. Answering this question proved to be impossible. The problems are that the report contained numerous predictions, and didn't specify which sources of climate data

should be used. Did the predictions come true? The answer depends on the choice of which prediction to test and which data set you test it against -- "a feast for cherry pickers" (3)

You can only evaluate the accuracy of predictions or diagnoses when the prediction unambiguously stated what was being predicted and when it would happen.

Combining Groups

When comparing two groups, the groups must be defined as part of the study design. If the groups are defined by the data, many comparisons are being made implicitly and ending the results cannot be interpreted.

Austin and Goldwasser demonstrated this problem(4). They looked at the incidence of hospitalization for heart failure in Ontario (Canada) in twelve groups of patients defined by their astrological sign (based on their birthday). People born under the sign of Pisces happened to have the highest incidence of heart failure. They then did a simple statistics test to compare the incidence of heart failure among people born under Pisces with the incidence of heart failure among all others (born under all the other eleven signs, combined into one group). Taken at face value, this comparison showed that the difference in incidence rates is very unlikely to be due to chance (the P value was 0.026). Pisces have a "statistically significant" higher incidence of heart failure than do people born in the other eleven signs.

The problem is that the investigators didn't test really one hypothesis; they tested twelve. They only focused on Pisces after looking at the incidence of heart failure for people born under all twelve astrological signs. So it isn't fair to compare that one group against the others, without considering the other eleven implicit comparisons. After correcting for those multiple comparisons, there was no significant association between astrological sign and heart failure.

Multiple regression, logistic regression, etc.

Fitting a multiple regression model provides even more opportunities to try multiple analyses:

• Try including or excluding possible confounding variables.

• Try including or excluding interactions.

- Change the definition of the outcome variable.

- Transform the outcome or any of the independent variables to logarithms or reciprocals or something else.

Unless these decisions were made in advance, the results of multiple regression (or multiple logistic or proportional hazards regression) cannot be interpreted at face value.

Chapter 38 of Intuitive Biostatistics(8) explains this problem of overfitting, as does Babyok (5).

The garden of forking paths

In some cases, you first look at the data (and perhaps do a preliminary analysis) and then decide what test to run next depending on those values. Gelman calls this "the garden of forking paths" and states that it is a form of multiple comparisons (10).

Publication Bias

Editors prefer to publish papers that report results that are statistically significant. Interpreting published results becomes problematic when studies with "not significant" conclusions are abandoned, while the ones with "statistically significant" results get published. This means that the chance of observing a 'significant' result in a published study can be much greater than 5% even if the null hypotheses are all true.

Turner demonstrated this kind of selectivity -- called publication bias -- in industry-sponsored investigations of the efficacy of antidepressant drugs (6). Between 1987 and 2004, the Food and Drug Administration (FDA) reviewed 74 such studies, and categorized them as "positive", "negative" or "questionable". The FDA reviewers found that 38 studies showed a positive result (the antidepressant worked). All but one of these studies was published. The FDA reviewers found that the remaining 36 studies had negative or questionable results. Of these, 22 were not published, 11 were published with a 'spin' that made the results seem somewhat positive, and only 3 of these negative studies were published with clear negative findings.

The problem is a form of multiple comparisons. Many studies are done, but only some are published, and these are selected because they show "desired" results.

Bottom line

Statistical analyses can be interpreted at face value only when all steps are planned, all planned analyses are published, and all the results are considered when reaching conclusions. These simple rules are violated in many ways in common statistical practice.

If you try hard enough, eventually 'statistically significant' findings will emerge from any reasonably complicated data set. This is called *data torture* (6) or *P-hacking (9)*. When reviewing results, you often can't even correct for the number of ways the data were analyzed since the number of possible comparisons was not defined in advance, and is almost unlimited. When results were analyzed many ways without a plan, the results simply cannot be interpreted. At best, you can treat the findings as an hypothesis to be tested in future studies with new data.

References

1. Vickers, A., What is a p value anyway, 2009. ISBN: 978-0321629302.

2. Lee, K. L., J. F. McNeer, C. F. Starmer, P. J. Harris, and R. A. Rosati. 1980. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation 61, (3) (Mar): 508-15

3. Pielke, R. Prometheus: Forecast verification for climate scient, part 3. Retrieved April 20, 2008.

4. Austin, P. C., and M. A. Goldwasser. 2008. Pisces did not have increased heart failure: Data-driven comparisons of binary proportions between levels of a categorical variable can result in incorrect statistical significance levels. Journal of Clinical Epidemiology 61, (3) (Mar): 295-300.

5. Babyak, M.A.. What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. Psychosomatic Medicine (2004) vol. 66 (3) pp. 411

6. Mills, J. L. 1993. Data torturing. New England Journal of Medicine 329, (16): 1196.

7. Turner, E. H., A. M. Matthews, E. Linardatos, R. A. Tell, and R. Rosenthal. 2008. Selective publication of antidepressant trials and its influence on apparent efficacy. The New England Journal of Medicine 358, (3) (Jan 17): 252-60.

8. Motulsky, H.J. (2010). [Intuitive Biostatistics](#), 2nd edition. Oxford University Press. ISBN=978-0-19-973006-3.

9. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22, 1359–1366 (2011).

10. Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or 'p-hacking' and the research hypothesis was posited ahead of time. Downloaded January, 30, 2014.

4.1.6   How to report statistical results

The guidelines below are an opinionated guide about how to present data and analyses. Of course, you also need to report details of experimental design, including blinding and randomization.

Overall

- Every statistical paper should report all methods (including those used to process and analyze the data) completely enough so someone else could reproduce the work exactly.

- Every figure and table should present the data clearly (and not be exaggerated in a way to emphasize your conclusion).

- All the results should be reported completely enough that no one wonders what you actually did.

The analyses before the analyses

- Did you decide to normalize? Remove outliers? Transform to logarithms? Smooth? Remove a baseline? Justify these decisions, and report enough details so anyone could start with your data and get exactly the same results. State whether these calculations were preplanned or only decided upon after seeing the data.

- If outliers were eliminated, say how many there were, what criteria you used to identify them, and whether these criteria were chosen in advance as part of the experimental design.

Sample size

- Report how you chose sample size[108].

- Explain exactly what was counted when reporting sample size. When you say n=3, do you mean three different animals, three different assays on tissue from one animal, one assay from tissue pooled from three animals, three repeat counts in a gamma counter from a preparation made from one run of an experiment...?

- State whether you choose sample size in advance[110], or adjusted sample size in an ad hoc manner as you saw the results accumulate.

- If the sample sizes of the groups are not equal, explain why.

Avoid P-hacking

- For each analysis (usually for each figure and table), state whether every step in data analysis followed a preplanned protocol or not. If you only decided to remove outliers after seeing the data, say so. If you only decided to use a nonparametric test after seeing the data, say so. If you only decided to analyze the logarithms of the data after viewing the data, say so.

- If you don't show every analysis you did, at least describe and enumerate them.

- If you started with one sample size and ended with another sample size, explain exactly how you decided to add additional samples and/or eliminate samples. State whether these decisions were based on a preset protocol, or were decided during the course of the experiment.

Graphing data

- Present data clearly. Focus on letting the reader see the data, and not only your conclusions.

- When possible, graph the individual data, not a summary of the data. If there are too many values to show in scatter plots, consider box-and-whisker plots or frequency distributions.

- If you choose to plot means with error bars, graph standard deviation[42] error bars which show variability, rather than standard error of the mean[49] error bars, which do not.

Statistical methods

- State the full name of the test. Don't say "t test", say "paired t test".

- Identify the program of the program that did the calculations (including detailed version number, which for GraphPad Prism might be 7.01).

- State all options you selected. Repeated measures? Correcting for unequal variances? Robust regression? Constraining parameters? Sharing parameters? Report enough detail so anyone could start with your data and get precisely the same results you got.

Reporting effect size

- The most important result of most experiments is an effect size. How big was the difference (or ratio or percent increase)? Or how strongly were two variables correlated? In almost all cases, you can summarize this effect size with a single value and should report this effect with a confidence interval, usually the 95% interval. This is by far the most important finding to report in a paper and its abstract.

- Consider showing a graph of effect sizes (i.e. differences or ratios) with 95% confidence intervals.

Reporting P values

- When possible, report the P value as a number with a few digits of precision, not an inequality. For example say "the P value was 0.0234" rather than "P < 0.05".

- If there is any possible ambiguity, clearly state the null hypothesis the P value tests. If you don't know the null hypothesis, then you shouldn't report a P value (since every P value tests a null hypothesis)!

- When comparing two groups, state if the P value is one- or two-sided[75] (which is the same as one- or two-tailed). If one-sided, state that you predicted the direction of the effect before collecting data (and recorded this prediction), and recorded that decision and prediction. If you didn't make this decision and prediction before collecting data, you should not report a one-sided P value.

Reporting statistical hypothesis testing (significance)

- Statistical hypothesis testing is used to make a firm decision based on a single P value. One use is choosing between the fit of two alternative models. If the P value is less than a preset threshold you pick one model, otherwise the other. When doing this, state both models, the method you are using to choose between them, the preset threshold P value, and the model you chose. Perhaps also report the goodness of fit of both models.

- When comparing groups, you don't always make a decision based on the result. If you *are* making a crisp decision, report the threshold P value, whether the computed P value was greater or less than the threshold, and the accompanying decision. If you are *not* making a decision, report the effect with its confidence interval, and perhaps a P value. If you are not making a decision based on that P value, then it doesn't really matter whether or not the P value was less than a threshold or not, and the whole idea of statistical hypothesis testing isn't really useful.

- The word "significant" has two related meanings, so has caused lots of confusion in science. The two bullet points above demonstrate that the results of statistical hypothesis testing can (and in my opinion should) be reported without using the word "significant". If you do choose to use the word "significant" in this context, always precede it with "statistically", so there is no confusion.

- Never use the word "significant" when discussing the clinical or physiological impact of a result. Instead use words like "large", "substantial", and "clinically relevant". Using "significant" in this context just leads to confusion.

Multiple comparisons

- Multiple comparisons must be handled thoughtfully, and all steps must be documented. Note that the [problem of multiple comparisons is widespread][117], and isn't just an issue when doing follow-up tests after ANOVA.

- State whether or not all comparisons were planned, and all planned comparisons were reported. If you report unplanned comparisons or omit some comparisons, the results must be identified as preliminary.

- If you used any correction for multiple comparisons, explain the details.

- If you report multiplicity adjusted P values, point out clearly that these P values were adjusted.

Other guides to presenting statistical results

1. Curtis, M. J., Bond, R. A., Spina, D., Ahluwalia, A., Alexander, S. P. A., Giembycz, M. A., et al. (2015). [Experimental design and analysis and their reporting: new guidance for publication in BJP]. Br J Pharmacol, 172(14), 3461–3471.

2. Altman DG, Gore SM, Gardner MJ, Pocock SJ (1983). Statistical guidelines for contributors to medical journals. Brit Med J 286: 1489–1493.

4.1.7   Ordinal, interval and ratio variables

Many statistics books begin by defining the different kinds of variables you might want to analyze. This scheme was developed by S. Stevens and published in 1946.

Definitions

A **categorical** variable, also called a nominal variable, is for mutually exclusive, but not ordered, categories. For example, your study might compare five different genotypes. You can code the five genotypes with numbers if you want, but the order is arbitrary and any calculations (for example, computing an average) would be meaningless.

An **ordinal** variable, is one where the order matters but not the difference between values. For example, you might ask patients to express the amount of pain they are feeling on a scale of 1 to 10. A score of 7 means more pain than a score of 5, and that is more than a score of 3. But the difference between the 7 and the 5 may not be the same as that between 5 and 3. The values simply express an order. Another example would be movie ratings, from * to *****.

An **interval** variable is a one where the difference between two values is meaningful. The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.

A **ratio** variable, has all the properties of an interval variable, but also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable. Variables like height, weight, enzyme activity are ratio variables. Temperature, expressed in F or C, is not a ratio variable. A temperature of 0.0 on either of those scales does not mean 'no heat. However, temperature in Kelvin is a ratio variable, as 0.0 Kelvin really does mean 'no heat'. Another counter example is pH. It is not a ratio variable, as pH=0 just means 1 molar of H+. and the definition of molar is fairly arbitrary. A pH of 0.0 does not mean 'no acidity' (quite the opposite!). When working with ratio variables, but not interval variables, you can look at the ratio of two measurements. A weight of 4 grams is twice a weight of 2 grams, because weight is a ratio variable. A temperature of 100 degrees C is not twice as hot as 50 degrees C, because temperature C is not a ratio variable. A pH of 3 is not twice as acidic as a pH of 6, because pH is not a ratio variable.

The categories are not as clear cut as they sound. What kind of variable is color? In some experiments, different colors would be regarded as nominal. But if color is quantified by wavelength, then color would be considered a ratio variable. The classification scheme really is somewhat fuzzy.

What is OK to compute

OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution	Yes	Yes	Yes	Yes
median and percentiles	No	Yes	Yes	Yes
sum or difference	No	No	Yes	Yes
mean, standard deviation, standard error of the mean	No	No	Yes	Yes
ratio, or coefficient of variation	No	No	No	Yes

Does it matter?

It matters if you are taking an exam in statistics, because this is the kind of concept that is easy to test for.

Does it matter for data analysis? The concepts are mostly pretty obvious, but putting names on different kinds of variables can help prevent mistakes like taking the average of a group of postal (zip) codes, or taking the ratio of two pH values. Beyond that, putting labels on the different kinds of variables really doesn't really help you plan your analyses or interpret the results.

4.1.8   The need for independent samples

Statistical tests are based on the assumption that each subject (or each experimental unit) was sampled independently of the rest. Data are independent when any random factor that causes a value to be too high or too low affects only that one value. If a random factor (one that you didn't account for in the analysis of the data) can affect more than one value, but not all of the values, then the data are not independent.

The concept of independence can be difficult to grasp. Consider the following three situations.

- You are measuring blood pressure in animals. You have five animals in each group, and measure the blood pressure three times in each animal. You do not have 15 independent measurements. If one animal has higher blood pressure than the rest, all three measurements in that animal are likely to be high. You should average the three measurements in each animal. Now you have five mean values that are independent of each other.

- You have done a biochemical experiment three times, each time in triplicate. You do not have nine independent values, as an error in preparing the reagents for one experiment could affect all three triplicates. If you average the triplicates, you do have three independent mean values.

- You are doing a clinical study and recruit 10 patients from an inner-city hospital and 10 more patients from a suburban clinic. You have not independently sampled 20 subjects from one population. The data from the 10 inner-city patients may be more similar to each other than to the data from the suburban patients. You have sampled from two populations and need to account for that in your analysis.

4.1.9 Intuitive Biostatistics (the book)
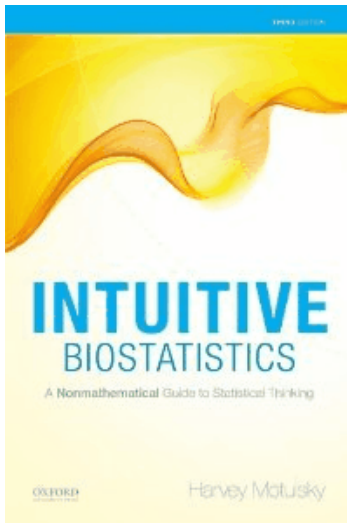
H.J. Motulsky, Intuitive Biostatistics, ISBN: 978-0199946648, 3rd edition 2014

Table of contents

Excerpts

Reviews

Intuitive Biostatistics is a beautiful book that has much to teach experimental biologists of all stripes. Unlike other statistics texts I have seen, it includes extensive and carefully crafted discussions of the perils of multiple comparisons, warnings about common and avoidable mistakes in data analysis, a review of the assumptions that apply to various tests, an emphasis on confidence intervals rather than P values, explanations as

> *to why the concept of statistical significance is rarely needed in scientific work, and a clear explanation of nonlinear regression (commonly used in labs; rarely explained in statistics books).*
>
> *In fact, I am so pleased with Intuitive Biostatistics that I decided to make it the reference of choice for my postdoctoral associates and graduate students, all of whom depend on statistics, and most of whom need a closer awareness of precisely why. Motulsky has written thoughtfully, with compelling logic and wit. He teaches by example what one may expect of statistical methods and perhaps just as importantly, what one may not expect of them. He is to be congratulated for this work, which will surely be valuable and perhaps even transformative for many of the scientists who read it.*
>
> —Bruce Beutler, 2011 Nobel Laureate, Physiology or Medicine,   Director, Center for the Genetics of Host Defense, UT Southwestern Medical Center

If you like the style of this guide, you'll also appreciate the introductory text I wrote: *Intuitive Biostatistics*.

Overview

Intuitive Biostatistics is both an introduction and review of statistics. Compared to other books, it has:

- Breadth rather than depth. It is a guidebook, not a cookbook.

- Words rather than math. It has few equations.

- Explanations rather than recipes. This book presents few details of statistical methods and only a few tables required to complete the calculations.

Who is it for?

I wrote Intuitive Biostatistics for three audiences:

- Medical (and other) professionals who want to understand the statistical portions of journals they read. These readers don't need to analyze any data, but need to understand analyses published by others. I've tried to explain the big picture, without getting bogged down in too many details.

- Undergraduate and graduate students, post-docs and researchers who will analyze data. This book explains general principles of data analysis, but it won't teach you how to do statistical calculations or how to use any particular statistical program. It makes a great companion to the more traditional statistics texts and to the documentation of statistical software.

- Scientists who consult with statisticians. Statistics often seems like a foreign language, and this text can serve as a phrase book to bridge the gap between scientists and statisticians. Sprinkled throughout the book are "Lingo" sections that explain statistical terminology, and point out when statistics gives ordinary words very specialized meanings (the source of much confusion).

4.1.10   Essential Biostatistics (the book)
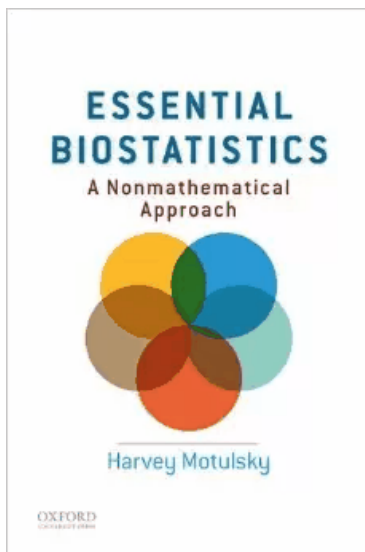
H.J. Motulsky
Essential Biostatistics,
ISBN: 978-0199365067
$20

See the table of contents and two full chapters at

Some ways in which this book is unique

- It doesn't explain how to calculate any statistical tests. In fact, it only includes two equations.

- Chapter 1 is a fun chapter that explains how common sense can lead you astray and why we therefore need to understand statistical principles.

- Chapter 2 is a unique approach to appreciating the complexities of probability.

- I introduce statistical thinking with Chapter 4, which explains the confidence interval of a proportion. This lets me explain the logic of generalizing from sample to population using a confidence interval before having to deal with concepts about how to quantify the scatter.

- I explain comparing groups with confidence intervals (Chapter 12) before explaining P values (Chapter 13) and statistical significance (Chapters 14 and 15). This way I could delay as long as possible dealing with the confusing

www.essentialbiostatistics.com

concept of a P value and the overused word "significant".

- Chapter 16 explains how common Type I errors are, and the difference between a significance level and the false discovery rate.

- Chapter 19 explains all common statistical tests as a series of tables.

- I include topics often omitted from introductory texts, but that I consider to be essential, including: multiple comparisons, the false discovery rate, p-hacking, lognormal distributions, geometric mean, normality tests, outliers and nonlinear regression.

- Nearly every chapter has a Lingo section explaining how statistical terminology can be misunderstood.

- Nearly every chapter includes a Common Mistakes section, and Chapter 25 explains more general mistakes to avoid.

"Essential Biostatistics distills the essence of university-level biostatistics topics in accessible concise language that is engaging and thought-provoking. Students and practitioners of biostatistics will find Intuitive Biostatistics: The Essentials to be an excellent resource that provides clarity on major statistical concepts and procedures while also illuminating erroneous statistical conclusions many fall prey to. This text would be an excellent companion to a traditional biostatistics book."
--Derek Webb, Bemidji State University

"The author does a great job explaining why we use statistics rather than getting bogged down explaining how we calculate statistics. I find it refreshing to step back from the calculations to see the larger context of why we use statistics in science."
--Dean W. Coble, Stephen F. Austin State University

"I really like the clear and humorous style, the wealth of examples, and the discussions of the limits and pitfalls. This is a wonderful book."
--*Naji Younes, George Washington University*

"Motulsky seems to have done the impossible, again. He has taken his already great textbook and extracted the bare-bones necessary for the reader to enjoy a lively, easy-to-read introduction to the concepts of biostatistics. In addition, Motulsky provides the reader with a discussion of common mistakes and how to avoid them. This is invaluable for a true understanding of biostatistics. Essential Biostatistics should be required reading for all beginning biology or biostatistics students. It provides foundational material for interpreting statistical analysis."
--*Philip Hejduk, University of Texas at Arlington*

4.2   The Gaussian distribution

"*Everybody believes in the [Gaussian distribution]:*

the experimenters, because they think it can be

proved by mathematics; and the mathematicians,

because they believe it has been established by

observation."

W. Lippmann

4.2.1   Importance of the Gaussian distribution

Statistical tests analyze a particular set of data to make more general conclusions. There are several approaches to doing this, but the most common is based on assuming that data in the population have a certain

distribution. The distribution used most commonly by far is the bell-shaped Gaussian distribution, also called the Normal distribution. This assumption underlies many statistical tests such as t tests and ANOVA, as well as linear and nonlinear regression.

When reading in other books about the Gaussian distribution, two statistical terms might be confusing because they sound like ordinary words:

- In statistics, the word "normal" is another name for a Gaussian, bell-shaped, distribution. In other contexts, of course, the word "normal" has very different meanings (absence of disease or common).

- Statisticians refer to the scatter of points around the line or curve as "error". This is a different use of the word than is used ordinarily. In statistics, the word "error" simply refers to deviation from the average. The deviation is usually assumed to be due to biological variability or experimental imprecision, rather than a mistake (the usual use of the word "error").

4.2.2   Origin of the Gaussian distribution

The Gaussian distribution emerges when many independent random factors act in an additive manner to create variability. This is best seen by an example.

Imagine a very simple "experiment". You pipette some water and weigh it. Your pipette is supposed to deliver 10 microliter of water, but in fact delivers randomly between 9.5 and 10.5 microliters. If you pipette one thousand times and create a frequency distribution histogram of the results, it will look like the figure below.
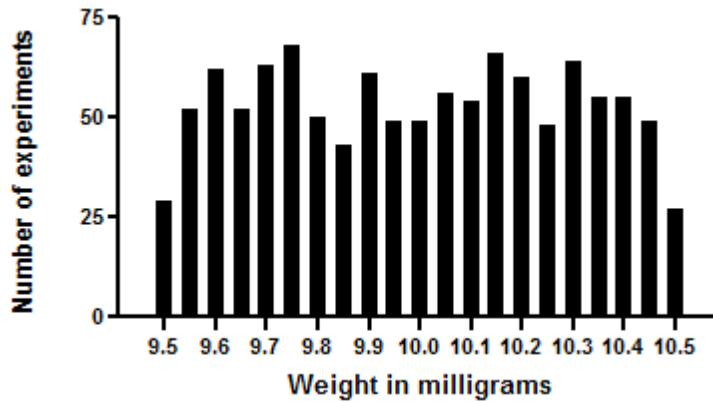
The average weight is 10 milligrams, the weight of 10 microliters of water (at least on earth). The distribution is flat, with no hint of a Gaussian distribution.

Now let's make the experiment more complicated. We pipette twice and weigh the result. On average, the weight will now be 20 milligrams. But you expect the errors to cancel out some of the time. The figure below is what you get.
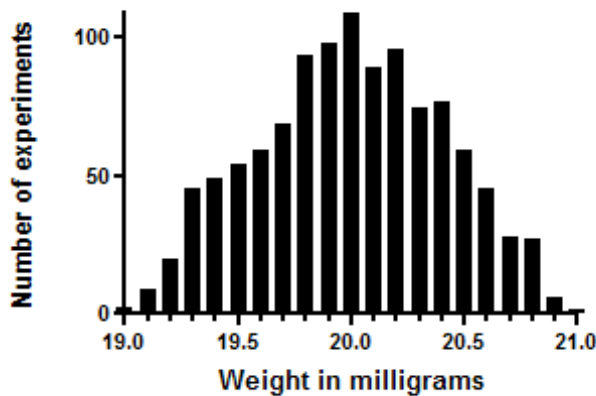
Each pipetting step has a flat random error. Add them up, and the distribution is not flat. For example, you'll get weights near 21 mg only if both pipetting steps err substantially in the same direction, and that is rare.

Now let's extend this to ten pipetting steps, and look at the distribution of the sums.
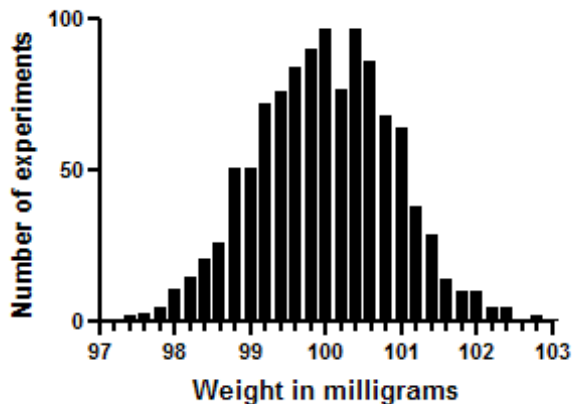
The distribution looks a lot like an ideal Gaussian distribution. Repeat the experiment 15,000 times rather than 1,000 and you get even closer to a Gaussian distribution.
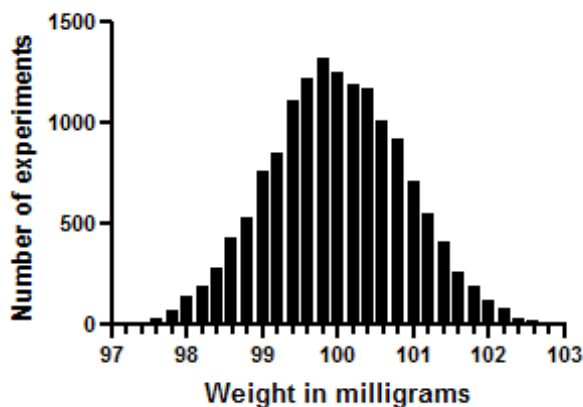
This simulation demonstrates a principle that can also be mathematically proven. Scatter will approximate a Gaussian distribution if your experimental scatter has numerous sources that are additive and of nearly equal weight, and the sample size is large.

The Gaussian distribution is a mathematical ideal. Few biological distributions, if any, really follow the Gaussian distribution. The Gaussian distribution extends from negative infinity to positive infinity. If the weights in the example above really were to follow a Gaussian

distribution, there would be some chance (albeit very small) that the weight is negative. Since weights can't be negative, the distribution cannot be exactly Gaussian. But it is close enough to Gaussian to make it OK to use statistical methods (like t tests and regression) that assume a Gaussian distribution.

4.2.3   The Central Limit Theorem of statistics

The Gaussian distribution plays a central role in statistics because of a mathematical relationship known as the Central Limit Theorem. To understand this theorem, follow this imaginary experiment:

1. Create a population with a known distribution (which does not have to be Gaussian).

2. Randomly pick many samples of equal size from that population. Tabulate the means of these samples.

3. Draw a histogram of the frequency distribution of the means.

The central limit theorem says that if your samples are large enough, the distribution of means will follow a Gaussian distribution even if the population is not Gaussian. Since most statistical tests (such as the t test and ANOVA) are concerned only with differences between means, the Central Limit Theorem lets these tests work well even when the populations are not Gaussian. For this to be valid, the samples have to be reasonably large. How large is that? It depends on how far the population distribution differs from a Gaussian distribution. Assuming the population doesn't have a really unusual distribution, a sample size of 10 or so is generally enough to invoke the Central Limit Theorem.

To learn more about why the ideal Gaussian distribution is so useful, read about the Central Limit Theorem in any statistics text.

4.3   Standard Deviation and Standard Error of the Mean

Rather than show raw data, many scientists

present results as mean plus or minus the

standard deviation (SD) or standard error (SEM).
This section helps you understand what these
values mean.

4.3.1   Key concepts: SD

What is the SD?

The standard deviation (SD) quantifies variability or scatter, and it is
expressed in the same units as your data.

How to interpret the SD when the data are Gaussian

If the data are sampled from a Gaussian distribution, then you expect
68% of the values to lie within one SD of the mean and 95% to lie within
two SD of the mean. This figure shows 250 values sampled from a
Gaussian distribution. The shaded area covers plus or minus one SD from
the mean, and includes about two-thirds of the values. The dotted lines
are drawn at the mean plus or minus two standard deviations, and about
95% of the values lie within those limits.
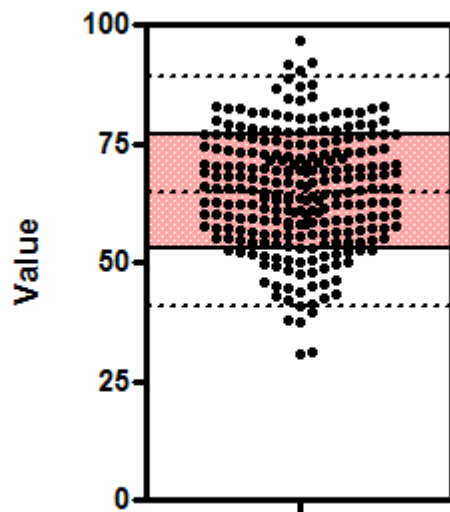
The graph that follows shows the relationship between the standard deviation and a Gaussian distribution. The area under a probability distribution represents the entire population, so the area under a portion of a probability distribution represents a fraction of the population. In the graph on the left, the green (shaded) portion extends from one SD below the mean to one SD above the mean. The green area is about 68% of the total area, so a bit more than two thirds of the values are in the interval mean plus or minus one SD. The graph on the right shows that about 95% of values lie within two standard deviations of the mean.
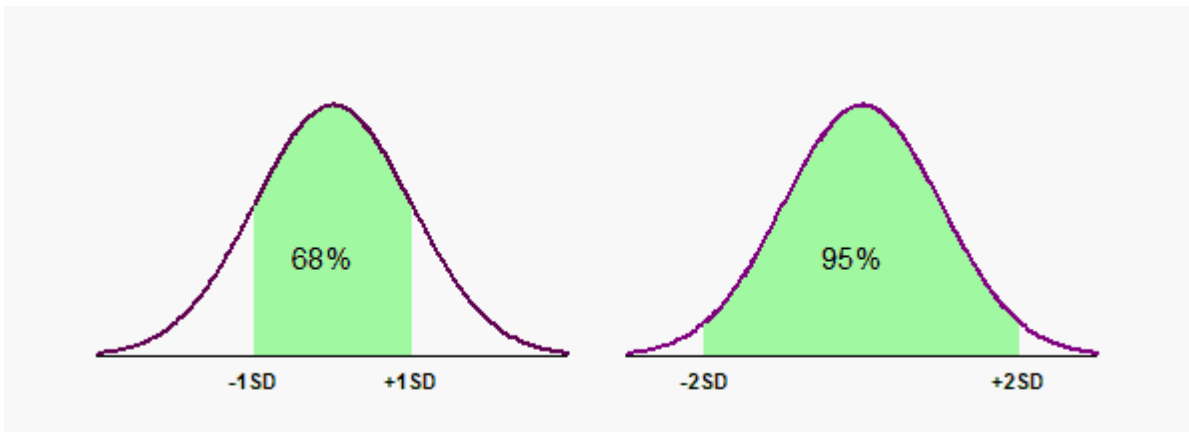
How to interpret the SD when the data are not Gaussian

The figure below shows three sets of data, all with exactly the same mean and SD. The sample on the left is approximately Gaussian. The other two samples are far from Gaussian yet have precisely the same mean (100) and standard deviation (35).
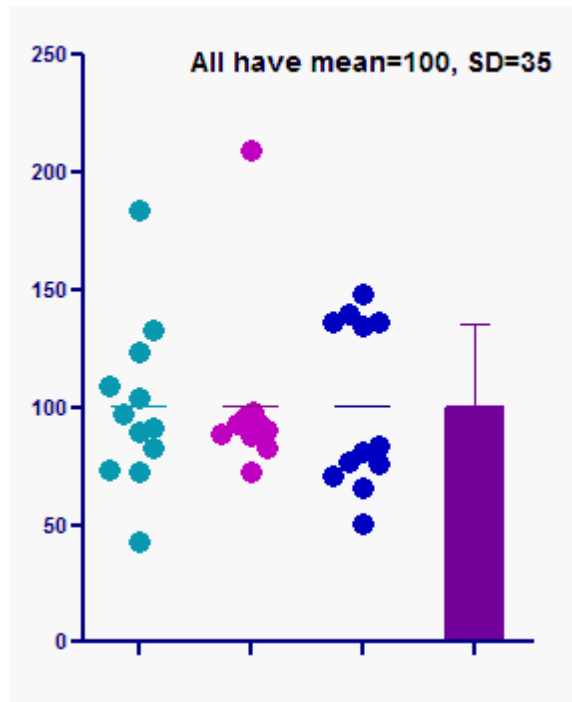
This graph points out that interpreting the mean and SD the usual way can be misleading if you assume the data are Gaussian, but that assumption isn't true.

The SD can still be interpreted without assuming a Gaussian distribution. The Chebyshev theorem states that even if data are not sampled from a Gaussian distribution:

- At least 75% of the values must lie within two standard deviations of the mean

- At least 89% of the values must be within three standard deviations

How to report standard deviations

Many people report a mean and a standard deviation something like this: "115±10 mmHg", with a footnote or statement in the Methods section defining the second value as a standard deviation.

Some (1,2) say that because the standard deviation is a single value that quantifies scatter, it should not follow a plus/minus symbol but instead should appear like this: "115 mmHg (SD 10)".

1. Curran-Everett D, Benos D. Guidelines for reporting statistics in journals published by the American Physiological Society. AJP - Gastrointestinal and Liver Physiology. 2004 Aug 1;287(2):G307.

2. Ludbrook J. The presentation of statistics in Clinical and Experimental Pharmacology and Physiology. Clin Exp Pharmacol Physiol. 2008 Oct 1;35(10):1271–4; authorreply1274.

4.3.2   Computing the SD

How is the SD calculated?

1. Compute the square of the difference between each value and the sample mean.

2. Add those values up.

3. Divide the sum by N-1. This is called the variance.

4. Take the square root to obtain the Standard Deviation.

Why n-1?

Why divide by n-1 rather than N in the third step above? In step 1, you compute the difference between each value and the mean of those values. You don't know the true mean of the population; all you know is the mean of your sample. Except for the rare cases where the sample mean happens to equal the population mean, the data will be closer to the sample mean than it will be to the true population mean. So the value you compute in step 2 will probably be a bit smaller (and can't be larger) than what it would be if you used the true population mean in step 1. To make up for this, we divide by n-1 rather than n.

But why n-1? If you knew the sample mean, and all but one of the values, you could calculate what that last value must be. Statisticians say there are n-1 degrees of freedom.

More about n vs. n-1.

But I've seen equations with n, not n-1, in the denominator!

The n-1 equation is used in the common situation where you are analyzing a sample of data and wish to make more general conclusions. The SD computed this way (with N-1 in the denominator) is your best guess for the value of the SD in the overall population.

If you simply want to quantify the variation in a particular set of data, and don't plan to extrapolate to make wider conclusions, compute the SD using N in the denominator. The resulting SD is the SD of those particular values, but will most likely underestimate the SD of the population from which those points were drawn.

The goal of science is always to generalize, so the equation with n in the denominator should not be used when analyzing scientific data. The only example I can think of where it might make sense to use n (not n-1) in the denominator is in quantifying the variation among exam scores. But much better would be to show a scatterplot of every score, or a frequency distribution histogram.

Prism always computes the SD using n-1.

How many values do you need to compute a SD?

The SD quantifies scatter, so clearly you need more than one value! Is two values enough? Many people believe it is not possible to compute a SD from only two values. But that is wrong. The equation that calculates the SD works just fine when you have only duplicate (n=2) data.

Are the results valid? There is no mathematical reason to think otherwise, but I answered the question with simulations. I simulated ten thousand data sets with n=2 and each data point randomly chosen from a Gaussian distribution. Since all statistical tests are actually based on the variance (the square of the SD), I compared the variance computed from the duplicate values with the true variance. The average of the 10,000 variances of simulated data was within 1% of the true variance from which the data were simulated. This means that the SD computed from duplicate data is a valid assessment of the scatter in your data. It is equally likely to be too high or too low, but is likely to be pretty far from the true SD[47].

Calculating the SD with Excel

Excel can compute the SD from a range of values using the STDEV() function. For example, if you want to know the standard deviation of the values in cells B1 through B10, use this formula in Excel:

    =STDEV(B1:B10)

That function computes the SD using n-1 in the denominator. If you want to compute the SD using N in the denominator (see above) use Excel's STDEVP() function.

Is the SD the same as the SEM?

No![51]

4.3.3   How accurately does a SD quantify scatter?

The SD of a sample is not the same as the SD of the population

It is straightforward to calculate the standard deviation from a sample of values. But how accurate is the standard deviation? Just by chance you may have happened to obtain data that are closely bunched together, making the SD low. Or you may have happened to obtain data that are far more scattered than the overall population, making the SD high. The SD of your sample may not equal, or even be close to, the SD of the population.

The 95% CI of the SD

You can express the precision of any computed value as a 95% confidence interval (CI). It's not done often, but it is certainly possible to compute a CI for a SD. We'll discuss confidence intervals more in the next section[59] which explains the CI of a mean. Here we are discussing the CI of a SD, which is quite different.

Interpreting the CI of the SD is straightforward. You must assume that your data were randomly and independently[32] sampled from a Gaussian[37] distribution. You compute the SD and its CI from that one sample, and use it to make an inference about the SD of the entire population. You can be 95% sure that the CI of the SD contains the true overall standard deviation of the population.

How wide is the CI of the SD? Of course the answer depends on sample size (N), as shown in the table below.

N	95% CI of SD
2	0.45*SD to 31.9*SD
3	0.52*SD to 6.29*SD
5	0.60*SD to 2.87*SD
10	0.69*SD to 1.83*SD
25	0.78*SD to 1.39*SD
50	0.84*SD to 1.25*SD
100	0.88*SD to 1.16*SD
500	0.94*SD to 1.07*SD
1000	0.96*SD to 1.05*SD
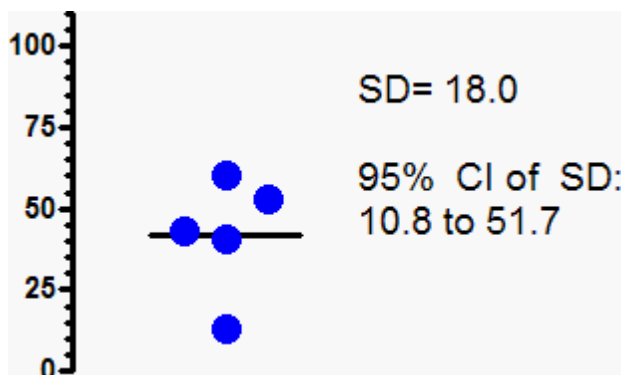
The standard deviation computed from the five values shown in the graph above is 18.0. But the true standard deviation of the population from which the values were sampled might be quite different. Since N=5, the 95% confidence interval extends from 10.8 (0.60*18.0) to 51.7 (2.87*18.0). When you compute a SD from only five values, the upper 95% confidence limit for the SD is almost five times the lower limit.

Most people are surprised that small samples define the SD so poorly. Random sampling can have a huge impact with small data sets, resulting in a calculated standard deviation quite far from the true population standard deviation.

Note that the confidence intervals are not symmetrical. Why? Since the SD is always a positive number, the lower confidence limit can't be less than zero. This means that the upper confidence interval usually extends further above the sample SD than the lower limit extends below the sample SD. With small samples, this asymmetry is quite noticeable.

If you want to compute these confidence intervals yourself, use these Excel equations (N is sample size; alpha is 0.05 for 95% confidence, 0.01 for 99% confidence, etc.):

Lower limit: `=SD*SQRT((N-1)/CHIINV((alpha/2), N-1))`

Upper limit: `=SD*SQRT((N-1)/CHIINV(1-(alpha/2), N-1))`

4.3.4   Key concepts: SEM

What is the SEM?

The standard error of the mean (SEM) quantifies the precision of the mean. It is a measure of how far your sample mean is likely to be from the true population mean. It is expressed in the same units as the data.

GraphPad Prism uses the abbreviation SEM, but some prefer (insist on) the abbreviation SE (1, 2).

Is the SEM larger or smaller than the SD?

The SEM is always smaller than the SD. With large samples, the SEM is much smaller than the SD.

How do you interpret the SEM?

Although scientists often present data as mean and SEM, interpreting what the SEM means is not straightforward. It is much easier to interpret the 95% confidence interval, which is calculated from the SEM.

With large samples (say greater than ten), you can use these rules-of-thumb:

The 67% confidence interval extends approximately one SEM in each direction from the mean.

The 95% confidence interval extends approximately two SEMs from the mean in each direction.

The multipliers are not actually 1.0 and 2.0, but rather are values that come from the t distribution and depend on sample size. With small samples, and certainly when N is less than ten, those rules of thumb are not very accurate.

Is the SEM the same as the SD?

No! [51]

Are all standard errors the standard error of a mean?

No. Statistical computations can compute a standard error for almost any parameter computed from a sample of data. Prism can compute the standard error of a slope in linear regression, and any parameter (i.e. rate constants) from nonlinear regression. The abbreviation SE applies to any standard error, including the standard error of the mean in many journals. The abbreviation SEM always applies to the standard error of the mean.

References

1. Curran-Everett D, Benos D. Guidelines for reporting statistics in journals published by the American Physiological Society. AJP - Gastrointestinal and Liver Physiology. 2004 Aug 1;287(2):G307.

2. Ludbrook J. The presentation of statistics in Clinical and Experimental Pharmacology and Physiology. Clin Exp Pharmacol Physiol. 2008 Oct 1;35(10):1271–4; authorreply1274.

4.3.5 Computing the SEM

How is the SEM calculated?

The SEM is calculated by dividing the SD by the square root of N. This relationship is worth remembering, as it can help you interpret published data.

If the SEM is presented, but you want to know the SD, multiply the SEM by the square root of N.

Calculating the SEM with Excel

Excel does not have a function to compute the standard error of a mean. It is easy enough to compute the SEM from the SD, using this formula.

=STDEV()/SQRT(COUNT())

For example, if you want to compute the SEM of values in cells B1 through B10, use this formula:

=STDEV(B1:B10)/SQRT(COUNT(B1:B10))

The COUNT() function counts the number of numbers in the range. If you are not worried about missing values, you can just enter N directly. In that case, the formula becomes:

=STDEV(B1:B10)/SQRT(10)

4.3.6 The SD and SEM are not the same

It is easy to be confused about the difference between the standard deviation (SD) and the standard error of the mean (SEM). Here are the key differences:

- The SD quantifies scatter — how much the values vary from one another.

- The SEM quantifies how precisely you know the true mean of the population. It takes into account both the value of the SD and the sample size.

- Both SD and SEM are in the same units -- the units of the data.

- The SEM, by definition, is always smaller than the SD.

- The SEM gets smaller as your samples get larger. This makes sense, because the mean of a large sample is likely to be closer to the true population mean than is the mean of a small sample. With a huge sample, you'll know the value of the mean with a lot of precision even if the data are very scattered.

- The SD does not change predictably as you acquire more data. The SD you compute from a sample is the best possible estimate of the SD of the overall population. As you collect more data, you'll assess the SD of the population with more precision. But you can't predict whether the SD from a larger sample will be bigger or smaller than the SD from a small sample. (This is not strictly true. It is the variance -- the SD squared -- that doesn't change predictably, but the change in SD is trivial and much much smaller than the change in the SEM.)

Note that standard errors can be computed for almost any parameter you compute from data, not just the mean. The phrase "the standard error" is a bit ambiguous. The points above refer only to the standard error of the mean.

4.3.7   Advice: When to plot SD vs. SEM

If you create a graph with error bars, or create a table with plus/minus values, you need to decide whether to show the SD, the SEM, or something else.

Often, there are better alternatives to graphing the mean with SD or SEM.

If you want to show the variation in your data

If each value represents a different individual, you probably want to show the variation among values. Even if each value represents a different lab experiment, it often makes sense to show the variation.

If you are plotting a column graph fewer than 100 or so values per data set, create a scatter plot that shows every value. What better way to show the variation among values than to show every value? If your data set has more than 100 or so values, a scatter plot becomes messy. Alternatives are to show a box-and-whiskers plot, a frequency distribution (histogram), or a cumulative frequency distribution.

If you are plotting XY data, especially with multiple treatment groups, plotting every replicate can lead to a messy graph. It can be a good first step, so you see your data fully. But then change to mean and error bar when presenting the data.

If you want to plot mean and error bar, the SD quantifies variability among replicates. So does a graph of median with interquartile range or full range. When plotting a graph with error bars, be sure to explain how the error bars were computed in the figure itself or in its legend.

If you want to show how precisely you have determined the mean

If your goal is to compare means with a t test or ANOVA, or to show how closely our data come to the predictions of a model, you may be more interested in showing how precisely the data define the mean than in showing the variability. In this case, the best approach is to plot the 95% confidence interval of the mean (or perhaps a 90% or 99% confidence interval).

What about the standard error of the mean (SEM)? Graphing the mean with an SEM error bars is a commonly used method to show how well you know the mean, The only advantage of SEM error bars are that they are shorter, but SEM error bars are harder to interpret than a confidence interval. Nonetheless, SEM error bars are the standard in many fields.

Whatever error bars you choose to show, be sure to state your choice. Noticing whether or not the error bars overlap tells you less than you might guess.

If you want to create persuasive propaganda

If your goal is to emphasize small and unimportant differences in your data, show your error bars as SEM, and hope that your readers think they are SD

If our goal is to cover-up large differences, show the error bars as the standard deviations for the groups, and hope that your readers think they are a standard errors.

This approach was advocated by Steve Simon in his excellent weblog. Of course he meant it as a joke. If you don't understand the joke, review the differences between SD and SEM.

4.3.8 Alternatives to showing the SD or SEM

If you want to show the variation in your data

If each value represents a different individual, you probably want to show the variation among values. Even if each value represents a different lab experiment, it often makes sense to show the variation.

With fewer than 100 or so values, create a scatter plot that shows every value. What better way to show the variation among values than to show every value? If your data set has more than 100 or so values, a scatter plot becomes messy. Alternatives are to show a box-and-whiskers plot, a frequency distribution (histogram), or a cumulative frequency distribution.

What about plotting mean and SD? The SD does quantify variability, so this is indeed one way to graph variability. But a SD is only one value, so is a pretty limited way to show variation. A graph showing mean and SD error bar is less informative than any of the other alternatives, but takes no less space and is no easier to interpret. I see no advantage to plotting a mean and SD rather than a column scatter graph, box-and-wiskers plot, or a frequency distribution.

Of course, if you do decide to show SD error bars, be sure to say so in the figure legend so no one will think it is a SEM.

If you want to show how precisely you have determined the mean

If your goal is to compare means with a t test or ANOVA, or to show how closely our data come to the predictions of a model, you may be more interested in showing how precisely the data define the mean than in showing the variability. In this case, the best approach is to plot the 95% confidence interval of the mean (or perhaps a 90% or 99% confidence interval).

What about the standard error of the mean (SEM)? Graphing the mean with an SEM error bars is a commonly used method to show how well you know the mean, The only advantage of SEM error bars are that they are shorter, but SEM error bars are harder to interpret than a confidence interval.

Whatever error bars you choose to show, be sure to state your choice. Noticing whether or not the error bars overlap tells you less than you might guess.[286]

4.4   The lognormal distribution and geometric mean and SD

Lognormal distributions are very common in
biology but very rare in statistics books.

4.4.1   The lognormal distribution

Key facts about the lognormal distribution

- A Gaussian distribution emerges when variation is caused by <u>multiple sources of scatter which add together</u>[38] . In contrast, a lognormal distribution emerges when variation is caused by multiple sources of scatter which are *multiplicative*.

- All values in a lognormal distribution are positive. Negative values and zeroes are not possible in a lognormal distribution.

- Lognormal distributions are common in biology.

- The lognormal distribution is asymmetrical. Lots of values are very similar, while a small fraction of the values are much larger. You can see this in the left panel of the graph below.
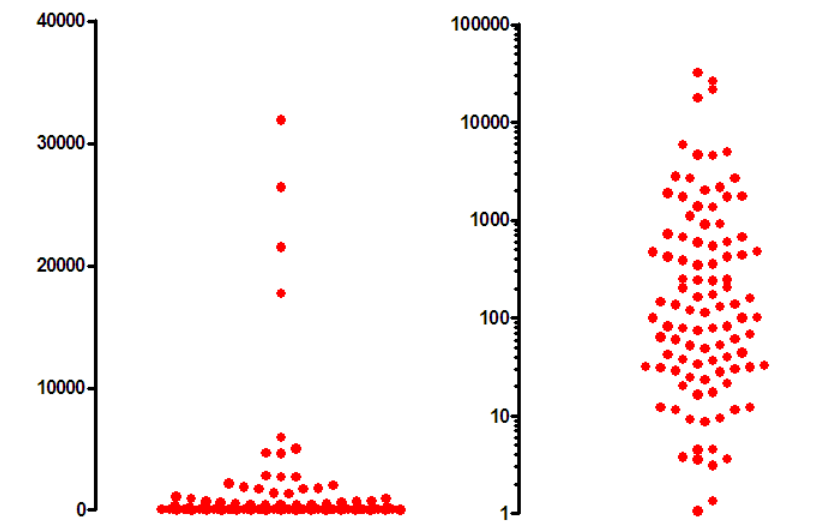
- When plotted on a logarithmic axis, as shown on the right panel of the graph above, the distribution is symmetrical.

- If you are not aware of lognormal distributions, you might be tempted to remove the highest four values in the example above as outliers, since they look like they are not part of the overall distribution. If you run the values through an outlier detection algorithm that assumes sampling from a Gaussian distribution, outliers will probably be identified (the highest four values in the example above.

- If you try to compare means with a t test or ANOVA, you are likely to find that the P value is high and the confidence intervals are wide. T tests and ANOVA assume that the values were sampled from a Gaussian distribution. You will lose power if you try to use those methods to compare means of data set sampled from a lognormal distribution.

- The logarithm of all the values from a lognormal distribution is Gaussian.

How to cope with lognormal distributions

Analyzing data from a lognormal distribution is easy. Simply transform the data by taking the logarithm of each value. These logarithms are expected to have a Gaussian distribution, so can be analyzed by t tests, ANOVA, etc.

4.4.2   The geometric mean and geometric SD factor

Key facts about the geometric mean

- Prism computes the geometric mean by computing the logarithm of all values, then calculating the mean of the logarithms, and finally taking the antilog.

- Prism uses base 10 (common) logarithms, and then takes ten to the power of the mean of the logarithms to get the geometric mean. Some programs use natural logs and then use the exponential function to convert back.

- Using logarithms and antilogarithms is equivalent to multiplying all the values together and taking that product to the 1/n power, where n is the number of values. You'll see this definition in some books.

- The geometric mean cannot be computed if any values are zero or negative.

- The geometric mean is in the same units as the data and the arithmetic mean.

- The geometric mean is never larger than the arithmetic mean.

- If the data are sampled from a <u>lognormal distribution</u>[55] , the geometric mean is probably the best way to express the center of the distribution.

Geometric SD factor

Prism (introduced in Prism 7) reports a Geometric SD factor when you request a geometric mean. It also can plot the geometric mean and its geometric SD factor on some graphs.

Key facts about the geometric SD factor:

- The term geometric SD is not commonly used. It was introduced by Kirkwood (1).

- How the geometric SD is computed: First, transform all the values to logarithms, compute the sample SD of those log values, and then take the antilogarithm of that SD.

- The geometric SD factor has no units. It is a unitless ratio.

- You can't add the geometric SD to the geometric mean (or any other value), and makes equally no sense to ever subtract the geometric SD from the geometric mean. The geometric SD is a value you always multiply or divide by. This is very different than a ordinary SD which has the same units as the data, so can be added to or subtracted from the mean.

- The range from (the geometric mean divided by the geometric SD factor) to (the geometric mean multiplied by the geometric SD factor) will contain about two thirds of the values if the data are sampled from

a lognormal distribution. Similarly, the range from (the mean minus the SD) to (the mean plus the SD) will contain about two thirds of the values when data are sampled from a Gaussian distribution.

- It is rare to see publications show the geometric SD. It is common to see a result reported as "The mean is 3.2 ± 1.2 (SD)". However it is currently rare to report that the geometric mean is 4.3 *÷ 1.14. Instead of a "plus or minus" symbol, I entered a "times or divided by" symbols.

- While it seems odd to express an error as "multiplied or divided by", it is really no stranger than "plus or minus".

Example
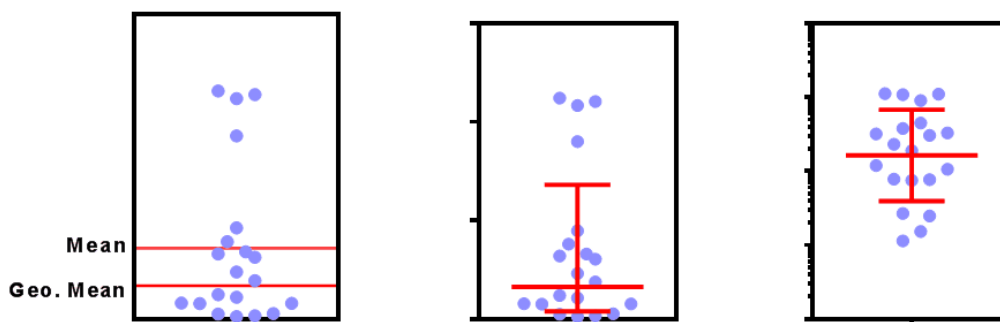
The graph above plots 20 values sampled from a lognormal distribution. The graph on the left shows you that the mean and geometrical mean are very different. The middle graph plots the geometric mean with error bars computed as the geometric mean times or divided by the geometric SD factor. The graph shows the same thing with a logarithmic Y axis. Now the distribution looks symmetrical, and the error bars seem to extend the same distance in each direction. But the ends of the error bars are at the same Y value in the middle and right graph. The right graph uses a logarithmic axis.

Reference

1. Kirkwood, TBL (1979). "Geometric means and measures of dispersion". Biometrics 35: 908–9.

4.5   Confidence intervals

How sure are you? That is a fundamental question when analyzing data, and confidence intervals are the way to answer it.

4.5.1   Key concepts: Confidence interval of a mean

What is the confidence interval of a mean?

The confidence interval (CI) of a mean tells you how precisely you have determined the mean.

For example, you measure weight in a small sample (N=5), and compute the mean. That mean is very unlikely to equal the population mean. The size of the likely discrepancy depends on the size and variability of the sample.

If your sample is small and variable, the sample mean is likely to be quite far from the population mean. If your sample is large and has little scatter, the sample mean will probably be very close to the population mean. Statistical calculations combine sample size and variability (standard deviation) to generate a CI for the population mean. As its name suggests, the CI is a range of values.

What assumptions are made in interpreting a CI of a mean?

To interpret the confidence interval of the mean, you must assume that all the values were independently[32] and randomly sampled from a population whose values are distributed according to a Gaussian[37] distribution. If you accept those assumptions, there is a 95% chance that the 95% CI contains the true population mean. In other words, if you generate many 95% CIs from many samples, you can expect the 95% CI to include the true population mean in 95% of the cases, and not to include the population mean value in the other 5%.

How is it possible that the CI of a mean does not include the true mean

The upper panel below shows ten sets of data (N=5), randomly drawn from a Gaussian distribution with a mean of 100 and a standard deviation of 35. The lower panel shows the 95% CI of the mean for each sample.
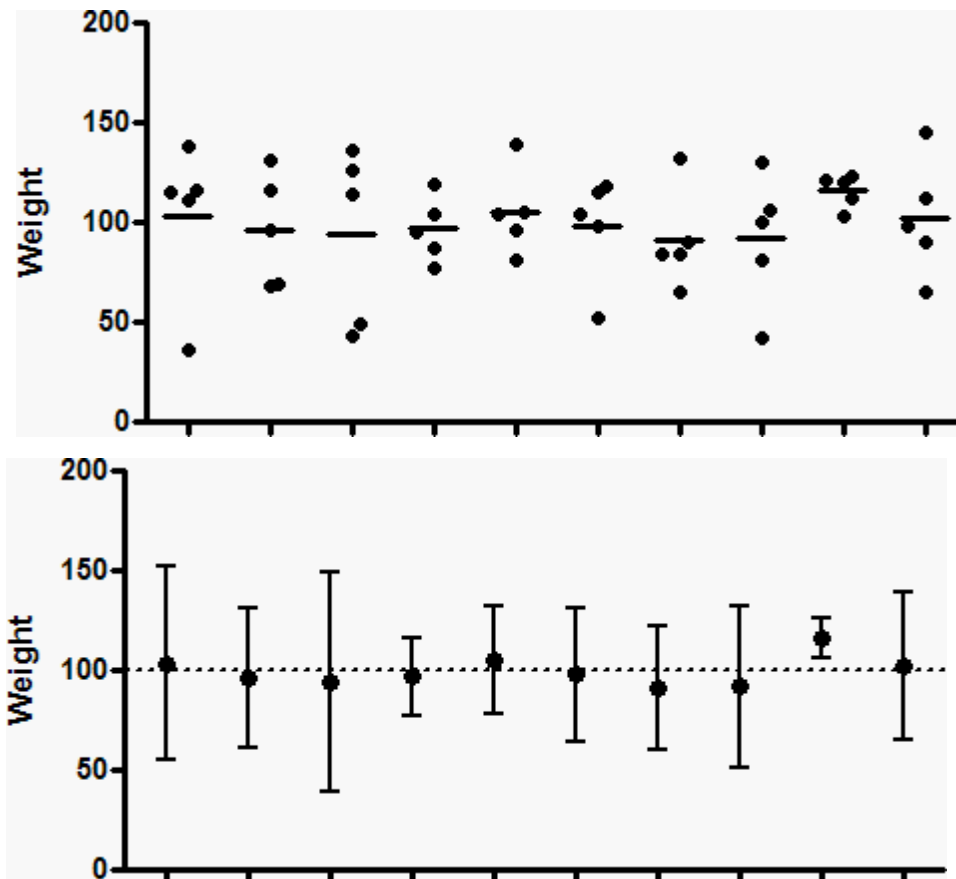
Because these are simulated data, we know the exact value of the true population mean (100), so can ask whether or not each confidence interval includes that true population mean. In the data set second from the right in the graphs above, the 95% confidence interval does not include the true mean of 100 (dotted line).

When analyzing data, you don't know the population mean, so can't know whether a particular confidence interval contains the true population mean or not. All you know is that there is a 95% chance that the confidence interval includes the population mean, and a 5% chance that it does not.

How is the confidence interval of a mean computed?

The confidence interval of a mean is centered on the sample mean, and extends symmetrically in both directions. That distance equals the SE of the mean times a constant from the t distribution. The value of that constant depends only on sample size (N) as shown below.

N	Multiplier
2	12.706
3	4.303
5	2.776
10	2.262
25	2.064
50	2.010
100	1.984
500	1.965
N	=TINV(0.05,N-1)

The samples shown in the graph above had five values. So the lower confidence limit from one of those samples is computed as the mean minus 2.776 times the SEM, and the upper confidence limit is computed as the mean plus 2.776 times the SEM.

The last line in the table above shows you the equation to use to compute the multiplier in Excel. The newer syntax is =T.INV.2T(0.005, N-1).

A common rule-of-thumb is that the 95% confidence interval is computed from the mean plus or minus two SEMs. With large samples, that rule is very accurate. With small samples, the CI of a mean is much wider than suggested by that rule-of-thumb.

4.5.2   Interpreting a confidence interval of a mean

A confidence interval does not quantify variability

A 95% confidence interval is a range of values that you can be 95% certain contains the true mean of the population. This is not the same as a range that contains 95% of the values. The graph below emphasizes this distinction.
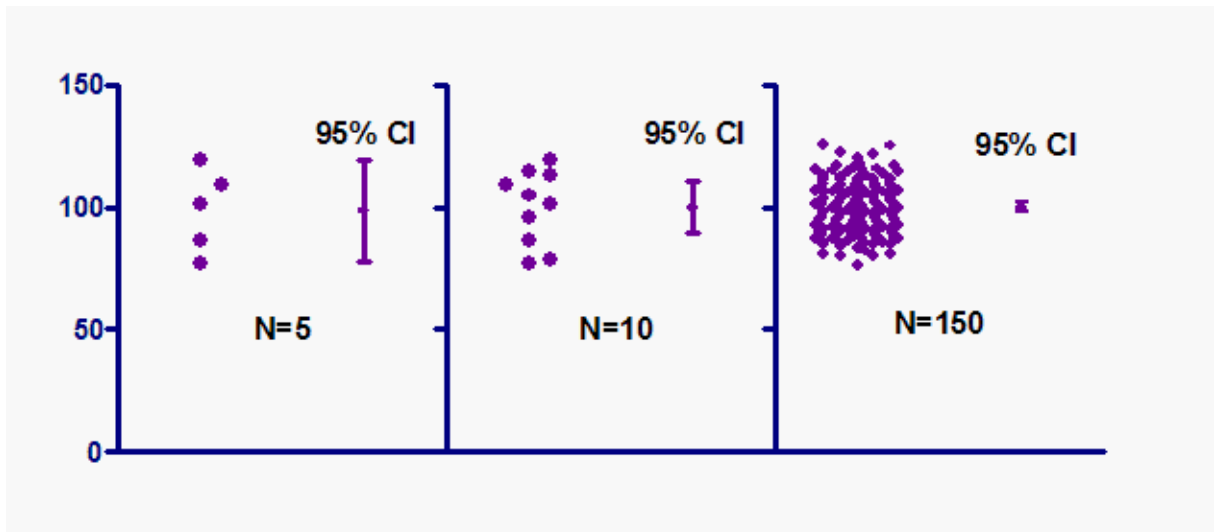
The graph shows three samples (of different size) all sampled from the same population.

With the small sample on the left, the 95% confidence interval is similar to the range of the data. But only a tiny fraction of the values in the large sample on the right lie within the confidence interval. This makes sense. The 95% confidence interval defines a range of values that you can be 95% certain contains the population mean. With large samples, you know that mean with much more precision than you do with a small sample, so the confidence interval is quite narrow when computed from a large sample.

> Don't view a confidence interval and misinterpret it as the range that contains 95% of the values.

A 95% chance of what?

It is correct to say that there is a 95% chance that the confidence interval you calculated contains the true population mean. It is not quite correct to say that there is a 95% chance that the population mean lies within the interval.

What's the difference?

The population mean has one value. You don't know what it is (unless you are doing simulations) but it has one value. If you repeated the experiment, that value wouldn't change (and you still wouldn't know what it is). Therefore it isn't strictly correct to ask about the probability that the population mean lies within a certain range.

In contrast, the confidence interval you compute depends on the data you happened to collect. If you repeated the experiment, your confidence interval would almost certainly be different. So it is OK to ask about the probability that the interval contains the population mean.

It is not quite correct to ask about the probability that the population mean is within the interval. It either is in the interval or it isn't. There is no chance about it. What you can say is that if you perform this kind of experiment many times, the confidence intervals would not all be the same, you would expect 95% of them to contain the population mean, you would expect 5% of the confidence intervals to not include the population mean, and you would never know whether the interval from a particular experiment contained the population mean or not.

Nothing special about 95%

While confidence intervals are usually expressed with 95% confidence, this is just a tradition. Confidence intervals can be computed for any desired degree of confidence.

People are often surprised to learn that 99% confidence intervals are wider than 95% intervals, and 90% intervals are narrower. But this makes perfect sense. If you want more confidence that an interval contains the true parameter, then the intervals will be wider. If you want to be 100.000% sure that an interval contains the true population, it has to contain every possible value so be very wide. If you are willing to be only 50% sure that an interval contains the true value, then it can be much narrower.

4.5.3 Other confidence intervals

The concept of confidence intervals is general. You can calculate the 95% CI for almost any value you compute when you analyze data. We've already discussed the CI of a SD[47]. Other confidence intervals computed by Prism include:

- The difference between two group means

- A proportion

- The ratio of two proportions

- The best-fit slope of linear regression

- The best-fit value of an EC50 determined by nonlinear regression

- The ratio of the median survival times of two groups

- The median of a set of values.

The concept is the same for all these cases. You collected data from a small sample and analyzed the data. The values you compute are 100% correct for that sample, but are affected by random scatter. A confidence interval tells you how precisely you have determined that value. Given certain assumptions (which we list with each analysis later in this book), you can be 95% sure that the 95% CI contains the true (population) value.

The fundamental idea of statistics is to analyze a sample of data, and make quantitative inferences about the population from which the data were sampled. Confidence intervals are the most straightforward way to do this.

4.5.4 Advice: Emphasize confidence intervals over P values

Many statistical analyses generate both P values and confidence intervals. Many scientists report the P value and ignore the confidence interval.

I think this is a mistake.

Interpreting P values is tricky[71] . Interpreting confidence intervals, in contrast, is quite simple. You collect some data, do some calculations to

quantify a difference (or ratio, or best-fit value...), and report that value along with a confidence interval to show how precise that value is.

The underlying theory is identical for confidence intervals and P values. So if both are interpreted correctly, the conclusions are identical. But that is a big 'if'', and I agree with the following quote (JM Hoenig and DM Heisey, The American Statistician, 55: 1-6, 2001):

> "... imperfectly understood confidence intervals are more useful and less dangerous than incorrectly understood P values and hypothesis tests."

4.5.5   One sided confidence intervals

Typically, confidence intervals are expressed as a two-sided range. You might state, for example, with 95% confidence, that the true value of a parameter such as mean, EC50, relative risk, difference, etc., lies in a range between two values. We call this interval "two sided" because it is bounded by both lower and upper confidence limits.

In some circumstances, it can make more sense to express the confidence interval in only one direction – to either the lower or upper confidence limit. This can best be illustrated by following an example.

A recent study was performed to evaluate the effectiveness of a new drug in the eradication of Heliobacter pylori infection, and to determine whether or not it was inferior to the standard drug. (This example was adapted from one presented in reference 1). The eradication rate for the new drug was 86.5% (109/126) compared with 85.3% (110/129) for patients treated with the standard therapy.

In this study, the difference between the eradication rates of the two treatments was 1.2%. The 95% confidence interval extends at the lower limit for the new drug from an eradication rate of 7.3% worse than standard drug, to the upper limit with an eradication rate of 9.7% better.

If we assume that the subjects of the study are representative of a larger population, this means there is a 95% chance that this range of values includes the true difference of the eradication rates of the two drugs. Splitting the remaining 5%, there is an additional 2.5% chance that the new treatment increases the eradication rate by more than 9.7%, and a 2.5% chance that the new treatment decreases the eradication rate by more than 7.3%.

In this case, our goal is to show that the new drug is not worse than the old one. So we can combine our 95% confidence level with the 2.5% upper limit, and say that there is a 97.5% chance that the eradication rate with the new drug is no more than 7.3% worse than the eradication rate with standard drug.

It is conventional, however, to state confidence intervals with 95%, not 97.5%, confidence. We can easily create a one-sided 95% confidence interval. To do this, we simply compute a 90% two-sided confidence interval instead of 95%.

The 90% CI for difference in eradication rate extends from -5.9% to 8.4%. Since we are less confident that it includes the true value, it doesn't extend as far as 95% interval. We can restate this to say that the 95% confidence interval is greater than -5.9%. Thus, we are 95% sure that the new drug has an eradication rate not more than 5.9% worse than that of the standard drug.

In this example of testing noninferiority, it makes sense to express a one-sided confidence interval as the lower limit only. In other situations, it can make sense to express a one-sided confidence limit as an upper limit only. For example, in toxicology you may care only about the upper confidence limit.

GraphPad Prism does not compute one-sided confidence intervals directly. But, as the example shows, it is easy to create the one-sided intervals yourself. Simply ask Prism to create a 90% confidence interval for the value you care about. If you only care about the lower limit, say that you are 95% sure the true value is higher than that (90%) lower limit. If you only care about the upper limit, say that you are 95% sure the true value is lower than the (90%) upper limit.

Reference

1. S. J. Pocock, The pros and cons of noninferiority trials, Fundamental & Clinical Pharmacology, 17: 483-490 (2003).

4.5.6   Compare confidence intervals, prediction intervals, and tolerance intervals

When you fit a parameter to a model, the accuracy or precision can be expressed as a confidence interval, a prediction interval or a tolerance interval. The three are quite distinct. Prism only reports confidence intervals.

The discussion below explains the three different intervals for the simple case of fitting a mean to a sample of data (assuming sampling from a Gaussian distribution). The same ideas can be applied to intervals for any best-fit parameter determined by regression.

Confidence interval

Confidence intervals tell you about how well you have determined the mean. Assume that the data really are randomly sampled from a Gaussian distribution. If you do this many times, and calculate a confidence interval of the mean from each sample, you'd expect about 95 % of those intervals to include the true value of the population mean. The key point is that the confidence interval tells you about the likely location of the true population parameter.

Prediction interval

Prediction intervals tell you where you can expect to see the next data point sampled. Assume that the data really are randomly sampled from a Gaussian distribution. Collect a sample of data and calculate a prediction interval. Then sample one more value from the population. If you do this many times, you'd expect that next value to lie within that prediction interval in 95% of the samples.The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.

Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter. So a prediction interval is always wider than a confidence interval.

Before moving on to tolerance intervals, let's define that word 'expect' used in defining a prediction interval. It means there is a 50% chance that you'd see the value within the interval in more than 95% of the samples, and a 50% chance that you'd see the value within the interval in less than 95% of the samples.

Tolerance interval

What if you want to be 95% sure that the interval contains 95% of the values? Or 90% sure that the interval contains 99% of the values? Those latter questions are answered by a tolerance interval. To compute, or understand, a tolerance interval you have to specify two different percentages. One expresses how sure you want to be, and the other expresses what fraction of the values the interval will contain. If you set the first value (how sure) to 50%, then a tolerance interval is the same as

a prediction interval. If you set it to a higher value (say 90% or 99%) then the tolerance interval is wider.

4.5.7   Confidence interval of a standard deviation

A confidence interval can be computed for almost any value computed from a sample of data, including the standard deviation.

The SD of a sample is not the same as the SD of the population

It is straightforward to calculate the standard deviation from a sample of values. But how accurate is that standard deviation? Just by chance you may have happened to obtain data that are closely bunched together, making the SD low. Or you may have randomly obtained values that are far more scattered than the overall population, making the SD high. The SD of your sample does not equal, and may be quite far from, the SD of the population.

Confidence intervals are not just for means

Confidence intervals are most often computed for a mean. But the idea of a confidence interval is very general, and you can express the precision of any computed value as a 95% confidence interval (CI). Another example is a confidence interval of a best-fit value from regression, for example a confidence interval of a slope.

The 95% CI of the SD

The sample SD is just a value you compute from a sample of data. It's not done often, but it is certainly possible to compute a CI for a SD. GraphPad Prism does not do this calculation, but a free GraphPad QuickCalc does.

Interpreting the CI of the SD is straightforward. If you assume that your data were randomly and independently sampled from a Gaussian distribution, you can be 95% sure that the CI contains the true population SD.

How wide is the CI of the SD? Of course the answer depends on sample size (n). With small samples, the interval is quite wide as shown in the table below.

n	95% CI of SD
2	0.45*SD to 31.9*SD

3   0.52*SD to 6.29*SD

5   0.60*SD to 2.87*SD

10   0.69*SD to 1.83*SD

25   0.78*SD to 1.39*SD

50   0.84*SD to 1.25*SD

100   0.88*SD to 1.16*SD

500   0.94*SD to 1.07*SD

1000 0.96*SD to 1.05*SD

Example

Data: 23, 31, 25, 30, 27
Mean: 27.2
SD: 3.35

The sample standard deviation computed from the five values is 3.35. But the true standard deviation of the population from which the values were sampled might be quite different. From the n=5 row of the table, the 95% confidence interval extends from 0.60 times the SD to 2.87 times the SD. Thus the 95% confidence interval ranges from 0.60*3.35 to 2.87*3.35, from 2.01 to 9.62. When you compute a SD from only five values, the upper 95% confidence limit for the SD is almost five times the lower limit.

Most people are surprised that small samples define the SD so poorly. Random sampling can have a huge impact with small data sets, resulting in a calculated standard deviation quite far from the true population standard deviation.

Note that the confidence interval is not symmetrical around the computed SD. Why? Since the SD is always a positive number, the lower confidence limit can't be less than zero. This means that the upper confidence interval usually extends further above the sample SD than the lower limit extends below the sample SD. With small samples, this asymmetry is quite noticeable.

Computing the Ci of a SD with Excel

These Excel equations compute the confidence interval of a SD. n is sample size; alpha is 0.05 for 95% confidence, 0.01 for 99% confidence, etc.:

Lower limit: =SD*SQRT((n-1)/CHIINV((alpha/2), n-1))

Upper limit: =SD*SQRT((n-1)/CHIINV(1-(alpha/2), n-1))

These equations come from page 197-198 of Sheskin (reference below).

Reference

David J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Fourth Edition, IBSN:1584888148.

4.6 P Values

Almost every statistical tests generates a P value (or several). Yet many scientists don't really understand what P values are. This section explains the principles, and also the difference between one- and two-tail P values.

4.6.1 What is a P value?

Suppose that you've collected data from two samples of animals treated with different drugs. You've measured an enzyme in each animal's plasma, and the means are different. You want to know whether that difference is due to an effect of the drug – whether the two populations have different means.

Observing different sample means is not enough to persuade you to conclude that the populations have different means. It is possible that the populations have the same mean (i.e., that the drugs have no effect on the enzyme you are measuring) and that the difference you observed between sample means occurred only by chance. There is no way you can

ever be sure if the difference you observed reflects a true difference or if it simply occurred in the course of random sampling. All you can do is calculate probabilities.

The first step is to state the **null hypothesis**, that really the treatment does not affect the outcome you are measuring (so all differences are due to random sampling).

The P value is a probability, with a value ranging from zero to one, that answers this question (which you probably never thought to ask):

> In an experiment of this size, if the populations really have the same mean, what is the probability of observing at least as large a difference between sample means as was, in fact, observed?

4.6.2   The most common misinterpretation of a P value

Many people misunderstand what a P value means. Let's assume that you compared two means and obtained a P value equal to 0.03.

Correct definitions of this P value:

> There is a 3% chance of observing a difference as large as you observed even if the two population means are identical (the null hypothesis is true).

>   or

> Random sampling from identical populations would lead to a difference smaller than you observed in 97% of experiments, and larger than you observed in 3% of experiments.

Wrong:

> ~~There is a 97% chance that the difference you observed reflects a real difference between populations, and a 3% chance that the difference is due to chance.~~

This latter statement is a common mistake. If you have a hard time understanding the difference between the correct and incorrect definitions, read this [Bayesian perspective](#)[87].

4.6.3   More misunderstandings of P values

Kline (1) lists commonly believed fallacies about P values, which I summarize here:

Fallacy: P value is the probability that the result was due to sampling error

The P value is computed assuming the null hypothesis is true. In other words, the P value is computed based on the assumption that the difference was due to sampling error. Therefore the P value cannot tell you the probability that the result is due to sampling error.

Fallacy: The P value Is the probability that the null hypothesis is true

Nope. The P value is computed assuming that the null hypothesis is true, so cannot be the probability that it is true.

Fallacy: 1-P is the probability that the alternative hypothesis is true

If the P value is 0.03, it is very tempting to think: If there is only a 3% probability that my difference would have been caused by random chance, then there must be a 97% probability that it was caused by a real difference. But this is wrong!

What you can say is that if the null hypothesis were true, then 97% of experiments would lead to a difference smaller than the one you observed, and 3% of experiments would lead to a difference as large or larger than the one you observed.

Calculation of a P value is predicated on the assumption that the null hypothesis is correct. P values cannot tell you whether this assumption is correct. P value tells you how rarely you would observe a difference as larger or larger than the one you observed if the null hypothesis were true.

The question that the scientist must answer is whether the result is so unlikely that the null hypothesis should be discarded.

Fallacy: 1-P is the probability that the results will hold up when the experiment is repeated

If the P value is 0.03, it is tempting to think that this means there is a 97% chance of getting 'similar' results on a repeated experiment. Not so.

Fallacy: A high P value proves that the null hypothesis is true.

No. A high P value means that if the null hypothesis were true, it would not be surprising to observe the treatment effect seen in this experiment. But that does not prove the null hypothesis is true.

Fallacy: The P value is the probability of rejecting the null hypothesis

You reject the null hypothesis (and deem the results statistically significant) when a P value from a particular experiment is less than the significance level α, which you (should have) set as part of the experimental design. So if the null hypothesis is true, α is the probability of rejecting the null hypothesis.

The P value and α are not the same. A P value is computed from each comparison, and is a measure of the strength of evidence. The significance level α is set once as part of the experimental design.

1. RB Kline, *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research, 2004,* IBSN:1591471184

4.6.4   How reproducible are P values? (not very)

P values are much less reproducible than you probably would guess.

If the null hypothesis is true, then the distribution of P values is uniform. Half the P values will be less than 0.50, 5% will be less than 0.05, etc.

But what if the null hypothesis is false? Most people would expect the P values to be fairly reproducible in that case. Not so.

To demonstrate this, I ran some simulations using GraphPad Prism. I sampled from two Gaussian populations with the SD equal to 5.0. The two populations means differed by 5.0. Here are four simulated experiments. Only random sampling separates the four experiments. The four P values vary considerably.
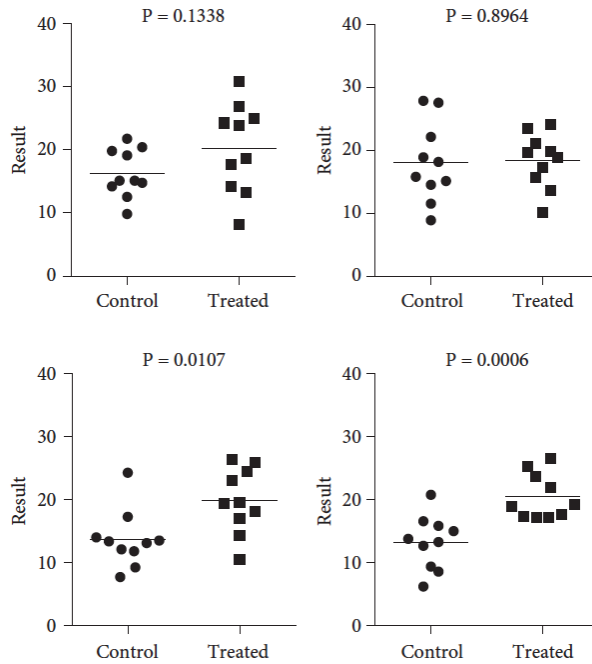
And here is the distribution of P values from 2,500 such simulated experiments. The X axis plots various values of P values. The Y axis shows how frequently that range of values occurred in the 2500 simulated experiments.
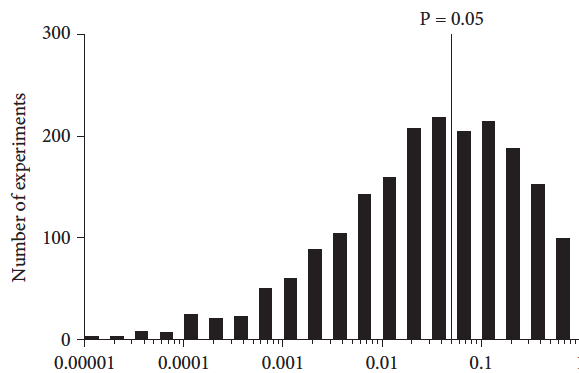
Leaving out the 2.5% highest and lowest P values, the middle 95% of the P values range from 0.0001517 to 0.6869—a range covering more than three orders of magnitude!

In these simulations, the median P value was about 0.05. I also ran simulations with a smaller SD which of course resulted in smaller P values, but the range still covered more than three orders of magnitude.

The idea for these simulations came from Boos and Stefanski (1) who demonstrated via simulations of a number of situations that P values from repeated simulated experiments vary over more than three orders of magnitude.

Cumming (2) asked if the P value from one particular experiment equals 0.05, what the P value from a repeat experiment is likely to be (taking into account only random sampling and assuming there are no glitches with either experiment). Surprisingly, the results don't depend on sample size. Of course, there is a 50% chance that the repeat P value will be greater than 0.05. But it is surprising (I think) that there is a 20% chance that the P value from a repeat experiment will be greater than 0.38 and a 5% chance it will be greater than 0.82.

There are [many reasons to not rely on P values](#)[71] when analyzing data. This page shows that beyond any conceptual reasons for avoiding P values is a very practical reason: P values are not very reproducible.

The figures above are Figure 15.1 from the third edition of [Intuitive Biostatistics](#).

1.   Boos, D. D. & Stefanski, L. A. [P-Value Precision and Reproducibility](#). The American Statistician 65, 213–221 (2011).

2.   Cumming, G. [Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better](#). Perspectives on Psychological Science 3, 286–300 (2008).

4.6.5   One-tail vs. two-tail P values

When comparing two groups, you must distinguish between one- and two-tail P values. Some books refer to one-sided and two-sided P values, which mean the same thing.

What does one-tail mean?

It is easiest to understand the distinction in context. So let's imagine that you are comparing the mean of two groups (with an unpaired t test). Both one- and two-tail P values are based on the same null hypothesis, that two populations really are the same and that an observed discrepancy between sample means is due to chance.

A two-tailed P value answers this question:

> Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far apart as (or further than) you observed in this experiment with either group having the larger mean?

To interpret a one-tail P value, you must predict which group will have the larger mean before collecting any data. The one-tail P value answers this question:

> Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far apart as (or further than) observed in this experiment with the specified group having the larger mean?

If the observed difference went in the direction predicted by the experimental hypothesis, the one-tailed P value is half the two-tailed P value (with most, but not quite all, statistical tests).

When is it appropriate to use a one-tail P value?

A one-tailed test is appropriate when previous data, physical limitations, or common sense tells you that the difference, if any, can only go in one direction. You should only choose a one-tail P value when both of the following are true.

• You predicted which group will have the larger mean (or proportion) before you collected any data. If you only made the "prediction" after seeing the data, don't even think about using a one-tail P value.

• If the other group had ended up with the larger mean – even if it is quite a bit larger – you would have attributed that difference to chance and called the difference 'not statistically significant'.

Here is an example in which you might appropriately choose a one-tailed P value: You are testing whether a new antibiotic impairs renal function, as measured by serum creatinine. Many antibiotics poison kidney cells, resulting in reduced glomerular filtration and increased serum creatinine. As far as I know, no antibiotic is known to decrease serum creatinine, and it is hard to imagine a mechanism by which an antibiotic would increase the glomerular filtration rate. Before collecting any data, you can state that there are two possibilities: Either the drug will not change the mean

serum creatinine of the population, or it will increase the mean serum creatinine in the population. You consider it impossible that the drug will truly decrease mean serum creatinine of the population and plan to attribute any observed decrease to random sampling. Accordingly, it makes sense to calculate a one-tailed P value. In this example, a two-tailed P value tests the null hypothesis that the drug does not alter the creatinine level; a one-tailed P value tests the null hypothesis that the drug does not increase the creatinine level.

The issue in choosing between one- and two-tailed P values is not whether or not you expect a difference to exist. If you already knew whether or not there was a difference, there is no reason to collect the data. Rather, the issue is whether the direction of a difference (if there is one) can only go one way. You should only use a one-tailed P value when you can state with certainty (and before collecting any data) that in the overall populations there either is no difference or there is a difference in a specified direction. If your data end up showing a difference in the "wrong" direction, you should be willing to attribute that difference to random sampling without even considering the notion that the measured difference might reflect a true difference in the overall populations. If a difference in the "wrong" direction would intrigue you (even a little), you should calculate a two-tailed P value.

How Prism reports one-tail P values

When you ask Prism to report a one-tail P value, it assumes the actual difference or effect went in the direction you predicted, so the one-sided P value reported by Prism is always smaller (almost always, exactly half of) the two-tail P value.

If, in fact, the observed difference or effect goes in the opposite direction to what you predicted, the one-sided P value reported by Prism is wrong. The actual one-tail P value will equal 1.0 minus the reported one. For example, if the reported one-tail P value is 0.04 and the actual difference is in the opposite direction to what you predicted, then the actual one-sided P value is 0.96.

What if you didn't predict the direction of the difference or effect before collecting data?

If you didn't predict the direction of the effect before collecting data, you should not be reporting one-sided P values. It is cheating to say "well, I

would have predicted...". If you didn't record the prediction, then you should not use a one-sided P value.

What if there are not two directions to the test?

The concept of one- and two-tail P values only makes sense for hypotheses where there are two directions to the effect, an increase or a decrease. If you are comparing three or more groups (ANOVA), then the concept of one- and two-tail P value makes no sense, and Prism doesn't ask you to make this choice.

How to convert between one- and two-tail P values

If the actual effect went in the direction you predicted:

• The one-tail P value is half the two-tail P value.

• The two-tail P value is twice the one-tail P value (assuming you correctly predicted the direction of the difference).

This rule works perfectly for almost all statistical tests. Some tests (Fisher's test) are not symmetrical, so these rules are only approximate for these tests.

If the actual effect went in the opposite direction to what you predicted:

• The one-tail P value equals 1.0 minus half the two-tail P value.

4.6.6   Advice: Use two-tailed P values

If in doubt, choose a two-tail P value. Why?

• The relationship between P values and confidence intervals is easier to understand with two-tail P values.

• Some tests compare three or more groups, which makes the concept of tails inappropriate (more precisely, the P values have many tails). A two-tail P value is more consistent with the P values reported by these tests.

- Choosing a one-tail P value can pose a dilemma. What would you do if you chose to use a one-tail P value, observed a large difference between means, but the "wrong" group had the larger mean? In other words, the observed difference was in the opposite direction to your experimental hypothesis. To be rigorous, you must conclude that the difference is due to chance, even if the difference is huge. While tempting, it is not fair to switch to a two-tail P value or to reverse the direction of the experimental hypothesis. You avoid this situation by always using two-tail P value.

4.6.7 Advice: How to interpret a small P value

Before you interpret the P value

Before thinking about P values, you should:

- Review the science. If the study was not designed well, then the results probably won't be informative. It doesn't matter what the P value is.

- Review the assumptions of the analysis you chose to make sure you haven't violated any assumptions. We provide an analysis checklist for every analysis that Prism does. If you've violated the assumptions, the P value may not be meaningful.

Interpreting a small P value

A small P value means that the difference (correlation, association,...) you observed would happen rarely due to random sampling. There are three possibilities:

- The null hypothesis of no difference is true, and a rare coincidence has occurred. You may have just happened to get large values in one group and small values in the other, and the difference is entirely due to chance. How likely is this? The answer to that question, surprisingly, is **not** the P value. Rather, the answer depends on the scientific background of the experiment. [87]

- The null hypothesis is false. There truly is a difference (or correlation, or association...) that is large enough to be scientifically interesting.

- The null hypothesis is false. There truly is a difference (or correlation, or association...), but that difference is so small that it is scientifically boring. The difference is real, but trivial.

Deciding between the last two possibilities is a matter of scientific judgment, and no statistical calculations will help you decide.

Using the confidence interval to interpret a small P value

If the P value is less than 0.05, then the 95% confidence interval will not contain zero (when comparing two means). To interpret the confidence interval in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference between means that you consider to be scientifically important or scientifically trivial. This section assumes you are comparing two means with a t test, but it is straightforward to use these same ideas in other contexts.

There are three cases to consider:

- **The confidence interval only contains differences that are trivial.** Although you can be 95% sure that the true difference is not zero, you can also be 95% sure that the true difference between means is tiny and uninteresting. The treatment had an effect, but a small one.

- **The confidence interval only includes differences you would consider to be important.** Since even the low end of the confidence interval represents a difference large enough that you consider it to be scientifically important, you can conclude that there is a difference between treatment means and that the difference is large enough to be scientifically relevant.

- **The confidence interval ranges from a trivial to an important difference.** Since the confidence interval ranges from a difference that you think would be scientifically trivial to one you think would be important, you can't reach a strong conclusion. You can be 95% sure that the true difference is not zero, but you cannot conclude whether the size of that difference is scientifically trivial or important.

4.6.8   Advice: How to interpret a large P value

Before you interpret the P value

Before thinking about P values, you should:

- Assess the science. If the study was not designed well, then the results probably won't be informative. It doesn't matter what the P value is.

- Review the assumptions of the analysis you chose to make sure you haven't violated any assumptions. We provide an analysis checklist for every analysis that Prism does. If you've violated the assumptions, the P value may not be meaningful.

Interpreting a large P value

If the P value is large, the data do not give you any reason to conclude that the overall means differ. Even if the true means were equal, you would not be surprised to find means this far apart just by chance. This is not the same as saying that the true means are the same. You just don't have convincing evidence that they differ.

Using the confidence interval to interpret a large P value

How large could the true difference really be? Because of random variation, the difference between the group means in this experiment is unlikely to be equal to the true difference between population means. There is no way to know what that true difference is. The uncertainty is expressed as a 95% confidence interval. You can be 95% sure that this interval contains the true difference between the two means. When the P value is larger than 0.05, the 95% confidence interval will start with a negative number (representing a decrease) and go up to a positive number (representing an increase).

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference that would be scientifically important or scientifically trivial. There are two cases to consider:

- **The confidence interval ranges from a decrease that you would consider to be trivial to an increase that you also consider to be trivial.** Your conclusions is pretty solid. Either the treatment has no

effect, or its effect is so small that it is considered unimportant. This is an informative negative experiment.

- **One or both ends of the confidence interval include changes you would consider to be scientifically important.** You cannot make a strong conclusion. With 95% confidence you can say that either the difference is zero, not zero but is scientifically trivial, or large enough to be scientifically important. In other words, your data really don't lead to any solid conclusions.

4.6.9   Decimal formatting of P values

Every analysis that reports P values lets you choose the decimal format used to report P values. Each analysis that computes P values gives you these choices:

- APA (American Psychological Association) style, which shows three digits but omits the leading zero (.123). P values less than 0.001 shown as "< .001". All P values less than 0.001 are summarized with three asterisks, with no possibility of four asterisks.

- NEJM (New England Journal of Medicine) style, which shows three digits and includes the leading zero (0.123). P values less than 0.001 shown as "< .001". All P values less than 0.001 are summarized with three asterisks, with no possibility of four asterisks.

- GraphPad style which reports four digits after the decimal point with a leading zero (0.1234). P values less than 0.0001 shown as "< .0001". P values less than 0.001 are summarized with three asterisks, and P values less than 0.0001 are summarized with four asterisks.

- Choose how many digits you want to see after the decimal point, up to 15. P values less than 0.001 are given three asterisks, and P values less than 0.0001 are given four asterisks. You'll choose this option in each analysis individually. But if you do choose this option, the exact appearance of the P value (decimals or scientific notation) will depend on a setting in the Analysis tab of the Prism preferences dialog. This is where you can also set the default number of digits after the decimal.

P value reporting

When reporting "full" P value ("Show N digits after the decimal ..."):

○ Always use scientific format ("1.234e-6")

◉ Use scientific format when the P value is less than: 1e-6   ∨

○ Always use decimal format ("0.0000000000345")

Default # of digits after the decimal: 5 ⬍

See how Prism summarizes P values using asterisks, i[86].e. ***.

4.6.10 How Prism computes P values from statistical ratios

Calcuations built-in to Prism

GraphPad Prism report exact P values with most statistical calculations using these algorithms, adapted from sections 6.2 and 6.4 of Numerical Recipes.

```
PFromF(F_Ratio, DF_Numerator, DF_Denominator) =
BetaI(DF_Denominato /2, DF_Numerator/2, DF_Denominator / (DF_Denominat

PFromT(T_Ratio, DF) = BetaI(DF /2, 1/2, DF / (DF + T_Ratio^2))

PFromZ(Z_Ratio) = PFromT(Z_Ratio, Infinity)

PFromR(R_Value) = PFromT(|R_Value| / SQRT((1 - R_Value^2)/DF) , DF)

PFromChi2(Chi2_Value, DF) = GammaQ(DF / 2, Chi2Value /2)
```

Note that BetaI is the incomplete beta function, and GammaQ is the incomplete gamma function. The variable names should all be self-explanatory.

Calculations with newer versions of Excel

If you want to compute P values using newer (2010 and later) Excel, use these functions:

P value from F   =F.DIST.RT (F, DFn, DFd)

P value from t (two tailed)	=T.DIST.2T(t, df)
P value from Chi Square	=CHISQ.DIST.RT(ChiSquare, DF)
P value from z (two tailed)	=2*(1.0-NORM.S.DIST(z,TRUE))

Calculations with older versions of Excel

If you want to compute P values using older (pre 2010) Excel, use these functions:

P value from F	=FDIST (F, DFn, DFd)
P value from t (two tailed)	=TDIST (t, df, 2) (The third argument, 2, specifies a two-tail P value.)
P value from Chi Square	=CHIDIST (ChiSquare, DF)
P value from z (two tailed)	=2*(1.0-NORMSDIST(z))

Reference

Numerical Recipes 3rd Edition: The Art of Scientific Computing, by William H. Press, Saul A. Teukolsky, William T. Vetterling, IBSN:0521880688.

4.7   Hypothesis testing and statistical significance

"Statistically significant". That phrase is commonly misunderstood. Before analyzing data and presenting statistical results, make sure you

understand what statistical 'significance' means and doesn't mean.

4.7.1 Statistical hypothesis testing

Much of statistical reasoning was developed in the context of quality control where you need a definite yes or no answer from every analysis. Do you accept or reject the batch? The logic used to obtain the answer is called hypothesis testing.

First, define a threshold P value before you do the experiment. Ideally, you should set this value based on the relative consequences of missing a true difference or falsely finding a difference. In practice, the threshold value (called alpha) is almost always set to 0.05 (an arbitrary value that has been widely adopted).

Next, define the null hypothesis. If you are comparing two means, the null hypothesis is that the two populations have the same mean. When analyzing an experiment, the null hypothesis is usually the opposite of the experimental hypothesis. Your experimental hypothesis -- the reason you did the experiment -- is that the treatment changes the mean. The null hypothesis is that two populations have the same mean (or that the treatment has no effect).

Now, perform the appropriate statistical test to compute the P value.

- If the P value is less than the threshold, state that you "reject the null hypothesis" and that the difference is "statistically significant".

- If the P value is greater than the threshold, state that you "do not reject the null hypothesis" and that the difference is "not statistically significant". You cannot conclude that the null hypothesis is true. All you can do is conclude that you don't have sufficient evidence to reject the null hypothesis.

4.7.2 Asterisks

Once you have set a threshold significance level (usually 0.05), every result leads to a conclusion of either "statistically significant" or not "statistically significant". Some statisticians feel very strongly that the only acceptable conclusion is significant or 'not significant', and oppose use of adjectives or asterisks to describe values levels of statistical significance.

Many scientists are not so rigid, and so prefer to use adjectives such as "very significant" or "extremely significant". Prism uses this approach as shown in the tables below. These definitions are not entirely standard. If you report the results in this way, you should define the symbols in your figure legend.

Prior to Prism 7, the scheme below was always used. Now it is used if you choose GP formatting[82] or if you ask for four or more digits after the decimal point.

P value	Wording	Summary
< 0.0001	Extremely significant	****
0.0001 to 0.001	Extremely significant	***
0.001 to 0.01	Very significant	**
0.01 to 0.05	Significant	*
≥ 0.05	Not significant	ns

If you choose APA or NEJM formatting for P values[82], Prism uses this scheme (note the absence of ****).

P value	Wording	Summary
< 0.001	Very significant	***
0.001 to 0.01	Very significant	**
0.01 to 0.05	Significant	*
≥ 0.05	Not significant	ns

Prism stores the P values in double precision (about 12 digits of precision), and uses that value (not the value you see displayed) when it decides how many asterisks to show. So if the P value equals 0.05000001, Prism will display "0.0500" and label that comparison as "ns".

[Decimal formatting of P values.](#)[82]

4.7.3 The false discovery rate and statistical signficance

Interpreting low P values is not straightforward

Imagine that you are screening drugs to see if they lower blood pressure. You use the usual threshold of P<0.05 as defining statistical significance. Based on the amount of scatter you expect to see and the minimum change you would care about, you've chosen the sample size for each experiment to have 80% [power](#)[97] to detect the difference you are looking for with a P value less than 0.05.

If you do get a P value less than 0.05, what is the chance that the drug truly works?

The answer is: It depends on the context of your experiment. Let's start with the scenario where based on the context of the work, you estimate there is a 10% chance that the drug actually has an effect. What happens when you perform 1000 experiments? Given your 10% estimate, the two column totals below are 100 and 900. Since the power is 80%, you expect 80% of truly effective drugs to yield a P value less than 0.05 in your experiment, so the upper left cell is 80. Since you set the definition of statistical significance to 0.05, you expect 5% of ineffective drugs to yield a P value less than 0.05, so the upper right cell is 45.

	Drug really works	Drug really doesn't work	Total
P<0.05, "significant"	80	45	125
P>0.05, "not significant"	20	855	875
Total	100	900	1000

In all, you expect to see 125 experiments that yield a "statistically significant" result, and only in 80 of these does the drug really work. The other 45 experiments yield a "statistically significant" result but are false positives or false discoveries. The false discovery rate (abbreviated FDR) is 45/125 or 36%. Not 5%, but 36%. This is also called the False Positive Rate (FPR).

The table below, from chapter 12 of [Essential Biostatistics][35] , shows the FDR for this and three other scenarios.

	Prior Probability	FDR for P<0.05	FDR for 0.045 < P < 0.050
Comparing randomly assigned groups in a clinical trial prior to treatment	0%	100%	100%
Testing a drug that might possibly work	10%	36%	78%
Testing a drug with 50:50 chance of working	50%	6%	27%
Positive controls	100%	0%	0%

Each row in the table above is for a different scenario defined by a different prior (before collecting data) probability of there being a real effect. The middle column shows the expected FDR (also called FPR) as calculated above. This column answers the question: "If the P value is less than 0.05, what is the chance that there really is no effect and the result is just a matter of random sampling?". Note this answer is not 5%. The FDR is quite different than alpha, the threshold P value used to define statistical significance.

The right column, determined by simulations, asks a slightly different question based on work by Colquhoun(1,2).: "If the P value is just a little bit less than 0.05 (between 0.045 and 0.050), what is the chance that there really is no effect and the result is just a matter of random sampling?" These numbers are much higher. Focus on the third row where the prior probability is 50%. In this case, if the P value is just barely under 0.05 there is a 27% chance that the effect is due to chance. Note: 27%, not 5%! And in a more exploratory situation where you think the prior probability is 10%, the false discovery rate for P values just barely lower than 0.05 is 78%. In this situation, a statistically significant result (defined conventionally) means almost nothing.

Bottom line: You can't interpret statistical significance (or a P value) in a vacuum. Your interpretation depends on the context of the experiment. The false discovery rate can be much higher than the value of alpha

(usually 5%). Interpreting results requires common sense, intuition, and judgment.

Reference

1. Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, 1(3), 140216–140216. http://doi.org/10.1098/rsos.140216

2. Colquhoun, D (2019). The False Positive Risk: A Proposal Concerning What to Do About p-Values. The American Statistician, Volume 73, supplement 1.

4.7.4   A legal analogy: Guilty or not guilty?

The statistical concept of 'significant' vs. 'not significant' can be understood by comparing to the legal concept of 'guilty' vs. 'not guilty'.

In the American legal system (and much of the world) a criminal defendant is presumed innocent until proven guilty. If the evidence proves the defendant guilty beyond a reasonable doubt, the verdict is 'guilty'. Otherwise the verdict is 'not guilty'. In some countries, this verdict is 'not proven', which is a better description. A 'not guilty' verdict does not mean the judge or jury concluded that the defendant is innocent -- it just means that the evidence was not strong enough to persuade the judge or jury that the defendant was guilty.

In statistical hypothesis testing, you start with the null hypothesis (usually that there is no difference between groups). If the evidence produces a small enough P value, you reject that null hypothesis, and conclude that the difference is real. If the P value is higher than your threshold (usually 0.05), you don't reject the null hypothesis. This doesn't mean the evidence convinced you that the treatment had no effect, only that the evidence was not persuasive enough to convince you that there is an effect.

4.7.5   Advice: Avoid the concept of 'statistical significance' when possible

The term "significant" is seductive and easy to misinterpret, because the statistical use of the word has a meaning entirely distinct from its usual meaning. Just because a difference is statistically significant does not

mean that it is biologically or clinically important or interesting. Moreover, a result that is not statistically significant (in the first experiment) may turn out to be very important.

Using the conventional definition with alpha=0.05, a result is said to be statistically significant when a difference that large (or larger) would occur less than 5% of the time if the populations were, in fact, identical.

The entire construct of 'hypothesis testing' leading to a conclusion that a result is or is not 'statistically significant' makes sense in situations where you must make a firm decision based on the results of one P value. While this situation occurs in quality control, it doesn't really happen in other situations. Usually, as with clinical trials, decisions are made based on several kinds of evidence. In basic research, it is rare to make a decision based on one experiment.

If you do not need to make a decision based on one P value, then there is no need to declare a result "statistically significant" or not. Simply report the P value as a number, without using the term 'statistically significant'. Better, simply report the confidence interval, without a P value.

4.7.6   Advice: Avoid the word "significant"

The word "significant" has two meanings in research:

- A P value is less than a preset threshold (often 0.05) so reject the null hypothesis and state the results are deemed statistically significant.

- A difference (or ratio or correlation...) is large enough that you think the result has biological (or scientific or practical) relevance.

To avoid this ambiguity, don't use the S word. If the P value is less than a threshold, say so. If an observed effect is large enough to be relevant, say so. The word "significant" is never needed.

4.7.7   Advice: Don't P-Hack

Statistical results can only be interpreted at face value when every choice in data analysis was performed exactly as planned and documented as

part of the experimental design. This rule is commonly broken in some research fileds. Instead, analyses are often done as shown below:
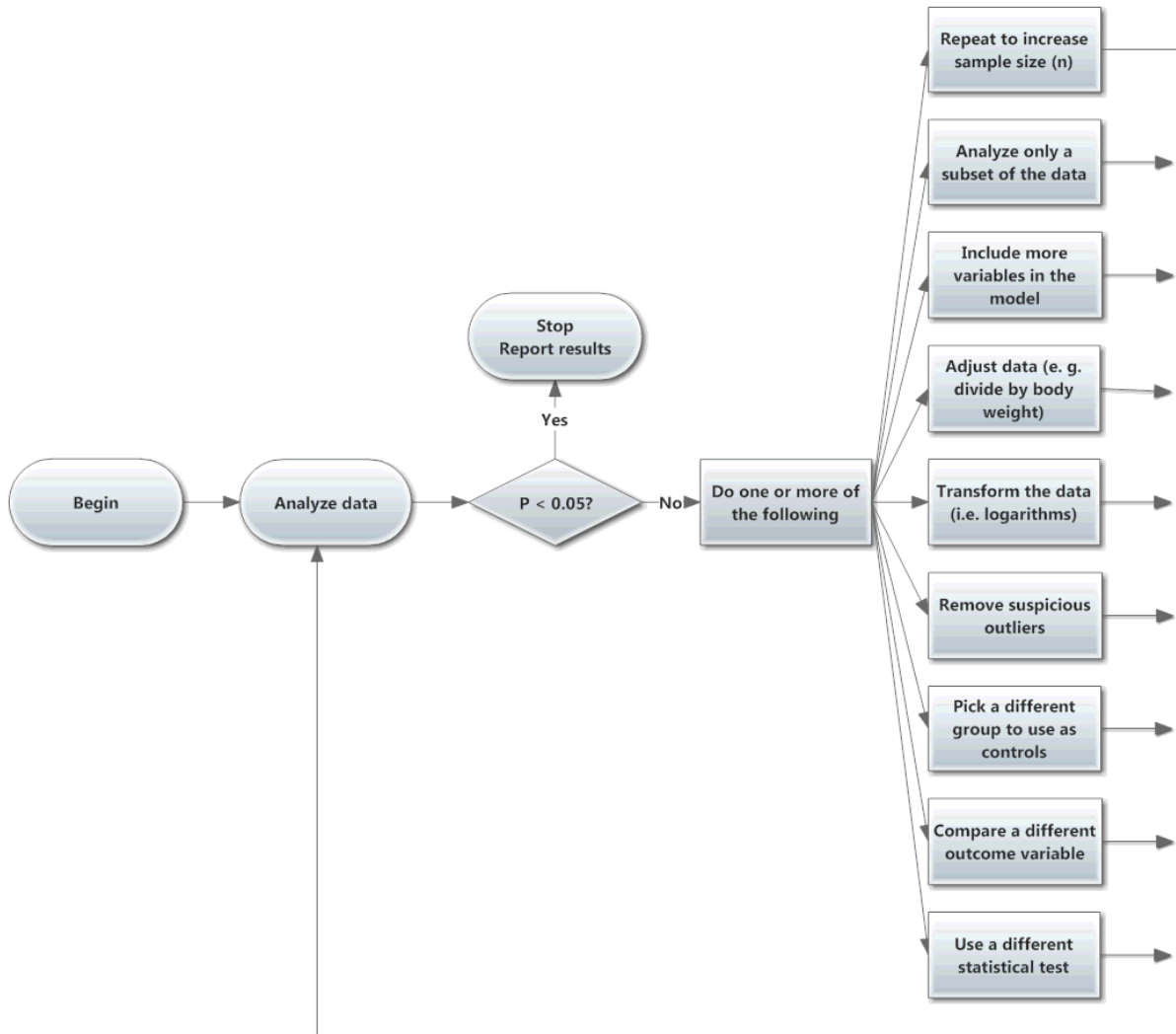
Collect and analyze some data. If the results are not statistically significant but show a difference or trend in the direction you expected, collect some more data and reanalyze. Or try a different way to analyze the data: remove a few outliers; transform to logarithms; try a nonparametric test; redefine the outcome by normalizing (say, dividing by each animal's weight); use a method to compare one variable while adjusting for differences in another; the list of possibilities is endless. Keep trying until you obtain a statistically significant result or until you run out of money, time, or curiosity.

The results from data collected this way cannot be interpreted at face value. Even if there really is no difference (or no effect), the chance of

finding a "statistically significant" result exceeds 5%. The problem is that you introduce bias when you choose to collect more data (or analyze the data differently) only when the P value is greater than 0.05. If the P value was less than 0.05 in the first analysis, it might be larger than 0.05 after collecting more data or using an alternative analysis. But you'd never see this if you only collected more data or tried different data analysis strategies when the first P value was greater than 0.05.

The term P-hacking was coined by Simmons et al (1) who also use the phrase, "too many investigator degrees of freedom". This is a general term that encompasses [dynamic sample size](#)[93] collection, [HARKing](#)[95], and more. There are three kinds of P-hacking:

- The first kind of P-hacking involves changing the actual values analyzed. Examples include *ad hoc* sample size selection, switching to an alternate control group (if you don't like the first results and your experiment involved two or more control groups), trying various combinations of independent variables to include in a multiple regression (whether the selection is manual or automatic), trying analyses with and without outliers, and analyzing various subgroups of the data.

- The second kind of P-hacking is reanalyzing a single data set with different statistical tests. Examples: Try parametric and nonparametric tests. Analyze the raw data, then try analyzing the logarithms of the data.

- The third kind of P-hacking is the garden of forking paths (2). This happens when researchers performed a reasonable analysis given their assumptions and their data, but would have done other analyses that were just as reasonable had the data turned out differently.

Exploring your data can be a very useful way to generate hypotheses and make preliminary conclusions. But all such analyses need to be clearly labeled, and then retested with new data.

Reference

1. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). [False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant](#). Psychological Science, 22(11), 1359–1366.

2. Gelman, A., & Loken, E. (2013). <u>The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or 'p-hacking' and the research hypothesis was posited ahead of time</u>. Unpublished as of Jan. 2016

4.7.8   Advice: Don't keep adding subjects until you hit 'significance'.

A commonly used approach leads to misleading results

This approach is tempting, but wrong (so shown crossed out):

~~Rather than choosing a sample size before beginning a study, simply repeat the statistical analyses as you collect more data, and then:~~

- ~~If the result is not statistically significant, collect some more data, and reanalyze.~~

- ~~If the result is statistically significant, stop the study.~~

The problem with this approach is that you'll keep going if you don't like the result, but stop if you do like the result. The consequence is that the chance of obtaining a "significant" result if the null hypothesis were true is a lot higher than 5%.

Simulations to demonstrate the problem

The graph below illustrates this point via simulation. We simulated data by drawing values from a Gaussian distribution (mean=40, SD=15, but these values are arbitrary). Both groups were simulated using exactly the same distribution. We picked N=5 in each group and computed an unpaired t test and recorded the P value. Then we added one subject to each group (so N=6) and recomputed the t test and P value. We repeated this until N=100 in each group. Then we repeated the entire simulation three times. These simulations were done comparing two groups with identical population means. So any "statistically significant" result we obtain must be a coincidence -- a Type I error.

The graph plots P value on the Y axis vs. sample size (per group) on the X axis. The green shaded area at the bottom of the graph shows P values less than 0.05, so deemed "statistically significant".
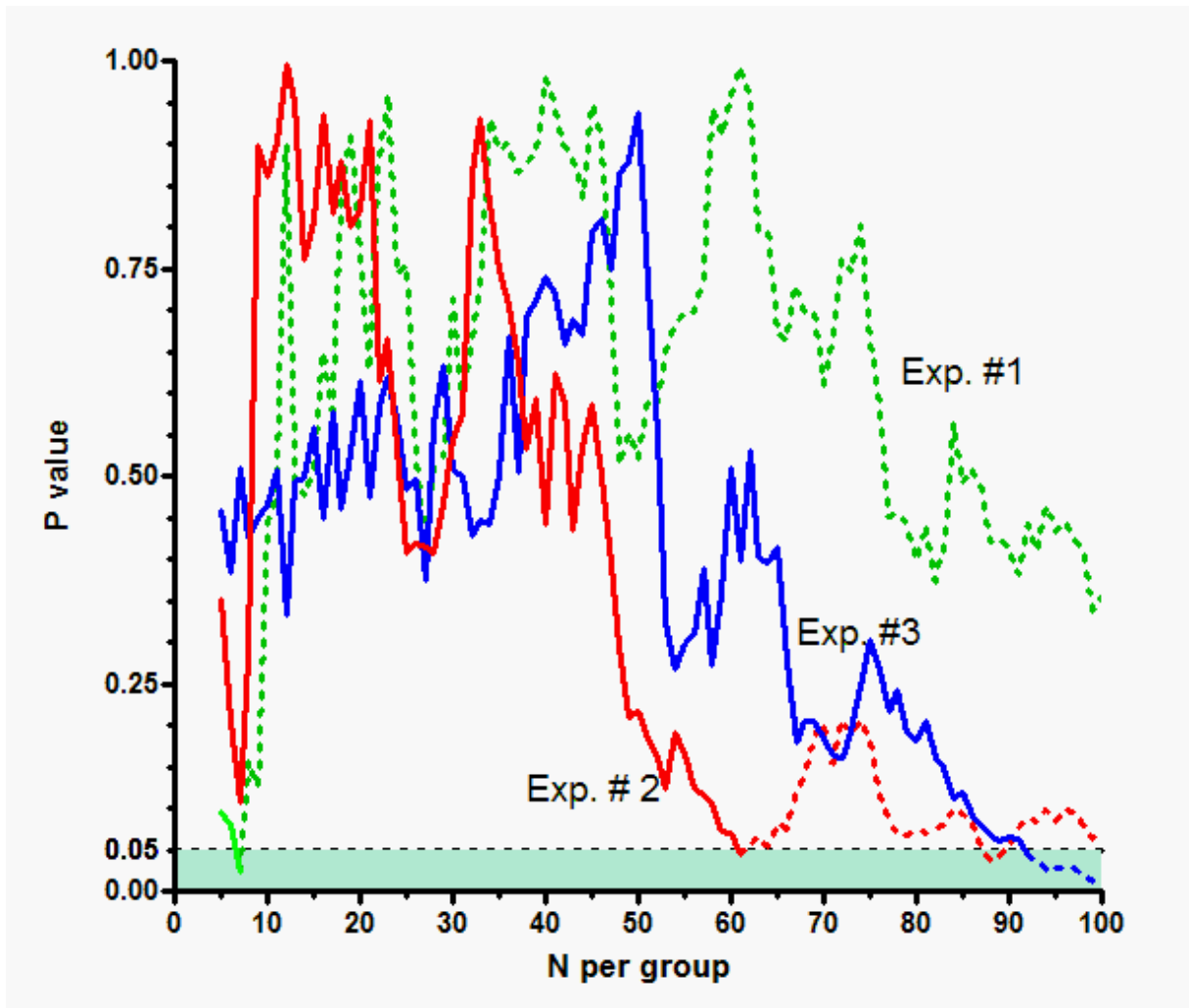
Experiment 1 (green) reached a P value less than 0.05 when N=7, but the P value is higher than 0.05 for all other sample sizes. Experiment 2 (red) reached a P value less than 0.05 when N=61 and also when N=88 or 89. Experiment 3 (blue) curve hit a P value less than 0.05 when N=92 to N=100.

If we followed the sequential approach, we would have declared the results in all three experiments to be "statistically significant". We would have stopped when N=7 in the first (green) experiment, so would never have seen the dotted parts of its curve. We would have stopped the second (red) experiment when N=6, and the third (blue) experiment when N=92. In all three cases, we would have declared the results to be "statistically significant".

Since these simulations were created for values where the true mean in both groups was identical, any declaration of "statistical significance" is a

Type I error. If the null hypothesis is true (the two population means are identical) we expect to see this kind of Type I error in 5% of experiments (if we use the traditional definition of alpha=0.05 so P values less than 0.05 are declared to be significant). But with this sequential approach, all three of our experiments resulted in a [Type I error.](#)[100] If you extended the experiment long enough (infinite N) all experiments would eventually reach statistical significance. Of course, in some cases you would eventually give up even without "statistical significance". But this sequential approach will produce "significant" results in far more than 5% of experiments, even if the null hypothesis were true, and so this approach is invalid.

Bottom line

It is important that you choose a sample size and stick with it. You'll fool yourself if you stop when you like the results, but keep going when you don't. The alternative is using specialized sequential or adaptive methods that take into account the fact that you analyze the data as you go. To learn more about these techniques, look up 'sequential' or 'adaptive' methods in advanced statistics books.

4.7.9   Advice: Don't HARK

Hypothesizing After the Result is Known (HARKing, Kerr 1998) is when you analyze the data many different ways (say different subgroups), discover an intriguing relationship, and then publish the data so it appears that the hypothesis was stated before the data were collected. [This cartoon from XKCD](#) shows the problem:
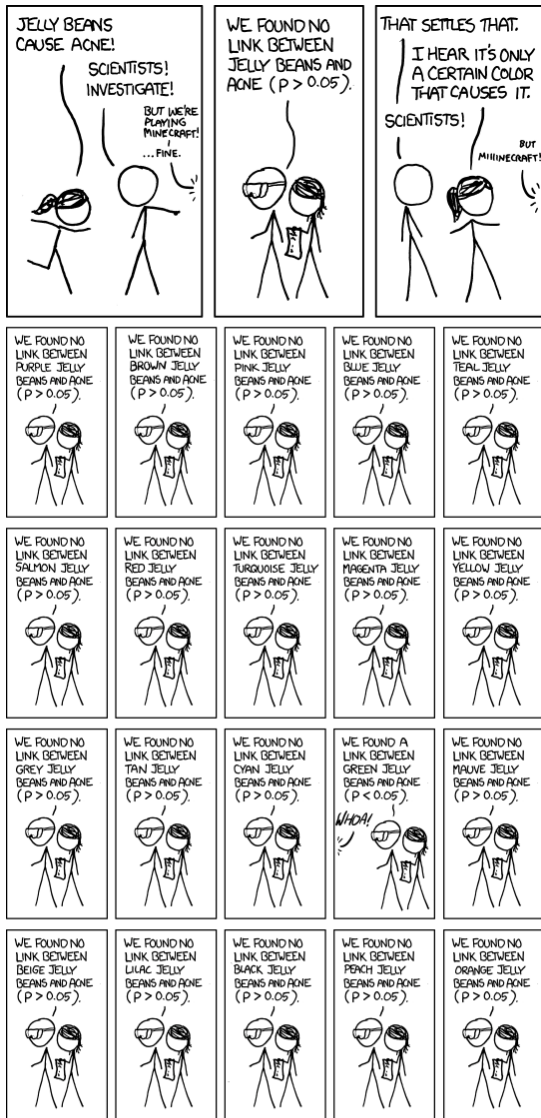
4.8 Statistical power

If there really is a difference (or correlation or association), you might not find it. it depends on the power of your experiment. This section explains what power means. Note that Prism does not provide any tools to compute power. Nonetheless, understanding power is essential to interpreting statistical results properly.

4.8.1 Key concepts: Statistical Power

Definitions of power and beta

Even if the treatment really does affect the outcome, you might not obtain a statistically significant difference in your experiment. Just by chance, your data may yield a P value greater than 0.05 (or whatever value, alpha, you use as your cutoff).

Let's assume we are comparing two means with a t test. Assume that the two means truly differ by a particular amount, and that you perform many experiments with the same sample size. Each experiment will have different values (by chance) so each t test will yield different results. In some experiments, the P value will be less than alpha (usually set to 0.05), so you call the results statistically significant. In other experiments, the P value will be greater than alpha, so you will call the difference not statistically significant.

If there really is a difference (of a specified size) between group means, you won't find a statistically significant difference in every experiment. Power is the fraction of experiments that you expect to yield a "statistically significant" P value. If your experimental design has high power, then there is a high chance that your experiment will find a "statistically significant" result if the treatment really works.

The variable beta is defined to equal 1.0 minus power (or 100% - power %). If there really is a difference between groups, then beta is the probability that an experiment like yours will yield a "not statistically significant" result.

How much power do I need?

The power is the chance that an experiment will result in a "statistically significant" result given some assumptions. How much power do you need? These guidelines might be useful:

- If the power is less than 50% to detect some effect that you think is worth detecting, then the study is really not helpful.

- Many investigators choose sample size to obtain a 80% power. This is arbitrary, but commonly used.

- Ideally, your choice of acceptable power should depend on the consequence of making a [Type II error] [100].

GraphPad StatMate

GraphPad Prism does not compute statistical power or sample size, but the companion program GraphPad StatMate does.

4.8.2 An analogy to understand statistical power

Looking for a tool in a basement

The concept of statistical power is a slippery one. Here is an analogy that might help (courtesy of John Hartung, SUNY HSC Brooklyn).

You send your child into the basement to find a tool. He comes back and says "it isn't there". What do you conclude? Is the tool there or not? There is no way to be sure.

So let's express the answer as a probability. The question you really want to answer is: "What is the probability that the tool is in the basement"? But that question can't really be answered without knowing the prior probability and using Bayesian thinking. We'll pass on that, and instead ask a slightly different question: "If the tool really is in the basement, what is the chance your child would have found it"?

The answer depends on the answers to these questions:

- How long did he spend looking? If he looked for a long time, he is more likely to have found the tool.

- How big is the tool? It is easier to find a snow shovel than the tiny screw driver you use to fix eyeglasses.

- How messy is the basement? If the basement is a real mess, he was less likely to find the tool than if it is super organized.

So if he spent a long time looking for a large tool in an organized basement, there is a high chance that he would have found the tool if it were there. So you can be quite confident of his conclusion that the tool isn't there. If he spent a short time looking for a small tool in a messy basement, his conclusion that "the tool isn't there" doesn't really mean very much.

Analogy with sample size and power

So how is this related to computing the power of a completed experiment? The question about finding the tool, is similar to asking about the power of a completed experiment. Power is the answer to this question: If an effect (of a specified size) really occurs, what is the chance that an experiment of a certain size will find a "statistically significant" result?

- The time searching the basement is analogous to sample size. If you collect more data you have a higher power to find an effect.

- The size of the tool is analogous to the effect size you are looking for. You always have more power to find a big effect than a small one.

- The messiness of the basement is analogous to the standard deviation of your data. You have less power to find an effect if the data are very scattered.

If you use a large sample size looking for a large effect using a system with a small standard deviation, there is a high chance that you would have obtained a "statistically significant effect" if it existed. So you can be quite confident of a conclusion of "no statistically significant effect". But if you use a small sample size looking for a small effect using a system with

a large standard deviation, then the finding of "no statistically significant effect" really isn't very helpful.

4.8.3   Type I, II (and III) errors

Type I and Type II errors

When you make a conclusion about whether an effect is statistically significant, you can be wrong in two ways:

- You've made a **type I error** when there really is no difference (association, correlation..) overall, but random sampling caused your data to show a statistically significant difference (association, correlation...). Your conclusion that the two groups are really different (associated, correlated) is incorrect.

- You've made a **type II error** when there really is a difference (association, correlation) overall, but random sampling caused your data to not show a statistically significant difference. So your conclusion that the two groups are not really different is incorrect.

Type 0 and Type III errors

Additionally, there are two more kinds of errors you can define:

- You've made a **type 0 error** when you get the right answer, but asked the wrong question! This is sometimes called a **type III error,** although that term is usually defined differently (see below).

- You've made a **type III error** when you correctly conclude that the two groups are statistically different, but are wrong about the direction of the difference. Say that a treatment really increases some variable, but you don't know this. When you run an experiment to find out, random sampling happens to produce very high values for the control subjects but low values for the treated subjects. This means that the mean of the treated subjects is lower (on average) in the treated group, and enough lower that the difference is statistically significant. You'll correctly reject the null hypothesis of no difference and correctly conclude that the treatment significantly altered the outcome. But you conclude that the treatment lowered the value on average, when in fact the treatment (on average, but not in your subjects) increases the

value. Type III errors are very rare, as they only happen when random chance leads you to collect low values from the group that is really higher, and high values from the group that is really lower.

4.8.4   Using power to evaluate 'not significant' results

Example data

Motulsky et al. asked whether people with hypertension (high blood pressure) had altered numbers of $alpha_2$-adrenergic receptors on their platelets (Clinical Science 64:265-272, 1983). There are many reasons to think that autonomic receptor numbers may be altered in hypertension. We studied platelets because they are easily accessible from a blood sample. The results are shown here:

Variable	Hypertensive	Control
Number of subjects	18	17
Mean receptor number (receptors per cell)	257	263
Standard Deviation	59.4	86.6

The two means were almost identical, so of course a t test computed a very high P value. We concluded that there is no statistically significant difference between the number of $alpha_2$ receptors on platelets of people with hypertension compared to controls. When we published this nearly 30 years ago, we did not go further.

These negative data can be interpreted in terms of confidence intervals or using power analyses. The two are equivalent and are just alternative ways of thinking about the data.

Interpreting not significant results using a confidence interval

All results should be accompanied by confidence intervals showing how well you have determined the differences (ratios, etc.) of interest. For our example, the 95% confidence interval for the difference between group means extends from -45 to 57 receptors/platelet. Once we accept the assumptions of the t test analysis, we can be 95% sure that this interval contains the true difference between mean receptor number in the two groups. To put this in perspective, you need to know that the average number of receptors per platelet is about 260.

The interpretation of the confidence interval must be in a scientific context. Here are two very different approaches to interpreting this confidence interval.

- The CI includes possibilities of a 20% change each way. A 20% change is huge. With such a wide CI, the data are inconclusive. Could be no change. Could be big decrease. Could be big increase.

- The CI tells us that the true difference is unlikely to be more than 20% in each direction. Since we are only interested in changes of 50%, we can conclude that any difference is, at best, only 20% or so, which is biologically trivial. These are solid negative results.

Both statements are sensible. It all depends on how you would interpret a 20% change. Statistical calculations can only compute probabilities. It is up to you to put these in a scientific context. As with power calculations, different scientists may interpret the same results differently.

Interpreting not significant results using power analysis

What was the power of this study to find a difference (if there was one)? The answer depends on how large the difference really is. Here are the results shown as a graph (created with GraphPad StatMate).
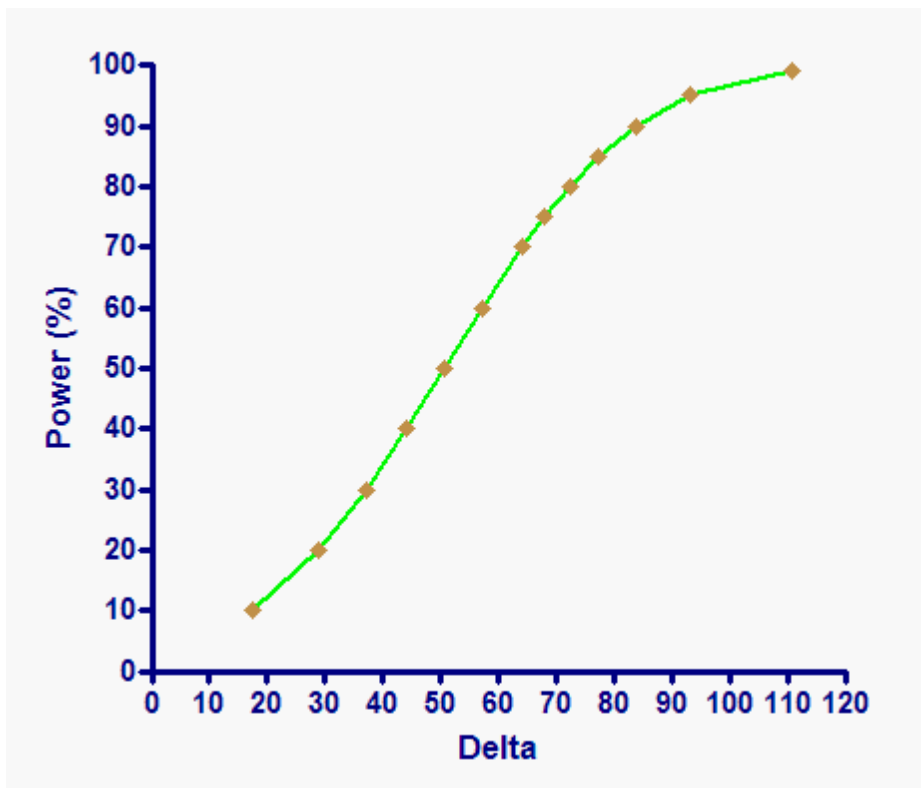
All studies have a high power to detect "big" differences and a low power to detect "small" differences, so power graph all have the same shape. Interpreting the graph depends on putting the results into a scientific context. Here are two alternative interpretations of the results:

- We really care about receptors in the heart, kidney, brain and blood vessels, not the ones in the platelets (which are much more accessible). So we will only pursue these results (do more studies) if the difference was 50%. The mean number of receptors per platelet is about 260, so we would only be seriously interested in these results if the difference exceeded half of that, or 130. From the graph above, you can see that this study had extremely high power to detect a difference of 130 receptors/platelet. In other words, if the difference really was that big, this study (given its sample size and variability) would almost certainly have found a statistically significant difference. Therefore, this study gives convincing negative results.

- Hey, this is hypertension. Nothing is simple. No effects are large. We've got to follow every lead we can. It would be nice to find differences of 50% (see above) but realistically, given the heterogeneity of hypertension, we can't expect to find such a large difference. Even if the difference was only 20%, we'd still want to do

follow up experiments. Since the mean number of receptors per platelet is 260, this means we would want to find a difference of about 50 receptors per platelet. Reading off the graph (or the table), you can see that the power of this experiment to find a difference of 50 receptors per cell was only about 50%. This means that even if there really were a difference this large, this particular experiment (given its sample size and scatter) had only a 50% chance of finding a statistically significant result. With such low power, we really can't conclude very much from this experiment. A reviewer or editor making such an argument could convincingly argue that there is no point publishing negative data with such low power to detect a biologically interesting result.

As you can see, the interpretation of power depends on how large a difference you think would be scientifically or practically important to detect. Different people may reasonably reach different conclusions. Note that it doesn't help at all to look up the power of a study to detect the difference we actually observed. This is a [common misunderstanding][105].

Comparing the two approaches

Confidence intervals and power analyses are based on the same assumptions, so the results are just different ways of looking at the same thing. You don't get additional information by performing a power analysis on a completed study, but a power analysis can help you put the results in perspective

The power analysis approach is based on having an alternative hypothesis in mind. You can then ask what was the probability that an experiment with the sample size actually used would have resulted in a statistically significant result if your alternative hypothesis were true.

If your goal is simply to understand your results, the confidence interval approach is enough. If your goal is to criticize a study of others, or plan a future similar study, it might help to also do a power analysis.

Reference

1. Motulsky HJ, O'Connor DT, Insel PA. Platelet alpha 2-adrenergic receptors in treated and untreated essential hypertension. Clin Sci (Lond). 1983 Mar;64(3):265-72.

4.8.5   Why doesn't Prism compute the power of tests

Post-hoc power analyses are rarely useful

Some programs report a power value as part of the results of t tests and other statistical comparisons. Prism does not do so, and this page explains why.

It is never possible to answer the question "what is the power of this experimental design?". That question is simply meaningless. Rather, you must ask "what is the power of this experimental design to detect an effect of a specified size?". The effect size might be a difference between two means, a relative risk, or some other measure of treatment effect.

Which effect size should you calculate power for? How large a difference should you be looking for? These are not statistical questions, but rather scientific questions. It only makes sense to do a power analysis when you think about the data scientifically. It makes sense to compute the power of a study design to detect an effect that is the smallest effect you'd care about. Or it makes sense to compute the power of a study to find an effect size determined by a prior study.

When computing statistical comparisons, some programs augment their results by reporting the power to detect the effect size (or difference, relative risk, etc.) actually observed in that particular experiment. The result is sometimes called *observed power*, and the procedure is sometimes called a *post-hoc power analysis* or *retrospective power analysis*.

Many (perhaps most) statisticians (and I agree) think that these computations are useless and misleading. If your study reached a conclusion that the difference is not statistically significant, then -- by definition-- its power to detect the effect actually observed is very low. You learn nothing new by such a calculation. It can be useful to compute the power of the study to detect a difference that would have been scientifically or clinically worth detecting. It is not worthwhile to compute the power of the study to detect the difference (or effect) actually observed.

Observed power is directly related to P value

Hoenig and Helsey (2001) pointed out that the observed power can be computed from the observed P value as well as the value of alpha you choose (usually 0.05). When the P value is 0.05 (assuming you define statistical significance to mean P<0.05, so have set alpha to 0.05), then the power must be 50%. If the P value is smaller than 0.05, the observed power is greater than 50%. If the P value is greater than 0.05, then the observed power is less than 50%. The observed power conveys no new information. The figure below (from Helsey, 2001) shows the relationship between P value and observed power of an unpaired t test, when alpha is set to 0.05.
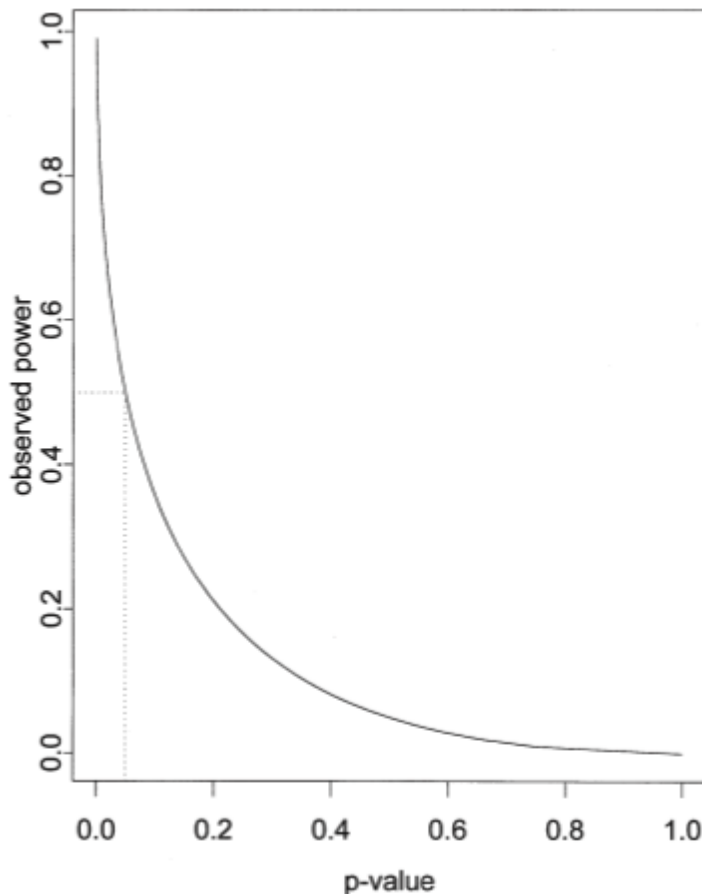
References

SN Goodman and JA Berlin, [The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting the Results](), Annals Internal Medicine 121: 200-206, 1994.

Hoenig JM, Heisey DM, 1710, The abuse of power, The American Statistician. February 1, 2001, 55(1): 19-24. doi:10.1198/000313001300339897.

Lenth, R. V. (2001), Some Practical Guidelines for Effective Sample Size Determination, The American Statistician, 55, 187-193

M Levine and MHH Ensom, Post Hoc Power Analysis: An Idea Whose Time Has Passed, Pharmacotherapy 21:405-409, 2001.

Thomas, L, Retrospective Power Analysis, Conservation Biology Vol. 11 (1997), No. 1, pages 276-280

4.8.6   Advice: How to get more power

If you are not happy with the power of your study, consider this list of approaches to increase power (abridged from Bausell and Li).

The best approach to getting more power is to collect more, or higher quality, data by:

- Increasing sample size. If you collect more data, you'll have more power.

- Increasing sample size for the group that is cheaper (or less risky). If you can't add more subjects to one group because it is too expensive, too risky, or too rare, add subjects to the other group.

- Reduce the standard deviation of the values (when comparing means) by using a more homogeneous group of subjects, or by improving the laboratory techniques.

You can also increase power, by making some compromises:

- Increase your choice for alpha. Alpha is the threshold P value below which you deem the results "statistically significant". While this is traditionally set at 0.05, you can choose another value. If you raise alpha, say to 0.10, you'll increase the power of the study to find a real difference while also increasing the chance of falsely finding a "significant" difference.

- Decide you only care about a larger difference or effect size. All studies have higher power to detect a large difference than a small one.

Reference

1. R. Barker Bausell, Yu-Fang Li, *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*, IBSN:0521809169.

4.9   Choosing sample size

How big a sample do you need? The answer, of course, is "it depends". This section explains what it depends on. Note that Prism does not do any sample size calculations, and this material is here for general interest.

4.9.1   Overview of sample size determination

The four questions

Many experiments and clinical trials are run with too few subjects. An underpowered study is a wasted effort because even substantial treatment effects are likely to go undetected. Even if the treatment substantially changed the outcome, the study would have only a small chance of finding a "statistically significant" effect.

When planning a study, therefore, you need to choose an appropriate sample size. The required sample size depends on your answers to these questions:

• How scattered do you expect your data to be?

• How willing are you to risk mistakenly finding a difference by chance?

• How big a difference are you looking for?

• How sure do you need to be that your study will detect a difference, if it exists? In other words, how much statistical power do you need?

The first question requires that you estimate the standard deviation you expect to see. If you can't estimate the standard deviation, you can't compute how many subjects you will need. If you expect lots of scatter, it is harder to discriminate real effects from random noise, so you'll need lots of subjects.

The second question is answered with your definition of statistical significance. Almost all investigators choose the 5% significance level, meaning that P values less than 0.05 are considered to be "statistically significant". If you choose a smaller significance level (say 1%), then you'll need more subjects.

The third and fourth questions are trickier. Everyone would prefer to plan a study that can detect very small differences, but this requires a large sample size. And everyone wants to design a study with lots of power, so it is quite certain to return a "statistically significant" result if the treatment actually works, but this too requires lots of subjects.

An alternative approach to sample size calculations

Rather than asking you to answer those last two questions, StatMate presents results in a table so you see the tradeoffs between sample size, power, and the effect size you can detect. You can look at this table, consider the time, expense and risk of your experiment, and decide on an appropriate sample size. Note that StatMate does not directly answer the question "how many subjects do I need?" but rather answers the related question "if I use N subjects, what information can I learn?". This approach to sample size calculations was recommended by Parker and Berman (1).

In some cases, StatMate's calculations may convince you that it is impossible to find what you want to know with the number of subjects you are able to use. This can be very helpful. It is far better to cancel such an experiment in the planning stage, than to waste time and money on a futile experiment that won't have sufficient power. If the experiment involves any clinical risk or expenditure of public money, performing such a study can even be considered unethical.

Also...

One benefit of larger sample size is you have more power to detect a specified effect, or with constant power can detect smaller effect sizes. But there is another reason to choose larger sample sizes when possible.

With larger samples, you can better assess teh distribution of the data. Is the assumption of sampling from a Gaussian, or lognormal, distribution reasonable? With larger samples, it is easier to assess

Reference

1. R. A. Parker and N. G. Berman, Sample Size: More than Calculations, Am. Statistician 57:166-170, 2003.

4.9.2   Why choose sample size in advance?

The appeal of choosing sample size as you go

To many, calculating sample size before the study starts seems like a nuisance. Why not do the analyses as you collect data? If your results are not statistically significant, then collect some more data, and reanalyze. If your results are statistically significant result, then stop the study and don't waste time or money on more data collection.

The problem with this approach is that you'll keep going if you don't like the result, but stop if you do like the result. The consequence is that the chance of obtaining a "significant" result if the null hypothesis were true is a lot higher than 5%.

Simulation to show the dangers of not choosing sample size in advance

The graph below illustrates this point via simulation. We simulated data by drawing values from a Gaussian distribution (mean=40, SD=15, but these values are arbitrary). Both groups were simulated using exactly the same distribution. We picked N=5 in each group and computed an unpaired t test and recorded the P value. Then we added one subject to each group (so N=6) and recomputed the t test and P value. We repeated this until N=100 in each group. Then we repeated the entire simulation three times. These simulations were done comparing two groups with identical population means. So any "statistically significant" result we obtain must be a coincidence -- a Type I error.

The graph plots P value on the Y axis vs. sample size (per group) on the X axis. The greenish shaded area at the bottom of the graph shows P values less than 0.05, so deemed "statistically significant".
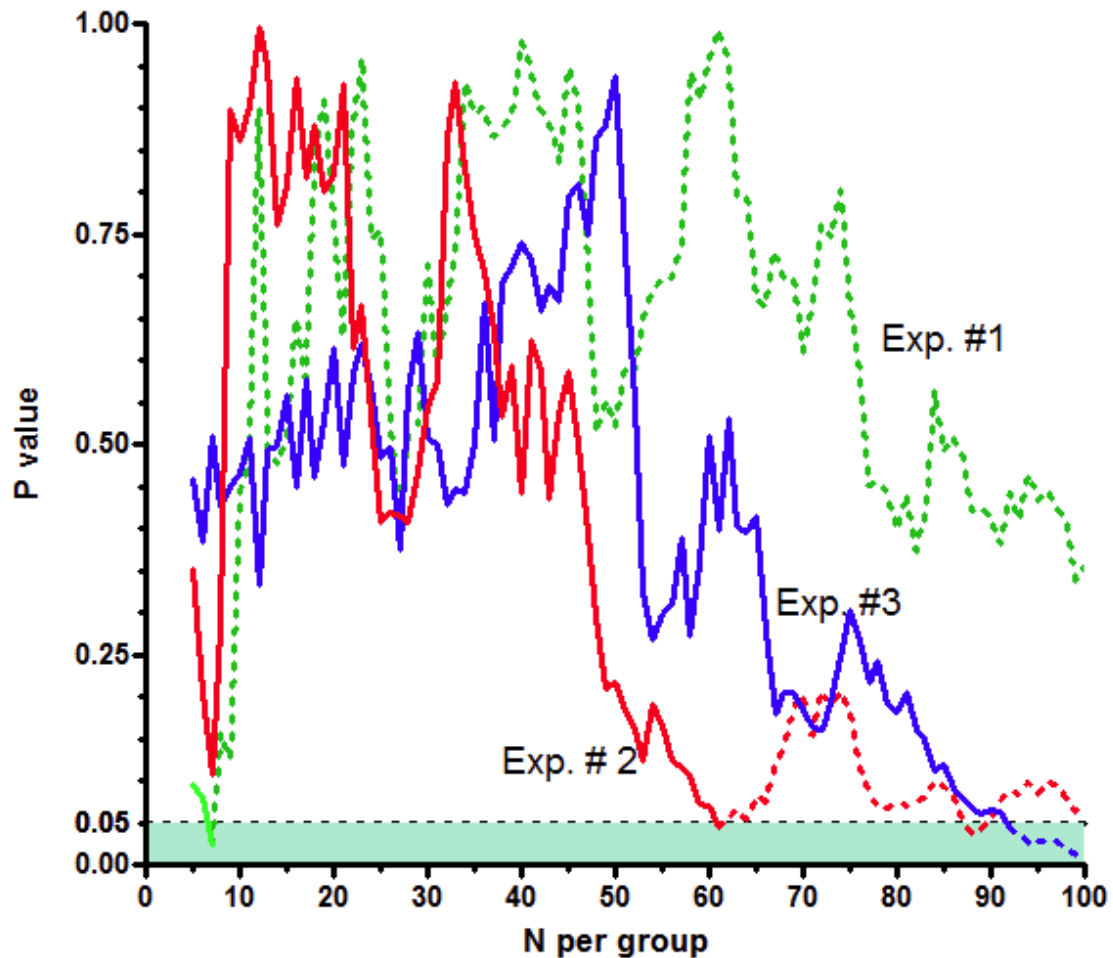
The green curve shows the results of the first simulated set of experiments. It reached a P value less than 0.05 when N=7, but the P value is higher than 0.05 for all other sample sizes. The red curve shows the second simulated experiment. It reached a P value less than 0.05 when N=61 and also when N=88 or 89. The blue curve is the third experiment. It has a P value less than 0.05 when N=92 to N=100.

If we followed the sequential approach, we would have declared the results in all three experiments to be "statistically significant". We would have stopped when N=7 in the green experiment, so would never have seen the dotted parts of its curve. We would have stopped the red experiment when N=6, and the blue experiment when N=92. In all three cases, we would have declared the results to be "statistically significant".

Since these simulations were created for values where the true mean in both populations was identical, any declaration of "statistical significance" is a Type I error. If the null hypothesis is true (the two population means are identical) we expect to see this kind of Type I error in 5% of experiments (if we use the traditional definition of alpha=0.05 so P values less than 0.05 are declared to be significant). But with this sequential approach, all three of our experiments resulted in a Type I error. If you extended the experiment long enough (infinite N) all experiments would eventually reach statistical significance. Of course, in some cases you would eventually give up even without "statistical significance". But this sequential approach will produce "significant" results in far more than 5% of experiments, even if the null hypothesis were true, and so this approach is invalid.

Bottom line

It is important that you choose a sample size and stick with it. You'll fool yourself if you stop when you like the results, but keep going when you don't. If experiments continue when results are not statistically significant, but stop when the results are statistically significant, the chance of mistakenly concluding that results are statistical significant is far greater than 5%.

There are some special statistical techniques for analyzing data sequentially, adding more subjects if the results are ambiguous and stopping if the results are clear. Look up 'sequential medical trials' in advanced statistics books to learn more.

4.9.3   Choosing alpha and beta for sample size calculations

Standard approach

When computing sample size, many scientists use standard values for alpha and beta. They always set alpha to 0.05, and beta to 0.20 (which allows for 80% power).

The advantages of the standard approach are that everyone else does it too and it doesn't require much thinking. The disadvantage is that it doesn't do a good job of deciding sample size

Choosing alpha and beta for the scientific context

When computing sample size, you should pick values for alpha and power according to the experimental setting, and on the consequences of making a Type I or Type II error ().

Let's consider four somewhat contrived examples. Assume you are running a screening test to detect compounds that are active in your system. In this context, a Type I error is concluding that a drug is effective, when it really is not. A Type II error is concluding that a drug is ineffective, when it fact it is effective. But the consequences of making a Type I or Type II error depend on the context of the experiment. Let's consider four situations.

- A. Screening drugs from a huge library of compounds with no biological rationale for choosing the drugs. You know that some of the "hits" will be false-positives (Type I error) so plan to test all those "hits" in another assay. So the consequence of a Type I error is that you need to retest that compound. You don't want to retest too many compounds, so can't make alpha huge. But it might make sense to set it to a fairly high value, perhaps 0.10. A Type II error occurs when you conclude that a drug has no statistically significant effect, when in fact the drug is effective. But in this context, you have hundreds of thousands of more drugs to test, and you can't possibly test them all. By choosing a low value of power (say 60%) you can use a smaller sample size. You know you'll miss some real drugs, but you'll be able to test many more with the same effort. So in this context, you can justify setting alpha to a high value. Summary: low power, high alpha.

- B. Screening selected drugs, chosen with scientific logic. The consequences of a Type I error are as before, so you can justify setting alpha to 0.10. But the consequences of a Type II error are more serious here. You've picked these compounds with some care, so a Type II error means that a great drug might be overlooked. In this context, you want to set power to a high value. Summary: high power, high alpha.

- C. Test carefully selected drugs, with no chance for a second round of testing. Say the compounds might be unstable, so you can only use them in one experiment. The results of this experiment -- the list of hits and misses -- will be used to do a structure-activity relationship which will then be used to come up with a new list of compounds for the chemists to synthesize. This will be a expensive and time-consuming task, so a lot is riding on this experiment, which can't easily be

repeated. In this case, the consequences of both a Type I and Type II error are pretty bad, so you set alpha to a small value (say 0.01) and power to a large value (perhaps 99%). Choosing these values means you'll need a larger sample size, but the cost is worth it here. Summary: high power, low alpha.

- D. Rethink scenario C. The sample size required for scenario C may be too high to be feasible. You simply can't run that many replicates. After talking to your colleagues, you decide that the consequence of making a Type I error (falsely concluding that a drug is effective) is much worse than making a Type II error (missing a real drug). One false hit may have a huge impact on your structure-activity studies, and lead the chemists to synthesize the wrong compounds. Falsely calling a drug to be inactive will have less severe consequences. Therefore you choose a low value of alpha and also a low power. Summary: low power, low alpha.

Bottom line

These scenarios are contrived, and I certainly am not in a position to tell anyone how to design their efforts to screen for drugs. But these scenarios make the point that you should choose values for alpha and power after carefully considering the consequences of making a Type I and Type II error. These consequences depend on the scientific context of your experiment. It doesn't really make sense to just use standard values for alpha and power.

4.9.4   What's wrong with standard values for effect size?

The appeal of using standard effect sizes

Computing sample size requires that you decide how large a difference you are looking for -- how large a difference (association, correlation..) would be scientifically interesting. You'll need a large sample size if your goal is to find tiny differences. You can get by with smaller samples, if you are only looking for larger differences.

In a very influential book (1) , Jacob Cohen makes some recommendations for what to do when you don't know what effect size you are looking for. He limits these recommendations to the behavioral sciences (his area of expertise), and warns that all general recommendations are more useful in some circumstances than others. Here are his guidelines for an unpaired t test:

- A "small" difference between means is equal to one fifth the standard deviation.

- A "medium" effect size is equal to one half the standard deviation.

- A "large" effect is equal to 0.8 times the standard deviation.

So if you are having trouble deciding what effect size you are looking for (and therefore are stuck and can't determine a sample size), Cohen would recommend you choose whether you are looking for a "small", "medium", or "large" effect, and then use the standard definitions.

The problem with standard effect sizes

Russell Lenth (2) argues that you should avoid these "canned" effect sizes, and I agree. You must decide how large a difference you care to detect based on understanding the experimental system you are using and the scientific questions you are asking. Cohen's recommendations seem a way to avoid thinking about the point of the experiment. It doesn't make sense to only think about the difference you are looking at in terms of the scatter you expect to see (anticipated standard deviation), without even considering what the mean value might be.

If you choose standard definitions of alpha (0.05), power (80%), and effect size (see above), then there is no need for any calculations. If you accept those standard definitions for all your studies (that use an unpaired t test to compare two groups), then all studies need a sample size of 26 in each group to detect a large effect, 65 in each group to detect a medium effect, 400 in each group to detect a small effect.

Bottom line

Choosing standard effect sizes is really the same as picking standard sample sizes.

References

1. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 1988, ISBN=978-0805802832

2. R. V. Lenth, R. V. (2001), "Some Practical Guidelines for Effective Sample Size Determination,'' The American Statistician, 55, 187-193. A preliminary draft was posted as a pdf file.

4.9.5 Sample size for nonparametric tests

The problem of choosing sample size for data to be analyzed by nonparametric tests

Nonparametric tests are used when you are not willing to assume that your data come from a Gaussian distribution. Commonly used nonparametric tests are based on ranking values from low to high, and then looking at the distribution of sum-of-ranks between groups. This is the basis of the Wilcoxon rank-sum (test one group against a hypothetical median), Mann-Whitney (compare two unpaired groups), Wilcoxon matched pairs (compare two matched groups), Kruskal-Wallis (three or more unpaired groups) and Friedman (three or more matched groups).

When calculating a nonparametric test, you don't have to make any assumption about the distribution of the values. That is why it is called nonparametric. But if you want to calculate necessary sample size for a study to be analyzed by a nonparametric test, you must make an assumption about the distribution of the values. It is not enough to say the distribution is not Gaussian, you have to say what kind of distribution it is. If you are willing to make such an assumption (say, assume an exponential distribution of values, or a uniform distribution) you should consult an advanced text or use a more advanced program to compute sample size.

A useful rule-of-thumb

Most people choose a nonparametric test when they don't know the shape of the underlying distribution. Without making an explicit assumption about the distribution, detailed sample size calculations are impossible. Yikes!

But all is not lost! Depending on the nature of the distribution, the nonparametric tests might require either more or fewer subjects. But they never require more than 15% additional subjects if the following two assumptions are true:

- You are looking at reasonably high numbers of subjects (how high depends on the nature of the distribution and test, but figure at least a few dozen)

- The distribution of values is not really unusual (doesn't have infinite tails, in which case its standard deviation would be infinitely large).

So a general rule of thumb is this (1):

> If you plan to use a nonparametric test, compute the sample size required for a parametric test and add 15%.

Reference

Erich L. Lehmann, Nonparametrics : Statistical Methods Based on Ranks, Revised, 1998, ISBN=978-0139977350, pages 76-81.

4.10   Multiple comparisons

Multiple comparisons are everywhere, and understanding multiple comparisons is key to understanding statistics.

4.10.1 The problem of multiple comparisons

4.10.1.1 The multiple comparisons problem

Review of the meaning of P value and alpha

Interpreting an individual P value is straightforward. Consider the simple case of comparing two means. Assuming the null hypothesis is true, the P value is the probability that random subject selection alone would result in a difference in sample means (or a correlation or an association...) at least as large as that observed in your study.

Alpha is a threshold that you set in advance. If the P value is less than alpha, you deem the comparison "statistically significant". If you set alpha to 5% and if the null hypothesis is true, there is a 5% chance of randomly selecting subjects such that you erroneously infer a treatment effect in the population based on the difference observed between samples

Multiple comparisons

Many scientific studies test multiple hypotheses. Some studies can generate hundreds, or even thousands of comparisons.

Interpreting multiple P values is difficult. If you test several independent null hypotheses and leave the threshold at 0.05 for each comparison, the chance of obtaining at least one "statistically significant" result is greater than 5% (even if all null hypotheses are true). This graph shows the problem. The probability at least one "significant" comparison is computed from the number of comparisons (N) on the X axis using this equation: $100(1.00 - 0.95^N)$.
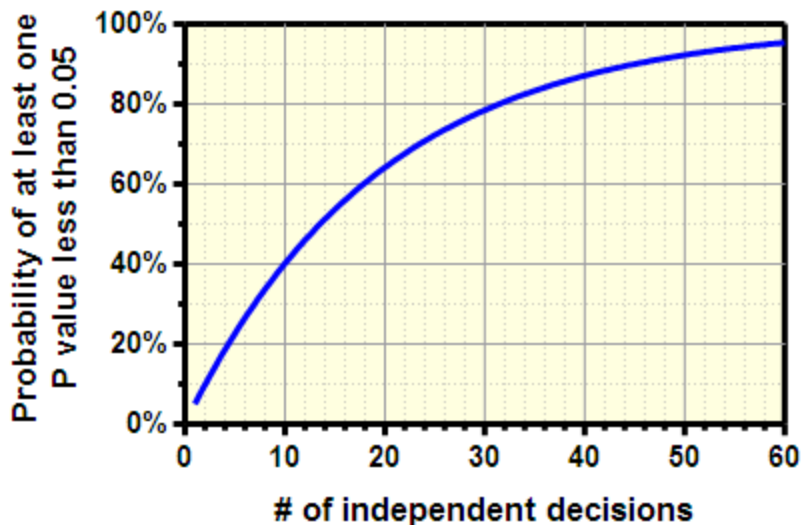
Remember the unlucky number 13. If you perform 13 independent comparisons, your chances are about 50% of obtaining at least one 'significant' P value (<0.05) just by chance.

The graph above (and the equation that generated it) assumes that the comparisons are independent. In other words, it assumes that the chance of any one comparison having a small P value is not related to the chance of any other comparison having a small P value. If the comparisons are not independent, it really is impossible to compute the probability shown the the graph.

Example

Let's consider an example. You compare control and treated animals, and you measure the level of three different enzymes in the blood plasma. You perform three separate t tests, one for each enzyme, and use the traditional cutoff of alpha=0.05 for declaring each P value to be significant. Even if the treatment doesn't actually do anything, there is a

14% chance that one or more of your t tests will be "statistically significant".

If you compare 10 different enzyme levels with 10 t tests, the chance of obtaining at least one "significant" P value by chance alone, even if the treatment really does nothing, is 40%. Finally, imagine that you test 100 different enzymes, at 10 time points, with 12 pre treatments... If you don't correct for multiple comparisons, you are almost certain to find that some of them are 'significant', even if really all null hypotheses are true.

You can only correct for comparisons you know about

When reading a study, you can only account for multiple comparisons when you know about all the comparisons made by the investigators. If they report only "significant" differences, without reporting the total number of comparisons, it is not possible to properly evaluate the results. Ideally, all analyses should be planned before collecting data, and all should be reported[20].

Learn more

Multiple comparisons is a big problem, affecting interpretation of almost all statistical results. Learn more from a review by Berry (1), excerpted below, or from chapter 22 and 23 of Intuitive Biostatistics(2).

> "*Most scientists are oblivious to the problems of multiplicities. Yet they are everywhere. In one or more of its forms, multiplicities are present in every statistical application. They may be out in the open or hidden. And even if they are out in the open, recognizing them is but the first step in a difficult process of inference. Problems of multiplicities are the most difficult that we statisticians face. They threaten the validity of every statistical conclusion.* " (1)

1.Berry, D. A. (2007). The difficult and ubiquitous problems of multiplicities. Pharmaceutical Statistics , 6, 155-160

2. Motulsky, H.J. (2010). Intuitive Biostatistics, 3rd edition. Oxford University Press. ISBN=978-0-19-994664-8.

4.10.1.2 Lingo: Multiple comparisons

Multiple comparison test applies whenever you make several comparisons at once.

Post test is sometimes used interchangeably with multiple comparison test (above) but sometimes as a short form of post-hoc test (below).

Post-hoc test is used for situations where you decide which comparisons you want to make after looking at the data. You didn't plan ahead.

Planned comparison tests require that you focus in on a few scientifically sensible comparisons. You can't decide which comparisons to do after looking at the data. The choice must be based on the scientific questions you are asking, and be chosen when you design the experiment.

Orthogonal comparison. When you only make a few comparison, the comparisons are called "orthogonal" when the each comparison is among different groups. Comparing Groups A and B is orthogonal to comparing Groups C and D, because there is no information in the data from groups A and B that is relevant when comparing Groups C and D. In contrast, comparing A and B is not orthogonal to comparing B and C.

Multiple comparisons procedures are used to cope with a set of comparisons at once. They analyze a *family* of comparisons.

When you set the customary significance level of 5% (or some other value) to apply to the entire family of comparisons, it is called a *familywise* error rate. When that significance level applies to only one comparison at a time (no correction for multiple comparisons), it is called a *per-comparison* error rate.

4.10.2 Three approaches to dealing with multiple comparisons

.

4.10.2.1 Approach 1: Don't correct for multiple comparisons

.

4.10.2.1.1 When it makes sense to not correct for multiple comparisons

Multiple comparisons can be accounted for with <u>Bonferroni and other corrections</u>[128], or by the approach of controlling the <u>False Discover Rate</u>[138]. But these approaches are not always needed. Here are three situations were special calculations are not needed.

Account for multiple comparisons when interpreting the results rather than in the calculations

Some statisticians recommend never correcting for multiple comparisons while analyzing data (1,2). Instead report all of the individual P values and confidence intervals, and make it clear that no mathematical correction was made for multiple comparisons. This approach requires that all comparisons be reported. When you interpret these results, you need to informally account for multiple comparisons. If all the null hypotheses are true, you'd expect 5% of the comparisons to have uncorrected P values less than 0.05. Compare this number to the actual number of small P values.

Following ANOVA, the unprotected Fishers Least Significant Difference test[127] follows this approach.

Corrections for multiple comparisons may not be needed if you make only a few planned comparisons

The term *planned comparison* is used when:

- You focus in on a few scientifically sensible comparisons rather than every possible comparison.

- The choice of which comparisons to make was part of the experimental design.

- You did not succumb to the temptation to do more comparisons after looking at the data.

It is important to distinguish between comparisons that are preplanned and those that are not (post hoc). It is not a planned comparison if you first look at the data, and based on that peek decide to make only two comparisons. In that case, you implicitly compared all the groups.

If you only make a few planned comparisons, some statistical texts recommend setting the significance level (or the meaning of the confidence interval) for each individual comparison without correction for multiple comparisons. In this case, the 5% traditional significance level applies to each individual comparisons, rather than the whole family of comparisons.

The logic of not correcting for multiple comparisons seems to be that some statisticians think this extra power is a deserved bonus for planning the experiment carefully and focussing on only a few scientifically sensible comparisons. Kepel and Wickles advocate this approach (reference below). But they also warn it is not fair to "plan" to make all comparisons, and thus not correct for multiple comparisons.

Corrections for multiple comparisons are not needed when the comparisons are complementary

Ridker and colleagues (3) asked whether lowering LDL cholesterol would prevent heart disease in patients who did not have high LDL concentrations and did not have a prior history of heart disease (but did have an abnormal blood test suggesting the presence of some inflammatory disease). They study included almost 18,000 people. Half received a statin drug to lower LDL cholesterol and half received placebo.

The investigators primary goal (planned as part of the protocol) was to compare the number of "end points" that occurred in the two groups, including deaths from a heart attack or stroke, nonfatal heart attacks or strokes, and hospitalization for chest pain. These events happened about half as often to people treated with the drug compared to people taking placebo. The drug worked.

The investigators also analyzed each of the endpoints separately. Those taking the drug (compared to those taking placebo) had fewer deaths, and fewer heart attacks, and fewer strokes, and fewer hospitalizations for chest pain.

The data from various demographic groups were then analyzed separately. Separate analyses were done for men and women, old and young, smokers and nonsmokers, people with hypertension and without, people with a family history of heart disease and those without. In each of 25 subgroups, patients receiving the drug experienced fewer primary endpoints than those taking placebo, and all these effects were statistically significant.

The investigators made no correction for multiple comparisons for all these separate analyses of outcomes and subgroups. No corrections were needed, because the results are so consistent. The multiple comparisons each ask the same basic question a different way (does the drug prevent disease?), and all the comparisons point to the same conclusion – people

taking the drug had less cardiovascular disease than those taking placebo.

References

1. Rothman, K.J. (1990). No adjustments are needed for multiple comparisons .Epidemiology, 1: 43-46.

2. D. J. Saville, Multiple Comparison Procedures: The Practical Solution. The American Statistician, 44:174-180, 1990

3. Ridker. Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. N Engl J Med (2008) vol. 359 pp. 3195

4.10.2.1.2 Example: Planned comparisons

What are planned comparisons?

The term *planned comparison* is used when you focus in on a few scientifically sensible comparisons. You don't do every possible comparison. And you don't decide which comparisons to do after looking at the data. Instead, you decide -- as part of the experimental design -- to only make a few comparisons.

Some statisticians recommend not correcting for multiple comparisons when you make only a few *planned* comparisons. The idea is that you get some bonus power as a reward for having planned a focussed study.

Prism always corrects for multiple comparisons, without regard for whether the comparisons were planned or post hoc. But you can get Prism to do the planned comparisons for you once you realize that a planned comparison is identical to a Bonferroni corrected comparison for selected pairs of means, when there is only one pair to compare.

Example data with incorrect analysis

In the graph below, the first column shows control data, and the second column shows data following a treatment. The goal of the experiment is to see if the treatment changes the measured activity (shown on the Y axis). To make sure the vehicle (solvent used to dissolve the treatment) isn't influencing the result, the experiment was performed with another

control that lacked the vehicle (third column). To make sure the experiment is working properly, nonspecific (blank) data were collected and displayed in the fourth column.
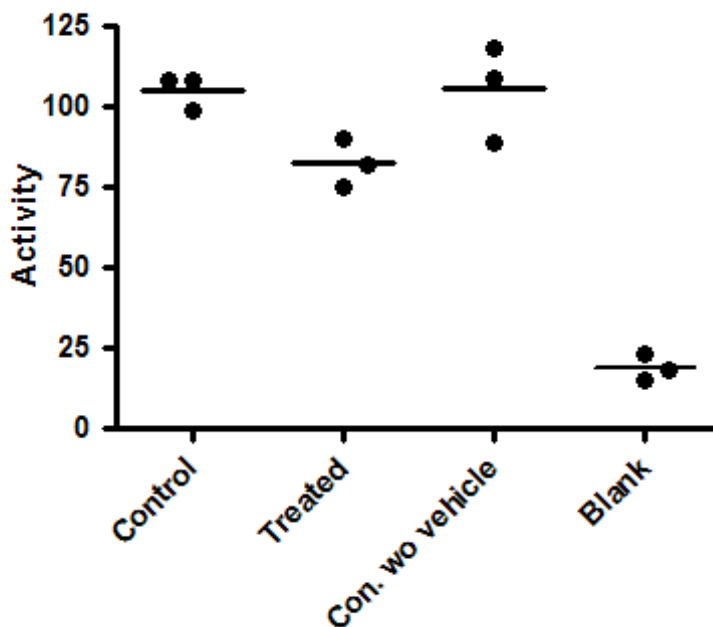
Here are the results of one-way ANOVA and Tukey multiple comparison tests comparing every group with every other group.

One-way analysis of variance

P value	P<0.0001
P value summary	***
Are means signif. different? (P < 0.05)	Yes
Number of groups	4
F	62.69
R squared	0.9592

ANOVA Table	**SS**	**df**	**MS**
Treatment (between columns)	15050	3	5015
Residual (within columns)	640	8	80
Total	15690	11	

Tukey's Multiple Comparison Test	**Mean Diff.**	**q**	**P value**	**95% CI of diff**
Control vs Treated	22.67	4.389	**P > 0.05**	-0.7210 to 46.05
Control vs Con. wo vehicle	-0.3333	0.06455	P > 0.05	-23.72 to 23.05
Control vs Blank	86.33	16.72	P < 0.001	62.95 to 109.7
Treated vs Con. wo vehicle	-23	4.454	P > 0.05	-46.39 to 0.3877
Treated vs Blank	63.67	12.33	P < 0.001	40.28 to 87.05

Con. wo vehicle vs Blank	86.67	16.78	P < 0.001	63.28 to 110.1

The overall ANOVA has a very low P value, so you can reject the null hypothesis that all data were sampled from groups with the same mean. But that really isn't very helpful. The fourth column is a negative control, so of course has much lower values than the others. The ANOVA P value answers a question that doesn't really need to be asked.

Tukey's multiple comparison tests were used to compare all pairs of means (table above). You only care about the first comparison -- control vs. treated -- which is not statistically significant (P>0.05).

These results don't really answer the question your experiment set out to ask. The Tukey multiple comparison tests set the 5% level of significance to the entire family of six comparisons. But five of those six comparisons don't address scientifically valid questions. You expect the blank values to be much lower than the others. If that wasn't the case, you wouldn't have bothered with the analysis since the experiment hadn't worked. Similarly, if the control with vehicle (first column) was much different than the control without vehicle (column 3), you wouldn't have bothered with the analysis of the rest of the data. These are control measurements, designed to make sure the experimental system is working. Including these in the ANOVA and post tests just reduces your power to detect the difference you care about.

Example data with planned comparison

Since there is only one comparison you care about here, it makes sense to only compare the control and treated data.

From Prism's one-way ANOVA dialog, choose the Bonferroni comparison between selected pairs of columns, and only select one pair.
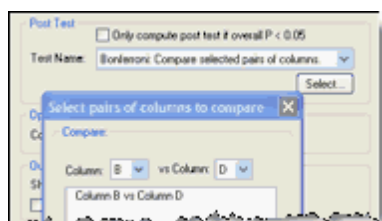
The difference is statistically significant with P<0.05, and the 95% confidence interval for the difference between the means extends from 5.826 to 39.51.

When you report the results, be sure to mention that your P values and confidence intervals are not corrected for multiple comparisons, so the P values and confidence intervals apply individually to each value you report and not to the entire family of comparisons.

In this example, we planned to make only one comparison. If you planned to make more than one comparison, choose the Fishers Least Significant Difference approach to performing multiple comparisons. When you report the results, be sure to explain that you are doing planned comparisons so have not corrected the P values or confidence intervals for multiple comparisons.

Example data analyzed by t test

The planned comparisons analysis depends on the assumptions of ANOVA, including the assumption that all data are sampled from groups with the same scatter. So even when you only want to compare two groups, you use data in all the groups to estimate the amount of scatter within groups, giving more degrees of freedom and thus more power.

That assumption seems dubious here. The blank values have less scatter than the control and treated samples. An alternative approach is to ignore the control data entirely (after using the controls to verify that the experiment worked) and use a t test to compare the control and treated data. The t ratio is computed by dividing the difference between the means (22.67) by the standard error of that difference (5.27, calculated from the two standard deviations and sample sizes) so equals 4.301. There are six data points in the two groups being compared, so four degrees of freedom. The P value is 0.0126, and the 95% confidence interval for the difference between the two means ranges from 8.04 to 37.3.

How planned comparisons are calculated

First compute the standard error of the difference between groups 1 and 2. This is computed as follows, where N_1 and N_2 are the sample sizes of the two groups being compared (both equal to 3 for this example) and $MS_{residual}$ is the residual mean square reported by the one-way ANOVA (80.0 in this example):

$$SE_{Difference} = \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right) \cdot MS_{Residual}}$$

For this example, the standard error of the difference between the means of column 1 and column 2 is 7.303.

Now compute the t ratio as the difference between means (22.67) divided by the standard error of that difference (7.303). So t=3.104. Since the MS_{error} is computed from all the data, the number of degrees of freedom is the same as the number of residual degrees of freedom in the ANOVA table, 8 in this example (total number of values minus number of groups). The corresponding P value is 0.0146.

The 95% confidence interval extends from the observed mean by a distance equal to SE of the difference (7.303) times the critical value from the t distribution for 95% confidence and 8 degrees of freedom (2.306). So the 95% confidence interval for the difference extends from 5.826 to 39.51.

4.10.2.1.3 Fisher's Least Significant Difference (LSD)

Fishers Least Significant Difference (LSD) test in Prism

Following one-way (or two-way) analysis of variance (ANOVA), you may want to explore further and compare the mean of one group with the mean of another. One way to do this is by using Fisher's Least Significant Difference (LSD) test.

Key facts about Fisher's LSD test

- The Fishers LSD test is basically a set of individual t tests. It is only used as a followup to ANOVA.

- Unlike the Bonferroni, Tukey, Dunnett and Holm methods, Fisher's LSD does **not** correct for multiple comparisons.

- If you choose to use the Fisher's LSD test, you'll need to account for multiple comparisons when you interpret the data, since the computations themselves do not correct for multiple comparisons.

- The only difference between a set of t tests and the Fisher's LSD test, is that t tests compute the pooled SD from only the two groups being

compared, while the Fisher's LSD test computes the pooled SD from all the groups (which gains power but depends on the assumption that all groups are sampled from populations with the same SD).

- Prism performs the *unprotected* LSD test. *Unprotected* simply means that calculations are reported regardless of the results of the ANOVA. The unprotected Fisher's LSD test is essentially a set of t tests, without any correction for multiple comparisons.

- Prism does not perform a *protected* Fisher's LSD test. *Protection* means that you only perform the calculations described above when the overall ANOVA resulted in a P value less than 0.05 (or some other value set in advance). This first step sort of controls the false positive rate for the entire family of comparisons. While the protected Fisher's LSD test is of historical interest as the first multiple comparisons test ever developed, it is no longer recommended. It pretends to correct for multiple comparisons, but doesn't do so very well.

- [How it works](#)[378].

4.10.2.2 Approach 2: Control the Type I error rate for the family of comparisons

.

4.10.2.2.1 What it means to control the Type I error for a family

Let's consider what would happen if you did many comparisons, and determined whether each result is 'significant' or not. Also assume that we are 'mother nature' so know whether a difference truly exists or not in the populations from which the data were sampled.

In the table below, the top row represents the results of comparisons where the null hypothesis is true -- the treatment really doesn't work. Nonetheless, some comparisons will mistakenly yield a 'significant' conclusion. The second line shows the results of comparisons where there truly is a difference. Even so, you won't get a 'significant' result in every experiment.

A, B, C and D represent numbers of comparisons, so the sum of A+B+C+D equals the total number of comparisons you are making. You can't make this table from experimental data because this table is an overview of many experiments.

	"Significant"	"Not significant"	Total
No difference. Null hypothesis true	A	B	A+B
A difference truly exists	C	D	C+D
Total	A+C	B+D	A+B+C+D

In the table above, alpha is the expected value of A/(A+B). If you set alpha to the usual value of 0.05, this means you expect 5% of all comparisons done when the null hypothesis is true (A+B) to be statistically significant (in the first column). So you expect A/(A+B) to equal 0.05.

The usual approach to correcting for multiple comparisons is to set a stricter threshold to define statistical significance. The goal is to set a strict definition of significance such that -- if all null hypotheses are true -- there is only a 5% chance of obtaining one or more 'significant' results by chance alone, and thus a 95% chance that *none* of the comparisons will lead to a 'significant' conclusion. The 5% applies to the entire experiment, so is sometimes called an *experimentwise error rate* or *familywise error rate (*the two are synonyms*)*.

Setting a stricter threshold for declaring statistical significance ensures that you are far less likely to be mislead by false conclusions of 'statistical significance'. But this advantage comes at a cost: your experiment will have less power to detect true differences.

The methods of [Bonferroni][131], [Tukey, Dunnett][136], [Dunn][137], [Holm][133] (and more) all use this approach.

4.10.2.2.2 Multiplicity adjusted P values

If you choose the Bonferroni, Tukey, Dunnett or Dunn (nonparametric) multiple comparisons test, Prism can compute a *multiplicity adjusted P*

value for each comparison. This is a choice on the Options tab of the ANOVA dialog. It is checked by default.

Key facts about multiplicity adjusted P values

- A separate adjusted P values is computed for each comparison in a family of comparisons.

- The value of each adjusted P value depends on the entire family. The adjusted P value for one particular comparison would have a different value if there were a different number of comparisons or if the data in the other comparisons were changed.

- Because the adjusted P value is determined by the entire family of comparisons, it cannot be compared to an individual P value computed by a t test or Fishers Least Significant Difference test.

- Choosing the compute adjusted P values won't change Prism's reporting of statistical significance. Instead Prism will report an additional set of results -- the adjusted P value for each comparison.

- Multiplicity adjusted P values are not reported by most programs. If you choose to report adjusted P values, be sure to explain that they are *multiplicity adjusted P values*, and to give a reference. Avoid ambiguous terms such as *exact P values*.

What are multiplicity adjusted P values?

Before defining adjusted P values, let's review the meaning of a P value from a single comparison. The P value is the answer to two equivalent questions:

- If the null hypothesis were true, what is the chance that random sampling would result in a difference this large or larger?

- What is the smallest definition of the threshold (alpha) of statistical significance at which this result would be statistically significant?

The latter form of the question is less familiar, but equivalent to the first. It leads to a definition of the adjusted P value, which is the answer to this question:

- What is the smallest significance level, when applied to the entire family of comparisons, at which this particular comparison will be deemed statistically significant?

The idea is pretty simple. There is nothing special about significance levels of 0.05 or 0.01... You can set the significance level to any probability you want. The adjusted P value is the smallest familywise significance level at which a particular comparison will be declared statistically significant as part of the multiple comparison testing.

Here is a simple way to think about it. You perform multiple comparisons twice. The first time you set the familywise significance level to 5%. The second time, you set it to 1% level. If a particular comparison is statistically significant by the first calculations (5% significance level) but is not for the second (1% significance level), its adjusted P value must be between 0.01 and 0.05, say 0.0323.

Learn more about adjusted P values

Three places to learn about adjusted P values:

- Wright defines these adjusted P values and argues for their widespread use (S.P. Wright. Adjusted P-values for simultaneous inference. Biometrics 48:1005-1013,1992).

- Multiple Comparisons and Multiple Tests (Text and Workbook Set) by Peter H. Westfall, Randall D. Tobias, Dror Romm, 2000, IBSN:1580258336.

- Adjusted P values are computed by SAS's PROC MULTTEST statement. However, the SAS documentation does not do a good job of explaining adjusted P values.

4.10.2.2.3 Bonferroni and Sidak methods

Bonferroni and Sidak tests in Prism

Prism can perform Bonferroni and Sidak multiple comparisons tests as part of several analyses:

- Following one-way ANOVA. This makes sense when you are comparing selected pairs of means, with the selection based on experimental design. Prism also lets you choose Bonferroni tests when comparing

every mean with every other mean. We don't recommend this. Instead, choose the [Tukey test](#)[136] if you want to compute confidence intervals for every comparison or the [Holm-Šídák test](#)[133] if you don't.

- Following two-way ANOVA. If you have three or more columns, and wish to compare means within each row (or three or more rows, and wish to compare means within each column), the situation is much like one-way ANOVA. The Bonferroni test is offered because it is easy to understand, but we don't recommend it. If you enter data into two columns, and wish to compare the two values at each row, then we recommend the Bonferroni method, because it can compute confidence intervals for each comparison. The alternative is the Holm-Šídák method, which has more power, but doesn't compute confidence intervals.

- As part of the analysis that [performs many t tests at once](#)[330].

- To [analyze a stack of P values](#)[642].

Key facts about the Bonferroni and Šídák methods

- The inputs to the Bonferroni and Šídák (the letter Š is pronounced "Sh") methods are a list of P values, so these methods can be used whenever you are doing multiple comparisons. They are not limited to use as followup tests to ANOVA.

- It only makes sense to use these methods in situations for which a specialized test has not been developed. For example, use the [Tukey method](#)[136] when comparing every mean with every other mean, and use [Dunnett's method](#)[136] to compare every mean with a control mean. But use Bonferroni or Šídák when you select a set of means to compare.

- The Bonferroni and Šídák methods can determine statistical significance, compute adjusted P value, and also compute confidence intervals.

- The Šídák method has a bit more power than the Bonferroni method.

- The Šídák method assumes that each comparison is independent of the others. If this assumption is independence cannot be supported, choose the Bonferroni method, which does not assume independence.

- The Bonferroni method is used more frequently, because it is easier to calculate (which doesn't matter when a computer does the work), easier to understand, and much easier to remember.

- Prism 5 and earlier offered the Bonferroni method, but not the Šídák method.

- The Bonferroni method is sometimes called the Bonferroni-Dunn method. And the Šídák method is sometimes called the Bonferroni-Šídák method.

References

1. H Abdi. The Bonferonni and Šidák Corrections for Multiple Comparisons. In N.J. Salkind (Ed.), 2007, Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage. pp. 103-107.

2. DJ Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Fifth edition, 2011, ISBN=978-7-1398-5801-1

4.10.2.2.4 The Holm-Sidak method

The Holm-Šídák test in Prism

Prism can perform the Holm multiple comparisons test as part of several analyses:

- Following one-way ANOVA. This makes sense when you are comparing selected pairs of means, with the selection based on experimental design. Prism also lets you choose Bonferroni tests when comparing every mean with every other mean. We don't recommend this. Instead, choose the [Tukey test][136] if you want to compute confidence intervals for every comparison or the [Holm-Šídák test][133] if you don't.

- Following two-way ANOVA. If you have three or more columns, and wish to compare means within each row (or three or more rows, and wish to compare means within each column), the situation is much like one-way ANOVA. The Bonferroni test is offered because it is easy to understand, but we don't recommend it. If you enter data into two columns, and wish to compare the two values at each row, then we recommend the Bonferroni method, because it can compute confidence intervals for each comparison. The alternative is the Holm-Šídák method, which has more power, but doesn't compute confidence intervals.

- As part of the analysis that [performs many t tests at once][330].

- To [analyze a stack of P values][642].

-

Key facts about the Holm test

- The input to the Holm method is a list of P values, so it is not restricted to use as a followup test to ANOVA.

- The Holm multiple comparison test can calculate [multiplicity adjusted P values][129], if you request them (2).

- The Holm multiple comparison test cannot compute confidence intervals for the difference between means.

- The method is also called the *Holm step-down method*.

- Although usually attributed to Holm, in fact this method was first described explicitly by Ryan (3) so is sometimes called the *Ryan-Holm step down method*.

- Holm's method has more power than the Bonferroni or Tukey methods (4). It has less power than the Newman-Keuls method, but that method is not recommended because it does not really control the familywise

significance level as it should, except for the special case of exactly three groups (4).

- The Tukey and Dunnett multiple comparisons tests are used only as followup tests to ANOVA, and they take into account the fact that the comparisons are intertwined. In contrast, Holm's method can be used to analyze any set of P values, and is not restricted to use as a followup test after ANOVA.

- The Šídák modification of the Holm test makes it a bit more powerful, especially when there are many comparisons.

- Note that Šídák's name is used as part of two distinct multiple comparisons methods, the Holm-Šídák test and the Šídák test related to the Bonferroni test[131].

- How it works.[378]

References:

1. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6 (2): 65–70.

2. Aickin, M. & Gensler, H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. American journal of public health 86, 726–728 (1996).

3. Ryan TA. Significance tests for proportions, variances, and other statistics. Psychol. Bull. 1960; 57: 318-28

4. MA Seaman, JR Levin and RC Serlin, New Developments in pairwise multiple comparisons: Some powerful and practicable procedures, Psychological Bulletin 110:577-586, 1991.

5. SA Glantz, *Primer of Biostatistics*, 2005, ISBN=978-0071435093.

4.10.2.2.5 Tukey and Dunnett methods

Tukey and Dunnett tests in Prism

Prism can perform either Tukey or Dunnett tests as part of one- and two-way ANOVA. Choose to assume a Gaussian distribution and to use a multiple comparison test that also reports confidence intervals. If you choose to compare every mean with every other mean, you'll be choosing a Tukey test. If you choose to compare every mean to a control mean, Prism will perform the Dunnett test.

Key facts about the Tukey and Dunnett tests

- The Tukey and Dunnet tests are only used as followup tests to ANOVA. They cannot be used to analyze a stack of P values.

- The Tukey test compares every mean with every other mean. Prism actually computes the Tukey-Kramer test, which allows for the possibility of unequal sample sizes.

- The Dunnett test compares every mean to a control mean.

- Both tests take into account the scatter of all the groups. This gives you a more precise value for scatter (Mean Square of Residuals) which is reflected in more degrees of freedom. When you compare mean A to mean C, the test compares the difference between means to the amount of scatter, quantified using information from all the groups, not just groups A and C. This gives the test more power to detect differences, and only makes sense when you accept the assumption that all the data are sampled from populations with the same standard deviation, even if the means are different.

- The results are a set of decisions: "statistically significant" or "not statistically significant". These decisions take into account multiple comparisons.

- It is possible to compute [multiplicity adjusted P values][129] for these tests.

- Both tests can compute a confidence interval for the difference between the two means. This confidence interval accounts for multiple comparisons. If you choose 95% intervals, then you can be 95% confident that all of the intervals contain the true population value.

- Prism reports the q ratio for each comparison. By historical tradition, this q ratio is computed differently for the two tests. For the Dunnett test, q is the difference between the two means (D) divided by the standard error of that difference (computed from all the data): q=D/SED. For the Tukey test, q=sqrt(2)*D/SED. Because of these different definitions, the two q values cannot be usefully compared. The only reason to look at these q ratios is to compare Prism's results with texts or other programs. Note that this use of the variable q is distinct from the use of q when using the FDR approach.

- Different tables (or algorithms) are used for the Tukey and Dunnett tests to determine whether or not a q value is large enough for a difference to be declared to be statistically significant. This calculation depends on the value of q, the number of groups being compared, and the number of degrees of freedom.

- Read the details of how these (and other) tests are calculated here. We use the original single step Dunnett method, not the newer step-up or step-down methods.

4.10.2.2.6 Dunn's multiple comparisons after nonparametric ANOVA

If you choose nonparametric ANOVA, the Multiple Comparisons tab lets you choose:

- No multiple comparisons

- Compare the mean rank of each group with the mean rank of every other group

- Compare the mean rank of each group to the mean rank of a control group you specify

- Compare selected pairs of columns.

In all cases, you won't have any choice of method. Prism will use Dunn's method[382] at a significance level alpha you choose (usually 0.05).

4.10.2.2.7 Newman-Keuls method

Prism offers the Newman-Keuls test (sometimes called the Student-Newman-Keuls test) for historical reasons, but we don't recommend ever using it.

This test is a powerful way to compare all pairs of means, reporting statistical significance but not confidence intervals or multiplicity adjusted P values. The problem is it is too powerful, and so it does not maintain the family-wise error rate at the specified level(2). In some cases, the chance of a Type I error can be greater than the alpha level you specified.

1. MA Seaman, JR Levin and RC Serlin, Psychological Bulletin 110:577-586, 1991.

4.10.2.3 Approach 3: Control the False Discovery Rate (FDR)

.

4.10.2.3.1 What it means to control the FDR

Defining the FDR

Here again is the table from the <u>previous page</u>[128] predicting the results from many comparisons. The only difference, is that I changed the term "statistically signfiicant" to "discovery" because that is more commonly used with the false discovery rate approach.

	"Discovery"	"Not a discovery"	Total
No difference. **Null hypothesis true**	A	B	A+B
A difference truly exists	C	D	C+D
Total	A+C	B+D	A+B+C+D

The top row represents the results of comparisons where the null hypothesis is true -- the treatment really doesn't work. Nonetheless, some comparisons will mistakenly yield a P value small enough so that comparison is deemed a "discovery".

The second row shows the results of comparisons where there truly is a difference. Even so, you won't get a P value small enough to call that finding a "discovery" in every experiment.

A, B, C and D represent numbers of comparisons, so the sum of A+B+C+D equals the total number of comparisons you are making.

Of course, you can only make this table in theory. If you collected actual data, you'd never know if the null hypothesis is true or not, so could not assign results to row 1 or row 2.

The usual approach to statistical significance and multiple comparisons asks the question:

If the null hypothesis is true what is the chance of getting "statistically significant" results?

The False Discovery Rate (FDR) answers a different question:

If a comparison is a "discovery", what is the chance that the null hypothesis is true?

In the table, above the False Discovery rate is the ratio A/(A+C).

Controlling the FDR with Q

When dealing with multiple comparisons, you may want to set a FDR value (usually called Q) and then use that value when deciding which comparisons are "discoveries" and which are not with the intention that the actual false discovery rate is no higher than Q.

If you are only making a single comparison, you can't do this without defining the prior odds and using Bayesian reasoning[87] . But if you have many comparisons, simple methods let you control the FDR approximately. You can set the desired value of Q, and the FDR method will decide if each P value is small enough to be designated a "discovery". If you set Q to 10%, you expect about 90% of the discoveries (in the long run) to truly reflect actual differences, while no more than 10% are false positives. In other words, you expect A/(A+C) to equal 10% (the value you set for Q).

q values or adjusted P values

There are two ways to think about the false discovery rate.

• You enter a value for Q (note the uppercase; the desired false discovery rate) and, using that definition, the program tells you which

comparisons are discoveries and which are not. In Prism, you enter Q as a percentage.

- For each comparison, the program computes a q value (note the lower case). This value is also called an adjusted P value. The way to think about this value is as follows. If you had set Q above to this value, then the comparison you are looking at now would be right at the border of being a discovery or not. Prism reports q as a decimal fraction.

4.10.2.3.2 Key facts about controlling the FDR

Prism uses the concept of False Discovery Rate as part of our method to define outliers ([from a stack of values](#)[161], or [during nonlinear regression](#)). Prism also can use the FDR method when calculating [many t tests at once](#)[330], when [analyzing a stack of P values computed elsewhere](#)[642], and as a multiple comparisons method following one-, two, or three-way ANOVA.

Key facts about the False Discovery Rate approach

- This approach first computes a P value for each comparison. When used as a followup to ANOVA, the comparisons are done using the Fisher Least Significant Different approach (which by itself does not correct for multiple comparisons but does pool the variances to increase the number of degrees of freedom). When used to analyze a set of t tests, each t test is first computed individually. When analyzing a set of P values, of course you enter these P values directly.

- The goal is explained [here](#)[138]. You enter Q, the desired false discovery rate (as a percentage), and Prism then tells you which P values are low enough to be called a "discovery", with the goal of ensuring that no more than Q% of those "discoveries" are actually false positives.

- Prism let's you choose [one of three algorithms](#)[386] for deciding which P values are small enough to be a "discovery". The Benjamini and Hochberg method was developed first so is more standard. The Benjamani, Krieger and Yekutieli method have more power, so is

preferred. The method of Benjamini & Yekutieli makes fewer assumptions, but has much less power.

- This FDR approach does not use the concept or phrase "statistically significant" when a P value is small, but instead uses the term "discovery". (Some authors use terminology differently.)

- The FDR approach cannot compute confidence intervals to accompany each comparison.

- Q (note the upper case) is a value you enter as the desired FDR. Prism also computes q (lower case) for each comparison. This value q is the value of Q at which this particular comparison would be right on the border of being classified as a discovery or not. The value q depends not only on that one comparison, but on the number of comparisons in the family and the distribution of P values.

- The q values Prism reports are FDR-*adjusted* p values, not [FDR-*corrected* P values](). This is a subtle distinction.

- If all the null hypotheses are true, there will be only a Q% chance that you find one or more discoveries (where Q is the false discovery rate you chose).

- If all the P values are less than your chosen value of Q (correcting for the fact that P values are fractions and Q is a percentage), then all the comparisons will be flagged as discoveries. (This rule is not true when you choose the method of Benjamini & Yekutieli).

- If all the P values are greater than your chosen value of Q, then no comparison will be flagged as a discovery.

- The q values are generally larger than the corresponding P value. The exception is the q value for the comparison with the largest P value can have a q value equal to the P value.

- The value of q is set by the P value for that comparison as well as the other P values and the number of comparisons. The value you enter for Q does not impact the computation of q.

- The algorithms in Prism control the FDR, not the pFDR (which won't be explained here).

- The q values determined by these methods tend to be higher (and are never lower) than the adjusted P values computed when using the usual multiple comparisons methods (Bonferroni, etc.).

- Great nonmathematical review: Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology*, *67*(8), 850–857.

4.10.2.3.3 Pros and cons of the three methods used to control the FDR

Prism offers three methods to control the FDR that differ in power, simplicity and assumptions.

Original method of Benjamini and Hochberg (1).

This method was developed first, and is still the standard. It assumes that "test statistics are independent or positive dependent". This seems to mean that while it is OK that some of the comparisons are positively correlated (if one is low, the others tend to be low), the method does not work well if some comparisons are negatively correlated (if one is low, others tend to be high).

We offer this method because it is the standard.

Two-stage step-up method of Benjamini, Krieger and Yekutieli (2).

This method relies on the same assumption as the Benjamini and Hochberg method, but it is a more clever method. It first examines the distribution of P values to estimate the fraction of the null hypotheses that are actually true. It then uses this information to get more power when deciding when a P value is low enough to be called a discovery.

The only downside of this method is that the math is a bit more complicated, so it is harder to use if you were doing the calculations by hand.

The improved adaptive method of Benjamini, Krieger and Yekutieli has more power than the Benjamini and Hochberg method, while making the same assumptions, so we recommend it.

The paper that describes this metnod (2) describes several methods. Prism uses the method defined in section 6 , the two-stage linear step-up procedure.

Corrected method of Benjamini & Yakutieli (3)

This method requires no assumptions about how the various comparisons correlate with each other. But the price of this is that is has less power, so identifies fewer comparisons as being a discovery. Another way of saying this is that the method is very conservative.

References

1. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289–300 (1995).

2. Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93, 491–507 (2006).

3. Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165–1188.

4.11 Testing for equivalence

4.11.1 Key concepts: Equivalence

Why test for equivalence?

Usually statistical tests are used to look for differences. But sometimes your goal is to prove that two sets of data are equivalent. A conclusion of "no statistically significant difference" is not enough to conclude that two treatments are equivalent. You've really need to rethink how the test is set up.

In most experimental situations, your goal is to show that one treatment is better than another. But in some situations, your goal is just the opposite -- to prove that one treatment is indistinguishable from another, that any difference is of no practical consequence. This can either be the entire goal of the study (for example to show that a new formulation of a drug works as well as the usual formulation) or it can just be the goal for analysis of a control experiment to prove that a system is working as expected, before moving on to asking the scientifically interesting questions.

Standard statistical tests cannot be used to test for equivalence

Standard statistical tests cannot be used to test for equivalence.

A conclusion of "no statistically significant difference" between treatments, simply means that you don't have strong enough evidence to persuade you that the two treatments lead to different outcomes. That is not the same as saying that the two outcomes are equivalent.

A conclusion that the difference is "statistically significant" means you have strong evidence that the difference is not zero, but you don't know whether the difference is large enough to rule out the conclusion that the two treatments are functionally equivalent.

You must decide how large a difference has to be to in order to be considered scientifically or clinically relevant.

In any experiment, you expect to almost always see some difference in outcome when you apply two treatments. So the question is not whether the two treatments lead to *exactly* the same outcome. Rather, the question is whether the outcomes are *close enough* to be clinically or scientifically indistinguishable. How close is that? There is no way to answer that question generally. The answer depends on the scientific or clinical context of your experiment.

To ask questions about equivalence, you first have to define a range of treatment effects that you consider to be scientifically or clinically trivial. This is an important decision that must be made totally on scientific or clinical grounds.

You can test for equivalence using either a confidence interval or P value approach

Statistical methods have been developed for testing for equivalence. You can use either a [confidence interval or a P value approach](#)[145].

4.11.2 Testing for equivalence with confidence intervals or P values

Before you can test for equivalence, you first have to define a range of treatment effects that you consider to be scientifically or clinically trivial. You must set this range based on scientific or clinical judgment -- statistical analyses can't help.

If the treatment effect you observed is outside this zone of scientific or clinical indifference, then clearly you can't conclude the treatments are equivalent.

If the treatment effect does lie within the zone of clinical or scientific indifference, then you can ask whether the data are tight enough to make a strong conclusion that the treatments are equivalent.

Testing for equivalence with confidence intervals.

The figure below shows the logic of how to test for equivalence with confidence intervals. The horizontal axis shows the absolute value of the treatment effect (difference between mean responses). The filled circles show the observed effect, which is within the zone of indifference. The horizontal error bars show the one-sided 95% confidence intervals, which show the largest treatment effect consistent with the data (with 95% confidence).
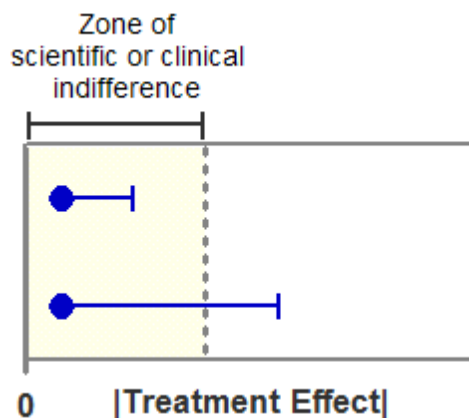
In the experiment shown on top, even the limit of the confidence interval lies within the zone of indifference. You can conclude (with 95% confidence) that the two treatments are equivalent.

In the experiment shown on the bottom, the confidence interval extends beyond the zone of indifference. Therefore, you cannot conclude that the treatments are equivalent. You also cannot conclude that the treatments are not equivalent, as the observed treatment is inside the zone of indifference. With data like these, you simply cannot make any conclusion about equivalence.

Testing for equivalence using statistical hypothesis testing

Thinking about statistical equivalence with confidence intervals (above) is pretty straightforward. Applying the ideas of statistical hypothesis testing to equivalence is much trickier.

Statistical hypothesis testing starts with a null hypothesis, and then asks if you have enough evidence to reject that null hypothesis. When you are looking for a difference, the null hypothesis is that there is no difference. With equivalence testing, we are looking for evidence that two treatments are equivalent. So the "null" hypothesis, in this case, is that the treatments are not equivalent, but rather that the difference is just barely large enough to be outside the zone of scientific or clinical indifference.

In the figure above, define the null hypothesis to be that the true effect equals the effect denoted by the dotted line. Then ask: If that null hypothesis were true, what is the chance (given sample size and variability) of observing an effect as small or smaller than observed. If the P value is small, you reject the null hypothesis of nonequivalence, so conclude that the treatments are equivalent. If the P value is large, then the data are consistent with the null hypothesis of nonequivalent effects.

Since you only care about the chance of obtaining an effect so much lower than the null hypothesis (and wouldn't do the test if the difference were higher), you use a one-tail P value.

The graph above is plotted with the absolute value of the effect on the horizontal axis. If you plotted the treatment effect itself, you would have two dotted lines, symmetric around the 0 point, one showing a positive treatment effect and the other showing a negative treatment effect. You would then have two different null hypotheses, each tested with a one-tail test. You'll see this referred to as *Two One-Sided Tests Procedure (1, 2).*

The two approaches are equivalent

Of course, using the 95% confidence interval approach (using one-sided 95% confidence intervals) and the hypothesis testing approach (using one-sided 0.05 threshold for significance are completely equivalent, so always give the same conclusion. The confidence interval seems to me to be far more straightforward to understand.

Testing for equivalence with Prism

Prism does not have any built-in tests for equivalence. But you can use Prism to do the calculations:

1. Compare the two groups with a t test (paired or unpaired, depending on experimental design).

2. Check the option to create **90%** confidence intervals. That's right 90%, not 95%.

3. If the entire range of the **90%** confidence interval lies within the zone of indifference that you defined, then you can conclude with **95%** confidence that the two treatments are equivalent.

> Confused about the switch from 90% confidence intervals to conclusions with 95% certainty? Good. That means you are paying attention. It **is** confusing!

References

1. D.J. Schuirmann, A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability, J. Pharmacokinetics and pharmacodynamics, 115: 1567, 1987.

2. S. Wellek, Testing Statistical Hypotheses of Equivalence, Chapman and Hall/CRCm, 2010, ISBN: 978-1439808184.

4.12   Nonparametric tests

4.12.1   Key concepts: Nonparametric tests

ANOVA, t tests, and many statistical tests assume that you have sampled data from populations that follow a [Gaussian](#)[37] bell-shaped distribution.

Biological data never follow a Gaussian distribution precisely, because a Gaussian distribution extends infinitely in both directions, and so it includes both infinitely low negative numbers and infinitely high positive numbers! But many kinds of biological data follow a bell-shaped distribution that is approximately Gaussian. Because ANOVA, t tests, and other statistical tests work well even if the distribution is only approximately Gaussian (especially with large samples), these tests are used routinely in many fields of science.

An alternative approach does not assume that data follow a Gaussian distribution. In this approach, values are ranked from low to high, and the analyses are based on the distribution of ranks. These tests, called *nonparametric* tests, are appealing because they make fewer assumptions about the distribution of the data.

4.12.2   Advice: Don't automate the decision to use a nonparametric test

Don't use this approach:

~~First perform a normality test. If the P value is low, demonstrating that the data do not follow a Gaussian distribution, choose a nonparametric test. Otherwise choose a conventional test.~~

Prism does not use this approach, because the choice of parametric vs. nonparametric is more complicated than that.

- Often, the analysis will be one of a series of experiments. Since you want to analyze all the experiments the same way, you cannot rely on the results from a single normality test.

- Many biological variables follow [lognormal distributions](#)[55]. If your data are sampled from a lognormal distribution, the best way to analyze the data is to first transform to logarithms and then analyze the logs. It would be a mistake to jump right to nonparametric tests, without considering transforming.

- Other transforms can also be useful (reciprocal) depending on the distribution of the data.

- Data can fail a normality test because of the presence of an <u>outlier</u>[155]. Removing that outlier can restore normality.

- The decision of whether to use a parametric or nonparametric test is most important with small data sets (since the power of nonparametric tests is so low). But with small data sets, <u>normality tests</u>[245] have little power to detect nongaussian distributions, so an automatic approach would give you false confidence.

- With large data sets, normality tests can be too sensitive. A low P value from a normality test tells you that there is strong evidence that the data are not sampled from an ideal Gaussian distribution. But you already know that, as almost no scientifically relevant variables form an ideal Gaussian distribution. What you want to know is whether the distribution deviates enough from the Gaussian ideal to invalidate conventional statistical tests (that assume a Gaussian distribution). A normality test does not answer this question. With large data sets, trivial deviations from the idea can lead to a small P value.

The decision of when to use a parametric test and when to use a nonparametric test is a difficult one, requiring thinking and perspective. This decision should not be automated.

4.12.3 The power of nonparametric tests

Why not always use nonparametric tests? You avoid assuming that the data are sampled from a Gaussian distribution -- an assumption that is hard to be sure of. The problem is that nonparametric tests have lower <u>power</u>[97] than do standard tests. How much less power? The answer depends on sample size.

This is best understood by example. Here are some sample data, comparing a measurement in two groups, each with three subjects.

Control	Treated
3.4	1234.5
3.7	1335.7
3.5	1334.8

When you see those values, it seems obvious that the treatment drastically increases the value being measured.

But let's analyze these data with the [Mann-Whitney test](#)[304] (nonparametric test to compare two unmatched groups). This test only sees ranks. So you enter the data above into Prism, but the Mann Whitney calculations only see the ranks:

Control	Treated
1	4
3	6
2	5

The Mann-Whitney test then asks if the ranks were randomly shuffled between control and treated, what is the chance of obtaining the three lowest ranks in one group and the three highest ranks in the other group. The nonparametric test only looks at rank, ignoring the fact that the treated values aren't just higher, but are a whole lot higher. The answer, the two-tail P value, is 0.10. Using the traditional significance level of 5%, these results are not significantly different. This example shows that with N=3 in each group, the Mann-Whitney test can never obtain a P value less than 0.05. In other words, with three subjects in each group and the conventional definition of 'significance', the Mann-Whitney test has zero power.

With large samples in contrast, the Mann-Whitney test has almost as much power as the t test. To learn more about the relative power of nonparametric and conventional tests with large sample size, look up the term "Asymptotic Relative Efficiency" in an advanced statistics book.

4.12.4 Nonparametric tests with small and large samples

Small samples

Your decision to choose a parametric or nonparametric test matters the most when samples are small (say less than a dozen values).

If you choose a parametric test and your data do not come from a Gaussian distribution, the results won't be very meaningful. Parametric tests are not very robust to deviations from a Gaussian distribution when the samples are tiny.

If you choose a nonparametric test, but actually do have Gaussian data, you are likely to get a P value that is too large, as nonparametric tests have less power than parametric tests, and the difference is noticeable with tiny samples.

Unfortunately, normality tests have little power to detect whether or not a sample comes from a Gaussian population when the sample is tiny. Small samples simply don't contain enough information to let you make reliable inferences about the shape of the distribution in the entire population.

Large samples

The decision to choose a parametric or nonparametric test matters less with huge samples (say greater than 100 or so).

If you choose a parametric test and your data are not really Gaussian, you haven't lost much as the parametric tests are robust to violation of the Gaussian assumption, especially if the sample sizes are equal (or nearly so).

If you choose a nonparametric test, but actually do have Gaussian data, you haven't lost much as nonparametric tests have nearly as much power as parametric tests when the sample size is large.

Normality tests work well with large samples, which contain enough data to let you make reliable inferences about the shape of the distribution of the population from which the data were drawn. But normality tests don't answer the question you care about. What you want to know is whether the distribution differs enough from Gaussian to cast doubt on the usefulness of parametric tests. But normality tests answer a different question. Normality tests ask the question of whether there is evidence that the distribution differs from Gaussian. But with huge samples, normality testing will detect tiny deviations from Gaussian, differences small enough so they shouldn't sway the decision about parametric vs. nonparametric testing.

Summary

	Large samples (>100 or so)	Small samples (<12 or so)
Parametric tests on nongaussian data	OK. Tests are robust.	Misleading. Not robust.

Nonparametric tests on Gaussian data	OK. Tests have good power.	Misleading. Too little power.
Usefulness of normality testing	A bit useful.	Not very useful.

4.12.5   Advice: When to choose a nonparametric test

Choosing when to use a nonparametric test is not straightforward. Here are some considerations:

- **Off-scale values.** With some kinds of experiments, one, or a few, values may be "off scale" -- too high or too low to measure. Even if the population is Gaussian, it is impossible to analyze these data with a t test or ANOVA. If you exclude these off scale values entirely, you will bias the results. If you estimate the value, the results of the t test depend heavily on your estimate. The solution is to use a nonparametric test. Assign an arbitrary low value to values that are too low to measure, and an arbitrary high value to values too high to measure. Since the nonparametric tests only analyze ranks, it will not matter that you don't know one (or a few) of the values exactly, so long as the numbers you entered gave those values the correct rank.

- **Transforming can turn a nongaussian distribution into a Gaussian distribution**. If you are sure the data do not follow a Gaussian distribution, pause before choosing a nonparametric test. Instead, consider transforming the data, perhaps using logarithms or reciprocals. Often a simple transformation will convert non-Gaussian data to a Gaussian distribution. Then analyze the transformed values with a conventional test.

- **Noncontinuous data.** The outcome is a rank or score with only a few categories. Clearly the population is far from Gaussian in these cases. The problem with using nonparametric tests is that so many values will tie for the same rank. Nonparametric tests have special corrections built-in to deal with tied ranks, but I am not sure how well those work when there are lots of tied ranks. An alternative would be to do a [chi-square test](#)[545].

- **Small samples.** If you have tiny samples (a few subjects in each group), the nonparametric tests have little or no power[149] to find a significant difference.

- **Normality tests** should not be used[148] to automatically decide whether or not to use a nonparametric test. But they can help you make the decision.

- You really should choose your statistical test as part of the experimental design. If you try this test, then that test, until you get a result you like, you are likely to be mislead.

4.12.6 Lingo: The term "nonparametric"

The term nonparametric is used inconsistently.

Nonparametric method or nonparametric data?

The term *nonparametric* characterizes an analysis method. A statistical test can be nonparametric or not, although the distinction is not as crisp as you'd guess.

It makes no sense to describe data as being nonparametric, and the phrase "nonparametric data" should never ever be used. The term *nonparametric* simply does not describe data, or distributions of data. That term should only be used to describe the method used to analyze data.

Which methods are nonparametric?

Methods that analyze ranks are uniformly called nonparametric. These tests are all named after their inventors, including: Mann-Whitney, Wilcoxon, Kruskal-Wallis, Friedman, and Spearman.

Beyond that, the definition gets slippery.

What about modern statistical methods including randomization, resampling and bootstrapping? These methods do not assume sampling from a Gaussian distribution. But they analyze the actual data, not the ranks. Are these methods nonparametric? Wilcox and Manly have each written texts about modern methods, but they do not refer to these methods as "nonparametric". Four texts of nonparametric statistics (by Conover, Gibbons, Lehmann, and Daniel) don't mention randomization,

resampling or bootstrapping at all, but the texts by Hollander and Wasserman do.

What about chi-square test, and Fisher's exact test? Are they nonparametric? Daniel and Gibbons include a chapter on these tests their texts of nonparametric statistics, but Lehmann and Hollander do not.

What about survival data? Are the methods used to create a survival curve (Kaplan-Meier) and to compare survival curves (logrank or Mantel-Haenszel) nonparametric? Hollander includes survival data in his text of nonparametric statistics, but the other texts of nonparametric statistics don't mention survival data at all. I think everyone would agree that fancier methods of analyzing survival curves (which involve fitting the data to a model) are not nonparametric.

References

W. J. Conover. *Practical Nonparametric Statistics* (Wiley Series in Probability and Statistics)ISBN:0471160687.

Wayne W. Daniel, *Applied Nonparametric Statistics,* IBSN:0534919766.

Jean Dickinson Gibbons and Subhabrata Chakraborti, *Nonparametric Statistical Inference*, Fourth Edition: Revised and Expanded (Statistics: A Series of Textbooks and Monographs), IBSN:0824740521.

Myles Hollander and Douglas A. Wolfe, *Nonparametric Statistical Methods*, 2nd Edition, IBSN:0471190454.

Erich L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, IBSN:013997735X.

Bryan F.J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Second Edition, IBSN:0412721309.

Larry Wasserman, *All of Nonparametric Statistics* (Springer Texts in Statistics), IBSN:1441920447.

Rand R. Wilcox, *Applying Contemporary Statistical Techniques,* IBSN:0127515410.

4.13   Outliers

When analyzing data, you'll sometimes find that one value is far from the others. Such a value is called an *outlier*, a term that is usually not defined rigorously. This section discusses the basic ideas of identifying outliers. Look elsewhere to learn how to identify outliers in Prism [from a column of data](#)[248], or [while fitting a curve with nonlinear regression](#).

4.13.1   An overview of outliers

What is an outlier?

When analyzing data, you'll sometimes find that one value is far from the others. Such a value is called an *outlier*, a term that is usually not defined rigorously.

Approach to thinking about outliers

When you encounter an outlier, you may be tempted to delete it from the analyses. First, ask yourself these questions:

- Was the value entered into the computer correctly? If there was an error in data entry, fix it.

- Were there any experimental problems with that value? For example, if you noted that one tube looked funny, you can use that as justification to exclude the value resulting from that tube without needing to perform any calculations.

- Could the outlier be caused by biological diversity? If each value comes from a different person or animal, the outlier may be a correct value. It is an outlier not because of an experimental mistake, but rather because that individual may be different from the others. This may be the most exciting finding in your data!

If you answered "no" to all three questions, you are left with two possibilities.

- The outlier was due to chance. In this case, you should keep the value in your analyses. The value came from the same distribution as the other values, so should be included.

- The outlier was due to a mistake: bad pipetting, voltage spike, holes in filters, etc. Since including an erroneous value in your analyses will give invalid results, you should remove it. In other words, the value comes from a different population than the other values, and is misleading.

The problem, of course, is that you can never be sure which of these possibilities is correct.

Robust methods

Some statistical tests are designed so that the results are not altered much by the presence of one or a few outliers. Such tests are said to be *robust*. When you use a robust method, there is less reason to want to exclude outliers.

Most nonparametric tests compare the distribution of ranks. This makes the test robust because the largest value has the largest rank, but it doesn't matter how large that value is.

Other tests are robust to outliers because rather than assuming a Gaussian distribution, they assume a much wider distribution where outliers are more common (so have less impact).

4.13.2 Advice: Beware of identifying outliers manually

A common practice is to visually inspect the data, and remove outliers by hand. The problem with this approach is that it is arbitrary. It is too easy to keep points that help the data reach the conclusion you want, and to remove points that prevent the data from reaching the conclusion you want.
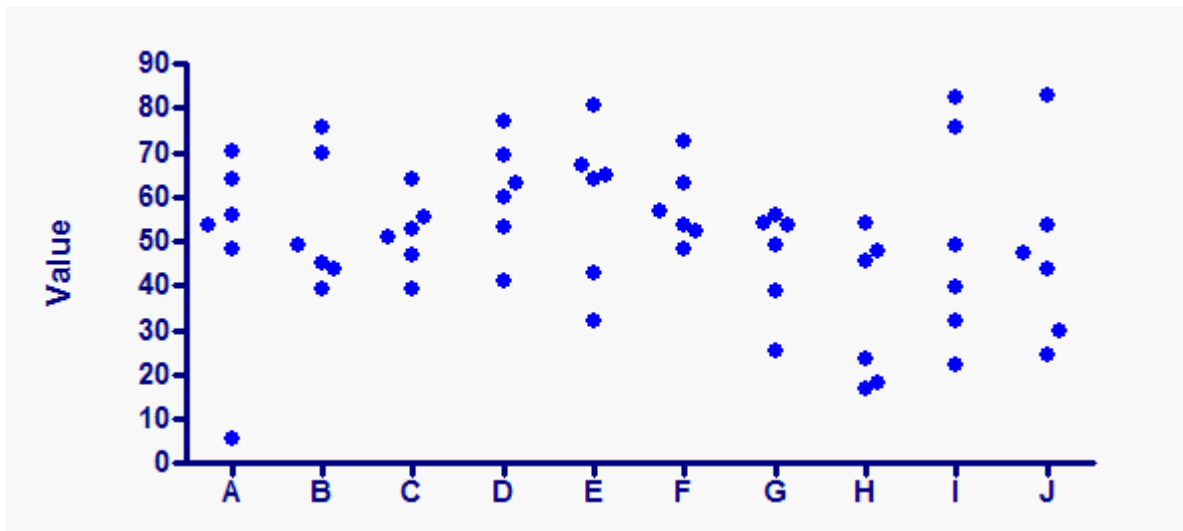
The graph above was created via simulation. The values in all ten data sets are randomly sampled from a Gaussian distribution with a mean of 50 and a SD of 15. But most people would conclude that the lowest value in data set A is an outlier. Maybe also the high value in data set J. Most people are unable to appreciate random variation, and tend to find 'outliers' too often.

4.13.3 Advice: Beware of lognormal distributions

The Grubbs' and ROUT outlier tests are both based on the assumption that the data, except the potential outlier(s), are sampled from a Gaussian distribution.

But what if the underlying distribution is not Gaussian? Then the outlier tests are misleading. A common situation is sampling from a lognormal distribution.

The graph below shows four data sets sampled from lognormal distributions.

Raw data
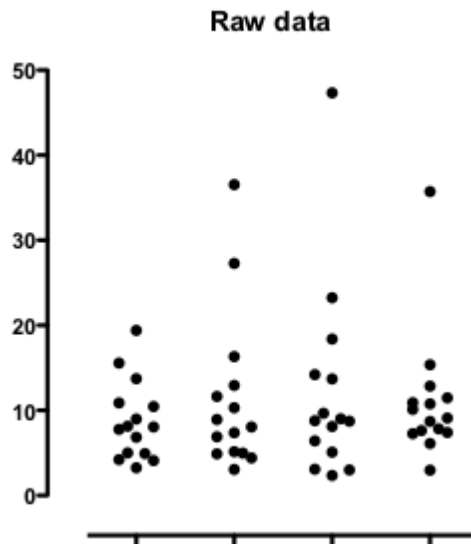
Three of those data sets seem to include an outlier, and indeed Grubbs' outlier test identified outliers in three of the data sets.

But these data are not sampled from a Gaussian distribution with an outlier. Rather they are sampled from a lognormal distribution. Transform all the values to their logarithms, and the distribution becomes Gaussian:

After transforming to logarithms
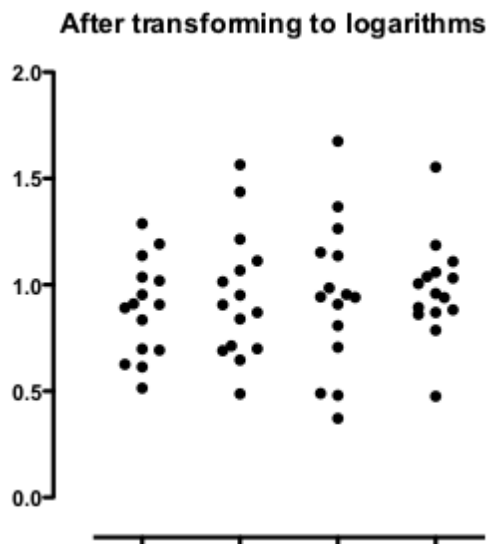
The apparent outliers are gone. Grubbs' test finds no outliers. The extreme points only appeared to be outliers because extremely large values are common in a lognormal distribution but are rare in a Gaussian distribution. If you don't realize the distribution was lognormal, an outlier test would be very misleading.

4.13.4 How it works: Grubb's test

What can an outlier tests do?

No mathematical calculation can tell you for sure whether the outlier came from the same, or a different, population than the others. Statistical calculations, however, can answer this question:

> If the values really were all sampled from a Gaussian distribution, what is the chance that you would find one value as far from the others as you observed?

If this probability is small, then you will conclude that the outlier is not from the same distribution as the other values. Assuming you answered no to all three questions above, you have justification to exclude it from your analysis.

Statisticians have devised several methods for detecting outliers. All the methods first quantify how far the outlier is from the other values. This can be the difference between the outlier and the mean of all points, the difference between the outlier and the mean of the remaining values, or the difference between the outlier and the next closest value. Next, standardize this value by dividing by some measure of scatter, such as the SD of all values, the SD of the remaining values, or the range of the data. Finally, compute a P value answering this question: If all the values were really sampled from a Gaussian population, what is the chance of randomly obtaining an outlier so far from the other values? If the P value is small, you conclude that the deviation of the outlier from the other values is statistically significant, and most likely from a different population.

How Grubbs's test works

Grubbs' test is one of the most popular ways to define outliers, and is quite easy to understand. This method is also called the ESD method (extreme studentized deviate).

The first step is to quantify how far the outlier is from the others. Calculate the ratio Z as the difference between the outlier and the mean divided by the SD. If Z is large, the value is far from the others. Note that you calculate the mean and SD from *all* values, including the outlier.

$$Z = \frac{|\text{mean} - \text{value}|}{SD}$$

You'll sometimes see this value referred to as G instead of Z.

Since 5% of the values in a Gaussian population are more than 1.96 standard deviations from the mean, your first thought might be to conclude that the outlier comes from a different population if Z is greater than 1.96. This approach only works if you know the population mean and SD from other data. Although this is rarely the case in experimental science, it is often the case in quality control. You know the overall mean and SD from historical data, and want to know whether the latest value matches the others. This is the basis for quality control charts.

When analyzing experimental data, you don't know the SD of the population. Instead, you calculate the SD from the data. The presence of an outlier increases the calculated SD. Since the presence of an outlier increases both the numerator (difference between the value and the mean) and denominator (SD of all values), Z can not get as large as you may expect. For example, if N=3, Z cannot be larger than 1.155 for any set of values. More generally, with a sample of N observations, Z can never get larger than:

$$(N - 1)/\sqrt{N}$$

Grubbs and others have tabulated critical values for Z which have been tabulated. The critical value increases with sample size, as expected. If your calculated value of Z is greater than the critical value in the table, then the P value is less than 0.05.

Note that the Grubbs' test only tests the most extreme value in the sample. If it isn't obvious which value is most extreme, calculate Z for all values, but only calculate a P value for Grubbs' test from the largest value of Z.

Prism can compute Grubbs' test with as few as three values in a data set.

How to interpret the P value

If the P value is less than 0.05, it means that there is less than a 5% chance that you'd encounter an outlier so far from the others (in either direction) by chance alone, if all the data were really sampled from a single Gaussian distribution.

Note that the 5% probability (or whatever value of alpha you choose) applies to the entire data set. If your dataset has 100 values, and all are sampled from a Gaussian distribution, there is a 5% chance that the largest (or smallest) value will be declared to be an outlier by Grubbs' test. If you performed outliers tests on lots of data sets, you'd expect this kind of mistake in 5% of data sets.

Don't get confused and think that the 5% applies to each data point. If there are 100 values in the data set all drawn from a Gaussian distribution, there is a 5% chance that Grubbs test will identify the value furthest from the mean as an outlier. This is different than concluding (mistakenly) that you expect 5 of the values (5% of the total) to be mistakenly declared to be outliers.

References

- B Iglewicz and DC Hoaglin. How to Detect and Handle Outliers (Asqc Basic References in Quality Control, Vol 16) Amer Society for Quality Control, 1993.

- V Barnett, T Lewis, V Rothamsted. Outliers in Statistical Data (Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics) John Wiley & Sons, 1994.

4.13.5 How it works: ROUT method

The basics of ROUT

The ROUT method was developed as a method to identify outliers from nonlinear regression. [Learn more about the ROUT method.](#)

Briefly, it first fits a model to the data using a robust method where outliers have little impact. Then it uses a new outlier detection method, based on the false discovery rate, to decide which points are far enough from the prediction of the model to be called outliers.

When you ask Prism to detect outliers in a stack of column data, it simply adapts this method. It considers the values you entered to be Y values, and fits the model Y= M, where M is a robust mean. [If you want to do this with Prism's nonlinear regression analysis, you'd need to assign arbirtrary X values to each row, and then fit to the model Y = X*0 + M.)

This method can detect any number of outliers (up to 30% of the sample size).

Prism can perform the ROUT test with as few as three values in a data set.

What is Q?

The ROUT method is based on the False Discovery Rate (FDR), so you specify Q, which is the maximum desired FDR. The interpretation of Q depends on whether there are any outliers in the data set.

When there are no outliers (and the distribution is entirely Gaussian), Q is very similar to alpha. Assuming all data come from a Gaussian distribution, Q is the chance of (falsely) identifying one or more outliers..

When there are outliers in the data, Q is the maximum desired false discovery rate. If you set Q to 1%, then you are aiming for no more than 1% of the identified outliers to be false (are in fact just the tail of a Gaussian distribution) and at least 99% to be actual outliers (from a different distribution).

Comparing ROUT to Grubbs' method

I performed simulations[165] to compare the Grubbs' and ROUT methods of detecting outliers. Briefly, the data were sampled from a Gaussian distribution. In most cases, outliers (drawn from a uniform distribution with specified limits) were added. Each experimental design was simulated 25,000 times, and I tabulated the number of simulations with zero, one, two, or more than two outliers.

When there are no outliers, the ROUT and Grubbs' tests perform almost identically. The value of Q specified for the ROUT method is equivalent to the value of alpha you set for the Grubbs' test.

When there is a single outlier, the Grubb's test is slightly better able to detect it. The ROUT method has both more false negatives and false

positives. In other words, it is slightly more likely to miss the outlier, and is also more likely to find two outliers even when the simulation only included one. This is not so surprising, as Grubbs' test was designed to detect a single outlier. While the difference between the two methods is clear, it is not substantial.

When there are two outliers in a small data set, the ROUT test does a much better job. The iterative Grubbs' test is subject to masking, while the ROUT test is not. Whether or not masking is an issue depends on how large the sample is and how far the outliers are from the mean of the other values. In situations where masking is a real possibility, the ROUT test works *much* better than Grubbs' test. For example, when n=10 with two outliers, the Grubbs test never found both outliers and missed both in 98.8% of the simulations (in the remaining 1.2% of simulations, the Grubbs' test found one of the two outliers). In contrast, the ROUT method identified both outliers in 92.8% of those simulations, and missed both in only 6% of simulations.

Summary:

- Grubbs' is slightly better than the ROUT method for the task it was designed for: Detecting a single outlier from a Gaussian distribution.

- The ROUT method is much better than the iterative Grubbs' test at detecting two outliers in some situations.

Reference

Motulsky HM and Brown RE, Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate, BMC Bioinformatics 2006, 7:123. Download from http://www.biomedcentral.com/1471-2105/7/123.

4.13.6 The problem of masking

The figure below shows two data sets, identical except for one data point. Clearly, the data set on the right has two outliers, and the one on the left has only one. This conclusion is not at all subtle.

Outliers detected by Grubbs test (α=0.05) are shown in red.
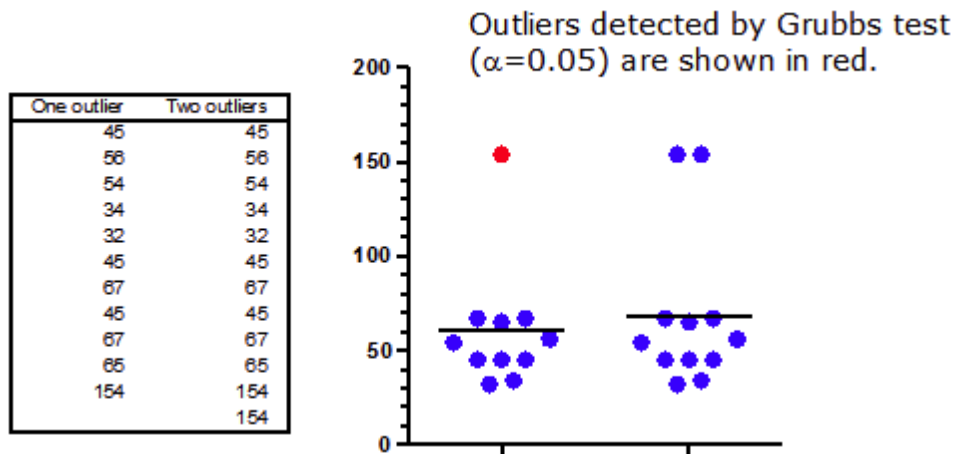
([Download the Prism file](#).)

The results of Grubbs' outlier test are surprising. That test (with alpha set to 5%, but the same results are obtained with alpha set to 1%) does identify the outlier in the data set on the left. No surprise there. But Grubbs' test doesn't find any outliers in the data set on the right. The presence of the second outlier prevents the outlier test from finding the first one. This is called *masking*.

[Grubbs' outlier test](#)[159] computes a ratio Z by first calculating the difference between the possible outlier and the mean, and then dividing that difference by the standard deviation. If Z is large enough (considering the sample size), that point is declared to be an outlier. Note that the mean and standard deviation are computed from all the data, including the suspected outlier in the calculations. As the table below shows, the presence of the second outlier (in a small data set) inflates the standard deviation, and so decreases the value of Z to below the threshold used to define an outlier.

	Left (one outlier)	**Right (two outliers)**
Mean	60.364	68.167
SD	33.384	41.759
Z	2.8048	2.0554
n	11	12

Critical Z to define outlier (alpha=5%)	2.3547	2.4116
Critical Z to define outlier (alpha=1%)	2.5641	2.6357

4.13.7   Simulations to compare the Grubbs' and ROUT methods

Goal

Since the ROUT method is not yet a standard method, we did simulations to compare it to the Grubbs method. We compared the two methods for data with no outliers, with one outlier and with two outliers.

- All simulations assumed a Gaussian distribution with a mean of 100 and SD of 15 for the bulk of the values.

- A specified number of outliers were added. These were selected from a uniform distribution whose limits are specified.

- How the false discovery rate (FDR) was computed: For each simulated data set, the FDR was defined to be 0.0 if no outliers were detected. If any outliers were detected, the FDR for that simulation is the fraction of outliers that are false -values that were simulated from the Gaussian distribution, and were not included as outliers by the simulation. The overall FDR is the average of these individual FDR values over the simulations.

- In each case, 25,000 simulations were done.

Details of the simulations

The table below shows the ten simulated experimental designs, which differ in sample size (n), the number of outliers included in the sample, and the range of values from which those outliers were selected.

Design	n	# of outliers	Outlier range
A	100	0	
B	10	0	
C	10	1	50-75
D	10	1	100-125
E	100	1	100-125
F	100	1	50-75
G	100	2	50-75
H	100	2	100-125
I	10	2	50-75
J	25	2	50-75

Here are the results. Each set of simulated data was analyzed by both the Grubbs and ROUT methods.

| | Design | # Outliers | Analysis method | Number of outliers identified | | | | |
				0	1	2	>2	FDR
1	A	0	Grubbs 5%	95.104%	4.69%	0.19%	0.20%	4.90%
2	A	0	Rout 5%	94.31%	4.68%	0.74%	0.10%	5.69%
3	A	0	Grubbs 1%	99.10%	0.90%	0.00%	0.00%	0.90%
4	A	0	Rout 1%	98.70%	1.21%	0.00%	0.08%	1.21%
5	B	0	Grubbs 5%	94.99%	5.01%	0.00%	0.00%	5.01%
6	B	0	Rout 5%	95.13%	3.87%	0.98%	0.02%	4.87%
7	B	0	Grubbs 1%	98.92%	1.08%	0.00%	0.00%	1.08%
8	B	0	Rout 1%	98.65%	1.14%	0.21%	0.00%	1.35%
9	C	1	Grubbs 1%	74.33%	25.41%	0.26%	0.00%	0.13%
10	C	1	Rout 1%	78.11%	21.29%	0.60%	0.00%	0.31%
11	D	1	Grubbs 1%	5.50%	93.51%	0.99%	0.00%	0.50%
12	D	1	Rout 1%	15.38%	84.01%	0.60%	0.00%	0.30%

13	D	1	Grubbs 5%	0.20%	94.86%	4.75%	0.18%	2.51%
14	D	1	Rout 5%	2.30%	94.96%	2.70%	0.04%	2.73%
15	E	1	Grubbs 1%	0.00%	98.94%	1.05%	0.01%	0.53%
16	E	1	Rout 1%	0.00%	97.92%	1.94%	0.14%	1.07%
17	F	1	Grubbs 1%	43.94%	55.47%	0.57%	0.02%	0.40%
18	F	1	Rout 1%	47.08%	51.16%	1.63%	0.11%	1.05%
19	G	2	Grubbs 1%	39.70%	29.84%	30.72%	0.38%	0.16%
20	G	2	Rout 1%	29.08%	26.61%	40.37%	1.88%	0.82%
21	G	2	Grubbs 5%	10.82%	21.29%	6 4.23%	3.66%	1.40%
22	G	2	Rout 5%	7.52%	15.50%	66.54%	10.43%	3.96%
23	H	2	Grubbs 1%	0.00%	0.00%	98.89%	1.11%	0.37%
24	H	2	Rout 1%	0.00%	0.00%	97.57%	2.43%	0.84%
25	I	2	Grubbs 5%	98.80%	1.20%	0.00%	0.00%	0.00%
26	I	2	Rout 5%	6.06%	0.97%	92.80%	0.16%	0.05%
27	I	2	Rout 1%	27.46%	2.58%	69.95%	0.01%	0.004%
28	J	2	Grubbs 5%	49.16%	7.86%	40.85%	2.14%	0.737%
29	J	2	Rout 5%	24.57%	13.27%	57.46%	0.71%	1.74%
30	J	2	Grubbs 1%	90.21%	3.51%	6.20%	0.72%	0.24%
31	J	2	Rout 1%	54.47%	15.08%	29.46%	0.98%	0.36%

Results

When there are no outliers

When the simulations added no outliers to the data sets, the ROUT and Grubbs' tests perform almost identically. The value of Q specified for the ROUT method is equivalent to the value of alpha you set for the Grubbs' test. If you set alpha to 0.05 or Q to 5%, then you'll detect a single outlier in about 5% of simulations, even though all data in these simulations came from a Gaussian distribution.

When there is one outlier

When the simulations include a single outlier not from the same Gaussian distribution as the rest, the Grubb's test is slightly better able to detect it. The ROUT method has both more false negatives and false positives. It is slightly more likely to miss the outlier, and is also more likely to find two outliers even when the simulation actually only included one.

This is not so surprising, as Grubbs' test was really designed to detect a single outlier (although it can be used iteratively to detect more). While the difference between the two methods is consistent, it is not substantial.

When there are two outliers

When simulations include two outliers in a small data set, the ROUT test does a much better job. The iterative Grubbs' test is subject to masking[163], while the ROUT test is not. Whether or not masking is an issue depends on how large the sample is and how far the outliers are from the mean of the other values. In situations where masking is a real possibility, the ROUT test works much better than Grubbs' test. For example, when n=10 with two outliers (experimental design I), the Grubbs test never found both outliers and missed both outliers in 98.8% of the simulations. In the remaining 1.2% of simulations, the Grubbs' test found one of the two outliers. In contrast, the ROUT method identified both outliers in 92.8% of those simulations, and missed both in only 6% of simulations.

Reminder. Don't delete outliers without thinking.

One an outlier (or several outliers) is detected, stop and think. Don't just delete it.

Think about the assumptions. Both the Grubbs' and ROUT methods assume that the data (except for any outlers) are sampled from a Gaussian distribution. If that assumption is violated, the "outliers" may be from the same distribution as the rest. Beware of lognormal distributions. These distributions have values in the tails that will often be incorrectly flagged as outliers by methods that assume a Gaussian distribution.

Even if the value truly is an outlier from the rest, it may be a important value. It may not be a mistake. It may tell you about biological variability.

Conclusion

Grubbs' is slightly better than the ROUT method for the task it was designed for: Detecting a single outlier from a Gaussian distribution.

The Grubbs' test is much worse than the ROUT method at detecting two outliers. I can't imagine any scientific situation where you know for sure that there are either no outliers, or only one outlier, with no possibility of two or more outliers. Whenever the presence of two (or more) outliers is possible, we recommend that the ROUT method be used instead of the Grubbs' test.

More details, with links to the Prism file used to do these simulations

4.14   Analysis checklists

All statistical analysis is based on a set of assumptions. These checklists help you review the assumptions, and make sure you have picked a useful test. The checklists appear twice in this guide: Once here with all the checklists together, and again as part of the explanation for each individual test.

Unpaired t test[170]

Paired t test[172]

Mann-Whitney test[176]

Wilcoxon matched pairs test[177]

One-way ANOVA[179]

Repeated measures one-way ANOVA[181]

Kruskal-Wallis test[184]

Friedman's test[185]

Two-way ANOVA[187]

Repeated measures two-way ANOVA[189]

Contingency tables[191]

Survival analysis[192]

Outliers[194]

4.14.1 Unpaired t test

The unpaired t test compares the means of two unmatched groups, assuming that the values follow a Gaussian distribution. Read elsewhere to learn about choosing a t test[263], and interpreting the results[276].

✔ **Are the populations distributed according to a Gaussian distribution?**

The unpaired t test assumes that you have sampled your data from populations that follow a Gaussian distribution. This assumption matters less with large samples due to the Central Limit Theorem[41].

Prism can perform normality tests as part of the Column Statistics[201] analysis. Learn more[239].

If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values (perhaps to logs or reciprocals) to make the distributions more Gaussian. Another choice is to use the Mann-Whitney nonparametric test instead of a t test. A final option is to use t test anyway, knowing that it is fairly robust to violations of a Gaussian distribution with large samples.

✔ **Do the two populations have the same variances?**

The unpaired t test assumes that the two populations have the same variances (and thus the same standard deviation).

 Prism tests for equality of variance with an F test. The P value from this test answers this question: If the two populations really have the same variance, what is the chance that you would randomly select samples whose ratio of variances is as far from 1.0 (or further) as observed in your experiment? A small P value suggests that the variances are different.

 Don't base your conclusion solely on the F test. Also think about data from other similar experiments. If you have plenty of previous data that convinces you that the variances are really equal, ignore the F test (unless the P value is really tiny) and interpret the t test results as usual.

 In some contexts, finding that populations have different variances may be as important as finding different means.

✔ Are the data unpaired?

 The unpaired t test works by comparing the difference between means with the standard error of the difference, computed by combining the standard errors of the two groups. If the data are paired or matched, then you should choose a paired t test instead. If the pairing is effective in controlling for experimental variability, the paired t test will be more powerful than the unpaired test.

✔ Are the "errors" independent?

 The term "error" refers to the difference between each value and the group mean. The results of a t test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✔ Are you comparing exactly two groups?

 Use the t test only to compare two groups. To compare three or more groups, use one-way ANOVA[405] followed by multiple comparison tests. It is not appropriate to perform several t tests, comparing two groups at a time. Making multiple comparisons increases the chance of finding a

statistically significant difference by chance and makes it difficult to interpret P values and statements of statistical significance. Even if you want to use planned comparisons to avoid correcting for multiple comparisons, you should still do it as part of one-way ANOVA to take advantage of the extra degrees of freedom that brings you.

✔ Do both columns contain data?

If you want to compare a single set of experimental data with a theoretical value (perhaps 100%) don't fill a column with that theoretical value and perform an unpaired t test. Instead, use a one-sample t test[256] .

✔ Do you really want to compare means?

The unpaired t test compares the means of two groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the two distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✔ If you chose a one-tail P value, did you predict correctly?

If you chose a one-tail P value[75] , you should have predicted which group would have the larger mean before collecting any data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the P value reported by Prism and state that P>0.50.

4.14.2 Paired t test

The paired t test compares the means of two matched groups, assuming that the distribution of the before-after differences follows a Gaussian distribution.

✔ Are the differences distributed according to a Gaussian distribution?

The paired t test assumes that you have sampled your pairs of values from a population of pairs where the difference between pairs follows a Gaussian distribution.

While this assumption is not too important with large samples, it is important with small sample sizes. Test this assumption with Prism[293].

Note that the paired t test, unlike the unpaired t test, does **not** assume that the two sets of data (before and after, in the typical example) are sampled from populations with equal variances.

✔ Was the pairing effective?

The pairing should be part of the experimental design and not something you do after collecting data. Prism tests the effectiveness of pairing by calculating the Pearson correlation coefficient, r, and a corresponding P value. If the P value is small, the two groups are significantly correlated. This justifies the use of a paired test.

If this P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based solely on this one P value, but also on the experimental design and the results of other similar experiments.

✔ Are the pairs independent?

The results of a paired t test only make sense when the pairs are independent[32] – that whatever factor caused a difference (between paired values) to be too high or too low affects only that one pair. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six pairs of values, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may cause the after-before differences from one animal to be high or low. This factor would affect two of the pairs, so they are not independent.

✔ Are you comparing exactly two groups?

Use the t test only to compare two groups. To compare three or more matched groups, use repeated measures one-way ANOVA followed by post tests. It is not appropriate[117] to perform several t tests, comparing two groups at a time.

✔ If you chose a one-tail P value, did you predict correctly?

If you chose a one-tail P value, you [should have predicted](#)[75] which group would have the larger mean before collecting data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the reported P value and state that P>0.50.

✔ Do you care about differences or ratios?

The paired t test analyzes the differences between pairs. With some experiments, you may observe a very large variability among the differences. The differences are larger when the control value is larger. With these data, you'll get more consistent results if you perform a [ratio t test](#)[299].

4.14.3 Ratio t test

The ratio t test compares the means of two matched groups, assuming that the distribution of the logarithms of the before/after ratios follows a Gaussian distribution.

✔ Are the log(ratios) distributed according to a Gaussian distribution?

The ratio t test assumes that you have sampled your pairs of values from a population of pairs where the log of the ratios follows a Gaussian distribution.

While this assumption is not too important with large samples, it is important with small sample sizes. [Test this assumption with Prism](#)[293].

✔ Was the pairing effective?

The pairing should be part of the experimental design and not something you do after collecting data. Prism tests the effectiveness of pairing by calculating the Pearson correlation coefficient, r, between the logarithms of the two columns of data. If the corresponding P value. If the P value is small, the two groups are significantly correlated. This justifies the use of a paired test.

If this P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based solely on this one P value, but also on the experimental design and the results of other similar experiments.

✔ Are the pairs independent?

The results of a ratio t test only make sense when the pairs are independent[32] – that whatever factor caused a rato (of paired values) to be too high or too low affects only that one pair. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six pairs of values, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may cause the after-before differences from one animal to be high or low. This factor would affect two of the pairs, so they are not independent.

✔ Are you comparing exactly two groups?

Use the t test only to compare two groups. To compare three or more matched groups, transform the values to their logarithms, and then use repeated measures one-way ANOVA followed by post tests. It is not appropriate[117] to perform several t tests, comparing two groups at a time.

✔ If you chose a one-tail P value, did you predict correctly?

If you chose a one-tail P value, you should have predicted[75] which group would have the larger mean before collecting data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the reported P value and state that P>0.50.

✔ Do you care about differences or ratios?

The ratio t test analyzes the logarithm of the ratios of paired values. The assumption is that the ratio is a consistent measure of experimental effect. With many experiments, you may observe that the difference between pairs is a consistent measure of effect, and the ratio is not. In these cases, use a paired t test[291], not the ratio t test.

4.14.4 Mann-Whitney test

The Mann-Whitney test[309] is a nonparametric test that compares the distributions of two unmatched groups. It is sometimes said to compare medians, but this is not always true[313].

✓ **Are the "errors" independent?**

The term "error" refers to the difference between each value and the group median. The results of a Mann-Whitney test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent[32] if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✓ **Are the data unpaired?**

The Mann-Whitney test works by ranking all the values from low to high, and comparing the mean rank in the two groups. If the data are paired or matched, then you should choose a Wilcoxon matched pairs test instead.

✓ **Are you comparing exactly two groups?**

Use the Mann-Whitney test only to compare two groups. To compare three or more groups, use the Kruskal-Wallis test followed by post tests. It is not appropriate to perform several Mann-Whitney (or t) tests, comparing two groups at a time.

✓ **Do the two groups follow data distributions with the same shape?**

If the two groups have distributions with similar shapes, then you can interpret the Mann-Whitney test as comparing medians. If the distributions have different shapes, you really cannot interpret[309] the results of the Mann-Whitney test.

✔️ **Do you really want to compare medians?**

The Mann-Whitney test compares the medians of two groups (<u>well, not exactly</u>[313]). It is possible to have a tiny P value – clear evidence that the population medians are different – even if the two distributions overlap considerably.

✔️ **If you chose a one-tail P value, did you predict correctly?**

If you chose a one-tail P value, you should have predicted which group would have the larger median before collecting any data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the P value reported by Prism and state that P>0.50. <u>One- vs. two-tail P values.</u>

✔️ **Are the data sampled from non-Gaussian populations?**

By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions, but there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to give you a small P value), and that difference is quite noticeable with small sample sizes.

4.14.5 Wilcoxon matched pairs test

The Wilcoxon test is a nonparametric test that compares two paired groups. Read elsewhere to learn about <u>choosing a t test</u>[263], and <u>interpreting the results</u>[324].

✔️ **Are the pairs independent?**

The results of a Wilcoxon test only make sense when the pairs are <u>independent</u>[32] – that whatever factor caused a difference (between paired values) to be too high or too low affects only that one pair. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six pairs of values, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may cause the after-before differences from one animal to be high or low. This factor

would affect two of the pairs (but not the other four), so these two are not independent.

✓ Is the pairing effective?

If the P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based solely on this one P value, but also on the experimental design and the results you have seen in other similar experiments.

✓ Are you comparing exactly two groups?

Use the Wilcoxon test only to compare two groups. To compare three or more matched groups, use the Friedman test followed by post tests. It is not appropriate[117] to perform several Wilcoxon tests, comparing two groups at a time.

✓ If you chose a one-tail P value, did you predict correctly?

If you chose a one-tail P value[75], you should have predicted which group would have the larger median before collecting any data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the P value reported by Prism and state that P>0.50.

✓ Are the data clearly sampled from non-Gaussian populations?

By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions. But there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to give you a small P value), especially with small sample sizes. Furthermore, Prism (along with most other programs) does not calculate confidence intervals when calculating nonparametric tests. If the distribution is clearly not bell-shaped, consider transforming the values (perhaps to logs or reciprocals) to create a Gaussian distribution and then using a t test.

✓ Are the differences distributed symmetrically?

The Wilcoxon test first computes the difference between the two values in each row, and analyzes only the list of differences. The Wilcoxon test

does not assume that those differences are sampled from a Gaussian distribution. However it does assume that the differences are distributed symmetrically around their median.

4.14.6 One-way ANOVA

One-way ANOVA compares the means of three or more unmatched groups. Read elsewhere to learn about <u>choosing a test</u>[406], and <u>interpreting the results</u>[425].

✔ **Are the populations distributed according to a Gaussian distribution?**

One-way ANOVA assumes that you have sampled your data from populations that follow a Gaussian distribution. While this assumption is not too important with large samples due to the <u>Central Limit Theorem</u>[41], it is important with small sample sizes (especially with unequal sample sizes). Prism can test for violations of this assumption, but normality tests have limited utility.

If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values (perhaps to logs or reciprocals) to make the distributions more Gaussian. Another choice is to use the Kruskal-Wallis nonparametric test instead of ANOVA. A final option is to use ANOVA anyway, knowing that it is fairly robust to violations of a Gaussian distribution with large samples.

✔ **Do the populations have the same standard deviation?**

One-way ANOVA assumes that all the populations have the same standard deviation (and thus the same variance). This assumption is not very important when all the groups have the same (or almost the same) number of subjects, but is very important when sample sizes differ.

InStat tests for equality of variance with two tests: The Brown-Forsythe test and Bartlett's test. The P value from these tests answer this question: If the populations really have the same variance, what is the chance that you'd randomly select samples whose variances are as different from one another as those observed in your experiment. A small P value suggests that the variances are different.

Don't base your conclusion solely on these tests. Also think about data from other similar experiments. If you have plenty of previous data that convinces you that the variances are really equal, ignore these tests (unless the P value is really tiny) and interpret the ANOVA results as usual. Some statisticians recommend ignoring tests for equal variance altogether if the sample sizes are equal (or nearly so).

In some experimental contexts, finding different variances may be as important as finding different means. If the variances are different, then the populations are different -- regardless of what ANOVA concludes about differences between the means.

✓ Are the data unmatched?

One-way ANOVA works by comparing the differences among group means with the pooled standard deviations of the groups. If the data are matched, then you should choose repeated-measures ANOVA instead. If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA.

✓ Are the "errors" independent?

The term "error" refers to the difference between each value and the group mean. The results of one-way ANOVA only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✓ Do you really want to compare means?

One-way ANOVA compares the means of three or more groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✔ Is there only one factor?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group, with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments.

Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. These data need to be analyzed by two-way ANOVA[455], also called two factor ANOVA.

✔ Is the factor "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Type II ANOVA, also known as random-effect ANOVA, assumes that you have randomly selected groups from an infinite (or at least large) number of possible groups, and that you want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment. Type II random-effects ANOVA is rarely used, and Prism does not perform it.

✔ Do the different columns represent different levels of a grouping variable?

One-way ANOVA asks whether the value of a single variable differs significantly among three or more groups. In Prism, you enter each group in its own column. If the different columns represent different variables, rather than different groups, then one-way ANOVA is not an appropriate analysis. For example, one-way ANOVA would not be helpful if column A was glucose concentration, column B was insulin concentration, and column C was the concentration of glycosylated hemoglobin.

4.14.7 Repeated measures one-way ANOVA

Repeated measures one-way ANOVA compares the means of three or more matched groups. Read elsewhere to learn about choosing a test[406], and interpreting the results[437].

✔ Was the matching effective?

 The whole point of using a repeated-measures test is to control for experimental variability. Some factors you don't control in the experiment will affect all the measurements from one subject equally, so will not affect the difference between the measurements in that subject. By analyzing only the differences, therefore, a matched test controls for some of the sources of scatter.

 The matching should be part of the experimental design and not something you do after collecting data. Prism tests the effectiveness of matching with an F test (distinct from the main F test of differences between columns). If the P value for matching is large (say larger than 0.05), you should question whether it made sense to use a repeated-measures test. Ideally, your choice of whether to use a repeated-measures test should be based not only on this one P value, but also on the experimental design and the results you have seen in other similar experiments.

✔ Are the subjects independent?

 The results of repeated-measures ANOVA only make sense when the subjects are independent. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six rows of data, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may affect the measurements from one animal. Since this factor would affect data in two (but not all) rows, the rows (subjects) are not independent.

✔ Is the random variability distributed according to a Gaussian distribution?

 Repeated-measures ANOVA assumes that each measurement is the sum of an overall mean, a treatment effect (the average difference between subjects given a particular treatment and the overall mean), an individual effect (the average difference between measurements made in a certain subject and the overall mean) and a random component. Furthermore, it assumes that the random component follows a Gaussian distribution and that the standard deviation does not vary between individuals (rows) or treatments (columns). While this assumption is not too important with large samples, it can be important with small sample sizes. Prism does not test for violations of this assumption.

✓ Is there only one factor?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group, with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments.

Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. Similarly, there are two factors if you wish to compare the effect of drug treatment at several time points. These data need to be analyzed by two-way ANOVA, also called two-factor ANOVA.

✓ Is the factor "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Type II ANOVA, also known as random-effect ANOVA, assumes that you have randomly selected groups from an infinite (or at least large) number of possible groups, and that you want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment. Type II random-effects ANOVA is rarely used, and Prism does not perform it.

WIth repeated measures, Prism can fit a mixed effects model. This model assumes the differences among subjects (or litters...) is random. But it assumes the factor that defines which column each value is entered into is fixed.

✓ Can you accept the assumption of circularity or sphericity?

Repeated-measures ANOVA assumes that the random error truly is random. A random factor that causes a measurement in one subject to be a bit high (or low) should have no affect on the next measurement in the same subject. This assumption is called *circularity* or *sphericity*. It is closely related to another term you may encounter, *compound symmetry*.

Repeated-measures ANOVA is quite sensitive to violations of the assumption of circularity. If the assumption is violated, the P value will be too low. One way to violate this assumption is to make the repeated

measurements in too short a time interval, so that random factors that cause a particular value to be high (or low) don't wash away or dissipate before the next measurement. To avoid violating the assumption, wait long enough between treatments so the subject is essentially the same as before the treatment. When possible, also randomize the order of treatments.

You only have to worry about the assumption of circularity when you perform a repeated-measures experiment, where each row of data represents repeated measurements from a single subject. It is impossible to violate the assumption with randomized block experiments, where each row of data represents data from a matched set of subjects.

If you cannot accept the assumption of sphericity, you can specify that on the Parameters dialog. In that case, Prism will take into account possible violations of the assumption (using the method of Geisser and Greenhouse) and report a higher P value.

✔ If any values are missing, was that due to a random event?

Starting with Prism 8, repeated measures data can be calculated with missing values by fitting a mixed model. But the results can only be interpreted if the reason for the value being missing is random. If a value is missing because it was too high to measure (or too low), then it is not missing randomly. If values are missing because a treatment is toxic, then the values are not randomly missing.

4.14.8 Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric test that compares three or more unpaired or unmatched groups.Read elsewhere to learn about [choosing a test](#)[406], and [interpreting the results](#)[448].

✔ Are the "errors" independent?

The term "error" refers to the difference between each value and the group median. The results of a Kruskal-Wallis test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have nine values in each of three groups, but these were obtained from two animals in each group

(in triplicate). In this case, some factor may cause all three values from one animal to be high or low.

✓ Are the data unpaired?

 If the data are paired or matched, then you should consider choosing the Friedman test instead. If the pairing is effective in controlling for experimental variability, the Friedman test will be more powerful than the Kruskal-Wallis test.

✓ Are the data sampled from non-Gaussian populations?

 By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions, but there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to detect a true difference), especially with small sample sizes. Furthermore, Prism (along with most other programs) does not calculate confidence intervals when calculating nonparametric tests. If the distribution is clearly not bell-shaped, consider transforming the values (perhaps to logs or reciprocals) to create a Gaussian distribution and then using ANOVA.

✓ Do you really want to compare medians?

 The Kruskal-Wallis test compares the medians of three or more groups. It is possible to have a tiny P value – clear evidence that the population medians are different – even if the distributions overlap considerably.

✓ Are the shapes of the distributions identical?

 The Kruskal-Wallis test does not assume that the populations follow Gaussian distributions. But it does assume that the shapes of the distributions are identical. The medians may differ – that is what you are testing for – but the test assumes that the shapes of the distributions are identical. If two groups have very different distributions, consider transforming the data to make the distributions more similar.

4.14.9 Friedman's test

Friedman's test is a nonparametric test that compares three or more paired groups.

✔ Was the matching effective?

The whole point of using a repeated-measures test is to control for experimental variability. Some factors you don't control in the experiment will affect all the measurements from one subject equally, so they will not affect the difference between the measurements in that subject. By analyzing only the differences, therefore, a matched test controls for some of the sources of scatter.

The matching should be part of the experimental design and not something you do after collecting data. Prism does not test the adequacy of matching with the Friedman test.

✔ Are the subjects (rows) independent?

The results of a Friedman test only make sense when the subjects (rows) are independent – that no random factor has affected values in more than one row. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six rows of data obtained from three animals in duplicate. In this case, some random factor may cause all the values from one animal to be high or low. Since this factor would affect two of the rows (but not the other four), the rows are not independent.

✔ Are the data clearly sampled from non-Gaussian populations?

By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions, but there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to give you a small P value), especially with small sample sizes. Furthermore, Prism (along with most other programs) does not calculate confidence intervals when calculating nonparametric tests. If the distribution is clearly not bell-shaped, consider transforming the values (perhaps to logs or reciprocals) to create a Gaussian distribution and then using repeated-measures ANOVA.

✔ Is there only one factor?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group, with a drug treatment

group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments.

Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. Similarly, there are two factors if you wish to compare the effect of drug treatment at several time points. These data need to be analyzed by two-way ANOVA, also called two-factor ANOVA.

4.14.10 Two-way ANOVA

Two-way ANOVA, also called two-factor ANOVA, determines how a response is affected by two factors. For example, you might measure a response to three different drugs in both men and women. In this example, drug treatment is one factor and gender is the other. Read elsewhere to learn about [choosing a test][456], and [interpreting the results.][488]

✓ Are the populations distributed according to a Gaussian distribution?

Two-way ANOVA assumes that your replicates are sampled from Gaussian distributions. While this assumption is not too important with large samples, it is important with small sample sizes, especially with unequal sample sizes. Prism does not test for violations of this assumption. If you really don't think your data are sampled from a Gaussian distribution (and no transform will make the distribution Gaussian), you should consider performing nonparametric two-way ANOVA. Prism does not offer this test.

ANOVA also assumes that all sets of replicates have the same SD overall, and that any differences between SDs are due to random sampling.

✓ Are the data unmatched?

Standard two-way ANOVA works by comparing the differences among group means with the pooled standard deviations of the groups. If the data are matched, then you should choose repeated-measures ANOVA instead. If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA.

✓ Are the "errors" independent?

The term "error" refers to the difference between each value and the mean of all the replicates. The results of two-way ANOVA only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six replicates, but these were obtained from two animals in triplicate. In this case, some factor may cause all values from one animal to be high or low.

✓ Do you really want to compare means?

Two-way ANOVA compares the means. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✓ Are there two factors?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments. Prism has a separate analysis for one-way ANOVA.

Some experiments involve more than two factors. For example, you might compare three different drugs in men and women at four time points. There are three factors in that experiment: drug treatment, gender and time. These data need to be analyzed by three-way ANOVA, also called three-factor ANOVA. Prism does not perform three-way ANOVA.

✓ Are both factors "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Different calculations are needed if you randomly selected groups from an infinite (or at least large) number of possible

groups, and want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment.

4.14.11 Repeated measures two-way ANOVA

Two-way ANOVA, also called two-factor ANOVA, determines how a response is affected by two factors. "Repeated measures" means that one of the factors was repeated. For example you might compare two treatments, and measure each subject at four time points (repeated). Read elsewhere to learn about choosing a test[456], graphing the data[513], and interpreting the results[506].[488]

✔ Are the data matched?

If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA. Also check that your choice in the experimental design tab matches how the data are actually arranged. If you make a mistake, and the calculations are done assuming the wrong factor is repeated, the results won't be correct or useful.

✔ Are there two factors?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments. Prism has a separate analysis for one-way ANOVA.

Some experiments involve more than two factors. For example, you might compare three different drugs in men and women at four time points. There are three factors in that experiment: drug treatment, gender and time. These data need to be analyzed by three-way ANOVA, also called three-factor ANOVA.

✔ Are both factors "fixed" rather than "random"?

While Prism assumes the participants in repeated measures are chosen randomly, it assumes that the treatments or categories designated by rows or data set columns are fixed. This means you are asking about how those particular treatments or categories affect the results. Different calculations would be needed if you randomly selected the

treatments or categories from an infinite (or at least large) number of possible treatments or categories, and want to reach conclusions about differences among ALL the treatments or categories, even the ones you didn't include in this experiment. Prism does not handle this situation.

✔ Can you accept the assumption of sphericity?

A random factor that causes a measurement in one subject to be a bit high (or low) should have no effect on the next measurement in the same subject. This assumption is called **circularity** or **sphericity**. It is closely related to another term you may encounter in advanced texts, **compound symmetry**.

You only have to worry about the assumption of circularity when your experiment truly is a repeated-measures experiment, with measurements from a single subject. Circularity is unlikely to be an issue with randomized block experiments where you used a matched set of subjects (or a matched set of experiments)

Repeated-measures ANOVA is quite sensitive to violations of the assumption of circularity. If the assumption is violated, the P value will be too low. You'll violate this assumption when the repeated measurements are made too close together so that random factors that cause a particular value to be high (or low) don't wash away or dissipate before the next measurement. To avoid violating the assumption, wait long enough between treatments so the subject is essentially the same as before the treatment. Also randomize the order of treatments, when possible.

✔ Consider alternatives to repeated measures two-way ANOVA.

Two-way ANOVA may not answer the questions your experiment was designed to address. [Consider alternatives.](#)

✔ If any values are missing, was that due to a random event?

Starting with Prism 8, repeated measures data can be calculated with missing values by fitting a mixed model. But the results can only be interpreted if the reason for the value being missing is random. If a value is missing because it was too high to measure (or too low), then it is not missing randomly. If values are missing because a treatment is toxic, then the values are not randomly missing.

4.14.12 Contingency tables

Contingency tables summarize results where you compared two or more groups and the outcome is a categorical variable (such as disease vs. no disease, pass vs. fail, artery open vs. artery obstructed). Read elsewhere to learn about relative risks & odds ratios[557], sensitivity & specificity[560], and interpreting P values[553].

✔ **Are the subjects independent?**

The results of a chi-square or Fisher's test only make sense if each subject (or experimental unit) is independent of the rest. That means that any factor that affects the outcome of one subject only affects that one subject. Prism cannot test this assumption. You must think about the experimental design. For example, suppose that the rows of the table represent two different kinds of preoperative antibiotics and the columns denote whether or not there was a postoperative infection. There are 100 subjects. These subjects are not independent if the table combines results from 50 subjects in one hospital with 50 subjects from another hospital. Any difference between hospitals, or the patient groups they serve, would affect half the subjects but not the other half. You do not have 100 independent observations. To analyze this kind of data, use the Mantel-Haenszel test or logistic regression. Beginning with version 8.3.0, Prism offers both simple logistic regression and multiple logistic regression.

✔ **Are the data unpaired?**

In some experiments, subjects are matched for age and other variables. One subject in each pair receives one treatment while the other subject gets the other treatment. These data should be analyzed by special methods such as McNemar's test[569]. Paired data should not be analyzed by chi-square or Fisher's test.

✔ **Is your table really a contingency table?**

To be a true contingency table, each value must represent numbers of subjects (or experimental units). If it tabulates averages, percentages, ratios, normalized values, etc. then it is not a contingency table and the results of chi-square or Fisher's tests will not be meaningful. If you've entered observed values on one row (or column) and expected values on another, you do not have a contingency table, and should use a separate analysis[563] designed for those kind of data.

✔ **Does your table contain only data?**

 The chi-square test is not only used for analyzing contingency tables. It can also be used to compare the observed number of subjects in each category with the number you expect to see based on theory. Prism cannot do this kind of chi-square test. It is not correct to enter observed values in one column and expected in another. When analyzing a contingency table with the chi-square test, Prism generates the expected values from the data – you do not enter them.

✔ **Are the rows or columns arranged in a natural order?**

 If your table has two columns and more than two rows (or two rows and more than two columns), Prism will perform the chi-square test for trend as well as the regular chi-square test. The results of the test for trend will only be meaningful if the rows (or columns) are arranged in a natural order, such as age, duration, or time. Otherwise, ignore the results of the chi-square test for trend and only consider the results of the regular chi-square test.

4.14.13 Survival analysis

Survival curves plot the results of experiments where the outcome is time until death. Usually you wish to compare the survival of two or more groups. Read elsewhere to learn about interpreting survival curves[587], and comparing two[592] (or more than two[600]) survival curves.

✔ **Are the subjects independent?**

 Factors that influence survival should either affect all subjects in a group or just one subject. If the survival of several subjects is linked, then you don't have independent observations. For example, if the study pools data from two hospitals, the subjects are not independent, as it is possible that subjects from one hospital have different average survival times than subjects from another. You could alter the median survival curve by choosing more subjects from one hospital and fewer from the other. To analyze these data, use Cox proportional hazards regression, which Prism cannot perform.

✓ Were the entry criteria consistent?

Typically, subjects are enrolled over a period of months or years. In these studies, it is important that the starting criteria don't change during the enrollment period. Imagine a cancer survival curve starting from the date that the first metastasis was detected. What would happen if improved diagnostic technology detected metastases earlier? Even with no change in therapy or in the natural history of the disease, survival time will apparently increase. Here's why: Patients die at the same age they otherwise would, but are diagnosed when they are younger, and so live longer with the diagnosis. (That is why airlines have improved their "on-time departure" rates. They used to close the doors at the scheduled departure time. Now they close the doors ten minutes before the "scheduled departure time". This means that the doors can close ten minutes later than planned, yet still be "on time". It's not surprising that "on-time departure" rates have improved.)

✓ Was the end point defined consistently?

If the curve is plotting time to death, then there can be ambiguity about which deaths to count. In a cancer trial, for example, what happens to subjects who die in a car accident? Some investigators count these as deaths; others count them as censored subjects. Both approaches can be justified, but the approach should be decided before the study begins. If there is any ambiguity about which deaths to count, the decision should be made by someone who doesn't know which patient is in which treatment group.

If the curve plots time to an event other than death, it is crucial that the event be assessed consistently throughout the study.

✓ Is time of censoring unrelated to survival?

The survival analysis is only valid when the survival times of censored patients are identical (on average) to the survival of subjects who stayed with the study. If a large fraction of subjects are censored, the validity of this assumption is critical to the integrity of the results. There is no reason to doubt that assumption for patients still alive at the end of the study. When patients drop out of the study, you should ask whether the reason could affect survival. A survival curve would be misleading, for example, if many patients quit the study because they were too sick to

come to clinic, or because they stopped taking medication because they felt well.

✔ Does average survival stay constant during the course of the study?

Many survival studies enroll subjects over a period of several years. The analysis is only meaningful if you can assume that the average survival of the first few patients is not different than the average survival of the last few subjects. If the nature of the disease or the treatment changes during the study, the results will be difficult to interpret.

✔ Is the assumption of proportional hazards reasonable?

The logrank test is only strictly valid when the survival curves have proportional hazards. This means that the rate of dying in one group is a constant fraction of the rate of dying in the other group. This assumption has proven to be reasonable for many situations. It would not be reasonable, for example, if you are comparing a medical therapy with a risky surgical therapy. At early times, the death rate might be much higher in the surgical group. At later times, the death rate might be greater in the medical group. Since the hazard ratio is not consistent over time (the assumption of proportional hazards is not reasonable), these data should not be analyzed with a logrank test.

✔ Were the treatment groups defined before data collection began?

It is not valid to divide a single group of patients (all treated the same) into two groups based on whether or not they responded to treatment (tumor got smaller, lab tests got better). By definition, the responders must have lived long enough to see the response. And they may have lived longer anyway, regardless of treatment. When you compare groups, the groups must be defined before data collection begins.

4.14.14 Outliers

If the outlier test identifies one or more values as being an outlier, ask yourself these questions:

✔ Was the outlier value entered into the computer incorrectly?

If the "outlier" is in fact a typo, fix it. It is always worth going back to the original data source, and checking that outlier value entered into Prism is

actually the value you obtained from the experiment. If the value was the result of calculations, check for math errors.

✓ Is the outlier value scientifically impossible?

Of course you should remove outliers from your data when the value is completely impossible. Examples include a negative weight, or an age (of a person) that exceed 150 years. Those are clearly errors, and leaving erroneous values in the analysis would lead to nonsense results.

✓ Is the assumption of a Gaussian distribution dubious?

Both the Grubbs' and ROUT tests assume that all the values are sampled from a Gaussian distribution, with the possible exception of one (or a few) outliers from a different distribution. If the underlying distribution is not Gaussian, then the results of the outlier test is unreliable. It is especially important to [beware of lognormal distributions][157]. If the data are sampled from a lognormal distribution, you expect to find some very high values which can easily be mistaken for outliers. Removing these values would be a mistake.

✓ Is the outlier value potentially scientifically interesting?

 If each value is from a different animal or person, identifying an outlier might be important. Just because a value is not from the same Gaussian distribution as the rest doesn't mean it should be ignored. You may have discovered a polymorphism in a gene. Or maybe a new clinical syndrome. Don't throw out the data as an outlier until first thinking about whether the finding is potentially scientifically interesting.

✓ Does your lab notebook indicate any sort of experimental problem with that value

It is easier to justify removing a value from the data set when it is not only tagged as an "outlier" by an outlier test, but you also recorded problems with that value when the experiment was performed.

✓ Do you have a policy on when to remove outliers?

Ideally, removing an outlier should not be an *ad hoc* decision. You should follow a policy, and apply that policy consistently.

✔ **If you are looking for two or more outliers, could *masking* be a problem?**

[Masking][163] is the name given to the problem where the presence of two (or more) outliers, can make it harder to find even a single outlier.

If you answered no to all those questions...

If you've answered no to all the questions above, there are two possibilities:

- The suspect value came from the same Gaussian population as the other values. You just happened to collect a value from one of the tails of that distribution.

- The suspect value came from a different distribution than the rest. Perhaps it was due to a mistake, such as bad pipetting, voltage spike, holes in filters, etc.

If you knew the first possibility was the case, you would keep the value in your analyses. Removing it would be a mistake.

If you knew the second possibility was the case, you would remove it, since including an erroneous value in your analyses will give invalid results.

The problem, of course, is that you can never know for sure which of these possibilities is correct. An outlier test cannot answer that question for sure. Ideally, you should create a lab policy for how to deal with such data, and follow it consistently.

If you don't have a lab policy on removing outliers, here is suggestion: Analyze your data both with and without the suspected outlier. If the results are similar either way, you've got a clear conclusion. If the results are very different, then you are stuck. Without a consistent policy on when you remove outliers, you are likely to only remove them when it helps push the data towards the results you want.

5   STATISTICS WITH PRISM 8

5.1   Getting started with statistics with Prism

Links to places to get started.

5.1.1   What happened to the "Column statistics" analysis?

Before Prism 8, there was one analysis called "Column Statistics". Its functions have now been divided into three analyses, all new with Prism 8:

- [Descriptive statistics](#)[201]

- [Normality and lognormality tests](#)[239]

- [One-sample t test and Wilcoxon test](#)[254]

If you open a file created with an earlier version of Prism that used the Column Statistics analysis, its results will be fine in Prism 8. You'll be able to bring up the dialog and inspect and change the choices. But for new files, you won't be able to use that older, overloaded analysis, and instead should use one or more of the new analyses.

5.1.2   Statistical analyses with Prism

Key concepts: Statistical analyses with Prism

- To analyze data, start from a data table (or graph, or green results table), and click the Analyze button.

- Prism ignores any selection you have made on the data table. If you want to analyze only certain data sets, you can choose that on the Analyze Data dialog.

- Prism remembers the links between data, analyses and graphs. If you change (or replace) the data, the analyses and graphs will update automatically.

- The best way to learn about analyses is to choose tutorial data sets.

From the User Guide

How to analyze data with Prism

Creating chains of analyses

Changing an analysis

Frozen and orphaned analysis results

Excluding data points from an analysis

Embedding results on a graph

Hooking to analysis and info constants

Color coding key results

Simulating data and Monte Carlo analyses

Prism can plot and analyze simulated data, as well as data you enter.

Transforming, normalizing, etc.

5.1.3   Guided examples: Statistical analyses

Guided examples

These examples will guide you through most of Prism's statistical analyses.

Descriptive statistics

Compare two groups

Wilcoxon matched pairs test[322]

Categorical outcomes

Contingency table analysis[547]

Survival analysis[574]

Diagnostic lab tests

ROC curve[628]

Bland-Altman plot[638]

Analysis checklists

After completing each analysis, click the Analysis checklist button in the Interpret section of the Prism toolbar to review a list of questions that will help you interpret the results.
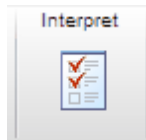
Here are links to a few of the analysis checklists, to view as examples.

Analysis checklist: Unpaired t test[170]

Analysis checklist: Survival analysis[192]

Analysis checklist: Repeated measures two-way ANOVA[189]

5.2   Descriptive statistics

.

This section explains how to analyze columns of numbers to compute descriptive statistics, compare the mean or median to a hypothetical value, and test for normality
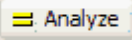
5.2.1   How to: Descriptive statistics
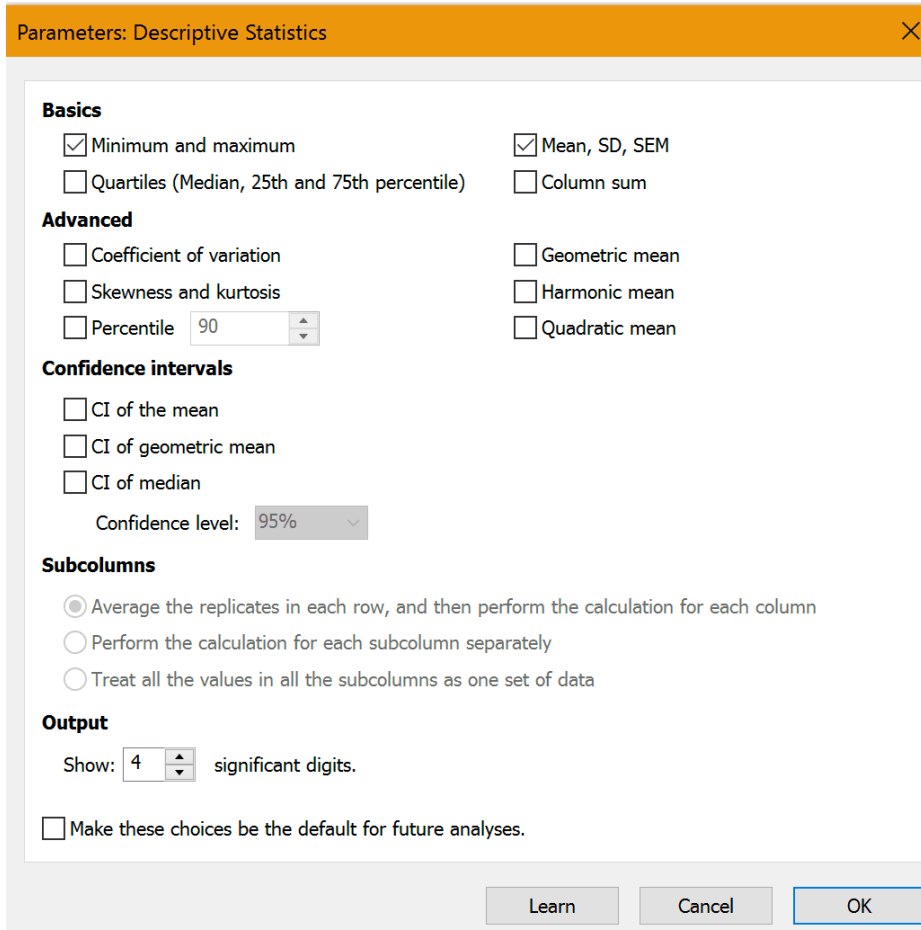
Entering data for descriptive statistics

Descriptive statistics by column are most often used with data entered on data tables formatted for Column data. If you want to experiment, create a Column data table and choose the sample data set: One-way ANOVA, ordinary.

You can also choose the descriptive statistics analysis from data entered onto XY or Grouped data tables.

Choose the descriptive statistics analysis

Click ⊒ Analyze and choose Descriptive statistics from the list of analyses for column data.

Prism's descriptive statistics analysis computes descriptive statistics of each data set. Tests for normality, and testing whether the mean of a column is different than a hypothetical value, are now in separate analyses (all were together in one analysis through Prism 7).

Choose analysis options

Basics

Learn more about quartiles[205], median[209], SD[42], and SEM[49] .

Advanced

Learn more about the [coefficient of variation][206], [skewness and kurtosis][211], [geometric mean][209], [harmonic mean, and quadratic mean.][214]

Confidence intervals

Choose to report the CI of the mean, the geometric mean or the median.

Subcolumns

The choices for subcolumn will not be available when you analyze data entered on table formatted for column data, which have no subcolumns. If your data are on a table formatted for XY or grouped data with subcolumns, choose to compute column statistics for each subcolumn individually or to average the subcolumns and compute columns statistics on the means.

If the data table has subcolumns for entry of mean and SD (or SEM) values, Prism calculates column statistics for the means, and ignores the SD or SEM values you entered.

5.2.2 Analysis checklist: Descriptive statistics

Value	Meaning
Minimum	The smallest value.
[25th percentile][205]	25% of values are lower than this.
[Median][209]	Half the values are lower; half are higher.
[75th percentile][205]	75% of values are lower than this.
Maximum	The largest value.
[Mean][209]	The average.
[Standard Deviation][42]	Quantifies variability or scatter.

Value	Meaning
Standard Error of Mean[49]	Quantifies how precisely the mean is known.
95% confidence interval[59]	Given some assumptions, there is a 95% chance that this range includes the true overall mean.
Coefficient of variation[206]	The standard deviation divided by the mean.
Geometric mean[209]	Compute the logarithm of all values, compute the mean of the logarithms, and then take the antilog of that mean. It is a better measure of central tendency when data follow a lognormal distribution (long tail).
Harmonic mean[214]	Compute the reciprocal of all values, compute the mean of the reciprocals, and then take the reciprocal of that mean.
Quadratic mean[214]	Compute the square of all values, compute the mean of the squares, and then take the square root of that mean.
Skewness[211]	Quantifies how symmetrical the distribution is. A distribution that is symmetrical has a skewness of 0.
Kurtosis[211]	Quantifies whether the tails of the data distribution matches the Gaussian distribution. A Gaussian distribution has a kurtosis of 0.

5.2.3   Interpreting results: Quartiles and the interquartile range

What are percentiles?

Percentiles are useful for giving the relative standing of an individual in a group. Percentiles are essentially normalized ranks. The 80th percentile is a value where you'll find 80% of the values lower and 20% of the values higher. Percentiles are expressed in the same units as the data.

The median

The median is the 50th percentile. Half the values are higher; half are lower. Rank the values from low to high. If there are an odd number of points, the median is the one in the middle. If there are an even number of points, the median is the average of the two middle values.

Quartiles

Quartiles divide the data into four groups, each containing an equal number of values. Quartiles are divided by the 25th, 50th, and 75th percentile, also called the first, second and third quartile. One quarter of the values are less than or equal to the 25th percentile. Three quarters of the values are less than or equal to the 75th percentile.

Interquartile range

The difference between the 75th and 25th percentile is called the interquartile range. It is a useful way to quantify scatter.

Computing percentiles

Computing a percentile other than the median is not straightforward. Believe it or not, there are at least eight different methods to compute percentiles. Here is another explanation of different methods (scroll down to "plotting positions").

Prism computes percentile values by first evaluating this expression:

```
R = P * (n + 1)/100
```

P is the desired percentile (25 or 75 for quartiles) and n is the number of values in the data set. The result is the rank that corresponds to the percentile value. If there are 68 values, the 25th percentile corresponds to a rank equal to:

```
0.25 * 69 = 17.25
```

Prism (since version 5) interpolates one quarter of the way between the 17th and 18th value. This is the method most commonly used in stats programs. It is definition 6 in Hyndman and Fan (1) . With this method, the percentile of any point is k/(n+1), where k is the rank (starting at 1) and n is the sample size. This is not the same way that Excel computes percentiles, so percentiles computed by Prism and Excel will not match when sample sizes are small.

Beware of percentiles of tiny data sets. Consider this example: What is the 90th percentile of six values? Using the formula above, R equals 6.3. Since the largest value has a rank of 6, it is not really possible to compute a 90th percentile. Prism reports the largest value as the 90th percentile. A similar problem occurs if you try to compute the 10th percentile of six values. R equals 0.7, but the lowest value has a rank of 1. Prism reports the lowest value as the 10th percentile.

Note that there is no ambiguity about how to compute the median. All definitions of percentiles lead to the same result for the median.

Five-number summary

The term *five-number summary* is used to describe a list of five values: the minimum, the 25th percentile, the median, the 75th percentile, and the maximum. These are the same values plotted in a box-and-whiskers plots (when the whiskers extend to the minimum and maximum; Prism offers other ways to define the whiskers).

Reference

1. R.J. and Y. Fan, Sample quantiles in statistical packages, The American Statistician, 50: 361-365, 1996

5.2.4   Interpreting results: Mean, SD, SEM

Mean

The mean is the average. Add up the values, and divide by the number of values.

Standard Deviation

The standard deviation[42] (SD) quantifies variability. It is expressed in the same units as the data. It is often abbreviated as s. Prism computes the

SD using a denominator of n-1, so computes what is sometimes called [the sample SD rather than the population SD](#)[51] .

Standard Error of the Mean and Confidence Interval of the mean

The Standard Error of the Mean (SEM) quantifies the precision of the mean. It is a measure of how far your sample mean is likely to be from the true population mean. It is expressed in the same units as the data.

Learn about the [difference between SD and SEM](#)[51] and [when to use each](#)[52] .

The SEM is used to compute the confidence interval of the mean, and this CI is [easier to interpret](#)[62] . If the data are sampled from a Gaussian distribution, you can be 95% certain that the interval encloses the population mean.

Variance

The variance equals the SD squared, and therefore is expressed in the units of the data squared. Mathematicians like to think about variances because they can partition variances into different components -- the basis of ANOVA. In contrast, it is not correct to partition the SD into components. Because variance units are usually impossible to think about, most scientists avoid reporting the variance of data, and stick to standard deviations. Prism does not report the variance.

5.2.5   Interpreting results: Median and its CI

The median is the 50th percentile. Half the values are greater than (or equal to) the median and half are smaller.

The confidence interval of the median is computed by a standard method explained well in Zar (pages 548-549), based on the binomial distribution.

Four notes:

- The confidence interval of the median is not symmetrical around the median.

- You do not need to assume that the population distribution is symmetrical in order to interpret the confidence interval.

- The confidence interval begins and ends with values in the data set. No interpolation.

- Even if you ask for 95% confidence level, the actual confidence level will usually be different (especially with small samples) and Prism reports this.

95% CI of median	
Actual confidence level	96.88%
Lower confidence limit	23.
Upper confidence limit	54.

J.H. Zar, Biostatistical Analysis, Fifth edition 2010, ISBN: 0131008463.

5.2.6   Interpreting results: Coefficient of Variation

The *coefficient of variation* (CV), also known as "relative variability", equals the standard deviation divided by the mean. It can be expressed either as a fraction or a percent.

What is the advantage of reporting CV? The only advantage is that it lets you compare the scatter of variables expressed in different units. It wouldn't make sense to compare the SD of blood pressure with the SD of pulse rate, but it might make sense to compare the two CV values.

Notes:

- It only makes sense to report CV for variables, such as mass or enzyme activity, where "0.0" is defined to really mean zero. A weight of zero means no weight. An enzyme activity of zero means no enzyme activity. These are called *ratio variables*. It can be helpful to express variation of ratio variables (weights or enzyme activities...) as the CV. In contrast, a temperature of "0.0" does not mean zero temperature (unless measured in degrees Kelvin), so it would be meaningless to report a CV of values expressed as degrees C.

- It never makes sense to calculate the CV of a variable expressed as a logarithm because the definition of zero is arbitrary. The logarithm of 1

equals 0, so the log will equal zero whenever the actual value equals 1. By changing units, you'll redefine 1.0 in the original scale, so redefine zero on the log scale, and so redefine the CV. The CV of a logarithm is, therefore, meaningless. For example, it makes no sense to compute the CV of a set of pH values. pH is measured on a log scale (it is the negative logarithm of the concentration of hydrogen ions). A pH of 0.0 does not mean 'no pH', and certainly doesn't mean 'no acidity' (quite the opposite). Therefore it makes no sense to compute the CV of pH.

- When computing the CV, Prism computes the SD as the sample SD (using n-1 as the denominator) not the population SD (using n in the denominator).

5.2.7   Interpreting results: Geometric mean and its CI

How Prism computes the geometric mean

Compute the logarithm of all values, compute the mean of the logarithms, and then take the antilog. Prism uses base 10 (common) logarithms, and then takes ten to the power of the mean of the logarithms to get the geometric mean. This is equivalent to multiplying all the values together and taking that product to the 1/n power, where n is the number of values.

Geometric means are often used to average ratios.

It makes sense to use the geometric mean when the set of logarithms of the data form a symmetrical approximately Gaussian distribution.

How Prism computes the geometric SD

First, transform all the values to logarithms, compute the sample SD of those log values, and then take the antilogarithm of that SD. Prism uses base 10 (common) logarithms, and then takes ten to the power of the mean of the logarithms to get the geometric mean.

The geometric SD factor has no units. It is a unitless ratio.

It makes no sense to add the geometric SD to the geometric mean (or any other value), and makes equally no sense to ever subtract the geometric SD from the geometric mean. The geometric SD is a value you

always multiply or divide by. The range from (the geometric mean divided by the geometric SD factor) to (the geometric mean multiplied by the geometric SD factor) will contain about two thirds of the values if the data are sampled from a lognormal distribution. Similarly, the range from (the mean minus the SD) to (the mean plus the SD) will contain about two thirds of the values when data are sampled from a Gaussian distribution.

More about the geometric SD. [56]

How to report the geometric mean and SD

While it is common to see a data sampled from a Gaussian distribution reported as, "The mean is 3.2 ± 1.2 (SD)", it is currently rare to report data sampled from a lognormal distribution reported as, "The geometric mean is 4.3 *÷ 1.14." But that kind of reporting makes sense. Instead of using a symbol meaning "plus or minus" which makes sense for data sampled from a Gaussian distribution, use symbols meaning "times or divided by" when reporting results from data sampled from a lognormal distribution.

Example
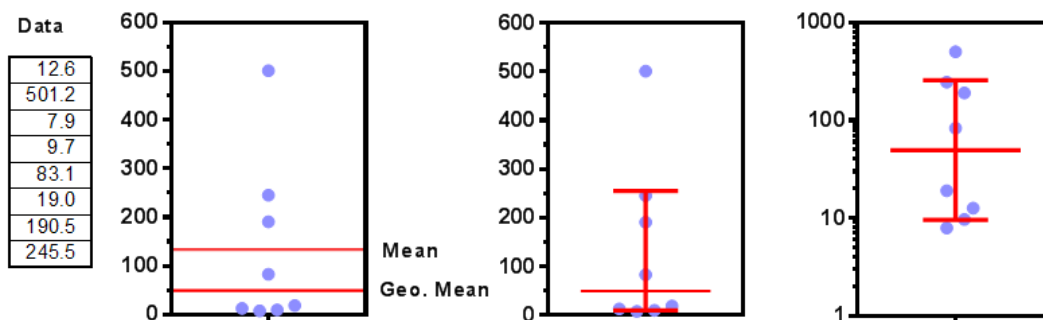
The example above shows eight values (so you can do the calculations yourself, if you want to). The geometric mean is 49.55 and the geometric SD factor is 5.15. The left graph shows the data with lines denoting the mean and geometric mean. The middle graph shows how Prism plots the geometric mean and geometric SD. The upper error bar extends up to the geometric mean times the geometric SD factor (49.55 * 5.15 = 255.2). The lower error bars extends down to the geometric mean divided by the

geometric SD factor (49.55 / 5.15 = 9.62). The right graph shows the data, the geometric mean, and the geometric SD plotted on a logarithmic axis. The log SD error bars appear visually symmetrical on a log axis, even though numerically they are very asymmetrical.

5.2.8   Interpreting results: Skewness

Key facts about skewness

Skewness quantifies how symmetrical the distribution is.

- A symmetrical distribution has a skewness of zero.

- An asymmetrical distribution with a long tail to the right (higher values) has a positive skew.

- An asymmetrical distribution with a long tail to the left (lower values) has a negative skew.

- The skewness is unitless.

- Any threshold or rule of thumb is arbitrary, but here is one: If the skewness is greater than 1.0 (or less than -1.0), the skewness is substantial and the distribution is far from symmetrical.

How skewness is computed

Skewness has been defined in multiple ways. The steps below explain the method used by Prism, called g1 (the most common method). It is identical to the skew() function in Excel.

1. We want to know about symmetry around the sample mean. So the first step is to subtract the sample mean from each value, The result will be positive for values greater than the mean, negative for values that are smaller than the mean, and zero for values that exactly equal the mean.

2. To compute a unitless measures of skewness, divide each of the differences computed in step 1 by the standard deviation of the values. These ratios (the difference between each value and the mean divided

by the standard deviation) are called z ratios. By definition, the average of these values is zero and their standard deviation is 1.

3. For each value, compute z^3. Note that cubing values preserves the sign. The cube of a positive value is still positive, and the cube of a negative value is still negative.

4. Average the list of z^3 by dividing the sum of those values by n-1, where n is the number of values in the sample. If the distribution is symmetrical, the positive and negative values will balance each other, and the average will be close to zero. If the distribution is not symmetrical, the average will be positive if the distribution is skewed to the right, and negative if skewed to the left. Why n-1 rather than n? For the same reason[45] that n-1 is used when computing the standard deviation.

5. Correct for bias. For reasons that I do not really understand, that average computed in step 4 is biased with small samples -- its absolute value is smaller than it should be. Correct for the bias by multiplying the mean of z^3 by the ratio n/(n-2). This correction increases the value if the skewness is positive, and makes the value more negative if the skewness is negative. With large samples, this correction is trivial. But with small samples, the correction is substantial.

 More on skewness and kurtosis

5.2.9   Interpreting results: Kurtosis

Kurtosis

Kurtosis quantifies whether the tails of the data distribution matches the Gaussian distribution.

• A Gaussian distribution has a kurtosis of 0.

• A distribution with fewer values in the tails than a Gaussian distribution has a negative kurtosis.

• A distribution with more values in the tails (or values further out in the tails) than a Gaussian distribution has a positive kurtosis.

- Kurtosis has no units.

- Although it is commonly thought to measure the shape of the peak, kurtosis actually tells you virtually nothing about the shape of the peak. Its only unambiguous interpretation is in terms of the values in the tail. Essentially it measures the presence of outliers (1).

- The value that Prism reports is sometimes called the **excess kurtosis** since the expected kurtosis for a Gaussian distribution is 0.0.

- An alternative definition of kurtosis is computed by adding 3 to the value reported by Prism. With this definition, a Gaussian distribution is expected to have a kurtosis of 3.0.

How Kurtosis is computed

1. Subtract the sample mean from each value, The result will be positive for values greater than the mean, negative for values that are smaller than the mean, and zero for values that exactly equal the mean.

2. Divide each of the differences computed in step 1 by the standard deviation of the values. These ratios (the difference between each value and the mean divided by the standard deviation) are called z ratios. By definition, the average of these values is zero and their standard deviation is 1.

3. For each value, compute z^4. In case that doesn't render well, that is z to the fourth power. All these values are positive.

4. Average that list of values by dividing the sum of those values by n-1, where n is the number of values in the sample. Why n-1 rather than n? For the same reason[45] that n-1 is used when computing the standard deviation.

5. With a Gaussian distribution, you expect that average to equal 3. Therefore, subtract 3 from that average. Gaussian data are expected to have a kurtosis of 0. This value (after subtracting 3) is sometimes called *excess kurtosis*.

Why don't values in the middle of the distribution affect the kurtosis very much?

Because the z values are taken to the fourth power, only large z values (so only values far from the mean) have a big impact on the kurtosis. If one value has a z value of 1 and another has a z value of 2, the second value will have 16 times more impact on the kurtosis (because 2 to the fourth power is 16). If one value has a z value of 1 and another has a z value of 3 (so is three times further from the mean), the second value will have 81 times more impact on the kurtosis (because 3 to the fourth power is 81). Accordingly, values near the mean (especially those less than one SD from the mean) have very little impact on the kurtosis, while values far from the mean have a huge impact. For this reason, the kurtosis does not quantify peakedness and does not really quantify the shape of the bulk of the distribution. Rather kurtosis quantifies the overall impact of points far from the mean.

Reference

1. Westfall, P. H. (2014). Kurtosis as Peakedness, 1905–2014. R.I.P. The American Statistician, 68(3), 191–195.

5.2.10 Harmonic, quadratic, trimmed, and winsorized mean

Harmonic mean

Prism computes the harmonic mean and its confidence interval by first transforming all the values to their reciprocals, and then computing the mean of those reciprocals and the CI of that mean. The harmonic mean is the reciprocal of that mean. If the values are all positive, larger numbers effectively get less weight than lower numbers. The harmonic means is most often used to find the average of a set of rates or velocities.

It makes sense to use the harmonic mean when the set of reciprocals of the data form a symmetrical approximately Gaussian distribution.

Quadratic mean

Prism computes the quadratic mean and its confidence interval by first transforming all the values to their square (value multiplied by itself), and then computing the mean of those squared values and the CI of that mean. The quadratic mean is the square root of that mean. The quadratic mean is also called the root mean square.

It makes sense to use the quadratic mean when the set of squares of the data form a symmetrical approximately Gaussian distribution.

Trimmed and Winsorized means

The idea of trimmed or Winsorized means is to not let the largest and smallest values have much impact. Before calculating a trimmed or Winsorized mean, you first have to choose how many of the largest and smallest values to ignore or down weight. If you set K to 1, the largest and smallest values are treated differently. If you set K to 2, then the two largest and two smallest values are treated differently. K must be set in advance. Sometimes K is set to 1, other times to some small fraction of the number of values, so K is larger when you have lots of data.

To compute a trimmed mean, simply delete the K smallest and K largest observations, and compute the mean of the remaining data.

To compute a Winsorized mean, replace the K smallest values with the value at the K+1 position, and replace the k largest values with the value at the N-K-1 position. Then take the mean of the data. .

The advantage of trimmed and Winsorized means is that they are not influenced by one (or a few) very high or low values. Prism does not compute these values.

Mode

The mode is the value that occurs most commonly. It is not useful with measured values assessed with at least several digits of accuracy, as most values will be unique. It can be useful with variables that can only have integer values. While the mode is often included in lists like this, the mode doesn't always assess the center of a distribution. Imagine a medical survey where one of the questions is "How many times have you had surgery?" In many populations, the most common answer will be zero, so that is the mode. In this case, some values will be higher than the mode, but none lower, so the mode is not a way to quantify the center of the distribution.

5.3   Row statistics

5.3.1 Overview: Side-by-side replicates

When entering data into tables formatted for XY or Grouped data, replicates go into side-by-side subcolumns. Prism then can plot these individually, or plot mean and error bar.

You can also format the table to enter mean, SD or SEM, and N. This is useful if you have already averaged the data in another program or if you have more than 52 replicates. Otherwise, it is best to enter the raw data into Prism, so you can plot every replicate.

Prism can take your raw data, and create graphs with mean (or median) and error bars (defined in several ways). There is no need to run an analysis to compute the SD or SEM. But if you want to see the descriptive stats for each set of replicates, use the [Row Means and Totals](#)[216] analysis.

5.3.2 Row means and totals
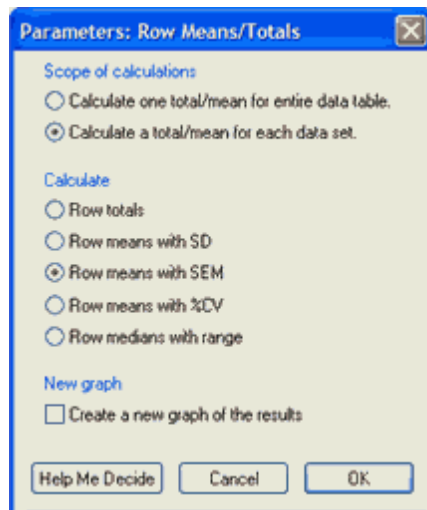
Calculate a total/mean for each data set

If you enter data onto XY or two-way tables with replicate Y values in subcolumns, and want to view a table of mean and SD (or SEM) values, click Analyze and choose to do a built-in analysis. Then choose Row means/totals, and choose one total/mean for each data set.

Note that you rarely need this analysis. Prism will automatically create graphs with the mean and SD (or SEM). You don't have to choose any analyses to make these graphs. Prism computes the error bars

automatically. Use settings on the Format Graph dialog (double-click on any symbol to see it) to plot individual points or to choose SD, SEM, 95% CI or range error bars. The only purpose of this analysis is if you want to see the SD or SEM values.

Calculate a total/mean for the entire data table

This choice is used rarely, but it helps you consolidate a larger table into a single data set.

If the data were entered onto a single subcolumn for each data set, then there is no ambiguity.

But what if you entered data on a table with subcolumns? In this case, the calculated total/mean values are based on the mean value of each data set for corresponding row. For example, let's calculate "Row means with SD" for three datasets with 3, 2 and 3 replicates in each. Here are the data for the first row.

```
Data set A: (2, 3, 4)
Data set B: (4, 6)
Data set C: (7, 8, 9)
```

Prism will first compute the mean values for each data set, which are 3, 5 and 8. It then computes the grand mean of those three values (and their standard deviation) so the results are Mean = 5.333, SD = 2.517, N = 3.

If Prism simply looked at those data as eight independent values, the mean would be 5.375, but since the values in different data sets are unlikely to be independent, Prism does not do this calculation.

Note:

- When Prism computes the grand mean, it does not account for the fact that data sets A and C are in triplicate, while data set B is only in duplicate.

- If you entered your data as mean, n and SD or SEM, these calculations use only the mean value you entered and ignore n and SD or SEM.

5.4 Frequency Distributions

5.4.1 Visualizing scatter and testing for normality without a frequency distribution

Viewing data distributions

Before creating a frequency distribution, think about whether you actually need to create one.

In many cases, plotting a column scatter graph is all you need to do to see the distribution of data.

- The graph on the left is a column scatter plot (with line drawn at the mean) made from the "Frequency distribution" sample data.

- The graph in the middle is a box-and-whiskers graph of the same data, showing the values lower than the 2.5th percentile and greater than the 97.5 th percentile as circles. Note that Prism offers several choices for how to define the whiskers in this kind of plot.

- The graph on the right is a violin plot, new with Prism 8. It shows the median and quartiles as the box-and-whisker plot does, but also shows a smoothed frequency distribution to give you a sense for how the data are distributed.

All three graphs were created by Prism directly from the column data table, with no analysis needed.
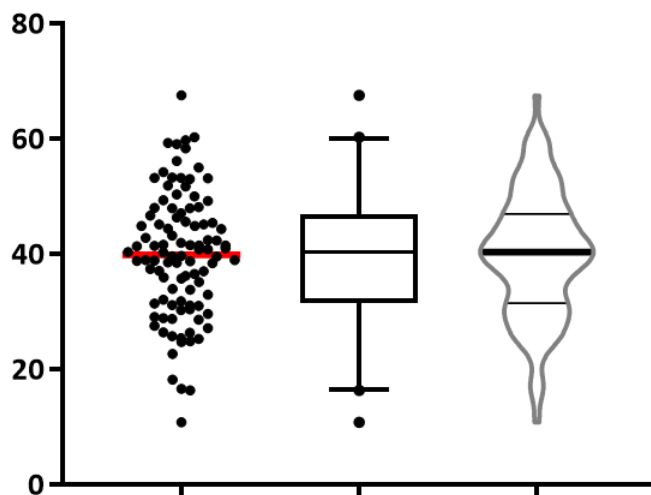
Testing for normality

Prism can [test for normality](#)⌐ [239] as part of the column statistics analysis. You don't have to create a frequency distribution, and then fit a Gaussian distribution.

5.4.2   How to: Frequency distribution

1. Enter data

Choose a Column table, and a column scatter graph. If you are not ready to enter your own data, choose the sample data set: Frequency distribution data and histogram.

2. Choose the analysis

Click Analyze and then choose Frequency distribution from the list of analyses for Column data.

3. Choose analysis options

Cumulative?

In a *frequency distribution*, each bin contains the number of values that lie within the range of values that define the bin. In a *cumulative distribution*, each bin contains the number of values that fall within *or below* that bin. By definition, the last bin contains the total number of values. The graph below shows a frequency distribution on the left, and a cumulative distribution of the same data on the right, both plotting the number of values in each bin.
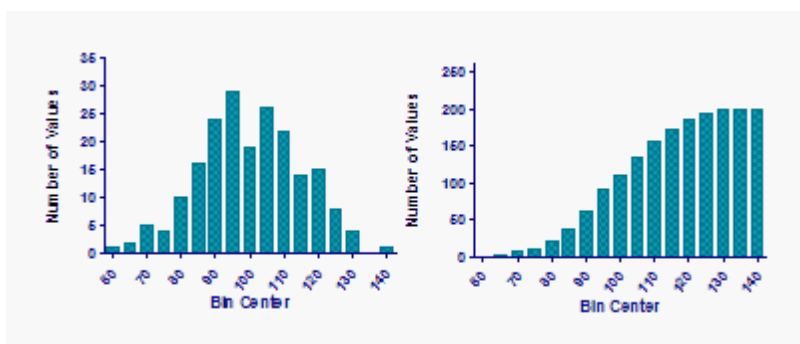
The main advantage of cumulative distributions is that you don't need to decide on a bin width. Instead, you can tabulate the exact cumulative distribution as shown below. The data set had 250 values, so this exact cumulative distribution has 250 points, making it a bit ragged. When you choose to tabulate a cumulative frequency distributions as percentages rather than fractions, those percentages are really percentiles and the resulting graph is sometimes called a *percentile plot*.
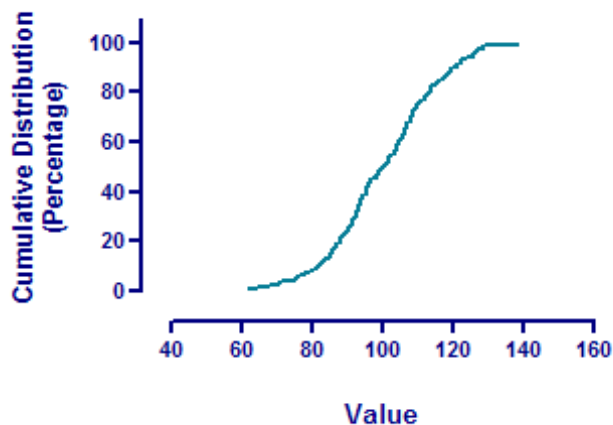
Relative or absolute frequencies?

Select Relative frequencies to determine the fraction (or percent) of values in each bin, rather than the actual number of values in each bin. For example, if 15 of 45 values fall into a bin, the relative frequency is 0.33 or 33%.

If you choose both cumulative and relative frequencies, you can plot the distribution using a probabilities axis. When graphed this way, a Gaussian distribution is linear.

Bin width

If you chose a cumulative frequency distributions, we suggest that you choose to create an exact distribution. In this case, you don't choose a bin width as each value is plotted individually.

To create an ordinary frequency distribution, you must decide on a bin width. If the bin width is too large, there will only be a few bins, so you will not get a good sense of how the values distribute. If the bin width is too low, many bins might have only a few values (or none) and so the number of values in adjacent bins can randomly fluctuate so much that you will not get a sense of how the data are distributed.

How many bins do you need? Partly it depends on your goals. And partly it depends on sample size. If you have a large sample, you can have more bins and still have a smooth frequency distribution. One rule of thumb is aim for a number of bins equal to the log base 2 of sample size. Prism uses this as one of its two goals when it generates an automatic bin width (the other goal is to make the bin width be a round number).

The figures below show the same data with three different bin widths. The graph in the middle displays the distribution of the data. The one on the left has too little detail, while the one on the right has too much detail.
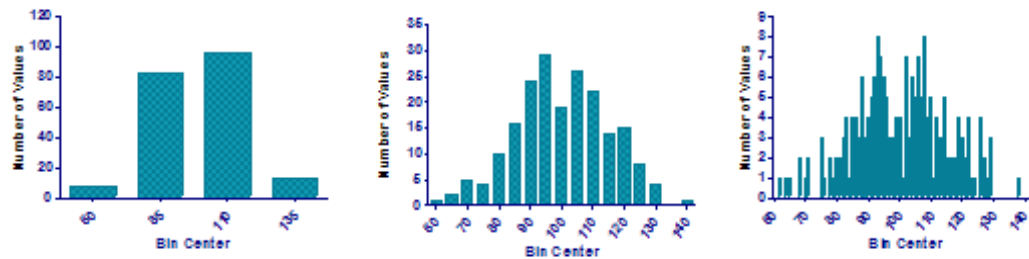
Bin range

In addition to deciding on the bin width, which controls the number of bins, you can also choose the center of the first bin. This can be important. Imagine that your data are percentages, running from 0 to 100. There is no possibility of a value that is less than 0 (negative) or greater than 100. Let's say you want the bin width to be 10, to make 10 bins. If the first bin is centered at 0, it will contain values between -5 and 5, the next bin will contain values between 5 and 15, the next between 15 and 25, etc. Since negative values are impossible, the first bin actually includes values only between 0 and 5, so its effective bin width is half the other bin widths. Also note, there are eleven bins that contain data, not ten.

If you instead make the first bin centered at 5, it will contain values between 0 and 10, the next bin contains values from 10 to 20, etc. Now, all bins truly contain the same range of values, and all the data are contained within ten bins.

A point on the border goes with the bin holding the larger values. So if one bin goes from 3.5 to 4.5 and the next from 4.5 to 5.5, a value of 4.5 ends up in that second bin (from 4.5 to 5.5).

Replicates

If you entered replicate values, Prism can either place each replicate into its appropriate bin, or average the replicates and only place the mean into a bin.

All values too small to fit in the first bin are omitted from the analysis. You can also enter an upper limit to omit larger values from the analysis.

How to graph

See [these examples][223].

> Prism can only make frequency distributions from numerical data. It can handle categorical data, but only if the categories are entered as values.

5.4.3 Graphing tips: Frequency distributions

At the bottom of the frequency distribution analysis dialog, you can choose among several ways to graph the resulting data. These are all shown below, using 'frequency distribution' sample data set.

Graphs of frequency distributions

If you don't create a cumulative distribution, Prism gives you three choices illustrated below: XY graph with points, XY graph with spikes (bars). or a bar graph
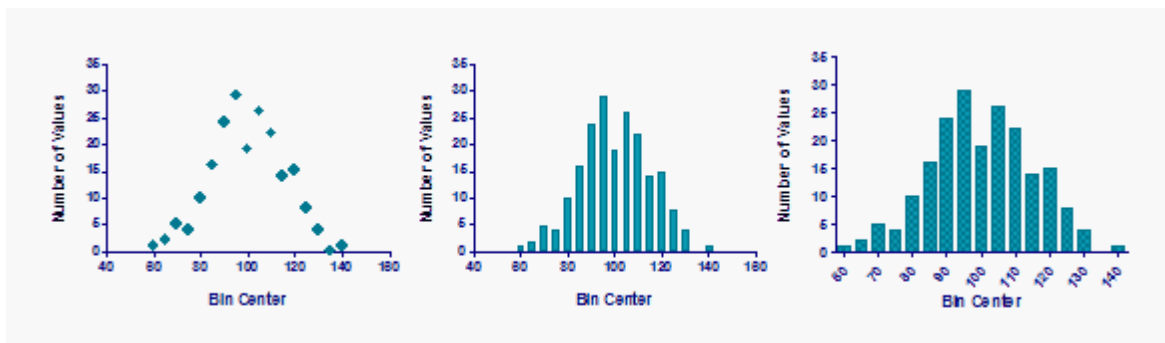
The last two graphs look very similar, but the graph on the right is a bar graph, while the one in the middle is an XY graph plotting bars or spikes instead of symbols. The graph in the middle has X values so you can [fit a Gaussian distribution][225] to it. The graph on the right has no X values (just category names, which happen to be numbers), so it is not possible to fit a curve.

The term *histogram* is used inconsistently. We use the term to mean a graph of a frequency distribution which is usually a bar graph. Some people use the term *histogram* to refer to any bar graph, even those that don't plot frequency distributions.

Graphs of cumulative frequency distributions

If you choose a cumulative frequency distribution that tabulates the actual number of values (rather than fractions or percents), Prism can only create one kind of graph:
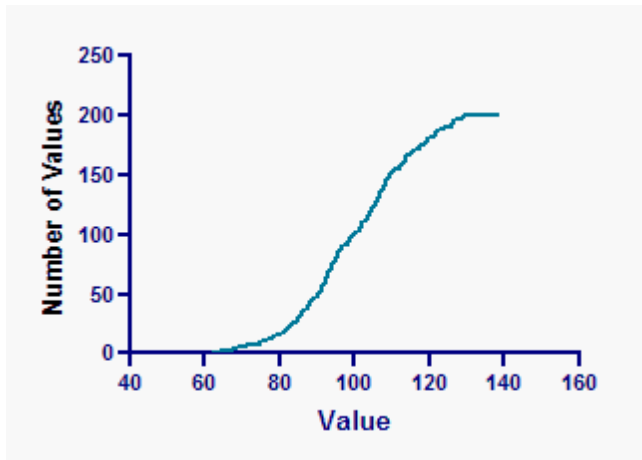
If you choose to tabulate the results as fractions or percentages, then Prism also offers you (from the bottom part of the Parameters dialog for frequency distributions) the choice of plotting on a probability axis. If your data were drawn from a Gaussian distribution, they will appear linear when the cumulative distribution is plotted on a probability axis. Prism uses standard values to label the Y axis, and you cannot adjust these. This graph is very similar to a Q-Q plot.
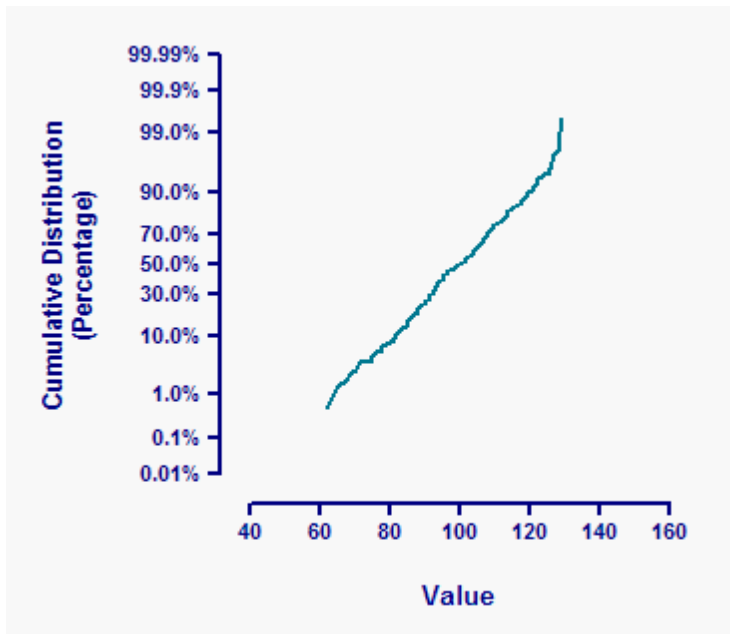
5.4.4   Fitting a Gaussian distribution to a frequency distribution

Why fit a Gaussian distribution to your data?

Does your data follow a Gaussian distribution? One way to answer that question is to <u>perform a normality test</u>[201] on the raw data. Another approach is to examine the frequency distribution or the cumulative frequency distribution.

Fitting a Gaussian distribution

Prism can superimpose a frequency distribution over the histogram.
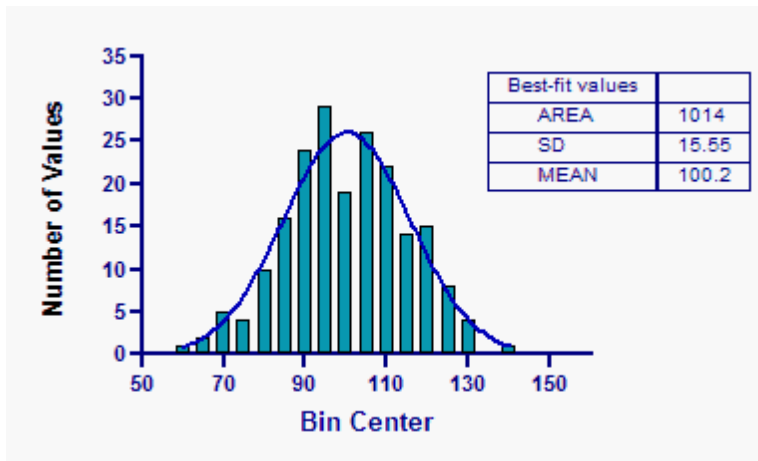
Follow these steps:

1. In the frequency distribution dialog, choose to create the frequency distribution (not a cumulative distribution). Also choose to plot the data as an XY graph of histogram spikes.
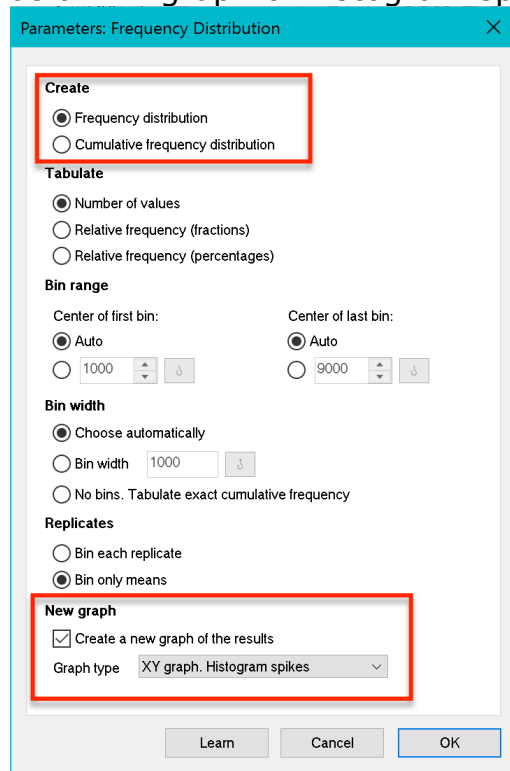
2. Go to the new graph.

3. Click Analyze, and choose nonlinear regression. On the first tab of the model, choose the Gaussian family of equations and then the Gaussian

equation. All the other choices on the nonlinear regression dialog can be left to their default settings.

The results depend to some degree on which value you picked for bin width, so we recommend fitting the cumulative distribution as explained below.

Fitting a cumulative Gaussian distribution

Prism can superimpose a cumulative Gaussian distribution over a graph of the cumulative distribution of the data. The cumulative Gaussian distribution has a sigmoidal shape.
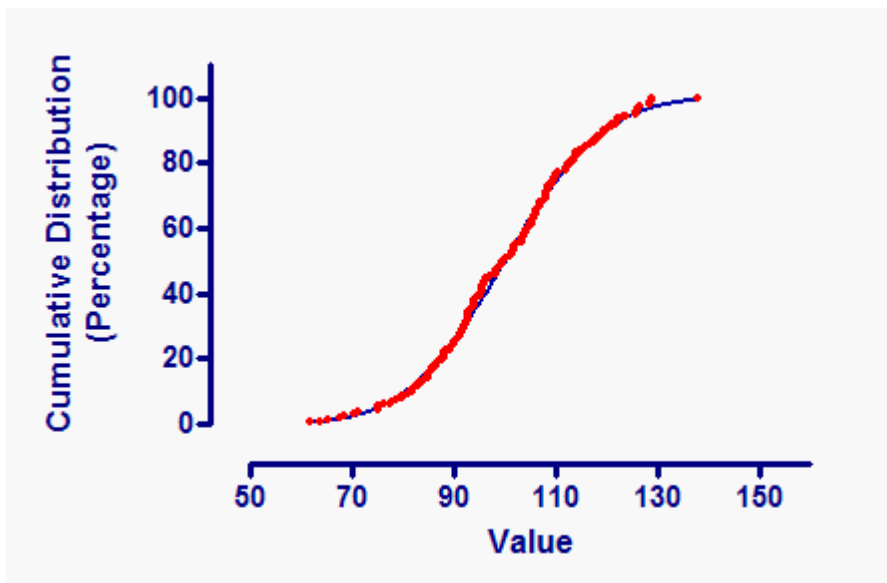
Follow these steps.

1. In the frequency distribution dialog, choose to create the cumulative frequency distribution. Also choose to plot the data as an XY graph of

points.

Parameters: Frequency Distribution

Create
- ○ Frequency distribution
- ● Cumulative frequency distribution

Tabulate
- ● Number of values
- ○ Relative frequency (fractions)
- ○ Relative frequency (percentages)

Bin range

Center of first bin: Center of last bin:
- ● Auto   ● Auto
- ○ 0   ○ 9000

Bin width
- ● Choose automatically
- ○ Bin width 1000
- ○ No bins. Tabulate exact cumulative frequency

Replicates
- ○ Bin each replicate
- ● Bin only means

New graph
- ☑ Create a new graph of the results

Graph type   XY graph. Points

[Learn]   [Cancel]   [OK]

2. Go to the new graph.

3. Click Analyze, choose nonlinear regression, and choose the one of the cumulative Gaussian models from the selection of Gaussian models. Prism offers separate models to use for data expressed as percentages, fractions or number of observations. With the last choice, you should constrain N to a constant value equal to the number of values. You can leave all other choices set to their default values.

The graph shown above the cumulative distribution of the sample data (in percents) fit to the cumulative Gaussian curve. The observed distribution is plotted with red circles and the fit distribution is a blue curve. The two are superimposed, so hard to distinguish.

Plotting on a probability axis

Below, the same graph is plotted using a probability Y axis. To do this, double-click on the Y axis to bring up the Format Axis dialog, drop down the choices for scale in the upper right corner, and choose "Probability (0..100%). The cumulative Gaussian distribution is linear when plotted on probability axes. At the top right of the graph, the cumulative distribution

is a bit higher than predicted by a Gaussian distribution. This discrepancy is greatly exaggerated when you plot on a probability axis.
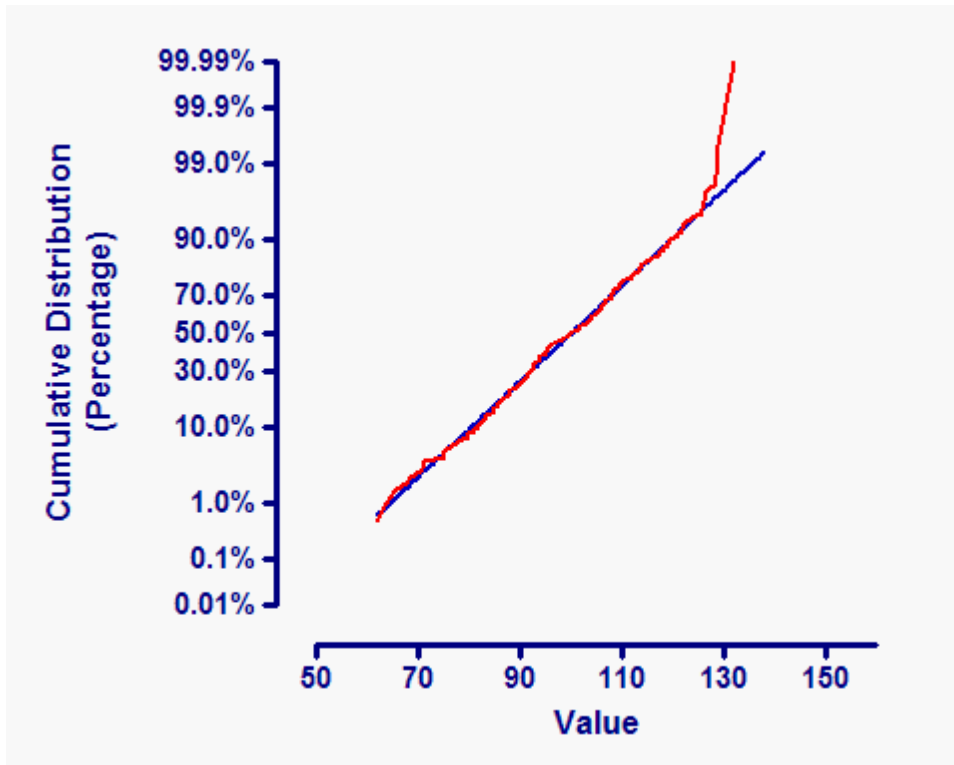
5.5   Analyzing curves

5.5.1   Smoothing, differentiating and integrating curves

A single Prism analysis smooths a curve and/or converts a curve to its derivative or integral.
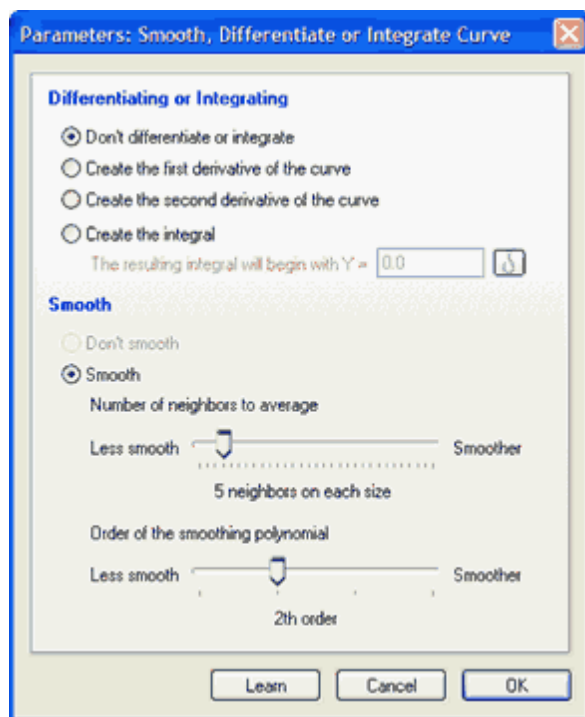
Finding the derivative or integral of a curve

The first **derivative** is the steepness of the curve at every X value. The derivative is positive when the curve heads uphill and is negative when the curve heads downhill. The derivative equals zero at peaks and troughs in the curve. After calculating the numerical derivative, Prism can smooth the results, if you choose.

The **second derivative** is the derivative of the derivative curve. The second derivative equals zero at the inflection points of the curve.

The **integral** is the cumulative area between the curve and the line at Y=0, or some other value you enter.

Notes:

- Prism cannot do symbolic algebra or calculus. If you give Prism a series of XY points that define a curve, it can compute the numerical derivative (or integral) from that series of points. But if you give Prism an equation, it cannot compute a new equation that defines the derivative or integral.

- This analysis integrates a curve, resulting in another curve showing cumulative area. Don't confuse with a separate Prism analysis that computes a single value for the [area under the curve](#) [233].

Smoothing a curve

If you import a curve from an instrument, you may wish to smooth the data to improve the appearance of a graph. Since you lose data when you smooth a curve, you should not smooth a curve prior to nonlinear regression or other analyses. Smoothing is not a method of data analysis, but is purely a way to create a more attractive graph.

Prism gives you two ways to adjust the smoothness of the curve. You choose the number of neighboring points to average and the 'order' of the smoothing polynomial. Since the only goal of smoothing is to make the curve look better, you can simply try a few settings until you like the appearance of the results. If the settings are too high, you lose some peaks which get smoothed away. If the settings are too low, the curve is not smooth enough. The right balance is subjective -- use trial and error.

The results table has fewer rows than the original data.

Don't analyze smoothed data

Smoothing a curve can be misleading. The whole idea is to reduce the "fuzz" so you can see the actual trends. The problem is that you can see "trends" that don't really exist. The three graphs in the upper row below are simulated data. Each value is drawn from a Gaussian distribution with a mean of 50 and a standard deviation of 10. Each value is independently drawn from that distribution, without regard to the previous values. When you inspect those three graphs, you see random scatter around a horizontal line, which is exactly how the data were generated.
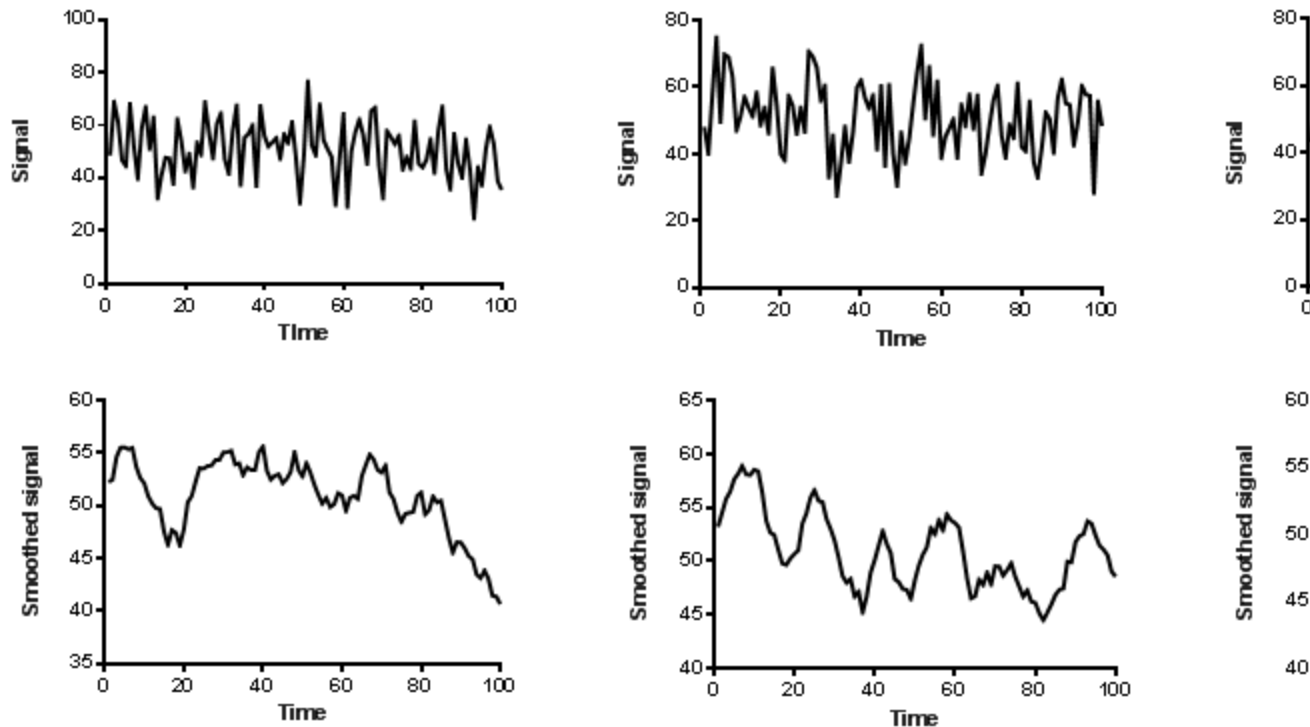
The bottom three graphs above show the same data after smoothing (averaging 10 values on each side, and using a second order smoothing polynomial). When you look at these graphs, you see trends. The first one tends to trend down. The second one seems to oscillate in a regular way. The third graph tends to increase. All these trends are artefacts of smoothing. Each graph shows the same data as the graph just above it.

Smoothing the data creates the impression of trends by ensuring that any large random swing to a high or low value is amplified, while the point-to-point variability is muted. A key assumption of correlation, linear regression and nonlinear regression is that the data are independent of each other. With smoothed data, this assumption is not true. If a value happens to be super high or low, so will the neighboring points after smoothing. Since random trends are amplified and random scatter is muted, any analysis of smoothed data (that doesn't account for the smoothing) will be invalid.

Mathematical details

- The first derivative is calculated as follows (x, and Y are the arrays of data; x' and y' are the arrays that contain the results).

```
x'[i] = (x[i+1] + x[i]) / 2
y' at x'[i] = (y[i+1] – y[i]) / (x[i+1] – x[i])
```

- The second derivative is computed by running that algorithm twice, to essentially compute the first derivative of the first derivative.

- Prism uses the [trapezoid rule](#)[233] to integrate curves. The X values of the results are the same as the X values of the data you are analyzing. The first Y value of the results equals a value you specify (usually 0.0). For other rows, the resulting Y value equals the previous result plus the area added to the curve by adding this point. This area equals the difference between X values times the average of the previous and this Y value.

- Smoothing is done by the method of Savistsky and Golay (1).

- If you request that Prism both both smooth and convert to a derivative (first or second order) or integral, Prism does the steps sequentially. First it creates the derivative or integral, and then it smooths.

Reference

1. A. Savitzky and M.J.E. Golay, (1964). [Smoothing and Differentiation of Data by Simplified Least Squares Procedures](#). Analytical Chemistry 36 (8): 1627–1639

5.5.2   Area under the curve

How to: Area under the curve

The area under the curve is an integrated measurement of a measurable effect or phenomenon. It is used as a cumulative measurement of drug effect in pharmacokinetics and as a means to compare peaks in chromatography.

Note that Prism also computes the area under a Receiver Operator Characteristic (ROC) curve as part of the [separate ROC analysis](#)[627].

Start from a data or results table that represents a curve. Click Analyze and choose Area under the curve from the list of XY analyses.
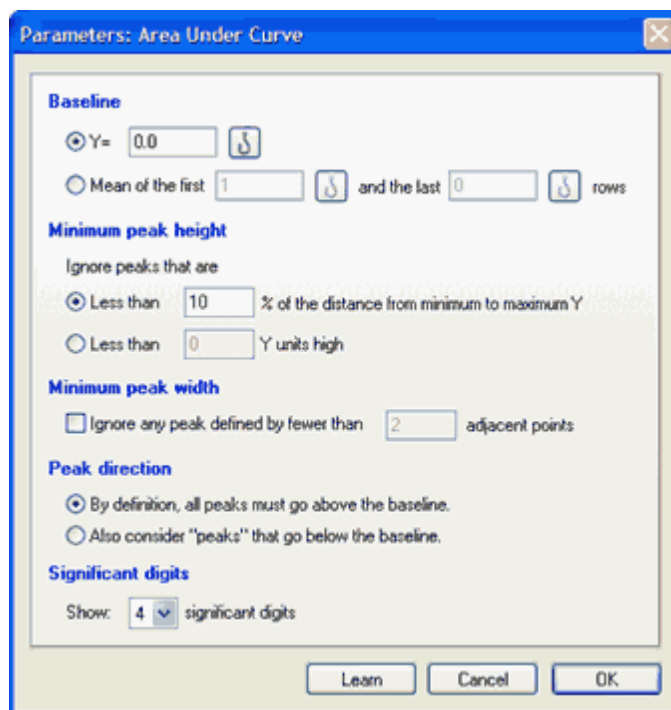
Interpreting area-under-the-curve results

If your data come from chromatography or spectroscopy, Prism can break the data into separate regions and determine the highest point (peak) of each. Prism can only do this, however, if the regions are clearly defined: the signal, or graphic representation of the effect or phenomenon, must go below the baseline between regions and the peaks cannot overlap.

For each region, Prism shows the area in units of the X axis times units of the Y axis. Prism also shows each region as a fraction of the total area under all regions combined. The area is computed using the trapezoid rule. It simply connects a straight line between every set of adjacent points defining the curve, and sums up the areas beneath these areas.

Next, Prism identifies the peak of each region. This is reported as the X and Y coordinates of the highest point in the region and the two X coordinates that represent the beginning and end of the region.

Prism may identify more regions than you are interested in. In this case, go back to the Parameters dialog box and enter a larger value for the minimum width of a region and/or the minimum height of a peak.

Limitations of this analysis

Note these limitations:

- The baseline must be horizontal.

- There is no smoothing or curve fitting.

- Prism will not separate overlapping peaks. The program will not distinguish two adjacent peaks unless the signal descends all the way to the baseline between those two peaks. Likewise, Prism will not identify a peak within a shoulder of another peak.

- If the signal starts (or ends) above the baseline, the first (or last) peak will be incomplete. Prism will report the area under the tails it "sees".

- Prism does not extrapolate back to X=0, if your first X value is greater than zero.

- Prism does not extrapolate beyond the highest X value in your data set, so does not extrapolate the curve down to the baseline.

- Prism no longer insists that the X values be equally spaced. When it sums the areas of the trapezoids, it is fine if some are fatter than others.

How Prism computes area under the curve

Prism computes the area under the curve using the trapezoid rule, illustrated in the figure below.
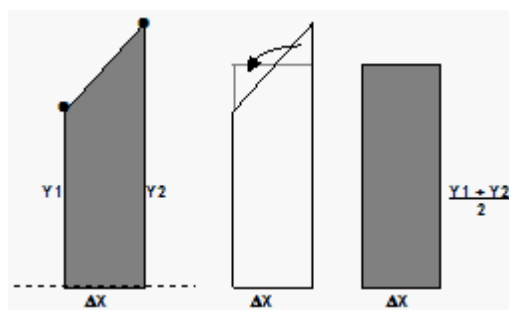
In Prism, a curve (created by nonlinear regression) is simply a series of connected XY points, with equally spaced X values. Prism can compute

area under the curve also for XY tables you enter, and does not insist that the X values be equally spaced. The left part of the figure above shows two of these points and the baseline as a dotted line. The area under that portion of the curve, a trapezoid, is shaded. The middle portion of the figure shows how Prism computes the area. The two triangles in the middle panel have the same area, so the area of the trapezoid on the left is the same as the area of the rectangle on the right (whose area is easier to calculate). The area, therefore, is $\Delta X*([(Y1+Y2)/2]-Baseline]$. Prism uses this formula repeatedly for each adjacent pair of points defining the curve.

The area is computed using the baseline you specify and the curve between two X values. Which X values?

- If all your data points are larger than the baseline, the AUC calculations start at the lowest X value in your data set and end at the largest X value. Note that Prism does not extend the curve beyond the X range of your data.

- If the Y values at the lowest X values are below your baseline: Prism finds the smallest X value in your data associated with a Y value greater than the baseline. It draws a line between that point and the point with the next smallest X value in your data set. It then uses linear interpolation to find where that line crosses the baseline, and uses that interpolated value as the first X value to compute the AUC.

- If the Y values at the largest X values are below your baseline: Prism finds the largest X value in your data associated with a Y value greater than the baseline. It draws a line between that point and the point with the next largest X value in your data set. It then uses linear interpolation to find where that line crosses the baseline, and uses that interpolated value as the last X value to compute the AUC.

The standard error and confidence interval of the AUC

If you enter data with replicate Y values, or as Mean and SD or SEM, Prism reports a SE and confidence interval for the AUC using the method described by Gagnon (1). If you entered replicate Y values in subcolumns, Prism assumes these are replicate measurements in one experiment. If each subcolumn is in fact a different repeated experiment, Prism does not compute one AUC per subcolumn, and then average those values. The 95% confidence interval equals the AUC plus or minus 1.96 times the SE. It uses the z distribution (so always 1.96) rather than the t distribution

(where the value would depend on sample size) because this was used in references 1-3. With more than a few dozen points defining the curve, the t and z methods will be nearly indistinguishable.

Prism does not compare peaks to provide a confidence interval for the difference or the corresponding P value. But you can get Prism to do this with a bit of work:

1. Create a new Grouped table, formatted for entry of mean, sem and n. You will enter values only into the first row of this table.

2. Enter the AUC values as means.

3. Enter the SE of the AUC values as "SEM".

4. Define the df for each group as the number of data points for that group minus the number of concentrations.

5. For n, enter one more than the df. When Prism does the t tests, it will subtract 1 from the entered n to obtain the df, which will now be correct.

6. Click analyze and choose the t test if you want to compare two AUCs, or one-way ANOVA if you want to compare three or more.

More details and full example.

What counts as a peak?

By default, Prism only considers points above the baseline to be part of peaks, so only reports peaks that stick above the baseline. You can choose to consider peaks that go below the baseline.

By default, Prism ignores any peaks whose height is less than 10% of the distance from minimum to maximum Y value, but you can change this definition in the area under the curve parameters dialog. You can also tell it to ignore peaks that are very narrow.

Total peak area vs. total area vs. net area

Prism reports the area under the peaks in two or three ways:

- Total Area. This sums positive peaks, negative peaks, peaks that are not high enough to count, and peaks that are too narrow to count. The only choice you make in the analysis dialog that affects the definition of total area is the definition of the baseline.

- Total Peak Area. The sum of the peaks you asked Prism to consider. This value is affected by several choices in the analysis dialog: The definition of baseline, your choice about including or ignoring negative peaks, and your definition of peaks too small to count.

- Net Area. You'll only see this value if you ask Prism to define peaks below the baseline as peaks. It is the difference computed by subtracting the area of peaks below the baseline from the area of peaks above the baseline.

Reference

1. Robert C. Gagnon and John J. Peterson, Estimation of Confidence Intervals for Area Under the Curve from Destructively Obtained Pharmacokinetic Data, Journal of Pharmacokinetics and Pharmacodynamics, 26: 87-102, 1998.

2. Bailer A. J. (1988). Testing for the equality of area under the curves when using destructive measurement techniques. Journal of Pharmacokinetics and Biopharmaceutics, 16(3):303-309.

3. Jaki T. and Wolfsegger M. J. (2009). A theoretical framework for estimation of AUCs in complete and incomplete sampling designs. Statistics in Biopharmaceutical Research, 1(2):176-184.

5.6   Normality (and lognormality) tests

Prism can test for normality as part of the Column Statistics analysis. It can also test for normality of residuals from nonlinear regression, as part of the nonlinear regression analysis.

5.6.1   How to: Normality test

Analyzing column data

1. Create a Column data table so each data set is in a single Y column.

2. Click Analyze, look at the list of Column analyses, and choose normality tests.

3. Prism offers four options for testing for normality. Choose[240] one, or more than one, of these options. You may also choose to test for lognormality[241] and to compare[242] normal and lognormal distributions.

Parameters: Normality and Lognormality Tests   ✕

Which distribution(s) to test?
☑ Normal (Gaussian) distribution

☐ Lognormal distribution

☐ Compute the relative likelihood of sampling from a Gaussian (normal) vs. a lognormal distribution (assuming no other possibilities)

Methods to test distribution(s)
☑ Anderson-Darling test

☑ D'Agostino-Pearson omnibus normality test

☑ Shapiro-Wilk normality test

☑ Kolmogorov-Smirnov normality test with Dallal-Wilkinson-Lilliefor P value

Graphing options
☑ Create a QQ plot

Subcolumns
◉ Average the replicates in each row, and then perform the calculation for each column

○ Perform calculations on each subcolumn separatelly

○ Treat all the values in all subcolumns as single set of data

Calculations
Significance level (alpha)   0.05

Output
P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0 ⌄

Show 4 ▲▼ significant digits.

☐ Make these choices the default for future analyses.

Learn   Cancel   OK

Analyzing normality of residuals from nonlinear regression

A residual is the distance of a point from the best-fit curve. One of the assumptions of linear and nonlinear regression is that the residuals follow a Gaussian distribution. You can test this with Prism. When setting up the nonlinear regression, go to the Diagnostics tab, and choose one (or more than one) of the normality tests.

Analyzing normality of residuals from linear regression

Prism's linear regression analysis does not offer the choice of testing the residuals for normality. But this limitation is easy to work around. Run nonlinear regression, choose a straight line model, and you'll get the same results as linear regression with the opportunity to choose normality testing. This is just one of many reasons to fit straight lines using the nonlinear regression analysis.

5.6.2   Choosing a normality test

Prism offers four normality tests. Why is there more than one way to test normality? There are many ways a distribution can deviate from a Gaussian distribution, so different normality tests give different results.

We recommend the **D'Agostino-Pearson** normality test. It first computes the skewness and kurtosis[211] to quantify how far the distribution is from Gaussian in terms of asymmetry and shape. It then calculates how far each of these values differs from the value expected with a Gaussian distribution, and computes a single P value from the sum of these discrepancies. It is a versatile and powerful normality test, and is recommended. Note that D'Agostino developed several normality tests. The one used by Prism is the "omnibus K2" test.

An alternative is the **Anderson-Darling test**. It computes the P value by comparing the cumulative distribution of your data set against the ideal cumulative distribution of a Gaussian distribution. It takes into account the discrepancies at all parts of the cumulative distribution curve (unlike the Kolmogorov-Smirnov test, see below).

Another alternative is the **Shapiro-Wilk** normality test. We prefer the D'Agostino-Pearson test for two reasons. One reason is that, while the Shapiro-Wilk test works very well if every value is unique, it does not work as well when several values are identical. The other reason is that

the basis of the test is hard to understand. There are several ways to compute the Shapiro-Wilk test. Prism uses the method of Royston (1).

Earlier versions of Prism offered only the **Kolmogorov-Smirnov** test. We still offer this test (for consistency) but no longer recommend it. It computes a P value from a single value: the largest discrepancy between the cumulative distribution of the data and a cumulative Gaussian distribution. This is not a very sensitive way to assess normality, and we now agree with this statement[1]: *"The Kolmogorov-Smirnov test is only a historical curiosity. It should never be used." (2).* Note that both this test and the Anderson-Darline test compare the actual and ideal cumulative distributions. The distinction is that Anderson-Darling considers the discrepancies at all parts of the curve, and Kolmogorov-Smirnov only look at the largest discrepancy.

The Kolmogorov-Smirnov method as originally published assumes that you know the mean and SD of the overall population (perhaps from prior work). When analyzing data, you rarely know the overall population mean and SD. You only know the mean and SD of your sample. To compute the P value, therefore, Prism uses the Dallal and Wilkinson approximation to Lilliefors' method (3). Since that method is only accurate with small P values, Prism simply reports "P>0.10" for large P values. In case you encounter any discrepancies, you should know that we fixed a bug in this test many years ago in Prism 4.01 and 4.0b.

Reference

1. P Royston, Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 44, No. 4 (1995), pp. 547-551

2. RB D'Agostino, "Tests for Normal Distribution" in *Goodness-Of-Fit Techniques* edited by RB D'Agostino and MA Stephens, Macel Dekker, 1986.

3. Dallal GE and Wilkinson L (1986), "An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality," The American Statistician, 40, 294-296.

5.6.3   Testing for lognormality

Prism can test the fit of your distribution to a lognormal distribution, using four different lognormality tests. It does so simply. It first computes the

logarithm of all the values, and then does the normality test(s) on the logarithms.

If any of the values are zero or negative, the distribution cannot possibly be lognormal, so Prism won't perform the lognormality tests.

5.6.4   Comparing normal and lognormal distributions

Prism compares the normal and lognormal distributions using a likelihood test, and computes the relative likelihood that the data were sampled from each. Notes:

- A lognormal distribution only contains positive numbers. Negative values and zeroes are impossible in lognormal distributions. If any values are zero or negative Prism does test for lognormality.

- Prism only fits those two distributions, and gives the percentage chance that each is more likely to be the distribution from which the data were sampled. Of course, there are an infinite number of other distributions the data could be sampled from. Prism only asks which is more likely, normal or lognormal. It won't notice if neither is very likely!

- Lognormal distributions are common in biology, so you'd think it would be common to ask whether data are more likely to be sampled from normal (Gaussian) or lognormal distributions. In fact, this comparison is done rarely. Prism (as of 2017) seems to be unique in making this test simple.

- For math details, see section 6.7.2 of Burnham and Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd edition. Basically, Prism fits a normal or lognormal distribution using maximum likelihood method, and then compares the two likelihoods. They point out that this is equivalent to comparing the AIC of the two fits.

- Don't rely entirely on the results of the likelihood comparison. Also look at the tests for normality and lognormality.

- Don't forget to look at graphs of the data distribution. Use the frequency distribution analysis to plot a frequency distribution histogram. Always look at the data before looking at statistical results. Also, the normality test analysis can create two QQ plots, one assuming a normal distribution and the other assuming a lognormal distribution. A QQ plot made with the appropriate assumption should be nearly linear.

5.6.5   QQ plot

When you run a normality test on column data or on residuals, Prism (new with Prism 8) can plot a QQ plot. There are multiple ways to label the axes of such graphs. Prism plots the actual Y values on the horizontal axis, and the predicted Y values (assuming sampling from a Gaussian distribution) on the Y axis. If the data were sampled from a Gaussian (normal) distribution, you expect the points to follow a straight line that matches the line of identity (which Prism shows).

Example QQ plot:
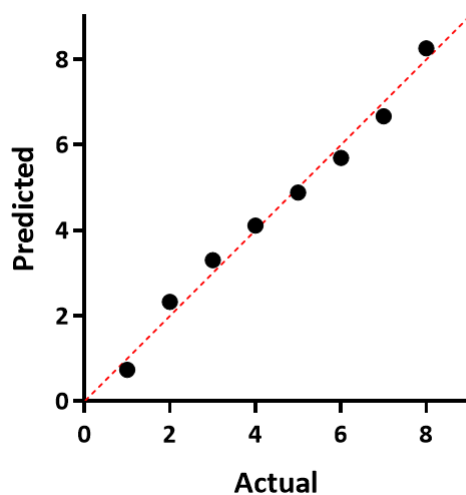
5.6.6   Interpreting results: Normality tests

What question does the normality test answer?

The normality tests all report a P value. To understand any P value, you need to know the null hypothesis. In this case, the null hypothesis is that all the values were sampled from a population that follows a Gaussian distribution.

The P value answers the question:

If that null hypothesis were true, what is the chance that a random sample of data would deviate from the Gaussian ideal as much as these data do?

Prism also uses the traditional 0.05 cut-off to answer the question whether the data passed the normality test. If the P value is greater than

0.05, the answer is Yes. If the P value is less than or equal to 0.05, the answer is No.

What should I conclude if the P value from the normality test is high?

All you can say is that the data are not inconsistent with a Gaussian distribution. A normality test cannot prove the data were sampled from a Gaussian distribution. All the normality test can do is demonstrate that the deviation from the Gaussian ideal is not more than you'd expect to see with chance alone. With large data sets, this is reassuring. With smaller data sets, the normality tests don't have much power to detect modest deviations from the Gaussian ideal.

What should I conclude if the P value from the normality test is low?

The null hypothesis is that the data are sampled from a Gaussian distribution. If the P value is small enough, you reject that null hypothesis and so accept the alternative hypothesis that the data are not sampled from a Gaussian population. The distribution could be close to Gaussian (with large data sets) or very far form it. The normality test tells you nothing about the alternative distributions.

If you P value is small enough to declare the deviations from the Gaussian idea to be "statistically significant", you then have four choices:

- The data may come from another identifiable distribution. If so, you may be able to transform your values to create a Gaussian distribution. For example, if the data come from a lognormal distribution, transform all values to their logarithms.

- The presence of one or a few outliers might be causing the normality test to fail. Run an outlier test. Consider excluding the outlier(s).

- If the departure from normality is small, you may choose to do nothing. Statistical tests tend to be quite robust to mild violations of the Gaussian assumption.

- Switch to nonparametric tests that don't assume a Gaussian distribution. But the decision to use (or not use) nonparametric tests is a big decision. It should not be based on a single normality test and should not be automated[148].

5.6.7   Q&A: Normality tests

Expand all answers	Collapse all answers

⊟ **Why the term "normality"?**

Because Gaussian distributions are also called Normal distributions.

⊟ **Which normality test is best?**

Prism offers four normality tests (offered as part of the Column Statistics analysis):

 We recommend using the D'Agostino-Pearson omnibus test. The Shapiro-Wilk test also works very well if every value is unique, but does not work well when there are ties. The basis of the test is hard for nonmathematicians to understand. For these reasons, we prefer the D'Agostino-Pearson test, even though the Shapiro-Wilk test works well in most cases.

The Kolmogorov-Smirnov test, with the Dallal-Wilkinson-Lilliefor corrected P value, is included for compatibility with older versions of Prism, but is not recommended.

⊟ **Why do the different normality tests give different results?**

All three tests ask how far a distribution deviates from the Gaussian ideal. Since the tests quantify deviations from Gaussian using different methods, it isn't surprising they give different results. The fundamental problem is that these tests do not ask which of two defined distributions (say, Gaussian vs. exponential) better fit the data. Instead, they compare Gaussian vs. not Gaussian. That is a pretty vague comparison. Since the different tests approach the problem differently, they give different results.

⊟ **How many values are needed to compute a normality test?**

The Kolmogorov-Smirnov test requires 5 or more values. The Shapiro-Wilk test requires 3 or more values. The D'Agostino test requires 8 or

more values, as does the Anderson-Darling test.

What question does the normality test answer?

The normality tests all report a P value. To understand any P value, you need to know the null hypothesis. In this case, the null hypothesis is that all the values were sampled from a Gaussian distribution. The P value answers the question:

If that null hypothesis were true, what is the chance that a random sample of data would deviate from the Gaussian ideal as much as these data do?

What cut-off does Prism use when deciding whether or not a data set passed a normality test?

You set the threshold in the analysis dialog. The default is to use the traditional 0.05 cut-off. If P<0.05, the data do not pass the normality test. If P> 0.05, the data do pass the normality test. This cut-off, of course, is totally arbitrary.

So it tells me whether a data set is Gaussian?

No. A population has a distribution that may be Gaussian or not. A sample of data cannot be Gaussian or not Gaussian. That term can only apply to the entire population of values from which the data were sampled.

Are any data sets truly sampled from ideal Gaussian distributions?

Probably not. In almost all cases, we can be sure that the data were not sampled from an ideal Gaussian distribution. That is because an ideal Gaussian distribution includes some very low negative numbers and some super high positive values.Those values will comprise a tiny fraction of all the values in the Gaussian population, but they are part of the distribution. When collecting data, there are constraints on the possible values. Pressures, concentrations, weights, enzyme activities, and many other variables cannot have negative values, so cannot be sampled from perfect Gaussian distributions. Other variables can be negative, but have

physical or physiological limits that don't allow super large values (or have extremely low negative values).

⊟ But don't t tests, ANOVA, and regression assume Gaussian distributions?

Yes, but plenty of simulations have shown that these tests work well even when the population is only approximately Gaussian.

⊟ So do the normality tests figure out whether the data are close enough to Gaussian to use one of those tests?

Not really. It is hard to define what "close enough" means, and the normality tests were not designed with this in mind.

⊟ Isn't the whole point of a normality test to decide when to use nonparametric tests?

No. Deciding whether to use a parametric or nonparametric test is a hard decision that <u>should not be automated based on a normality test</u>[148].

⊟ How should I interpet the K2, KS or W values reported by the normality test?

Each normality test reports an intermediate value that it uses to compute the P value. Unfortunately, there is no obvious way to interpret K2 (computed by the D'Agostino test), KS (computed by the Kolmogorov-Smirnov test), or W (computed by Shapiro-Wilk test). As far as I know, there is no straightforward way to use these values to decide if the deviation from normality is severe enough to switch away from parametric tests. Prism only reports these values so you can compare results with texts and other programs.

⊟ How useful are normality tests?

Not very useful, in most situations. With small samples, the normality tests don't have much power to detect nongaussian distributions. With large samples, it doesn't matter so much if data are nongaussian, since the t tests and ANOVA are fairly robust to violations of this standard.

What you would want is a test that tells you whether the deviations from the Gaussian ideal are severe enough to invalidate statistical methods that assume a Gaussian distribution. But normality tests don't do this.

References

1 RB D'Agostino, "Tests for Normal Distribution" in Goodness-Of-Fit Techniques edited by RB D'Agostino and MA Stepenes, Macel Decker, 1986.

Parts of this page are excerpted from Chapter 24 of Motulsky, H.J. (2010). Intuitive Biostatistics, 2nd edition. Oxford University Press. ISBN=978-0-19-973006-3.

5.7   Identifying outliers

Prism can identify outliers in each column using either the Grubbs' or ROUT method. Outlier detection can be a useful way to screen data for problems, but it can also be misused.

5.7.1   How to: Identify outliers

Identifying outliers in a stack of data is simple. Click Analyze from a Column data table, and then choose *Identify outliers* from the list of analyses for Column data. Prism can perform outlier tests with as few as three values in a data set.
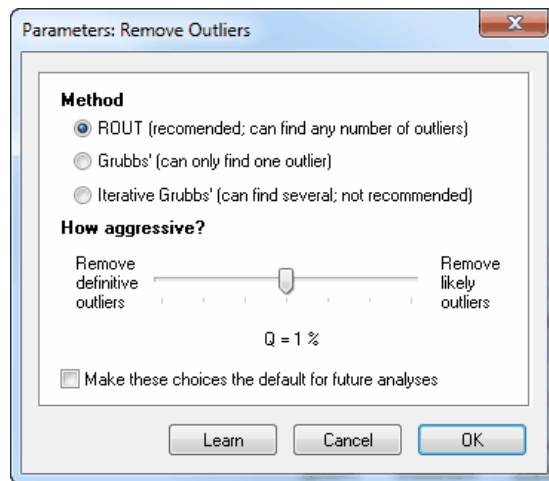
Note: This page explains how to identify an outlier from a stack of values in a data table formatted for Column data. Prism can also identify outliers during nonlinear regression.

Which method?

Prism offers three methods for identifying outliers:

ROUT

We developed the ROUT method[161] to detect outliers while fitting a curve with nonlinear regression. Prism adapts this method to detecting outliers from a stack of values in a column data table. The ROUT method can identify one or more outliers.

Grubbs' method

Grubbs' test[159] is probably the most popular method to identify an outlier. This method is also called the ESD method (Extreme Studentized Deviate). It can only identify one outlier in each data set. Prism uses the two-sided Grubbs' test, which means it will detect a value much larger than the rest, or a value much smaller than the rest.

Iterative Grubbs'

While it was designed to detect one outlier, Grubbs' method is often extended to detect multiple outliers. This is done using a simple method. If an outlier is found, it is removed and the remaining values are tested with Grubbs' test again. If that second test finds an outlier, then that value is removed, and the test is run a third time ...

While Grubb's test does a good job of finding one outlier in a data set, it does not work so well with multiple outliers. The presence of a second outlier in a small data set can prevent the first one from being detected. This is called *masking*. Grubbs' method identifies an outlier by calculating the difference between the value and the mean, and then dividing that difference by the standard deviation of all the values. When that ratio is too large, the value is defined to be an outlier. The problem is that the standard deviation is computed from all the values, including the outliers. With two outliers, the standard deviation can become large, which reduces that ratio to a value below the critical value used to define outliers. See an example of masking[163].

Recommendation

- If you somehow knew for sure that the data set had either no outliers or one outlier, then choose Grubbs' test.

- If you want to allow for the possibility of more than one outlier, choose the ROUT method. Compare the Grubbs' and ROUT methods.

- Avoid the iterative Grubbs' method.

- When you create a box-and-whiskers plot with Prism, you can choose to show Tukey whiskers, which shows points individually when their distance from the median exceeds 1.5 times the interquartile range (difference between the 75th and 25th percentiles). Some people define these points to be outliers We did not implement this method of outlier detection in Prism (beyond creating box-and-whiskers plots) because it seems to not be widely used, and has no real theoretical basis. Let us know if you'd like us to include this method of detecting outliers.

How aggressive?

There is no way to cleanly separate outliers from values sampled from a Gaussian distribution. There is always a chance that some true outliers will be missed, and that some "good points" will be falsely identified as outliers. You need to decide how aggressively to define outliers. The choice is a bit different depending on which method of outlier detection you choose.

Grubbs's test. Choose alpha.

With the Grubbs' test, you specify alpha. This has an interpretation familiar from any tests of statistical significance. If there are no outliers, alpha is the chance of mistakenly identifying an outlier.

Note that alpha applies to the entire experiment, not to each value. Assume that you set alpha to 5% and test a data set with 1000 values, all sampled from a Gaussian distribution. There is a 5% chance that the most extreme value will be identified as an outlier. That 5% applies to the entire data set, no matter how many values it has. It would be a mistake to multiply 5% by the sample size of 1000, and conclude that you'd expect 50 outliers to be identified.

Alpha is two-tailed, because the Grubbs test in Prism identifies outliers that are either "too large" or "too small".

Rout method. Choose Q.

The ROUT method is based on the False Discovery Rate (FDR), so you specify Q, which is the maximum desired FDR.

When there are no outliers (and the distribution is Gaussian), Q can be interpreted just like alpha. When all the data are sampled from a Gaussian distribution (so no outliers are present), Q is the chance of identifying one or more outliers.

When there are outliers in the data, Q is the desired maximum false discovery rate. If you set Q to 1%, then you are aiming for no more than 1% of the identified outliers to be false (are in fact just the tail of a Gaussian distribution) and thus for at least 99% identified outliers to actually be outliers (from a different distribution). If you set Q to 5%, then you are expecting no more than 5% of the identified outliers to be false and for at least 95% of the identified outliers to be real.

Recommendation

The trade-off is clear. If you set alpha or Q too high, then many of the identified "outliers" will be actually be data points sampled from the same Gaussian distribution as the others. If you set alpha or Q too low, then you won't identify all the outliers.

There are no standards for outlier identification. We suggest that you start by setting Q to 1% or alpha to 0.01.

How Prism presents the results

The results are presented on three pages:

- Cleaned data (outliers removed). You could use this page as the input to another analysis, such as a t test or one-way ANOVA.

- Outliers only.

- Summary. This page lists the number of outliers detected in each data set.

5.7.2   Analysis checklist: Outliers

If the outlier test identifies one or more values as being an outlier, ask yourself these questions:

✔ **Was the outlier value entered into the computer incorrectly?**

If the "outlier" is in fact a typo, fix it. It is always worth going back to the original data source, and checking that outlier value entered into Prism is actually the value you obtained from the experiment. If the value was the result of calculations, check for math errors.

✔ **Is the outlier value scientifically impossible?**

Of course you should remove outliers from your data when the value is completely impossible. Examples include a negative weight, or an age (of a person) that exceed 150 years. Those are clearly errors, and leaving erroneous values in the analysis would lead to nonsense results.

✔ **Is the assumption of a Gaussian distribution dubious?**

Both the Grubbs' and ROUT tests assume that all the values are sampled from a Gaussian distribution, with the possible exception of one (or a few) outliers from a different distribution. If the underlying distribution is not Gaussian, then the results of the outlier test is unreliable. It is especially important to beware of lognormal distributions[157]. If the data are sampled from a lognormal distribution, you expect to find some very high values

which can easily be mistaken for outliers. Removing these values would be a mistake.

✓ Is the outlier value potentially scientifically interesting?

If each value is from a different animal or person, identifying an outlier might be important. Just because a value is not from the same Gaussian distribution as the rest doesn't mean it should be ignored. You may have discovered a polymorphism in a gene. Or maybe a new clinical syndrome. Don't throw out the data as an outlier until first thinking about whether the finding is potentially scientifically interesting.

✓ Does your lab notebook indicate any sort of experimental problem with that value

It is easier to justify removing a value from the data set when it is not only tagged as an "outlier" by an outlier test, but you also recorded problems with that value when the experiment was performed.

✓ Do you have a policy on when to remove outliers?

Ideally, removing an outlier should not be an *ad hoc* decision. You should follow a policy, and apply that policy consistently.

✓ If you are looking for two or more outliers, could *masking* be a problem?

Masking[163] is the name given to the problem where the presence of two (or more) outliers, can make it harder to find even a single outlier.

If you answered no to all those questions...

If you've answered no to all the questions above, there are two possibilities:

- The suspect value came from the same Gaussian population as the other values. You just happened to collect a value from one of the tails of that distribution.

- The suspect value came from a different distribution than the rest. Perhaps it was due to a mistake, such as bad pipetting, voltage spike, holes in filters, etc.

If you knew the first possibility was the case, you would keep the value in your analyses. Removing it would be a mistake.

If you knew the second possibility was the case, you would remove it, since including an erroneous value in your analyses will give invalid results.

The problem, of course, is that you can never know for sure which of these possibilities is correct. An outlier test cannot answer that question for sure. Ideally, you should create a lab policy for how to deal with such data, and follow it consistently.

If you don't have a lab policy on removing outliers, here is suggestion: Analyze your data both with and without the suspected outlier. If the results are similar either way, you've got a clear conclusion. If the results are very different, then you are stuck. Without a consistent policy on when you remove outliers, you are likely to only remove them when it helps push the data towards the results you want.

5.8   One sample t test and Wilcoxon signed rank test

You've measured a variable in one group, and the means (or median) is not the same as expected by theory (or by the null hypothesis). Is that due to chance? Or does it tell you the mean (or median) of the values is really different from the hypothetical value?

5.8.1   How to: One-sample t test and Wilcoxon signed rank test

In prior versions of Prism, the one-sample t test and the Wilcoxon rank sum tests were computed as part of Prism's Column Statistics analysis. Starting with Prism 8, there is a separate analysis in Prism for these tests.

1. Create a Column data table.

2. Enter each data set in a single Y column. So all values from each group are stacked into a column. Prism will perform a one-sample t test (or Wilcoxon rank sum test) on each column you enter.

3. Click Analyze, look in the list of Column analyses, and choose one-sample t and Wilcoxon test.

5.8.2 Experimental design tab

Choose test

Choose one of two tests:

- A one-sample t test compares the mean of a single column of numbers against a hypothetical mean that you provide. It assumes the values were sampled from a Gaussian distribution.

- The [nonparametric][148] Wilcoxon signed rank test compares the median of a single column of numbers against a hypothetical median. Don't confuse it with the [Wilcoxon matched pairs test][321] which compares two paired or matched groups.

Hypothetical value

Enter the hypothetical value to which you wish to compare the mean (t test) or median (Wilcoxon test). This value is often 0, or 100 (when values are percentages), or 1.0 (when values are ratios).

With the Wilcoxon test, you also need to specify what to do when a data value exactly equals the hypothetical values. [Understand the two choices][257].

5.8.3 Options tab

Subcolumns

Usually, you'll run this analysis on values placed on a Column table, and these choices won't be available. But if you enter data on a grouped table with subcolumns, Prism gives you three choices: To analyze each subcolumn separately, to average the values in the subcolumns on each row and then analyze that stack of means, or to treat all the values as one sample of data (paying no attention to which row any values are on).

Calculation

Enter a value for alpha, the threshold P value to define statistical signficance.

Output

Choose P value format and number of digits to use when displaying results.

5.8.4   Interpreting results: One-sample t test

A one-sample t test compares the mean of a single column of numbers against a hypothetical mean that you provide.

The P value answers this question:

If the data were sampled from a Gaussian population with a mean equal to the hypothetical value you entered, what is the chance of randomly selecting N data points and finding a mean as far (or further) from the hypothetical value as observed here?

If the P value is large[81], the data do not give you any reason to conclude that the population mean differs from the hypothetical value you entered. This is not the same as saying that the true mean equals the hypothetical value. You just don't have evidence of a difference.

If the P value is small[79] (usually defined to mean less than 0.05), then it is unlikely that the discrepancy you observed between sample mean and hypothetical mean is due to a coincidence arising from random sampling. You can reject the idea that the difference is a coincidence, and conclude instead that the population has a mean different than the hypothetical value you entered. The difference is statistically significant. But is the difference scientifically important? The confidence interval helps you decide[79].

Prism also reports the 95% confidence interval for the difference between the actual and hypothetical mean. You can be 95% sure that this range includes the true difference.

Assumptions

The one sample t test assumes that you have sampled your data from a population that follows a Gaussian distribution. While this assumption is not too important with large samples, it is important with small sample sizes, especially when N is less than 10. If your data do not come from a Gaussian distribution, you have three options. Your best option is to transform the values to make the distribution more Gaussian, perhaps by transforming all values to their reciprocals or logarithms. Another choice is to use the Wilcoxon signed rank nonparametric test instead of the t test. A final option is to use the t test anyway, knowing that the t test is fairly robust to departures from a Gaussian distribution with large samples.

The one sample t test also assumes that the "errors" are independent[32] . The term "error" refers to the difference between each value and the group mean. The results of a t test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption.

How the one-sample t test works

Prism calculates the t ratio by dividing the difference between the actual and hypothetical means by the standard error of the mean.

A P value is computed from the t ratio and the numbers of degrees of freedom (which equals sample size minus 1).

5.8.5 Interpreting results: Wilcoxon signed rank test

The nonparametric[148] Wilcoxon signed rank test compares the median of a single column of numbers against a hypothetical median. Don't confuse it with the Wilcoxon matched pairs test[321] which compares two paired or matched groups.

Interpreting the confidence interval

The signed rank test compares the median of the values you entered with a hypothetical population median you entered. Prism reports the difference between these two values, and the confidence interval of the difference. Prism subtracts the median of the data from the hypothetical

median, so when the hypothetical median is higher, the result will be positive. When the hypothetical median is lower, the result will be negative

Since the nonparametric test works with ranks, it is usually not possible to get a confidence interval with exactly 95% confidence. Prism finds a close confidence level, and reports what it is. So you might get a 96.2% confidence interval when you asked for a 95% interval.

Interpreting the P value

The P value answers this question:

> If the data were sampled from a population with a median equal to the hypothetical value you entered, what is the chance of randomly selecting N data points and finding a median as far (or further) from the hypothetical value as observed here?

If the P value is small[79], you can reject the idea that the difference is a due to chance and conclude instead that the population has a median distinct from the hypothetical value you entered.

If the P value is large[81], the data do not give you any reason to conclude that the population median differs from the hypothetical median. This is not the same as saying that the medians are the same. You just have no compelling evidence that they differ. If you have small samples, the Wilcoxon test has little power. In fact, if you have five or fewer values, the Wilcoxon test will always give a P value greater than 0.05, no matter how far the sample median is from the hypothetical median.

Assumptions

The Wilcoxon signed rank test does not assume that the data are sampled from a Gaussian distribution. However it does assume that the data are distributed symmetrically around the median. If the distribution is asymmetrical, the P value will not tell you much about whether the median is different than the hypothetical value.

Like all statistical tests, the Wilcoxon signed rank test assumes that the errors are independent[32]. The term "error" refers to the difference between each value and the group median. The results of a Wilcoxon test only make sense when the scatter is random – that any factor that causes a value to be too high or too low affects only that one value.

How the Wilcoxon signed rank test works

1. Calculate how far each value is from the hypothetical median.

2. Ignore values that exactly equal the hypothetical value. Call the number of remaining values N.

3. Rank these distances, paying no attention to whether the values are higher or lower than the hypothetical value.

4. For each value that is lower than the hypothetical value, multiply the rank by negative 1.

5. Sum the positive ranks. Prism reports this value.

6. Sum the negative ranks. Prism also reports this value.

7. Add the two sums together. This is the sum of signed ranks, which Prism reports as W.

If the data really were sampled from a population with the hypothetical median, you would expect W to be near zero. If W (the sum of signed ranks) is far from zero, the P value will be small.

With fewer than 200 values, Prism computes an exact P value, using a method explained in Klotz(2). With 200 or more values, Prism uses a standard approximation that is quite accurate.

Prism calculates the confidence interval for the discrepancy between the observed median and the hypothetical median you entered using the method explained on page 234-235 of Sheskin (1) and 302-303 of Klotz (2).

How Prism deals with values that exactly equal the hypothetical median

What happens if a value is identical to the hypothetical median?

When Wilcoxon developed this test, he recommended that those data simply be ignored. Imagine there are ten values. Nine of the values are distinct from the hypothetical median you entered, but the tenth is identical to that hypothetical median (to the precision recorded). Using Wilcoxon's original method, that tenth value would be ignored and the

other nine values would be analyzed.This is how InStat and previous versions of Prism (up to version 5) handle the situation.

Pratt(3,4) proposed a different method that accounts for the tied values. Prism 6 and later offer the choice of using this method.

Which method should you choose? Obviously, if no value equals the hypothetical median, it doesn't matter. Nor does it matter much if there is, for example, one such value out of 200.

It makes intuitive sense that data should not be ignored, and so Pratt's method must be better. However, Conover (5) has shown that the relative merits of the two methods depend on the underlying distribution of the data, which you don't know.

Why results in Prism 6 and later can be different than from previous versions of Prism

Results from Prism 6 can differ from prior versions because starting with version 6, Prism does exact calculations in two situations where Prism 5 did approximate calculations. All versions of Prism report whether it uses an approximate or exact methods.

- Prism 6 and later can perform the exact calculations much faster than did Prism 5, so does exact calculations with some sample sizes that earlier versions of Prism could only do approximate calculations.

- If two values are the same, prior versions of Prism always used the approximate method. Prism 6 and later use the exact method unless the sample is huge.

Another reason for different results between Prism 6 and prior versions is if a value exactly matches the hypothetical value you are comparing against. Prism 6 introduced a new option (method of Pratt) which will give different results than prior versions did. See the previous section.

References

1. D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, fourth edition.

2. JH Klotz, A computational approach to statistics, 2006, self-published book, Chapter 15.2 The Wilcoxon Signed Rank Test.

3. Pratt JW (1959) <u>Remarks on zeros and ties in the Wilcoxon signed rank procedures.</u> Journal of the American Statistical Association, Vol. 54, No. 287 (Sep., 1959), pp. 655-667

4. Pratt, J.W. and Gibbons, J.D. (1981), Concepts of Nonparametric Theory, New York: Springer Verlag.

5. WJ Conover, <u>On Methods of Handling Ties in the Wilcoxon Signed-Rank Test</u>, Journal of the American Statistical Association, Vol. 68, No. 344 (Dec., 1973), pp. 985-988

5.9   t tests, Mann-Whitney and Wilcoxon matched pairs test

You've measured a variable in two groups, and the means (and medians) are distinct. Is that due to chance? Or does it tell you the two groups are really different?

5.9.1   Running a t test (or related nonparametric test)

5.9.1.1   Entering data for a t test

Setting up the data table

From the Welcome (or New Table and graph) dialog, choose the Column tab.

If you aren't ready to enter your own data, choose one of the sample data sets.

If you want to enter data, note that there are two choices. You can enter raw data or summary data (as mean, SD or SEM, and n).

Entering raw data

[Enter the data for each group into a separate column](#)[272]. The two groups do not have be the same size (it's OK to leave some cells empty). If the data are unpaired, it won't make sense to enter any row titles.

If the data are matched, so each row represents a different subject of experiment, then [you may wish to use row titles](#)[291] to identify each row.

Enter mean and SD or SEM

Prism can compute an unpaired t test (but not a paired t test, and not nonparametric comparisons) [with data entered as mean, SD (or SEM), and n](#)[274]. This can be useful if you are entering data from another program or publication.

From the Column tab of the Welcome dialog, choose that you wish to enter and plot error values computed elsewhere. Then choose to enter mean, n, and either SD, SEM or %CV (coefficient of variation). Entering sample size (n) is essential. It is not possible to compute a t test if you only enter the mean and SD or SEM without n.

Even though you made your choice on the Column tab of the Welcome dialog, Prism will show you a Grouped data table. Enter your data on the first row of this table.

Q&A: Entering data for t tests and related tests

Is it possible to define the two groups with a grouping variable?

Some programs expect (or allow) you to enter all the data into one column, and enter a grouping variable into a second column to define which rows belong to which treatment group. Prism does not use this way to organize data. Instead, the two groups must be defined by two columns. Enter data for one group into column A and the other group into column B.

Can I enter data in lots of columns and then choose two to compare with a t test?

Yes. After you click Analyze, you'll see a list of all data sets on the right side of the dialog. Select the two you wish to compare.

Can I enter data as mean, SD (or SEM) and N?

Yes. Follow [this example][274] to see how. With data entered this way, you can only choose an unpaired t test. It is impossible to run a paired t test or a nonparametric test from data entered as mean, SD (or SEM) and N.

Can I enter data for many t tests on one table, and ask Prism to run them all at once?

[Yes!][330]

5.9.1.2 Experimental design tab: t tests

Prism offers seven related tests that compare two groups. To choose among these tests, answer three questions in the *Experimental Design* tab of the t test parameters dialog:

Experimental design: unpaired or paired

Choose a paired test when the columns of data are matched. That means that values on the same row are related to each other.

Here are some examples:

- You measure a variable in each subject before and after an intervention.

- You recruit subjects as pairs, matched for variables such as age, ethnic group, and disease severity. One of the pair gets one treatment; the other gets an alternative treatment.

- You run a laboratory experiment several times, each time with a control and treated preparation handled in parallel.

- You measure a variable in twins or child/parent pairs.

Matching should be determined by the experimental design, and definitely should not be based on the variable you are comparing. If you are comparing blood pressures in two groups, it is OK to match based on age or postal code, but it is not OK to match based on blood pressure.

Assume Gaussian distribution?

[Nonparametric tests](#)[148], unlike t tests, are not based on the assumption that the data are sampled from a [Gaussian distribution](#)[37]. But nonparametric tests have [less power](#)[149]. Deciding when to use a nonparametric test is [not straightforward](#)[152].

Choose test

After defining the experimental design, and the general approach (parametric or nonparametric), you need to decide exactly what test you want Prism to perform.

Parametric, not paired

Decide whether to accept the assumption that the two samples come from populations with the same standard deviations (same variances). This is a standard assumption of the unpaired t test. If don't wish to make this assumption, Prism will perform the [unequal variance (Welch) unpaired t test](#)[282].

Parametric, paired

Choose the paired t test (which is standard in this situation) or the [ratio t test](#)[299] (which is less standard). Choose the paired t test when you expect the *differences* between paired values to be a consistent measure of treatment effect. Choose the ratio paired t test when you expect the *ratio* of paired values to be a consistent measure of treatment effect.

Nonparametric, not paired

Prism offers two choices: The Mann-Whitney test (which Prism has always offered) and the [Kolmogorov-Smirnov test](#)[317] (which is new). It is hard to offer guidelines for choosing one test vs. the other except to follow the tradition of your lab or field. The main difference is that the Mann-Whitney test has more power to detect a difference in the median, but the Kolmogorov-Smirnov test has more power to detect differences in the shapes of the distributions.

| | Mann-Whitney test | Kolmogorov-Smirnov test |

Power to detect a shift in the median	More power	Less power
Power to detect differences in the shape of the distributions	Less power	More power

Nonparametric, paired

In this case there is no choice. Prism will perform the Wilcoxon matched pairs test.

5.9.1.3   Options tab: t tests

The second tab of the parameters dialog for t tests and nonparmaetric tests is labeled Options. The choices on this tab vary a bit depending on which test you chose on the first tab.

Calculations

The default choices for the calculation options will be fine for most people (two-tailed P values, 95% confidence intervals, and difference computed as the first column minus the second).

- [One- or two-tailed P value](#)[75]. Choose a two-tailed P value, unless you have a strong reason not to.

- Report differences as. This determines the sign of the difference between means or medians that Prism reports. Do you want to subtract the second mean from the first, or the first from the second?

- Confidence level. 95% is standard, but you can pick other degrees of confidence.

Graphing options

None of the options are selected by default. They can be useful to view the data with more depth, but none are essential to beginners.

- Graph differences (paired). The paired t test and Wilcoxon matched pairs test first compute the difference between the two values on each

row. This option creates a table and graph showing this list of differences

- Graph ranks (nonparametric). The Mann-Whitney test first ranks all the values from low to high, and then compares the mean rank of the two groups. This option creates a table and graph showing those ranks. The Wilcoxon first computes the difference between each pair, and then ranks the absolute value of those differences, assigning negative values when the difference is negative.

- Graph correlation (paired). Graph one variable vs. the other to visually assess how correlated they are.Graph residuals. This option is only offered for unpaired data. To create the new residuals table, Prism computes the difference between each value and the mean (or median) of that column. Inspecting a graph of residuals can help you assess the assumption that all the data are sampled from populations with the same SD.

- Graph confidence interval of differences between means

Additional results

These four choices are not selected by default. The second choice (AIC) is for special purposes. The other three might be useful even to beginners.

- Descriptive statistics. Check this option, and Prism will create a new table of descriptive statistics for each data set.

- Also compare models using AICc. Most people will not want to use this, as it is not standard. The unpaired t test essentially compares the fit of two models to the data (one shared mean, vs. two separate group means). The t test calculations are equivalent to the extra sum-of-squares F test. When you check this option, Prism will report the usual t test results, but will also compare the fit of the two models by AICc, and report the percentage chance that each model is correct.

- Nonparametric tests. Compute the 95% CI for the difference between medians (Mann-Whitney) or the median of the paired differences (Wilcoxon). You can only trust this confidence interval if you make an additional assumption not required to interpret the P value. For the Mann-Whitney test, you must assume that the two populations have

the same shape (whatever it is). For the Wilcoxon test, you must assume that the distribution of differences is symmetrical. Statistical analyses are certainly more useful when reported with confidence intervals, so it is worth thinking about whether you are willing to accept those assumptions. <u>Calculation details.</u>

- Wilcoxon test. What happens when the two matching values in a row are identical? Prism 5 handled this as Wilcoxon said to when he created the test. Prism offers the option of <u>using the Pratt method instead</u>[329]. If your data has lots of ties, it is worth reading about the two methods and deciding which to use.

5.9.1.4   Residuals tab: t tests

Why residuals?

Prism 8 introduced the ability to plot residual plots with t tests, provided that you entered raw data and not averaged data as mean, n and SD or SEM.

Many scientists thing of residual as values that are obtained with regression. But the t test is really regression in disguise. It fits a model. One of the assumptions of t tests is that the residuals from that model are sampled from a Gaussian distribution. A residual plot helps you assess this assumption.

Which graph to create?

Prism can make three kinds of residual plots. For t tests, since there are only two groups, the first two choices are not super useful, and the QQ plot is the most useful way to plot residuals.

- Residual plot. The X axis is the actual value of the value (unpaired tests) or difference (paired test). The Y axis is the residual. This lets you spot residuals that are much larger or smaller than the rest.

- Homoscedasticity plot. The X axis is the actual value of the value (unpaired tests) or difference (paired test). The Y axis is the absolute value of the residual.This lets you check whether larger values are associated with bigger residuals (larger absolute value).

- QQ plot. The X axis is the actual residual. The Y axis is the predicted residual, computed from the percentile of the residual (among all residuals) and assuming sampling from a Gaussian distribution. ANOVA

assumes a Gaussian distribution of residuals, and this graph lets you check that assumption.

Diagnostics for residuals

• Are the residuals Gaussian? Prism runs four normality tests on the residuals. The residuals from both groups are pooled and entered into one set of normality tests.

How residuals are computed

Residuals with t tests and related tests are simple to understand.

• Unpaired t test. A residual is computed for each value. Each residual is the difference between the entered value and the mean of all values for that group. A residual is positive when the corresponding value is greater than the sample mean, and is negative when the value is less than the sample mean.

• Mann-Whitney test. A residual is computed for each value. Each residual is the difference between an entered value and the median of all values for that group. A residual is positive when the corresponding value is greater than the sample median, and is negative when the value is less than the sample median.

• Paired t test. A residual is calculated for each pair. First a difference is computed for each pair. The residual equals the actual difference between the pairs minus the mean of all such difference in the data sets.

• Ratio t test. A residual is calculated for each pair. A ratio is computed for each pair. The residual equals the logarithm of that ratio minus the mean of the logarithms of all the ratios.

• Wilcoxon matched pairs test. A residual is calculated for each pair. First a difference is computed for each pair. The residual equals the actual difference between the pairs minus the median of all such difference in the data sets.

5.9.1.5   Q&A: Choosing a test to compare two groups

If I have data from three or more groups, is it OK to compare two groups at a time with a t test?

No. You should analyze all the groups at once with <u>one-way ANOVA</u>[425], and then follow up with <u>multiple comparison tests</u>[117]. The only exception is when some of the 'groups' are really controls to prove the assay worked, and are not really part of the experimental question you are asking.

I know the mean, SD (or SEM) and sample size for each group. Which tests can I run?

You can <u>enter data</u>[272] as mean, SD (or SEM) and N, and Prism can compute an unpaired t test or the Welch t test. Prism cannot perform an paired test, as that requires analyzing each pair. It also cannot do any nonparametric tests, as these require ranking the data.

I only know the two group means, and don't have the raw data and don't know their SD or SEM. Can I run a t test?

No. The t test compares the difference between two means and compares that difference to the standard error of the difference, computed from the standard deviations and sample size. If you only know the two means, there is no possible way to do any statistical comparison.

Can I use a normality test to make the choice of when to use a nonparametric test?

It is <u>not a good idea</u>[148] to base your decision solely on the normality test. Choosing when to use a nonparametric test is not a straightforward decision, and you can't really automate the process.

I want to compare two groups. The outcome has two possibilities, and I know the fraction of each possible outcome in each group. How can I compare the groups?

Not with a t test. Enter your data into a <u>contingency table</u>[545] and analyze with <u>Fisher's</u>[551] exact test.

I want to compare the mean survival time in two groups. But some subjects are still alive so I don't know how long they will live. How can I do a t test on survival times?

You should use special methods designed to [compare survival curves]572. Don't run a t test on survival times.

I don't know whether it is ok to assume equal variances. Can't a statistical test tell me whether or not to use the Welch t test?

While that sounds like a good idea, [in fact it is not]. The decision really should be made as part of the experimental design and not based on inspecting the data.

I don't know whether it is better to use the regular paired t test or the ratio test. Is it ok to run both, and report the results with the smallest P value?

No. The results of any statistical test can only be interpreted at face value when the choice of analysis method was part of the experimental design.

Should I use the Welch test routinely because it is always possible the two populations have different standard deviations.

Ruxton (1) and Delacre (2) make a strong case that this is a good idea.

References

1. Ruxton. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology (2006) vol. 17 (4) pp. 688

2. Delacre, M., Lakens, D.L., and Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. Rips 30: 92–10.

5.9.1.6   The advantage of pairing

This set of graphs shows the importance of designing experiments where pairing or matching is part of the experimental design, and of accounting for that pairing when analyzing the data.

Left	Right
87	43
23	14
32	44
54	52
31	21
29	29
65	31
75	42
89	54
95	63

Left	Right
87	21
23	63
32	44
54	52
31	43
29	42
65	31
75	29
89	54
95	14
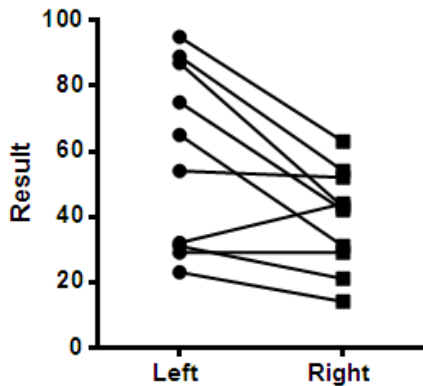
Paired t test:
P = 0.0126
Mean difference = 18.7
95% CI: 5.081 to 32.32
SD of differences = 19.04
SEM of differences = 6.02
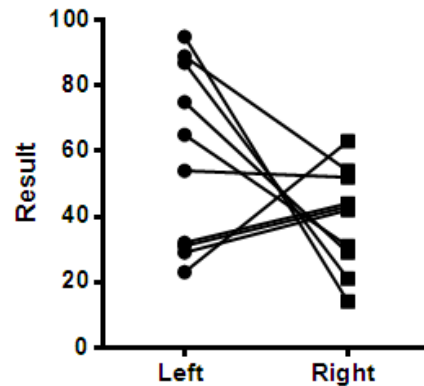
Paired t test:
P = 0.1677
Mean difference = 18.7
95% CI: -9.492 to 46.89
SD of differences = 39.41
SEM of differences = 12.46

These data compare a result in the left and right eye of the same person. The two data tables show two different sets of results, and the figure below show the data and results.

The data for the left eye is the same in both cases. The data for the right eye differs. Actually, the values are the same values, but the order is different. Since the values are the same, an unpaired t test would look identical results for both experiments. A bar graph showing the mean and SD (or SEM) of each group would also be identical for both groups.

The before-after graph, which shows the pairing, looks very different for the two experiments, and the results of a paired t test are very different. The experiment on the left shows a consistent difference between left and

right, with a small P value. The experiment on the right leads to no clear conclusion.

This example makes these points:

- When the experiment had a paired design, it is really important to do a paired test.

- When the experiment has a paired design, it is important to use a before-after graph to show the results. A graph showing only the mean and SD (or SEM) separately for each eye would not really give you a good idea of what's going on.

- It doesn't really help to report the mean and SD (or SEM) of each treatment (left and right in the experiments shown above). These results are identical for the two experiments shown above. Instead, it makes sense to show the mean difference with its SD, SEM or confidence interval.

5.9.2   Unpaired t test

5.9.2.1   How to: Unpaired t test from raw data

1. Create data table and enter data
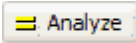
From the Welcome (or New Table and graph) dialog, choose the Column tab.

Choose to enter replicate values stacked in columns. Or, if you are not ready to enter your own data, choose sample data and choose: t test - unpaired.

Enter the data for each group into a separate column. The two groups do not have to have the same number of values, and it's OK to leave some cells empty.

Tabl...	A	B	
One...	Male	Female	
	Y	Y	
1	54	43	
2	23	34	
3	45	65	
4	54	77	
5	45	46	
6		65	
7			
8			

2. Choose the unpaired t test

1. From the data table, click 🔲 Analyze on the toolbar.

2. Choose t tests from the list of column analyses.

3. On the first (Experimental Design) tab of t test dialog, make these choices:

   - Experimental design: Unpaired

   - Assume Gaussian distribution: Yes.

   - Choose test: Unpaired t test. Choose the Welch's correction if you don't want to assume the two sets of data are sampled from populations with equal variances, and you are willing to accept the loss of power that comes with that choice. That choice is used rarely, so don't check it unless you are quite sure.

4. On the options tab, make these choices:

   - Choose a one- or two-sided P value[75]. If in doubt, choose a two-tail P value.

   - Choose the direction of the differences. This choice only affects the sign of the difference and the confidence interval of the difference, without affecting the P value.

   - Choose a confidence level. Leave this set to 95%, unless you have a good reason to change it.

3. Review the results

The t test investigates the likelihood that the difference between the means of the two groups could have been caused by chance. So the most important results are the 95% confidence interval for that difference and the P value.

Learn more about [interpreting][276] and [graphing][284] the results.

Before accepting the results, [review the analysis checklist][170].

5.9.2.2   How to: Unpaired t test from averaged data

1. Enter data

From the Welcome (or New Table and graph) dialog, choose the Column tab.

Choose to enter and plot error values computed elsewhere.

Enter the data all on one row.

Table format: Two-way	A Control			B Treated		
	Mean	SD	N	Mean	SD	N
1 Title	34.5	11.3	12	46.5	7.3	14
2 Title						

2. Choose the unpaired t test

1. From the data table, click **Analyze** on the toolbar.

2. Choose t tests from the list of column analyses.

   3. On the first (Experimental Design) tab of t test dialog, make these choices:

   • Experimental design: Unpaired

   • Assume Gaussian distribution: Yes.

   • Choose test: Unpaired t test. Choose the Welch's correction if you don't want to assume the two sets of data are sampled from populations with equal variances, and you are willing to accept the loss of power that comes with that choice. That choice is used rarely, so don't check it unless you are quite sure.

4. On the options tab, make these choices:

- Choose a <u>one- or two-sided P value</u>[75]. If in doubt, choose a two-tail P value.

- Choose the direction of the differences. This choice only affects the sign of the difference and the confidence interval of the difference, without affecting the P value.

- Choose a confidence level. Leave this set to 95%, unless you have a good reason to change it.

3. Review the results

The t test investigates the likelihood that the difference between the means of the two groups could have been caused by chance. So the most important results are the 95% confidence interval for that difference and the P value.

Learn more about <u>interpreting the results of a t test</u>[276].

Before accepting the results, <u>review the analysis checklist</u>[170].

4. Polish the graph
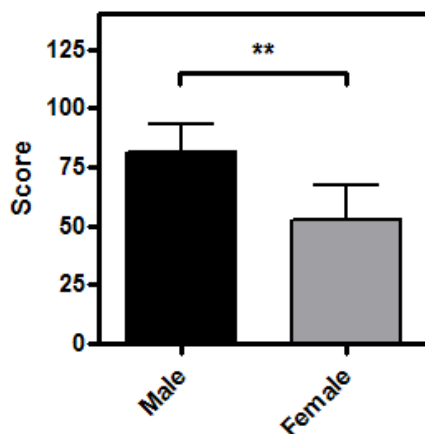
- Be sure to mention on the figure, figure legend or methods section whether the error bars represent SD or SEM ([what's the difference?](#)[49]).

- To add the [asterisks representing significance level](#)[86] copy from the results table and paste onto the graph. This creates a live link, so if you edit or replace the data, the number of asterisks may change (or change to 'ns'). Use the drawing tool to add the line below the asterisks, then right-click and set the arrow heads to "half tick down".

- To make your graph simple to understand, we strongly recommend avoiding log axes, starting the Y axis at any value other than zero, or having a discontinuous Y axis.

5.9.2.3 Interpreting results: Unpaired t

Difference between means

The unpaired t test compares the means of two groups. Note that this test is often called independent samples t test.

The most useful result is the confidence interval for the difference between the means. The point of the experiment was to see how far apart the two means are. The confidence interval tells you how precisely you know that difference. If the [assumptions of the analysis are true](#)[170], you can be 95% confident that the 95% confidence interval contains the true difference between the means.

For many purposes, this confidence interval is all you need. Note that you can change the sign of the differences in the Options tab of the t test dialog, where you can tell Prism to subtract column B from A, or A from B.

Prism also reports the difference between the two means with the standard error of that difference.

P value

The P value is used to ask whether the difference between the mean of two groups is likely to be due to chance. It answers this question:

If the two populations really had the same mean, what is the chance that random sampling would result in means as far apart (or more so) than observed in this experiment?

It is traditional, but not necessary and often not useful, to use the P value to make a simple statement about whether or not the difference is "[statistically significant][84] ".

You will interpret the results differently depending on whether the P value is [small][79] or [large][81] .

t ratio

To calculate a P value for an unpaired t test, Prism first computes a t ratio. The t ratio is the difference between sample means divided by the standard error of the difference, calculated by combining the SEMs of the two groups. If the difference is large compared to the SE of the difference, then the t ratio will be large (or a large negative number), and the P value is small. The sign of the t ratio indicates only which group had the larger mean. The P value is derived from the absolute value of t. Prism reports the t ratio so you can compare with other programs, or examples in text books. In most cases, you'll want to focus on the confidence interval and P value, and can safely ignore the value of the t ratio.

For the unpaired t test, the number of degrees of freedom (df) equals the total sample size minus 2. Welch's t test (a modification of the t test which doesn't assume equal variances) calculates df from a complicated equation.

F test for unequal variance

The unpaired t test depends on the assumption that the two samples come from populations that have identical standard deviations (and thus identical variances). Prism tests this assumption using an F test.

First compute the standard deviations of both groups, and square them both to obtain variances. The F ratio equals the larger variance divided by the smaller variance. So F is always greater than (or possibly equal to) 1.0.

The P value then asks:

  If the two populations really had identical variances, what is the chance of obtaining an F ratio this big or bigger?

Don't mix up the P value testing for equality of the variances (standard deviations) of the groups with the P value testing for equality of the

means. That latter P value is the one that answers the question you most likely were thinking about when you chose the t test.

R squared from unpaired t test

Prism, unlike most statistics programs, reports a R^2 value as part of the unpaired t test results. It quantifies the fraction of all the variation in the samples that is accounted for by a difference between the group means. If $R^2=0.36$, that means that 36% of all the variation among values is attributed to differences between the two group means, leaving 64% of the variation that comes from scatter among values within the groups.

If the two groups have the same mean, then none of the variation between values would be due to differences in group means so R^2 would equal zero. If the difference between group means is huge compared to the scatter within the group, then almost all the variation among values would be due to group differences, and the R^2 would be close to 1.0.

5.9.2.4   What to do when the groups have different standard deviations?

The t test assumes equal variances

The standard unpaired t test (but not the Welch t test) assumes that the two sets of data are sampled from populations that have identical standard deviations, and thus identical variances, even if their means are distinct.

Testing whether two groups are sampled from populations with equal variances

As part of the t test analysis, Prism tests this assumption using an F test to compare the variance of two groups. Note that a bug in very earlier versions of Prism and InStat gave a P value for the F test that was too small by a factor of two.

Don't mix up the P value testing for equality of the standard deviations of the groups with the P value testing for equality of the means. That latter P value is the one that answers the question you most likely were thinking about when you chose the t test or one-way ANOVA. The P value that tests for equality of variances answers this question:

If the populations really had identical standard deviations, what is the chance of observing as large a discrepancy among sample standard deviations as occurred in the data (or an even larger discrepancy)?

What to do if the variances differ

If the P value is small, you reject the null hypothesis that both groups were sampled from populations with identical standard deviations (and thus identical variances).

Then what? There are five possible answers.

- Conclude that the populations are different. In many experimental contexts, the finding of different standard deviations is as important as the finding of different means. If the standard deviations are different, then the populations are different regardless of what the t test concludes about differences between the means. Before treating this difference as a problem to workaround, think about what it tells you about the data. This may be the most important conclusion from the experiment! Also consider whether the group with the larger standard deviation is heterogeneous. If a treatment was applied to this group, perhaps it only worked on about half of the subjects.

- Transform your data. In many cases, transforming the data can equalize the standard deviations. If that works, you can then run the the t test on the transformed results. Logs are especially useful. (See Chapter 46 of Intuitive Biostatistics for an example). The log transform is appropriate when data are sampled from a lognormal distribution. In other situations, a reciprocal or square root transform may prove useful. Ideally, of course, the transform should have been planned as part of the experimental design.

- Ignore the result. With equal, or nearly equal, sample size (and moderately large samples), the assumption of equal standard deviations is not a crucial assumption. The t test work pretty well even with unequal standard deviations. In other words, the t test is robust to violations of that assumption so long as the sample size isn't tiny and the sample sizes aren't far apart. If you want to use ordinary t tests, run some simulations with the sample size you are actually using and the difference in variance you are expecting, to see how far off the t test results are.

- Go back and rerun the t test, checking the option to do the Welch t test that allows for unequal variance. While this sounds sensible, Moser and Stevens (1) have shown that it isn't. If you use the F test to compare variances to decide which t test to use (regular or Welch), you will have increased your risk of a Type I error. Even if the populations are identical, you will conclude that the populations are different more than 5% of the time. Hayes and Cai reach the same conclusion (2). The Welch test must be specified as part of the experimental design.

- Use a permutation test. No GraphPad program offers such a test. The idea is to treat the observed values as a given, and to ask about the distribution of those values to the two groups. Randomly shuffle the values between the two groups, maintaining the original sample size. What fraction of those shuffled data sets have a difference between means as large (or larger) than observed. That is the P value. When the populations have different standard deviations, this test still produces reasonably accurate P values (Good, reference below, page 55). The disadvantage of these tests is that they don't readily yield a confidence interval. Learn more in Wikipedia, or Hyperstat.

What about switching to the nonparametric Mann-Whitney test? At first glance, this seems to be a good solution to the problem of unequal standard deviations. But it isn't! The Mann-Whitney test tests whether the distribution of ranks is different. If you know the standard deviations are different, you already know that the distributions are different. What you may still want to know is whether the means or medians are distinct. But when the groups have different distributions, nonparametric tests do not test whether the medians differ. This is a common misunderstanding.

How to avoid the problem

None of the solutions above are great. It is better to avoid the problem.

One approach to avoiding the problem is to think clearly about the distribution of your data, and transform the data as part of routine data processing. If you know a system creates lognormal data, analyze the logarithms always.

Another solutions is to always use the unequal variance (Welch) t test. As mentioned above, it is not a good idea to first test for unequal standard deviations, and use that results as the basis to decide whether to use the ordinary or modified (unequal variance, Welch) t test. But does it make

sense to always use the modified test? Ruxton makes a strong case that this is the best thing to do (3). So does Delacre (4). You lose some power when the standard deviations are, in fact, equal but gain power in the cases where they are not.

The Welch t test makes a strange set of assumptions. What would it mean for two populations to have the same mean but different standard deviations? Why would you want to test for that? Swailowsky points out that this situation simply doesn't often come up in science (5). I prefer to think about the unequal variance t test as a way to create a confidence interval. Your prime goal is not to ask whether two populations differ, but to quantify how far apart the two means are. The unequal variance t test reports a confidence interval for the difference between two means that is usable even if the standard deviations differ.

References

1. Moser, B.K. and G.R. Stevens Homogeneity of Variance in the Two Sample Means Test, The American Statistician, 1992;46(1):19-22.

2. Hayes and Cai. Further evaluating the conditional decision rule for comparing two independent means. Br J Math Stat Psychol (2007)

3. Ruxton. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology (2006) vol. 17 (4) pp. 688

4. Delacre, M., Lakens, D.L., and Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. Rips 30: 92–10.

5. S.S. Sawilowsky. Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means With Different Variances. J. Modern Applied Statistical Methods (2002) vol. 1 pp. 461-472

6. P.I. Good and J.W. Hardin, Common Errors in Statistics: (and How to Avoid Them), 2003, IBSN:0471460680.

5.9.2.5   The unequal variance Welch t test

Two unpaired t tests

When you choose to compare the means of two nonpaired groups with a t test, you have two choices:

- Use the standard unpaired t test. It assumes that both groups of data are sampled from Gaussian populations with the same standard deviation.

- Use the unequal variance t test, also called the Welch t test. It assues that both groups of data are sampled from Gaussian populations, but does not assume those two populations have the same standard deviation.

The usefulness of the unequal variance t test

To interpret any P value, it is essential that the null hypothesis be carefully defined. For the unequal variance t test, the null hypothesis is that the two population means are the same but the two population variances may differ. If the P value is large, you don't reject that null hypothesis, so conclude that the evidence does not persuade you that the two population means are different, even though you assume the two populations may have different standard deviations. What a strange set of assumptions. What would it mean for two populations to have the same mean but different standard deviations? Why would you want to test for that? Swailowsky points out that this situation simply doesn't often come up in science (1).

I think the unequal variance t test is more useful when you think about it as a way to create a confidence interval. Your prime goal is not to ask whether two populations differ, but to quantify how far apart the two means are. The unequal variance t test reports a confidence interval for the difference between two means that is usable even if the standard deviations differ.

How the unequal variance t test is computed

Both t tests report both a P value and confidence interval. The calculations differ in two ways:

Calculation of the standard error of the difference between means

The t ratio is computed by dividing the difference between the two sample means by the standard error of the difference between the two means. This standard error is computed from the two standard deviations and sample sizes. When the two groups have the same sample size, the standard error is identical for the two t tests. But when the two groups have different sample sizes, the t ratio for the Welch t test is different than for the ordinary t test. This standard error of the difference is also used to compute the confidence interval for the difference between the two means.

Calculation of the df

For the ordinary unpaired t test, df is computed as the total sample size (both groups) minus two. The df for the unequal variance t test is computed by a complicated formula that takes into account the discrepancy between the two standard deviations. If the two samples have identical standard deviations, the df for the Welch t test will be identical to the df for the standard t test. In most cases, however, the two standard deviations are not identical and the df for the Welch t test is smaller than it would be for the unpaired t test. The calculation usually leads to a df value that is not an integer. Prism reports and uses this fractional value for df. Many programs, including Prism 5, as well as InStat and our QuickCalc all round the df down to next lower integer. For this reason, the P value reported by Prism can be a bit smaller than the P values reported by other programs.

When to chose the unequal variance (Welch) t test

Deciding when to use the unequal variance t test is not straightforward.

It seems sensible to first test whether the variances are different, and then choose the ordinary or Welch t test accordingly. In fact, this is not a good plan. You should decide to use this test as part of the experimental planning.

What about always choosing the Welch test? Ruxton (2) and Delacre (3) make a strong case that this is a good idea. You lose some power when the standard deviations are, in fact, equal but gain power in the cases where they are not.

Reference

1. S.S. Sawilowsky. Fermat, Schubert, Einstein, and Behrens-Fisher: The Probable Difference Between Two Means With Different Variances. J. Modern Applied Statistical Methods (2002) vol. 1 pp. 461-472

2. Ruxton. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behavioral Ecology (2006) vol. 17 (4) pp. 688

3. Delacre, M., Lakens, D.L., and Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. Rips 30: 92–10.

5.9.2.6 Graphing tips: Unpaired t

Points or bars?
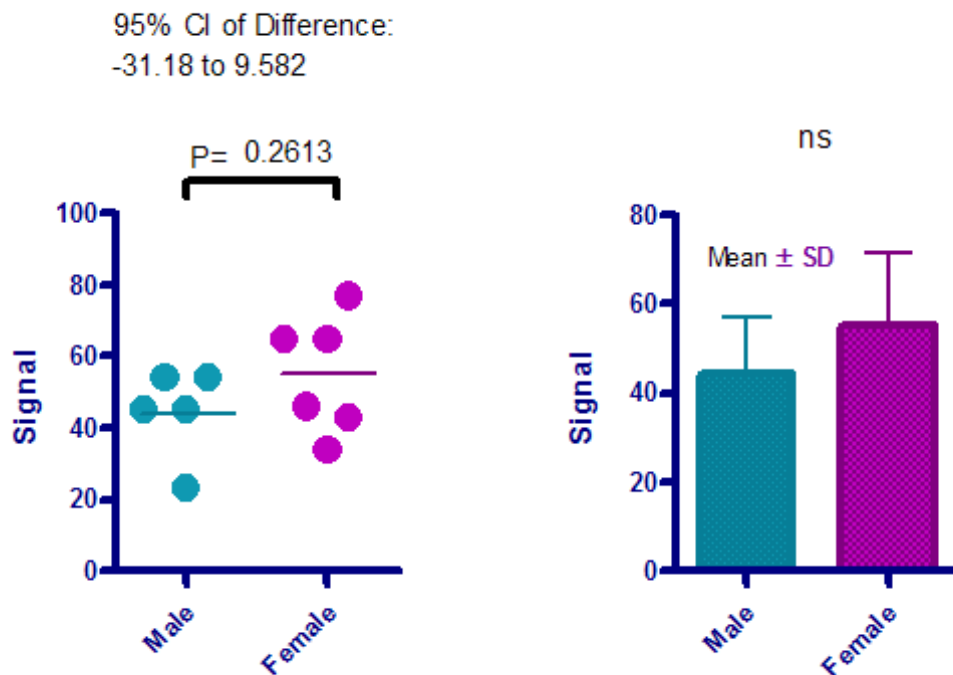
The graphs above plot the sample data for an unpaired t test. We prefer the graph on the left which shows each individual data point. This shows more detail, and is easier to interpret, than the bar graph on the right.

Graphing tips

- The scatter plot shows a horizontal line at the mean. If you choose the nonparametric Mann-Whitney test, you'll probably want to plot the median instead (a choice in the Format Graph dialog). Prism lets you turn off the horizontal line altogether.

- The horizontal line with caps is easy to draw. Draw a line using the tool in the Draw section of the toolbar. Then double click that line to bring up the Format Object dialog, where you can add the caps.

- The text objects "P=" and "95% CI of Difference" were created separately than the values pasted from the results. Click the text "T" button, then click on the graph and type the text.

- Don't forget to state somewhere how the error bars are calculated. We recommend plotting the mean and SD if you analyze with an unpaired t test, and the median and Interquartile range if you use the nonparametric Mann-Whitney test.

- If you choose a bar graph, don't use a log scale on the Y axis. The whole point of a bar graph is that viewers can compare the height of the bars. If the scale is linear (ordinary), the relative height of the bars is the same as the ratio of values measured in the two groups. If one bar is twice the height of the other, its value is twice as high. If the axis is logarithmic, this relationship does not hold. If your data doesn't show well on a linear axis, either show a table with the values, or plot a graph with individual symbols for each data point (which work fine with a log axis).

- For the same reason, make sure the axis starts at Y=0 and has no discontinuities. The whole idea of a bar graph is to compare height of bars, so don't do anything that destroys the relationship between bar height and value.

Including results on the graph

You can copy and paste any results from the results table onto the graph. The resulting embedded table is linked to the results. If you edit the data, Prism will automatically recalculate the results and update the portions pasted on the graph.

The graph on the left shows the exact P value. The graph on the right just shows the summary of significance ("ns" in this case, but one or more asterisks with different data). I recommend you show the exact P value.

The most useful information from an unpaired t test is the confidence interval for the difference between the two means, and this range is pasted onto the graph on the left.

5.9.2.7   **Advice: Don't pay much attention to whether error bars overlap**

When two SEM error bars overlap

When you view data in a publication or presentation, you may be tempted to draw conclusions about the statistical significance of differences between group means by looking at whether the error bars overlap. It turns out that examining whether or not error bars overlap tells you less than you might guess. However, there is one rule worth remembering:

> When SEM bars for the two groups overlap, and the sample sizes are equal, you can be sure the difference between the two means is not statistically significant (P>0.05).

When two SEM error bars do not overlap

The opposite is not true. Observing that the top of one standard error (SE) bar is under the bottom of the other SE error bar does not let you conclude that the difference is statistically significant. The fact that two SE error bars do **not** overlap does not let you make any conclusion about statistical significance. The difference between the two means might be statistically significant or the difference might not be statistically significant. The fact that the error bars do not overlap doesn't help you distinguish the two possibilities.

Other kinds of error bars

SD error bars

If the error bars represent standard deviation rather than standard error, then no conclusion is possible. The difference between two means might be statistically significant or the difference might not be statistically significant. The fact that the SD error bars do or do not overlap doesn't help you distinguish between the two possibilities.

Confidence interval error bars

Error bars that show the 95% confidence interval (CI) are wider than SE error bars. It doesn't help to observe that two 95% CI error bars overlap, as the difference between the two means may or may not be statistically significant.

Useful rule of thumb: If two 95% CI error bars do not overlap, and the sample sizes are nearly equal, the difference is statistically significant with a P value much less than 0.05 (Payton 2003).

With multiple comparisons following ANOVA, the significance level usually applies to the entire family of comparisons. With many comparisons, it takes a much larger difference to be declared "statistically significant". But the error bars are usually graphed (and calculated) individually for each treatment group, without regard to multiple comparisons. So the rule above regarding overlapping CI error bars does not apply in the context of multiple comparisons.

Summary of rules of thumb (assuming equal, or nearly equal, sample size and no multiple comparisons)

Type of error bar	Conclusion if they overlap	Conclusion if they don't overlap
SD	No conclusion	No conclusion
SEM	P > 0.05	No conclusion
95% CI	No conclusion	P < 0.05 (assuming no multiple comparisons)

There are two ways to think about this. If what you really care about is statistical significance, then pay no attention to whether error bars overlap or not. But if what you really care about is the degree to which the two distributions overlap, pay little attention to P values and conclusions about statistical significance.

Unequal sample sizes

The rules of thumb listed above are true only when the sample sizes are equal, or nearly equal.

Here is an example where the rule of thumb about confidence intervals is not true (and sample sizes are very different).

Sample 1: Mean=0, SD=1, n=10

Sample 2: Mean=3, SD=10, n=100

The confidence intervals do not overlap, but the P value is high (0.35).

And here is an example where the rule of thumb about SE is not true (and sample sizes are very different).

Sample 1: Mean=0, SD=1, n=100, SEM=0.1

Sample 2: Mean 3, SD=10, n=10, SEM=3.33

The SEM error bars overlap, but the P value is tiny (0.005).

5.9.2.8   Analysis checklist: Unpaired t test

The unpaired t test compares the means of two unmatched groups, assuming that the values follow a Gaussian distribution. Read elsewhere to learn about [choosing a t test][263], and [interpreting the results][276].

✔ Are the populations distributed according to a Gaussian distribution?

The unpaired t test assumes that you have sampled your data from populations that follow a Gaussian distribution. This assumption matters less with large samples due to the [Central Limit Theorem][41].

Prism can perform normality tests as part of the [Column Statistics](#)[201] analysis. [Learn more](#)[239].

If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values (perhaps to logs or reciprocals) to make the distributions more Gaussian. Another choice is to use the Mann-Whitney nonparametric test instead of a t test. A final option is to use t test anyway, knowing that it is fairly robust to violations of a Gaussian distribution with large samples.

✔ **Do the two populations have the same variances?**

The unpaired t test assumes that the two populations have the same variances (and thus the same standard deviation).

Prism tests for equality of variance with an F test. The P value from this test answers this question: If the two populations really have the same variance, what is the chance that you would randomly select samples whose ratio of variances is as far from 1.0 (or further) as observed in your experiment? A small P value suggests that the variances are different.

Don't base your conclusion solely on the F test. Also think about data from other similar experiments. If you have plenty of previous data that convinces you that the variances are really equal, ignore the F test (unless the P value is really tiny) and interpret the t test results as usual.

In some contexts, finding that populations have different variances may be as important as finding different means.

✔ **Are the data unpaired?**

The unpaired t test works by comparing the difference between means with the standard error of the difference, computed by combining the standard errors of the two groups. If the data are paired or matched, then you should choose a paired t test instead. If the pairing is effective in controlling for experimental variability, the paired t test will be more powerful than the unpaired test.

✔ Are the "errors" independent?

 The term "error" refers to the difference between each value and the group mean. The results of a t test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✔ Are you comparing exactly two groups?

 Use the t test only to compare two groups. To compare three or more groups, use one-way ANOVA[405] followed by multiple comparison tests. It is not appropriate to perform several t tests, comparing two groups at a time. Making multiple comparisons increases the chance of finding a statistically significant difference by chance and makes it difficult to interpret P values and statements of statistical significance. Even if you want to use planned comparisons to avoid correcting for multiple comparisons, you should still do it as part of one-way ANOVA to take advantage of the extra degrees of freedom that brings you.

✔ Do both columns contain data?

 If you want to compare a single set of experimental data with a theoretical value (perhaps 100%) don't fill a column with that theoretical value and perform an unpaired t test. Instead, use a one-sample t test[256] .

✔ Do you really want to compare means?

 The unpaired t test compares the means of two groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the two distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✔️ **If you chose a one-tail P value, did you predict correctly?**

If you chose a [one-tail P value](#)[75] , you should have predicted which group would have the larger mean before collecting any data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the P value reported by Prism and state that P>0.50.

5.9.3   Paired or ratio t test

5.9.3.1   How to: Paired t test

1. Enter data

From the Welcome (or New Table and graph) dialog, choose the Column tab.

If you are not ready to enter your own data, choose sample data and choose: t test - Paired.

Enter the data for each group into a separate column, with matched values on the same row. If you leave any missing values, that row will simply be ignored. Optionally, enter row labels to identify the source of the data for each row (i.e. subject's initials).
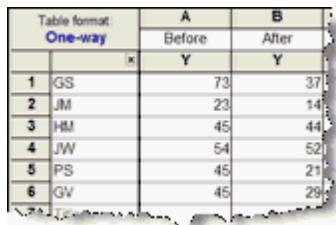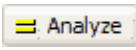
Table format One-way	A Before	B After
x	Y	Y
1 GS	73	37
2 JM	23	14
3 HU	45	44
4 JW	54	52
5 PS	45	21
6 GV	45	29

2. Choose the paired t test

1. From the data table, click 🟰 Analyze on the toolbar.

2. Choose t tests from the list of column analyses.

3. On the first (Experimental Design) tab of t test dialog, make these choices:

   • Experimental design: Paired

- Assume Gaussian distribution: Yes.

- Choose test: Paired t test

4. On the options tab, make these choices:

- Choose a [one- or two-sided P value](#)[75]. If in doubt, choose a two-tail P value.

- Choose the direction of the differences. This choice only affects the sign of the difference and the confidence interval of the difference, without affecting the P value.

- Choose a confidence level. Leave this set to 95%, unless you have a good reason to change it.

3. Review the results

The t test investigates the likelihood that the difference between the means of the two groups could have been caused by chance. So the most important results are the 95% confidence interval for that difference and the P value.

Learn more about interpreting the [results of a paired t test](#)[294].

Before accepting the results, [review the analysis checklist](#)[172].

4. Polish the graph
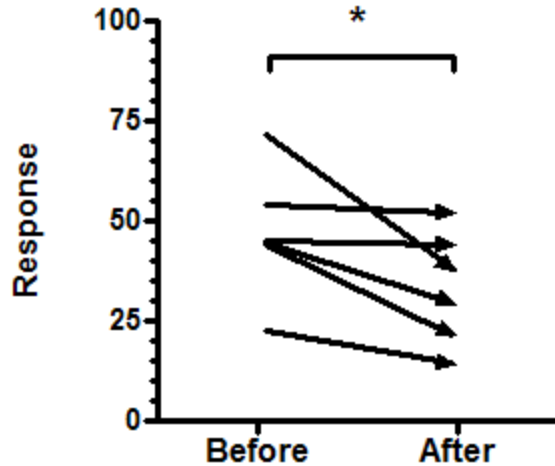
- A before-after graph shows all the data. This example plots each subject as an arrow to clearly show the direction from 'before' to 'after', but you may prefer to plot just lines, or lines with symbols.

- Avoid using a bar graph, since it can only show the mean and SD of each group, and not the individual changes.

- To add the asterisks representing significance level[86] copy from the results table and paste onto the graph. This creates a live link, so if you edit or replace the data, the number of asterisks may change (or change to 'ns'). Use the drawing tool to add the line below the asterisks, then right-click and set the arrow heads to "half tick down".

- Read more about graphing a paired t test[298].

5.9.3.2   Testing if pairs follow a Gaussian distribution

The paired t test assumes that you have sampled your pairs of values from a population of pairs where the difference between pairs follows a Gaussian distribution. If you want to test this assumption with a normality test, you need to go through some extra steps:

1. On the Options tab of the t test dialog, choose the option to graph the differences.

2. View the results table (part of the t test results) showing the differences. Click Analyze and choose Column statistics.

3. Choose the normality test(s) you want. We recommend D'Agostino's test. Note that none of the normality tests are selected by default, so you need to select at least one.

4. If the P value for the normality test is low, you have evidence that your pairs were not sampled from a population where the differences follow a Gaussian distribution. Read more about interpreting normality tests[243].

If your data fail the normality test, you have two options. One option is to transform the values (perhaps to logs or reciprocals) to make the distributions of differences follow a Gaussian distribution. Another choice is to use the Wilcoxon matched pairs nonparametric test instead of the t test.

Note that the assumption is about the set of *differences*. The paired t test does not assume that the two sets of data are each sampled from a Gaussian distribution, but only that the differences are consistent with a Gaussian distribution.

5.9.3.3 **Interpreting results: Paired t**

Confidence Interval

The paired t test compares the means of two paired groups, so look first at the difference between the two means. Prism also displays the confidence interval for that difference. If the assumptions of the analysis are true[172], you can be 95% sure that the 95% confidence interval contains the true difference between means.

P value

The P value is used to ask whether the difference between the mean of two groups is likely to be due to chance. It answers this question:

If the two populations really had the same mean, what is the chance that random sampling would result in means as far apart (or more so) than observed in this experiment?

It is traditional, but not necessary and often not useful, to use the P value to make a simple statement about whether or not the difference is "statistically significant[84]".

You will interpret the results differently depending on whether the P value is [small][79] or [large][81] .

t ratio

The paired t test compares two paired groups. It calculates the difference between each set of pairs and analyzes that list of differences based on the assumption that the differences in the entire population follow a Gaussian distribution.

First, Prism calculates the difference between each set of pairs, keeping track of sign. The t ratio for a paired t test is the mean of these differences divided by the standard error of the differences. If the t ratio is large (or is a large negative number) the P value will be small. The direction of the differences (Column A minus B, or B minus A) is set in the Options tab of the t test dialog.

The number of degrees of freedom equals the number of pairs minus 1. Prism calculates the P value from the t ratio and the number of degrees of freedom.

Test for adequate pairing

The whole point of using a paired experimental design and a paired test is to control for experimental variability. Some factors you don't control in the experiment will affect the before and the after measurements equally, so they will not affect the difference between before and after. By analyzing only the differences, a paired test corrects for those sources of scatter.

If pairing is effective, you expect the before and after measurements to vary together. Prism quantifies this by calculating the Pearson correlation coefficient, r. From r, Prism calculates a P value that answers this question:

> If the two groups really are not correlated at all, what is the chance that randomly selected subjects would have a correlation coefficient as large (or larger) as observed in your experiment? The P value has one-tail, as you are not interested in the possibility of observing a strong negative correlation.

If the pairing was effective, r will be positive and the P value will be small. This means that the two groups are significantly correlated, so it made sense to choose a paired test.

If the P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based on this one P value, but also on the experimental design and the results you have seen in other similar experiments.

If r is negative, it means that the pairing was counterproductive! You expect the values of the pairs to move together – if one is higher, so is the other. Here, the opposite is true – if one has a higher value, the other has a lower value. Most likely this is just a matter of chance. If r is close to -1, you should review your experimental design, as this is a very unusual result.

Descriptive statistics

The analysis tab of descriptive statistics summarizes only the data that was used for the paired t test. If you had any data in one column, but not the other, those values are not included in the descriptive statistics results that are included with the paired t test. But of course, the general descriptive statistics analysis analyzes all the data.

5.9.3.4 Analysis checklist: Paired t test

The paired t test compares the means of two matched groups, assuming that the distribution of the before-after differences follows a Gaussian distribution.

✔ **Are the differences distributed according to a Gaussian distribution?**

The paired t test assumes that you have sampled your pairs of values from a population of pairs where the difference between pairs follows a Gaussian distribution.

While this assumption is not too important with large samples, it is important with small sample sizes. Test this assumption with Prism[293].

Note that the paired t test, unlike the unpaired t test, does **not** assume that the two sets of data (before and after, in the typical example) are sampled from populations with equal variances.

✔ Was the pairing effective?

 The pairing should be part of the experimental design and not something you do after collecting data. Prism tests the effectiveness of pairing by calculating the Pearson correlation coefficient, r, and a corresponding P value. If the P value is small, the two groups are significantly correlated. This justifies the use of a paired test.

 If this P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based solely on this one P value, but also on the experimental design and the results of other similar experiments.

✔ Are the pairs independent?

 The results of a paired t test only make sense when the pairs are independent[32] – that whatever factor caused a difference (between paired values) to be too high or too low affects only that one pair. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six pairs of values, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may cause the after-before differences from one animal to be high or low. This factor would affect two of the pairs, so they are not independent.

✔ Are you comparing exactly two groups?

 Use the t test only to compare two groups. To compare three or more matched groups, use repeated measures one-way ANOVA followed by post tests. It is not appropriate[117] to perform several t tests, comparing two groups at a time.

✔ If you chose a one-tail P value, did you predict correctly?

 If you chose a one-tail P value, you should have predicted[75] which group would have the larger mean before collecting data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the reported P value and state that P>0.50.

✔ **Do you care about differences or ratios?**

The paired t test analyzes the differences between pairs. With some experiments, you may observe a very large variability among the differences. The differences are larger when the control value is larger. With these data, you'll get more consistent results if you perform a ratio t test[299].

5.9.3.5   Graphing tips: Paired t

Paired t test or Wilcoxon matched pairs test
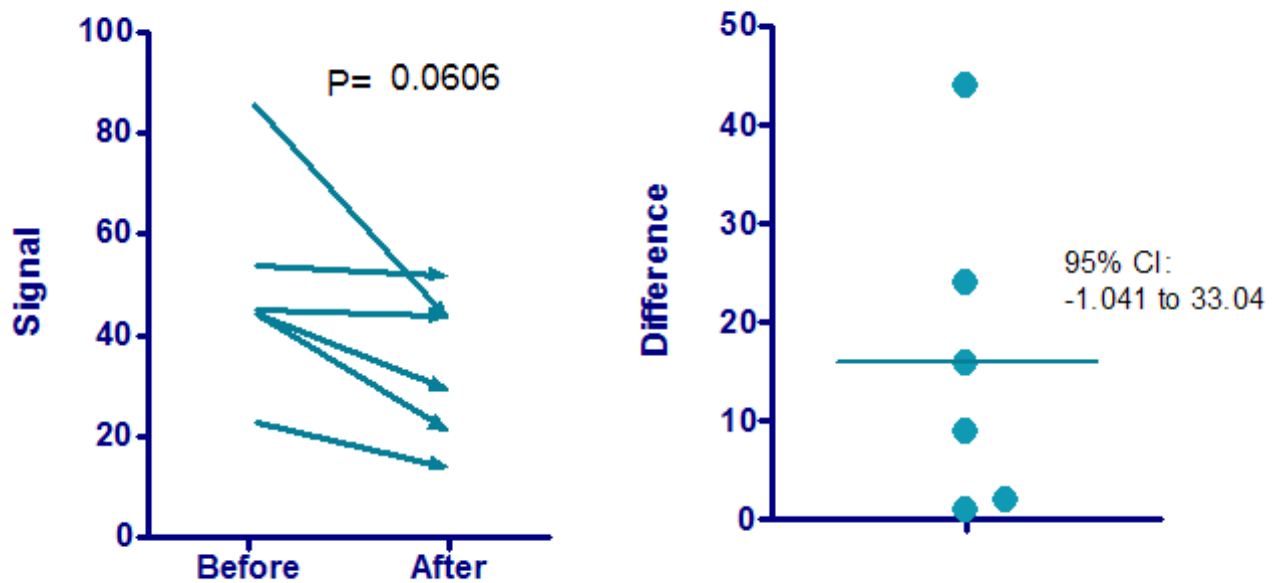
The graph above shows the sample data for a paired t test. Note the following:

- Since the data are paired, the best way to show the data is via a before after graph, as shown on the left. A bar graph showing the average value before and the average value after really doesn't properly display the results from a paired experiment.

- The graph uses arrows to show the sequence from Before to After. You may prefer to just show the lines with no arrowheads. Choose in the Format Graph dialog.

- The P value is copy and pasted from the paired t test analysis.

- The paired t test first computes the difference between pairs. The graph on the right shows these differences. These values can be computed using the Remove Baseline analysis, but there is no need to do so. On the options tab of the analysis dialog, check the option to graph the differences.

- The confidence interval for the difference between means shown on the right graph was copy and pasted from the paired t test results.

5.9.3.6 Paired or ratio t test?

Paired vs. ratio t tests

The paired t test analyzes the *differences* between pairs. For each pair, it calculates the difference. Then it calculates the average difference, the 95% CI of that difference, and a P value testing the null hypothesis that the mean difference is really zero.

The paired t test makes sense when the *difference* is consistent. The control values might bounce around, but the difference between treated and control is a consistent measure of what happened.

With some kinds of data, the difference between control and treated is not a consistent measure of effect. Instead, the differences are larger when the control values are larger. In this case, the *ratio* (treated/control) may be a much more consistent way to quantify the effect of the treatment.

Analyzing ratios can lead to problems because ratios are intrinsically asymmetric – all decreases are expressed as ratios between zero and one; all increases are expressed as ratios greater than 1.0. Instead it makes more sense to look at the logarithm of ratios. Then no change is zero (the logarithm of 1.0), increases are positive and decreases are negative.

A ratio t test averages the logarithm of the ratio of treated/control and then tests the null hypothesis that the population mean of that set of logarithms is really zero.

Because the ratio t test works with logarithms, it cannot be computed if any value is zero or negative. If all the values are negative, and you really want to use a ratio t test, you could transform all the values by taking their absolute values, and doing the ratio t test on the results. If some values are negative and some are positive, it makes no sense really to think that a ratio would be a consistent way to quantify effect.

How the ratio t test calculations work
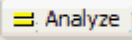
1. Transform all the values to their logarithm: Y=log(Y).

2. Perform a paired t test on the logarithms.

3. The antilogarithm of the difference between logarithms is the geometric mean of the ratios.

4. Calculate the antilogarithm of each confidence limit of the difference between the means of the logarithms. The result is the 95% confidence interval of the geometric mean of the ratios. Be sure to match the base of the logarithm and antilogarithm transform. If step 1 used common (base 10) logs, then this step should take 10 to the power of each confidence limit.

5.9.3.7   How to: Ratio t test

Prism can easily perform a ratio t test from two columns of data entered into a Column data table.

1. Create a column data table and enter two columns of data, with matched values on the same row. For example:

Control	Treated
4.2	8.7
2.5	4.9
6.5	13.1

2. From the data table, click ⊒ Analyze on the toolbar.

3. Choose t tests from the list of column analyses.

4. On the first (Experimental Design) tab of t test dialog, make these choices:

   • Experimental design: Paired

   • Assume Gaussian distribution: Yes. (Actually you are assuming a lognormal distribution of differences.)

   • Choose test: Ratio paired t test

5. On the second tab of the t test dialog, choose to compute Treated - Control, rather than Control - Treated. Note that even though the ratio t test computes a ratio, the choice on the dialog is worded as if the values were subtracted.

5.9.3.8   Interpreting results: Ratio t test

Example

You measure the Km of a kidney enzyme (in nM) before and after a treatment. Each experiment was done with renal tissue from a different animal.

Control	Treated	Difference	Ratio
4.2	8.7	4.5	2.09
2.5	4.9	2.4	1.96
6.5	13.1	6.6	2.02

Using a conventional paired t test, the 95% confidence interval for the mean difference between control and treated Km value is -0.72 to 9.72, which includes zero. The P value 0.07. The difference between control and treated is not consistent enough to be statistically significant. This makes sense because the paired t test looks at differences, and the differences are not very consistent.

The ratios are much more consistent, so it makes sense to perform the ratio t test. The geometric mean of the ratio treated/control is 2.02, with a 95% confidence interval ranging from 1.88 to 2.16. The data clearly show that the treatment approximately doubles the Km of the enzyme.

Analyzed with a paired t test, the results were ambiguous. But when the data are analyzed with a ratio t test, the results are very persuasive – the treatment doubled the Km of the enzyme.

The P value is 0.0005, so the effect of the treatment is highly statistically significant.

The P value answers this question:

> If there really were no differences between control and treated values, what is the chance of obtaining a ratio as far from 1.0 as was observed? If the P value is small, you have evidence that the ratio between the paired values is not 1.0.

Descriptive statistics

The analysis tab of descriptive statistics summarizes only the data that was used for the paired t test. If you had any data in one column, but not the other, those values are not included in the descriptive statistics results that are included with the paired t test. But of course, the general descriptive statistics analysis analyzes all the data.

5.9.3.9   Analysis checklist: Ratio t test

> The ratio t test compares the means of two matched groups, assuming that the distribution of the logarithms of the before/after ratios follows a Gaussian distribution.

✔ Are the log(ratios) distributed according to a Gaussian distribution?

The ratio t test assumes that you have sampled your pairs of values from a population of pairs where the log of the ratios follows a Gaussian distribution.

While this assumption is not too important with large samples, it is important with small sample sizes. [Test this assumption with Prism](#)[293].

✔ Was the pairing effective?

The pairing should be part of the experimental design and not something you do after collecting data. Prism tests the effectiveness of pairing by calculating the Pearson correlation coefficient, r, between the logarithms of the two columns of data. If the corresponding P value. If

the P value is small, the two groups are significantly correlated. This justifies the use of a paired test.

 If this P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based solely on this one P value, but also on the experimental design and the results of other similar experiments.

✔ Are the pairs independent?

 The results of a ratio t test only make sense when the pairs are independent[32] – that whatever factor caused a rato (of paired values) to be too high or too low affects only that one pair. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six pairs of values, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may cause the after-before differences from one animal to be high or low. This factor would affect two of the pairs, so they are not independent.

✔ Are you comparing exactly two groups?

 Use the t test only to compare two groups. To compare three or more matched groups, transform the values to their logarithms, and then use repeated measures one-way ANOVA followed by post tests. It is not appropriate[117] to perform several t tests, comparing two groups at a time.

✔ If you chose a one-tail P value, did you predict correctly?

 If you chose a one-tail P value, you should have predicted[75] which group would have the larger mean before collecting data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the reported P value and state that P>0.50.

✔ Do you care about differences or ratios?

 The ratio t test analyzes the logarithm of the ratios of paired values. The assumption is that the ratio is a consistent measure of experimental effect. With many experiments, you may observe that the difference

between pairs is a consistent measure of effect, and the ratio is not. In these cases, use a [paired t test][291], not the ratio t test.

5.9.4 Mann-Whitney or Kolmogorov-Smirnov test

5.9.4.1 Choosing between the Mann-Whitney and Kolmogorov-Smirnov tests

Both the Mann-Whitney and the Kolmogorov-Smirnov tests are nonparametric tests to compare two unpaired groups of data. Both compute P values that test the null hypothesis that the two groups have the same distribution. But they work very differently:

- The [Mann-Whitney test][309] first ranks all the values from low to high, and then computes a P value that depends on the discrepancy between the mean ranks of the two groups.

- The [Kolmogorov-Smirnov test][317] compares the cumulative distribution of the two data sets, and computes a P value that depends on the largest discrepancy between distributions.

Here are some guidelines for choosing between the two tests:

- The KS test is sensitive to any differences in the two distributions. Substantial differences in shape, spread or median will result in a small P value. In contrast, the MW test is mostly sensitive to changes in the median.

- The MW test is used more often and is recognized by more people, so choose it if you have no idea which to choose.

- The MW test has been extended to handle tied values. The KS test does not handle ties so well. If your data are categorical, so has many ties, don't choose the KS test.

- Some fields of science tend to prefer the KS test over the MW test. It makes sense to follow the traditions of your field.

5.9.4.2 How to: MW or KS test

1. Enter data

From the Welcome (or New Table and graph) dialog, choose the Column tab.

If you are not ready to enter your own data, choose sample data and choose: t test - unpaired.

Enter the data for each group into a separate column. The two groups do not have to have the same number of values, and it's OK to leave some cells empty. Since the data are unmatched, it makes no sense to enter any row titles.
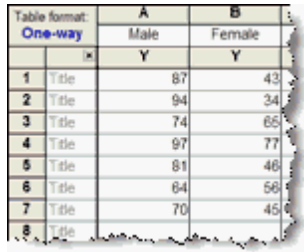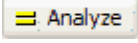
Table format: One-way	A Male	B Female
	Y	Y
1 Title	87	43
2 Title	94	34
3 Title	74	65
4 Title	97	77
5 Title	81	46
6 Title	64	56
7 Title	70	45
8 Title		

Note: The KS test works by comparing two cumulative frequency distributions. But you enter two stacks of data. Prism creates the frequency distributions as part of its analysis, but you can not enter frequency distributions and get Prism to compare them with the KS test.

2. Choose a test

1. From the data table, click **Analyze** on the toolbar.

2. Choose t tests from the list of column analyses.

3. On the t test dialog, choose the an unpaired experimental design, and that you do not wish to assume a Gaussian distribution.
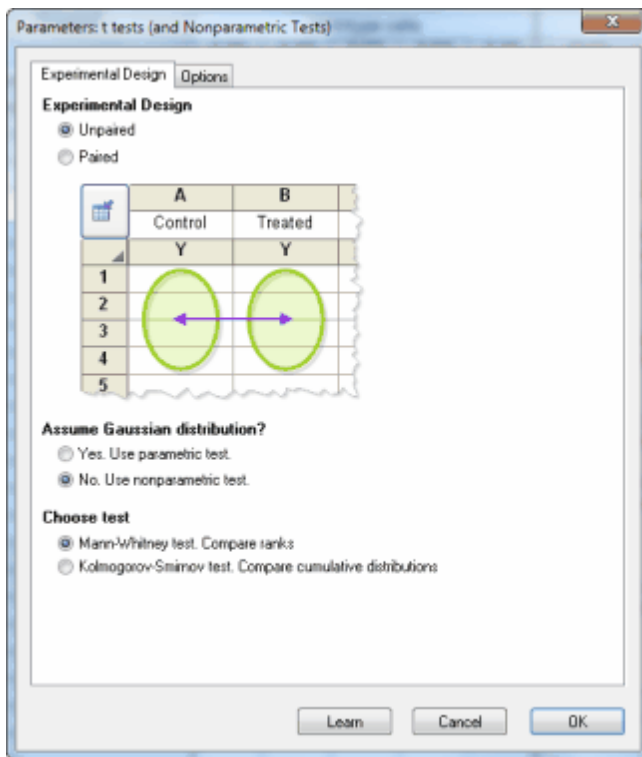
4. At the bottom of the first tab, choose either the Mann-Whitney (MW) or the Kolmogorov-Smirnov (KS) test. Here are some guidelines[304].

3. Choose options

These options, with the exception of the option to tabulate descriptive statistics, only apply to the MW test and not the KS test.

- Choose a <u>one- or two-tail P value</u>[75]. If in doubt, choose a two-tail P value.

- If you chose the option below to compute the 95% CI of the difference between medians, specify how Prism will compute that difference (A-B or B-A).

- The MW test works by comparing ranks. Check an option to graph those ranks.

- Create a table of descriptive statistics for each column.

- Compute the 95% confidence interval for the difference between medians. This calculation is only meaningful if you assume that the two population distributions have the same shape.

4. Review the results

Learn more about interpreting the [results of a Mann-Whitney test](#)[309].

Before accepting the results, review the [analysis checklist](#)[176].

5. Polish the graph
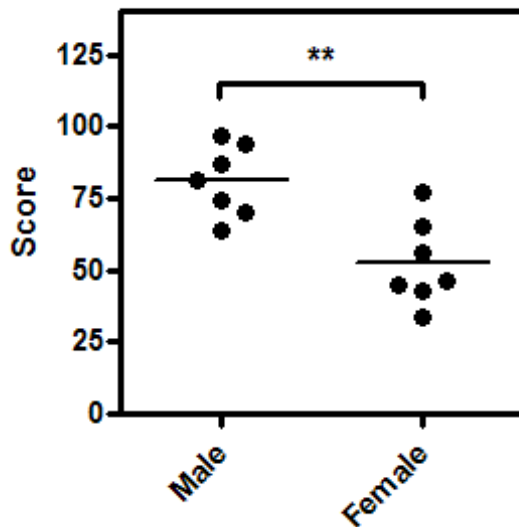
Graphing notes:

- A scatter plot shows every point. If you have more than several hundred points, a scatter plot can become messy, so it makes sense to plot a box-and-whiskers graph instead. We suggest avoiding bar graphs, as they show less information than a scatter plot, yet are no easier to comprehend.

- The horizontal lines mark the medians. Set this choice (medians rather than means) on the Welcome dialog, or change on the Format Graph dialog.

- To add the [asterisks representing significance level](#)[86] copy from the results table and paste onto the graph. This creates a live link, so if you edit or replace the data, the number of asterisks may change (or change to 'ns'). Use the drawing tool to add the line below the asterisks, then right-click and set the arrow heads to "half tick down".

5.9.4.3   Interpreting results: Mann-Whitney test

How it works

The Mann-Whitney test, also called the Wilcoxon rank sum test, is a nonparametric test that compares two unpaired groups. To perform the Mann-Whitney test, Prism first ranks all the values from low to high, paying no attention to which group each value belongs. The smallest number gets a rank of 1. The largest number gets a rank of n, where n is the total number of values in the two groups. Prism then averages the ranks in each group, and reports the two averages. If the means of the ranks in the two groups are very different, the P value will be small.

P value

You can't interpret a P value until you know the null hypothesis being tested. For the Mann-Whitney test, the null hypothesis is a bit hard to understand. The null hypothesis is that the distributions of both groups are identical, so that there is a 50% probability that an observation from a value randomly selected from one population exceeds an observation randomly selected from the other population.

The P value answers this question:

> If the groups are sampled from populations with identical distributions, what is the chance that random sampling would result in the mean ranks being as far apart (or more so) as observed in this experiment?

In most cases (including when ties are present), Prism calculates an exact P value(2). If your samples are large (the smaller group has more than 100 values), it approximates the P value from a Gaussian approximation. Here, the term Gaussian has to do with the distribution of sum of ranks and does not imply that your data need to follow a Gaussian distribution. The approximation is quite accurate with large samples and is standard (used by all statistics programs).

Note that Prism computes the exact P value much faster than did old versions, so does so with moderate size data sets where Prism 5 would have used an approximate method. It computes an exact P value when

the size of the smallest sample is less than or equal to 100, and otherwise computes an approximate one (with such large samples, the approximation is excellent).

If the P value is small, you can reject the null hypothesis that the difference is due to random sampling, and conclude instead that the populations are distinct.

If the P value is large, the data do not give you any reason to reject the null hypothesis. This is not the same as saying that the two populations are the same. You just have no compelling evidence that they differ. If you have small samples, the Mann-Whitney test has [little power](#)[149]. In fact, if the total sample size is seven or less, the Mann-Whitney test will always give a P value greater than 0.05 no matter how much the groups differ.

Mann-Whitney U and U'

Prism reports the value of the Mann-Whitney U value, in case you want to compare calculations with those of another program or text. To compute the U value, pick one value from group A and also pick a value from group B. Record which group has the larger value. Repeat for all values in the two groups. Total up the number of times that the value in A is larger than B, and the number of times the value in B is larger than the value in A. The smaller of these two values is U. The larger of the two values is U' (see below).

When computing U, the number of comparisons equals the product of the number of values in group A times the number of values in group B. If the null hypothesis is true, then the value of U should be about half that value. If the value of U is much smaller than that, the P value will be small. The smallest possible value of U is zero. The largest possible value is half the product of the number of values in group A times the number of values in group B.

Some programs also report U', but Prism doesn't. It can be easily computed as n1*n2 - U, where n1 and n2 are the two sample sizes, and U is reported by Prism. Prism defines U to be the smaller of the two values, so U' is the larger of the two. Some programs define U and U' based on which data set is entered first, so may reverse the definitions of U and U' for some analyses.

The difference between medians and its confidence interval

The Mann-Whitney test compares the distributions of ranks in two groups. If you assume that both populations have distributions with the same shape (which doesn't have to be Gaussian), it can be viewed as a comparison of two medians. Note that if you don't make this assumption, the Mann-Whitney test does not compare medians.

Prism reports the difference between medians only if you check the box to compare medians (on the Options tab). It reports the difference in two ways. One way is the obvious one -- it subtracts the median of one group from the median of the other group. The other way is to compute the Hodges-Lehmann estimate. Prism systematically computes the difference between each value in the first group and each value in the second group. The Hodges-Lehmann estimate is the median of this set of differences. Many think it is the best estimate for the difference between population medians.

Prism computes the confidence interval for the difference using the method explained on page 521-524 of Sheskin (1) and 312-313 of Klotz (3). This method is based on the Hodges-Lehmann method.

Since the nonparametric test works with ranks, it is usually not possible to get a confidence interval with exactly 95% confidence. Prism finds a close confidence level, and reports what it is. For example, you might get a 96.2% confidence interval when you asked for a 95% interval. Prism reports the confidence level it uses, which is as close as possible to the level you requested. When reporting the confidence interval, you can either report the precise confidence level ("96.2%") or just report the confidence level you requested ("95%"). I think the latter approach is used more commonly.

Prism computes an exact confidence interval when the smaller sample has 100 or fewer values, and otherwise computes an approximate interval. With samples this large, this approximation is quite accurate.

Tied values in the Mann-Whitney test

The Mann-Whitney test was developed for data that are measured on a continuous scale. Thus you expect every value you measure to be unique. But occasionally two or more values are the same. When the Mann-Whitney calculations convert the values to ranks, these values tie for the

same rank, so they both are assigned the average of the two (or more) ranks for which they tie.

Prism uses a standard method to correct for ties when it computes U (or the sum of ranks; the two are equivalent).

Unfortunately, there isn't a standard method to get a P value from these statistics when there are ties. When the smaller sample has 100 or fewer values, Prism computes the exact P value, even with ties(2). It tabulates every possible way to shuffle the data into two groups of the sample size actually used, and computes the fraction of those shuffled data sets where the difference between mean ranks was as large or larger than actually observed. When the samples are large (the smaller group has more than 100 values), Prism uses the approximate method, which converts U or sum-of-ranks to a Z value, and then looks up that value on a Gaussian distribution to get a P value.

Why Prism 6 and later can report different results than prior versions

There are two reasons why Prism 6 and later can report different results than prior versions:

• Exact vs. approximate P values. When samples are small, Prism computes an exact P value. When samples are larger, Prism computes an approximate P value. This is reported in the results. Prism 6 is much (much!) faster at computing exact P values, so will do so with much larger samples. It does the exact test whenever the smaller group has fewer than 100 values.

• How to handle ties? If two values are identical, they tie for the same rank. Prism 6, unlike most programs, computes an exact P value even in the presence of ties. Prism 5 and earlier versions always computed an approximate P value, and different approximations were used in different versions. Details.

Reference

1. DJ Sheskin, *Handbook of parametric and nonparametric statistical procedures*, 4th edition, 2007, ISBN=1584888148.

2. Ying Kuen Cheung and Jerome H. Klotz, The Mann-Whitney Wilcoxon distribution using linked lists, Statistical Sinica 7:805-813, 1997.

3. JH Klotz, A computational Approach to Statistics, 2006,
http://www.stat.wisc.edu/~klotz/Book.pdf

5.9.4.4 The Mann-Whitney test doesn't really compare medians

You'll sometimes read that the Mann-Whitney test compares the medians
of two groups. But this is not exactly true, as this example
demonstrates.
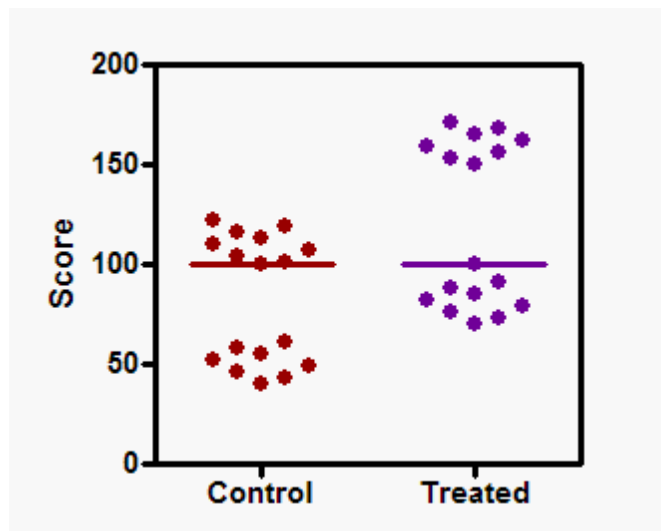
The graph shows each value obtained from control and treated subjects.
The two-tail P value from the Mann-Whitney test is 0.0288, so you
conclude that there is a statistically significant difference between the
groups. But the two medians, shown by the horizontal lines, are identical.
The Mann-Whitney test ranked all the values from low to high, and then
compared the mean ranks. The mean of the ranks of the control values is
much lower than the mean of the ranks of the treated values, so the P
value is small, even though the medians of the two groups are identical.

It is also not entirely correct to say that the Mann-Whitney test asks
whether the two groups come from populations with different
distributions. The two groups in the graph below clearly come from
different distributions, but the P value from the Mann-Whitney test is high
(0.46). The standard deviation of the two groups is obviously very
different. But since the Mann-Whitney test analyzes only the ranks, it
does not see a substantial difference between the groups.
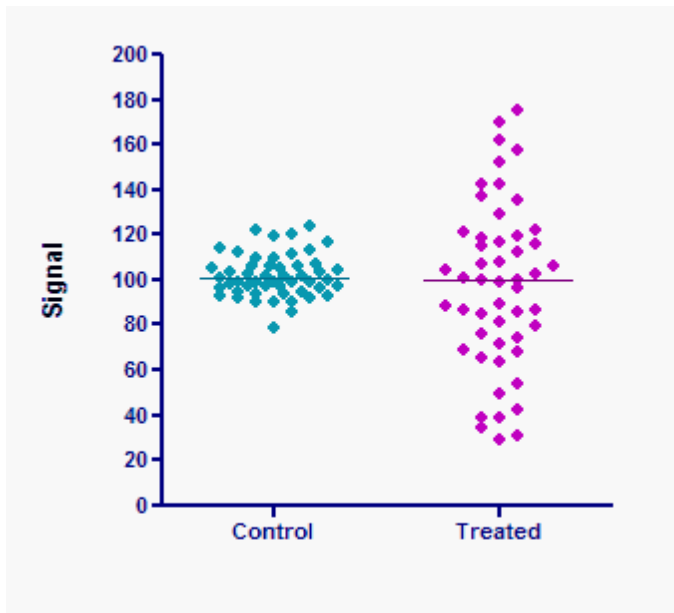
The Mann-Whitney test compares the mean ranks -- it does not compare medians and does not compare distributions. More generally, the P value answers this question: What is the chance that a randomly selected value from the population with the larger mean rank is greater than a randomly selected value from the other population?

If you make an additional assumption -- that the distributions of the two populations have the same shape, even if they are shifted (have different medians) -- then the Mann-Whiteny test can be considered a test of medians. If you accept the assumption of identically shaped distributions, then a small P value from a Mann-Whitney test leads you to conclude that the difference between medians is statistically significant. But Michael J. Campbell pointed out, "However, if the groups have the same distribution, then a shift in location will move medians and means by the same amount and so the difference in medians is the same as the difference in means. Thus the Mann-Whitney test is also a test for the difference in means."

The Kruskal-Wallis test is the corresponding nonparametric test for comparing three or more groups. Everything on this page about the Mann-Whitney test applies equally to the Kruskal-Wallis test.

 1. A. Hart. Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ (2001) vol. 323 (7309) pp. 391

5.9.4.5   Analysis checklist: Mann-Whitney test

The Mann-Whitney test[309] is a nonparametric test that compares the distributions of two unmatched groups. It is sometimes said to compare medians, but this is not always true[313].

✔ Are the "errors" independent?

The term "error" refers to the difference between each value and the group median. The results of a Mann-Whitney test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent[32] if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✔ Are the data unpaired?

The Mann-Whitney test works by ranking all the values from low to high, and comparing the mean rank in the two groups. If the data are paired or matched, then you should choose a Wilcoxon matched pairs test instead.

✔ Are you comparing exactly two groups?

Use the Mann-Whitney test only to compare two groups. To compare three or more groups, use the Kruskal-Wallis test followed by post tests. It is not appropriate to perform several Mann-Whitney (or t) tests, comparing two groups at a time.

✔ Do the two groups follow data distributions with the same shape?

If the two groups have distributions with similar shapes, then you can interpret the Mann-Whitney test as comparing medians. If the distributions have different shapes, you really cannot interpret[309] the results of the Mann-Whitney test.

✓ Do you really want to compare medians?

The Mann-Whitney test compares the medians of two groups ([well, not exactly](#)[313]). It is possible to have a tiny P value – clear evidence that the population medians are different – even if the two distributions overlap considerably.

✓ If you chose a one-tail P value, did you predict correctly?

If you chose a one-tail P value, you should have predicted which group would have the larger median before collecting any data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the P value reported by Prism and state that P>0.50. [One- vs. two-tail P values.](#)

✓ Are the data sampled from non-Gaussian populations?

By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions, but there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to give you a small P value), and that difference is quite noticeable with small sample sizes.

5.9.4.6   Why the results of Mann-Whitney test can differ from prior versions of Prism

The results of the Mann-Whitney test will not always match the results reported by older (5 and earlier) versions of Prism for two reasons.

- Exact vs. approximate P values. When samples are small, Prism computes an exact P value. When samples are larger, Prism computes an approximate P value. This is reported in the results. Prism is much (much!) faster at computing exact P values, so will do so with much larger samples. It does the exact test whenever the smaller group has fewer than 100 values.

- How to handle ties? If two values are identical, they tie for the same rank. Prism 6 and later, unlike most programs, computes an exact P value even in the presence of ties. Prism 5 an earlier versions always computed an approximate P value, and different approximations were used in different versions. [Details.](#) Bergmann and colleagues (1)

published a challenging example data set (with two variations) for the Mann-Whitney test, and Prism matches the results of StatXact, which they considered to be the gold standard (correct).

- Bugs. This page lists fixed bugs in Prism regarding this test.

1. Bergmann R, Ludbrook J. Different outcomes of the Wilcoxon—Mann—Whitney test from different statistics packages. The American …. 2000;54(1):72–7.

5.9.4.7   Interpreting results: Kolmogorov-Smirnov test

Key facts about the Kolmogorov-Smirnov test

- The two sample Kolmogorov-Smirnov test is a nonparametric test that compares the cumulative distributions of two data sets(1,2).

- The test is nonparametric. It does not assume that data are sampled from Gaussian distributions (or any other defined distributions).

- The results will not change if you transform all the values to logarithms or reciprocals or any transformation. The KS test report the maximum difference between the two cumulative distributions, and calculates a P value from that and the sample sizes. A transformation will stretch (even rearrange if you pick a strange transformation) the X axis of the frequency distribution, but cannot change the maximum distance between two frequency distributions.

- Converting all values to their ranks also would not change the maximum difference between the cumulative frequency distributions (pages 35-36 of Lehmann, reference 2). Thus, although the test analyzes the actual data, it is equivalent to an analysis of ranks. Thus the test is fairly robust to outliers (like the Mann-Whitney test).

- The null hypothesis is that both groups were sampled from populations with identical distributions. It tests for any violation of that null hypothesis -- different medians, different variances, or different distributions.

- Because it tests for more deviations from the null hypothesis than does the Mann-Whitney test, it has less power to detect a shift in the median

but more power to detect changes in the shape of the distributions (Lehmann, page 39).

- Since the test does not compare any particular parameter (i.e. mean or median), it does not report any confidence interval.

- Don't use the Kolmogorov-Smirnov test if the outcome (Y values) are categorical, with many ties. Use it only for ratio or interval data, where ties are rare.

- The concept of one- and two-tail P values only makes sense when you are looking at an outcome that has two possible directions (i.e. difference between two means). Two cumulative distributions can differ in lots of ways, so the concept of tails is not really appropriate. the P value reported by Prism essentially has many tails. Some texts call this a two-tail P value.

Interpreting the P value

The P value is the answer to this question:

> If the two samples were randomly sampled from identical populations, what is the probability that the two cumulative frequency distributions would be as far apart as observed? More precisely, what is the chance that the value of the Komogorov-Smirnov D statistic would be as large or larger than observed?

If the P value is small, conclude that the two groups were sampled from populations with different distributions. The populations may differ in median, variability or the shape of the distribution.

Graphing the cumulative frequency distributions

The KS test works by comparing the two cumulative frequency distributions, but it does not graph those distributions. To do that, go back to the data table, click Analyze and choose the Frequency distribution analysis. Choose that you want to create cumulative distributions and tabulate relative frequencies.

Don't confuse with the KS normality test

It is easy to confuse the two sample Kolmogorov-Smirnov test (which compares two groups) with the one sample Kolmogorov-Smirnov test,

also called the Kolmogorov-Smirnov goodness-of-fit test, which tests whether one distribution differs substantially from theoretical expectations.

The one sample test is most often used as a normality test to compare the distribution of data in a single dat aset with the predictions of a Gaussian distribution. Prism [performs this normality test](#)[239] as part of the Column Statistics analysis.

Comparison with the Mann-Whitney test

The Mann-Whitney test is also a nonparametric test to compare two unpaired groups. The Mann-Whitney test works by ranking all the values from low to high, and comparing the mean rank of the values in the two groups.

How Prism computes the P value

Prism first generates the two cumulative relative frequency distributions, and then asks how far apart those two distributions are at the point where they are furthest apart. Prism uses the method explained by Lehmann (2). This distance is reported as *Kolmogorov-Smirnov D*.

The P value is computed from this maximum distance between the cumulative frequency distributions, accounting for sample size in the two groups. With larger samples, an excellent approximation is used (2, 3).

An exact method is used when the samples are small, defined by Prism to mean when the number of permutations of n1 values from n1+n2 values is less than 60,000, where n1 and n2 are the two sample sizes. Thus an exact test is used for these pairs of group sizes (the two numbers in parentheses are the numbers of values in the two groups):

```
(2, 2), (2, 3) ... (2, 346)
(3, 3), (3, 4) ... (3, 69)
(4, 4), (4, 5) ... (4, 32)
(5, 5), (5, 6) ... (5, 20)
(6, 6), (6, 7) ... (6, 15)
(7, 7), (7, 8) ... (7, 12)
(8, 8), (8, 9), (8, 10)
(9, 9)
```

Prism accounts for ties in its exact algorithm (developed in-house). It systematically shuffles the actual data between two groups (maintaining sample size). The P value it reports is the fraction of these reshuffled data

sets where the D computed from the reshuffled data sets is greater than or equal than the D computed from the actual data.

References

1. Kirkman, T.W. (1996) Statistics to Use: Kolmogorov-Smirnov test. (Accessed 10 Feb 2010)

2. Lehmann, E. (2006), Nonparametrics: Statistical methods based on ranks. ISBN: 978-0387352121

3. WH Press, et. al, Numerical Recipes, third edition, Cambridge Press, ISBN: 0521880688

5.9.4.8   Analysis checklist: Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test[317] is a nonparametric test that compares the distributions of two unmatched groups.

✔ Are the values independent?

The results of a Kolmogorov-Smirnov test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the values are not independent[32] if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✔ Are the data unpaired?

The Kolmogorov-Smirnov test works by comparing the cumulative frequency distributions of the two groups.It does not account for any matching or pairing. If the data are paired or matched, consider using a Wilcoxon matched pairs test instead.

✔ Are you comparing exactly two groups?

Use the Kolmogorov-Smirnov test only to compare two groups. To compare three or more groups, use the Kruskal-Wallis test followed by

post tests. It is not appropriate to perform several Kolmogorov-Smirnov tests, comparing two groups at a time without doing some correction for multiple comparisons.

✔ Are the data sampled from non-Gaussian populations?

By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions, but there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to give you a small P value), especially with small sample sizes.

✔ Have you entered raw data f(and not frequency distributions)?

The Kolmogorov-Smirnov test compares two cumulative frequency distributions. Prism creates these distributions from raw data. Prism cannot run the Kolmogorov-Smirnov test from distributions you enter, only from raw data entered into two columns of a Column data table.

5.9.5   Wilcoxon matched pairs test

5.9.5.1   "The Wilcoxon test" can refer to several statistical tests

Wilcoxon's name is used to describe several statistical tests.

- The [Wilcoxon matched-pairs signed-rank test](322) is a nonparametric method to compare before-after, or matched subjects. It is sometimes called simply the Wilcoxon matched-pairs test.

- The [Wilcoxon signed rank test](257) is a nonparametric test that compares the median of a set of numbers against a hypothetical median.

- The Wilcoxon rank sum test is a nonparametric test to compare two unmatched groups. It is equivalent to the [Mann-Whitney test](309).

- The [Gehan-Wilcoxon test](592) is a method to compare survival curves.

The first two tests listed above are related. The matched-pairs signed-rank test works by first computing the difference between each set of matched pairs, and then using the Wilcoxon signed rank test to ask if the median of these differences differs from zero. Often the term "Wilcoxon

signed rank" test is used to refer to either test. This is not really confusing as it is usually obvious whether the test is comparing one set of numbers against a hypothetical median, or comparing a set of differences between matched values against a hypothetical median difference of zero.

5.9.5.2   How to: Wilcoxon matched pairs test

1. Enter data
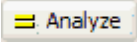
From the Welcome (or New Table and graph) dialog, choose the Column tab.

If you are not ready to enter your own data, choose sample data and choose: t test - Paired.

Enter the data for each group into a separate column, with matched values on the same row. If you leave any missing values, that row will simply be ignored. Optionally, enter row labels to identify the source of the data for each row (i.e. subject's initials).

Table format: One-way	A Before	B After
✕	Y	Y
1 GS	73	37
2 JM	23	14
3 HM	45	44
4 JW	54	52
5 PS	45	21
6 GV	45	29

2. Choose the Wilcoxon matched pairs test

1. From the data table, click ⇆ Analyze on the toolbar.

2. Choose t tests from the list of column analyses.

   3. On the first (Experimental Design) tab of t test dialog, make these choices:

- Experimental design: Paired

- Assume Gaussian distribution:No

- Choose test: Wilcoxon matched pairs test

4. On the options tab, make these choices:

- Choose a <u>one- or two-sided P value</u>[75]. If in doubt, choose a two-tail P value.

- Choose the direction of the differences. This choice only affects the sign of the difference and the confidence interval of the difference, without affecting the P value.

- Choose a confidence level. Leave this set to 95%, unless you have a good reason to change it.

- Choose which graphs to make. Graph differences? Graph correlation?

- <u>Choose how to handle</u>[329] rows where both values are identical.

3. Review the results

Learn more about interpreting the results of <u>Wilcoxon's matched pairs test</u>[324].

Before accepting the results, <u>review the analysis checklist</u>[177].

4. Polish the graph

- A before-after graph shows all the data. This example plots each subject as an arrow to clearly show the direction from 'before' to 'after', but you may prefer to plot just lines, or lines with symbols.

- Avoid using a bar graph, since it can only show the mean and SD of each group, and not the individual changes.

- To add the [asterisks representing significance level][86] copy from the results table and paste onto the graph. This creates a live link, so if you edit or replace the data, the number of asterisks may change (or change to 'ns'). Use the drawing tool to add the line below the asterisks, then right-click and set the arrow heads to "half tick down".

5.9.5.3   Results: Wilcoxon matched pairs test

Interpreting the P value

The Wilcoxon test is a nonparametric test that compares two paired groups. Prism first computes the differences between each set of pairs and ranks the absolute values of the differences from low to high. Prism then sums the ranks of the differences where column A was higher (positive ranks), sums the ranks where column B was higher (it calls these negative ranks), and reports the two sums. If the average sums of ranks are very different in the two groups, the P value will be small.

The P value answers this question:

If the median difference in the entire population is zero (the treatment is ineffective), what is the chance that random sampling would result in a median change as far from zero (or further) as observed in this experiment?

If the P value is small, you can reject the idea that the difference is due to chance, and conclude instead that the populations have different medians.

If the P value is large, the data do not give you any reason to conclude that the overall medians differ. This is not the same as saying that the medians are the same. You just have no compelling evidence that they differ. If you have small samples, the Wilcoxon test has little power to detect small differences.

How the P value is calculated

If there are fewer than 200 pairs, Prism calculates an exact P value. See more details in the page about the Wilcoxon signed rank test[257]. Prism 6 and later can do this even if there are ties. With more than 200 pairs, it calculates the P value from a Gaussian approximation. The term Gaussian, as used here, has to do with the distribution of sum of ranks and does not imply that your data need to follow a Gaussian distribution.

How Prism deals with pairs that have exactly the same value

What happens if some of the subjects have exactly the same value before and after the intervention (same value in both columns)?

When Wilcoxon developed this test, he recommended that those data simply be ignored. Imagine there are ten pairs. Nine of the pairs have distinct before and after values, but the tenth pair has identical values so the difference equals zero. Using Wilcoxon's original method, that tenth pairs would be ignored and the other nine pairs would be analyzed.This is how InStat and previous versions of Prism (up to version 5) handle the situation.

Pratt(1,2) proposed a different method that accounts for the tied values. Prism 6 and later offers the choice of using this method.

Which method should you choose? Obviously, if no pairs have identical before and after values, it doesn't matter. Nor does it matter much if there is, for example, only one such identical pair out of 200.

It makes intuitive sense that data should not be ignored, and so Pratt's method must be better. However, Conover (3) has shown that the relative merits of the two methods depend on the underlying distribution of the data, which you don't know.

95% Confidence interval for the median difference

Prism can compute a 95% confidence interval for the median of the paired differences (choose on the options tab). This can only be interpreted when you assume that the distribution of differences is symmetrical. Prism 6 and later uses the method explained in page 234-235 of Sheskin (Fourth Edition) and 302-303 of Klotz.

Test for effective pairing

The whole point of using a paired test is to control for experimental variability. Some factors you don't control in the experiment will affect the before and the after measurements equally, so they will not affect the difference between before and after. By analyzing only the differences, therefore, a paired test corrects for these sources of scatter.

If pairing is effective, you expect the before and after measurements to vary together. Prism quantifies this by calculating the nonparametric Spearman correlation coefficient, r_s. From r_s, Prism calculates a P value that answers this question: If the two groups really are not correlated at all, what is the chance that randomly selected subjects would have a correlation coefficient as large (or larger) as observed in your experiment? The P value is one-tail, as you are not interested in the possibility of observing a strong negative correlation.

If the pairing was effective, r_s will be positive and the P value will be small. This means that the two groups are significantly correlated, so it made sense to choose a paired test.

If the P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based on this one P value, but also on the experimental design and the results you have seen in other similar experiments (assuming you have repeated the experiments several times).

If r_s is negative, it means that the pairing was counterproductive! You expect the values of the pairs to move together – if one is higher, so is

the other. Here the opposite is true – if one has a higher value, the other has a lower value. Most likely this is just a matter of chance. If r_s is close to -1, you should review your procedures, as the data are unusual.

Why results might differ from those reported by earlier versions of Prism

Results from Prism 6 and later can differ from prior versions because Prism now does exact calculations in two situations where Prism 5 did approximate calculations. All versions of Prism report whether it uses an approximate or exact methods.

- Prism can perform the exact calculations much faster than did Prism 5, so does exact calculations with some sample sizes that earlier versions of Prism could only do approximate calculations.

- If the before-after differences for two pairs are the same, prior versions of Prism always used the approximate method. Prism 6 uses the exact method unless the sample is huge.

Prism reports whether it uses an approximate or exact method, so it is easy to tell if this is the reason for different results.

Descriptive statistics

The analysis tab of descriptive statistics summarizes only the data that was used for the Wilcoxon test. If you had any data in one column, but not the other, those values are not included in the descriptive statistics results that are included with the paired t test. But of course, the general descriptive statistics analysis analyzes all the data.

Reference

1. Pratt JW (1959) Remarks on zeros and ties in the Wilcoxon signed rank procedures. Journal of the American Statistical Association, Vol. 54, No. 287 (Sep., 1959), pp. 655-667

2. Pratt, J.W. and Gibbons, J.D. (1981), Concepts of Nonparametric Theory, New York: Springer Verlag.

3. WJ Conover, On Methods of Handling Ties in the Wilcoxon Signed-Rank Test, Journal of the American Statistical Association, Vol. 68, No. 344 (Dec., 1973), pp. 985-988

5.9.5.4   Analysis checklist: Wilcoxon matched pairs test

The Wilcoxon test is a nonparametric test that compares two paired groups. Read elsewhere to learn about [choosing a t test](#)[263], and [interpreting the results](#)[324].

✔ **Are the pairs independent?**

The results of a Wilcoxon test only make sense when the pairs are [independent](#)[32] – that whatever factor caused a difference (between paired values) to be too high or too low affects only that one pair. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six pairs of values, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may cause the after-before differences from one animal to be high or low. This factor would affect two of the pairs (but not the other four), so these two are not independent.

✔ **Is the pairing effective?**

If the P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based solely on this one P value, but also on the experimental design and the results you have seen in other similar experiments.

✔ **Are you comparing exactly two groups?**

Use the Wilcoxon test only to compare two groups. To compare three or more matched groups, use the Friedman test followed by post tests. It is [not appropriate](#)[117] to perform several Wilcoxon tests, comparing two groups at a time.

✔ **If you chose a one-tail P value, did you predict correctly?**

If you chose a [one-tail P value](#)[75], you should have predicted which group would have the larger median before collecting any data. Prism does not ask you to record this prediction, but assumes that it is correct. If your prediction was wrong, then ignore the P value reported by Prism and state that P>0.50.

✔ Are the data clearly sampled from non–Gaussian populations?

 By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions. But there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to give you a small P value), especially with small sample sizes. Furthermore, Prism (along with most other programs) does not calculate confidence intervals when calculating nonparametric tests. If the distribution is clearly not bell-shaped, consider transforming the values (perhaps to logs or reciprocals) to create a Gaussian distribution and then using a t test.

✔ Are the differences distributed symmetrically?

 The Wilcoxon test first computes the difference between the two values in each row, and analyzes only the list of differences. The Wilcoxon test does not assume that those differences are sampled from a Gaussian distribution. However it does assume that the differences are distributed symmetrically around their median.

5.9.5.5   How to handle rows where the before and after values are identical

The Wilcoxon matched pairs test is a nonparametric test to compare two paired groups.

Like the paired t test, the first step in calculating this test is to subtract one paired value from the other. If the values are before and after a treatment, the difference is the change with treatment.

The next step is to rank the absolute value of those differences.

But what happens if, for one particular pair of values, the two values are identical, so the before value is identical to the after value.

When Wilcoxon developed this test, he recommended that those data simply be ignored. Imagine there are ten pairs of values. In nine pairs, the before and after values are distinct, but in the tenth pair those two values are identical (to the precision recorded). Using Wilcoxon's original method, that tenth pair would be ignored and the data from the other nine pairs would be analyzed.This is how InStat and Prism (up to version 5) handle the situation.

Pratt(1) proposed a different method that accounts for the tied values. Prism 6 and later offers the choice of using this method.

Which method should you choose? Obviously, if there are no ties among paired values (no differences equal to zero), it doesn't matter. Nor does it matter much if there is, for example, one such pair out of 200.

It makes intuitive sense that data should not be ignored, and that Pratt's method is better. Connover (2) has shown that the relative merits of the two methods depend on the underlying distribution of the data, which you don't know.

1. Pratt, J.W. and Gibbons, J.D. (1981), Concepts of Nonparametric Theory, New York: Springer Verlag.

2. WJ Conover, On Methods of Handling Ties in the Wilcoxon Signed-Rank Test, Journal of the American Statistical Association, Vol. 68, No. 344 (Dec., 1973), pp. 985-988

5.9.6 Multiple t tests

5.9.6.1 How to: Multiple t tests

Distinguish the t test analysis from the multiple t test analysis

- The t test (and nonparametric) analysis compares two data set columns. Each set of replicate values are usually entered into a column, although Prism can also enter replicates entered into side-by-side subcolumns all on one row.

- The multiple t test analysis performs many unpaired t tests at once -- one per row. Replicates are entered into side-by-side subcolumns.

- It assumes there is no pairing of values on one row. Y2 in column A is not matched to Y2 of column B.

- It also assumes no pairing or matching of values stacked in a subcolumn. The value in Column A, replicate Y2 in row 1 is not matched to the value in Column A, replicate Y2 or row 2. (Violating this assumption won't change the results.)

How to perform a multiple t test analysis with Prism

1. Create a Grouped data table. Format the table either for entry of replicate values into subcolumns, or for entry of mean, SD (or SEM) and n.

2. Enter the data on two data set columns. One unpaired t test will be performed on each row of data.

3. Click Analyze, and choose "Multiple t tests -- one per row" from the list of analyses for Grouped data.

4. [Choose how to compute each test, and when to flag a comparison for further analysis.](#) 332

If you want to copy/paste (or export) the P values to another table or program

Note a potential problem when copying the P values (or adjusted P values) to another program. If any P values (or adjusted P values) show something like "<0.000001", when you copy and paste that value will be seen as text and may get ignored. To avoid this issue you need to do two things.

- In the Analysis tab of Preferences, choose either to always report P values in scientific format, or at least to report tiny P values in scientific format.
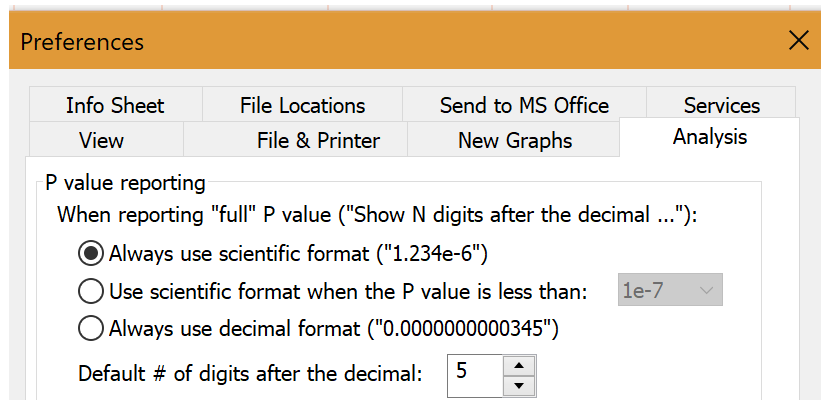
- In the parameters dialog for the multiple t test analysis, choose to report P values with N digits after the decimal and enter a large value for N.

Output

Show this many significant digits (for everything except P values): 4

P value style: Show N digits after the decimal. ns, *, **, ***, ∨ N = 12

Once you have made both of these choices, all P values will be numbers and the less than "<" symbol won't appear to confuse things.

5.9.6.2   Options for multiple t tests
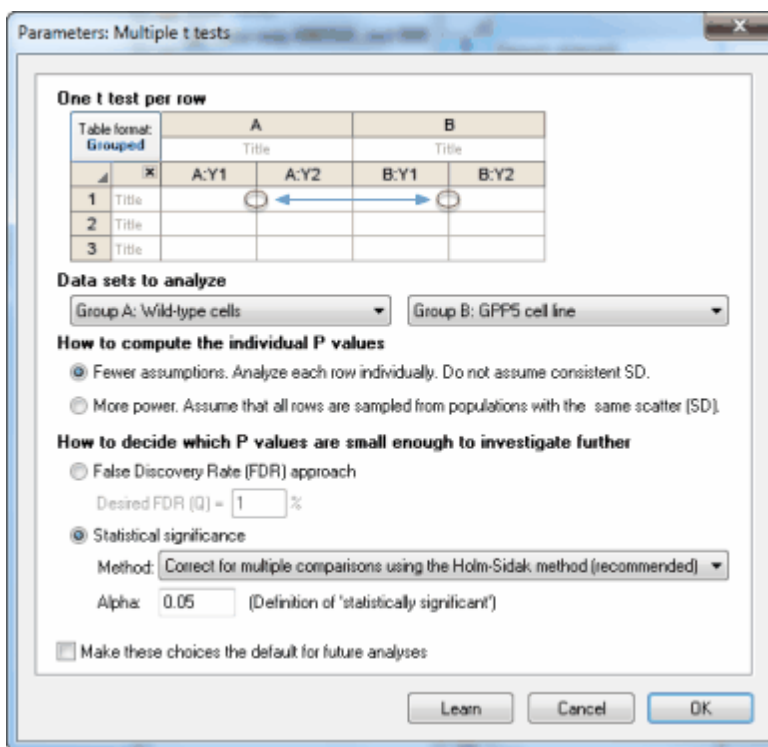
How to compute the individual P values

Prism computes an unpaired t test for each row, and reports the corresponding two-tailed P value. There are two ways it can do this calculation.

- **Fewer assumptions.** With this choice, each row is analyzed individually. The values in the other rows have nothing at all to do with how the values in a particular row are analyze. There are fewer df, so less power, but you are making fewer assumptions. Note that while you are not assuming that data on different rows are sampled from populations with identical standard deviations, you are assuming that data from the two columns on each row are sampled from populations with the same standard deviation. This is the standard assumption of an unpaired test -- that the two samples being compared are sampled from populations with identical standard deviations.

- **More power.** You assume that all the data from both columns and all the rows are sampled from populations with identical standard deviations. Your sample SDs vary, of course. But the assumption is that this variation is random, and really all the data from all rows comes from populations with the same SD. This is the assumption of homoscedasticity. Prism therefore computes one pooled SD, as it would by doing two-way ANOVA. This gives you more degrees of freedom and thus more power. Note the pooled SD is for both data set columns for all rows. So the data in each row will influence the P value not only for that row, but also for every other row.

Choosing between these options is not always straightforward. Certainly if the data in the different rows represent different quantities, perhaps measured in different units, then there would be no reason to assume that the scatter is the same in all. So if the different rows represent different gene products, or different measures of educational achievement (to pick two very different examples), then choose the "few assumptions" choice. If the different rows represent different conditions, or perhaps different brain regions, and all the data are measurements of the same outcome, then it might make sense to assume equal standard deviation and choose the "more power" option.

If in doubt, choose fewer assumptions.

How to decide which P values are small enough to investigate further

When performing a whole bunch of t tests at once, the goal is usually to come up with a subset of comparisons where the difference seems substantial enough to be worth investigating further. Prism offers two approaches to decide when a two-tailed P value is small enough to make that comparison worthy of further study.

One approach is based on the familiar idea of statistical significance.

The other choice is to base the decision on the False Discovery Rate (FDR; recommended). The whole idea of controlling the FDR is quite different than that of declaring certain comparisons to be "statistically significant".   This method doesn't use the term "significant" but rather the term "discovery". You set Q, which is the desired maximum percent of "discoveries" that are false discoveries. In other words, it is the maximum desired FDR.

Of all the rows of data flagged as "discoveries", the goal is that no more than Q% of them will be false discoveries (due to random scatter of data) while at least 100%-Q% of the discoveries are true differences between population means. Read more about FDR. [138] Prism offers three methods to control the FDR.

How to deal with multiple comparisons

If you chose the False Discovery Rate approach, you need to choose a value for Q, which is the acceptable percentage of discoveries that will prove to be false. Enter a percentage, not a fraction. If you are willing to accept 5% of discoveries to be false positives, enter 5 not 0.05. You also need to choose which method to use [386].

If you choose to use the approach of statistical significance, you need to make an additional decision about multiple comparisons. You have three choices:

- Correct for multiple comparisons using the Holm-Šídák method [133] (recommended). You specify the significance level, alpha, you want to apply to the entire family of comparisons. The definition of "significance" is designed so that if all the null hypotheses were true for every single row, the chance of declaring one or more row's comparison to be significant is alpha.

- Correct for multiple comparisons using the Šídák-Bonferroni method (not recommended). The Bonferroni method is much simpler to understand and is better known than the Holm-Šídák method, but it has no other advantages. The Holm-Šídák method has more power, and we recommend it. Note that if you choose the Bonferroni approach, Prism always uses the Šídák-Bonferroni method [131], often just called the Šídák method, which has a bit more power than the plain Bonferroni

(sometimes called Bonferroni-Dunn) approach -- especially when you are doing many comparisons.

- Do not correct for multiple comparisons (not recommended). Each P value is interpreted individually without regard to the others. You set a value for the significance level, alpha, often set to 0.05. If a P value is less than alpha, that comparison is deemed to be "statistically significant". If you use this approach, understand that you'll get a lot of false positives (you'll get a lot of "significant" findings that turn out not to be true). That's ok in some situations, like drug screening, where the results of the multiple t tests are used merely to design the next level of experimentation.

Volcano plot

New with Prism 8, Prism creates a volcano plot of your data. The X axis is the difference between means for each row. The Y axis plots the P value. Actually it plots the negative logarithm of the P value. So if P=0.01, log(P)=-2, and -log(P)=2, which is plotted. So rows with larger differences are further to either edge of the graph and rows with smaller P values are plotted higher on the graph.

Prism automatically places a vertical grid line at X=0 (no difference) and a horizontal grid line at Y=-log(alpha). Points above this horizontal grid line have P values less than the alpha you chose.

5.9.6.3   Interpreting results: Multiple t tests

The results are presented on two pages.

t tests page

This page shows the t test for each row. The P value for each comparison is shown, and also the multiplicity adjusted P value[129], which is higher.

If you chose the statistical significance approach, the first column is labeled "Significant?" and provides a "Yes" or "No" to indicate if the comparison was declared to be statistically significant, after any adjustments you requested for multiple comparisons. In prior versions of Prism, this column either contained a single asterisk (to indicate that a comparison was declared statistically significant) or not. However, it never showed several asterisks regardless of the corresponding P value.

To avoid confusion, this notation was replaced with a simple "Yes" or "No".

If you chose the FDR approach, then the first column is labeled "Discovery?" and contains an asterisk for those rows deemed to be "discoveries" and otherwise is blank.

Significant results page or Discoveries page

This page only shows the rows whose results meet the definition of a "statistically significant result" or a "discovery" based on your choices in the dialog. The rows are sorted by P value, with the smallest P value on top.

What is n?

In most cases, you have complete data on each row and the number of t tests equals the number of rows, so there is no ambiguity.

In some cases, it may not be possible to compute the t test for some rows (because the SD of the replicates is 0.0, or because there is only one value for each group). Prism only counts the rows for which a t test could be computed. If a t test could not be computed for some rows, Prism points this out in a floating note.

5.9.6.4   Volcano plot from multiple t tests

What is a volcano plot?

When you run multiple t tests, Prism (starting with version 8) automatically creates what is known as a *volcano plot*.

Volcano Plot
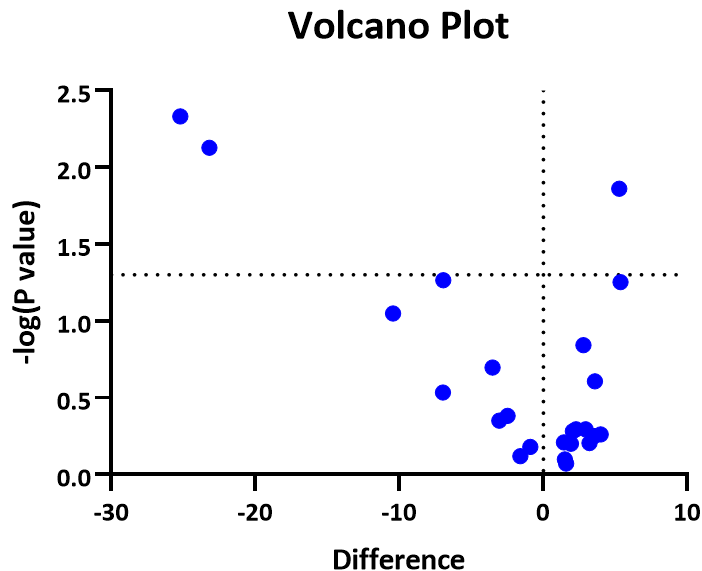
Each dot represents one row in your data table.

The X axis plots the difference between means. A dotted grid line is shown at X=0, no difference.

The value plotted on the Y axis depends on your choices.

- If you chose the statistical significance approach without correcting for multiple comparisons, then the Y value in the volcano plot is minus one times the logarithm of the P value. So if the P value is 0.01, then the logarithm (base 10) is -2, and the value plotted on the Y axis is 2.0.

- If you chose the statistical significance approach with correction for multiple comparisons, the Y axis is the minus logarithm of the multiplicity adjusted P value.

- If you chose the FDR approach, the Y value plots the minus logarithm of the q ratio.

Notes

- All logarithms are base 10.

- Prism automatically draws a grid line at X=0 (no difference) and at Y= -log(alpha) if you used the method of statistical significance and -log(Q) if you chose the method of FDR.

- The Volcano plot is created automatically and Prism does not offer the choice to not create it. But it is easy enough to delete.

- You'll often see volcano plots where X is the ratio between the two means. But the t test looks at differences, not ratios. If you care about ratios, consider transforming all the values to logarithms and then run the multiple t test analysis. This takes advantage of the fact that the difference between two logs is equal to the logarithm of their ratio.

- The Wikipedia article on Volcano plots will teach you more.

5.10   Nested t tests and ANOVA

Enter topic text here.

5.10.1 Overview of nested factors

Enter topic text here.

5.10.1.1 Example of a nested design with two treatments

Experimental design

You measure a variable in control and treated rats. There were three rats in each group, and you measured four technical replicates in each rat. Note that the four values stacked in each subcolumn are in arbitrary order. There is no time course or any other meaning to order of the four rows.

	Group A			Group B		
	Control			Treated		
	Rat 1	Rat 2	Rat 3	Rat 4	Rat 5	Rat 6
1	18	19	23	21	28	27
2	23	16	26	24	32	28
3	22	13	27	31	31	24
4	19	14	22	24	33	21

Why is this nested?

This design is said to be *nested* because each rat was either a control or treated animal. You cannot ask whether some rats respond better to treatments than others, because each rat only got one of the alternative treatments. Rats are said to be nested within treatment.

This is also called a *hierarchical* design. *Hierarchical* and *nested* are synonyms for describing this kind of design.

Incorrect analysis: t test of all data

It is tempting, but not appropriate, to treat these data as n=12 in each group.

	Group A Control	Group B Treated
1	18	21
2	23	24
3	22	31
4	19	24
5	19	28
6	16	32
7	13	31
8	14	33
9	23	27
10	26	28
11	27	24
12	22	21

If you run a t test on these data, the 95% confidence interval for the difference in means runs from 3.2 to 10.5 and the two-sided P value testing the null hypothesis of sampling from a population where the means are identical is 0.0008. This seems to be convincing evidence that the treatment increased the outcome variable. But theses results are not meaningful.

Why not? Because the t test assumes that each value provides independent information. These data have three independent rats for each treatment, but do not have twelve independent measurements of outcome. The replicates for each animal are closer to each other than to the values obtained from other animals. Another way to say this is that when you combine the three replicate rats and the four technical replicates within each rat, the 12 resulting values are *pseudoreplicates*. Analyzing pseudoreplicates as if they were actual replicates gives you a confidence interval that is too narrow and a P value that is too small.

Incorrect analysis: two-way ANOVA

The data as shown at the top of this page appear to be set up for two-way ANOVA. But running a two-way ANOVA would lead to incorrect or misleading results. Two-way ANOVA would assume that the kind of rat whose data is in row 2 is somehow exposed to control and treated conditions. But in fact the order of the values in each subcolumn is arbitrary, so testing the rows as a "factor" in two-way ANOVA would make no sense.

Alternative analysis without nested t test (works only when sample sizes are equal)

If there are no missing values, you can analyze the data with a t test. The first step is to average the technical replicates for each rat. Then enter those means on a new table, and then compare the two sets of means with an unpaired t test. Note that the three rats in each group are stacked in a column for the t test, but are side by side in subcolumns for the nested t test.

Unpaired t test	
P value	0.1058
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
t, df	t=2.081, df=4
How big is the difference?	
Mean of column A	20.17
Mean of column B	27.00
Difference between means (B - A) ± SEM	6.833 ± 3.283
95% confidence interval	-2.282 to 15.95
R squared (eta squared)	0.5200

	Group A	Group B
	Control	Treated
1	20.50	25.0
2	15.50	31.0
3	24.50	25.0

So long as there are no missing values, this gives the correct result. With this correct result, the 95% confidence interval for the difference runs from -2.3 to 15.9 and the P value is 0.1058. Note that the conclusion here (no evidence of treatment effect) is very different than the incorrectly conclusion you'd reach from analysis of the pseudoreplicates.

Nested t test in Prism

Prism 8 introduces a new analysis, Nested t test[345], that does the analysis in one step and can handle missing values. It assumes that the subcolumn means are sampled from a Gaussian population of subcolumn

means, and that the replicates within the subcolumns are sampled from a Gaussian population. The two Gaussian populations will generally have different standard deviations, and Prism calculates (estimates) both and reports them as both SD and Variance.

5.10.1.2 Example of a nested design with three treatments

Experimental design

This example comes from the excellent on-line statistics text titled How to Avoid and Detect Statistical Malpractice. The study assesses the effect of two methods of disease vector control versus no control on the packed cell volume (PCV) of cattle. Three herds of cows are randomly assigned to each of the three treatments. Blood samples are taken from four cows from each herd, and packed red cell volume is tabulated.

The factor we care about is treatment. The nested factor is herd. Each herd only gets one treatment, so herd is nested within treatment.

	Group A Traps + pour-ons			Group B Pour-on			Group C No vector control		
	Herd 1	Herd 2	Herd 3	Herd 4	Herd 5	Herd 6	Herd 7	Herd 8	Herd 9
1	28	32	27	25	26	25	21	19	18
2	26	27	25	24	28	26	19	18	20
3	27	28	29	27	29	24	17	23	19
4	31	29	27	23	27	23	20	20	18

Why is this nested?

This design is said to be *nested* because each herd was either a control or treated in one way. You cannot ask whether some herds respond better to treatments than others, because each herd only got one of the alternative treatments. Herds are said to be nested within treatment.

This is also called a *hierarchical* design. *Hierarchical* and *nested* are synonyms for describing this kind of design.

Incorrect analysis: one-way ANOVA of all data

It is tempting, but not appropriate, to treat these data as n=12 in each of the three treatment group.

	Group A Traps + pour-ons	Group B Pour-on	Group C No vector control
1	28	25	21
2	32	26	19
3	27	25	18
4	26	24	19
5	27	28	18
6	25	26	20
7	27	27	17
8	28	29	23
9	29	24	19
10	31	23	20
11	29	27	20
12	27	23	18

If you run one-way ANOVA on these data, the results will not be meaningful. ANOVA assumes that each value provides independent information. Three herds provided information for each treatment, with data from four cows in each herd. You do not have twelve independent measurements for each treatment. The replicates for each cow from one herd are closer to each other than to cows in other herds.

If you analyze the data this way, the P value is almost certainly going to be too small and the confidence intervals for the differences between means will be too narrow.

Incorrect analysis: two-way ANOVA

The data as shown at the top of this page appear to be set up for two-way ANOVA. But running a two-way ANOVA would lead to incorrect or misleading results. Two-way ANOVA would assume that all the cows whose data are in row 2 are somehow related to each other. But they are not. But in fact the order of the values in each subcolumn is arbitrary, so testing cow as a "factor" in two-way ANOVA would make no sense.

Alternative analysis without nested ANOVA (but only when sample sizes are equal)

If there are no missing values, you can analyze the data with one-way ANOVA. The first step is to average the technical replicates for each herd. Then enter those means on a new table, and then compare the three sets of means with one-way ANOVA.

	Group A	Group B	Group C
	Traps + pour-ons	Pour-on	No vector control
1	28.00	24.75	19.25
2	29.00	27.50	20.00
3	27.00	24.50	18.75

Note that the three herds in each treatment group are stacked in a column for the one-way ANOVA, but are side by side in subcolumns for the nested t test.

With no missing data, one-way ANOVA of these mean data will be identical to nested one-way ANOVA of the entire data set.

5.10.1.3 Lingo: Nested t test and Nested one-way ANOVA

Prism uses a unique set of terms.

- When there are two data sets, Prism offers the *nested t test.* We chose that name because it best describes the use of this test. You want to compare two treatment groups. There is another nested variable that needs to be accounted for, but that is not the primary goal. The term is not widely used and is sometimes used in a very different sense to mean a two-sample or indepependent sample t test.

- When there are three or more data sets, Prism offers *nested one-way ANOVA.* We use that name because it best describes the use of this

test. You want to compare three or more treatment groups. There is another nested variable that needs to be accounted for, but that is not the primary goal.

Other books and programs generally use the term *nested two-way ANOVA* in both cases, because one factor (rat in this example) is nested within another factor (treatment). The analysis accounts for two factors (one nested) so the analysis is named nested two-way ANOVA.

Another term for *nested* ANOVA is *hierarchical* ANOVA.

5.10.1.4 How Prism performs the nested t test and one-way ANOVA

Prism fits a mixed effects model. It treats the main factor (that defines the data set columns) as a fixed factor, and the nested factor as a random factor. Review the difference.[499]

If there are no missing values, this analysis gives identical results to a simple t test or one-way ANOVA where only the mean of each subcolumn is presented to the analysis. When there are missing values, there is no shortcut.

If your statistical consultants want to know exactly how Prism computes a nested t test, show them this faq which gives an example in Prism along with corresponding R and SAS code.

5.10.1.5 If P is high, should you pool?

In the nested t test example[345], the P value for Rats(Treatment) was 0.0239. Using the usual definition of statistical significance as $P<0.05$, you can reject the null hypothesis that all the rats with each treatment group have the same mean.

In the nested one-way ANOVA example, the P value for Herd(Treatment) is 0.1231. Since this is greater than the traditional threshold of 0.05, you

cannot reject the null hypotheses that all the herds within each treatment group have the same mean packed cell volume.

Because that P value is "high", should you conclude that there is no difference between the herds, so pool the data and run a regular t test? Difficult question.

- The attraction of this approach is that the confidence interval for the difference between means will be narrower and the P value will be smaller.

- The problem with this approach is that a high P value does not prove the herds (for this example) have identical means, it just says you don't have strong evidence that the means are not identical.

- Some statisticians suggest never pooling, seeing pooling as essentially a trick to get a lower P value for the main comparison. Others statisticians cautiously recommend pooling, but only when the P value for the nested factor is quite high (perhaps greater than 0.25 or even greater than 0.75).

- Prism does not facilitate pooling, and we do not recommend it.

5.10.2 Nested t test

How to: Nested t test[345]

Interpreting results: Nested t test[350]

Another example of a nested t test[352]

Analysis checklist: Nested t test[354]

5.10.2.1 How to: Nested t test

1. Data entry

This is the tutorial data provided with Prism for the nested t test. You are comparing two teaching methods (data set columns), each used in three rooms (subcolumns), with an outcome measured in between 4-6 students per rooms (rows).

If you are not using the tutorial data, create a Nested table, and set it up the number of subcolumns corresponding to the number of actual replicates you have. In the example, there are three rooms used for each teaching method, so create the table with three subcolumns.

Enter the data with the technical replicates stacked. In the example, you tested between 4 and 6 students in each room and these are stacked in each subcolumn. In a lab example, you might apply each alternative treatment to three rats (subcolumns) and measure something in each rat a number of times (technical replicates, stacked in the subcolumn)

If you want to label the subcolumns appropriately ("Room1"... below), double click on a subcolumn header to bring up the dialog where you enter all the subcolumn titles.

	Group A			Group B		
	Teaching method A			Teachng method B		
	Room 1	Room 2	Room 3	Room 4	Room 5	Room 6
1	21	18	35	26	38	31
2	26	25	28	34	44	41
3	33	26	32	27	34	34
4	22	24	36		45	35
5		21	38		38	38
6		25				46
7						

Notes:

- The term "technical replicates" may not always apply. If you are studying three hospitals in each group, with four doctors in each hospital, stack the information for each hospital in a subcolumn, with each doctor on a different row.

- Note that the replicates are stacked. This is different than Prism usually works. We set it up this way for two reasons. First, it lets you label the subcolumns ("Room 1", "Room 2"...above). Second, it matches the way most texts do this analysis. If you entered data with technical replicates (students in this example) side by side with different replicates (rooms in this example) in different rows, Prism's nested t test analysis would give meaningless results.

- The order of values in each subcolumn is arbitrary. You could randomly scramble the data in each subcolumn and the results won't change. The values in row 2 are not at all matched to each other.

- The order of the subcolumns doesn't matter. If you swapped the data for Room 2 and Room 3, the result would be the same. There is no connection between, say, the second subcolumn for the control data and the second column for the treated data.

- In this example, note that the subcolumns do not have the same number of values. The nested t tests works fine with unequal sample sizes.

- We use the name "nested t test" because it best describes the use of this test. Most books call this nested two-way ANOVA, because one factor (room in this example) is nested within another factor (teaching method).

- This example comes from Table 18.4 of Maxwell and Delaney (3rd edition). They label the second set of three classrooms 1, 2 and 3 rather than 4, 5 and 6. This is because of a quirk in how these data are analyzed in SPSS. There is no matching. The second classroom for the first teaching method (classroom 2) is not at all matched to the second classroom for the second teaching method (which we call classroom 5, but the text also calls classroom 2).

- Prism cannot run the nested t test with huge data sets and presents an message telling you so. How huge is huge? Details here.

2. Run the analysis

Click Analyze and then choose nested t test from the list of Grouped analyses.

On the first tab, enter the names of the two factors:

Parameters: Nested t test   ✕

Experimental Design   Options   Residuals

Data arrangement

	Group A		Group B	
	Hospital 1		Hospital 2	
	Doctor 1	Doctor 2	Doctor 3	Doctor 4
1				
2				
3				

Factor names

Name the factor that defines the columns:

    For the example: "Hospital"

> Treatment

Name the factor that defines the subcolumns:

    For the example: "Doctor"

> Rats

> To be sure you are doing the right analysis, imagine how the meaning of the data would change if you randomly scrambled the order of values in any subcolumn.
>
> For nested t test, the order of values in a subcolumn stack is irrelevant, and randomly scrambling that order won't affect the results.

Learn   Cancel   OK

The second tab offers options:

Parameters: Nested t test ✕

Experimental Design | Options | Residuals

Calculations

☐ Swap direction of comparisons (A-B) vs. (B-A).

Confidence level: 95% ▾

| Definition of statistical significance: P < 0.05 |

Graphing options

☑ Graph difference between means with confidence interval.

☑ Plot the data for each subcolumn.

Additional results

☐ Descriptive statistics & normality tests for each subcolumn.

Output

Show this many significant digits (for everything except P values): 4 ⏶⏷

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 ▾   N = 5 ⏶⏷

☐ Make options on this tab be the default for future nested t tests.

Learn | Cancel | OK

The third tab gives several choices for plotting residuals.

[Interpreting the results](#)³⁵⁰

5.10.2.2 Interpreting results: Nested t test

Important results

P value

Nested t test	
P value	0.1477
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.792, df=4
F, DFn, Dfd	3.210, 1, 4

The results are presented to be similar to an unpaired t test[276]. The P value tests the null hypothesis that the two treatment means are identical. The P value can be computed from either a t ratio (matching a t test) or a F ratio (since these kind of data are often analyzed with nested ANOVA). The t ratio is the square root of the F ratio (since the denominator df is zero) so the P value is the same either way. We show both so you can match other programs or texts

Confidence interval

How big is the difference?	
Mean of column A	35.63
Mean of column B	27.48
Difference between means (B - A) ± SEM	8.148 ± 4.548
95% confidence interval	-4.480 to 20.78

The most important result is the 95% confidence interval for the difference between two means. You can ask for 90% or 99% confidence intervals in the dialog if you prefer.

Other results (that most scientists will ignore)

Random effects	SD	Variance
Variation within subcolumns	5.160	26.63
Variation among subcolumns	4.501	20.26
Do the subcolumns differ?		
Chi-square, df	9.004, 1	
P value	0.0027	
P value summary	**	
Is there significant difference between subcolumns (P < 0.1)?	Yes	
Goodness of fit		
Degrees of freedom	25	
REML criterion	86.22	
Data analyzed		
Number of treatments (columns)	2	
Number of subjects (subcolumns)	6	
Total number of values	29	

The nested t test fits a *mixed model*. It is called mixed because the values stacked in subcolumns and the choice of subcolumns are assumed to be random, while the treatment (teaching method, for this example) is fixed. This means we care about testing these two teaching methods, but the choice of schools and students within schools is random. We don't care about those particular students or those particular schools. The model fits the variation within subcolumns and between subcolumns and reports these as both the variance and the standard deviation (which is the square root of variance). Prism reports these so you can compare with other programs or texts. Those values are not so helpful.

 You put the students into different subcolumns because you expect the results in different classrooms to be distinct. Prism tests the null hypothesis that in fact all the subcolumns (schools) within a column (teaching method) are identical. Here that P value is 0.0027 so you can reject that null hypothesis. This test is of marginal usefulness.

Prism optionally reports the REML goodness of fit to match other programs and books. There is no point trying to interpret it.

5.10.2.3 Another example of a nested t test

Example data

This example analyzes the data shown earlier[338] with replicate rats, each with multiple measurements (technical replicates).

	Group A			Group B		
	Control			Treated		
	Rat 1	Rat 2	Rat 3	Rat 4	Rat 5	Rat 6
1	18	19	23	21	28	27
2	23	16	26	24	32	28
3	22	13	27	31	31	24
4	19	14	22	24	33	21

Results of nested t test

	Nested t test Tabular results		
1	**Table Analyzed**	Nested	
2	Column B	Treated	
3	vs.	vs.	
4	Column A	Control	
5			
6	**Nested t test**		
7	P value	0.1058	
8	P value summary	ns	
9	Significantly different (P < 0.05)?	No	
10	One- or two-tailed P value?	Two-tailed	
11	t, df	t=2.081, df=4	
12	F, DFn, Dfd	4.332, 1, 4	
13			
14	**How big is the difference?**		
15	Mean of column A	27.00	
16	Mean of column B	20.17	
17	Difference between means (B - A) ± SEM	6.833 ± 3.283	
18	95% confidence interval	-2.282 to 15.95	
19			
20	**Random effects**	**SD**	**Variance**
21	Variation within subcolumns	2.915	8.500
22	Variation among subcolumn means	3.747	14.04
23			
24	**Do the subcolumns differ (within each column)?**		
25	Chi-square, df	9.244, 1	
26	P value	0.0024	
27	P value summary	**	
28	Is there significant difference between subcolumns (P < 0.05)?	Yes	
29			
30	**Data analyzed**		
31	Number of treatments (columns)	2	
32	Number of subjects (subcolumns)	6	
33	Total number of values	24	
34			

The 95% confidence interval for the difference between control and treated means range sfrom -2.3 to 15.9. The P value testing the null hypothesis of no difference between the two populations is 0.11. These data do not provide convincing evidence of a treatment effect.

The P value testing the null hypothesis that all the subcolumns (within each column) are identical is 0.0024. That provides evidence that the rats vary from one another.

5.10.2.4 Analysis checklist: Nested t test

The [nested t test][350] compares the means of two unmatched groups, where there is a nested factor within those treatment groups.

✓ **Are the residuals distributed according to a Gaussian distribution?**

The nested t test assumes that the residuals (variation among tecnhical replicates in many cases) are sampled from a Gaussian distribution. This assumption matters less with large samples due to the [Central Limit Theorem][41].

The third tab of the nested t test dialog lets you plot the residuals in several ways to assess their normality.

✓ **Does the variation within each of the subcolumns have the same variance?**

The nested t test assumes that the data in each subcolumn are sampled from populations with the same SD (same variance). Prism does not test this, but you can look at the data to see if this is badly violated.

Consider running the ANOVA on the logarithms of the values. In some cases this makes the variances much closer to being equal.

✓ **Is the variation among subcolumn means Gaussian?**

The nested t test assumes that variation among subcolumn means is Gaussian and also that the replicates within the subcolumns are Gaussian.

✓ **Are you comparing exactly two groups?**

Use the t test only to compare two groups. It is not appropriate to perform several nested t tests, comparing two treatment groups at a time.

✔️**Do both columns contain data?**

If you want to compare a single set of experimental data with a theoretical value (perhaps 100%) don't fill a column with that theoretical value and perform an unpaired t test. Instead, use a one-sample t test[256].

✔️**Do you really want to compare means?**

The nested t test compares the means of two groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

5.10.3 Nested one-way ANOVA

Enter topic text here.

5.10.3.1 How to: Nested one-way ANOVA

This example comes from the excellent on-line statistics text titled How to Avoid and Detect Statistical Malpractice. The study assesses the effect of two methods of disease vector control versus no control on the packed cell volume (PCV) of cattle. Three herds of cows are randomly assigned to each of the three treatments. Blood samples are taken from four cows from each herd, and packed red cell volume is tabulated.

The factor we care about is treatment. The nested factor is herd. Each herd only gets one treatment, so herd is nested within treatment.

1. Data entry

Create a Nested table, and set it up the number of subcolumns corresponding to the number of actual replicates you have. In the example, there are three herds in each treatment group, so create the table with three subcolumns.

Enter the data with the data from replicate cows stacked.

To label the subcolumns appropriately ("Herd1"... below), double click on a subcolumn header to bring up the dialog where you enter subcolumn titles.

	Group A Traps + pour-ons			Group B Pour-on			Group C No vector control		
	Herd 1	Herd 2	Herd 3	Herd 4	Herd 5	Herd 6	Herd 7	Herd 8	Herd 9
1	28	32	27	25	26	25	21	19	18
2	26	27	25	24	28	26	19	18	20
3	27	28	29	27	29	24	17	23	19
4	31	29	27	23	27	23	20	20	18

Notes:

- Note that the replicates are stacked. This is different than Prism usually works. We set it up this way for two reasons. First, it lets you label the subcolumns ("Herd 1", "Herd 2"...above). Second, it matches the way most texts do this analysis. If the technical replicates were side by side with different rats in different rows, Prism's nested t test analysis would give meaningless results.

- The order of values in each subcolumn is arbitrary. You could randomly scramble the data in each subcolumn and the results won't change. The values in row 2 are not at all matched to each other.

- The order of the subcolumns doesn't matter. If you swapped the data for Herd 2 and Herd 3, the result would be the same. There is no connection between, say, the second subcolumn for the control data and the second column for the treated data.

- In this example, every subcolumn has the same number of values (4). Nested one-way ANOVA does not require this and works fine with unequal sample sizes.

- We use the name "nested one-way ANOVA" because it best describes the use of this test. Most books call this nested two-way ANOVA, because one factor (herd in this example) is nested within another factor (treatment).

- Prism cannot run nested ANOVA with huge data sets and presents an message telling you so. How huge is huge? [Details here](#)

2. Run the analysis

Click Analyze and then choose nested one-way ANOVA from the list of Grouped analyses.

On the first tab, choose whether or not to generate a graph of the grand mean and 95% confidence interval for each data set, and whether or not to display the goodness of fit in the tabular results.

The second tab offers options for multiple comparisons, and the third tab gives several choices for plotting residuals.

[Interpreting the results](#)[358]

5.10.3.2 Interpreting results: Nested one-way ANOVA

Main results

	Nested 1way ANOVA		
1	**Table Analyzed**	Nested ANOVA	
2	Data sets analyzed	A-C	
3			
4	**Nested one-way ANOVA**		
5	P value	0.0003	
6	P value summary	***	
7	Significantly different (P < 0.05)?	Yes	
8	F, DFn, Dfd	43.21, 2, 6	
9			
10	**Random effects**	**SD**	**Variance**
11	Variation within subcolumns	1.724	2.972
12	Variation among subcolumn means	0.8036	0.6458
13			
14	**Do the subcolumns differ (within each column)?**		
15	Chi-square, df	1.089, 1	
16	P value	0.2967	
17	P value summary	ns	
18	Is there significant difference between subcolumns (P < 0.05)?	No	
19			
20	**Data analyzed**		
21	Number of treatments (columns)	3	
22	Number of subjects (subcolumns)	9	
23	Total number of values	36	
24			

- Nested ANOVA. This P value tests the null hypothesis that the three columns (treatments in this example) are the same on average.

- Random effects. The variation among subcolumn means and within subcolumns, expressed as both standard deviations (easier for scientists) and variances (more familiar to statisticians).

- Do the subcolumns differ (within each column). This P value tests the null hypothesis data for all subcolumns are sampled from populations with identical standard deviations.

- Goodness of fit (optional). If you choose this option, Prism reports the number of df and the REML criterion, which will only be meaningful to a statistician familiar with mixed models.

Multiple comparisons

Multiple comparisons are done as they are for one-way ANOVA. For this example, it makes sense to compare both treatments to the control in the third column with Dunnett's test. Here are the results.

	Nested 1way ANOVA Multiple comparisons								
1	Compare column means (main column effect)								
2									
3	Number of families	1							
4	Number of comparisons per family	2							
5	Alpha	0.05							
6									
7	Dunnett's multiple comparisons test	Mean Diff.	95.00% CI of diff.	Significant?	Summary	Adjusted P Value			
8									
9	Group C vs. Traps + pour-ons	-8.667	-11.42 to -5.912	Yes	***	0.0002			
10	Group C vs. Pour-on	-6.250	-9.005 to -3.495	Yes	**	0.0011			
11									
12									
13	Test details	Mean 1	Mean 2	Mean Diff.	SE of diff.	N1	N2	q	DF
14									
15	Group C vs. Traps + pour-ons	19.33	28.00	-8.667	0.9623	12	12	9.007	6
16	Group C vs. Pour-on	19.33	25.58	-6.250	0.9623	12	12	6.495	6

5.10.3.3 Analysis checklist: Nested one-way ANOVA

[Nested one-way ANOVA][358] compares the means of three or more unmatched groups, where there is a nested factor within those treatment groups.

✔ **Are the residuals distributed according to a Gaussian distribution?**

Nested ANOVA assumes that the residuals (variation among technical replicates in many cases) are sampled from a Gaussian distribution. This assumption matters less with large samples due to the [Central Limit Theorem][41] .

The third tab of the nested t test dialog lets you plot the residuals in several ways to assess their normality.

✓ Does the variation within each of the subcolumns have the same variance?

Nested ANOVA assumes that the data in each subcolumn are sampled from populations with the same SD (same variance). Prism does not test this, but you can look at the data to see if this is badly violated.

Consider running the ANOVA on the logarithms of the values. In some cases this makes the variances much closer to being equal.

✓ Is the variation among subcolumn means Gaussian?

Nested one-way ANOVA assumes that variation among subcolumn means is Gaussian and also that the replicates within the subcolumns are Gaussian.

✓ Do you really want to compare means?

The nested ANOVA compares the means of three or more groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✓ Is the main factor "fixed" rather than "random"?

Prism assumes the groups (data sets) are fixed factors. In other words, Prism tests for differences among the means of the particular groups you have collected data from. It is also possible that group (data sets) represents a random factor. This happens when you have randomly selected groups from an infinite (or at least large) number of possible groups, and that you want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment. Prism cannot perform nested ANOVA when the main factor is random.

Note that Prism does assume that the nested factor is random.

✔️ **Do the different columns represent different levels of a grouping variable?**

Nested one-way ANOVA asks whether the value of a single variable differs significantly among three or more groups. In Prism, you enter each group in its own column. If the different columns represent different variables, rather than different groups, then one-way ANOVA is not an appropriate analysis. For example, one-way ANOVA would not be helpful if column A was glucose concentration, column B was insulin concentration, and column C was the concentration of glycosylated hemoglobin.

.

5.11   Multiple comparisons after ANOVA

Interpreting multiple comparisons after ANOVA is tricky. It is important to know exactly what statistical significance means in this situation.

5.11.1   Overview on followup tests after ANOVA

.

5.11.1.1   Which multiple comparisons tests does Prism offer?

Overview

The followup multiple comparison tests that Prism offer differ a depend on which kind of ANOVA you are using, so there are separate instructions for [one-way ANOVA](#)[412], [two-way ANOVA](#)[475], and [three-way ANOVA](#)[531]. Note that the multiple comparisons choices are on two tabs of each ANOVA dialog:

- The Multiple Comparisons tab is where you set your goal, and this is quite different for each kind of ANOVA.

- The Options tab is where you choose the test you want to use, and these choices are similar for the three kinds of ANOVA and this page provides an overview that pertains to all three kinds of ANOVA.

The Options tab of all three ANOVA dialogs offers three big choices, corresponding to the three headings below, and additional choices within.

Correct for multiple comparisons using statistical hypothesis testing

The choices for multiple comparisons that Prism makes available to you depends on three questions:

- Your goal. Which comparisons do you want to make? Answer this question, based on your experimental goals, on the multiple comparisons tab[412] of the one-way ANOVA dialog.

- Do you want to include confidence intervals with your results? Not all multiple comparisons tests can compute confidence intervals. Answer this question, which is a personal preference not really linked to particular experimental designs, on the options tab[414] of the one-way ANOVA dialog.

- Nonparametric?

Goal	Report CI as well as significance?	Method
Compare every mean to every other mean	Yes	Tukey[136] (recommended) Bonferroni[131] Sidak[131]
	No	Holm-Sidak[133] (preferred) Newman-Keuls[138] Dunn (nonparametric)[137]
Compare every mean to a control mean	Yes	Dunnettt[136] Sidak[131] Bonferroni[131]

	No	Holm-Sidak[133] Dunn (nonparametric)[137]
Compare selected pairs of means (up to 40)	Yes	Bonferroni-Dunn [131] Sidak-Bonferroni[131]
	No	Holm-Sidak[133] Dunn (nonparametric)[137]
Linear trend? Do column mean correlate with column order?	No	Test for linear trend[442]. Only available with one-way ANOVA.

One-way ANOVA without assuming equal variances

If you do one-way ANOVA and choose to not assume equal SDs in all the populations, Prism performs alternative forms of ANOVA[431], also alternative multiple comparisons tests[384].

Correct for multiple comparisons by controlling the False Discovery Rate

Prism offers three methods[142] to control the false discovery rate[140]. All decide which (if any) comparisons to label as "discoveries" and do so in a way that controls the false discovery rate to be less than a value Q you enter.

When you choose to control the False discovery rate, Prism first computes an exact P value for each comparison. For regular ANOVA, it uses the Fishers LSD method. For nonparametric ANOVA, it uses Dunn's method without correcting for multiple comparisons. For one-way ANOVA without assuming equal variances, it uses the Welch t test. Then it takes this set of P value, and uses the method to control the false discovery rate that you chose and reports which comparisons are large enough (which P values are small enough) to be tagged as "discoveries".

Don't correct for multiple comparisons. Each comparison stands alone.

For regular (parametric) ANOVA, Prism computes these with the Fisher LSD test.[127] For one-way ANOVA without assuming equal variances, it computes P values using the Welch t test instead.

With nonparametric ANOVA, Prism uses [uncorrected Dunn's test](#)[382] which does not correct for multiplicity.

These P values, which do not account for multiple comparisons, will be smaller than multiplicity adjusted P values. If you report these P values, explain that they are not adjusted for multiple comparisons.

5.11.1.2   Multiple comparison results

Where are the multiple comparison results?

Starting with version 8, Prism presents some results on multiple tabs. Multiple comparison results are always on a separate results tab than the main ANOVA results. Click on the Multiple Comparisons tab to see those results.
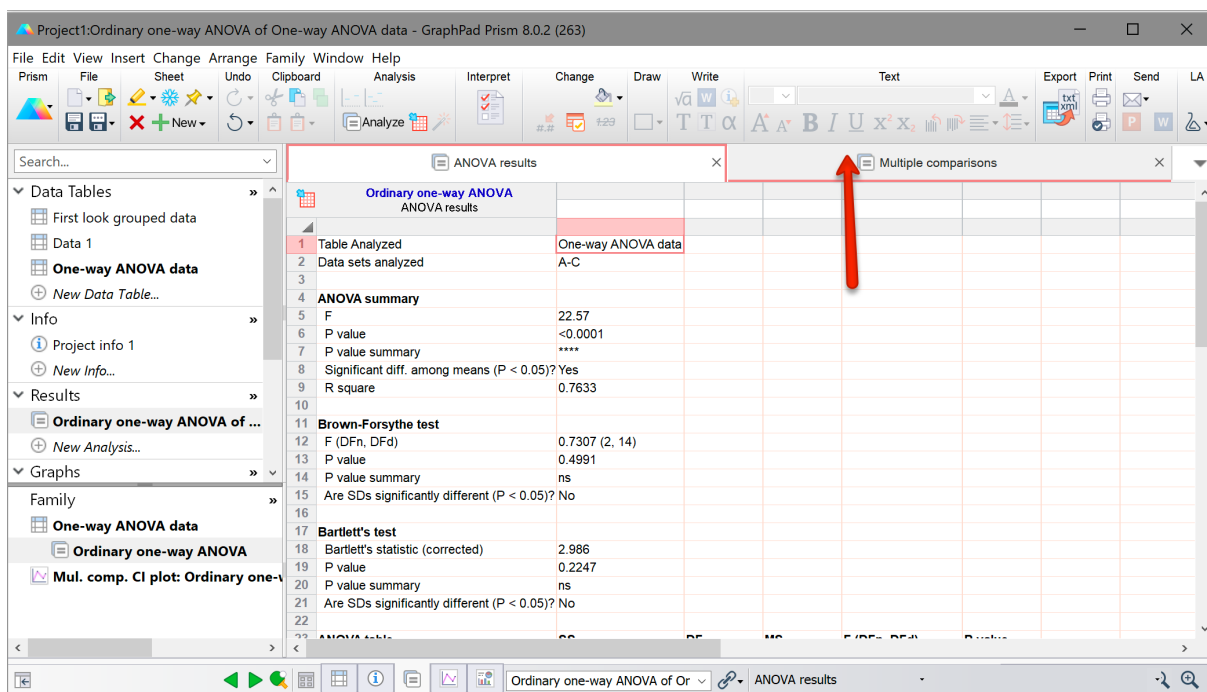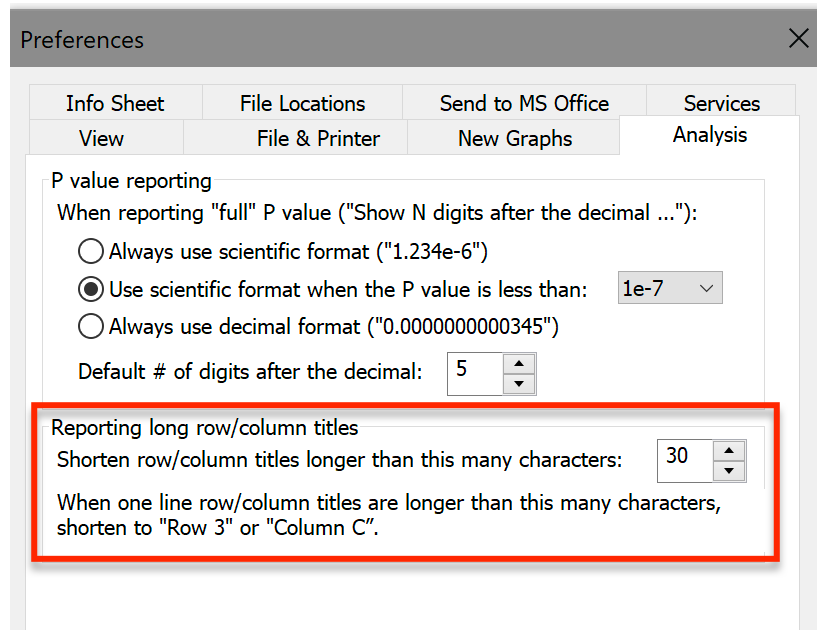

In Prism 8.2 the appearance of results tabs were updated, but are still found at the top of the results sheet.

When does Prism use column (or row) titles and when does it use it use column letters (or row numbers)?

When the titles are short, Prism will report the multiple comparison results using those titles, for example "Treated - Control". If the titles are on two or more lines or are long, Prism will report for example "Column B - Column A". How long is long? You set the threshold (for one line titles) in the Preferences dialog.

Preferences ✕

Info Sheet	File Locations	Send to MS Office	Services
View	File & Printer	New Graphs	Analysis

P value reporting
When reporting "full" P value ("Show N digits after the decimal ..."):

◯ Always use scientific format ("1.234e-6")

◉ Use scientific format when the P value is less than: 1e-7 ⌄

◯ Always use decimal format ("0.0000000000345")

Default # of digits after the decimal: 5 ▲▼

Reporting long row/column titles
Shorten row/column titles longer than this many characters: 30 ▲▼

When one line row/column titles are longer than this many characters, shorten to "Row 3" or "Column C".

5.11.1.3 Relationship between overall ANOVA and multiple comparisons tests

If the overall ANOVA finds a significant difference among groups, am I certain to find a significant post test?

If one-way ANOVA reports a P value of <0.05, you reject the null hypothesis that all the data come from populations with the same mean. In this case, it seems to make sense that at least one of the folow-up multiple comparisons tests will find a significant difference between pairs of means.

This is not necessarily true.

It is possible that the overall mean of group A and group B combined differs significantly from the combined mean of groups C, D and E. Perhaps the mean of group A differs from the mean of groups B through E. Scheffe's post test detects differences like these (but this test is not offered by Prism). If the overall ANOVA P value is less than 0.05, then Scheffe's test will definitely find a significant difference somewhere (if you look at the right comparison, also called contrast). The multiple comparisons tests offered by Prism only compare group means, and it is quite possible for the overall ANOVA to reject the null hypothesis that all

group means are the same yet for the followup tests to find no significant difference among group means.

If the overall ANOVA finds no significant difference among groups, are the multiple comparisons test results valid?

You may find it surprising, but all the multiple comparisons tests offered by Prism are valid even if the overall ANOVA did not find a significant difference among means. It is certainly possible that any of the tests offered by Prism can find significant differences even when the overall ANOVA showed no significant differences among groups. These tests are more focused, so have power to find differences between groups even when the overall ANOVA is not significant.

"An unfortunate common practice is to pursue multiple comparisons only when the null hypothesis of homogeneity is rejected." (Hsu, page 177)

"...these methods [e.g., Bonferroni, Tukey, Dunnet, etc.] should be viewed as substitutes for the omnibus test because they control alphaEW at thee desired level all by themselves. Requiring a significant omnibus test before proceeding to perform any of these analyses, as is sometimes done, only serves to lower alphaEW below the desired level (Bernhardson, 1975) and hence inappropriately decreases power" (Maxwell and Delaney, p. 236)

There are two exceptions, but both are tests not offered by Prism.

- Scheffe's test (not available in Prism) is intertwined with the overall F test. If the overall ANOVA has a P value greater than 0.05, then no post test using Scheffe's method will find a significant difference.

- Restricted Fisher's Least Significant Difference (LSD) test (not available in Prism). In this form of the LSD test, the multiple comparisons tests are performed only if the overall ANOVA finds a statistically significant difference among group means. But this restricted LSD test is outmoded, and no longer recommended. The LSD test in Prism[127] is unrestricted -- the results don't depend on the overall ANOVA P value and don't correct for multiple comparisons.

Are the results of the overall ANOVA useful at all? Or should I only look at multiple comparisons tests?

ANOVA tests the overall null hypothesis that all the data come from groups that have identical means. If that is your experimental question -- does the data provide convincing evidence that the means are not all identical -- then ANOVA is exactly what you want. More often, your experimental questions are more focused and answered by multiple comparison tests. In these cases, you can safely ignore the overall ANOVA results and jump right to the multiple comparisons results (some people disagree with this statement).

Note that the multiple comparison calculations all use the mean-square result from the ANOVA table. So even if you don't care about the value of F or the P value, the post tests still require that the ANOVA table be computed.

1. J. Hsu, Multiple Comparisons, Theory and Methods. ISBN: 0412982811.

2. SE Maxwell, HD Delaney, Designing Experiments and Analyzing Data: A Model Comparison Perspective, Second Edition, ISBN: 978-0805837186

5.11.1.4 Relationship between multiple comparisons tests and t tests

Fishers LSD method

The only difference a set of t tests and the Fisher's LSD test[127], is that t tests compute the pooled SD from only the two groups being compared, while the Fisher's LSD test computes the pooled SD from all the groups (which gains power). Note that unlike the Bonferroni, Tukey, Dunnett and Holm methods, Fisher's LSD does **not** correct for multiple comparisons.

Tukey, Dunnet, Bonferroni

Multiple comparisons use a familywise definition of alpha. The significance level doesn't apply to each comparison, but rather to the entire family of comparisons. In general, this makes it harder to reach significance. This is really the main point of multiple comparisons, as it reduces the chance

of being fooled by differences that are due entirely to random sampling. Here is an example:

Group 1	Group 2	Group 3
34	43	48
16	37	43
25	47	69

An unpaired t test comparing Groups 1 and 2 computes a P values of 0.0436, which is less than 0.05 so deemed statistically significant. But a Tukey multiple comparison test after ANOVA computes a multiplicity adjusted P value of 0.1789, which is not statistically significant.

In some cases, the effect of increasing the df (by pooling the SD of all groups, even when only comparing two) overcomes the effect of controlling for multiple comparisons. In these cases, you may find a 'significant' difference in a multiple comparisons test where you wouldn't find it doing a simple t test. Here is an example:

Group 1	Group 2	Group 3
34	43	48
38	45	49
29	47	47

Comparing groups 1 and 2 by unpaired t test yields a two-tail P value of 0.0164, while the Tukey multiple comparisons test calculates a multiplicity adjusted P value of 0.0073. If we set our threshold of 'significance' for this example to 0.01, the results are not 'statistically significant' with a t test but are statistically significant with the multiple comparisons test.

FDR approach

When you ask Prism to use the FDR approach (using any of the three methods), it first computes the P values using the Fisher LSD method (if you assume sampling from Gaussian distributions) or the <u>uncorrected</u>

Dunn's test[382] (nonparametric). These methods do not correct for multiple comparisons.

Then the three methods decide which of these P values are small enough to be called a "discovery", with the threshold depending on the distribution of P values, the number of P values, and which of the three methods you chose.

The first step, computing the P value is very close to what a conventional t test does. The second step, deciding which are "discoveries" is very different.

5.11.1.5 Correcting the main ANOVA P values for multiple comparisons?

Not considering followup multiple comparisons tests (post tests), how many P values does ANOVA compute in its main calculations?

- One-way ANOVA reports only one main P value.

- Two-way ANOVA reports three P value[488], one for each of the two factors and one for their interaction.

- Three-way ANOVA reports seven P values[538], one for each of the three factors, one for each of three two-way interaction and one for the three-way interaction.

Should one take multiple comparisons into account when interpreting the main P values from two- and three-way ANOVA? Statistical tradition has been to not do any correction, and GraphPad Prism follows this tradition (so only corrects for multiple comparisons for the followup tests that compare one treatment or cell with another.

Lakens argues (1) that a correction should be applied, to prevent too many false positives.

To correct for multiple comparisons of the main ANOVA P values in Prism, you should copy all the P values from the ANOVA results table and paste into one column of a Column table. If you did a three-way ANOVA, you would copy-paste seven P values into one new column. Then run the Analyze a stack of P values analysis[642] to correct for multiple

comparisons. You can correct for multiple comparisons using Bonferroni, Holm or by controlling the false discovery rate (FDR).

1. D. Lakens, Error Control in Exploratory ANOVA's: The How and the Why. Posted in his blog, The 20% statistician.

5.11.2 Interpreting results from multiple comparisons after ANOVA

.[371]

Confidence intervals from multiple comparisons tests[372]

Exact P values from multiple comparisons tests[373]

False Discovery Rate approach to multiple comparisons[374]

5.11.2.1 Statistical significance from multiple comparisons

If you correct for multiple comparisons using statistical hypothesis testing, one of the main results will be a decision for each comparison as to whether it is statistically significant or not. With all the methods except Fisher's LSD, these decisions correct for multiple comparisons. If all the data in all groups were really sampled from the same population, then there is a 5% (if you pick the traditional value for alpha) that any one (or more) of the comparisons would be designated as statistically significant. Note that the 5% probability is for the family of comparisons, not just for one comparison.

"Statistically significant" is not the same as "scientifically important".:

- It is better in many cases to focus on the the size of the difference and the precision of that value quantified as a confidence interval[372].

- Rather than just report which comparisons are, or are not, "statistically significant", Prism can report multiplicity adjusted P values[373] for many tests, and these can be more informative than a simple statement about which comparisons are statistically significant or not.

- Don't get mislead into focusing on whether whether or not error bars overlap. That doesn't tell you much about whether multiple

comparisons tests will be statistically significant or not. If two SE error bars overlap, you can be sure that a multiple comparison test comparing those two groups will find no statistical significance. However if two SE error bars do not overlap, you can't tell whether a multiple comparison test will, or will not, find a statistically significant difference. And if you plot SD error bars, rather than SEM, the fact that they do (or don't) overlap does not let you reach any conclusion about statistical significance. Details.

- With one-way ANOVA, you can choose to test for linear trend between column mean and column order and Prism will report the slope. Details here[442]. This test will tell you whether or not the trend is statistically significant or not.

5.11.2.2 Confidence intervals from multiple comparisons tests

If you don't need to make a decision from each comparison, you don't need each comparison to be reported as "statistically significant" or not. In this situation, ignore the conclusions about statistical significance and the P values. Instead focus on how large each difference is and how wide each confidence interval is. When thinking about confidence intervals, you need to think about how large a difference you would consider to be scientifically important. How small a difference would you consider to be scientifically trivial? Use scientific judgment and common sense to answer these questions. Statistical calculations cannot help, as the answers depend on the context and goals of the experiment.

If you do want to focus on confidence intervals, then make sure you pick a multiple comparisons method that can report them: The methods of Tukey, Dunnett, Bonferroni, and Sidak.

Note that the confidence intervals reported with multiple comparisons tests (except for Fisher's LSD) adjust for multiple comparisons. They are sometimes called *simultaneous confidence intervals*. If you make the usual assumptions and run lots of experiments, you expect that in 95% of these experiments *all 95%* confidence intervals will contain the true population values, which leaves 5% of experiments where *any one or more* of the intervals do not include the population value.

5.11.2.3 Exact P values from multiple comparisons tests

Prism reports exact P values from multiple comparisons tests using two different approaches that are quite different.

Multiplicity adjusted P values

Prism can compute a *multiplicity adjusted P value* for each comparison for many multiple comparison methds: The methods of Tukey, Dunnett, Bonferroni, Sidak, Dunn, and Holm. Since adjusted P values are not reported by most programs, and are not widely reported in scientific papers (yet), be sure <u>you fully understand what they mean</u>[129] before reporting these values.

A <u>multiplicity adjusted P value</u>[129] is the family-wise significance level at which that particular comparison would just barely be considered statistically significant. That is a hard concept to grasp. You can set the threshold of significance, for the whole family of comparisons, to any value you want. Usually, it is set to 0.05 or 0.01 or perhaps 0.10. But it can be set to any value you want, perhaps 0.0345. The adjusted P value is the smallest significance threshold, for the entire family of comparisons, at which this one comparison would be declared "statistically significant".

The adjusted P value for each comparison depends on all the data, not just the data in the two groups that P value compares. If you added one more comparison to the study (or took one away), all the adjusted P values would change. The adjusted P value can be thought of as a measure of the strength of evidence.

P values that don't correct for multiple comparisons

<u>Fisher's Least Significant Difference (LSD) test</u>[127] computes a P value (and confidence interval) for each comparison, without adjusting for multiple comparisons. The results will be similar to performing independent t tests for each comparison, except the Fishers LSD test uses all the data to compute a pooled standard deviation (rather than using the variation only in the two groups being compared). This will usually give it more power than independent t tests. When reporting P values from the Fishers LSD test, be sure to explain that these do not account for multiple comparisons, the reader must do so when evaluating the results.

The [uncorrected Dunn's test](#)[382] is the nonparametric test that computes a P value for each comparison without correcting for multiple comparisons.

Adjusted P values are very different than P values that don't account for multiple comparisons

Multiplicity adjusted P values, as the name suggests, accounts for multiple comparisons.

The Fisher LSD test and the uncorrected Dunn's test (nonparametric) do not account for multiple comparisons.

The "exact" P values computed by the two approaches are not the same. If you report either, be sure to be very explicit about exactly what P value you are reporting.

5.11.2.4 False Discovery Rate approach to multiple comparisons

When you choose the [False Discovery Rate (FDR) approach](#)[140] to multiple comparisons after ANOVA, Prism does the following:

1. Perform the comparisons you requested using the unprotected Fisher's LSD test. This results in a P value for each comparison. These P values do **not** correct for multiple comparisons. They are **not** multiplicity adjusted P values.

2. Use the FDR approach you chose ([Prism offers three variants](#)[142]) to decide which P values are small enough to be deemed "discoveries". This calculation depends on which method you chose, and the value of Q you chose (the desired false discovery rate, as a percentage).

3. For each comparison, also compute a q value. The q values will be different for each comparison. If you had set Q to this value (what Prism reported as q) then this comparison would have been right on the border of being a "discovery" or not.

Notes:

- When you choose the FDR approach, Prism will not report anything about statistical significance, and will not (cannot) report confidence

intervals or multiplicity adjusted P values. But it <u>does report q values</u>[140], which are similar to adjusted P values.

- While the FDR approach is often used to deal with many P values such as those computed by Prism's <u>multiple t test analysis</u>[330], they are not commonly used as followup tests for ANOVA.

- The variable q is used as part of the results of the FDR approach to multiple comparisons, and as part of the Tukey and Dunnett multiple comparisons tests. The three tests define the variable q differently so they should not be compared.

5.11.3   How the various multiple comparisons methods work

.

5.11.3.1 The pooled standard deviation

When computing most multiple comparisons tests (not nonparametric or repeated measures one-way ANOVA), Prism assumes that all the data are sampled from populations with the same standard deviation, so the multiple comparison tests use a pooled standard deviation from all the groups.

Prism does not report this pooled standard deviation, but it is easy to calculate. Find the MSerror, also called MSresidual, and take its square root (MS stands for Mean Square, one of the columns in the ANOVA table).

For one-way ANOVA or two-way ANOVA with no repeated measures, there is only one MSerror (or MSresidual) in the ANOVA table. For two-way ANOVA with repeated measures, see <u>this document</u> for the details of which MS value is used.

The pooled standard deviation is expressed in the same units as the data.

5.11.3.2 The SE of the difference between means

ANOVA and repeated measures ANOVA

For most multiple comparisons tests, the first step is to compute the standard error of the difference between two mean using the equation below, where n1 and n2 are the sample sizes of the two means being compared and MSerror is the appropriate mean-square value from the ANOVA table.

$$SE_{\text{Difference}} = \sqrt{MS_{Error}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

For one-way ANOVA or two-way ANOVA with no repeated measures, there is only one MSerror (or MSresidual) in the ANOVA table. For two-way ANOVA with repeated measures, see [this document](#) for the details of which MS value is used.

The equation above can be simplified a bit by first computing the [pooled standard deviation](#)[375]:

$$SE \text{ of difference} = \text{Pooled SD}\sqrt{\frac{1}{n1} + \frac{1}{n2}}$$

Note that the MSerrror (and the pooled standard deviation) are computed from all the data in all the groups. The SE of the difference between means will the be same for all pairs of means if the samples sizes are equal.

[Multiple comparisons after repeated measures ANOVA.](#)[436]

Multiple comparisons after fitting a mixed effects model

With repeated measures data, Prism lets you do repeated measures ANOVA or [fit a mixed effects model](#)[391]. The two are equivalent when there are no missing values. When there are missing values, repeated measures ANOVA is impossible but it is still possible to fit a mixed effects model.

For multiple comparisons, the standard error of the difference is computed as sqrt(**L*C*L'**), where **L** is a vector/matrix of contrasts and **C** is a variance-covariance matrix at best-fit values.

5.11.3.3 How the Tukey and Dunnett methods work

To compute the Tukey or Dunnett test, divide the difference between the means you are comparing with [the standard error of the difference](#)[376] and call the quotient q.

$$q = \frac{|Mean_1 - Mean_2|}{SEDifference}$$

The critical values of q for Tukey multiple comparisons is explained in reference 2. C code can be found [here](#). The critical values of q for Dunnett's test are calculated according to methods explained in reference 3 and an appendix in reference 4.

It is critical in all cases to use the correct number of degrees of freedom, which is the df that corresponds with the mean square value used.

Prism uses the single step (1955) Dunnett test, not the newer (1990s) step up or step down methods.

References

1. Scott E. Maxwell, Harold D. Delaney, [Designing Experiments and Analyzing Data: A Model Comparison Perspective](#), Second Edition,IBSN:0805837183

2. Margaret Diponzio, Copenhaver, and Burt Holland. [Computation of the distribution of the maximum studentized range statistic with application to multiple significance testing of simple effects](#). Journal of Statistical Computation and Simulation, 1563-5163, Volume 30, Issue 1, 1988, Pages 1 – 15

3. K.S. Kwong and W. Liu. [Calculation of critical values for Dunnett and Tamhane's step-up multiple test procedure](#). Statistics and Probability Letters, Volume 49, Number 4, 1 October 2000 , pp. 411-416(6)

4. J. Hsu, [Multiple Comparisons, Theory and Methods](#). ISBN: 0412982811.

5.11.3.4 How the Fisher LSD method works

The Fisher's LSD test works just like a t test[276] with one exception. A t test computes a pooled standard deviation from the two groups being compared. The Fisher LSD test uses the pooled standard deviation[375]. This pooled SD comes from all the groups, which gives it more degrees of freedom and more power.

Note that the Fisher's LSD test does not account for multiple comparisons. It is rarely recommended.

More details about the Fisher's LSD test.[127]

5.11.3.5 How the Holm-Sidak method works

How to decide which P values are small enough to be deemed statistically significant

Here is a brief description of how the Holm multiple comparison test works:

1. If you are using the test as a followup to ANOVA, P values for each comparison are computed as they are for the Fisher's LSD test[378]. These are not corrected for multiple comparisons.

2. The P values are ranked from smallest to largest.

3. Set a value for the significance level, alpha. This is often set to 5%.

4. Define K equal to the number of comparisons you are making.

5. Start with the smallest P value and set i=K. Ask: Is the smallest P value less than alpha/i?

    If No: Conclude that none of the comparisons are statistically significant, and you are done.

    If Yes: Conclude that this comparison is statistically significant, and continue.

6. The second to smallest P value is compared next. Set i=K-1. Is the P value less than alpha/i?

If No: Conclude that this comparison (and all with larger P values) is not statistically significant. You are done.

If Yes: Conclude that this comparison is statistically significant, and continue.

7. The third to smallest P value is compared next. Set i=K-2. Compare the P value to alpha/i...

8. Continue until you find a comparison that is not statistically significant.

Prism actually uses the Šídák modification, so computes the Holm-Šídák test. At steps 5-7 above, the P value is not compared to to alpha/i but rather to 1-(1-alpha)$^{(1/i)}$

Multiplicity adjusted P values from the Holm-Sidak method

To compute the adjusted P value, called Padj(i), from the P value, called P(i), use the equations below where, the P values are sorted so P(1) is the smallest, k is the number of comparisons (number of P values) and max is a function that returns the larger of two values.

```
PAdj(1) = 1 - (1 - P(1))^k
PAdj(2) = max(PAdj(1), 1 - (1 - P(2))^(k-1))
..........
PAdj(j) = max(PAdj(j-1), 1 - (1 - P(j))^(k-j+1))
..........
PAdj(k) = max(PAdj(k-1), 1 - (1 - P(k))^(k-k+1)) = max(PAdj(k-1), P(k
```

Note that in some cases successive adjusted P values will be identical, even when the original P values are not.

5.11.3.6 How the Bonferroni and Sidak methods work

The calculations use a pooled SD

The first step for the Bonferroni and Sidak tests used as a followup to ANOVA is to compute the Fisher LSD test. Note two important points:

• The P values from this test are not corrected for multiple comparisons, so the correction for multiple comparisons is done as a second step.

- The P values are computed from difference between the two means being compared and the overall [pooled SD](#)[375]. When you compare columns A and B, the values in columns C, D, E, etc. affect the calculation of the pooled SD so affect the P value for the comparison of A and B. Using a pooled SD makes sense if all the values are sampled from populations with the same SD, as use of the pooled SD gives the Bonferroni or Sidak test more degrees of freedom, and therefore more power.

How the Šídák multiple comparison test works

The logic is simple(1). If you perform three independent comparisons (with the null hypothesis actually true for each one), and use the conventional significance threshold of 5% for each comparison without correcting for multiple comparisons, what is the chance that one or more of those tests will be declared to be statistically significant? The best way to approach that question, is to ask the opposite question -- what is the chance that all three comparisons will reach a conclusion that the differences are not statistically significant? The chance that each test will be not significant is 0.95, so the chance that all three independent comparisons will be not statistically significant is 0.95*0.95*0.95, which equals 0.8574. Now switch back to the original question. The chance that one or more of the comparisons will be statistically significant is 1.0000 - 0.8574, which is 0.1426.

You can also start with the significance threshold that you want to apply to the entire family of comparisons, and use the Šídák-Bonferroni method to compute the significance threshold that you must use for each individual comparison.

Call the significance threshold for the family of comparisons, the familywise alpha, alphaFW, and the number of comparisons K. The significance threshold to use for each individual comparisons, the per comparison alpha (alphaPC), is defined to be:

$$alphaPC = 1.0 \ - \ (1.0 - alphaFW)^{1/K}$$

If you are making three comparisons, and wish the significance threshold for the entire family to be 0.05, then the threshold for each comparison is:

$$alphaPC = 1.0 \ - \ (1.0 - alphaFW)^{1/K} = 1.0 - (1.0 - 0.05)^{1/3} = 0.0170$$

If you are making ten comparisons, and wish the significance threshold for the entire family of comparisons to be 0.05, then the threshold for each comparison is:

$$\text{alphaPC} = 1.0 \ - \ (1.0 - \text{alphaFW})^{1/K} = 1.0 - (1.0 - 0.05)^{0.10} = \ 0.0051$$

How the Bonferroni multiple comparison test works

The Bonferroni method uses a simpler equation to answer the same questions as the Šídák method. If you perform three independent comparisons (with the null hypothesis actually true for each one), and use the conventional significance threshold of 5% for each one without correcting for multiple comparisons, what is the chance that one or more of those tests will be declared to be statistically significant?

The Bonferroni method simply multiplies the individual significance threshold (0.05) by the number of comparisons (3), so the answer is 0.15. This is close, but not the same as the more accurate calculations above, which computed the answer to be 0.1426. (With many comparisons, the product of the significance threshold times the number of comparisons can exceed 1.0; in this case, the result is reported as 1.0.)

To use the Bonferroni method to compute the significance threshold to use for each comparison (alphaPC) from the number of comparisons and the significance threshold you wish to apply to the entire family of comparisons (alphaFW), use this simple equation:

$$\text{alphaPC} = \text{alphaFW}/K$$

Let's say you set the significance threshold for the entire family of comparisons to 0.05 and that you are making three comparisons. The threshold for determining significance for any particular comparison is reduced to 0.05/3, or 0.0167. Note that this is a bit more strict than the result computed above for the Šídák method, 0.0170.

If you are making ten comparisons, the Bonferroni threshold for each comparisons is 0.05/10 = 0.0050. Again this is a bit more strict (smaller) than the value computed by the Šídák method above, which is 0.0051.

5.11.3.7 How the Dunn method for nonparametric comparisons works

Two forms of the Dunn's test

On Prism's option tab of the parameters dialog, you can choose two different form of Dunn's test.

Multiple comparisons test
- (●) Correct for multiple comparisons using statistical hypothesis testing. Recommended.
   Test: Dunn's
- (○) Correct for multiple comparisons by controlling the False Discovery Rate.
   Test: Two-stage step-up method of Benjamini, Krieger and Yekutieli (recommended)
- (○) Don't correct for multiple comparisons. Each comparison stands alone.
   Test: Uncorrected Dunn's test

Prism performs the Dunn's multiple comparison test(1), which is standard. One source is *Daniel's Applied nonparametric statistics*, second edition page 240-241. Some books and programs don't use Dunn's name, but simply refer to this test as the post test following a Kruskal-Wallis test, and don't give it an exact name.

Step 1. Calculate z

For ordinary (not matched, not repeated measures) nonparametric ANOVA: To compare group i and j, find the absolute value of the difference between the mean rank of group i and the mean rank of group j. If there are no ties, compute z by dividing this difference in mean ranks by the square root of $[(N*(N+1)/12)*(1/N_i + 1/N_j)]$. Here N is the total number of data points in all groups, and N_i and N_j are the number of data points in the two groups being compared. If there are ties, calculate z by dividing the difference in mean ranks by the square root of $[(N*(N+1) - Sum(T_i^3 - T_i) / (N - 1)) / 12 * (1/N_i + 1/N_j)$, where T_i is the number of ties in the i-th group of ties.

For repeated measures nonparametric ANOVA (Friedman's test): To compare treatment i and j, find the absolute value of the

difference between the mean rank of group i and the mean rank of group j. Calculate z by dividing this difference in mean ranks by the square root of [K(K+1)/(6N)]. Here N is the number of matched sets of data, which is the number of rows in the data table, and K is the number of treatment groups (number of columns).

Step 2. Calculate the uncorrected P value

Calculate the two-tail (two-sided) P value that corresponds to the z ratio you just calculated. This free web calculator can do the job.

If you chose the uncorrected Dunn's test, you are done. Compare the P value to alpha to decide whether or not the result is statistically significant.

Step 3. If you want to correct for multiple comparisons,, compute the multiplicity adjusted P value

Multiply the uncorrected P value computed in step 2 by K. If this product is less than 1.0, it is the multiplicity adjusted P value. If the product is greater than 1.0 the multiplicity adjusted P value is reported as > 0.9999

Compare the multiplicity adjusted P value to alpha to decide whether or not the result is statistically significant.

Notes

- When comparing two groups as part of Dunn's test, the ranks are those for all the values. Rank all the values, compute the mean rank in each group, and do the calculations above. Don't create ranks just for the values in those two groups.

- This method accounts for ties when computing the ranks, and thus when computing the mean ranks which are compared.

- It would seem sensible to base multiple comparisons tests after nonparametric ANOVA to be based on the Mann-Whitney or Wilcoxon method to compute P values, followed by Bonferroni or other method to correct for multiple comparisons. Prism does not offer this approach, because it is not commonly used (but we'd appreciate comments and suggestions).

Reference

1. O.J. Dunn, Technometrics, 5:241-252, 1964

5.11.3.8 How the Dunnett T3, Games and Howell, and Tamhane T2 tests work

Starting with Prism 8, one-way ANOVA allows you to specify that you don't wish to assume that all the groups are sampled from populations with the same SD (and thus the same variances). When you make this choice, Prism performs both Brown-Forsythe ANOVA and Welch ANOVA. When you make this choice, Prism offers a different set of multiple comparisons tests.

Hypothesis testing approach

If you choose the statistical hypothesis testing approach, Prism offers three tests: Dunnett T3, Games and Howell, and Tamhane T2. We recommend Dunnett T3 when sample sizes are small (<50 per group) and Games and Howell when samples are larger.

A multiple comparisons procedure starts by calculating the ratio of the difference between a pair of means divided by the standard error of that difference.

When you assume that all variances are equal (as is usual with ANOVA), the multiple comparisons procedures pool the SDs from all the groups, and uses the combined sample size of all groups when computing the degrees of freedom. This computed from the Mean Square Error (or Residual) in the ANOVA table. This gives a more precise measure of population scatter, which gives the procedures a bit more power.

If you don't assume that all variances are equal, then when comparing two means the multiple comparison procedure uses only the SD and sample sizes of those two groups. The ratio is the t ratio, just as if you were only comparing those two groups. The degrees of freedom are computed from a complicated equation that accounts for unequal sample size and unequal SDs. The P values are computed from t and df, accounting for multiple comparisons.

The three tests compute the same t ratio and the same df values. They differ in how they compute the P value from t and df.

The details are explained in a good review by Shingala(1). However, Shingala mistakenly says that the Dunnett's T3 test is designed (like the test Dunnett is best know for) for comparing each mean to a control mean. Not so. It is designed to compare all pairs of means(2).Tamhane (3) has recommended some modifications of these tests in some situations, but these modfifcations are not standard so Prism uses the original versions.

We recommend Dunnett T3 when sample sizes are small (<50 per group) and Games and Howell when samples are larger.

FDR approach

If you use one of the methods based on controlling the FDR, Prism first computes individual P values using the [Welch-correct t test][282], and uses those P values in the FDR method. It does not pool the standard deviations as is done with the Fisher LSD test, because it only makes sense to pool variances when you assume the population variances are equal.

Don't correct for multiple comparisons

If you choose to not correct for multiple comparisons, Prism computes the P values using a [t test with the Welch correction][282]. It does not pool the standard deviations as is done with the Fisher LSD test, because it only makes sense to pool variances when you assume the population variances are equal.

References

1. Shingala, M.C., and Rajyaguru, A. (2015). [Comparison of post hoc tests for unequal variance]. International Journal of New Technologies in Science and Engineering

2. Dunnett, C.W. (1980). [Pairwise Multiple Comparisons in the Unequal Variance Case]. Journal of the American Statistical Association 75: 796–800.

3. Tamhane, A.C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. Journal of the American Statistical Association 74: 471–480.

5.11.3.9 How the methods used to control the FDR work

Prism offers a choice of three algorithms for controlling the FDR. The three algorithms all work in a similar manner.

1. Calculate a P value for each comparison.

2. Rank the P values from low to high.

3. Start with the largest P value.

4. Compute a threshold for the largest P value. That threshold value depends on the number of P values being looked at. For the method of   Benjamini, Krieger and Yekutieli, the threshold also depends on the estimate of the number of true null hypotheses provided by the method.

5. If the P value is less than the threshold, then all P values are flagged as discoveries, and you are done. Otherwise, continue.

6. Go to the second largest P value.

7. Compute a threshold for the second highest P value. This threshold will be smaller than the threshold for the largest P value. Computing the threshold (see below) depends on the rank of the P value, the number of P values and (for the method of   Benjamini, Krieger and Yekutieli) the estimate of the number of true null hypotheses (computed by the method; nothing for you to think about).

8. If the P value is less than the threshold, that P value and all smaller ones are flagged as discoveries and you are done. Otherwise continue.

9. Go to the next P lower value.

10. Compute a threshold for this rank. It will be smaller than the previous threshold.

11. If the P value is less than the threshold, that P value and all smaller ones are flagged as discoveries and you are done. Otherwise repeat steps 9-10 until done.

The difference between the three methods is how they compute the threshold values. The table below gives the details, where Q is the desired false discovery rate (as a percentage), N is the number of P values in the set, and Ntrue is the number of the null hypotheses estimated to be true (part of the second method below). Define q to equal Q/100. This converts the value you enter as a percentage into a fraction.

Method	Threshold for smallest P value	Threshold for largest P value
Original method of Benjamini and Hochberg (1)	q/n	q
Two-stage step-up method of Benjamini, Krieger and Yekutieli (2)	$q/[(1+q)Ntrue]$	$[q/(1+q)] * (N / Ntrue)$
Corrected method of Benjamini & Yekutieli (3)	$q/[N* (1 + 1/2 + 1/3 + ... + 1/N)]$	$q /(1 + 1/2 + 1/3 + ... + 1/N)$

Notes:

- The variable q is defined to be Q/100, where Q is the desired false discovery rate (as a percentage) you enter.

- The thresholds for the P values between the smallest and largest are determined by a linear interpolation between those extremes.

- The threshold is computed as a fraction (not a percentage), to compare to a P value.

Here is a graph showing the thresholds for analyzing 20 P values (N=20), you set Q=5% and Ntrue=12 (computed by the BKY method from the data, and only applicable for the red line). You can see that the two-stage linear step-up method method of Benjamini, Krieger and Yekuteili (red) has largest thresholds so has the most power, and the corrected method of Benjamini & Yekutieli (green) has the least power. You can also see that the methods diverge the most when computing the threshold for the largest P values and nearly converge for smaller P values.
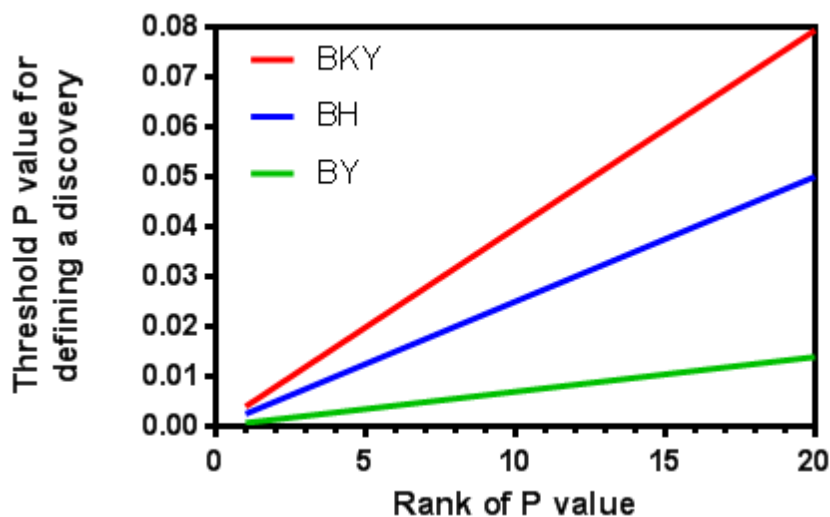
References

1. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289–300 (1995).

2. Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93, 491–507 (2006). We use the method defined in section 6 of this paper, the two-stage linear step-up procedure.

3. Benjamini, Y., & Yekutieli, D. (2001). [The control of the false discovery rate in multiple testing under dependency](#). Annals of statistics, 1165–1188.

5.11.3.10 Mathematical details

All calculations use standard methods detailed in Maxwell and Delaney (1). The details of how Prism does its calculations for controlling the Type I error are in [this eight-page document](#).

1. Scott E. Maxwell, Harold D. Delaney, [Designing Experiments and Analyzing Data: A Model Comparison Perspective](#), Second Edition, IBSN:0805837183

5.12   Repeated measures ANOVA and mixed model analysis

Repeated measures data can be analyzed by repeated meausres ANOVA or by fitting a mixed effects model.

5.12.1   What is repeated measures experimental design?

The advantage of repeated measures

The difference between ordinary and repeated measures ANOVA, is similar to the difference between unpaired and paired t tests. See the [advantages of pairing or matching](#)[270]. Since each participant or experiment acts as its own control, repeated measures design can do a better job of separating signal from noise, so this design usually has more power. Some participants may have larger measurements at all time points, and others may have lower measurements at all time points. Repeated measures ANOVA focuses on how much the Y value changes between treatments.

Repeated measures or randomized block?

The term *repeated measures* is used when you give treatments repeatedly to each animal or participant.

The term *randomized block* is used when you randomly assign treatments within each cluster (block) of matched subjects.

Imagine that you compare three different treatments. In a repeated measures design, you'd recruit say 10 participants (or use ten animals) and measure each of the participants (animals) after each of the treatments. With a randomized block design, you'd recruit ten clusters of four partipants each, matched for age, gender, postal code, etc. (or ten sets of four animals, with the four treated at the same time in adjacent cages...). Another example would be a laboratory experiment is run several times, each time with several treatments (or a control and several treatments) handled in parallel.

Any time the participants/animals/experiments are clustered in some way that can make their responses similar, other than the treatments you're comparing, you need to treat those clusters as randomized blocks. Examples include animals from the same litter, blood samples from the same draw, or mice tested in batches on the same day.

ANOVA works identically for repeated-measures and randomized block experiments, and Prism always uses the term *repeated-measures.*

Counting factors

One-, two- or three-way ANOVA? Two possible points of confusion:

- A design with three or more measurements of different treatments on the same participant is called repeated measures one-way ANOVA in Prism, because there really is only one factor, treatment as denoted by the data set columns. But you could argue there is a second factor too, subject, because each row represents a different subject (or block). In fact, you'll get the same results if you analyze with two-way ANOVA (without replicates) and one-way repeated measures ANOVA.

- Say you give some participants one treatment, others a different treatment, and others a third treatment. You then measure an outcome for each participant at three time points (before, during and after the. There are two factors: treatment and time. So you need two-way ANOVA, with repeated measures in one factor. Some people mistakenly ignore time and think this is a one-way ANOVA problem, forgetting that time is a factor.

5.12.2 Missing values in repeated measures ANOVA

What happens if a value is missing? With repeated measures ANOVA, you can't use any of the data for that subject. To do repeated measures ANOVA, you'd need to remove the data for that participant/animal/whatever entirely from the data table before running the ANOVA.

Beginning with Prism 8, Prism offers an alternative method to analyze repeated measures data: <u>fitting a mixed effects model</u>[391]. This analysis works fine even when there are some missing values. This mixed model choice is offered in the ANOVA parameters dialogs.

The results of fitting a mixed model with missing values will be meaningful, of course, only if the values are missing for random reasons. The results will probably be misleading if the values are missing because those participants were very sick, or those values were too high to measure (or too low to measure). If the data truly are repeated measures over time and all missing values are at the last time point or last few times points, they are unlikely to be missing at random, but rather missing because something happened to those participants (or animals...) over the course of the study.

5.12.3 The mixed model approach to analyzing repeated measures data

The problem: Repeated measures ANOVA cannot handle missing values

Repeated measures ANOVA calculations require complete data. If a value is missing for one partiicpant or animal, you'd need to ignore all data for that participant or animal. The only way to overcome this (using ANOVA) would be to impute what the values of the missing values probably were and then analyze without any missing values, correcting the results (reducing df) to account for the imputing. This is not a preferred method, and is not offered by Prism.

Prism offers fitting a mixed effects model to analyze repeated measures data with missing values.

Fitting a mixed effects model - the big picture

The mixed effects model approach is very general and can be used (in general, not in Prism) to analyze a wide variety of experimental designs. Many books have been written on the mixed effects model. Because of

this versatility, the mixed effects model approach (in general) is not for beginners.

Prism 8 fits the mixed effects model for repeated measures data. Prism uses a mixed effects model approach that gives the same results as repeated measures ANOVA if there are no missing values, and comparable results when there are missing values.

Prism uses the mixed effects model in only this one context. You don't have to, or get to, define a covariance matrix. You can't add a covariate. You can't compare alternative mixed effects models. You can't do mixed effects model regression.

When fitting a mixed effects model in Prism, think of it as repeated measures ANOVA that allows missing values.

Mixed? Fixed vs. random factors

Statistical calculations can deal with two kinds of factors.

- A factor is fixed when you wish to test for variation among the means of the particular groups from which you have collected data.

- A factor is random when you have randomly selected groups from an infinite (or at least large) number of possible groups, and that you want to reach conclusions about differences among all the groups, not just the groups from which you collected data.

ANOVA works by partitioning the total variation among values into different components. With repeated measures ANOVA, one of those components is variation among participants or blocks. In Prism, ANOVA treats all factors, including participant or block, as fixed factors.

As the name suggests, the mixed effects model approach fits a model to the data. The model is *mixed* because there are both fixed and random factors. When Prism does mixed-model analysis of repeated measures data, it assumes that the main factors (defined by the data set columns in one-way, and by data set columns and rows in two- and three-way) are fixed, but that subjects (or participants, or runs..) are random. You are not interested in variation among those particular participants, but want to know about variation among participants in general.

Results of the two methods

The results of repeated measures ANOVA and fitting a mixed effects model look quite different. Here are examples of the one-way repeated measures data (with no missing values) analyzed both ways.

Repeated measures ANOVA results

	RM one-way ANOVA					
1	Table Analyzed	Repeated measures one-way ANOVA data				
2						
3	Repeated measures ANOVA summary					
4	Assume sphericity?	No				
5	F	3.618				
6	P value	0.0873				
7	P value summary	ns				
8	Statistically significant (P < 0.05)?	No				
9	Geisser-Greenhouse's epsilon	0.5793				
10	R square	0.4750				
11						
12	Was the matching effective?					
13	F	10.45				
14	P value	0.0007				
15	P value summary	***				
16	Is there significant matching (P < 0.05)?	Yes				
17	R square	0.6464				
18						
19	ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
20	Treatment (between columns)	2175	3	725.0	F (1.738, 6.951) = 3.618	P=0.0873
21	Individual (between rows)	8372	4	2093	F (4, 12) = 10.45	P=0.0007
22	Residual (random)	2404	12	200.4		
23	Total	12951	19			
24						
25	Data summary					
26	Number of treatments (columns)	4				
27	Number of subjects (rows)	5				
28	Number of missing values	0				
29						

Mixed effects model results

	Mixed-effects analysis					
1	Table Analyzed	Repeated measures one-way ANOVA data				
2	Method	Mixed-effects model (Restricted max. likelihood)				
3	Assume sphericity?	No				
4						
5	Fixed effect	P value		P value summary	Statistically significant (P < 0.05)?	F (DFn, DFd)	Geisser-Greenhouse's epsilon
6	Treatment (between columns)	0.0873		ns	No	F (1.738, 6.951) = 3.618	0.5793
7						
8	Random effects	SD		Variance		
9	Individual (between rows)	21.75		473.1		
10	Residual	14.15		200.4		
11						
12	Was the matching effective?					
13	Chi-square, df	10.01, 1				
14	P value	0.0016				
15	P value summary	**				
16	Is there significant matching (P < 0.05)?	Yes				
17						
18	Data summary					
19	Number of treatments (columns)	4				
20	Number of subjects (rows)	5				
21	Number of missing values	0				
22						

Main results are the same

The main result is the P value that tests the null hypothesis that all the treatment groups have identical population means. That P value is 0.0873 by both methods (row 6 and repeated in row 20 for ANOVA; row 6 for mixed effects model). For these data, the differences between treatments are not statistically significant.

Multiple comparisons results are the same

Multiple comparisons following repeated measures ANOVA are computed from the pooled standard deviation, which is the square root of the mean square residuals.

Multiple comparisons following fitting a mixed effects model is much more complicated, based on matrix algebra. With no missing data, the two are equivalent.

Some results differ

One part of the results differ, the part that tests whether there was indeed variation among subjects. ANOVA tests this by having variation among subjects one of the variation components, and tests for its contribution with a F ratio and P value, which is 0.0007 (line 21 above). The mixed effects model compares the fit of a model where subjects are a random factor vs. a model that ignores difference between subjects. This results in a chi-square ratio and P value, which is 0.0016 (line 14 above). Because ANOVA assumes subjects are a fixed factor (you care about

those specific subjects) and the mixed effects model treats subjects as a random factor (you care about subjects in general), the two P values are usually not the same.

5.12.4 Sphericity and compound symmetry

Overview

One of the assumptions of repeated measures ANOVA is called *sphericity* or *circularity* (the two are synonyms). Prism lets you decide whether to accept this assumption. If you choose not to accept this assumption, Prism uses the method of Geisser and Greenhouse to correct for violations of the assumption.

Should you assume sphericity?

Sphericity is defined below, but here are some guidelines for answering Prism's question about whether to assume sphericity:

- If your experimental design relies on <u>matching rather than repeated measurements</u>[435], then you can assume sphericity, as violations are unlikely.

- If your experiment design is repeated measures (multiple measurements over time), we recommend that you do not assume sphericity. We follow the recommendation of Maxwell and Delaney(1).

Defining sphericity

The name is confusing. Don't try to intuit what the term *sphericity* means by thinking about *spheres*. Mathematical statistics books define the term in terms of matrix algebra. That makes it seem confusing. But, in fact, the concept is pretty easy to understand.

Here is the table of sample data from Prism (choose a Column table, and then choose sample data for repeated measures one-way ANOVA).

		A	B	C	D
		Control	Treatment 1	Treatment 2	Treatment 3
	⊗	Y	Y	Y	Y
1	GS	54	43	78	111
2	JM	23	34	65	99
3	HM	45	65	99	78
4	DR	54	77	79	90
5	PS	45	46	87	95

Each row represents data from one subject identified by the row title. Each column represents a different treatment. In this example, each of five subjects was given four sequential treatments. The data will be analyzed by repeated measures one-way ANOVA.

The assumption of sphericity states that the variance of the differences between treatment A and B equals the variance of the difference between A and C, which equals the variance of the differences between A and D, which equals the variance of the differences between B and D... Like all statistical assumptions, this assumption pertains to the populations from which the data were sampled, and not just to these particular data sets.

This is easier to see on a graph:
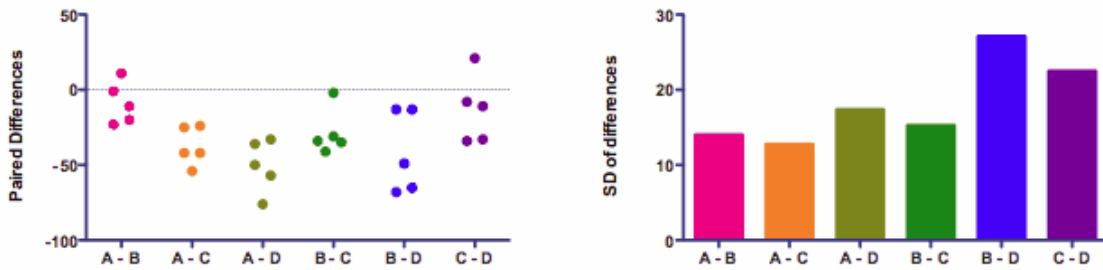
The left panel shows the differences. Each of the six columns represents the difference between two treatments. There are five subjects, so there are five dots for each difference.

The graph on the right shows the standard deviations. The assumption of sphericity states that the data were sampled from populations where these standard deviations are identical. (Most statistics books talk about variance, which is the square of the standard deviation. If the standard deviations are equal, so are the variances.) The standard deviations in the right panel above are not identical. That doesn't really matter. The assumption is about the population of values from which the data were sampled. In any particular samples, you expect some variation. Here the variation among the standard deviations is fairly small.

You might be surprised that the differences between nonadjacent columns are considered. Why should the difference between A and C matter? Or between A and D? The answer is that ANOVA, even repeated measures ANOVA, pays no attention to the order of the groups. Repeated measures ANOVA treats each row of values as a set of matched values. But the order of the treatments is simply not considered. If you randomly scrambled the treatment order of all subjects, the ANOVA results wouldn't change a bit (unless you choose a post test for trend).

References 2 and 3 below are clear, nonmathematical explanations of sphericity.

Compound symmetry

When you read about this topic, you will also encounter the term *compound symmetry*, which is based on the covariance matrix of the raw data (without computing paired differences). If the assumption of compound symmetry is valid for a data set, then so is the assumption of sphericity. But the reverse is not always true. It is possible, but rare, for

data to violate compound symmetry even when the assumption of sphericity is valid.

What happens when the sphericity assumption is violated?

The assumption of sphericity would be violated when the repeated measurements are made in too short a time interval, so that random factors that cause a particular value to be high (or low) don't wash away or dissipate before the next measurement. To avoid violating the assumption, wait long enough between treatments so the subject is essentially the same as before the treatment. When possible, also randomize the order of treatments.

If the assumption of sphericity is violated, and you don't account for this in the calculations, then the P value reported by repeated measures ANOVA will be too small. In other words, the Geisser-Greenhouse correction increases the P value.

Quantifying deviations from sphericity

Prism quantifies deviation from sphericity by calculating and reporting the value of [epsilon](#)[400].

It seems like Prism should be able to decide whether to correct for violations of sphericity based on the value of epsilon. However, using this value to decide how to analyze the data is not recommended(1).

Repeated measures ANOVA without assuming sphericity

Prism can use the method of Greenhouse and Geisser to adjust the results of the repeated measures ANOVA to account for the value of epsilon. The only thing this adjustment does is reduce the number of degrees of freedom, which increases the P value.

Notes:

• This method is sometimes attributed to Box.

• Geisser and Greenhouse also derived a *lower-bound correction*. This is a simpler method to calculate, but corrects too far. Prism does not use this method, but instead uses the *Geisser and Greenhouse epsilon hat* method.

- Huynh and Feldt have developed an alternative method to perform repeated measures ANOVA without assuming sphericity. Prism does not compute this method, as Maxwell and Delaney prefer (slightly) the Geisser and Greenhouse method (1).

- The correction works by decreasing the values of the degrees of freedom. These revised values can be fractional, and Prism computes P from the F ratio and these revised fractional degrees of freedom.

- The correction works the same way if you fit data by fitting the mixed-model. The model is fit the same as would have been if you hadn't asked for the correction. The DF are decreased, so the P values computed from the F ratios are larger.

When looking at a printed page of Prism results, how can you tell if sphericity was assumed?

If sphericity was not assumed, you'll see that Prism reports a value for Geisser-Greenhouse epsilon, and that fractional df values are used to compute a P value.

Repeated measures ANOVA summary						
Assume sphericity?	No					
F	15.76					
P value	0.003070					
P value summary	**					
Statistically significant (P < 0.05)?	Yes					
Geisser-Greenhouse's epsilon	0.5773					
R square	0.7976					
Was the matching effective?						
F	1.333					
P value	0.3134					
P value summary	ns					
Is there significant matching (P < 0.05)?	No					
R square	0.08253					
ANOVA table	SS	DF	MS	F (DFn, DFd)	P value	
Treatment (between columns)	8417	3	2806	F (1.732, 6.927) = 15.76	P = 0.003070	
Individual (between rows)	949.3	4	237.3	F (4, 12) = 1.333	P = 0.3134	
Residual (random)	2136	12	178.0			
Total	11503	19				

Reference

1. Scott E. Maxwell, Harold D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective,* Second Edition. IBSN:0805837183.

2. Andy Field, <u>A Bluffer's Guide to ... Sphericity</u>.

3. T. Baguley, <u>What is all this stuff about sphericity in my repeated measures ANOVA Output?</u>

5.12.5 Quantifying violations of sphericity with epsilon

Deviations from <u>sphericity</u>[395] in repeated measures ANOVA can be quantified by a value known as *epsilon*. There are two methods for calculating it. Based on a recommendation from Maxwell and Delaney (p 545, reference below), Prism uses the method of Greenhouse and Geisser. While this method might be a bit conservative and underestimate deviations from the ideal, the alternative method by Huynh and Feldt tends to go too far in the other direction.

If you choose not to assume sphericity in repeated measures ANOVA, Prism reports the value of epsilon. Its value can never be higher than 1.0, which denotes no violation of sphericity. The value of epsilon gets smaller with more violation of sphericity, but its value can never be lower than 1/(k - 1), where k is the number of treatment groups.

Number of treatments, k	Possible values of epsilon
3	0.5000 to 1.0000
4	0.3333 to 1.0000
5	0.2500 to 1.0000
6	0.2000 to 1.0000
7	0.1667 to 1.0000
8	0.1429 to 1.0000
9	0.1250 to 1.0000
10	0.1111 to 1.0000
11	0.1000 to 1.0000
12	0.0909 to 1.0000
13	0.0833 to 1.0000
14	0.0769 to 1.0000
15	0.0714 to 1.0000
20	0.0526 to 1.0000
25	0.0417 to 1.0000
50	0.0204 to 1.0000
k	$1/(k-1)$ to 1.0000

Reference

Scott E. Maxwell, Harold D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective,* Second Edition. IBSN:0805837183.

5.12.6 Analysis details for fitting the mixed model

Details for statistical consultants

We provide details to statistical consultants in two ways.

R and SAS code that does what Prism does

We have written FAQs that contain a Prism file along with corresponding R and SAS code for one-way repeated measures anova, and for two-way repeated measures ANOVA with repeated measures in both factors.

Math overview

Prism fits data to a mixed effects model $y = X\beta + Z\gamma + \varepsilon$ using a compound symmetry covariance matrix **V**, where **y** denotes the vector of responses, **X** is the known regressor matrix of fixed effects, β is the unknown fixed-effects parameter vector, **Z** is the known design matrix of random effects and γ is the vector of unknown random-effects parameters.

γ and ε are normally distributed with:

$E(\gamma) = E(\varepsilon) = 0$,

$Var(\gamma) = G$, $Var(\varepsilon) = R$, $Cov(\gamma, \varepsilon) = 0$,

where

$R = \sigma^2 * I$ (**I** is the identity matrix)

G is a diagonal matrix with variances of random effects on the main diagonal.

Therefore $V = R + Z*G*Z'$

The fitting is done with REstricted (REsidual) Maximum Likelihood (REML) that provides unbiased estimates of unknown variances σ^2 and **G**.

For multiple comparisons, the standard error of the difference is computed as sqrt(**L*C*L'**), where **L** is a vector/matrix of contrasts and **C** is a covariance matrix of $(\hat{\beta} - \beta, \hat{\gamma} - \gamma)$

$$C = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^{-}$$

where $^{-}$ denotes a generalized inverse.

The fitting of the mixed model is the same whether or not you selected the Geisser-Greenhouse correction. The Geisser-Greenhouse correction only changes the degrees of freedom used when calculating P values from F ratios. So all results except the P value will be the same with and without the correction.

5.12.7   How to report the methods used for the mixed model analysis

How scientists can think about Prism's mixed model analysis

Most scientists can think of the mixed model in Prism as repeated measures ANOVA that allows missing values.

What to say in Methods sections or Figure legends

Here is some suggested wording for explaining how repeated measures ANOVA with Prism are done (using the Mixed-model approach):

> Repeated measures ANOVA cannot handle missing values. We analyzed the data instead by fitting a mixed model as implemented in GraphPad Prism 8.0. This mixed model uses a compound symmetry covariance matrix, and is fit using Restricted Maximum Likelihood (REML). In the absence of missing values, this method gives the same P values and multiple comparisons tests as repeated measures ANOVA. In the presence of missing values (missing completely at random), the results can be interpreted like repeated measures ANOVA.

Also state whether or not you asked Prism to use the Geisser-Greenhouse correction.

5.12.8   "Not positive definite" ?

In some cases, Prism will say the following in a yellow floating note:

> "Some results are missing because estimated variance-covariance matrix is not positive definite."

Prism will still show some results, but probably not the ones that are most important to you!

The message won't mean much to most scientists. Neither will this alternative wording:

> "The Hessian matrix is non-invertible. It is not possible to take the square root of the matrix. "

Basically, this message means that Prism was not able to complete the analysis given your data and choices. Given your data, the model is too complicated.

Possible fixes or workarounds:

- It might help to go back to the RM Analysis tab of the parameter dialog and tell Prism to remove terms from the model when the variance would be zero or negative.

- If there are no missing values, use repeated measures ANOVA rather than fitting the mixed model.

- If both factors are repeated measures, try a model where only one factor is specified as repeated measures. If only one factor is repeated measures, consider ordinary ANOVA. The whole point of a repeated measures design is to have internal controls so variation among participants/animals/plates... is factored out of the analysis. But if there isn't much variation, the repeated measures analysis may not be needed. Graph your data and see how much the participants/animals/plates... differ from one another.

5.13 One-way ANOVA, Kruskal-Wallis and Friedman tests

You've measured a variable in three or more groups, and the means (and medians) are distinct. Is that due to chance? Or does it tell you the groups are really different? Which groups are different from which other groups?

5.13.1 How to: One-way ANOVA and mixed model

.

5.13.1.1 Entering data for one-way ANOVA and related tests

This page explains how to use Prism perform one-way ANOVA, repeated measures one-way ANOVA, the Kruskal-Wallis and Friedman tests.

Setting up the data table without repeated measures (completely randomized design)

From the Welcome (or New Table and graph) dialog, the Column tab.

If you aren't ready to enter your own data, choose one of the sample data sets.

If you want to enter data, note that there are two choices. You can enter raw data or summary data (as mean, SD or SEM, and n).

Enter replicate values stacked into columns

Enter the data for each group into a separate column. The two groups do not have be the same size (it's OK to leave some cells empty). If the data are unpaired, it won't make sense to enter any row titles.

A	B	C
Control	Treated	Treated+Antagonist
Y	Y	Y
54	87	45
23	98	39
45	64	51
54	77	49
45	89	50
47		55

Enter and plot error values calculated elsewhere

Prism can compute one-way ANOVA (but not repeated measures ANOVA, and not nonparametric comparisons) with data entered as mean, SD (or SEM), and n. This can be useful if you are entering data from another program or publication.

Create a Grouped table, and enter the data all on one row.

Table format: Two-way		A Control			B Low dose			C High dose		
	x	Mean	SD	N	Mean	SD	N	Mean	SD	N
1	Title	34.5	11.3	12	46.5	7.3	14	75.3	14.1	13

Setting up the data table with repeated measures design

From the Welcome (or New Table and graph) dialog, the Column tab.
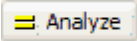
If you aren't ready to enter your own data, choose one of the tutorial data sets.

With repeated measures data, each row represents a different subject or experiment. Identify each row (optionally) using row titles.

Table format: One-way		A Control	B Treatment 1	C Treatment 2	D Treatment 3
	x	Y	Y	Y	Y
1	GS	54	43	78	111
2	JM	23	34	42	65
3	HM	45	65	99	105
4	JW	54	77	79	90
5	PS	45	46	75	86

Starting with Prism 8, it is ok to leave one or a few values blank (missing). The results will only be meaningful if the values are missing randomly. The results would not be meaningful if a value is missing because it would have been too large (or too small) to measure.

Run the ANOVA

1. From the data table, click ⇌ Analyze on the toolbar.

2. Choose one-way ANOVA from the list of column analyses.

3. [Choose the test](#)[406] you want to perform on the first tab.

4. Choose the multiple comparisons tests on the [Multiple Comparisons](#)[412] and [Options](#)[414] tabs of the one-way ANOVA dialog.

5.13.1.2 Experimental design tab: One-way ANOVA

Prism offers four related tests that compare three or more groups. Your choice of a test depends on these choices:
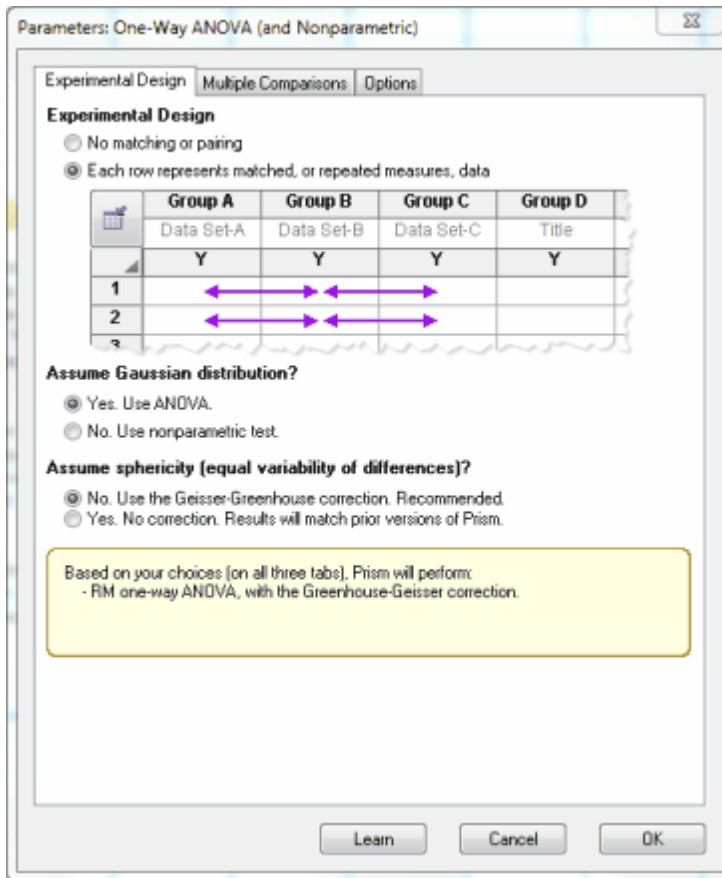
Experimental Design

Choose a repeated measures test when the columns of data are matched. Here are some examples:

- You measure a variable in each subject several times, perhaps before, during and after an intervention.

- You recruit subjects as matched groups, matched for variables such as age, ethnic group, and disease severity.

- You run a laboratory experiment several times, each time with several treatments handled in parallel. Since you anticipate experiment-to-experiment variability, you want to analyze the data in such a way that each experiment is treated as a matched set.

Matching should not be based on the variable you are comparing. If you are comparing blood pressures in three groups, it is OK to match based on age or zip code, but it is not OK to match based on blood pressure.

The term *repeated measures* applies strictly when you give treatments repeatedly to one subject (the first example above). The other two examples are called *randomized block experiments* (each set of subjects is called a block, and you randomly assign treatments within each block). The analyses are identical for repeated measures and randomized block experiments, and Prism always uses the term *repeated measures.*

Choose "no matching" if you have a completely randomized design.

Assume Gaussian distribution?

Nonparametric tests[148], unlike ANOVA are not based on the assumption that the data are sampled from a Gaussian distribution[37]. But nonparametric tests have less power[149], and report only P values but not confidence intervals. Deciding when to use a nonparametric test is not straightforward[152].

If no matching: Assume homoscedasticity?

One assumption underlying the usual ANOVA F test is homogeneity of variance. That means that each group is sampled from populations with the same variance (and thus the same standard deviation) even if the means differ.

Starting with Prism 8, you choose whether or not to assume equal population variances. If you choose not to make that assumption, Prism performs two alternative forms of ANOVA and reports both results. Both **Welch's ANOVA** and **Brown-Forsythe ANOVA** adjusts the calculations of the F ratio and degrees of freedom to adjust for heterogeneity of within-group variances. The P value can be interpreted in the same manner as in the analysis of variance table.

Notes:

- Why use these special forms of ANOVA rather than use a nonparametric Kruskal-Wallis test? Because while the Kruskal-Wallis test does not assume that the data are sampled from Gaussian distributions, it does assume that the dispersion or spread of the distributions are the same.

- As an alternative to these tests, consider transforming your data (logarithms, reciprocals, etc.) and analyzing the transformed values with ordinary ANOVA.

- This Brown-Forsythe test to compare means is distinct from another test also named Brown-Forstyhe that compares variances.

If repeated measures: Assume sphericity?

The concept of sphericity

The [concept of sphericity](#)[395] is tricky to understand. Briefly it means that you waited long enough between treatments for any treatment effect to wash away.This concept is not relevant if your data are not repeated measures, or if you choose a nonparametric test.

For each subject subtract the value in column B from the value in column A, and compute the standard deviation of this list of differences. Now do the same thing for the difference between column A and C, between B and C, etc. If the assumption of sphericity is true, all these standard deviations should have similar values, with any differences being due to chance. If there are large, systematic differences between these standard deviations, the assumption of sphericity is not valid.

How to decide whether to assume sphericity

If each row of data represents a set of matched observations, then there is no reason to doubt the assumption of sphericity. This is sometimes called a randomized block experimental design.

If each row of data represents a single subject given successive treatments, then you have a repeated measures experimental design. The assumption of sphericity is unlikely to be an issue if the order of treatments is randomized for each subject, so one subject gets treatments A then B then C, while another gets B, then A, then C... But if all subjects are given the treatments in the same order, it is better to not assume sphericity.

If you aren't sure, we recommend that you do not assume sphericity.

How your choice affects Prism's calculations

 If you choose to not assume sphericity, Prism will:

- Include the Geisser-Greenhouse correction when computing the repeated measures ANOVA P value. The resulting P value will be higher than it would have been without that correction.

- Quantify violations of sphericity by reporting [epsilon](#)[400].

- [Compute multiple comparisons tests differently.](#)[436]

If you ask Prism to assume sphericity, but in fact that assumption is violated, the P value from ANOVA will be too low. For that reason, if you are unsure whether or not to assume sphericity, we recommend that you check the option to *not* assume sphericity.

Test summary

Test	Matched	Nonparametric	Assume equal variances?
[Ordinary one-way ANOVA](#)[405]	No	No	Yes
[Welch's ANOVA and Brown-Forsythe ANOVA](#)[431]	No	No	No
[Repeated measures one-way ANOVA](#)[435]	Yes	No	NA
[Kruskal-Wallis test](#)[448]	No	Yes	NA
[Friedman test](#)[452]	Yes	Yes	NA

5.13.1.3 Repeated measures tab

Two ways to analyze repeated measures data

Prism can analyze repeated measures data in two ways:

- Repeated measures ANOVA

- Fitting a mixed effects model. This analysis works fine even when there are some missing values. The results will only be meaningful, of course,

if the values are missing for random reasons. For example, those results won't be helpful or meaningful if the values are missing because those participants were very sick, or those values were too high to measure (or too low to measure). Fitting a mixed model with missing values only makes sense when there is zero association between the treatments or time-points and the reason why some values are missing.

In general, fitting a mixed effects model is a much more versatile method. As implemented in Prism 8, the two are completely equivalent when there are no missing values. But the mixed effects model method can also fit data with missing values.

Analyze using which method

The repeated measures tab of the ANOVA dialog (same for one-, two- and three-way data) gives you three choices:

- Use repeated measures ANOVA always. If there are missing values, no results will be reported. This matches what Prism 7 and earlier did. Prism is not "smart enough" to remove all data for a participant with missing values, but you could exclude all those values and rerun the ANOVA.

- Fit a mixed effects model always. This will make all analyses be consistent, whether or not there are missing values. If there are no missing values, the key results will be the same as repeated measures ANOVA but the results will be presented in an format unfamiliar to those used to repeated measures ANOVA.

- Report the fit to a mixed effects model only when there are missing values, when repeated measures ANOVA is impossible. When there are no missing values, report the familiar repeated measures ANOVA results.

How to fit the mixed effects model any random factor is zero or negative

The whole point of repeated measures or mixed model analyses is that you have multiple response measurements on the same subject or when individuals are matched (twins or litters), so need to account for any correlation among multiple responses from the same subject. Mixed model analysis does this by estimating variances between subjects. In a simple mixed model, where only one variable is repeated, it's possible that this correlation for the particular data in your study is zero or even

negative (of course it is impossible for a variance to be negative, but it can happen with mixed effects models). You are given two choices for what Prism should do when this happens:

•Analyze as usual. If there are no missing values, this will match repeated measures ANOVA.

•Remove the subject factor from the model and refit. This approach will have one more degree of freedom and thus have a bit more power. This approach is better but means the mixed model results may not match repeated measures ANOVA results.

In more complicated models, where there is more than one repeated measures variable, there are even more possible variance estimates (generally interactions with subject), and any of those could turn out to be zero or negative. It's best to take these out as leaving them in can make the results unstable.

Defaults for future analyses

Check an option at the bottom of the Repeated Measures tab to make your choices the default for future analyses. Your default will apply to one-, two- and three-way ANOVA.

5.13.1.4 Multiple comparisons tab: One-way ANOVA

Overview of multiple comparisons choices

Multiple comparisons testing is chosen on two tabs of the analysis parameters dialog.

• The Multiple Comparisons tab specifies the questions you want the multiple comparisons tests to answer. This decision depends on the experimental design and will vary from experiment to experiment.

• The [next tab (Options)][414] drills down to choose a test. Those choices tend to be personal or lab preferences, so we put them on the options tab.

These choices should reflect the experimental plan. It is not ok to first look at the data, and then decide which comparisons you want Prism to calculate. If you first look at the data, then you effectively have made all possible comparisons.

How many comparisons?

Parameters: One-Way ANOVA (and Nonparametric)   ✕

Experimental Design | **Multiple Comparisons** | Options | Residuals

Followup tests

⦿ None.

◯ Compare the mean of each column with the mean of every other column.

◯ Compare the mean of each column with the mean of a control column.

     Control column: Column A : Control   ⌄

◯ Compare the means of preselected pairs of columns.

     Selected pairs: Select...

◯ Test for linear trend between column mean and left-to-right column order.

Which test?

Use choices on the Options tab to choose the test, and to set the defaults for future ANOVAs.

No multiple comparisons

Multiple comparisons are optional.

Compare each column mean with every other mean

This is probably the most commonly used comparison. Because it makes more comparisons that the other choices, it will have less power to detect differences. You'll choose the exact test on the Options tab, but Tukey's test is used most often.

Compare the mean of each column with the mean of a control column

It is common to only wish to compare each group to a control group, and not to every other group. This reduces the number of comparisons considerably (at least if there are many groups), and so increases the power to detect differences. You'll choose the exact test on the Options tab, but Dunnett's test is used most often.

Compare preselected pairs of columns

Comparing preselected pairs of column means reduces the number of comparisons, and so increases power. But you must have chosen the pairs of means to compare as part of the experimental design and your

scientific goals. If you looked at the data first, and then decided which pairs of means to compare, then you really compared all means. Read about these planned comparisons[123]. You can select up to 40 pairs, but the list may get truncated if you open the file on older versions of Prism).

Test for linear trend

The test for linear trend is a specialized test that only makes sense if the columns are arranged in a natural order (e.g. dose or time) and you want to test whether there is a trend[442] such that the column means tend to increase (or decrease) as you move from left to right across columns. The other multiple comparisons tests pay no attention at all to the order of the data sets. Note that while this choice is on the Multiple Comparisons tab, there really is only one comparison.

Choices Prism doesn't offer

Prism omits two choices that some other programs offer.

- Prism cannot do comparisons (called contrasts) that involve multiple groups -- for example, comparing the grand mean of groups A and B, with the grand mean of groups C, D and E. Scheffe's method is designed to handle these kinds of contrasts, but Prism does not offer it.

- While Prism can test for linear trend between column mean and column order, it cannot test for other trends (quadratic, etc.).

5.13.1.5 Options tab: Multiple comparisons: One-way ANOVA

There are two help screens for the Options tab for the one-way ANOVA dialog:

- This page explains the multiple comparisons options.

- A different page explains the graphing and output options[418].

Which multiple comparison test?

The list of tests available depends on the goal you specified on the Multiple Comparisons (second) tab and your choice about assuming equal SDs on the Experimental Design (first) tab.

Correct for multiple comparisons using statistical hypothesis testing

Compare every mean with every other mean

The available choices depend on whether you assume assume homoscedasticity (equal SDs, so equal variances) on the first tab of the ANOVA dialog.

If you assume homoscedasticity (equal SDs), you choices are:

- [Tukey test][136] (recommended)

- [Bonferroni][379]

- [Sidak][379]

- [Holm-Sidak][378]. This test is more powerful than the Tukey method (3), which means that it can sometimes find a statistically significant difference where the Tukey method cannot. This test cannot compute confidence intervals, and for this reason we prefer the Tukey test.

- Newman-Keuls. We offer this test only for compatibility with old versions of Prism, but we suggest you avoid it. The problem is that it does not maintain the family-wise error rate at the specified level(1). In some cases, the chance of a Type I error can be greater than the alpha level you specified.

If you do not assume homoscedasticity (equal SDs), [your choices][384] are:

- Games-Howell (recommended for large samples)

- Dunnett T3 (recommended when sample size per group is less than 50)

- Tamhane T2

All [three methods][384] can compute confidence intervals and multiplicity adjusted P values.

Compare a control mean with the other means

If you assume homoscedasticity (equal SDs), your choices are:

- [Dunnett's][136] (recommended)

- [Bonferroni][379]

- [Sidak][379]

- [Holm-Sidak][378]. Glantz says that Holm's test ought to have more power than Dunnett's test, but this has not (to his knowledge) been explored in depth(2). This test cannot compute confidence intervals or multiplicity adjusted P values, and for this reason we prefer the Dunnett test.

If you do not assume homoscedasticity (equal SDs), your choices are:

- [Dunnett's T3][136] (recommended)

- [Tamhane T2][384]

Compare the means of a prespecified pairs of columns

If you assume homoscedasticity (equal SDs), your choices are:

- [Bonferroni][379] (most commonly used)

- [Sidak][379] (more power, so recommended)

- [Holm-Sidak][378] (cannot compute confidence intervals)

If you do not assume homoscedasticity (equal SDs), [your choices][384] are:

- Games-Howell (recommended)

- Dunnett T3

- Tamhane T2

Correct for multiple comparisons by controlling the False Discovery Rate

Prism offers [three methods][142] to control the [false discovery rate][140]. All decide which (if any) comparisons to label as "discoveries" and do so in a way that controls the false discovery rate to be less than a value Q you enter.

The FDR approach is not often used as a followup test to ANOVA, but there is no good reason for that.

Don't correct for multiple comparisons. Each comparison stands alone.

If you choose this approach, Prism will perform [Fisher's Least Significant Difference (LSD) test][127].

This approach (Fisher's LSD) has much more power to detect differences. But it is more likely to falsely conclude that a difference is statistically significant. When you correct for multiple comparisons (which Fisher's LSD does not do), the significance threshold (usually 5% or 0.05) applies to the entire family of comparisons. With Fisher's LSD, that threshold applies separately to each comparison.

Only use the Fisher's LSD approach if you have a very good reason, and are careful to explain what you did when you report the results.

Multiple comparisons options

Swap direction of comparisons

The only affect of this option is to change the sign of all reported differences between means. A difference of 2.3 will be -2.3 if the option is checked. A difference of -3.4 will be 3.4 if you check the option. It is purely a personal preference that depends on how you think about the data.

Report multiplicity adjusted P value for each comparison

If you choose the Bonferroni, Tukey or Dunnett multiple comparisons test, Prism can also report [multiplicity adjusted P values][129]. If you check this option, Prism reports an adjusted P value for each comparison. These calculations take into account not only the two groups being compared, but the total number groups (data set columns) in the ANOVA, and the data in all the groups. With Dunnett's test, Prism can only report a multiplicity adjusted P value when it would be greater than 0.0001.

Otherwise it reports "<0.0001" (prior to Prism 8, Prism reported 0.0001 without the less-than symbol).

The multiplicity adjusted P value is the smallest significance threshold (alpha) for the entire family of comparisons at which a particular comparison would be (just barely) declared to be "statistically significant".

Until recently, multiplicity adjusted P values have not been commonly reported. If you choose to ask Prism to compute these values, take the time to be sure you understand what they mean. If you include these values in publications or presentations, be sure to explain what they are.

Confidence and significance level (or desired FDR)

By tradition, confidence intervals are computed for 95% confidence and statistical significance is defined using an alpha of 0.05. Prism lets you choose other values. If you choose to control the FDR, select a value for Q (in percent). If you set Q to 5%, you expect up to 5% of the "discoveries" to be false positives.

References

1. SA Glantz, Primer of Biostatistics, sixth edition, ISBN= 978-0071435093.

2. MA Seaman, JR Levin and RC Serlin, Psychological Bulletin 110:577-586, 1991.

5.13.1.6 Options tab: Graphing and output: One-way ANOVA

There are two help screens for the Options tab for the one-way ANOVA dialog:

- A different page explains the [multiple comparisons options](#)[414].

- This page explains the graphing and output options.

Graphing

Prism gives you options to create some extra graphs, each with its own extra page of results.

- If you chose a multiple comparison method that computes confidence intervals (Tukey, Dunnett, etc.) Prism can plot these confidence intervals.

- You can choose to plot the residuals. For ordinary ANOVA, each residual is the difference between a value and the mean value of that group. For repeated measures ANOVA, each residual is computed as the difference between a value and the mean of all values from that particular individual (row).

- If you chose the Kruskal-Wallis nonparametric test, Prism can plot the ranks of each value, since that is what the test actually analyzes.

- If you chose repeated measures ANOVA, Prism can plot the differences. If you have four treatments (A, B, C, D), there will be six set of differences (A-B, A-C, B-C, A-D, B-D, C-D). Seeing these differences graphed can give you a better feel for the data.

Additional results

- You can choose an extra page of results showing descriptive statistics for each column, similar to what the Column statistics analysis reports.

- Prism also can report the overall ANOVA comparison using the information theory approach (AICc), in addition to the usual P value. Prism fits two models to the data -- one where all the groups are sampled from populations with identical means, and one with separate means -- and tells you the likelihood that each is correct. This is not a standard way to view ANOVA results, but it can be informative.

Output

Choose [how you want P values reported](#)[82], and how many significant digits you need.

5.13.1.7 Residuals tab: One-way ANOVA

Why residuals?

Prism 8 introduced the ability to plot residual plots with ANOVA, provided that you entered raw data and not averaged data as mean, n and SD or SEM.

Many scientists thing of residual as values that are obtained with regression. But ANOVA is really regression in disguise. It fits a model. One of the assumptions of ANOVA is that the residuals from that model are sampled from a Gaussian distribution. A residual plot helps you assess this assumption.

Which graph to create?

Prism can make three kinds of residual plots.

- Residual plot. The X axis is the predicted value (or fitted value), the mean of the replicates of the data (but see below for repeated measures). The Y axis is the residual. This lets you spot residuals that are much larger or smaller than the rest.

- Homoscedasticity plot. The X axis is the predicted value (or fitted value), the mean of the replicates of the data (but see below for repeated measures). The Y axis is the absolute value of the residual.This lets you check whether larger values are associated with bigger residuals (large absolute value).

- QQ plot. The X axis is the actual residual. The Y axis is the predicted residual, computed from the percentile of the residual (among all residuals) and assuming sampling from a Gaussian distribution. ANOVA assumes a Gaussian distribution of residuals, and this graph lets you check that assumption.

Diagnostics for residuals

- Are the residuals clustered or heteroscedastic? ANOVA assumes each sample was randomly drawn from populations with the same standard deviation. Prism can test this assumption with two tests. The Brown-Forsythe test and the Barlett test. Both these tests compute a P value designed to answer this question:If the populations really have the same standard deviations, what is the chance that you'd randomly select samples whose standard deviations are as different from one another (or more different) as they are in your experiment?

- Are the residuals Gaussian? Prism runs four normality tests on the residuals. The residuals from all groups are pooled and then entered into one normality test.

How residuals are computed

Residuals with one-way ANOVA and related tests are simple to understand.

- One-way ANOVA. A residual is computed for each value. Each residual is the difference between a entered value and the mean of all values for that group. A residual is positive when the corresponding value is greater than the sample mean, and is negative when the value is less than the sample mean.

- One-way repeated measures ANOVA. This is harder to understand. The residual is calculated as Actual value - Predicted value, where Predicted value = predicted group mean + predicted subject (row) mean - predicted grand mean.

- Kruskal-Wallis test. A residual is computed for each value. Each residual is the difference between a entered value and the median of all values for that group. A residual is positive when the corresponding value is greater than the sample median, and is negative when the value is less than the sample median.

- Friedman matched pairs test. This is harder to understand than the others. The residual is calculated as Actual value - Predicted value, where Predicted value = predicted group median + predicted subject median - predicted grand median.

5.13.1.8 Q&A: One-way ANOVA

▼ **Is it possible to define the groups with a grouping variable?**

No. The groups must be defined by columns. Enter data for one group into column A, another group into column B, etc..

▼ **Can I enter data in lots of columns and then choose which to include in the ANOVA?**

Yes. After you click Analyze, you'll see a list of all data sets on the right side of the dialog. Select the ones you wish to compare.

▼ **Can I enter data as mean, SD (or SEM) and N?**

Yes. Follow [this example][405] to see how. It is impossible to run repeated measures ANOVA or a nonparametric test from data entered as mean, SD (or SEM) and N. You can only choose an ordinary one-way ANOVA.

▼ **If I have data from three or more groups, but I am particularly interested in comparing certain groups with other groups. Is it OK to compare two groups at a time with a t test?**

No. You should analyze all the groups at once with [one-way ANOVA][425], and then follow up with multiple comparison post tests. An [exception][123] is when some of the 'groups' are really controls to prove the assay worked, and are not really part of the experimental question you are asking.

⊟ **I know the mean, SD (or SEM) and sample size for each group. Which tests can I run?**

You can enter data as mean, SD (or SEM) and N, and Prism can compute one-way ANOVA. It is not possible to compute repeated measures ANOVA, or nonparametric ANOVA without access to the raw data.

☐ I only know the group means, and don't have the raw data and don't know their SD or SEM. Can I run ANOVA?

No. ANOVA compares the difference among group means with the scatter within the groups, taking into account sample size. If you only know the means, there is no possible way to do any statistical comparison.

☐ Can I use a normality test to make the choice of when to use a nonparametric test?

This is [not a good idea][148]. Choosing when to use a nonparametric test is not straightforward, and you can't really automate the process.

☐ I want to compare three groups. The dependent variable has two possible outcomes. How can I compare the groups?

Not with ANOVA. Enter your data into a [contingency table][545] and analyze with a [chi-square test][547].

☐ What does 'one-way' mean?

One-way ANOVA, also called one-factor ANOVA, determines how a response is affected by one factor. For example, you might measure a response to three different drugs. In this example, drug treatment is the factor. Since there are three drugs, the factor is said to have three levels.

If you measure response to three different drugs, and two time points, then you have two factors: drug and time. One-way ANOVA would not be helpful. Use two-way ANOVA instead.

If you measure response to three different drugs at two time points with subjects from two age ranges, then you have three factors: drug, time and age. Prism does not perform three-way ANOVA, but other programs do.

If there are only two levels of one factor --say male vs. female, or control vs. treated --, then you should use a t test. One-way ANOVA

is used when there are three or more groups (although the underlying math is the same for a t test and one-way ANOVA with two groups).

⊟ **What does 'repeated measures' mean? How is it different than 'randomized block'?**

The term *repeated-measures* strictly applies only when you give treatments repeatedly to each subject, and the term *randomized block* is used when you randomly assign treatments within each group (block) of matched subjects. The analyses are identical for repeated-measures and randomized block experiments, and Prism always uses the term repeated-measures.

⊟ **How should I decide whether or not to assume sphericity?**

This question only applies to repeated-measures ANOVA. These tips might help:

- Previous versions of Prism assumed [sphericity][395]. Check the option to assume sphericity to match results from older versions.

- If you ask Prism not to assume sphericity, the P values will be larger but probably more accurate. Confidence intervals of multiple comparisons tests will be computed differently. Some will be wider and some narrower than they would have been if you had assumed sphericity.

- We suggest that, if in doubt, you choose to not assume sphericity.

- It sounds sensible to measure deviations from sphericity (with epsilon), and then use that value to decide whether or not the ANOVA should assume sphericity. But statisticians have shown this approach works poorly. You need to decide based on experimental design, not based on the data.

⊟ **Can the overall ANOVA give a statistically significant result, while no multiple comparison test does?**

Yes.

5.13.2 One-way ANOVA results

5.13.2.1 Interpreting results: One-way ANOVA

One-way ANOVA compares three or more unmatched groups, based on the assumption that the populations are Gaussian.

P value

The P value tests the null hypothesis that data from all groups are drawn from populations with identical means. Therefore, the P value answers this question:

> If all the populations really have the same mean (the treatments are ineffective), what is the chance that random sampling would result in means as far apart (or more so) as observed in this experiment?

If the overall P value is large, the data do not give you any reason to conclude that the means differ. Even if the population means were equal, you would not be surprised to find sample means this far apart just by chance. This is not the same as saying that the true means are the same. You just don't have compelling evidence that they differ.

If the overall P value is small, then it is unlikely that the differences you observed are due to random sampling. You can reject the idea that all the populations have identical means. This doesn't mean that every mean differs from every other mean, only that at least one differs from the rest. Look at the results of post tests to identify where the differences are.

F ratio and ANOVA table

The P value is computed from the F ratio which is computed from the ANOVA table.

ANOVA partitions the variability among all the values into one component that is due to variability among group means (due to the treatment) and another component that is due to variability within the groups (also called residual variation). Variability within groups (within the columns) is quantified as the sum of squares of the differences between each value and its group mean. This is the residual sum-of-squares. Variation among groups (due to treatment) is quantified as the sum of the squares of the

differences between the group means and the grand mean (the mean of all values in all groups). Adjusted for the size of each group, this becomes the treatment sum-of-squares.

Each sum-of-squares is associated with a certain number of degrees of freedom (df, computed from number of subjects and number of groups), and the mean square (MS) is computed by dividing the sum-of-squares by the appropriate number of degrees of freedom. These can be thought of as variances. The square root of the mean square residual can be thought of as the pooled standard deviation.

The F ratio is the ratio of two mean square values. If the null hypothesis is true, you expect F to have a value close to 1.0 most of the time. A large F ratio means that the variation among group means is more than you'd expect to see by chance. You'll see a large F ratio both when the null hypothesis is wrong (the data are not sampled from populations with the same mean) and when random sampling happened to end up with large values in some groups and small values in others.

The P value is determined from the F ratio and the two values for degrees of freedom shown in the ANOVA table.

Tests for equal variances

ANOVA is based on the assumption that the data are sampled from populations that all have the same standard deviations. Prism tests this assumption with two tests. It computes the Brown-Forsythe test and also (if every group has at least five values) computes Bartlett's test. There are no options for whether to run these tests. Prism automatically does so and always reports the results.

Both these tests compute a P value designed to answer this question:

> If the populations really have the same standard deviations, what is the chance that you'd randomly select samples whose standard deviations are as different from one another (or more different) as they are in your experiment?

Bartlett's test

Prism reports the results of the "corrected" Barlett's test as explained in section 10.6 of Zar(1). Bartlett's test works great if the data really are sampled from Gaussian distributions. But if the distributions deviate even

slightly from the Gaussian ideal, Bartett's test may report a small P value even when the differences among standard deviations is trivial. For this reason, many do not recommend that test. That's why we added the test of Brown and Forsythe. It has the same goal as the Bartlett's test, but is less sensitive to minor deviations from normality. We suggest that you pay attention to the Brown-Forsythe result, and ignore Bartlett's test (which we left in to be consistent with prior versions of Prism).

Brown-Forsythe test

The Brown-Forsythe test is conceptually simple. Each value in the data table is transformed by subtracting from it the median of that column, and then taking the absolute value of that difference. One-way ANOVA is run on these values, and the P value from that ANOVA is reported as the result of the Brown-Forsythe test.

How does it work. By subtracting the medians, any differences between medians have been subtracted away, so the only distinction between groups is their variability.

Why subtract the median and not the mean of each group? If you subtract the column mean instead of the column median, the test is called the *Levene test for equal variances*. Which is better? If the distributions are not quite Gaussian, it depends on what the distributions are. Simulations from several groups of statisticians show that using the median works well with many types of nongaussian data. Prism only uses the median (Brown-Forsythe) and not the mean (Levene).

Interpreting the results

If the P value is small, you must decide whether you will conclude that the standard deviations of the populations are different. Obviously the tests of equal variances are based only on the values in this one experiment. Think about data from other similar experiments before making a conclusion.

If you conclude that the populations have different variances, you have four choices:

* Conclude that the populations are different. In many experimental contexts, the finding of different standard deviations is as important as the finding of different means. If the standard deviations are truly different, then the populations are different regardless of what ANOVA

concludes about differences among the means. This may be the most important conclusion from the experiment.

- Transform the data to equalize the standard deviations, and then rerun the ANOVA. Often you'll find that converting values to their reciprocals or logarithms will equalize the standard deviations and also make the distributions more Gaussian.

- Use the Welch or Brown-Forsythe versions of one-way ANOVA that do not assume that all standard deviations are equal.

- Switch to the nonparametric Kruskal-Wallis test. The problem with this is that if your groups have very different standard deviations, it is difficult to interpret the results of the Kruskal-Wallis test. If the standard deviations are very different, then the shapes of the distributions are very different, and the kruskal-Wallis results cannot be interpreted as comparing medians.

R squared

R^2 is the fraction of the overall variance (of all the data, pooling all the groups) attributable to differences among the group means. It compares the variability among group means with the variability within the groups. A large value means that a large fraction of the variation is due to the treatment that defines the groups. The R^2 value is calculated from the ANOVA table and equals the between group sum-of-squares divided by the total sum-of-squares. Some programs (and books) don't bother reporting this value. Others refer to it as $\eta 2$ (eta squared) rather than R^2. It is a descriptive statistic that quantifies the strength of the relationship between group membership and the variable you measured.

Reference

J.H. Zar, Biostatistical Analysis, Fifth edition 2010, ISBN: 0131008463.

5.13.2.2 Analysis checklist: One-way ANOVA

One-way ANOVA compares the means of three or more unmatched groups. Read elsewhere to learn about choosing a test[406], and interpreting the results[425].

Are the populations distributed according to a Gaussian distribution?

One-way ANOVA assumes that you have sampled your data from populations that follow a Gaussian distribution. While this assumption is not too important with large samples due to the Central Limit Theorem[41] , it is important with small sample sizes (especially with unequal sample sizes). Prism can test for violations of this assumption, but normality tests have limited utility.

If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values (perhaps to logs or reciprocals) to make the distributions more Gaussian. Another choice is to use the Kruskal-Wallis nonparametric test instead of ANOVA. A final option is to use ANOVA anyway, knowing that it is fairly robust to violations of a Gaussian distribution with large samples.

Do the populations have the same standard deviation?

One-way ANOVA assumes that all the populations have the same standard deviation (and thus the same variance). This assumption is not very important when all the groups have the same (or almost the same) number of subjects, but is very important when sample sizes differ.

InStat tests for equality of variance with two tests: The Brown-Forsythe test and Bartlett's test. The P value from these tests answer this question: If the populations really have the same variance, what is the chance that you'd randomly select samples whose variances are as different from one another as those observed in your experiment. A small P value suggests that the variances are different.

Don't base your conclusion solely on these tests. Also think about data from other similar experiments. If you have plenty of previous data that convinces you that the variances are really equal, ignore these tests (unless the P value is really tiny) and interpret the ANOVA results as usual. Some statisticians recommend ignoring tests for equal variance altogether if the sample sizes are equal (or nearly so).

In some experimental contexts, finding different variances may be as important as finding different means. If the variances are different, then the populations are different -- regardless of what ANOVA concludes about differences between the means.

✓ Are the data unmatched?

 One-way ANOVA works by comparing the differences among group means with the pooled standard deviations of the groups. If the data are matched, then you should choose repeated-measures ANOVA instead. If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA.

✓ Are the "errors" independent?

 The term "error" refers to the difference between each value and the group mean. The results of one-way ANOVA only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✓ Do you really want to compare means?

 One-way ANOVA compares the means of three or more groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✓ Is there only one factor?

 One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group, with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments.

 Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. These data need to be analyzed by two-way ANOVA[455], also called two factor ANOVA.

✔ Is the factor "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Type II ANOVA, also known as random-effect ANOVA, assumes that you have randomly selected groups from an infinite (or at least large) number of possible groups, and that you want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment. Type II random-effects ANOVA is rarely used, and Prism does not perform it.

✔ Do the different columns represent different levels of a grouping variable?

One-way ANOVA asks whether the value of a single variable differs significantly among three or more groups. In Prism, you enter each group in its own column. If the different columns represent different variables, rather than different groups, then one-way ANOVA is not an appropriate analysis. For example, one-way ANOVA would not be helpful if column A was glucose concentration, column B was insulin concentration, and column C was the concentration of glycosylated hemoglobin.

5.13.3 Welch and Brown-Forsythe ANOVA

Welch and Brown-Forsythe ANOVA compares three or more sets of unpaired measurements (data expressed using an [interval or ratio scale](#)[30]), assumed to be sampled from a Gaussian distribution but without assuming that the groups have equal variances.

5.13.3.1 Interpreting results: Welch and Brown-Forsythe tests

One-way ANOVA compares three or more unmatched groups, based on the assumption that the populations are Gaussian. The Welch and Brown-

Forsythe versions of one-way ANOVA do not assume that all the groups were sampled from populations with equal variances.

P value

The P value tests the null hypothesis that data from all groups are drawn from populations with identical means. Therefore, the P value answers this question:

> If all the populations really have the same mean (the treatments are ineffective), what is the chance that random sampling would result in means as far apart (or more so) as observed in this experiment?

If the overall P value is large, the data do not give you any reason to conclude that the means differ. Even if the population means were equal, you would not be surprised to find sample means this far apart just by chance. This is not the same as saying that the true means are the same. You just don't have compelling evidence that they differ.

If the overall P value is small, then it is unlikely that the differences you observed are due to random sampling. You can reject the idea that all the populations have identical means. This doesn't mean that every mean differs from every other mean, only that at least one differs from the rest. Look at the results of multiple comparisons tests to identify where the differences are.

How the Welch test works

Prism reports the W ratio, which is analogous to the F ratio of ordinary ANOVA. If sample size is equal for all groups, the value of W is identical to what the F ratio would have been in ordinary one-way ANOVA. If the sample sizes are not equal, W is not the same as F. Depending on the data, W can be larger or smaller than F.

Prism also reports the number of degrees of freedom for numerator and denominator. The numerator df is the same as it would have been with regular ANOVA. The denominator df is different, whether or not the sample sizes are adjusted.

The P value is computed from W using the same algorithm to compute a P value from F. Depending on the data, the P value from the Welch test can be larger or smaller than the P value from ordinary ANOVA.

How the Brown-Forsythe test works

Point of possible confusion: The Brown-Forsythe test here is a test for equality of means. It is distinct from another test by Brown and Forsythe to test equality of variances.

Prism reports the F* ratio, which is analogous to the F ratio of ordinary ANOVA.

Prism also reports the number of degrees of freedom for numerator and denominator. The numerator df is the same as it would have been with regular ANOVA. The denominator df is different.

The P value is computed from F* using the same algorithm to compute a P value from F. Depending on the data, the P value from the Brown-Forstyhe test can be larger or smaller than the P value from ordinary ANOVA.

Choosing between Welch and Brown-Forsythe tests

Glantz and colleagues (1) recommend using the Welch test in most situations, as it both has more power and maintains alpha at its desired level. They recommend Brown-Forsythe in one situation, when the data are skewed (not Gaussian).

1. SA Glantz, BK Slinker, TB Neilands, [Primer of Regression & Analysis of Variance](#), Third edition, 2016.

5.13.3.2 Analysis checklist: Welch and Brown-Forsythe tests

One-way ANOVA compares the means of three or more unmatched groups. Read elsewhere to learn about [choosing a test](#)[406], and [interpreting the results](#)[425].

✔️ **Are the populations distributed according to a Gaussian distribution?**

One-way ANOVA assumes that you have sampled your data from populations that follow a Gaussian distribution. While this assumption is not too important with large samples due to the [Central Limit Theorem](#)[41], it is important with small sample sizes (especially with unequal sample sizes). Prism can test for violations of this assumption, but normality tests have limited utility.

If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values (perhaps to logs or reciprocals) to make the distributions more Gaussian. Another choice is to use the Kruskal-Wallis nonparametric test instead of ANOVA. A final option is to use ANOVA anyway, knowing that it is fairly robust to violations of a Gaussian distribution with large samples.

✔ Are the data unmatched?

One-way ANOVA works by comparing the differences among group means with the pooled standard deviations of the groups. If the data are matched, then you should choose repeated-measures ANOVA instead. If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA.

✔ Are the "errors" independent?

The term "error" refers to the difference between each value and the group mean. The results of one-way ANOVA only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six values in each group, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all triplicates from one animal to be high or low.

✔ Do you really want to compare means?

One-way ANOVA compares the means of three or more groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✔ Is there only one factor?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group, with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments.

Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. These data need to be analyzed by [two-way ANOVA](#)[455], also called two factor ANOVA.

✓ Is the factor "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Type II ANOVA, also known as random-effect ANOVA, assumes that you have randomly selected groups from an infinite (or at least large) number of possible groups, and that you want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment. Type II random-effects ANOVA is rarely used, and Prism does not perform it.

✓ Do the different columns represent different levels of a grouping variable?

One-way ANOVA asks whether the value of a single variable differs significantly among three or more groups. In Prism, you enter each group in its own column. If the different columns represent different variables, rather than different groups, then one-way ANOVA is not an appropriate analysis. For example, one-way ANOVA would not be helpful if column A was glucose concentration, column B was insulin concentration, and column C was the concentration of glycosylated hemoglobin.

5.13.4   Repeated-measures one-way ANOVA or mixed model

5.13.4.1   What is repeated measures?

The advantage of repeated measures

The difference between ordinary and repeated measures ANOVA, is similar to the difference between unpaired and paired t tests. See the [advantages of pairing or matching](#)[270]. Since each participant or experiment acts as its own control, repeated measures design can do a better job of separating signal from noise, so this design usually has more power. Some participants may have larger measurements at all time points, and others may have lower measurements at all time points.

Repeated measures ANOVA focuses on how much the Y value changes between treatments.

Repeated measures or randomized block?

The term *repeated measures* is used when you give treatments repeatedly to each animal or participant.

The term *randomized block* is used when you randomly assign treatments within each group (block) of matched subjects.

Imagine that you compare three different treatments. In a repeated measures design, you'd recruit say 10 subjects (or use ten animals) and measure each of the subjects (animals) after each of the treatments. With a randomized block design, you'd recruit ten sets of four subject each, matched for age, gender etc. (or ten sets of four animals, with the four treated at the same time in adjacent cages...).

ANOVA works identically for repeated-measures and randomized block experiments, and Prism always uses the term *repeated-measures.*

One way? Or two way?

A design with three or more measurements on the same subject is called repeated measures one-way ANOVA in Prism, because there really is only one factor, denoted by the data set columns. But you could argue there is a second factor too, subject, because each row represents a different subject (or block). In fact, you'll get the same results if you analyze with two-way ANOVA (without replicates) and one-way repeated measures ANOVA.

5.13.4.2 Multiple comparisons after repeated measures one-way ANOVA

The use of multiple comparisons tests after repeated measures ANOVA is a tricky topic that many statistics texts avoid. We follow methods suggested by Maxwell and Delaney(1).

With one way ANOVA, Prism computes the multiple comparisons tests in two different ways, depending on whether you ask Prism (on the first tab of the ANOVA dialog) to assume sphericity[395].

If you assume sphericity

The multiple comparisons tests performed by Prism use the mean square residual for all comparisons. This is a pooled value that assess variability in all the groups. If you assume that variability really is the same in all groups (with any differences due to chance) this gives you more power. This makes sense, as you get to use data from all time points to assess variability, even when comparing only two times.

If you do not assume sphericity

If you check the option to not assume sphericity, Prism does two things differently.

• It applies the Geisser-Greenhouse correction when computing the P values for the main effect.

• It computes the multiple comparisons differently. For each comparison of two groups, it uses only the data in those two groups (essentially performing a paired t test). This makes sense when scatter increases with time, so later treatments give a more variable response than earlier treatments. It uses the method described on pages 552-555 of Maxwell(1).

When you choose not to assume sphericity, some multiple comparisons will have more power (and narrower confidence intervals) than they would if you did not assume sphericity. But others will have less power (and wider confidence intervals).

Reference

1. Scott E. Maxwell, Harold D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective,* Second Edition. IBSN:0805837183

5.13.4.3 Interpreting results: Repeated measures one-way ANOVA

Repeated-measures ANOVA compares the means of three or more matched groups. The term *repeated-measures* strictly applies only when you give treatments repeatedly to each subject, and the term *randomized block* is used when you randomly assign treatments within each group (block) of matched subjects. The analyses are identical for repeated-

measures and randomized block experiments, and Prism always uses the term repeated-measures.

P value

The P value answers this question:

> If all the populations really have the same mean (the treatments are ineffective), what is the chance that random sampling would result in means as far apart (or more so) as observed in this experiment?

If the overall P value is large, the data do not give you any reason to conclude that the means differ. Even if the true means were equal, you would not be surprised to find means this far apart just by chance. This is not the same as saying that the true means are the same. You just don't have compelling evidence that they differ.

If the overall P value is small, then it is unlikely that the differences you observed are due to random sampling. You can reject the idea that all the populations have identical means. This doesn't mean that every mean differs from every other mean, only that at least one differs from the rest. Look at the results of post tests to identify where the differences are.

Was the matching effective?

A repeated-measures experimental design can be very powerful, as it controls for factors that cause variability between subjects. If the matching is effective, the repeated-measures test will yield a smaller P value than an ordinary ANOVA. The repeated-measures test is more powerful because it separates between-subject variability from within-subject variability. If the pairing is ineffective, however, the repeated-measures test can be less powerful because it has fewer degrees of freedom.

Prism tests whether the matching was effective and reports a P value that tests the null hypothesis that the population row means are all equal. If this P value is low, you can conclude that the matching was effective. If the P value is high, you can conclude that the matching was not effective and should reconsider your experimental design.

F ratio and ANOVA table

The P values are calculated from the ANOVA table. With repeated-measures ANOVA, there are three sources of variability: between columns (treatments), between rows (individuals), and random (residual). The ANOVA table partitions the total sum-of-squares into those three components. It then adjusts for the number of groups and number of subjects (expressed as degrees of freedom) to compute two F ratios. The main F ratio tests the null hypothesis that the column means are identical. The other F ratio tests the null hypothesis that the row means are identical (this is the test for effective matching). In each case, the F ratio is expected to be near 1.0 if the null hypothesis is true. If F is large, the P value will be small.

If you don't accept the assumption of sphericity

If you checked the option to not accept the assumption of sphericity, Prism does two things differently.

- It applies the correction of Geisser and Greenhouse. You'll see smaller degrees of freedom, which usually are not integers. The corresponding P value is higher than it would have been without that correction.

- It reports the value of [epsilon][400], which is a measure of how badly the data violate the assumption of sphericity.

R^2

Prism reports two different R^2 values, computed by taking ratios of sum-of-squares (SS):

- To quantify how large the treatment effects are. There are two ways to compute this. Prism uses the method described by Keppel (1), in which R^2 is the variation due to treatment effects as a fraction of the sum of the variation due to treatment effects plus random variation. That text refers to the value as both R^2 and also eta squared, and states that this value an estimate of the partial omega squared. It is computed simply as the SS treatment divided by the sum of the SS treatment plus the SSresidual. Note that variation between subjects (SSindividual) is not part of the calculation. This R^2 is reported in the results section with the heading "Repeated measures ANOVA summary".

- To quantify how effecting the effectiveness of matching. This R^2 quantifies the fraction of total variation that is due to differences among

subjects. It is computed as SSindividual divided by the SStotal, and reported within the results section with the heading "Was the matching effective".

Multiple comparisons tests and analysis checklist

Learn about <u>multiple comparisons tests after repeated measures ANOVA</u>[436].

Before interpreting the results, <u>review the analysis checklist</u>[181].

(1) G Keppel and TD Wickens, <u>Design and Analysis</u>, Fourth Edition, 2004, ISBN: 0135159415

5.13.4.4 Interpreting results: Mixed model one-way

Fitting a mixed effects model to repeated-measures one-way data compares the means of three or more matched groups. The term *repeated-measures* strictly applies only when you give treatments repeatedly to each subject, and the term *randomized block* is used when you randomly assign treatments within each group (block) of matched subjects. The analyses are identical for repeated-measures and randomized block experiments, and Prism always uses the term repeated-measures.

Read about<u>using the mixed model to fit repeated measures da</u>[391]ta.

Fixed effect P value

There is one fixed effect in the model, the variable that determines which column each value was placed into. The mixed effects model results present a P value that answers this question:

> If all the populations really have the same mean (the treatments are ineffective), what is the chance that random sampling would result in means as far apart (or more so) as observed in this experiment?

If the overall P value is large, the data do not give you any reason to conclude that the means differ. Even if the true means were equal, you would not be surprised to find means this far apart just by chance. This is

not the same as saying that the true means are the same. You just don't have compelling evidence that they differ.

If the overall P value is small, then it is unlikely that the differences you observed are due to random sampling. You can reject the idea that all the populations have identical means. This doesn't mean that every mean differs from every other mean, only that at least one differs from the rest. Look at the results of post tests to identify where the differences are.

Random effects SD and variance

The mixed effects model treats the different subjects (participants, litters, etc) as a random variable. The residual random variation is also random. Prism presents the variation as both a SD and a variance (which is the SD squared). You, or more likely your statistical consultant, may be interested in these values to compare with other programs. The calculation of these values is complicated requiring matrix algebra.

Was the matching effective?

A repeated-measures experimental design can be very powerful, as it controls for factors that cause variability between subjects. If the matching is effective, the repeated-measures test will yield a smaller P value than an ordinary ANOVA. The repeated-measures test is more powerful because it separates between-subject variability from within-subject variability. If the pairing is ineffective, however, the repeated-measures test can be less powerful because it has fewer degrees of freedom.

Prism tests whether the matching was effective and reports a P value. This P value comes from a chi-square statistic that is computed by comparing the fit of the full mixed effects model to a simpler model without accounting for repeated measures. If this P value is low, you can conclude that the matching was effective. If the P value is high, you can conclude that the matching was not effective and should reconsider your experimental design.

Goodness of fit

 Prism optionally expresses the goodness-of-fit in a few ways. These will only be meaningful to someone who understand mixed effects models deeply. Most scientists will ignore these results or uncheck the option so they don't get reported.

If you don't accept the assumption of sphericity

If you checked the option to not accept the assumption of sphericity, Prism does two things differently.

- It applies the correction of Geisser and Greenhouse. You'll see smaller degrees of freedom, which usually are not integers. The corresponding P value is higher than it would have been without that correction.

- It reports the value of <u>epsilon</u>[400], which is a measure of how badly the data violate the assumption of sphericity.

Multiple comparisons tests and analysis checklist

Learn about <u>multiple comparisons tests after repeated measures ANOVA</u>[436].

Before interpreting the results, <u>review the analysis checklist</u>[181].

5.13.4.5 Interpreting results: Test for trend

.

5.13.4.5.1 Overview: Test for linear trend

What is the test for linear trend?

Prism can test for linear trend as part of the followup testing after one-way (but not two- or three-way) ANOVA. It is a choice on the Multiple Comparisons tab of the parameters dialog for one-way ANOVA.

This test makes sense when the columns represent ordered and equally spaced (or nearly so) groups. For example, the columns might represent age groups, or doses or times. The test for linear trend asks whether the column means increase (or decrease) systematically as the columns go from left to right. Prism does the calculations setting column A to 1, column B to 2 and so on. You cannot assign other values to the columns. If you enter values into the column titles, they will be ignored.

Alternative names are *testing for a linear contrast*, *post-test for trend*, and *test for linear trend.*

Are you sure you don't want to do regression?

Note that ANOVA calculations (except for this trend test) completely ignore the order of the columns. You could randomly scramble the column order, and get exactly the same ANOVA results. So if the columns represent time points or concentrations or anything that can be quantified, think hard about whether ANOVA is the best way to analyze your data. Rather than do ANOVA followed by a test for linear trend, you may want to fit the data with linear (or nonlinear) regression.

After you've entered your data for ANOVA, you'll need to rearrange (transpose) the data if you want to analyze the same data with regression.

1. From the column table used for ANOVA, click Analyze and choose the Transpose analysis.

2. Select that the results should be an XY results table. Choose sequential X values values if the columns are equallyd and you want to match the results o the test for trend. Enter your own X values if the values associated with each column are not equally spaced.

3. From the completed XY table, click Analyze and choose linear regression.

5.13.4.5.2 Results from test for trend

Interpreting the results

Slope

This is the change in the mean value per column as you move one column to the right. In other words, it assumes that the X values corresponding to column order are separated by 1 (we call this the span). Note that Prism 6 and earlier used a span of 1 when there were an odd number of columns, but a span of 2 when there were an even number of columns (this was not made clear in the results). With an even number of columns, therefore, the slope reported by Prism 6 (or earlier) is twice the slope that Prism now reports.

R square

Prism reports two different R^2 in the context of testing for linear trend after ANOVA.

- *The effect size R^2* is the fraction of the total variance accounted for by the linear trend. This was the only R^2 reported by Prism 6 which labeled it simply R^2.

- The *alerting R^2* is the fraction of the variance between group means that is accounted for by the linear trend. Because the variance between group means is always less than the total variance, the alerting R^2 is always higher than the effect size R^2.

P values

Prism reports two P values.

- Test for linear trend. The P value tests the null hypothesis that there is no linear trend between the population means and group order. It answers the question: If there really is no linear trend between column number and column mean, what is the chance that random sampling would result in a slope as far from zero (or further) than you obtained here? If the P value is small, conclude that there is a statistically significant linear trend. As you go from left to right in the data table, the column means tend to get higher (or lower).

- Test for nonlinear trend. If you have four or more groups (data set columns), then Prism also reports a P value testing nonlinear trend. The null hypothesis is that the entire relationship between the column means and column order is linear. A small P value tells you there is also a nonlinear trend.

5.13.4.5.3 How the test for trend works

How it works

The overall ANOVA table partitions the variation among values into a portion that is variation within groups and a portion that is between groups. The test for trend further divides the variation between group means into a portion that is due to a linear relationship between column mean and column order, and the rest that is due to a nonlinear relationship between column mean and column order. Prism computes an

F ratio as the ratio of the mean square for linear trend divided by the mean square within groups, and computes the P value from that.

The test for trend only "sees" the column means and does not "see" the individual values. Since it doesn't look at the raw data, the results don't match linear regression of the raw data (which would require you to transpose the data onto an XY table). Because the method accounts for sample size, the results also don't match linear regression of just column means vs column order either. The calculation Prism does is standard as a followup to ANOVA, but it isn't clear if there is any advantage this test for trend vs. simply computing linear regression on transposed data(3).

If there are any missing values, Prism fits a mixed-effects model[391].

References

1. DG Altman, Practical Statistics for Medical Research IBSN:0412276305

2. SE Maxwell and HD Delaney Designing Experiments and Analyzing Data: A Model Comparison Perspective, Second Edition. byISBN: 0805837183

3. Bradley Huitema, The Analysis of Covariance and Alternatives, pp 344-346, section "Trend analysis versus Regression analysis

5.13.4.6 Analysis checklist: Repeated-measures one way ANOVA

Repeated measures one-way ANOVA compares the means of three or more matched groups. Read elsewhere to learn about choosing a test[406] , and interpreting the results[437].

✔ Was the matching effective?

The whole point of using a repeated-measures test is to control for experimental variability. Some factors you don't control in the experiment will affect all the measurements from one subject equally, so will not affect the difference between the measurements in that subject. By analyzing only the differences, therefore, a matched test controls for some of the sources of scatter.

The matching should be part of the experimental design and not something you do after collecting data. Prism tests the effectiveness of

matching with an F test (distinct from the main F test of differences between columns). If the P value for matching is large (say larger than 0.05), you should question whether it made sense to use a repeated-measures test. Ideally, your choice of whether to use a repeated-measures test should be based not only on this one P value, but also on the experimental design and the results you have seen in other similar experiments.

✓ Are the subjects independent?

The results of repeated-measures ANOVA only make sense when the subjects are independent. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six rows of data, but these were obtained from three animals, with duplicate measurements in each animal. In this case, some factor may affect the measurements from one animal. Since this factor would affect data in two (but not all) rows, the rows (subjects) are not independent.

✓ Is the random variability distributed according to a Gaussian distribution?

Repeated-measures ANOVA assumes that each measurement is the sum of an overall mean, a treatment effect (the average difference between subjects given a particular treatment and the overall mean), an individual effect (the average difference between measurements made in a certain subject and the overall mean) and a random component. Furthermore, it assumes that the random component follows a Gaussian distribution and that the standard deviation does not vary between individuals (rows) or treatments (columns). While this assumption is not too important with large samples, it can be important with small sample sizes. Prism does not test for violations of this assumption.

✓ Is there only one factor?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group, with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments.

Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. Similarly, there

are two factors if you wish to compare the effect of drug treatment at several time points. These data need to be analyzed by two-way ANOVA, also called two-factor ANOVA.

✔ Is the factor "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Type II ANOVA, also known as random-effect ANOVA, assumes that you have randomly selected groups from an infinite (or at least large) number of possible groups, and that you want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment. Type II random-effects ANOVA is rarely used, and Prism does not perform it.

WIth repeated measures, Prism can fit a mixed effects model. This model assumes the differences among subjects (or litters...) is random. But it assumes the factor that defines which column each value is entered into is fixed.

✔ Can you accept the assumption of circularity or sphericity?

Repeated-measures ANOVA assumes that the random error truly is random. A random factor that causes a measurement in one subject to be a bit high (or low) should have no affect on the next measurement in the same subject. This assumption is called *circularity* or *sphericity*. It is closely related to another term you may encounter, *compound symmetry*.

Repeated-measures ANOVA is quite sensitive to violations of the assumption of circularity. If the assumption is violated, the P value will be too low. One way to violate this assumption is to make the repeated measurements in too short a time interval, so that random factors that cause a particular value to be high (or low) don't wash away or dissipate before the next measurement. To avoid violating the assumption, wait long enough between treatments so the subject is essentially the same as before the treatment. When possible, also randomize the order of treatments.

You only have to worry about the assumption of circularity when you perform a repeated-measures experiment, where each row of data represents repeated measurements from a single subject. It is

impossible to violate the assumption with randomized block experiments, where each row of data represents data from a matched set of subjects.

 If you cannot accept the assumption of sphericity, you can specify that on the Parameters dialog. In that case, Prism will take into account possible violations of the assumption (using the method of Geisser and Greenhouse) and report a higher P value.

✔ If any values are missing, was that due to a random event?

 Starting with Prism 8, repeated measures data can be calculated with missing values by fitting a mixed model. But the results can only be interpreted if the reason for the value being missing is random. If a value is missing because it was too high to measure (or too low), then it is not missing randomly. If values are missing because a treatment is toxic, then the values are not randomly missing.

5.13.5 Kruskal-Wallis test

5.13.5.1 Interpreting results: Kruskal-Wallis test

P value

The Kruskal-Wallis test is a nonparametric test that compares three or more unmatched groups. To perform this test, Prism first ranks all the values from low to high, paying no attention to which group each value belongs. The smallest number gets a rank of 1. The largest number gets a rank of N, where N is the total number of values in all the groups. The discrepancies among the rank sums are combined to create a single value called the Kruskal-Wallis statistic (some books refer to this value as H). A large Kruskal-Wallis statistic corresponds to a large discrepancy among rank sums.

The P value answers this question:

> If the groups are sampled from populations with identical distributions, what is the chance that random sampling would result in a sum of ranks as far apart (or more so) as observed in this experiment?

If your samples are small (even if there are ties), Prism calculates an exact P value. If your samples are large, it approximates the P value from a Gaussian approximation (based on the fact that the Kruskal-Wallis statistic H approximates a chi-square distribution. Prism labels the P

value accordingly as *exact* or *approximate*. Here, the term Gaussian has to do with the distribution of sum of ranks and does not imply that your data need to follow a Gaussian distribution. The approximation is quite accurate with large samples and is standard (used by all statistics programs). The exact calculations can be slow with large(ish) data sets or slow(ish) computers. You can cancel the calculations in that case, by clicking the cancel button on the progress dialog. If you cancel computation of the exact P value, Prism will instead show the approximate P value.

If the P value is small, you can reject the idea that the difference is due to random sampling, and you can conclude instead that the populations have different distributions.

If the P value is large, the data do not give you any reason to conclude that the distributions differ. This is not the same as saying that the distributions are the same. Kruskal-Wallis test has little power. In fact, if the total sample size is seven or less, the Kruskal-Wallis test will always give a P value greater than 0.05 no matter how much the groups differ.

Tied values

The Kruskal-Wallis test was developed for data that are measured on a continuous scale. Thus you expect every value you measure to be unique. But occasionally two or more values are the same. When the Kruskal-Wallis calculations convert the values to ranks, these values tie for the same rank, so they both are assigned the average of the two (or more) ranks for which they tie.

Prism uses a standard method to correct for ties when it computes the Kruskal-Wallis statistic.

There is no completely standard method to get a P value from these statistics when there are ties. Prism 6 and later handles ties differently than did prior versions. Prism will compute an exact P value with moderate sample sizes. Earlier versions always computed an approximate P value when there were ties. Therefore, in the presence of ties, Prism 6 and later may report a P value different than that reported by earlier versions of Prism or by other programs.

If your samples are small, Prism calculates an exact P value. If your samples are large, it approximates the P value from the chi-square distribution. The approximation is quite accurate with large samples. With

medium size samples, Prism can take a long time to calculate the exact P value. While it does the calculations, Prism displays a progress dialog and you can press Cancel to interrupt the calculations if an approximate P value is good enough for your purposes. Prism always reports whether the P value was computed exactly or via an approximation.

Dunn's test

Dunn's multiple comparisons test compares the difference in the sum of ranks between two columns with the expected average difference (based on the number of groups and their size).

For each pair of columns, Prism reports the P value as >0.05, <0.05, <0.01, or <0.001. The calculation of the P value takes into account the number of comparisons you are making. If the null hypothesis is true (all data are sampled from populations with identical distributions, so all differences between groups are due to random sampling), then there is a 5% chance that at least one of the post tests will have P<0.05. The 5% chance does not apply to each comparison but rather to the entire family of comparisons.

For more information on the post test, see *Applied Nonparametric Statistics* by WW Daniel, published by PWS-Kent publishing company in 1990 or *Nonparametric Statistics for Behavioral Sciences* by S. Siegel and N. J. Castellan, 1988. The original reference is O.J. Dunn, Technometrics, 5:241-252, 1964.

Prism refers to the post test as the Dunn's post test. Some books and programs simply refer to this test as the post test following a Kruskal-Wallis test, and don't give it an exact name.

Analysis checklist

Before interpreting the results, <u>review the analysis checklist</u>[184].

5.13.5.2 Analysis checklist: Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric test that compares three or more unpaired or unmatched groups.Read elsewhere to learn about <u>choosing a test</u>[406], and <u>interpreting the results</u>[448].

✔ Are the "errors" independent?

The term "error" refers to the difference between each value and the group median. The results of a Kruskal-Wallis test only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have nine values in each of three groups, but these were obtained from two animals in each group (in triplicate). In this case, some factor may cause all three values from one animal to be high or low.

✔ Are the data unpaired?

If the data are paired or matched, then you should consider choosing the Friedman test instead. If the pairing is effective in controlling for experimental variability, the Friedman test will be more powerful than the Kruskal-Wallis test.

✔ Are the data sampled from non-Gaussian populations?

By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions, but there are drawbacks to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to detect a true difference), especially with small sample sizes. Furthermore, Prism (along with most other programs) does not calculate confidence intervals when calculating nonparametric tests. If the distribution is clearly not bell-shaped, consider transforming the values (perhaps to logs or reciprocals) to create a Gaussian distribution and then using ANOVA.

✔ Do you really want to compare medians?

The Kruskal-Wallis test compares the medians of three or more groups. It is possible to have a tiny P value – clear evidence that the population medians are different – even if the distributions overlap considerably.

✔ Are the shapes of the distributions identical?

The Kruskal-Wallis test does not assume that the populations follow Gaussian distributions. But it does assume that the shapes of the distributions are identical. The medians may differ – that is what you are

testing for – but the test assumes that the shapes of the distributions are identical. If two groups have very different distributions, consider transforming the data to make the distributions more similar.

5.13.6 Friedman's test

5.13.6.1 Interpreting results: Friedman test

P value

The Friedman test is a nonparametric test that compares three or more matched or paired groups. The Friedman test first ranks the values in each matched set (each row) from low to high. Each row is ranked separately. It then sums the ranks in each group (column). If the sums are very different, the P value will be small. Prism reports the value of the Friedman statistic, which is calculated from the sums of ranks and the sample sizes. This value goes by several names. Some programs and texts call this value Q or T1 or FM. Others call it chi-square, since its distribution is approximately chi-square so the chi-square distribution is used to compute the P value.

The whole point of using a matched test is to control for experimental variability between subjects, thus increasing the power of the test. Some factors you don't control in the experiment will increase (or decrease) all the measurements in a subject. Since the Friedman test ranks the values in each row, it is not affected by sources of variability that equally affect all values in a row (since that factor won't change the ranks within the row).

The P value answers this question: If the different treatments (columns) really are identical, what is the chance that random sampling would result in sums of ranks as far apart (or more so) as observed in this experiment?

If the P value is small, you can reject the idea that all of the differences between columns are due to random sampling, and conclude instead that at least one of the treatments (columns) differs from the rest. Then look at post test results to see which groups differ from which other groups.

If the P value is large, the data do not give you any reason to conclude that the overall medians differ. This is not the same as saying that the medians are the same. You just have no compelling evidence that they differ. If you have small samples, Friedman's test has little power.

Exact or approximate P value?

With a fairly small table, Prism does an exact calculation. When the table is larger, Prism uses a standard approximation. To decide when to use the approximate method, Prism computes $(T!)^S$ (T factorial to the S power) where T is number of treatments (data sets) and S is the number of subjects (rows).When that value exceeds 10^9, Prism uses the approximate method. For example, if there are 3 treatments and 12 rows, then $(T!)^S$ equals 6^{12}, which equals 2.2×10^9, so Prism uses an approximate method.

The approximate method is sometimes called a Gaussian approximation. The term *Gaussian* has to do with the distribution of sum of ranks, and does not imply that your data need to be sampled from a Gaussian distribution. With medium size samples, Prism can take a long time to calculate the exact P value. You can interrupt the calculations if an approximate P value meets your needs.

The exact method works by examining all possible rearrangements of the values, keeping each value in the same row (same subject, since this is a repeated measures design) but allowing the column (treatment) assignment to vary.

If two or more values (in the same row) have the same value, previous versions of Prism were not able to calculate the exact P value, so Prism computed an approximate P value even with tiny samples. Prism 6 and later can compute an exact P value even in the presence of ties, so only uses an approximation when sample size is fairly large as explained above. This means that with some data sets, Prism will report different results than old versions did.

Dunn's post test

Following Friedman's test, Prism can perform Dunn's post test. For details, see Applied Nonparametric Statistics by WW Daniel, published by PWS-Kent publishing company in 1990 or Nonparametric Statistics for Behavioral Sciences by S Siegel and NJ Castellan, 1988. The original reference is O.J. Dunn, Technometrics, 5:241-252, 1964. Note that some books and programs simply refer to this test as the post test following a Friedman test and don't give it an exact name.

Dunn's post test compares the difference in the sum of ranks between two columns with the expected average difference (based on the number

of groups and their size). For each pair of columns, Prism reports the P value as >0.05, <0.05, <0.01, or < 0.001. The calculation of the P value takes into account the number of comparisons you are making. If the null hypothesis is true (all data are sampled from populations with identical distributions, so all differences between groups are due to random sampling), then there is a 5% chance that at least one of the post tests will have P<0.05. The 5% chance does not apply to each comparison but rather to the entire family of comparisons.

5.13.6.2 Analysis checklist: Friedman's test

Friedman's test is a nonparametric test that compares three or more paired groups.

✔ Was the matching effective?

The whole point of using a repeated-measures test is to control for experimental variability. Some factors you don't control in the experiment will affect all the measurements from one subject equally, so they will not affect the difference between the measurements in that subject. By analyzing only the differences, therefore, a matched test controls for some of the sources of scatter.

The matching should be part of the experimental design and not something you do after collecting data. Prism does not test the adequacy of matching with the Friedman test.

✔ Are the subjects (rows) independent?

The results of a Friedman test only make sense when the subjects (rows) are independent – that no random factor has affected values in more than one row. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six rows of data obtained from three animals in duplicate. In this case, some random factor may cause all the values from one animal to be high or low. Since this factor would affect two of the rows (but not the other four), the rows are not independent.

✔ Are the data clearly sampled from non-Gaussian populations?

By selecting a nonparametric test, you have avoided assuming that the data were sampled from Gaussian distributions, but there are drawbacks

to using a nonparametric test. If the populations really are Gaussian, the nonparametric tests have less power (are less likely to give you a small P value), especially with small sample sizes. Furthermore, Prism (along with most other programs) does not calculate confidence intervals when calculating nonparametric tests. If the distribution is clearly not bell-shaped, consider transforming the values (perhaps to logs or reciprocals) to create a Gaussian distribution and then using repeated-measures ANOVA.

✓ Is there only one factor?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group, with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments.

Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. Similarly, there are two factors if you wish to compare the effect of drug treatment at several time points. These data need to be analyzed by two-way ANOVA, also called two-factor ANOVA.

5.14   Two-way ANOVA

Two-way ANOVA, also called two-factor ANOVA, determines how a response is affected by two factors. For example, you might measure a response to three different drugs in both men and women. Drug treatment is one factor and gender is the other. Is the response affected by drug? By gender? Are the two intertwined? These are the kinds of questions that two-way ANOVA answers.

5.14.1 How to: Two-way ANOVA

Two-way ANOVA, also called two-factor ANOVA, determines how a response is affected by two factors. For example, you might measure a response to three different drugs in both men and women. In this example, drug treatment is one factor and gender is the other.

5.14.1.1 Notes of caution for statistical novices

Our goal with Prism has always been to make basic biostatistics very accessible and easy. Two-way ANOVA is pushing the limits of "basic biostatistics". Multiple comparisons after two-way ANOVA stretch this definition even more. If you haven't taken the time to really understand two-way ANOVA, it is quite easy to be misled by the results. Beware!

- Two-way ANOVA is not a topic that is easy to master. In addition to reading textbooks, also consider getting help.from someone with more experience.

- Prism also offers to fit a mixed model to repeated measures data. Understanding this fully is even more complicated.

- Before getting lost in the many choices for multiple comparisons, first articulate clearly the scientific goals of the study. Don't articulate your goals in terms of ANOVA (looking for interactions) and avoid the word "significant" which often leads to muddled thinking. Figure out what you really want to know. Then figure out the best statistical approach to getting the answer.

- In Prism, the two-way ANOVA analysis can be used when, as the name suggests, there are two factors. None, one, or both of the factors can be repeated measures.

5.14.1.2 Deciding which factor defines rows and which defines columns?

Two ways to enter data on a Grouped table

In a grouped table, each data set (column) represents a different level of one factor, and each row represents a different level of the other factor.

You need to decide which factor to define by rows, and which to define by data set columns. For example, if you are comparing men and women at three time points, there are two ways to organize the data:

	X Labels	A		B	
		Men		Women	
	X	A:Y1	A:Y2	B:Y1	B:Y2
1	Before	123	132	143	145
2	During	143	154	141	156
3	After	162	156	175	164

X Labels	A		B		C	
	Before		During		After	
X	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1 Men	123	132	143	154	162	156
2 Women	143	145	141	156	175	164

Your choice affects the appearance of graphs

The ANOVA results will be identical no matter which way you enter the data. But the choice defines how the graph will appear. If you enter data as shown in the first approach above, men and women will appear in bars of different color, with three bars of each color representing the three time points (left graph below). If you enter data using the second approach shown above, there will be one bar color and fill for Before, another for During, and another for After (right graph below). Men and Women appear as two bars of identical appearance.
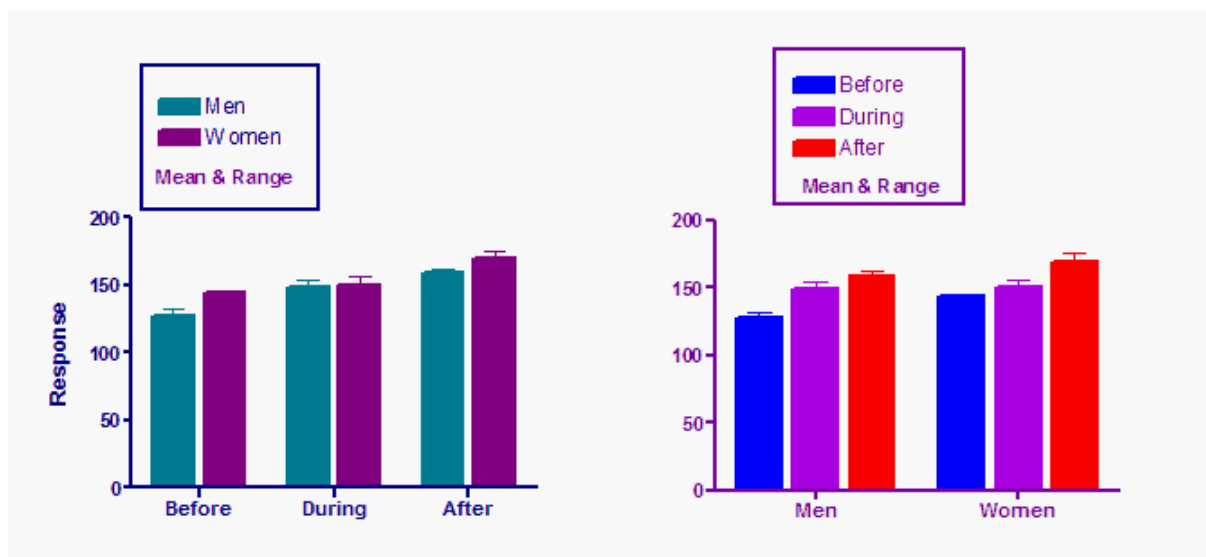
Use the transpose analysis to change your mind

What happens if after entering and analyzing your data using one of the choices above, you then realize you wish you had done it the other way? You don't need to reenter your data. Instead use Prism's transpose analysis, and then create a graph from the results table.

5.14.1.3 Entering data for two-way ANOVA (not repeated measures)

Groups are defined by rows and columns

Prism organizes data for two-way ANOVA differently than do most other programs.

Prism does not use grouping variables. Instead, use rows and columns to designate the different groups (levels) of each factor. Each data set (column) represents a different level of one factor, and each row represents a different level of the other factor.

Setting up the data table

From the Welcome (or New Data Table and Graph) dialog, choose the Grouped tab.

Entering raw data

Create a Grouped table with enough subcolumns to hold the maximum number of replicates you have.

Table format: Grouped		Group A — Wild-type cells					Group B — GPP5 cell line					Group C — GPP7 cell line				
	⊠	A:Y1	A:Y2	A:Y3	A:Y4	A:Y5	B:Y1	B:Y2	B:Y3	B:Y4	B:Y5	C:Y1	C:Y2	C:Y3	C:Y4	C:Y5
1	Serum starved	34	36	41		43	98	87	95	99	88	77	89	97	66	76
2	Normal culture	23	19	26	29	25	32	29	26	33	30	33	45	35	46	54

In the example above, the two rows encode the two levels of one factor (serum starved vs. normal culture) and the three data set columns encode the three levels of the other factor (cell line).

Entering averaged data

If you have already averaged your replicates in another program, you can choose to enter and plot the mean and SD (or SEM) and n. If your data has more than 256 replicates, this is the only way to enter data into Prism for two-way ANOVA.
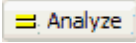
Table format: Grouped		Group A — Wild-type cells			Group B — GPP5 cell line			Group C — GPP7 cell line		
	⊠	Mean	SD	N	Mean	SD	N	Mean	SD	N
1	Serum starved	38.50	4.20	4	93.40	5.59	5	81.00	12.10	5
2	Normal culture	24.40	3.71	5	30.00	2.74	5	42.60	8.62	5

Note that repeated measures ANOVA requires raw data. This is not a quirk of Prism, but fundamental to repeated measures analyses. So if you enter mean, sample size and SD or SEM, you'll only be able to do ordinary (not repeated measures) ANOVA.

Entering single values

If you only have one value for each condition, create a Grouped table and choose to enter a single Y value (no subcolumns). In this case, Prism will only be able to compute ordinary (not repeated measures) ANOVA, and will assume that there is no interaction between the row and column factor. It cannot test for interaction without replicates, so simply assumes there is none. This may or may not be a reasonable assumption for your situation.

Run the ANOVA

1. From the data table, click ⟌ Analyze on the toolbar.

2. Choose Two-way ANOVA from the list of grouped analyses.

3. Since the data do not have repeated measures[461], don't check either option on the first tab (RM Design[469]). Without repeated measures, there are no choices available on the second (RM Analysis[472]) tab.

4. On the third (Factor Names[474]) tab, optionally name the grouping variables that define the rows and columns. For the example shown above, you might label the columns as "cell line" and the rows as "serum"

5. On the fourth (Multiple Comparisons[475]) tab, select your goal (if any) for multiple comparisons.

6. On the fifth (Options[480]) tab, choose the details of which multiple comparisons test you want.

7. On the last (Residuals[485]) tab, choose if and how you want to plot residuals, and if you want to test them for normality and equal variability (homoscedasticity).

Prism can't do two-way ANOVA with huge data sets

Prism cannot run ordinary (not repeated measures) two-way ANOVA with huge data sets and presents an message telling you so. How huge is huge? Details here.

5.14.1.4 Entering repeated measures data

The term *repeated-measures* refers to an experiment that collects multiple measurements of the dependent variable from each participant. The repeat can be across time (eg. pre/post), across different conditions (eg. high and low temperature), or across space (eg. left knee and right knee). The key issue is that the same participant has multiple responses.

The analysis of repeated measures data is identical to the analysis of randomized block experiments that use paired or matched subjects. In this case, think of the pair or match itself as the "participant." Prism can calculate repeated-measures two-way ANOVA when either one of the factors are repeated or matched (mixed effects) or when both factors are. In other words, Prism can handle these three situations with its two-way ANOVA analysis:

• Two between-subject variables (neither factor is repeated measures)

• One between-subject variable and one within subject variable

• Two within-subject variables (both factors are repeated measures)

One data table can correspond to four experimental designs

Prism uses a unique way to enter data. You use rows and columns to designate the different groups (levels) of each factor. Each data set (column) represents a different level of one factor, and each row represents a different level of the other factor. You need to decide which factor is defined by rows, and which by columns. Your choice will not affect the ANOVA results, but the choice is important[457] as it affects the appearance of graphs.

Table format: Grouped		A Control		B Treated	
	☒	A:Y1	A:Y2	B:Y1	B:Y2
1	Baseline	23	24	28	31
2	Dose 1	34	41	41	54
3	Dose 2	43	47	56	60

The table above shows example data testing the effects of three doses of drugs in control and treated animals.

These data could have come from four distinct experimental designs.

Not repeated measures

Perhaps one of the simplest explanations for the table above is that this experiment was performed using twelve animals, with one measurement per animal. In this case, each cell would represent a different animal. Of course, this would not be a repeated measures experiment.

Another situation could be that the experiment that generated the data above was performed using six animals. Let's say each treatment combination has only one animal and measurements were made in duplicate (two measurements per animal). In this situation, the value in row 1, column A:Y1 (23) came from the same animal as the value in row 1, column A:Y2 (24). These duplicate values for each animal are known as pseudo-replicates. You must be careful when analyzing data from this type of experiment with Prism due to the problem of pseudo-replication. Performing a standard two-way ANOVA in this scenario (with no data treatment) would provide misleading results. Prism would think that there were twelve total animals (as described above) - not six. Because of this, Prism would assume that the variation between pseudo-replicates represented variation *between* animals. However, in this experimental design, variation between pseudo-replicates actually represents variation *within* animals. This is an important problem, and there are a number of ways to deal with it (such as averaging the pseudo-replicates before analyzing the data). If your experimental design includes pseudo-replicates, be sure to understand how they differ from standard replicates and take appropriate action to account for them in your analysis.

Matched values are spread across a row

The experiment was done with six animals, two for each dose. The control values were measured first in all six animals. Then you applied a

treatment to all the animals and made the measurement again. In the table above, the value at row 1, column A, Y1 (23) came from the same animal as the value at row 1, column B, Y1 (28). The matching is by row.

Matched values are stacked into a subcolumn

The experiment was done with four animals. First each animal was exposed to a treatment (or placebo). After measuring the baseline data (dose=zero), you inject the first dose and make the measurement again. Then inject the second dose and measure again. The values in the first Y1 column (23, 34, and 43) were repeated measurements from the same animal. The other three subcolumns came from three other animals. The matching was by column.

When each subcolumn represents an animal or participant, Prism lets you label the subcolumns on the data table. Simply double click on the subcolumn heading (A:Y1).

Repeated measures in both factors

The experiment was done with two animals. First you measured the baseline (control, zero dose). Then you injected dose 1 and made the next measurement, then dose 2 and measured again. Then you gave the animal the experimental treatment, waited an appropriate period of time, and made the three measurements again. Finally, you repeated the experiment with another animal (Y2). So a single animal provided data from both Y1 subcolumns (23, 34, 43 and 28, 41, 56).

When each subcolumn represents an animal or participant, Prism lets you label the subcolumns on the data table. Simply double click on the subcolumn heading (A:Y1).

When do you specify which design applies to this experiment?

The example above shows that one grouped data set can represent four different experimental designs. You do not distinguish these designs when creating the data table. The data table doesn't "know" whether or not the data are repeated measures. You should take into account experimental design when choosing how to graph the data. And you must take it into account when performing two-way ANOVA. On the first tab of the two-way ANOVA dialog[469], you'll designate the experimental design.
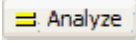
Lingo: "Repeated measures" vs. "randomized block" experiments

The term **repeated measures** is appropriate when you made repeated measurements from each subject.

Some experiments involve matching but not repeated measurements. The term **randomized-block** describes these kinds of experiments. For example, imagine that the three rows were three different cell lines. All the Y1 data came from one experiment, and all the Y2 data came from another experiment performed a month later. The value at row 1, column A, Y1 (23) and the value at row 1, column B, Y1 (28) came from the same experiment (same cell passage, same reagents). The matching is by row.

Randomized block data are analyzed identically to repeated-measures data. Prism always uses the term *repeated measures*, so you should choose repeated measures analyses when your experiment follows a randomized block design.

Run the ANOVA

1. From the data table, click [≡ Analyze] on the toolbar.

2. Choose Two-way ANOVA from the list of grouped analyses.

3. On the first tab (RM Design[469]) choose your experimental design.

4. On the second (RM Analysis[472]) tab, choose whether you want to run repeated measures ANOVA or a mixed model.

5. On the third (Factor Names[474]) tab, optionally name the grouping variables that define the rows and columns. For the example shown above, you might label the columns as "treatment" and the rows as "dose". Each matched set might be named "animal".

6. On the fourth (Multiple Comparisons[475]) tab, select your goal (if any) for multiple comparisons.

7. On the fifth (Options[480]) tab, choose the details of which multiple comparisons test you want.

8. On the last (Residuals[485]) tab, choose if and how you want to plot residuals, and if you want to test them for normality and equal variability (homoscedasticity).

Graphing data for two-way repeated measures ANOVA

When graphing data with repeated measures, choose a graph that connects symbols representing individual points so the graph shows the nature of the data. To do this, choose the "individual values" tab of the New Graph dialog, and then choose one of the two graph types on the right. The example below shows the graph to choose if matched values are spread across a row. The next graph to the right is for when matched values are stacked. If you have matching in both direction, you need to choose which to plot as no graph (in Prism anyway) can show matching in both grouping variables.
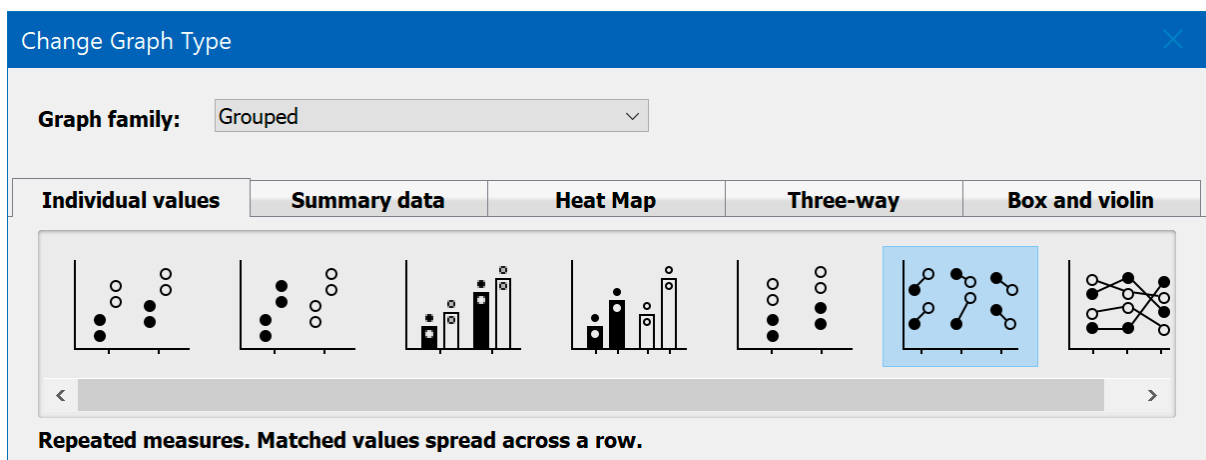
Prism can't run repeated measures two-way ANOVA with huge data sets

Prism cannot run repeated measures two-way ANOVA with huge data sets and presents an message telling you so. How huge is huge? Really big!

Each column represents a different time point, so matched values are spread across a row

Prism cannot analyze tables where: (number of rows)^2 * number of columns * number of subcolumns * (number of columns + number of subcolumns) > 268,435,456

Each row represents a different time point, so matched values are stacked into a subcolumn

Prism cannot analyze tables where: number of rows * (number of columns)^2 * number of subcolumns * (number of rows + number of subcolumns) > 268,435,456

Repeated measures by both factors

Prism can handle any table you can enter.

5.14.1.5 Missing values and two-way ANOVA

Missing values with ordinary (not repeated measures) ANOVA

Table format: Grouped		A					B				
		Wild-type cells					GPP5 cell line				
	✕	A:Y1	A:Y2	A:Y3	A:Y4	A:Y5	B:Y1	B:Y2	B:Y3	B:Y4	B:Y5
1	Serum starved	34	36	41		43	98	87	95	99	88
2	Normal culture	23	19	26	29	25	32	29	26	33	30

Note that one value is blank. It is fine to have some missing values, but you must have at least one value in each row for each data set.

The following table cannot be analyzed by two-way ANOVA because there are no data for treated women. Every cell must have at least one value.

Table format: Two-way		A			B		
		Control			Treated		
	✕	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Men	33.0	34.6		54.2		56.9
2	Women	65.3	59.4	54.3			

If you are entering mean, SD (or SEM) and n, it is fine if n is not always the same, but ANOVA won't work if you leave n blank or enter zero.

Missing values with repeated measures ANOVA

There are two different situation regarding missing values and repeated measures two-way ANOVA:

• Prism can compute repeated measures two-way ANOVA fine if treatment groups have different numbers of participants, but each participant (experiment, litter, ...) has data at each repeat.

• Prism (starting with Prism 8) can also do the equivalent of repeated measures two-way ANOVA if values at some repeats are missing, so long as not too many points are missing and they are missing completely at random. If some values are missing because they would have been too large to measure, then any results would be meaningless because the missing values actually represent the largest values in the

table. But if a few values are missing completely at random, Prism uses a can fit a [mixed effects model](#)[469].

5.14.1.6 Point of confusion: ANOVA with a quantitative factor

ANOVA with a quantitative factor

Two-way ANOVA is sometimes used when one of the factors is quantitative, such as when comparing time courses or dose response curves. In these situations one of the factors is dose or time that you have set to one of several (or many) discrete values.

ANOVA pays no attention to the order of your time points (or doses). Think about that. The whole point of your experiment may have been to look at a trend or at a dose-response relationship. But the ANOVA calculations completely ignores the order of the time points or doses. If you randomly scramble the time points or doses, two-way ANOVA would report identical results. ANOVA treats different time points, or different doses, exactly the same way it would treat different drugs, different genotypes, or different countries.

Since ANOVA ignores the entire point of the experiment when one of the factors is quantitative, consider using alternative (regression) approaches. In some cases, you don't have enough data or enough theory to fit a curve, so ANOVA might be a reasonable analysis.

Interpreting P values with a quantitative factor

Let's imagine you compare two treatments at six time points.

The two-way ANOVA will report three P values:

- One P value tests the null hypothesis that time has no effect on the outcome. In many situatons, you know that the outcome changes over time. That's why you did a time course. Since you expect a small P value for the effect of time, it doesn't tell you much.

- Another P value tests the null hypothesis that the treatment makes no difference, on average. This hypothesis might be worth testing in some situaitons. But in many situations, you expect no difference at early time points, and only care about differences at late time points. In these situations, testing the average treatment effect may not be so helpful.

- The third P value tests for interaction. The null hypothesis is that any difference between treatments is identical at all time points. But if you collect data at time zero, or at early time points, you don't expect to find any difference then. Your experiment really is designed to ask about later time points. In this situation, you expect an interaction, so finding a small P value for interaction does not help you understand your data. It is even less useful if the difference between treatments gets larger at some time points and then gets smaller at later time points.

Interpreting multiple comparisons tests with a quantitative factor

What about multiple comparisons tests?

Some scientists like to ask which is the lowest dose (or time) at which the change in response is statistically significant. Multiple comparisons tests can give you the answer, but the answer depends on sample size. Run more subjects, or more doses or time points for each curve, and the answer will change. With a large enough sample size (at each dose or time point), you will find a statistically significant (but biologically trivial) effect with a tiny dose or at a very early time point. With fewer replicates at each dose or time point, you won't see statistical significance until a larger dose or later time point. Since asking for the smallest dose that gives a "signficant" effect of does not ask a fundamental question about the system, the results may not be helpful helpful.

If you want to know the minimally effective dose, consider finding the minimum dose that causes an effect bigger than some threshold you set based on physiology (or some other scientific context). For example, find the minimum dose that raises the pulse rate by more than 10 beats per minute. That approach can lead to useful answers. Searching for the smallest dose that leads to a "significant" effect does not.

If you look at all the multiple comparisons tests (and not just ask which is the lowest dose or time point that gives a 'significant' effect), you can get results that make no sense. You might find that the difference is statistically significant at time points 3, 5, 6 and 9 but not at time points 1, 2, 4, 7, 8 and 10. How do you interpret that? Knowing at which doses or time points the treatment had a statistically significant rarely helps you understand the biology of the system and rarely helps you design new experiments.

Alternatives to two-way ANOVA

What is the alternative to two-way ANOVA?

If you have a repeated measures design, consider using this alternative to ANOVA, which Will G Hopkins calls within-subject modeling.

First, quantify the data for each subject in some biologically meaningful way. Perhaps this would be the area under the curve. Perhaps the peak level. Perhaps the time to peak. Perhaps you can fit a curve with nonlinear regression and determine a rate constant or a slope.

Now take these values (the areas or rate constants...) and compare between groups of subjects using a t test (if two treatments) or one-way ANOVA (if three or more). Unlike two-way ANOVA, this kind of analysis follows the scientific logic of the experiment, and so leads to results that are understandable and can lead you to the next step (designing a better experiment).

If you don't have a repeated measures design, you can still fit a curve for each treatment. Then compare slopes, or EC50s, or lag times as part of the linear or nonlinear regression.

Think hard about what your scientific goals are, and try to find a way to make the statistical testing match the scientific goals. In many cases, you'll find a better approach than using two-way ANOVA.

Test for trend

One of the choices for multiple comparisons tests following one-way ANOVA is a test for linear trend. This test, of course, does consider the order of the treatments. Other programs (but not Prism) offer polynomial post tests, which also take into account the treatment order.

5.14.1.7 Repeated measures design tab: Two-way ANOVA

Repeated measures defined

Repeated measures means that the data are matched. Here are some examples:

- You measure a dependent variable in each subject several times, perhaps before, during and after an intervention.

- You recruit subjects as matched groups, matched for variables such as age, ethnic group, and disease severity.

- You run a laboratory experiment several times, each time with several treatments handled in parallel. Since you anticipate experiment-to-experiment variability, you want to analyze the data in such a way that each experiment is treated as a matched set. Although you don't intend it, responses could be more similar to each other within an experiment than across experiments due to external factors like more humidity one day than another, or unintentional practice effects for the experimenter.

Matching should not be based on the variable you are comparing. If you are comparing blood pressures in three groups, it is OK to match based on age or zip code, but it is not OK to match based on blood pressure.

The term *repeated measures* applies strictly only when you give treatments repeatedly to one subject (the first example above). The other two examples are called *randomized block experiments* (each set of subjects is called a block, and you randomly assign treatments within each block). The analyses are identical for repeated measures and randomized block experiments, and Prism always uses the term *repeated measures.*

Matching by which factor(s)?

If your data are matched, choose which of the two factors are repeated measures, or if both factors are repeated measures. If one factor is repeated measures and the other is not, this analysis is also called mixed effects model ANOVA.

Choose carefully, as the results can be very misleading if you make a choice that doesn't correspond to the experimental design. The choices are:

 No matching. Use regular two-way ANOVA (not repeated measures).

Each column represents a different repeat, so matched values are spread across a row.

Each row represents a different time point, so matched values are stacked into a subcolumn.

Repeated measures by both factors.

Take your time when making this choice. We find that too many people think "repeated measures" but don't take the time to carefully specify which factor is repeated or if both factors are repeated. If your choice here doesn't match your actual experimental design, the results will probably be incorrect.

Assume sphericity?

With two-way repeated measures ANOVA, choose whether to assume sphericity[395]. If you do not assume sphericity, Prism uses the the Greenhouse-Geisser correction and calculates epsilon[400].

The assumption of sphericity states that the variance of the differences between treatment A and B equals the variance of the difference between A and C, which equals the variance of the differences between A and D, which equals the variance of the differences between B and D... (Or, when repeated measures are stacked in subcolumns, that the variance of the differences between treatment 1 and 2 equals the variance of the difference between 1 and 3...) Like all statistical assumptions, this assumption pertains to the populations from which the data were sampled, and not just to these particular data sets.

Note that if the factor with repeated measures has only two levels, then there is no reason to be concerned about violations of sphericity. For example if each subject is measured before and after a treatment, and there are four different treatments, there would be no need to worry about sphericity, since the repeated measures factor only has two levels (before and after). If you ask for the Greenhouse-Geisser correction with only two levels of the repeated measures factor, the results will be identical to what they would have been if you hadn't chosen that option and the reported value of epsilon will be 1.0000000.

5.14.1.8 Repeated measures analysis tab: Two-way ANOVA

Two ways to analyze repeated measures data

Prism can analyze repeated measures data in two ways:

• Repeated measures ANOVA

• Fitting a mixed effects model. This analysis works fine even when there are some missing values. The results will only be meaningful, of course, if the values are missing for random reasons. For example, those results won't be helpful or meaningful if the values are missing because those participants were very sick, or those values were too high to measure (or too low to measure). Fitting a mixed model with missing values only makes sense when there is zero association between the treatments or

time-points and the reason why some values are missing.

In general, fitting a mixed effects model is a much more versatile method. As implemented in Prism 8, the two are completely equivalent when there are no missing values. But the mixed effects model method can also fit data with missing values.

Analyze using which method

The repeated measures tab of the ANOVA dialog (same for one-, two- and three-way data) gives you three choices:

- Use repeated measures ANOVA always. If there are missing values, no results will be reported. This matches what Prism 7 and earlier did. Prism is not "smart enough" to remove all data for a participant with missing values, but you could exclude all those values and rerun the ANOVA.

- Fit a mixed effects model always. This will make all analyses be consistent, whether or not there are missing values. If there are no missing values, the key results will be the same as repeated measures ANOVA but the results will be presented in an format unfamiliar to those used to repeated measures ANOVA.

- Report the fit to a mixed effects model only when there are missing values, when repeated measures ANOVA is impossible. When there are no missing values, report the familiar repeated measures ANOVA results.

How to fit the mixed effects model any random factor is zero or negative

The whole point of repeated measures or mixed model analyses is that you have multiple response measurements on the same subject or when individuals are matched (twins or litters), so need to account for any correlation among multiple responses from the same subject. Mixed model analysis does this by estimating variances between subjects. In a simple mixed model, where only one variable is repeated, it's possible that this correlation for the particular data in your study is zero or even negative (of course it is impossible for a variance to be negative, but it can happen with mixed effects models). You are given two choices for what Prism should do when this happens:

- Analyze as usual. If there are no missing values, this will match repeated

measures ANOVA.

•Remove the subject factor from the model and refit. This approach will have one more degree of freedom and thus have a bit more power. This approach is better but means the mixed model results may not match repeated measures ANOVA results.

In more complicated models, where there is more than one repeated measures variable, there are even more possible variance estimates (generally interactions with subject), and any of those could turn out to be zero or negative. It's best to take these out as leaving them in can make the results unstable.

Defaults for future analyses

Check an option at the bottom of the Repeated Measures tab to make your choices the default for future analyses. Your default will apply to one-, two- and three-way ANOVA.

5.14.1.9 Factor names tab: Two-way ANOVA

Factor names

Entering descriptive names for the two factors will make it easier to interpret your results.

If the two rows are labeled Male and Female, enter "Sex" or "Gender" as the name for the row factor. If two columns are labeled Control and Treated, enter "Treatment" as the name for the column factor.

If you have repeated measures, you can change the name of the repeated measurement block from "subject" to "litter" or "experiment" or whatever you want.

5.14.1.10 Multiple comparisons tab: Two-way ANOVA
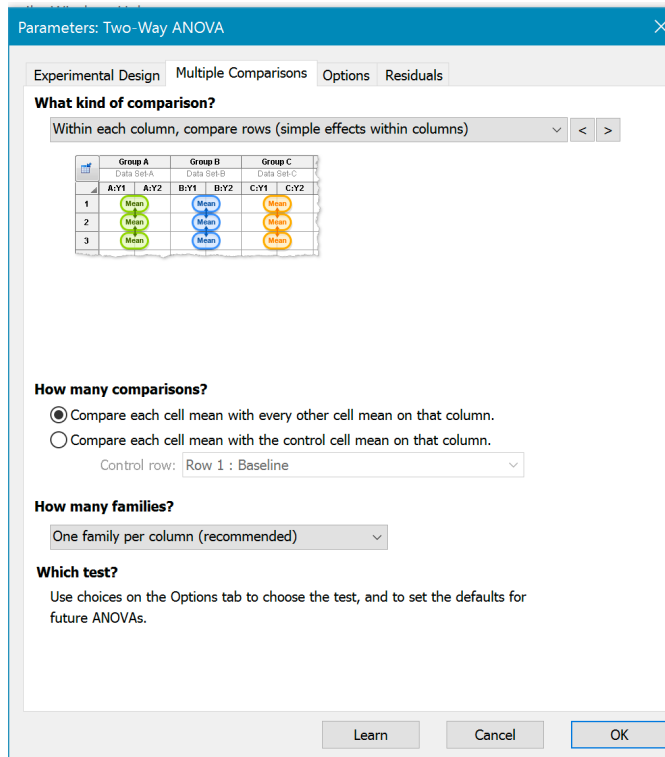
Choosing multiple comparisons for two-way ANOVA is not straightforward. Make this choice carefully, and after learning about two-way ANOVA. Consider getting help.

Which kind of comparison?

This is the most important decision. You need to pick a multiple comparison scheme that matches your scientific goal. The pictures, shown below and on the dialog, are probably more helpful than the explanations

The choices of comparisons (in the drop down) depend on the number of rows and columns in your data set.

Expand all	Collapse all

⊟ Compare each cell mean with the other cell mean in that row

This was the only choice in early versions of Prism, and is probably the most useful kind of multiple comparisons. This choice is available only if

there are exactly two columns. For each row, therefore, there are two cell means, and Prism compares these.

☐ Compare each cell mean with the other mean in that column

This choice is available only if there are exactly two rows.

☐ Simple effects. Within each row, compare columns.

This choice is only available if you have three or more columns of data. Within each row, Prism does multiple comparisons between cell means.

For each row, compare the mean of side-by-side replicates of one column with another. This only makes sense, so the choice is only available, only when there are three or more columns. You must decide whether each row becomes its own family of comparisons, or whether all the comparisons are defined to be one family.

☐ Simple effects. Within each column, compare rows.

Within each column, compare the mean of side by side replicates of one row with the mean of other rows. This choice is only available when you have three or more rows. You must decide whether each column becomes its own family of comparisons, or whether all the comparisons are defined to be one family.

☐ Main column effects

Testing for main column effects involves computing the mean of each data set column, and comparing those means. This makes sense (so the choice is available) only if there are data in three or more data set columns. If your data table has only two data set columns, then the main ANOVA computations give a P value for the effect of the variable that defines the columns, and no multiple comparison testing for column effects makes sense.

Main row effects

Testing for main row effects involves computing the mean value for each row, and then comparing those means. It only makes sense, so the choice is only available, when there are three or more rows. If your data table has only two rows, then the main ANOVA computations give a P value for the effect of the variable that defines the rows, and no multiple comparison testing for row effects makes sense.

Compare cell means regardless of rows and columns

Compare each cell means with every other cell mean, paying no attention to which row and column each cell mean is part of. This choice is not available when one factor is repeated measures, but is available when both factors are repeated measures.
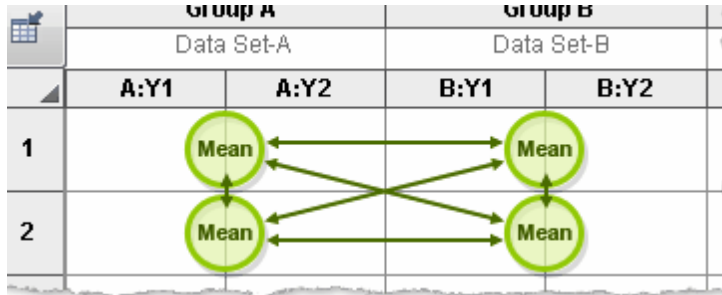
How many comparisons?

Do you want to compare each mean (in the set) with each other mean? Or only compare each mean to the first, control, mean? The latter approach makes fewer comparisons, so has more power. The choice should be based on experimental design and the scientific questions you are asking.

How many families? (Applies to simple effects only.)

Multiple comparisons take into account the number of comparisons in the family of comparisons. The significance level (alpha) applies to the entire family of comparisons. Similarly, the confidence level (usually 95%) applies to the entire family of intervals, and the multiplicity adjusted P values adjust each P value based on the number of comparisons in a family.

If you choose to look at Simple effects (defined above), the definition of *family* is not obvious, and Prism offers two choices:

- One family for all comparisons. With this choice, there is always one family of comparisons for all rows (or all columns).This approach has less power, because it applies a stricter correction for multiple comparisons. This makes sense because there are more comparisons in the family.

- One family per column (or per row). Define the comparisons for each column (or each row) to be its own family of comparisons. With this choice, there are fewer comparisons per family (but more families), so comparisons have more power. We recommend this choice unless you have strong reason to consider all the comparisons to be one family.

The results page will repeat your choices, so it is clear how to interpret the results.

Prism 5.04 and 5.0d use the first definition of family (and do not offer you a choice of the other definition). If you wish to compare results with Prism 5, note this bug in releases of Prism 5 prior to 5.04 (Windows) and 5.0d (Mac).

5.14.1.11 Options tab: Multiple comparisons: Two-way ANOVA

Correct for multiple comparisons using statistical hypothesis testing

Some of these methods let you compute confidence intervals and multiplicity adjusted P values, and some don't. We recommend one of the tests that compute confidence intervals and multiplicity adjusted P values for two reasons:

- Confidence intervals[59] are much easier for most to interpret than statements about statistical significance.

- [Multiplicity adjusted P values](#)[129] provide more information that simply knowing if a difference has been deemed statistically significant or not.

Recommended methods

The list of tests available depends on the goal you specified on the second tab. We recommend these tests because they can compute confidence intervals and multiplicity adjusted P values

- If you are comparing every row (or column) mean with every other row (or column) mean, we recommend the [Tukey test](#)[136].

- If you are comparing a control row (or column) mean with the other row (or column) means, we suggest the [Dunnett's test](#)[136].

- If you are comparing a bunch of independent comparisons, we recommend the [Sidak](#)[131] method, which is very similar to Bonferroni but has a tiny bit more power.

Other available methods

The Bonferroni and Sidak methods are offered for compatibility with other programs, but we see no advantages from choosing these tests.

If you don't care about seeing and reporting confidence intervals, you can gain a bit more power by choosing the [Holm-Šídák test](#)[133]. It is is more powerful than the Tukey method for comparing all pairs of means (3). That means that with some data sets, the Holm-Šídák method can find a statistically significant difference where the Tukey method cannot. Glantz says that Holm's test ought to have more power than Dunnett's test, but this has not (to his knowledge) been explored in depth(2)

Prism also offers the *Newman-Keuls test* (when comparing each mean with each other mean) for historical reasons (so files made with old versions of Prism will open) but we suggest you avoid it because it does not maintain the family-wise error rate at the specified level(1). In some cases, the chance of a Type I error can be greater than the alpha level you specified.

Correct for multiple comparisons by controlling the False Discovery Rate

Prism offers [three methods](#)[142] to control the [false discovery rate](#)[140]. All decide which (if any) comparisons to label as "discoveries" and do so in a way that controls the false discovery rate to be less than a value Q you enter.

The FDR approach is not often used as a followup test to ANOVA, but there is no good reason for that.

Don't correct for multiple comparisons. Each comparison stands alone.

If you choose this approach, Prism will perform [Fisher's Least Significant Difference (LSD) test](#)[127].

This approach (Fisher's LSD) has much more power to detect differences. But it is more likely to falsely conclude that a difference is statistically significant. When you correct for multiple comparisons (which Fisher's LSD does not do), the significance threshold (usually 5% or 0.05) applies to the entire family of comparisons. With Fisher's LSD, that threshold applies separately to each comparison.

Only use the Fisher's LSD approach if you have a very good reason, and are careful to explain what you did when you report the results.

Multiple comparisons options

Swap direction of comparisons

The only affect of this option is to change the sign of all reported differences between means. A difference of 2.3 will be -2.3 if the option is checked. A difference of -3.4 will be 3.4 if you check the option. It is purely a personal preference that depends on how you think about the data.

Report multiplicity adjusted P value for each comparison

If you choose the Bonferroni, Tukey or Dunnett multiple comparisons test, Prism can also report [multiplicity adjusted P values](#)[129]. If you check this option, Prism reports an adjusted P value for each comparison. These calculations take into account not only the two groups being compared, but the total number groups (data set columns) in the ANOVA, and the data in all the groups.

The multiplicity adjusted P value is the smallest significance threshold (alpha) for the entire family of comparisons at which a particular comparison would be (just barely) declared to be "statistically significant".

Until recently, multiplicity adjusted P values have not been commonly reported. If you choose to ask Prism to compute these values, take the time to be sure you understand what they mean. If you include these values in publications or presentations, be sure to explain what they are.

Confidence and significance level (or desired FDR)

By tradition, confidence intervals are computed for 95% confidence and statistical significance is defined using an alpha of 0.05. Prism lets you choose other values. If you choose to control the FDR, select a value for Q (in percent). If you set Q to 5%, you expect up to 5% of the "discoveries" to be false positives.

References

1. SA Glantz, Primer of Biostatistics, sixth edition, ISBN= 978-0071435093.

2. MA Seaman, JR Levin and RC Serlin, Psychological Bulletin 110:577-586, 1991.

5.14.1.12 Summary of multiple comparisons available (two-way)

Two rows or two columns

If the data table only has two columns (or two rows), then Prism compares the two values at each row (column) and uses either the Bonferroni or Holm method to correct for multiple comparisons.

More than two rows or columns

If there are more than two rows and columns, then you first need to choose how to define each family of comparisons[475] in the Experimental Design tab. Then you need to choose how to correct for multiple comparisons within each family by making choices in the Options tab. The choices for two-way ANOVA depend on two decisions:

- Your goal. Which comparisons do you want to make?

- Do you want confidence intervals (CI) included in your results? Not all multiple comparisons tests can compute confidence intervals.

Goal	CI?	Method
Compare every mean to every other mean	Yes	Tukey (preferred) Bonferroni
	No	Holm (preferred) Newman-Keuls
Compare every mean to a control mean	Yes	Dunnettt
	No	Holm

5.14.1.13 Options tab: Other choices

Graphing options

If you chose a multiple comparison method that computes confidence intervals (Tukey, Dunnett, etc.) Prism can plot these confidence intervals. This is a great way to show the key results of ANOVA. Consider showing a graph like this with your results.

95% Confidence Intervals (Sidak)
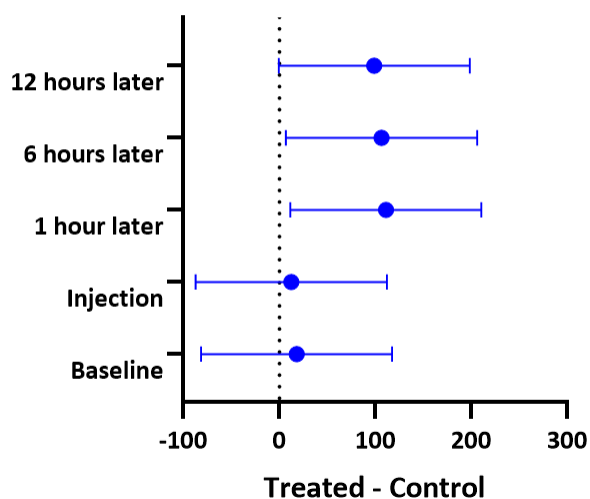
Additional results

- Report the results in paragraph (narrative) form, as well as the usual tabular results. This is not available when you fit a mixed model.

- Create a table of means showing the mean (or predicted LS mean if there are missing values) for each cell, as well as for each row and column, and grand means.

Output

Choose how you want P values reported, and how many significant digits you need for the other results.

5.14.1.14 Residual tab: Two-way ANOVA

Why residuals?

Prism 8 introduced the ability to plot residual plots with ANOVA, provided that you entered raw data and not averaged data as mean, n and SD or SEM.

Many scientists thing of residual as values that are obtained with regression. But ANOVA is really regression in disguise. It fits a model. One of the assumptions of ANOVA is that the residuals from that model are sampled from a Gaussian distribution. A residual plot helps you assess this assumption.

Which graph to create?

Prism can make three kinds of residual plots.

- Residual plot. The X axis is the predicted value. The Y axis is the residual. This lets you spot residuals that are much larger or smaller than the rest.

- Homoscedasticity plot. The X axis is the predicted value. The Y axis is the absolute value of the residual.This lets you check whether larger values are associated with bigger residuals (large absolute value).

- QQ plot. The X axis is the actual residual. The Y axis is the predicted residual, computed from the percentile of the residual (among all residuals) and assuming sampling from a Gaussian distribution. ANOVA assumes a Gaussian distribution of residuals, and this graph lets you check that assumption.

For ordinary ANOVA, the predicted values (used for residual and homoscedasticity plots) are simply the means of the replicates in a cell. For repeated measures ANOVA, the predicted value also takes into account the subject mean.

Diagnostics for residuals

- Are the residuals clustered or heteroscedastic? ANOVA assumes each sample was randomly drawn from populations with the same standard deviation. Prism can test this assumption with two tests. The Brown-Forsythe test and the Barlett test. Both these tests compute a P value designed to answer this question:If the populations really have the same standard deviations, what is the chance that you'd randomly select samples whose standard deviations are as different from one another (or more different) as they are in your experiment?

- Are the residuals Gaussian? Prism runs four normality tests on the residuals. The residuals from all groups are pooled and then entered into one set of normality tests.

5.14.1.15Q&A: Two-way ANOVA

I know the mean, SD (or SEM) and sample size for each group. Which tests can I run?

You can enter data as mean, SD (or SEM) and n, and Prism can compute two-way ANOVA. It is not possible to compute repeated measures ANOVA without access to the raw data.

I only know the group means, and don't have the raw data and don't know their SD or SEM. Can I run ANOVA?

Yes, two-way ANOVA is possible if you only have one value for each condition (no subcolumns). In this case, Prism will only be able to compute ordinary (not repeated measures) ANOVA, and will assume that

there is no interaction between the row and column factor. It cannot test for interaction without replicates, so simply assumes there is none. This may or may not be a reasonable assumption for your situation

I want to compare three groups. The dependent variable has two possible outcomes. How can I compare the groups?

Not with ANOVA. Enter your data into a [contingency table](#)[545] and analyze with a [chi-square test](#)[547].

What does 'two-way' mean?

Two-way ANOVA, also called *two-factor* ANOVA, determines how a continuous response is affected by two factors. For example, you might measure a response to three different drugs at two time points. The two factors are drug and time.

If you measure response to three different drugs at two time points with subjects from two age ranges, then you have three factors: drug, time and age so wouldwant to choose three-way ANOVA.

What does 'repeated measures' mean? How is it different than 'randomized block'?

The term *repeated-measures* strictly applies only when you give treatments repeatedly to each subject, and the term *randomized block* is used when you randomly assign treatments within each group (block) of matched subjects. The analyses are identical for repeated-measures and randomized block experiments, and Prism always uses the term repeated-measures.

What is a mixed-model design?

In the context of two-way ANOVA, a mixed-model is one where one factor is repeated measures and the other is not. Prism can analyze data where neither factor is repeated measures, one of the two factors is repeated measures, or when both factors are repeated measures. Prism 5 could not analyze data where both factors are repeated measures.

My two factors are genotype and treatment, and each animal was measured at several time points. Can Prism analyze this kind of "two way" ANOVA?

No. Your experimental design has three factors: genotype, treatment and time. If you wanted to use ANOVA, you'd need to use three-way ANOVA.

How does Prism define "cell mean"?

Prism uses the term "cell" to mean the data on one row for one data set. There may be multiple subcolumns side-by-side, so the cell mean is the average of the values in these subcolumns.

How many rows/columns can Prism handle?

Prism cannot run two-way ANOVA with huge data sets and presents an message telling you so. How huge is huge? [Details here](#).

5.14.2   Intepreting ordinary (not repeated measures) two-way ANOVA results

5.14.2.1   Interpreting results: Two-way ANOVA

Two-way ANOVA determines how a response is affected by two factors. For example, you might measure a response to three different drugs in both men and women.

Source of variation

Two-way ANOVA divides the total variability among values into four components. Prism tabulates the percentage of the variability due to interaction between the row and column factor, the percentage due to the row factor, and the percentage due to the column factor. The remainder of the variation is among replicates (also called residual variation).

These values (% of total variation) are called *standard omega squared* by Sheskin (equations 27.51 - 27.53, and R^2 by Maxwell and Delaney (page 295). Others call these values *eta squared* or the correlation ratio.

ANOVA table

The ANOVA table breaks down the overall variability between measurements (expressed as the sum of squares) into four components:

- Interactions between row and column. These are differences between rows that are not the same at each column, equivalent to variation between columns that is not the same at each row.

- Variability among columns.

- Variability among rows.

- Residual or error. Variation among replicates not related to systematic differences between rows and columns.

The ANOVA table shows how the sum of squares is partitioned into the four components. Most scientists will skip these results, which are not especially informative unless you have studied statistics in depth. For each component, the table shows sum-of-squares, degrees of freedom, mean square, and the F ratio. Each F ratio is the ratio of the mean-square value for that source of variation to the residual mean square (with repeated-measures ANOVA, the denominator of one F ratio is the mean square for matching rather than residual mean square). If the null hypothesis is true, the F ratio is likely to be close to 1.0. If the null hypothesis is not true, the F ratio is likely to be greater than 1.0. The F ratios are not very informative by themselves, but are used to determine P values.

P values

Two-way ANOVA partitions the overall variance of the outcome variable into three components, plus a residual (or error) term. Therefore it computes P values that test three null hypotheses (repeated measures two-way ANOVA adds yet another P value).

Interaction P value

The null hypothesis is that there is no interaction between columns (data sets) and rows. More precisely, the null hypothesis states that any systematic differences between columns are the same for each row and that any systematic differences between rows are the same for each column. Often the test of interaction is the most important of the three tests. If columns represent drugs and rows represent gender, then the null hypothesis is that the differences between the drugs are consistent for men and women.

The P value answers this question:

If the null hypothesis is true, what is the chance of randomly sampling subjects and ending up with as much (or more) interaction than you have observed?

The graph on the left below shows no interaction. The treatment has about the same effect in males and females. The graph on the right, in contrast, shows a huge interaction. the effect of the treatment is completely different in males (treatment increases the concentration) and females (where the treatment decreases the concentration). In this example, the treatment effect goes in the opposite direction for males and females. But the test for interaction does not test whether the effect goes in different directions. It tests whether the average treatment effect is the same for each row (each gender, for this example).
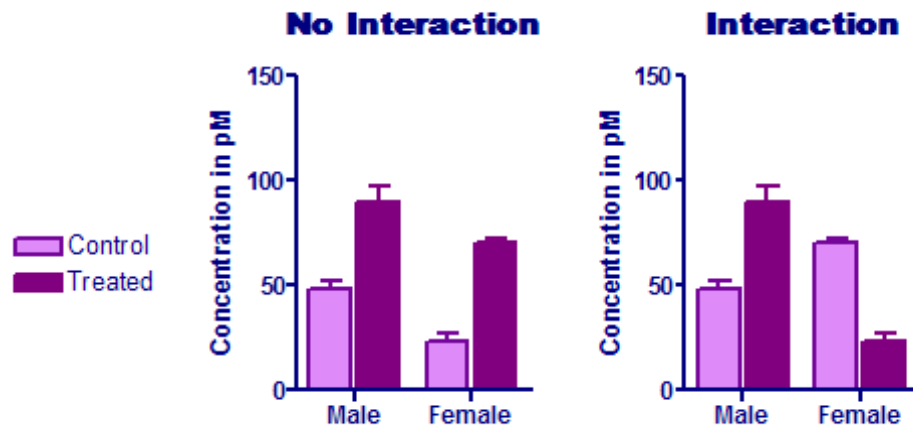
Testing for interaction requires that you enter replicate values or mean and SD (or SEM) and N. If you entered only a single value for each row/column pair, Prism assumes that there is no interaction, and continues with the other calculations. Depending on your experimental design, this assumption may or may not make sense.

If the test for interaction leads to statistically significant results, you probably won't learn anything of interest from the other two P values. In the example above, a statistically significant interaction means that the effect of the treatment (difference between treated and control) differs between males and females. In this case, it is really impossible to interpret the overall P value testing the null hypothesis that the treatment has no effect at all. Instead focus on the multiple comparison post tests. Is the effect statistically significant in males? How about females?

Column factor P value

The null hypothesis is that the mean of each column (totally ignoring the rows) is the same in the overall population, and that all differences we see between column means are due to chance. In the example graphed above, results for control and treated were entered in different columns (with males and females being entered in different rows). The null hypothesis is that the treatment was ineffective so control and treated values differ only due to chance. The P value answers this question: If the null hypothesis is true, what is the chance of randomly obtaining column means as different (or more so) than you have observed?

In the example shown in the left graph above, the P value for the column factor (treatment) is 0.0002. The treatment has an effect that is statistically significant.

In the example shown in the right graph above, the P value for the column factor (treatment) is very high (0.54). On average, the treatment effect is indistinguishable from random variation. But this P value is not meaningful in this example. Since the interaction P value is low, you know that the effect of the treatment is not the same at each row (each gender, for this example). In fact, for this example, the treatment has opposite effects in males and females. Accordingly, asking about the overall, average, treatment effect doesn't make any sense.

Row factor P value

The null hypothesis is that the mean of each row (totally ignoring the columns) is the same in the overall population, and that all differences we see between row means are due to chance. In the example above, the rows represent gender, so the null hypothesis is that the mean response is the same for men and women. The P value answers this question: If the null hypothesis is true, what is the chance of randomly obtaining row means as different (or more so) than you have observed?

In both examples above, the P value for the row factor (gender) is very low.

Data Summary Table

This small section on the results sheet provides a summary of:

- The number of columns (Column Factor)

- The number of rows (Row Factor)

- The number of values

Note that using the [Factor Names tab]474 to enter descriptive names for the Column Factor and Row Factor will display the entered descriptive names in the Data Summary Table. This feature was added for ordinary two-way ANOVA in Prism 8.2.

Multiple comparisons tests

Note that the three P values produced by two-way ANOVA are not corrected for the three comparisons. It would seem logical to do so, but this is not traditionally (ever?) done in ANOVA.

Multiple comparisons testing is one of the most confusing topics in statistics. Since Prism offers nearly the same multiple comparisons tests for one-way ANOVA and two-way ANOVA, we have [consolidated the information on multiple comparisons]117.

References

David J. Sheskin. Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition IBSN:1584884401.

5.14.2.2 Two-way ANOVA without replication

Big picture

Two-way ANOVA requires replication to do a sensible analysis. Before discussing the problem of ANOVA without replication, lets consider an example with replication (so no issue).

Table format: Grouped		Group A Control		Group B Treated	
	☒	A:1	A:2	B:1	B:2
1	Male	34	38	124	104
2	Female	145	163	241	222

The two values in side-by-side subcolumns are results from different animals. There are eight values so eight animals. Prism tests whether the effect of the treatment is the same in both genders -- whether there is an interaction. The P value for this example is quite high, so there is no evidence of interaction. The ANOVA table (not shown here) shows that

Prism is able to separately compute the sum-of-squares for interaction and the sum-of-squares between replicate animals (residual variation).

Source of Variation	% of total variation	P value	P value summary	Significant?
Interaction	0.0003096	0.9774	ns	No
Gender	68.69	0.0001	***	Yes
Treatment	29.95	0.0007	***	Yes

There are two situations where lack of replication is a problem.

Problem 1. No replication. Only one value for each set of conditions.

If you only have one value for each row/column cell, Prism cannot test for interaction so assumes there is none. For the example, this assumption is that the effect of treatment is the same in males and females. The P values are tiny for both the effects of Gender and the effect of Treatment.

Table format: Grouped	Group A Control	Group B Treated
1 Male	34	124
2 Female	145	241

Source of Variation	% of total variation	P value
Gender	60.02	0.0167
Treatment	39.94	0.0205

But what if the treatment really didn't work in females and the only value for treated females (241) value is high simply because that particular animal is an outlier? If that were true, there would have been an interaction -- the treatment would have worked better in one gender than the other. With only a single value in each cell, Prism can't test for this, and assumes no interaction.

The assumption of no interaction may or may not be a reasonable assumption for your situation. It is far better to include replication in your experimental design so Prism can test for interaction.

Problem 2. Pseudo-replication

This second problem is a bigger problem, because ANOVA gives results that can be misleading.

Given the same experimental goals as before, we'll do the experiment a bit differently using a different experimental design. We'll use only one animal for each row/column combination, and measure the response twice in each animal. Note there are only four animals tested in total. but there are eight values because each animal is measured twice. The data look exactly like the data on the top of the page with eight animals.

Table format: Grouped	Group A Control		Group B Treated	
◢ ☒	A:1	A:2	B:1	B:2
1 Male	34	38	124	104
2 Female	145	163	241	222

There are no data to assess animal-to-animal variation, so it is impossible to distinguish variation among animals from variation due to interaction. Accordingly, this is a poor experimental design.

There are three approaches to analyzing such data:

Assume no variation among animals (what Prism does)

When you analyze these data in Prism with two-way ANOVA, Prism is mixed up. It does the full ANOVA, so reports the same results shown above for the data from eight animals. It assumes that the variation between duplicate values tells you about variation among animals. It can't tell if the variation seen between treatment combinations is due to differences between animals or the effects of the treatment.

For example, the two values for treated females could be large because that one animal is a bit different than the others, quite apart from the effect of treatment. So the results only make sense if you assume that the variation seen in the duplicate measurements is the same as the variation between animals. Be aware, however, that when your experiment contains pseudo-replicates, this assumption often is not valid.

Assume no interaction and ignore pseudo replication (Prism can do with an extra step)

The second approach is to assume there is no interaction and therefore all variation not due to row and column effects is due to differences among animals. Prism doesn't offer this option directly, but you can run this analysis by following these steps:

1. Run the analysis: Row means with SD or SEM.

2. Choose to only compute the mean, without error bars.

3. From the resulting table, run two-way ANOVA.

Row stats	A Control	B Treated
×		
1 Male	36.000	114.000
2 Female	154.000	231.500

Source of Variation	% of total variation	P value	P value summary	Significant?
Gender	69.64	0.0014	**	Yes
Treatment	30.36	0.0020	**	Yes

Since there is no replication in the data table being analyzed, Prism assumes no interaction.

With this example, the two ways of analyzing the data give similar results. But with other examples, the results could be quite different. The two approaches use very different assumptions, so it is worth thinking through carefully.

Assume no interaction and fit a mixed effects model that accounts for pseudo replication (not available in Prism)

The best way to analyze these data would be to fit a mixed model that assumes no interaction but properly accounts for the replicated measurements in each animal. This approach is preferred because it analyzes all the data without making extra assumptions. However, Prism does not offer this analysis.

5.14.2.3 Graphing tips: Two-way ANOVA
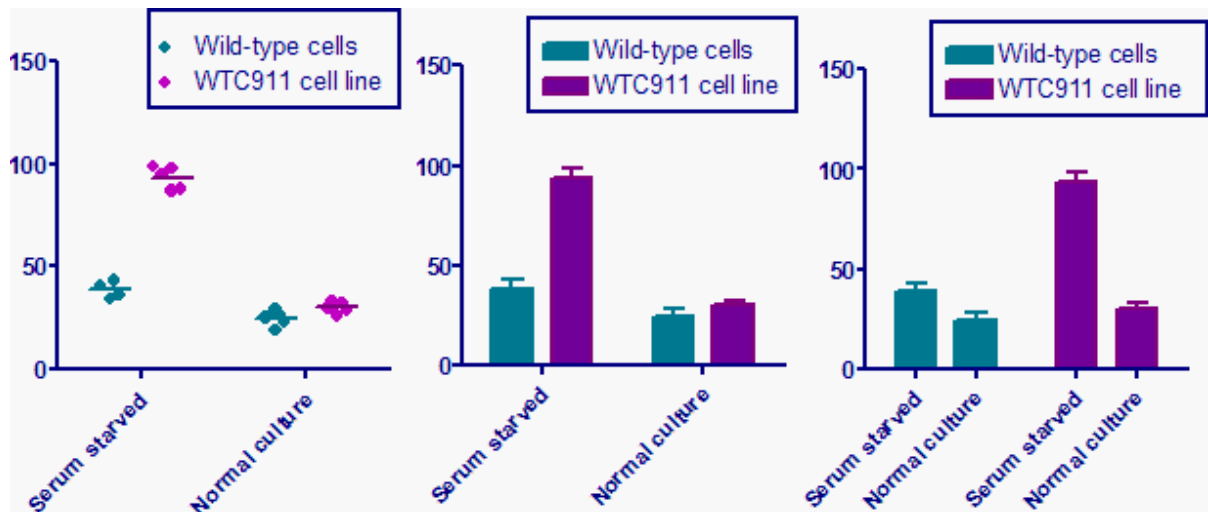
The graph above shows three ways to plot the sample data for two-way ANOVA.

The graphs on the left and middle interleave the data sets. This is set on the second tab of the Format Graphs dialog. In this case, the data sets are defined by the figure legend, and the groups (rows) are defined by the labels on the X axis.

The graph on the right has the data sets grouped. In this graph, the labels on the X axis show the row title -- one per bar. Use the "number format" choice in the Format Axes dialog to change this to Column titles -- one per set of bars. With this choice, there wouldn't be much point in also having the legend shown in the box, and you would need to define the side by side bars ("serum starved" vs "normal culture" for this example) in the figure legend.

The graph on the left has the appearance set as a column dot plot. The other two graphs have the appearance set as bars with error bars plotted from the mean and SD. I prefer the column dot plot as it shows all the data, without taking up more space and without being harder to interpret. If you have a huge number of replicates, use a violin plot instead.

Don't forget to include in the figure legend whether the error bars are SD or SEM or something different.

5.14.2.4 Beware of using multiple comparisons tests to compare dose-response curves or time courses

Does it make sense to use ANOVA multiple comparison tests to compare two dose-response curves at every dose (or two time course curves at every time point)?

No.

Two-way ANOVA can be used to compare two dose-response or time-course curves. The problem with this approach is that ANOVA treats the different doses (or time points) the same way it treats different species or different drugs. The fact that the different doses (or times) are sequential or numerical is ignored by ANOVA. You could randomly scramble the doses or times, and still get the same ANOVA results.

If you don't have enough data or enough theory to fit a curve, ANOVA might be a reasonable first-step in comparing curves. You get one P value testing the null hypothesis that all doses lead to the same effect, another P value testing the null hypothesis that all (both) treatments are indistinguishable, and a third testing whether there is interaction -- whether the difference between treatments is consistent at all doses. The first P value will always be tiny, and not very informative (of course the treatment does something). The second P value is the one you probably care most about, since it asks about differences between the two curves.

It is tempting to then run multiple comparison tests at each dose (or each time point) asking whether the difference between treatment groups is statistically significant. I don't see how these multiple comparison tests provide useful information. If you have two distinct dose-response curves, you expect to see tiny differences at low doses and large differences at intermediate doses. Does running a multiple comparison test at each dose help you understand your system? Does it help you design better experiments? I think the answer to both questions is almost always no.

What's the alternative? Use nonlinear regression to ask a focused question. In this case, use nonlinear regression to quantify the fold shift in the EC50 and its confidence interval, and to compute a P value that tests the null hypothesis that there was no shift. Details here.

Does it make sense to ask what is the lowest dose that produces a statistically significant difference?

No.

Some people want to focus on the low doses and ask: What is the lowest dose that produces a statistically significant difference between the two treatments? The term "significant" often clouds clear thinking, so let's translate that question to: What is the lowest dose where the data convince me that the difference between the two curves is due to the treatments and not due to chance? The answer depends, in part, on how many replicates you run at each dose. You could make that lowest-significant-dose be lower just by running more replicates. I don't see how helps you understand your system better, or how it helps you design better experiments.

The simulated data below demonstrate this point. Both graphs were simulated using a four-parameter variable slope dose response curve with the same parameters and same amount of random scatter. The graph on the left had three data points per dose (triplicates). The graph on the right had 24 replicates per dose.
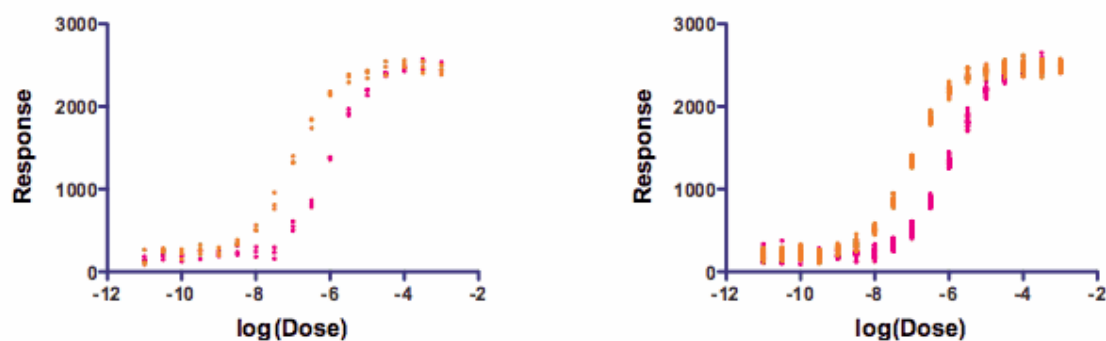
Prism file.

The data were analyzed with two-way ANOVA and the Bonferroni multiple comparison test.

For the graph on the left, the difference between the two data sets first became statistically significant (alpha =0.05 applied to the family of comparisons using Bonferroni) when the log(concentration) was -8.

In contrast, for the graph on the right, the difference first became statistically significant when the log(concentration) was -9. Concentrations between those two values (between 1nM and 10nM) caused a statistically significant effect in the right graph, but not the left.

I ran the simulations a few times, and the results were consistent, so this is not just a quirk of random numbers. Instead, it demonstrates that using more replicates allows smaller differences to be detected as "statistically significant".

By changing the experimental design, we could change the answer to the question: What is the lowest concentration where the response of the two drugs is statistically distinct? That suggests the question is not one worth asking.

5.14.2.5 How Prism computes two-way ANOVA

Two-way ANOVA calculations are quite standard, and these comments only discuss some of the ambiguities.

Model I (fixed effects) vs. Model II (random effects) ANOVA

To understand the difference between fixed and random factors, consider an example of comparing responses in three species at three times. If you were interested in those three particular species, then species is considered to be a fixed factor. It would be a random factor if you were interested in differences between species in general, and you randomly selected those three species. Time is considered to be a fixed factor if you chose time points to span the interval you are interested in. Time would be a random factor if you picked those three time points at random. Since this is not likely, time is almost always considered to be a fixed factor.

When both row and column variables are fixed factors, the analysis is called Model I ANOVA. When both row and column variables are random factors, the analysis is called Model II ANOVA. When one is random and one is fixed, it is termed mixed effects (Model III) ANOVA. With no repeated measures, Prism calculates only Model I two-way ANOVA. Since most experiments deal with fixed-factor variables, this is rarely a limitation.

Missing values

If some values are missing, two-way ANOVA calculations are challenging. David Lane gives a very clear explanation of the challenge of missing values in two-way ANOVA in his online text.

Prism uses the method detailed by Glantz and Slinker (1). This method converts the ANOVA problem to a multiple regression problem and then displays the results as ANOVA. Prism performs multiple regression three times — each time presenting columns, rows, and interaction to the multiple regression procedure in a different order. Although it calculates each sum-of-squares three times, Prism only displays the sum-of-squares for the factor entered last into the multiple regression equation. These are called Type III sum-of-squares. This article explains the difference between Type I, II and III sum-of-square. Type II sum-of-squares assumes no interaction. Type I and III differ only when there are missing values.

Prism cannot perform repeated-measures two-way ANOVA if any values are missing. It is OK to have different numbers of numbers of subjects in each group, so long as you have complete data (at each time point or dose) for each subject.

Data entered as mean, n and SD (or SEM)

If your data are balanced (same sample size for each condition), you'll get the same results if you enter raw data, or if you enter mean, SD (or SEM), and n. If your data are unbalanced, it is impossible to calculate precise results from data entered as mean, SD (or SEM), and n. Instead, Prism uses a simpler method called *analysis of unweighted means.* This method is detailed in LD Fisher and G vanBelle, *Biostatistics*, John Wiley, 1993. If sample size is the same in all groups, and in some other special cases, this simpler method gives exactly the same results as obtained by analysis of the raw data.

If you sample sizes are not all the same, these results will only be approximately correct. If your data are almost balanced (just one or a few missing values), the approximation is a good one. When data are unbalanced, you should enter individual replicates whenever possible and avoid entering mean, n and SD or SEM.

David Lane also discusses the method of unweighted means in his online text.

Single values without replicates

Prism can perform two-way ANOVA even if you have entered only a single replicate for each column/row pair. This kind of data does not let you test for interaction between rows and columns (random variability and interaction can't be distinguished unless you measure replicates). Instead, Prism assumes that there is no interaction and only tests for row and column effects. If this assumption is not valid, then the P values for row and column effects won't be meaningful.

Reference

SA Glantz and BK Slinker, *Primer of Applied Regression and Analysis of Variance*, McGraw-Hill, 1990.

5.14.2.6 Table of means (or predicted means)

Starting with Prism 8, Prism can report a table of means as part of two-way ANOVA. Choose to create this table in the Options tab of the two-way ANOVA dialog.

If there are no missing values

If the data are balanced (no missing values) you'll see the label "Means" at the top of this table which shows the cell means, the row means, the column means, and the grand mean. You could easily compute the values yourself.

If there are any missing values

If the data are unbalanced (there are any missing values), this table is labeled "Predicted means (LS Means)" for the cell, row, column and grand means. These predicted means, also known as "LSmeans", may not match the ones you compute by hand.

For ANOVA without repeated measures, LS means are easy to compute. Say you have a row with these values and want to compute the row mean.

Group A			Group B		
Control			Treated		
A:1	A:2	A:3	B:1	B:2	B:3
3	4	5	6	4	

There are two ways to compute the row mean. One way would be to add up all five values and divide by five. That value (4.4) is not what Prism reports. Prism averages the three values in the first column (4), then averages the two values in the second column (5), and finally averages those two values, reporting 4.5. This is the LS mean.

Cell, row, column and grand SD

Prism also reports the SD of each cell, which is straightforward. It also reports the SD for each row and column, and SD for the whole table (the grand SD). These are computed by first averaging the replicates in each cell and then computing the SD of those cell means. If a row has three data set columns, each with five subcolumns for replicates, the n for computing the SD is 3, not 15.

5.14.2.7 When there are only two rows

Big picture

Two-way ANOVA can be done with any number of rows in the data table, each representing different levels in one of the grouping variables, but is often done with only two rows. When there are two rows, Prism (starting with version 8) reports the difference between the two row means along with the 95% confidence interval for that difference.

When there is little interaction between the row and column factor, the overall difference between row means (with its confidence interval) may be a useful way to present the data. If there is strong interaction, then the difference between row means varies among columns, so the overall difference between row means may not be useful.

Calculation details

- If there are no missing values, the two row means are computed as the mean of the values in all the columns in that row. If there are missing values, the two row means are computed as the mean of the predicted means[501] for all the columns in that row.

- The difference between row means is computed as the mean of Row 1 minus the mean of Row 2.

- The standard error of the difference is computed. The equation below is used where MSresidual is the mean square of the residuals reported in the ANOVA table, and Na and Nb are the two sample sizes. If row 1 has

four columns and entry of data in triplicate with one missing value, Na would equal 11. When there are missing values, the computation of the SEdifference is more complicated.

$$SE_{difference} = \sqrt{MS_{residual}\left(\frac{1}{Na} + \frac{1}{Nb}\right)}$$

- The margin of error equals SEdifference multiplied by the critical value of the t distribution, using Na+Nb-2 degrees of freedom.

- The confidence interval equals the difference between means plus or minus the margin of error.

5.14.2.8 When there are two rows and two columns

Big picture

If there are only two rows and two columns, Prism (starting with version 8) reports the interaction confidence interval. The interaction is defined as (A1 - B1) - (A2 -B2), where each of those values is the mean of replicates in one of the four cells.

Calculation details ANOVA

- The standard error of the difference is computed using the equation below, where MSresiduals is the mean square of the residuals reported in the ANOVA table, and the four n values are the four sample sizes.

$$SE_{difference} = \sqrt{MS_{residuals}\left(\frac{1}{n_{A1}} + \frac{1}{n_{A2}} + \frac{1}{n_{B1}} + \frac{1}{n_{B2}}\right)}$$

- The margin of error equals SEdifference multiplied by the critical value of the t distribution, using the residual degrees of freedom reported in the ANOVA table.

- The confidence interval equals the difference between means plus or minus the margin of error.

Calculation details mixed model

In this case the computation of SE(differences) is [more complicated](#)[401]. General idea is using of respective contrast that defines this kind of comparison ((A1-B1) - (A2-B2)) and estimated covariance matrix.

5.14.2.9 Analysis checklist: Two-way ANOVA

Two-way ANOVA, also called two-factor ANOVA, determines how a response is affected by two factors. For example, you might measure a response to three different drugs in both men and women. In this example, drug treatment is one factor and gender is the other. Read elsewhere to learn about [choosing a test](#)[456], and [interpreting the results.](#)[488]

✓ Are the populations distributed according to a Gaussian distribution?

Two-way ANOVA assumes that your replicates are sampled from Gaussian distributions. While this assumption is not too important with large samples, it is important with small sample sizes, especially with unequal sample sizes. Prism does not test for violations of this assumption. If you really don't think your data are sampled from a Gaussian distribution (and no transform will make the distribution Gaussian), you should consider performing nonparametric two-way ANOVA. Prism does not offer this test.

ANOVA also assumes that all sets of replicates have the same SD overall, and that any differences between SDs are due to random sampling.

✓ Are the data unmatched?

Standard two-way ANOVA works by comparing the differences among group means with the pooled standard deviations of the groups. If the data are matched, then you should choose repeated-measures ANOVA instead. If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA.

Are the "errors" independent?

The term "error" refers to the difference between each value and the mean of all the replicates. The results of two-way ANOVA only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six replicates, but these were obtained from two animals in triplicate. In this case, some factor may cause all values from one animal to be high or low.

Do you really want to compare means?

Two-way ANOVA compares the means. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

Are there two factors?

One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group with a drug treatment group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments. Prism has a separate analysis for one-way ANOVA.

Some experiments involve more than two factors. For example, you might compare three different drugs in men and women at four time points. There are three factors in that experiment: drug treatment, gender and time. These data need to be analyzed by three-way ANOVA, also called three-factor ANOVA. Prism does not perform three-way ANOVA.

Are both factors "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Different calculations are needed if you randomly selected groups from an infinite (or at least large) number of possible

groups, and want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment.

5.14.3 Interpreting repeated measures two-way ANOVA results

5.14.3.1 Interpreting results: Repeated measures two-way ANOVA

Note there is a separate page for interpreting the fit of a mixed model.[512]

Are you sure that ANOVA is the best analysis?

Before interpreting the ANOVA results, first do a reality check. If one of the factors is a quantitative factor like time or dose, consider alternatives to ANOVA[467]. If one of the factors in ANOVA is dose (say 0, 10, 20 and 50 mg) or time (say 0, 10, 20, 30, 60 minutes), ANOVA treats these doses or time points just like it teats different species or different drugs, totally ignoring the fact that doses or time points are ordered.

Interpreting P values from repeated measures two-way ANOVA

When interpreting the results of two-way ANOVA, most of the considerations are the same whether or not you have repeated measures. So read the general page on interpreting two-way ANOVA results[488] first. Also read the general page on the assumption of sphericity[395], and assessing violations of that assumption with epsilon[400].

Repeated measures ANOVA has one additional row in the ANOVA table, "Subjects (matching)". This row quantifies how much of all the variation among the values is due to differences between subjects. The corresponding P value tests the null hypothesis that the subjects are all the same. If the P value is small, this shows you have justification for choosing repeated measures ANOVA. If the P value is high, then you may question the decision to use repeated measures ANOVA in future experiments like this one.

How the repeated measures ANOVA is calculated

Prism computes repeated-measures two-way ANOVA calculations using the standard method explained especially well in Glantz and Slinker (1).

If you have data with repeated measures in both factors, Prism uses methods from Chapter 12 of Maxwell and Delaney (2)

If you do not assume sphericity, Prism uses the the Greenhouse-Geisser correction and calculates epsilon[400]. If your repeated measures factor has only two levels, then the concept of sphericity doesn't apply. The results will be the same whether or not you chose to assume sphericity and the value of epsilon will be 1.00000.

Multiple comparisons tests

Multiple comparisons testing is one of the most confusing topics in statistics. Since Prism offers nearly the same multiple comparisons tests for one-way ANOVA and two-way ANOVA, we have consolidated the information on multiple comparisons[117].

Multiple comparisons after two-way repeated measures ANOVA can be computed in two ways.

Prism always computes the multiple comparison tests using a pooled error term (see page 583 of Maxwell and Delaney, 2). If only one factor is repeated measures, the number of degrees of freedom equals (n-1)(a-1) where n is the number of subjects and a is the number of levels of the repeated measures factor. If both factors are repeated measures, the number of degrees of freedom equals (n-1)(a-1)(b-1) where n is the number of subjects, a is the number of levels one factor, and b is the number of levels of the other factor. Another way to look at this is n is the number of subcolumns, a is the number of rows, and b is the number of data set columns. This extra power comes by an extra assumption that for every comparison you make, in the overall population from which the data were sampled the variation is the same for all those comparisons.

Some programs compute separate error term for each comparison. These comparisons have only n-1 degrees of freedom, so the confidence intervals are wider and the adjusted P values are higher. This approach does not assume that the variance is the same for all comparisons.

Reference

1. SA Glantz and BK Slinker, *Primer of Applied Regression and Analysis of Variance*, McGraw-Hill, second edition, 2000.

2. SE Maxwell and HD Delaney. *Designing Experiments and Analyzing Data*, second edition. Laurence Erlbaum, 2004.

5.14.3.2 ANOVA table in two ways RM ANOVA

Example data sets

To create the examples below, I entered data with two rows, three columns, and three side-by-side replicates per cell. There were no missing values, so 18 values were entered in all.

I analyzed the data four ways: assuming no repeated measures, assuming repeated measures with matched values stacked, assuming repeated measures with matched values spread across a row, and with repeated measures in both directions. The tables below are color coded to explain these designs. Each color within a table represents one subject. The colors are repeated between tables, but this means nothing.

Matched values on row. 6 subjects.

Table format: Grouped		Group A Wild-type cells			Group B GPP5 cell line			Group C GPP7 cell line		
	⊗	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3	C:Y1	C:Y2	C:Y3
1	Serum starved	34	36	41	98	87	95	77	89	97
2	Normal culture	23	19	26	32	29	26	33	45	35

Matched values stacked. 9 subjects

Table format: Grouped		Group A Wild-type cells			Group B GPP5 cell line			Group C GPP7 cell line		
	⊗	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3	C:Y1	C:Y2	C:Y3
1	Serum starved	34	36	41	98	87	95	77	89	97
2	Normal culture	23	19	26	32	29	26	33	45	35

Repeated measures in both factors. 3 subjects.

Table format: Grouped		Group A Wild-type cells			Group B GPP5 cell line			Group C GPP7 cell line		
	⊗	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3	C:Y1	C:Y2	C:Y3
1	Serum starved	34	36	41	98	87	95	77	89	97
2	Normal culture	23	19	26	32	29	26	33	45	35

ANOVA tables

The table below shows the ANOVA tables for the four analyses. The values below are all reported by Prism. I rearranged and renamed a bit so the four can be shown on one table.

ANOVA table	No repeated measures SS	DF	Matched values on row SS	DF	Matched values stacked SS	DF	RM in both factors SS	DF
Row x Column	1989.0	2	1989.0	2	1989.0	2	1989.0	2
Row Factor	8278.0	1	8278.0	1	8278.0	1	8278.0	1
Column Factor	4124.0	2	4124.0	2	4124.0	2	4124.0	2
Subjects (matching)			120.9	4	269.0	6	45.4	2
Column x Subjects							223.6	4
Row x Subjects							75.4	2
Residual	418.7	12	297.8	8	149.7	6	74.2	4
Total	**14809.7**	**17**	**14809.7**	**17**	**14809.7**	**17**	**14809.7**	**17**

Focus first on the sum-of-squares (SS) column with no repeated measures:

- The first row shows the interaction of rows and columns. It quantifies how much variation is due to the fact that the differences between rows are not the same for all columns. Equivalently, it quantifies how much variation is due to the fact that the differences among columns is not the same for both rows.

- The second row show the the amount of variation that is due to systematic differences between the two rows.

- The third row show the the amount of variation that is due to systematic differences between the columns.

- The second to the last row shows the variation not explained by any of the other rows. This is called residual or error.

- The last row shows the total amount of variation among all 18 values.

Now look at the SS columns for the analyses of the same data but with various assumptions about repeated measures.

- The total SS stays the same. This makes sense. This measures the total variation among the 18 values.

- The SS values for the interaction and for the systematic effects of rows and columns (the top three rows) are the same in all four analyses.

- The SS for residual is smaller when you assume repeated measures, as some of that variation can be attributed to variation among subjects. In the final columns, some of that variation can also be attributed to interaction between subjects and either rows or columns.

Now look at the DF values.

- The total DF (bottom row) is 17. This is the total number of values (18) minus 1. It is the same regardless of any assumptions about repeated measures.

- The df for interaction equals (Number of columns - 1) (Number of rows - 1), so for this example is 2*1=2. This is the same regardless of repeated measures.

- The df for the systematic differences among rows equals number of rows -1, which is 1 for this example. This is the same regardless of repeated measures.

- The df for the systematic differences among columns equals number of columns -1, whiich is 2 for this example. It is the same regardless of repeated measures.

- The df for subjects is the number of subjects minus number of treatments. When the matched values are stacked, there are 9 subjects and three treatments, so df equals 6. When the matched values are in the same row, there arr 6 subjects treated in two ways (one for each row), so df is 4. When there are repeated measures for both factors, this value equals the number of subjects (3) minus 1, so df=2.

Details on how the SS and DF are computed can be found in Maxwell and Delaney (1). Table 12.2 on page 576 explains the ANOVA table for repeated measures in both factors. But note they use the term "A x B x S" where Prism says "Residual". Table 12.16 on page 595 explains the ANOVA table for two way ANOVA with repeated measures in one factor. They say "B x S/A" where Prism says "residual", and say "S/A" where Prism says "subject".

Mean squares

Each mean square value is computed by dividing a sum-of-squares value by the corresponding degrees of freedom. In other words, for each row in the ANOVA table divide the SS value by the df value to compute the MS value.

F ratio

Each F ratio is computed by dividing the MS value by another MS value. The MS value for the denominator depends on the experimental design.

For two-way ANOVA with no repeated measures: The denominator MS value is always the MSresidual.

For two-way ANOVA with repeated measures in one factor (p 596 of Maxwell and Delaney):

- For interaction, the denominator MS is MSresidual

- For the factor that is not repeated measures, the denominator MS is MSsubjects

- For the factor that is repeated measures, the denominator MS is MSresidual

For two-way ANOVA with repeated measures in both factors (p 577 of Maxwell and Delaney): The MS for the denominator is the MS for the interaction of the factor being tested with subjects.

- For Row Factor, the denominator MS is for Interaction of Row factor x Subjects

- For Column Factor, the denominator MS is for Interaction of Column factor x Subjects

- For the Interaction:Row Factor x Column Factor, the denominator MS is for Residuals (also called the interaction of Row x Column x Subjects)

P values

Each F ratio is computed as the ratio of two MS values. Each of those MS values has a corresponding number of degrees of freedom. So the F ratio is associated with one number of degrees of freedom for the numerator and another for the denominator. Prism reports this as something like: $F_{(1, 4)} = 273.9$

Calculting a P value from F and the two degrees of freedom can be done with a free web calculator or with the =FDIST(F, dfn, dfd) Excel formula

1. SE Maxwell and HD Delaney. <u>Designing Experiments and Analyzing Data: A Model Comparison Perspective, Second Edition</u>. Routledge, 2003.

5.14.3.3 Interpreting results: Mixed model two-way

P values

When interpreting the results of fitting a mixed model, interpreting the P values is the same as two-way ANOVA. So read the general page on <u>interpreting two-way ANOVA results</u>[488] first. Also read the general page on the assumption of <u>sphericity</u>[395], and <u>assessing violations of that assumption with epsilon</u>[400].

Random effects SD and variance

The mixed effects model treats the different subjects (participants, litters, etc) as a random variable. The residual random variation is also random. The effect of all random variables is quantified with its variation. Prism presents the variation as both a SD and a variance (which is the SD squared). You, or more likely your statistical consultant, may be interested in these values to understand the relative variation among subjects responses (the subject variance) and within the repeated responses from the same subject (the residual variance). compare with other programs.

Was the matching effective?

A repeated-measures experimental design can be very powerful, as it controls for random factors that cause or unmeasured variability between subjects. If the matching is effective, the repeated-measures test will yield a smaller P value than an ordinary ANOVA. The repeated-measures test is more powerful because it separates between-subject variability from within-subject variability. If the pairing is ineffective, however, the repeated-measures test can be less powerful because it has fewer degrees of freedom.

Prism tests whether the matching was effective and reports a P value. This P value comes from a chi-square statistic that is computed by comparing the fit of the full mixed effects model to a simpler model without accounting for repeated measures. If this P value is low, you can conclude that the matching was effective. If the P value is high, you can conclude that the matching was not effective and should reconsider your experimental design for your next study.

Goodness of fit

Prism expresses the goodness-of-fit in a few ways. These will only be meaningful to someone who understand mixed effects models deeply. Most scientists will ignore these results or not check the option so they never get reported. But some journals may ask you to report at least one measure of goodness of fit.

If you don't accept the assumption of sphericity

If you checked the option to not accept the assumption of sphericity, Prism does two things differently.

- It applies the correction of Geisser and Greenhouse. You'll see smaller degrees of freedom, which usually are not integers. The corresponding P value is higher than it would have been without that correction.

- It reports the value of [epsilon][400], which is a measure of how badly the data violate the assumption of sphericity.

Multiple comparisons tests and analysis checklist

Learn about [multiple comparisons tests after repeated measures ANOVA][436].

Before interpreting the results, [review the analysis checklist][181].

5.14.3.4   Graphing tips: Repeated measures two-way ANOVA

Graphing two-way ANOVA with repeated measures by row

From the New Graph dialog, you can choose a graph designed for repeated measures by rows or by columns.

Customize the graph within the Format Graph dialog:

- The appearance (for all data sets) should be 'Before-After'.

- Plot either symbols and lines or lines only. Choose the latter if you want to plot arrows.

- The line style drop down lets you choose arrow heads.
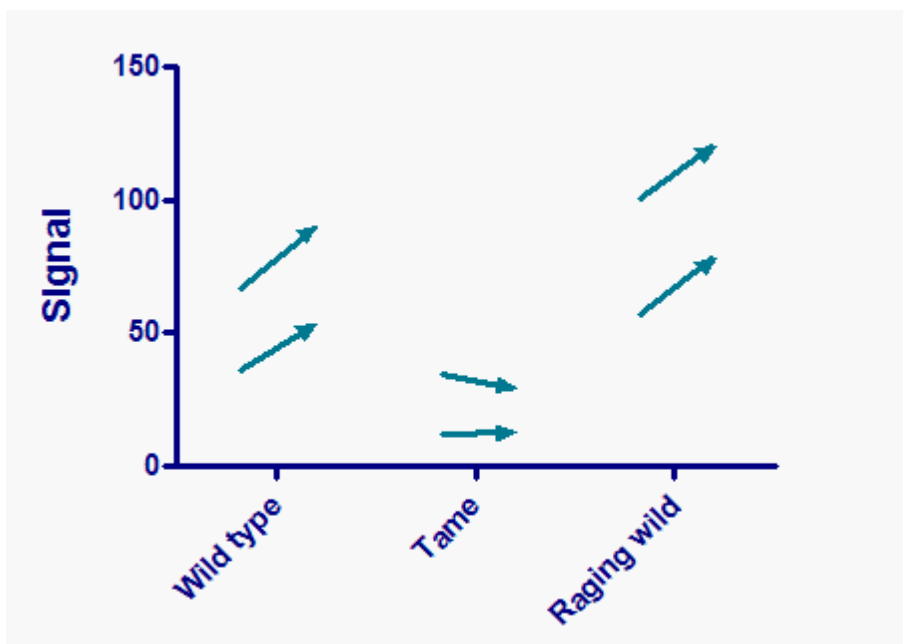
Graphing two-way ANOVA with repeated measures by column

From the New Graph dialog, you can choose a graph designed for repeated measures by rows. This is the second choice on the bottom row of graphs in the two-way tab.

Customize the graph within the Format Graph dialog:

- The appearance (for all data sets) should be "Each replicate".

- if you plot the replicates as 'Staggered", Prism will move them right or left to prevent overlap. In this example, none of the points overlap so 'Staggered' and 'Aligned' look the same.

- Check the option to plot 'one line for each subcolumn'.
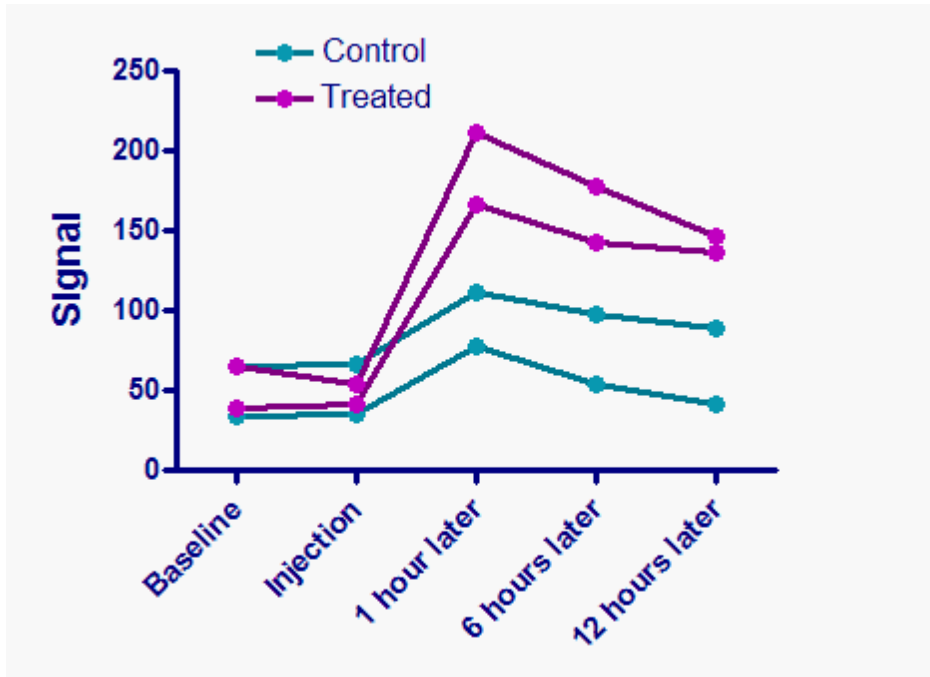
5.14.3.5 Table of means or predicted means

Starting with Prism 8, Prism can report a table of means as part of two-way ANOVA. Choose to create this table in the Options tab of the two-way ANOVA dialog.

If there are no missing values

If the data are balanced (no missing values) you'll see the label "Means" at the top of this table which shows the cell means, the row means, the column means, and the grand mean. You'll get the same means for ordinary or repeated measures ANOVA, and could easily compute the values yourself.

If there are any missing values

If the data are unbalanced (there are any missing values), this table is labeled "Predicted means (LS Means)" for the cell, row, column and

grand means. These predicted means, also known as "LSmeans", may not match the ones you compute by hand.

If you have repeated measures and missing values, the situation is complicated. Think about it this way. Prism fits a mixed model to your data, and that model is used to predict the cell means. If there are no missing values, those predictions always equal the actual means of the data. But if there are missing values, the predictions take into account the pattern of missing values, and the experimental design (repeated measures), so the predicted means do not equal the actual means and cannot easily be compute by hand.

Cell, row, column and grand SD

Prism also reports the SD of each cell, which is straightforward. It also reports the SD for each row and column, and SD for the whole table (the grand SD). These are computed by first averaging the replicates in each cell and then computing the SD of those cell means. If a row has three data set columns, each with five subcolumns for replicates, the n for computing the SD is 3, not 15.

5.14.3.6 Analysis checklist: Repeated measures two-way ANOVA (and mixed model)

Two-way ANOVA, also called two-factor ANOVA, determines how a response is affected by two factors. "Repeated measures" means that one of the factors was repeated. For example you might compare two treatments, and measure each subject at four time points (repeated). Read elsewhere to learn about choosing a test[456], graphing the data[513], and interpreting the results[506],[488].

✔ Are the data matched?

 If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA. Also check that your choice in the experimental design tab matches how the data are actually arranged. If you make a mistake, and the calculations are done assuming the wrong factor is repeated, the results won't be correct or useful.

✔ Are there two factors?

 One-way ANOVA compares three or more groups defined by one factor. For example, you might compare a control group with a drug treatment

group and a group treated with drug plus antagonist. Or you might compare a control group with five different drug treatments. Prism has a separate analysis for one-way ANOVA.

 Some experiments involve more than two factors. For example, you might compare three different drugs in men and women at four time points. There are three factors in that experiment: drug treatment, gender and time. These data need to be analyzed by three-way ANOVA, also called three-factor ANOVA.

✔ Are both factors "fixed" rather than "random"?

 While Prism assumes the participants in repeated measures are chosen randomly, it assumes that the treatments or categories designated by rows or data set columns are fixed. This means you are asking about how those particular treatments or categories affect the results. Different calculations would be needed if you randomly selected the treatments or categories from an infinite (or at least large) number of possible treatments or categories, and want to reach conclusions about differences among ALL the treatments or categories, even the ones you didn't include in this experiment. Prism does not handle this situation.

✔ Can you accept the assumption of sphericity?

 A random factor that causes a measurement in one subject to be a bit high (or low) should have no effect on the next measurement in the same subject. This assumption is called **circularity** or **sphericity**. It is closely related to another term you may encounter in advanced texts, **compound symmetry**.

 You only have to worry about the assumption of circularity when your experiment truly is a repeated-measures experiment, with measurements from a single subject. Circularity is unlikely to be an issue with randomized block experiments where you used a matched set of subjects (or a matched set of experiments)

 Repeated-measures ANOVA is quite sensitive to violations of the assumption of circularity. If the assumption is violated, the P value will be too low. You'll violate this assumption when the repeated measurements are made too close together so that random factors that cause a particular value to be high (or low) don't wash away or dissipate before the next measurement. To avoid violating the assumption, wait long enough between treatments so the subject is essentially the same

as before the treatment. Also randomize the order of treatments, when possible.

✔ **Consider alternatives to repeated measures two-way ANOVA.**

Two-way ANOVA may not answer the questions your experiment was designed to address. Consider alternatives.

✔ **If any values are missing, was that due to a random event?**

Starting with Prism 8, repeated measures data can be calculated with missing values by fitting a mixed model. But the results can only be interpreted if the reason for the value being missing is random. If a value is missing because it was too high to measure (or too low), then it is not missing randomly. If values are missing because a treatment is toxic, then the values are not randomly missing.

5.15   Three-way ANOVA

Three-way ANOVA, also called three-factor ANOVA, determines how a response is affected by three factors. For example, you might compare a response to drug vs. placebo in both men and women at two time points. Drug treatment is one factor, gender is the other, and time is the third. Is the response affected by drug? By gender? By time? Are the three intertwined? These are the kinds of questions that three-way ANOVA answers.

5.15.1   How to: Three-way ANOVA

.

5.15.1.1 Note of caution for statistical novices

Our goal with Prism has always been to make basic biostatistics very accessible and easy. Three-way ANOVA really is beyond "basic biostatistics". Multiple comparisons after three-way ANOVA stretch this definition even more. If you haven't taken the time to really understand three-way ANOVA, it is quite easy to be mislead by the results. Beware!

- Three-way ANOVA is not a topic that is easy to master. In addition to reading textbooks, also consider getting help from someone with more experience.

- Prism also offers to fit a mixed model to repeated measures data. Understanding this fully is even more complicated.

- Before getting lost in the many choices for multiple comparisons, first articulate clearly the scientific goals of the study. Don't articulate your goals in terms of ANOVA (looking for interactions). Figure out what you really want to know. Then figure out the best statistical approach to getting the answer.

- We can only provide limited technical support with understanding three-way ANOVA.

5.15.1.2 What is three-way ANOVA used for?

Visualizing three factors

Three-way ANOVA, also called three-factor ANOVA, determines how a response is affected by three factors, for example:

- Treated vs. control

- Male vs. female

- Pretreatment with low vs. high dose

This example has two levels of each of the three factors, so there are 2x2x2=8 different treatment groups. This diagram might help this make sense.
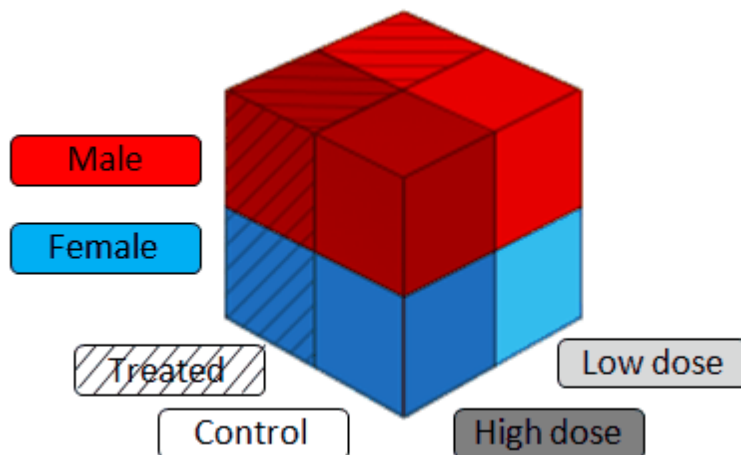
Seven null hypotheses

Three-way ANOVA tests seven null hypotheses so reports seven P values. Yes seven! Three-way ANOVA is complicated.

Three of the P values test main effects:

• Null hypothesis 1: On average, the measured value is the same in males and females. So this P value compares the red vs. the blue cubes above.

• Null hypothesis 2: On average, the measured value is the same for treated and control. This P value compares the striped vs solid cubes above.

• Null hypothesis 3: On average, the measured value is the same when the pretreatment is low or high dose. This P value compares the dark colored cubes with the light colored cubes above.

Three of the P values test two-way interactions, and one tests a three way interaction. Here are the null hypotheses:

- Null hypothesis 4: Pooling male and female, the effect of treatment vs. control is the same for pretreatment with low and high dose.

- Null hypothesis 5: Pooling treated and control, is the effect of pretreatment with low and high dose the same for males and females.

- Null hypothesis 6: Pooling pretreatment with low and high dose, the effect of treatment vs. control is the same for males and females.

- Null hypothesis 7: There is no three way interaction among all three factors. This one is hard to understand.

5.15.1.3 Three way ANOVA may not answer your scientific questions

GraphPad Prism can compute three-way ANOVA in certain circumstances. But before using three-way ANOVA, note that it often is much less useful than most scientists hope. When three way ANOVA is used to analyze data, the results often do not answer the questions the experiment was designed to ask. Let's work through an example:

The scientific goals and experimental design

A gene has been identified that is required for angiogenesis (growth of new blood vessels) under pathological conditions. The question is whether it also is active in the brain. Hypoxia (low oxygen levels) is known to provoke angiogenesis in the brain. So the question is whether angiogenesis (stimulated by hypoxia) will be reduced in animals created with that gene removed (knocked-out; KO) compared to normal (wild type, WT) animals. In other words, the goal is to find out whether there is a significant difference in vessels growth in the KO hypoxic mice compared to WT hypoxic mice.

The experimental design:

- Half the animals are wild-type. Half have the gene of interest knocked out.

- Half the animals are kept in normal air. Half are kept in hypoxic (low oxygen) conditions.

- Blood vessel number in a region of brain is measured at two time points (1 and 3 weeks).

What questions would three-way ANOVA answer?

The experiment has three factors: genotype (wild-type vs KO), oxygen (normal air vs. low oxygen) and time (1 and 3 weeks). So it seems logical to think that three-way ANOVA is the appropriate analysis. Three-way ANOVA will report seven P values (even before asking for multiple comparisons tests or contrasts). These P values test seven null hypotheses:

• Effect of genotype. The null hypothesis is that in both conditions (hypoxic or not) and at all time points, the average result in the wild-type animals equals the average affect in the KO animals. This isn't very useful. You don't expect the KO to be different in the normal air condition, so averaging that with hypoxia just muddles the picture. This P value is not helpful.

• Effect of hypoxia. The null hypothesis is that with both genotypes and all time points, the average result in normal air is identical to the average result in hypoxia. We already know hypoxia will provoke angiogenesis in WT animals. The point of the experiment is to see if hypoxia has a different affect in the KO animals. Combining the results of WT and KO animals together doesn't really make sense, so this P value is not helpful.

• Effect of time. The null hypothesis is that for both genotypes and both conditions (hypoxia or not), the average result at the two times points is the same. But we know already it takes time for angiogenesis to occur, so there will be more vessel growth at late times than at early time points in the normal animals treated with hypoxia. Combining both genotypes and both conditions doesn't really make sense. This P value is not helpful.

• Interaction of genotype and hypoxia. The null hypothesis is that the effect of hypoxia is the same in wild-type and KO animals at all both points. This sort of gets at the point of the study, and is the only one of seven P values that seems to answer the experimental question. But even this P value doesn't quite test the null hypothesis you care about. You really want to know if the two genotypes have different outcomes in the presence of hypoxia. Including the data collected under normal air will confuse the results, rather than clarify. Including the data at the earliest time point, before angiogenesis had a chance to begin also clouds the picture.

- Interaction of genotype and time. Under both conditions (hypoxia and not), the null hypothesis is that the difference between the two genotypes is consistent over time. Since the whole point of the experiment is to investigate the affect of hypoxia, it makes no sense really to average together the results from hypoxic animals with results from animals breathing regular air. This P value is not useful.

- Interaction of hypoxia and time. Averaging together both genotypes, the null hypothesis is that the effect of hypoxia is the same at all times. It really makes no sense to average together both genotypes, so this P value won't be useful.

- Three-way interaction of genotype, hypoxia and time. This P value is not useful, because it is too hard to figure out what null hypothesis it tests!

One alternative approach: Two-way ANOVA

Why were animals exposed to ordinary air included in the experiment? As a control. We don't expect much angiogenesis in the three week period for unstressed animals. The other half of the animals were exposed to hypoxia, which is known to provoke angiogenesis. The animals exposed to regular air are a control to show the experiment worked as expected. So I think it is reasonable to look at these results as a way to decide whether the experiment worked, and whether the hypoxic data are worth analyzing. If there was much angiogenesis in the animals exposed to regular air, you'd suspect some other toxin was present. Once you are sure the experiment worked, those data can be ignored in the final analysis.

By analyzing the data only from the hypoxic animals, we are down to two factors: genotype and time, so the data could be analyzed by two way ANOVA. Two-way ANOVA reports three P values from three null hypotheses:

- Effect of genotype. The null hypothesis is that pooling all time points, the average result in the wild-type animals equals the average affect in the KO animals. That gets at the experimental question, so is useful.

- Effect of time. The null hypothesis is that pooling both genotypes, the average result at the three times points is the same. But we know already there will be more vessel growth at late times than at early time points in the normal animals. We know that there are more blood

vessels at later times than earlier, so this P value is likely to be small, and that doesn't help answer the experimental question.

- Interaction of genotype and time. The null hypothesis is that the difference between the two genotypes is consistent at all time points. If the P value is large, you won't reject that hypothesis. In this case the P value for genotype answers the question the experiment was designed to ask. If the P value is small, you will reject the null hypothesis and conclude that the difference between genotypes is different at the various times. In this case, multiple comparison tests could compare the two genotypes at each time point individually.

Bottom line: With these data, considering half the experiment to be a control proving the methods worked vastly simplifies data analysis.

A statistician might object that those control data provide information about variability, so it isn't fair to ignore those data entirely. Someone skilled with R or SAS (etc.) could find a way to analyze all the data, to report P values that test the particular hypotheses of interest. But this is far from straightforward, and beyond the skills of most scientists. Blindly plugging the data into three-way ANOVA would not lead to results that answer the experimental question.

A better choice? Linear regression?

One problem with ANOVA (even two-way) is that it treats the three time points exactly as it would treat three species or treatment with three alternative drugs.

An alternative analysis approach would be to use regression. The simplest model is linear (and with only two time points, there would be no point fitting a more complicated model). Use linear regression to look at the rate of angiogenesis in hypoxic animals. Fit one slope to the WT animals and one to the KO animals, and compare the slopes.

This approach seems best to me. Each slope is understandable on its own as a measure of the rate of angiogenesis. The null hypothesis is understandable as well (the two rates are the same). The analysis seems much closer to the biological question, and the results will be much easier for nonstatisticians to interpret. Of course, it assumes that angiogenesis is linear over the time course studied, which may or may not be a reasonable assumption.

Summary

- Just because an experimental design includes three factors, don't assume that three-way ANOVA is the best analysis.

- Many experiments are designed with positive or negative controls. These are important, as they let you know whether everything worked as it should. If the controls gave unexpected results, it would not be worth analyzing the rest of the data. Once you've verified that the controls worked as expected, those control data can often be removed from the data used in the key analyses. This can vastly simplify data analysis.

- When a factor is dose or time, fitting a regression model often answers an experimental question better than does ANOVA.

5.15.1.4 Entering data for three-way ANOVA

Groups are defined by rows and columns

Prism organizes data for three-way ANOVA differently than do most other programs. Here is an example of data entry:

Table format: Grouped		Group A		Group B		Group C		Group D	
		Control Low- dose		Control High dose		Treated Low dose		Treated High dose	
	⊘	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2	D:Y1	D:Y2
1	Male								
2	Female								

- Note that Prism does not use grouping variables. Instead, you only enter the outcome variable and designate treatments or groups by choice of rows and columns.

- Rows 1 and 2 represent two levels of one factor (male vs. female) in the example. You can use more than two rows if this factor has more than three levels.

- Data set columns A and B vs. C and D represent the second factor, Control vs. Treated in the example.

- Data set columns A and C vs. B and D represent the third factor, low vs. high dose in the example.

- Only data in the first four data set columns (A-D) are considered in three-way ANOVA.

- The example shows two subcolumns, for duplicate data. But you can choose any number of subcolumns you want (up to 256).

Setting up the data table

From the Welcome (or New Data Table and Graph) dialog, choose the Grouped tab.

Entering raw data

If you are not ready to enter your own data, choose to use sample data and choose the three-way ANOVA sample data sets.
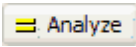
If you plan to enter your own data, create a table with the right number subcolumns to hold your replicated measurements.

You must enter two or more values for most combinations of conditions (alternatively, enter the mean, sample size and SD or SEM). Prism cannot compute three-way ANOVA with only a single value for each set of conditions.

Entering averaged data

If you have already averaged your replicates in another program, you can choose to enter and plot the mean and SD (or SEM) and n. If your data has more than 256 replicates, this is the only way to enter data into Prism for three-way ANOVA.

Running three-way ANOVA

1. From the data table, click ⊒ Analyze on the toolbar.

2. Choose Three-way ANOVA from the list of grouped analyses.

3. On the first tab (RM Design[527]) define your experimental design -- which factors are repeated measures.

4. If you have repeated measures, use the second (RM Analysis[527]) tab to specify whether you want to run repeated measures anova or fit a mixed model.

5. On the third (Factor Names[531]) tab, optionally name the grouping variables that define the rows and columns. For the example shown above, A&B vs C&D could be named "treatment"; A&C vs. B&D could be named "dose", the rows could be named "sex".

6. On the fourth (Multiple Comparisons[531]) tab, select your goal (if any) for multiple comparisons.

7. On the fifth (Options[532]) tab, choose the details of which multiple comparisons test you want.

8. On the sixth (Residuals[536]) tab, choose if and how you want to plot residuals, and if you want to test them for normality and equal variability (homoscedasticity).

Prism can't do two-way ANOVA with huge data sets

Prism cannot run three-way ANOVA with huge data sets and presents an message telling you so. How huge is huge? Details here.

5.15.1.5 Repeated measures design tab: Three-way ANOVA

Repeated measures by which factor(s)?

☐ Values on the same row in columns A and C (and B and D) represent a set of matched values.

☐ Values on the same row in columns A and B (and C and D) represent a set of matched values.

☑ Values stacked in subcolumns represent a set of matched values.

Assume sphericity (equal variability of differences)?

◉ No. Use the Geisser-Greenhouse correction. Recommended.

◯ Yes. No correction.

Repeated measures

Starting with Prism 8, three-way ANOVA can handle repeated measures. Three checkboxes let you specify which of the factors are, and are not, repeated measures. As you check and uncheck these options, look at the graphic on top of the screen which tries to show you the design you have selected.

Assume sphericity?

The concept of sphericity

The [concept of sphericity](#)[395] is tricky to understand. Briefly it means that you waited long enough between treatments for any treatment effect to wash away.This concept is not relevant if your data are not repeated measures.

If you aren't sure, we recommend that you do not assume sphericity.

How your choice affects Prism's calculations

If you choose to not assume sphericity, Prism will:

- Include the Geisser-Greenhouse correction when computing the repeated measures ANOVA P value. The resulting P value will be higher than it would have been without that correction.

- Quantify violations of sphericity by reporting [epsilon](#)[400].

- [Compute multiple comparisons tests differently.](#)[436]

If you ask Prism to assume sphericity, but in fact that assumption is violated, the P value from ANOVA will be too low. For that reason, if you are unsure whether or not to assume sphericity, we recommend that you check the option to *not* assume sphericity.

Random or fixed factors?

Read an advanced stats book to learn about random and fixed factors. Prism doesn't let you specify which factor(s) is/are random and which is/are fixed. It uses this rule:

- The row and column factors are considered to be fixed factors. You care about those exact levels of the factor. If there are three separate treatments, you care about those treatments. If there are four separate concentrations, you care about those concentrations (and have not selected those concentrations at random from all possible concentrations).

- When you choose repeated measures ANOVA, the participant or animal or subject is considered to be a random factor. You don't care about

these particular participants or animals or whatever. Instead, they are assumed to be randomly selected from a larger population that you care about.

5.15.1.6 Repeated measures analysis tab

Two ways to analyze repeated measures data

Prism can analyze repeated measures data in two ways:

- Repeated measures ANOVA

- Fitting a mixed effects model. This analysis works fine even when there are some missing values. The results will only be meaningful, of course, if the values are missing for random reasons. For example, those results won't be helpful or meaningful if the values are missing because those participants were very sick, or those values were too high to measure (or too low to measure). Fitting a mixed model with missing values only makes sense when there is zero association between the treatments or time-points and the reason why some values are missing.

In general, fitting a mixed effects model is a much more versatile method. As implemented in Prism 8, the two are completely equivalent when there are no missing values. But the mixed effects model method can also fit data with missing values.

Analyze using which method

The repeated measures tab of the ANOVA dialog (same for one-, two- and three-way data) gives you three choices:

- Use repeated measures ANOVA always. If there are missing values, no results will be reported. This matches what Prism 7 and earlier did. Prism is not "smart enough" to remove all data for a participant with missing values, but you could exclude all those values and rerun the ANOVA.

- Fit a mixed effects model always. This will make all analyses be consistent, whether or not there are missing values. If there are no missing values, the key results will be the same as repeated measures ANOVA but the results will be presented in an format unfamiliar to those

used to repeated measures ANOVA.

- Report the fit to a mixed effects model only when there are missing values, when repeated measures ANOVA is impossible. When there are no missing values, report the familiar repeated measures ANOVA results.

How to fit the mixed effects model any random factor is zero or negative

The whole point of repeated measures or mixed model analyses is that you have multiple response measurements on the same subject or when individuals are matched (twins or litters), so need to account for any correlation among multiple responses from the same subject. Mixed model analysis does this by estimating variances between subjects. In a simple mixed model, where only one variable is repeated, it's possible that this correlation for the particular data in your study is zero or even negative (of course it is impossible for a variance to be negative, but it can happen with mixed effects models). You are given two choices for what Prism should do when this happens:

•Analyze as usual. If there are no missing values, this will match repeated measures ANOVA.

•Remove the subject factor from the model and refit. This approach will have one more degree of freedom and thus have a bit more power. This approach is better but means the mixed model results may not match repeated measures ANOVA results.

In more complicated models, where there is more than one repeated measures variable, there are even more possible variance estimates (generally interactions with subject), and any of those could turn out to be zero or negative. It's best to take these out as leaving them in can make the results unstable.

Defaults for future analyses

Check an option at the bottom of the Repeated Measures tab to make your choices the default for future analyses. Your default will apply to one-, two- and three-way ANOVA.

5.15.1.7 Factor names tab: Three-way ANOVA

On this tab assign names to the three factors. The only reason to do this is to make the output easier to understand. For the example shown at the top of the dialog, the labels would be "Treatment, "Dose" and "Gender".

If you have repeated measures, you can change the name of the repeated measurement factor from "subject" to "litter" or "experiment" or whatever you want.

5.15.1.8 Multiple comparisons tab: Three-way ANOVA

Choose your goal for multiple comparisons. On the next tab, you can choose exactly which test to use.
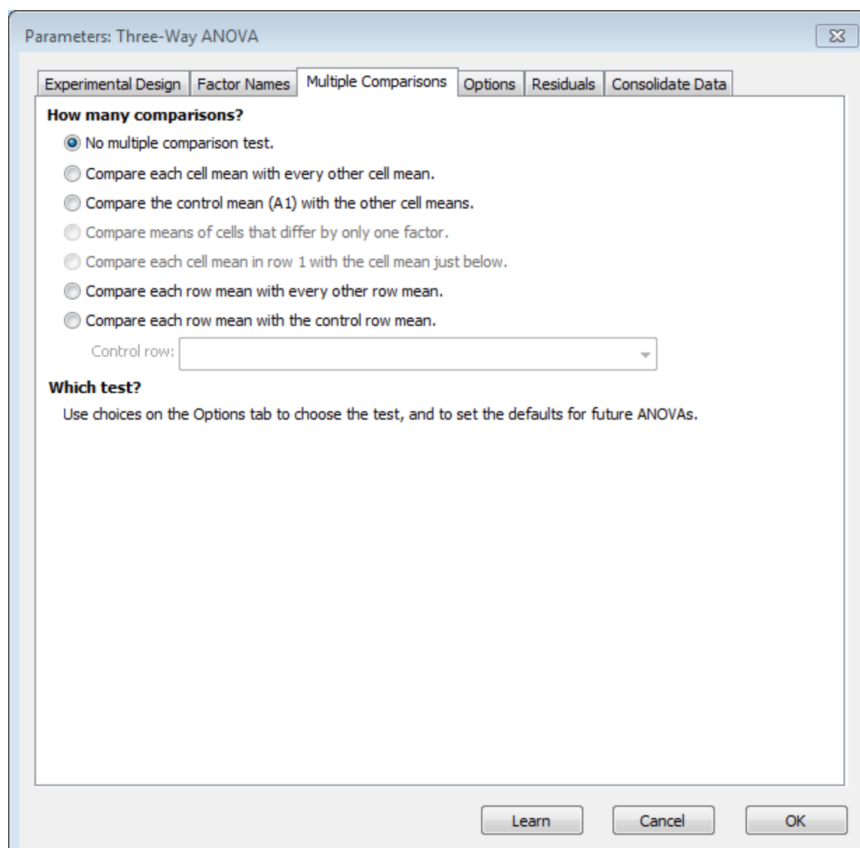
In many cases, the answers you really want are determined by multiple comparisons and not by the overall ANOVA analysis. Choose the comparisons you want to make on this tab, and then set the details on the next (Options[532]) tab.

Be thoughtful about which comparisons you need to make as part of your experimental design. If you make too many comparisons, the tests become stricter and it is harder to find anything.

If you have 2x2x2 ANOVA, then the number of comparisons are:

- Compare each cell mean with every other cell mean: 28 comparisons

- Compare the control mean with the other cell means: 7 comparisons

- Compare means that differ by only one factor: 12 comparisons

- Compare each mean in row 1 with the cell mean just below it: 4 comparisons

If you have 2 x 2 x k ANOVA, then the number of comparisons are:

- Compare each cell mean with every other cell mean: $8k^2-2k$ comparisons

- Compare the control mean with the other cell means: 2x2xk-1 comparisons

5.15.1.9 Options tab: Multiple comparisons: Three-way ANOVA

Multiple comparison approach

Correct for multiple comparisons using statistical hypothesis testing

Some of these methods let you compute confidence intervals and multiplicity adjusted P values, and some don't. We recommend one of the tests that compute confidence intervals and multiplicity adjusted P values for two reasons:

- Confidence intervals[59] are much easier for most to interpret than statements about statistical significance.

- [Multiplicity adjusted P values](#)[129] provide more information that simply knowing if a difference has been deemed statistically significant or not.

Methods than can compute confidence intervals and multiplicity adjusted P values

The list of tests available on this third tab of the dialog depends on the goal you specified on the second tab.

- If you are comparing every row (or column) mean with every other row (or column) mean, we recommend the [Tukey test](#)[136].

- If you are comparing a control row (or column) mean with the other row (or column) means, we suggest the [Dunnett's test](#)[136].

- If you are comparing a bunch of independent comparisons, we recommend the [Sidak](#)[131] method, which is very similar to Bonferroni but has a tiny bit more power.

Methods than cannot compute confidence intervals or multiplicity adjusted P values

If you don't care about seeing and reporting confidence intervals, you can gain a bit more power by choosing one of these tests. The list of tests available on this third tab of the dialog depends on the goal you specified on the second tab.

- If you are comparing every row or column mean with every other row or column mean, we recommend that you choose the [Holm-Šídák test](#)[133], which is more powerful than the Tukey method (3). That means that with some data sets, the Holm-Šídák method can find a statistically significant difference where the Tukey method cannot. We offer the Newman-Keuls test for historical reasons (so files made with old versions of Prism will open) but we suggest you avoid it because it does not maintain the family-wise error rate at the specified level(1). In some cases, the chance of a Type I error can be greater than the alpha level you specified.

- If you are comparing each column mean to a control mean, Prism only offers the [Holm-Šídák test](#)[133]. Glantz says that Holm's test ought to have more power than Dunnett's test, but this has not (to his knowledge) been explored in depth(2).

- If you are comparing a bunch of independent comparisons, Prism offers only the the Holm-Šídák test[133]

Correct for multiple comparisons by controlling the False Discovery Rate

Prism offers three methods[142] to control the false discovery rate[140]. All decide which (if any) comparisons to label as "discoveries" and do so in a way that controls the false discovery rate to be less than a value Q you enter.

The FDR approach is not often used as a followup test to ANOVA, but there is no good reason for that.

Don't correct for multiple comparisons. Each comparison stands alone.

If you choose this approach, Prism will perform Fisher's Least Significant Difference (LSD) test[127].

This approach (Fisher's LSD) has much more power to detect differences. But it is more likely to falsely conclude that a difference is statistically significant. When you correct for multiple comparisons (which Fisher's LSD does not do), the significance threshold (usually 5% or 0.05) applies to the entire family of comparisons. With Fisher's LSD, that threshold applies separately to each comparison.

Only use the Fisher's LSD approach if you have a very good reason, and are careful to explain what you did when you report the results.

Multiple comparisons

Swap direction of comparisons

The only affect of this option is to change the sign of all reported differences between means. A difference of 2.3 will be -2.3 if the option is checked. A difference of -3.4 will be 3.4 if you check the option. It is purely a personal preference that depends on how you think about the data.

Report multiplicity adjusted P value for each comparison

If you choose the Bonferroni, Tukey or Dunnett multiple comparisons test, Prism can also report multiplicity adjusted P values[129]. If you check this option, Prism reports an adjusted P value for each comparison. These calculations take into account not only the two groups being compared,

but the total number groups (data set columns) in the ANOVA, and the data in all the groups.

The multiplicity adjusted P value is the smallest significance threshold (alpha) for the entire family of comparisons at which a particular comparison would be (just barely) declared to be "statistically significant".

Until recently, multiplicity adjusted P values have not been commonly reported. If you choose to ask Prism to compute these values, take the time to be sure you understand what they mean. If you include these values in publications or presentations, be sure to explain what they are.

Confidence and significance level (or desired FDR)

By tradition, confidence intervals are computed for 95% confidence and statistical significance is defined using an alpha of 0.05. Prism lets you choose other values. If you choose to control the FDR, select a value for Q (in percent). If you set Q to 5%, you expect up to 5% of the "discoveries" to be false positives.

References

1. SA Glantz, Primer of Biostatistics, sixth edition, ISBN= 978-0071435093.

2. MA Seaman, JR Levin and RC Serlin, Psychological Bulletin 110:577-586, 1991.

5.15.1.10 Options tab: Graphing and output: Three-way ANOVA

Graphing options

If you chose a multiple comparison method that computes confidence intervals (Tukey, Dunnett, etc.) Prism can plot these confidence intervals.

Output

Choose [how you want P values reported](#) [82], and how many significant digits you need.

5.15.1.11 Residuals tab: Three-way ANOVA

Why residuals?

Prism 8 introduced the ability to plot residual plots with ANOVA, provided that you entered raw data and not averaged data as mean, n and SD or SEM.

Many scientists thing of residual as values that are obtained with regression. But ANOVA is really regression in disguise. It fits a model. One of the assumptions of ANOVA is that the residuals from that model are sampled from a Gaussian distribution. A residual plot helps you assess this assumption.

Which graph to create?

Prism can make three kinds of residual plots.

- Residual plot. The X axis is the predicted value. The Y axis is the residual. This lets you spot residuals that are much larger or smaller than the rest.

- Homoscedasticity plot. The X axis is the predicted value. The Y axis is the absolute value of the residual.This lets you check whether larger values are associated with bigger residuals (large absolute value).

- QQ plot. The X axis is the actual residual. The Y axis is the predicted residual, computed from the percentile of the residual (among all residuals) and assuming sampling from a Gaussian distribution. ANOVA assumes a Gaussian distribution of residuals, and this graph lets you check that assumption.

For ordinary ANOVA, the predicted values (used for residual and homoscedasticity plots) are simply the means of the replicates in a cell. For repeated measures ANOVA, the predicted value also takes into account the subject mean.

Diagnostics for residuals

- Are the residuals clustered or heteroscedastic? ANOVA assumes each sample was randomly drawn from populations with the same standard deviation. Prism can test this assumption with two tests. The Brown-Forsythe test and the Barlett test. Both these tests compute a P value designed to answer this question:If the populations really have the same standard deviations, what is the chance that you'd randomly

select samples whose standard deviations are as different from one another (or more different) as they are in your experiment?

- Are the residuals Gaussian? Prism runs four normality tests on the residuals. The residuals from all groups are pooled and then entered into one normality test.

5.15.1.12 Consolidate tab: Three-way ANOVA

This will consolidate the three-way data down to a two-way table.

If one of the factors has little impact, it may make sense to consolidate the data so it has two factors, rather than three.

When columns are combined, the values stay the same but there are more subcolumns. So if you choose to combine columns A and B (and C and D), you'll end up with half as many dataset columns, each with twice as many subcolumns. The Y1 subcolumn of column B becomes the Y3 subcolumn of the new column A.

5.15.2 Interpreting results: Three-way ANOVA

.

5.15.2.1 Interpreting results: Three-way ANOVA

Three-way ANOVA really is beyond "basic biostatistics". Multiple comparisons after three-way ANOVA stretch this definition even more. If you haven't taken the time to really understand three-way ANOVA, it is quite easy to be mislead by the results. Beware! It is not possible to understand three-way ANOVA only by reading these help screens.

Three-way ANOVA divides the total variability among values into eight components, the variability due to each of the factors (three components), due to each of the two-way interactions between two factors, due to three three-way interaction among all factors, and due to the variation among replicates (called residual or error variation). For each of those sources of variation, Prism reports the fraction of the variation attributed to that source, and (for all but the last) a P value testing the null hypothesis that the data are drawn from a population where that potential source of variation in fact contributes nothing to the overall variation among values.

Note that the eight P values produced by two-way ANOVA are not corrected for the eight comparisons. It would seem logical to do so, but this is not traditionally (ever?) done in ANOVA.

If your data have repeated measures in any (or all) of the factors, Prism can either do repeated measures ANOVA or fit a mixed effects model. With no missing values, the P values and multiple comparisons tests are identical. If there are missing values, the results can only be interpreted if the reason for the value being missing is random. If a value is missing because it was too high to measure (or too low), then it is not missing randomly. If values are missing because a treatment is toxic, then the values are not randomly missing.

It might help to review our information on interpreting [ANOVA table][508] and [mixed model][512] results for two-way ANOVA results.

Multiple comparisons testing is one of the most confusing topics in statistics. Since Prism offers nearly the same multiple comparisons tests for one-, two and three-way ANOVA, we have [consolidated the information on multiple comparisons][117].

5.15.2.2 Analysis checklist: Three-way ANOVA

Three-way ANOVA, also called three-factor ANOVA, determines how a response is affected by three factors. For example, you might measure a response to three different drugs, in men and women, with two different pretreatments. In this example, drug treatment is one factor, gender is the other, and pretreatment is the third. Read elsewhere to learn about [interpreting the results](#)[538].

✔ **Are the populations distributed according to a Gaussian distribution?**

Three-way ANOVA assumes that your replicates are sampled from Gaussian distributions. While this assumption is not too important with large samples, it is important with small sample sizes, especially with unequal sample sizes. Prism does not test for violations of this assumption. If you really don't think your data are sampled from a Gaussian distribution (and no transform will make the distribution Gaussian), you should consider performing nonparametric ANOVA. Prism does not offer this test.

ANOVA also assumes that all sets of replicates have the same SD overall, and that any differences between SDs are due to random sampling.

✔ **Are the data unmatched?**

Three-way ANOVA works by comparing the differences among group means with the pooled standard deviations of the groups. If subjects were given more than one treatment sequentially, or the experimental design worked with sets of matched subjects, then you should use repeated measures ANOVA. Prism cannot calculate three-way ANOVA with repeated measures in any factor.

✔ **Are the "errors" independent?**

The term "error" refers to the difference between each value and the mean of all the replicates. The results of three-way ANOVA only make sense when the scatter is random – that whatever factor caused a value to be too high or too low affects only that one value. Prism cannot test this assumption. You must think about the experimental design. For

example, the errors are not independent if you have six replicates, but these were obtained from two animals in triplicate. In this case, some factor may cause all values from one animal to be high or low.

✔ Do you really want to compare means?

Three-way ANOVA compares the means. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

✔ Are there three factors?

Don't mix up three way ANOVA with one way ANOVA with three groups. With three way ANOVA, there are three grouping variables, maybe gender, presence or absence of disease, and control vs. treated. With one-way ANOVA there is one grouping variable (perhaps treatment). If there are three alternative treatments, you need one-way ANOVA not three-way ANOVA.

✔ Are all three factors "fixed" rather than "random"?

Prism performs Type I ANOVA, also known as fixed-effect ANOVA. This tests for differences among the means of the particular groups you have collected data from. Different calculations are needed if you randomly selected groups from an infinite (or at least large) number of possible groups, and want to reach conclusions about differences among ALL the groups, even the ones you didn't include in this experiment.

5.16 Categorical outcomes

You've assessed an outcome with only two (or a few) possibilities. Survive or not. Metastasis or not. Graduate or not. Democrat, republican or independent. How can you express the precision

by which you know the proportions? How can you compare two or more groups?

5.16.1 The Confidence Interval of a proportion

5.16.1.1 How Prism can compute a confidence interval of a proportion

Example

When an experiment has two possible outcomes, the results are expressed as a proportion. Since your data are derived from random sampling, the true proportion in the overall population is almost certainly different than the proportion you observed. A 95% confidence interval quantifies the uncertainty.

For example, you look in a microscope at cells stained so that live cells are white and dead cells are blue. Out of 85 cells you looked at, 6 were dead. The fraction of dead cells is 6/85 = 0.0706.

The 95% confidence interval extends from 0.0263 to 0.1473. If you assume that the cells you observed were randomly picked from the cell suspension, and that you assessed viability properly with no ambiguity or error, then you can be 95% sure that the true proportion of dead cells in the suspension is somewhere between 2.63 and 14.73 percent.

How to compute the confidence interval with Prism

1. Create a new table formatted for parts of whole data.

2. Enter data only into the first two rows of column A. Enter the actual number of times each outcome occurred. For the example, enter 6 into the first row (number of blue dead cells) and 79 into the second row (number of white alive cells). Don't enter the total number of events or objects you examined. Prism will compute the total itself.

3. If you have more proportions that you wish to compute a confidence interval for, enter them into more columns of the data table.

4. Click Analyze, and choose the Fraction of Total analysis.

5. Choose to divide each value by its column total, and check the option to compute 95% confidence intervals. Choose whether you want to see the results as fractions of percentages.
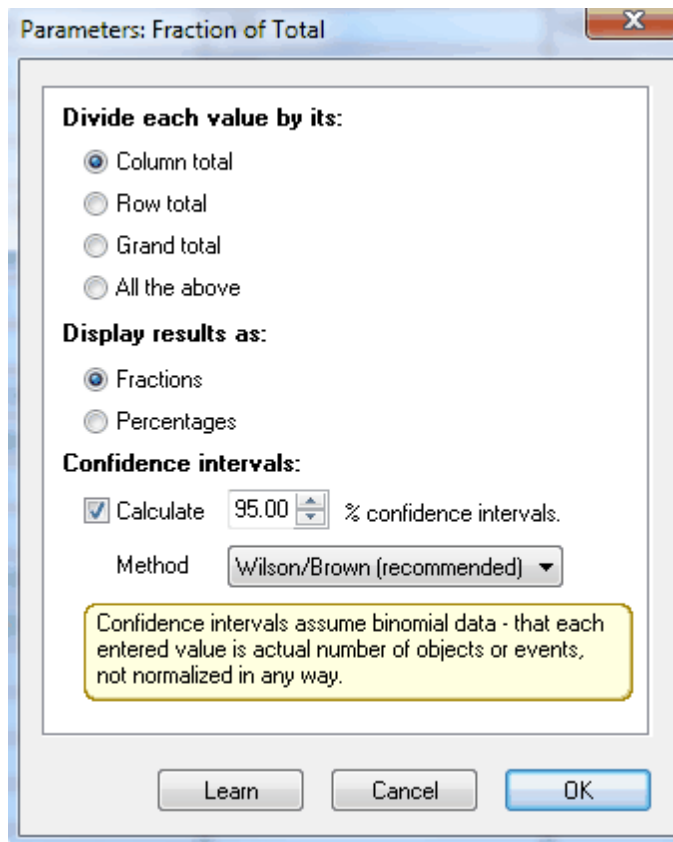
5.16.1.2 Three methods for computing the CI of a proportion

There are many methods to compute the confidence interval of a proportion. Your goal is to have a 95% confidence intervals, but in fact the actual confidence level (for all methods) depends on the precise values for the numerator and denominator.

Prism offers three methods to compute the confidence interval of a proportion:

- The so called "exact method" of Clopper and Pearson (1). This is the only method Prism 6 (and earlier) used. No matter what data you enter, this method always ensures that the actual confidence level is greater than the level you requested (usually 95%). But often the actual

confidence level is a lot higher. On average, therefore, these intervals have a greater confidence level than you requested so are wider than they need to be.

- The method of Wilson (2). With some data the actual confidence level is greater than what you requested, and for some data the actual confidence level is less. On the average, the actual confidence level equals the confidence level you requested. Wilson's method is great except when the probability is very close to 0 or 1. Note that while some programs offer a variation of this method with a continuity correction, but Prism does not.

- The hybrid Wilson/Brown method (3). In most cases this method uses the Wilson method with no modifications. The exception is when the proportion is very close to 0.00 or 1.00. In these cases, Brown prescribes using a Poisson approximation instead. Prism uses this approximation for the lower confidence limit when the numerator (N) equals 1 or 2, and when N=3 and the denominator (D) exceeds 50. It uses this approximation for the upper limit when N=D-1 or N=D-2 (or N=D-3, and N > 50). Note that the approximation is not needed when N=0. In that case the lower confidence limit is 0.0 and the upper limit is computed by Wilson's method. Similarly, when N=D, the upper confidence limit equals 1.00 and the lower limit is computed by Wilson's method. Brown and colleagues (3) call this hybrid method the modified Wilson method, but this name can be ambiguous because other modifications of Wilson's method have been proposed.

We recommend the third (Wilson/Brown hybrid) method. Use one of the first two methods only if you need to be sure that Prism's results match those of another program.

References

1. Clopper, C.; Pearson, E. S. (1934). "The use of confidence or fiducial limits illustrated in the case of the binomial". Biometrika 26: 404–413. doi:10.1093/biomet/26.4.404

2. Wilson, E. B. (1927). "Probable inference, the law of succession, and statistical inference". Journal of the American Statistical Association 22: 209–212. JSTOR 2276774.

3. Brown, L., Cai, T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statist. Sci, 16(2), 101–133.

5.16.1.3 The meaning of "95% confidence" when the numerator is zero

Interpreting a confidence interval is usually straightforward. But if the numerator of a proportion is zero, the interpretation is not so clear. In fact, the "95% confidence interval" really gives you 97.5% confidence. Here's why:

When the proportion does not equal zero, Prism reports the 95% confidence interval so that there is a 2.5% chance that the true proportion is less than the lower limit of the interval, and a 2.5% chance that the true proportion is higher than the upper limit. This leaves a 95% chance (100% -2.5% - 2.5%) that the interval includes the true proportion. When the numerator is zero, we know that the true proportion cannot be less than zero, so we only need to compute an upper confidence limit. Prism still calculates the upper limit so that there is a 2.5% chance that the true proportion is higher. Since the uncertainty only goes one way you'll actually have a 97.5% CI (100% - 2.5%). The advantage of calculating the "95%" confidence interval this way is that it is consistent with 95% CIs computed for proportions where the numerator is not zero.

If you don't care about consistency with other data, but want to really calculate a 95% CI, you can do that by computing a "90% CI". This is computed so that there is a 5% chance that the true proportion is higher than the upper limit. If the numerator is zero, there is no chance of the proportion being less than zero, so the "90% CI" really gives you 95% confidence.

5.16.1.4 What are binomial variables

The confidence interval for a proportion computes results that only make sense if the variable is binomial. There must be two possible outcomes, and you know the exact number of times each occurred.

If you know that an outcome happens 25% of the time, it makes no sense to compute the confidence interval for 25/100 unless you observed exactly 100 times and saw that outcome in 25 of them.

If you know that a treatment reduces white cell count by 33%, you cannot compute a confidence interval assuming a binomial distribution. Nor if a treatment decreases body weight by 25%.

5.16.2   Contingency tables

Contingency tables summarize results where you compared two or more groups and the outcome is a categorical variable (such as disease vs. no disease, pass vs. fail, artery open vs. artery obstructed).

5.16.2.1 Key concepts: Contingency tables

Contingency tables

Contingency tables summarize results where you compared two or more groups and the outcome is a categorical variable (such as disease vs. no disease, pass vs. fail, artery open vs. artery obstructed).

Contingency tables display data from these five kinds of studies:

- In a **cross-sectional** study, you recruit a single group of subjects and then classify them by two criteria (row and column). As an example, let's consider how to conduct a cross-sectional study of the link between electromagnetic fields (EMF) and leukemia. To perform a cross-sectional study of the EMF-leukemia link, you would need to study a large sample of people selected from the general population. You would assess whether or not each subject has been exposed to high levels of EMF. This defines the two rows in the study. You then check the subjects to see whether or not they have leukemia. This defines the two columns. It would not be a cross-sectional study if you selected subjects based on EMF exposure or on the presence of leukemia.

- A **prospective** study starts with the potential risk factor and looks forward to see what happens to each group of subjects. To perform a prospective study of the EMF-leukemia link, you would select one group of subjects with low exposure to EMF and another group with high exposure. These two groups define the two rows in the table. Then you would follow all subjects over time and tabulate the numbers

that get leukemia. Subjects that get leukemia are tabulated in one column; the rest are tabulated in the other column.

- A **retrospective** case-control study starts with the condition being studied and looks backwards at potential causes. To perform a retrospective study of the EMF-leukemia link, you would recruit one group of subjects with leukemia and a control group that does not have leukemia but is otherwise similar. These groups define the two columns. Then you would assess EMF exposure in all subjects. Enter the number with low exposure in one row, and the number with high exposure in the other row. This design is also called a case-control study.

- In an **experiment**, you manipulate variables. Start with a single group of subjects. Half get one treatment, half the other (or none). This defines the two rows in the study. The outcomes are tabulated in the columns. For example, you could perform a study of the EMF/leukemia link with animals. Half are exposed to EMF, while half are not. These are the two rows. After a suitable period of time, assess whether each animal has leukemia. Enter the number with leukemia in one column, and the number without leukemia in the other column. Contingency tables can also tabulate the results of some basic science experiments. The rows represent alternative treatments, and the columns tabulate alternative outcomes.

- Contingency tables also assess the accuracy of a **diagnostic test**. Select two samples of subjects. One sample has the disease or condition you are testing for, the other does not. Enter each group in a different row. Tabulate positive test results in one column and negative test results in the other.

For data from prospective and experimental studies, the top row usually represents exposure to a risk factor or treatment, and the bottom row is for controls. The left column usually tabulates the number of individuals with disease; the right column is for those without the disease. In case-control retrospective studies, the left column is for cases; the right column is for controls. The top row tabulates the number of individuals exposed to the risk factor; the bottom row is for those not exposed.

Logistic regression

Contingency tables analyze data where the outcome is categorical, and where there is one independent (grouping) variable that is also

categorical. If your experimental design is more complicated, you need to use logistic regression which is available in Prism. Logistic regression is used when the outcome is categorical, specifically when the outcome is binary (yes/no, alive/dead, pass/fail, etc.). In some cases, you may only have one independent variable (X variable) as a predictor for this outcome. In this case, you can use simple logistic regression. Additionally, if there are multiple independent variables, which can be categorical or numerical, you can use multiple logistic regression. To continue the example above, imagine you want to compare the incidence of leukemia in people who were, or were not, exposed to EMF, but want to account for gender, age, and family history of leukemia. You can't use a contingency table for this kind of analysis, but you *can* use logistic regression.

5.16.2.2 How to: Contingency table analysis

1. Create a contingency table

From the Welcome or New table dialog, choose the contingency tab.

 If you are not ready to enter your own data, choose one of the sample data sets.

2. Enter data

Most contingency tables have two rows (two groups) and two columns (two possible outcomes), but Prism lets you enter tables with any number of rows and columns.

You must enter data in the form of a contingency table. Prism cannot cross-tabulate raw data to create a contingency table.

For calculation of P values, the order of rows and columns does not matter. But it does matter for calculations of relative risk, odds ratio, etc. Use the sample data to see how the data should be organized.
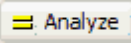
Be sure to enter data as a contingency table. The categories defining the rows and columns must be mutually exclusive, with each subject (or experimental unit) contributing to one cell only. In each cell, enter the number of subjects actually observed. Your results will be completely meaningless if you enter averages, percentages or rates. You must enter the actual number of subjects, objects, events. For this reason, Prism

won't let you enter a decimal point when entering values into a contingency table.

If your experimental design matched patients and controls, you should not analyze your data with contingency tables. Instead you should use McNemar's test[569].

If you want to compare an observe distribution of values with a distribution expected by theory, do not use a contingency table. Prism offers another analysis[563] for that purpose.

3. Analyze

From the data table, click ⇋ Analyze on the toolbar, and choose **Chi-square (and Fisher's exact) test**.

Main calculations for tables with two rows and two columns
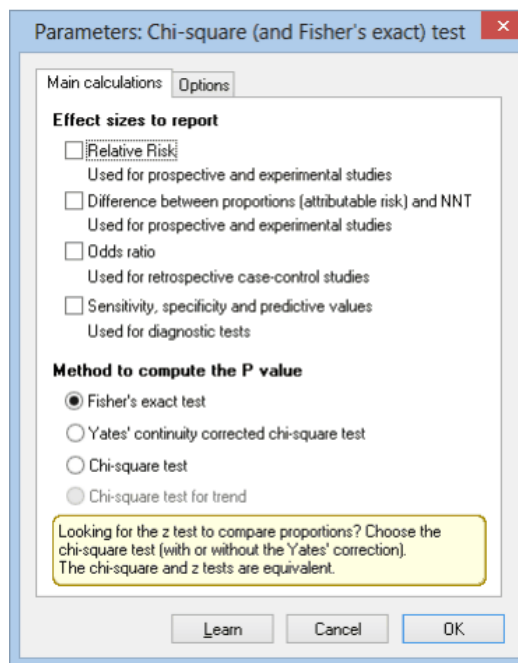
Your choice of effect sizes will depend on experimental design. Calculate an Odds ratio from retrospective case-control data, sensitivity (etc.) from a study of a diagnostic test, and relative risk and difference between proportions from prospective and experimental studies. All of these effect sizes apply only to 2x2 tables, so the choices will be gray if your table is larger.

If your table has two rows and two columns, we suggest you always choose Fisher's exact test[551] to calculate the P value.

Main calculations for tables with more than two rows and/or more than two columns

If your table has two columns and three or more rows, you can choose the chi-square test or the **chi-square test for trend**. This calculation tests whether there is a linear trend between row number and the fraction of subjects in the left column. It only makes sense when the rows are arranged in a natural order (such as by age, dose, or time), and are equally spaced. The test is also called the *Cochran-Armitage test for trend*. It is explained clearly, with equations and an example, on pages 261-265 Altman (2). You can find these pages at Google Books.

With contingency tables with more than two rows or columns, Prism always calculates the chi-square test. You have no choice. Extensions to Fisher's exact test have been developed for larger tables, but Prism doesn't offer them.

Options
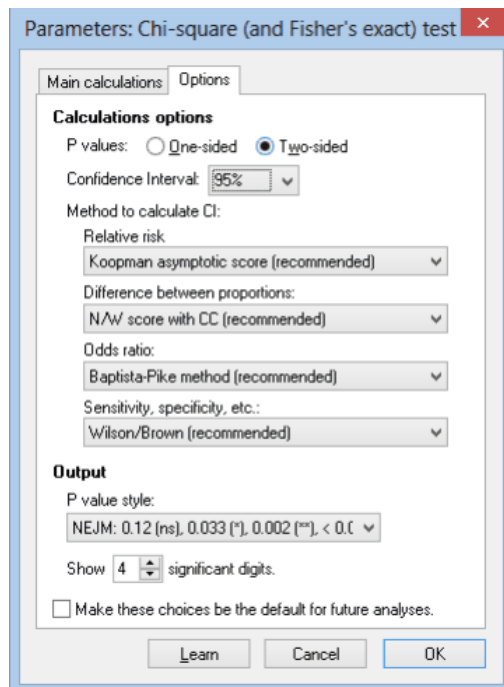
We suggest always choosing a [two-sided P value](#)[75] unless you have a strong reason to choose a one-sided P value.

Prism now offers a choice of method to use when computing confidence intervals.

- CI of the relative risk. The method of Katz is an approximation, and we suggest not using it except to maintain compatibility with analyses done with earlier versions of Prism. There are many ways to compute the CI of a relative risk (1), and many of these seem good. Prism now offers the Koopman asymptotic score method, which we recommend.

- CI for the difference between proportions. The asymptotic method used by earlier versions of Prism is an approximation, and we suggest not using it except for compatibility. Instead choose the Newcombe/Wilson method(2). We offer that method with and without the continuity correction, and recommend the variation with that correction.

- CI of the odds ratio. The method of Woolf used by Prism 6 and earlier is an approixmation, and we suggest you not use it except to maintain compatibility. There are many good methods to compute the CI of an odds ratio (2). Prism now offers the Baptista-Pike method, which we recommend.

- CI of the sensitivity, specificity, etc. The so called "exact method" of Clopper and Pearson produces wide confidence intervals, and we suggest you don't use it except for compatibility. Instead [choose the hybrid Wilson/Brown method](#)[542] (3).

You can also choose [how you want P values formatted](#)[82] .

4. Review the results

[Interpreting results: relative risk and odds ratio](#)[557]

[Interpreting results: sensitivity and specificity](#)[560]

[Interpreting results: P values (from contingency tables)](#)[553]

[Analysis checklist: Contingency tables](#)[191]

1. Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. Stat Methods Med Res. SAGE Publications; 2011 Oct 13.

2. Newcombe, R. G. R. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine, 17(8), 873–890.

3. Brown, L., Cai, T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statist. Sci, 16(2), 101–133.

5.16.2.3 Fisher's test or chi-square test?

If you entered data with two rows and two columns, you must choose the *chi-square test* (sometimes called the *chi-square test of homogeneity*) *or Fisher's exact test*.

Chi-square and Yates correction

In the days before computers were readily available, people analyzed contingency tables by hand, or using a calculator, using chi-square tests. This test works by computing the expected values for each cell if the relative risk (or odds' ratio) were 1.0. It then combines the discrepancies between observed and expected values into a chi-square statistic from which a P value is computed.

The chi-square test is only an approximation. The **Yates continuity correction** is designed to make the chi-square approximation better, but it over corrects so gives a P value that is too large (too 'conservative'). With large sample sizes, Yates' correction makes little difference, and the chi-square test works very well. With small sample sizes, chi-square is not accurate, with or without Yates' correction. Statisticians seem to disagree on whether or not to use Yates correction. Prism gives you the choice.

If the observed and expected values are all very close (within 0.25), the Yates correction sort of works backwards, and actually increases the value of chi-square and thus lowers the P value, rather than decreasing chi-square and increasing P. This is a rare occurrence, and only happens

when the relative risk or odds ratio is very close to 1.0. If you asked for the Yates correction, Prism does the Yates correction even in this case.

Fisher's test. Exactly correct answer to wrong question?

Fisher's exact test, as its name implies, always gives an exact P value and works fine with small sample sizes. Fisher's test (unlike chi-square) is very hard to calculate by hand, but is easy to compute with a computer. Most statistical books advise using it instead of chi-square test. If you choose Fisher's test, but your values are huge, Prism will override your choice and compute the chi-square test instead, which is very accurate with large values.

As its name implies, Fisher's exact test, gives an exactly correct answer no matter what sample size you use. But some statisticians conclude that Fisher's test gives the exact answer to the wrong question, so its result is also an approximation to the answer you really want. The problem is that the Fisher's test is based on assuming that the row and column totals are fixed by the experiment. In fact, the row totals (but not the column totals) are fixed by the design of a prospective study or an experiment, the column totals (but not the row totals) are fixed by the design of a retrospective case-control study, and only the overall N (but neither row or column totals) is fixed in a cross-sectional experiment. Ludbrook (1) points out that Fisher designed his exact test to analyze a unique experiment, and that experimental design is extremely rare.

Since the design of your study design is extremely unlikely to match the constraints of Fisher's test, you could question whether the exact P value produced by Fisher's test actually answers the question you had in mind.

If you enter huge numbers (the sum is greater than 1,000,000) Prism will perform the chi-square test even if you chose Fisher's test.

An alternative to Fisher's test is the **Barnard test**. Fisher's test is said to be 'conditional' on the row and column totals, while Barnard's test is not. Mehta and Senchaudhuri explain the difference and why Barnard's test has more power (2). Berger modified this test to one that is easier to calculate yet more powerful. Ludbrook discusses other exact methods that are appropriate to common experimental designs (1).

At this time, we do not plan to implement Bernard's or Berger's test in Prism or the exact tests mentioned by Ludbrook (1). There certainly does not seem to be any consensus that these tests are preferred. But let us

know if you would like to see these tests in a future version of Prism. Here is an online calculator that performs Berger's test.

References

1. Ludbrook, J. (2008). Analysis of 2 x 2 tables of frequencies: matching test to experimental design. International Journal of Epidemiology, 37, 1430 -1435.

2. Mehta, C. R. and Senchaudhur, P., Conditional versus Unconditional Exact Tests for Comparing Two Binomials. http://www.cytel.com/Papers/twobinomials.pdf

5.16.2.4 Interpreting results: P values from contingency tables

What question does the P value answer?

The P value from a Fisher's or chi-square test answers this question:

If there really is no association between the variable defining the rows and the variable defining the columns in the overall population, what is the chance that random sampling would result in an association as strong (or stronger) as observed in this experiment?

The *chi-square test for trend* is performed when there are two columns and more than two rows arranged in a natural order. It is also called the *Cochran-Armitage method*. The P value answers this question:

If there is no linear trend between row number and the fraction of subjects in the left column, what is the chance that you would happen to observe such a strong trend as a consequence of random sampling?

For more information about the chi-square test for trend, see the excellent text, Practical Statistics for Medical Research by D. G. Altman, published in 1991 by Chapman and Hall.

Don't forget that "statistically significant" is not the same as "scientifically important"[89] .

You will interpret the results differently depending on whether the P value is small[79] or large[81] .

Why isn't the P value always consistent with the confidence interval?

P values and confidence intervals are intertwined. If the P value is less than 0.05, then the 95% confidence interval cannot contain the value that defines the null hypothesis. (You can make a similar rule for P values < 0.01 and 99% confidence intervals, etc.)

This rule is not always upheld with Prism's results from contingency tables.

The P value computed from Fisher's test is exactly correct. However, the confidence intervals for the Odds ratio and Relative Risk are computed by methods that are only approximately correct. Therefore it is possible that the confidence interval does not quite agree with the P value.

For example, it is possible for results to show P<0.05 with a 95% CI of the relative risk that includes 1.0. (A relative risk of 1.0 means no risk, so defines the null hypothesis). Similarly, you can find P>0.05 with a 95% CI that does not include 1.0.

These apparent contradictions happens rarely, and most often when one of the values you enter equals zero.

How the P value is calculated

Calculating a chi-square test is standard, and explained in all statistics books.

The Fisher's test is called an "exact" test, so you would think there would be consensus on how to compute the P value. Not so!

While everyone agrees on how to compute one-sided (one-tail) P value, there are actually three methods to compute "exact" two-sided (two-tail) P value from Fisher's test. Prism computes the two-sided P value using the method of summing small P values. Most statisticians seem to recommend this approach, but some programs use a different approach.

If you want to learn more, SISA provides a detail discussion with references. Also see the section on Fisher's test in Categorical Data Analysis by Alan Agresti. It is a very confusing topic, which explains why different statisticians (and so different software companies) use different methods.

One-sided P values

Prism gives you the choice of reporting a [one-sided or two-sided P value][75].

With the chi-square test, the one-sided P value is half the two-sided P value. [Zar] points out (p.503, 5th edition) that there is one extremely rare situation where the one-sided P value can be misleading: If your experimental design is such that you chose both the row totals and the column totals.

Why we use the term "one-sided" and not "one-tailed"? To avoid confusion. The value of chi-square is always positive. To find the P value from chi-square, Prism calculates the probability (under the null hypothesis) of seeing that large a value of chi-square or even larger. So it only looks at the right tail of the chi-square distribution. But a chi-square value can be high when the deviation from the null hypothesis goes in either direction (positive or negative difference between proportions, relative risk greater than or less than 1). So the two-sided P value is actually computed from one tail of the chi-square distribution.

With Fisher's test, the definition of a one-sided P value is not ambiguous. But in most cases, the one-sided P value is not half the two-sided P value.

5.16.2.5 Interpreting results: Attributable risk

Attributable risk

Here are results from an experimental study:

	Progress	No Progress
AZT	76	399
Placebo	129	332

 In this example, disease progressed in 28% of the placebo-treated patients and in 16% of the AZT-treated subjects.

The difference between proportions (P1-P2), the attributable risk, is 28% - 16% = 12%.

The Number Needed to Treat (NNT)

The NNT is simply the reciprocal of the difference between the two proportions. In the example above, the difference between the two proportions is 0.12, so the NNT is 1/0.12= 8.3. For every eight people treated with AZT, you'd expect one more to progress than if all were treated with placebo. Especially, when the difference between proportions is a tiny fraction, it can be easier to understand the NNT than the difference between proportions.

In the example above, the drug is used to treat, so the name Number Needed to Treat is apt. In some cases, there is risk or harm, rather than treatment, and the term Number Needed to Harm (NNH) is used. In other nonclinical situations it is not clear which of two outcomes is better, so neither of those phrases really makes sense. Prism always uses the abbreviation NNT, but it is up to you to interpret the value in the context of the study.

How Prism computes the confidence interval of the attributable risk

On the Options tab of the Contingency table analysis dialog, Prism offers three methods to do the calculation, all explained in Newcombe (1).

- Asymptotic with continuity correction. This is the approximate method used by Prism 6 and earlier. We recommend using it only when needed for compatibility.

- Newcombe/Wilson score

- Newcombe/Wilson score with continuity correction. This is much better than the asymptotic method, so we recommend it. Whether or not you use the continuity correction matters little, but we offer the choice so results will match other programs.

Prism takes the reciprocal of both confidence limits and presents these as the confidence interval of the NNT.

If you choose the asymptotic method and some of the values are zero, Prism adds 0.5 to all cells before calculating the attributable risk and its confidence interval. Prism shows a floating note on the results page when it does this. In this case, we suggest you switch to the Newcombe/Wilson method.

1. Newcombe, R. G. R. (1998). <u>Interval estimation for the difference between independent proportions</u>: comparison of eleven methods. Statistics in Medicine, 17(8), 873–890.

5.16.2.6 Interpreting results: Relative risk

Relative risk

Here are results from an experimental study:

	Progress	No Progress
AZT	76	399
Placebo	129	332

 In this example, disease progressed in 28% of the placebo-treated patients and in 16% of the AZT-treated subjects.

The relative risk is 16%/28% = 0.57. A subject treated with AZT has 57% the chance of disease progression as a subject treated with placebo. The word "risk" is not always appropriate. Think of the relative risk as being simply the ratio of proportions.

How Prism computes the confidence interval of the relative risk

Prism computes the confidence interval of the relative risk using either the Method of Katz (reference 1, the only method used by Prism 6 and earlier) or the Koopman asymptotic score (2), which we recommend because it is more accurate. Choose on the Options tab of the Contingency table dialog. Fagerland (3) reviews the various methods available to compute this confidence interval.

If you choose the method of Katz and some of the values are zero, Prism adds 0.5 to all cells before calculating the relative risk and its confidence interval. Prism shows a floating note on the results page when it does this. In this case, we suggest you switch to the Koopman method.

The order of the two columns matters, the order of rows not so much

Note that it matters how you enter the data. The relative risk would have been different if you had entered the "progress" data in the example above into the second column and the "no progress" data into the first column. For each row, Prism computes the risk by dividing the value in the first column by the sum of the values in the two columns.

After computing the two risks (see prior paragraph), Prism computes the relative risk by dividing the risk in the second row by the risk in the first. But it also reports the reciprocal of that risk. So it really doesn't matter which order you entered the two rows.

References

1. Katz D, Baptista J, Azen SP and Pike MC. Obtaining confidence intervals for the risk ratio in cohort studies. Biometrics 1978; 34: 469–474.

2. Koopman PAR. Confidence intervals for the ratio of two binomial proportions. Biometrics 1984; 40: 513–517.

3. Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. Stat Methods Med Res. SAGE Publications; 2011 Oct 13.

5.16.2.7 Interpreting results: Odds ratio

Odds ratio

Here are the sample data for a case-control study (the first study to link smoking to lung cancer). The investigators chose to study a group of cases with lung cancer and a group of controls without lung cancer. They then asked whether each person had smoked or not (Doll and Hill, British Med. J, 1950, 739-748). The results were:

	Cases (lung cancer)	Control
Smoked	688	650
Never smoked	21	59

With a retrospective case-control data, direct calculations of the relative risk or the difference between proportions should not be performed, as the results are not meaningful. When designing this kind of study, you decide how many cases and controls to study. Those numbers don't have to be equal. Changing the ratio of cases to controls would also change the computed values for the relative risk and difference between proportions. For that reason, it makes no sense to compute or try to interpret these values from case-control data.

In contrast, changing the ratio of cases to controls does not change the expected value of the odds ratio. If the disease or condition you are studying is rare, you can interpret the Odds ratio as an approximation of the relative risk

For the sample data above, the odds of a case being a smoker is 688/21 or 32.8. The odds of a control being a smoker is 650/59 or 11.0. The odds ratio is 32.8/11.0, which is 3.0. Prism reports the value more precisely as 2.974 with a 95% confidence interval ranging from 1.787 to 4.950. You can interpret this odds ratio as a relative risk. The risk of a smoker getting lung cancer is about three times the risk of a nonsmoker getting lung cancer.

How Prism computes the confidence interval of the odds ratio

Prism computes the confidence interval of the odds ratio using computed either using the Woolf logit method (reference 1; the only method used by Prism 6 and earlier) or the Baptista-Pike method (2) which we recommend. Choose on the Options tab of the Contingency table dialog. Fagerland (3) reviews the various methods available to compute this confidence interval.

If any cell has a zero and you choose the Woolf method, Prism adds 0.5 to all cells before calculating the odds ratio and its confidence interval. In this case, we suggest you switch to the Baptista-Pike method.

References

1. Woolf B. On estimating the relation between blood group and disease. Ann Human Gene 1955; 19: 251–253.

2. Baptista J and Pike MC. Exact two-sided confidence limits for the odds ratio in a 2 2 table. J R Stat Soc C Appl Stat 1977; 26: 214–220.

3. Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. Stat Methods Med Res. SAGE Publications; 2011 Oct 13.

5.16.2.8 Interpreting results: Sensitivity and specificity

If your data represent evaluation of a diagnostic test, Prism reports the results in five ways:

Term	Meaning
Sensitivity	The fraction of those with the disease correctly identified as positive by the test.
Specificity	The fraction of those without the disease correctly identified as negative by the test.
Positive predictive value	The fraction of people with positive tests who actually have the condition.
Negative predictive value	The fraction of people with negative tests who actually don't have the condition.
Likelihood ratio	If you have a positive test, how many times more likely are you to have the disease? If the likelihood ratio equals 6.0, then someone with a positive test is six times more likely to have the disease than someone with a negative test. The likelihood ratio equals sensitivity/(1.0-specificity).

The sensitivity, specificity and likelihood ratios are properties of the test.

The positive and negative predictive values are properties of both the test and the population you test. If you use a test in two populations with different disease prevalence, the predictive values will be different. A test that is very useful in a clinical setting (high predictive values) may be almost worthless as a screening test. In a screening test, the prevalence of the disease is much lower so the predictive value of a positive test will also be lower.

Prism computes confidence intervals for all these values using a method you choose on the Options tab for computing the confidence interval of a proportion. Prism offers three methods[542]. We recommend the hybrid Wilson/Brown method.

5.16.2.9 Analysis checklist: Contingency tables

Contingency tables summarize results where you compared two or more groups and the outcome is a categorical variable (such as disease vs. no disease, pass vs. fail, artery open vs. artery obstructed). Read elsewhere to learn about relative risks & odds ratios[557], sensitivity & specificity[560], and interpreting P values[553].

✔ **Are the subjects independent?**

The results of a chi-square or Fisher's test only make sense if each subject (or experimental unit) is independent of the rest. That means that any factor that affects the outcome of one subject only affects that one subject. Prism cannot test this assumption. You must think about the experimental design. For example, suppose that the rows of the table represent two different kinds of preoperative antibiotics and the columns denote whether or not there was a postoperative infection. There are 100 subjects. These subjects are not independent if the table combines results from 50 subjects in one hospital with 50 subjects from another hospital. Any difference between hospitals, or the patient groups they serve, would affect half the subjects but not the other half. You do not have 100 independent observations. To analyze this kind of data, use the Mantel-Haenszel test or logistic regression. Beginning with version 8.3.0, Prism offers both simple logistic regression and multiple logistic regression.

✔ **Are the data unpaired?**

In some experiments, subjects are matched for age and other variables. One subject in each pair receives one treatment while the other subject gets the other treatment. These data should be analyzed by special methods such as McNemar's test[569]. Paired data should not be analyzed by chi-square or Fisher's test.

✓ Is your table really a contingency table?

To be a true contingency table, each value must represent numbers of subjects (or experimental units). If it tabulates averages, percentages, ratios, normalized values, etc. then it is not a contingency table and the results of chi-square or Fisher's tests will not be meaningful. If you've entered observed values on one row (or column) and expected values on another, you do not have a contingency table, and should use a [separate analysis][563] designed for those kind of data.

✓ Does your table contain only data?

The chi-square test is not only used for analyzing contingency tables. It can also be used to compare the observed number of subjects in each category with the number you expect to see based on theory. Prism cannot do this kind of chi-square test. It is not correct to enter observed values in one column and expected in another. When analyzing a contingency table with the chi-square test, Prism generates the expected values from the data – you do not enter them.

✓ Are the rows or columns arranged in a natural order?

If your table has two columns and more than two rows (or two rows and more than two columns), Prism will perform the chi-square test for trend as well as the regular chi-square test. The results of the test for trend will only be meaningful if the rows (or columns) are arranged in a natural order, such as age, duration, or time. Otherwise, ignore the results of the chi-square test for trend and only consider the results of the regular chi-square test.

5.16.2.10 Graphing tips: Contingency tables

Contingency tables are always graphed as bar graph. Your only choices are whether you want the bars to go horizontally or vertically, and whether you want the outcomes to be interleaved or grouped. These choices are available on the Welcome or New Table & Graph dialogs. You can change your mind on the Format Graph dialog, in the Graph Settings tab.

5.16.3   Compare observed and expected distributions

5.16.3.1 How to: Compare observed and expected distributions

This analysis compares the distribution you entered into a parts-of-whole table (observed distribution) with a theoretical distribution you enter into the dialog (expected distribution).

1. Enter the data onto a parts-of-whole table

Table format: Parts of whole	A # of seeds
	Y
1 Round and yellow	315
2 Round and green	108
3 Angular and yellow	101
4 Angular and green	32

Enter the actual number of objects or events. The results will be meaningless if you enter normalized values, rates or percentages. These are actual data from one of Mendel's famous experiments. I obtained the data from H. Cramer. *Mathematical methods of statistics.* Princeton University Press, 1999.

2. Enter the expected values

Click Analyze, and choose *Compare observed distribution with expected* in the Parts of whole section. These values were computed by multiplying a proportion predicted by Hardy-Weinberg Mendelian genetics (9/16 or 0.5625 for the first category) times the number of peas used in the experiment. You can also enter the percentages directly by selecting an option on the dialog.

Enter the expected values. You can choose to enter the actual number of objects or events expected in each category, in which case the total of the expected values must equal the total of the observed data you entered on the data table. Or you can choose to enter percentages, in which case they must total 100. In either case, it is ok to enter fractional values.

In this example, the expected values are not integers. That's ok. That is the average expectation if you did a large number of experiments. In any one experiment, of course, the number of peas of each category must be an integer. These values are computed based on Mendelian genetics. For

example, the theory predicts that 9/16 of peas would be in the first category. Multiply that fraction by the total number of peas used in this experiment to get the expected values.

3. Choose the test

If you entered more than two rows of data (as in the example above), you'll have no choice. Prism will perform the chi-square goodness-of-fit test.

If you entered only two rows of data, you can also choose the binomial test, which we strongly recommend. With only two categories, the chi-square test reports P values that are too small. This is a huge issue with small data sets, but the discepancy exists even with sample sizes in the hundreds. Use the binomial test.

4. Interpret the P value

The results table summarizes the data, reports the value of chi-square and its df (if you picked the chi-square test), and states the P value. The null hypothesis is that the observed data are sampled from a populations with the expected frequencies. The P value answers this question:

> Assuming the theory that generated the expected values is correct, what is the probability of observing such a large discrepancy (or larger) between observed and expected values?

A small P value is evidence that the data are not sampled from the distribution you expected. In this example, the P value is large (0.93) so the data provide no evidence of a discrepancy between the observed data and the expected values based on theory.

5.16.3.2 How the chi-square goodness of fit test works

The null hypothesis is that the observed data are sampled from a populations with the expected frequencies. The chi-square test combines the discrepancies between the observed and expected values.

How the calculations work:

1. For each category compute the difference between observed and expected counts.

2. Square that difference and divide by the expected count.

3. Add the values for all categories. In other words, compute the sum of (O-E)2/E.

4. Use a computer program to calculate the P value. You need to know that the number of degrees of freedom equals the number of categories minus 1.

The null hypothesis is that the observed data are sampled from a populations with the expected frequencies. The P value answers this question:

> Assuming the theory that generated the expected values is correct, what is the probability of observing such a large discrepancy (or larger) between observed and expected values?

A small P value is evidence that the data are not sampled from the distribution you expected.

The Yates' correction

When there are only two categories, some statisticians recommend using the Yates' correction. This would reduce the value of chi-square and so would increase the P value. With large sample sizes, this correction makes little difference. With small samples, it makes more difference. Statisticians disagree about when to use the Yates' correction, and Prism does not apply it.

With only two categories, it is better to use the binomial test[566], which gives an exact result instead of either form of the chi-square calculation, which is only an approximation.

5.16.3.3 The binomial test

When to use the binomial test rather than the chi-square test

The binomial test is an exact test to compare the observed distribution to the expected distribution when there are only two categories (so only two rows of data were entered). In this situation, the chi-square is only an approximation, and we suggest using the exact binomial test instead.

Example

Assume that your theory says that an event should happen 20% of the time. In fact, in an experiment with 100 repetitions, that event happened only 7 times. You expected the event to occur 20 times (20% of 100) but it only occurred 7 times. How rare a coincidence is that? That is the question the binomial test answers.

Create a parts-of-whole table, and enter 7 into row 1 and 93 into row 2, and label the rows if you like. Click Analyze, and choose Compare observed distribution with expected in the Parts of whole section. Enter the expected values (20 and 80) and choose the binomial test (rather than chi-square)

Prism reports both one- and two-tail P values.

One-tail P value

The one-tail P value (also called a one sided P value) is straightforward. The null hypothesis is that the expected results are from a theory that is correct. So the P value answers the question:

If the true proportion is 20%, what is the chance in 100 trials that you'll observe 7 or fewer of the events?

You need to include the "or fewer" because it would have been even more surprising if the number of events in 100 trials was any value less than seven.

The one-tail P value for this example is: 0.0003.

If the observed value is less than the expected value, Prism reports the one-tail P value which is the probability of observing that many events or fewer. If the observed value is greater than the expected value, Prism reports the one-tail P value which is the probability of observing that many events or more.

Two-tail P value

The two-tail P value is a bit harder to define. In fact, there are (at least) three ways to define it.

Prism uses the third definition below, and this is the P value Prism uses when it creates the summary (* or **...).

- Double the one-tail P value. Twice 0.0002769 equals 0.0005540 That seems sensible, but that method is not used. Unless the expected proportion is 50%, the asymmetry of the binomial distribution makes it unwise to simply double the one-tail P value.

- Equal distance from expected. The theory said to expect 20 events. We observed 7. The discrepancy is 13 (20-7). So the other tail of the distribution should be the probability of obtaining 20+13=33 events or more. The two-tailed P value, computed this way, is the probability of obtaining 7 or less (0.0002769; the same as the one-tail P value) plus the probability of obtaining 33 or more (0.001550441) which means the two-tail P value equals 0.00182743..

- [Method of small P values](). To define the second tail with this method, we don't go out the same distance but instead start the second tail at an equally unlikely value. The chance of observing exactly 7 out of 100 events when the true probability is 0.20 equals 0.000199023. The probability of obtaining 33 events (how the second tail was defined in the other method) is higher: 0.000813557. The chance of obtaining 34 events is also higher. But the chance of observing 35 events is a bit lower (0.000188947). The second tail, therefore, is defined as the chance of observing 35 or more events. That tail is 0.0033609. The two tail P value therefore is 0.00061307. This is the method that Prism uses.

The distinction between the second and third methods is subtle. The first tail is unambiguous. It starts at 7 and goes down to zero. The second tail is symmetrical, but there are two ways to define this. The second method is symmetrical around the counts. In other words, the border for that tail (33) is as far from the expected value of 20 as is the observed value of 7 (33-20=20-7). The third method is symmetrical regarding probabilities. Given the assumption that the true probability is 20% so we expect to observe 20, the chance of observing 7 events is about the same as the chance of observing 35. So the second tail is the probability of observing 35 or more events.

If the expected probability is 0.5, the binomial distribution is symmetrical and all three methods give the same result. When the expected probability is 0.5, then the binomial test is the same as the *sign test*.

5.16.3.4 McNemar's test

Overview of McNemar's test

In the usual kind of case-control study, the investigator compares a group of controls with a group of cases. As a group, the controls are supposed to be similar to the cases (except for the absence of disease). Another way to perform a case-control study is to match individual cases with individual controls based on age, gender, occupation, location and other relevant variables. This is the kind of study McNemar's test is designed for.

Displaying and analyzing data from matched case-control studies on an ordinary contingency table obscures the fact that the cases and controls were matched. Matching makes the experiment stronger, so the analysis ought to take it into account.

Example

Here are some sample data:

		Control		
		+	**-**	**Total**
Case	**+**	13	25	38
	-	4	92	96
	Total	17	117	134

The investigators studied 134 cases and 134 matched controls, for a total of 268 subjects. Each entry in the table represents one pair (a case and a control). The + and - labels refer to people who were, or were not, exposed to the putative risk factor or exposure.

This is not a contingency table, so the usual analyses of contingency tables would not be helpful. It turns out that the odds ratio can be computed quite simply. The 13 pairs in which both cases and controls were exposed to the risk factor provide no information about the association between risk factor and disease. Similarly, the 92 pairs in which neither case nor control were exposed to risk factor provide no information. The odds ratio is calculated as the ratio of the other two values: pairs in which the case was exposed to the risk factor but the control was not divided by pairs in the control was exposed to the risk

factor but the case was not. In this example, the odds ratio for the association between risk factor and disease is 25/4 = 6.25. The equation for the confidence interval is complicated (see page 286 of S. Selvin, Statistical Analysis of Epidemiologic Data, 2nd edition). The 95% confidence interval for the odds ratio ranges from 2.158 to 24.710.

Computing the P value with Prism using the Binomial test

When you read about McNemar's test, most books explain how to do a chi-square calculation. Prism won't do that, but we offer a free web calculator that does. The binomial test asks the same question, but is more accurate, especially with small studies. Follow these steps with Prism:

1. Create a parts-of-whole data table.

2. Enter the numbers of discordant pairs in the first two rows of column A. For the example, enter 25 and 4.

3. Click Analyze and choose the analysis that compares observed and expected counts.

4. Choose to enter the expected values as percentages, and enter 50 as both expected percentages.

5. Choose the binomial test, rather than the chi-square test.

6. For the sample data, the P value is less than 0.0001. The P value answers this question: If there really were no association between disease and risk factor, what is the chance that the two values entered into this analysis would be as far apart as they are, or even further?

Computing the P value with QuickCalcs using McNemar's test

GraphPad's free web QuickCalc computes McNemar's test using a chi-square approximation. Call the two discrepant numbers (25 and 4) R and S. QuickCalc computes chi-square using this equation:

$$\chi^2 = \frac{\left(\left|R-S\right|-1\right)^2}{R+S}$$

For this example, chi-square=13.79, which has one degree of freedom. The two-tailed P value is 0.0002. If there were really no association between risk factor and disease, there is a 0.02 percent chance that the observed odds ratio would be so far from 1.0 (no association).

The equation above uses the Yates' correction (the "-1" in the equation above). Sometimes this correction is shown as "- 0.5". If you choose the chi-square approach with Prism, no Yates' correction is applied at all. Rather than choosing the chi-square approach (which is an approximation) and worrying about whether to apply the Yates' correction, and which correction to use to, we recommend that you choose the binomial test, which is an exact test.

5.16.3.5 Don't confuse with related analyses

The chi-square goodness of fit test can easily be confused with other tests. Here are some distinctions to avoid any confusion.

Relationship to the chi-square analysis of contingency tables

Note that the chi-square test is used in two quite different contexts.

One use is to compare the observed distribution with an expected distribution generated by theory.

Another use is to analyze a contingency table[545]. In this analysis, the expected values are computed from the data, and not from an external theory.

Relationship to normality tests

Normality tests compare the observed distribution of a continuous variable, with a theoretical distribution generated by the Gaussian distribution. Prism offers three ways to do this comparison, all offered as part of the Column statistics analysis.[200]

Relationship to the Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test[317] can be used as a nonparametric method to compare two groups of continuous data. It compares the two observed cumulative frequency distributions, and does not compare either observed distribution to an expected distribution.

5.16.3.6 Analysis Checklist: Comparing observed and expected distributions

The chi-square and binomial tests compare an observed categorical distribution with a theoretical distribution.

✓ **Are the values entered the exact number of objects or events ?**

The results can be interpreted only if you entered the actual number of objects or events. The results will be meaningless if you enter normalized values, rates or percentages.

✓ **Do the expected values come from theory?**

The whole point of this analysis is to compare an observed distribution with a distribution expected by theory. It does not compare two observed distributions.

5.17   Survival analysis

Survival curves plot the results of experiments where the outcome is time until death (or some other one-time event). Prism can use the Kaplan-Meier method to create survival curves from raw data, and can compare survival curves.

5.17.1   How to: Survival analysis

Enter topic text here.

5.17.1.1 Key concepts. Survival curves

In many clinical and animal studies, the outcome is survival time. The goal of the study is to determine whether a treatment changes survival. Prism creates survival curves, using the product limit method of Kaplan

and Meier, and compares survival curves using both the logrank test and the Gehan-Wilcoxon test.

Censored data

Creating a survival curve is not quite as easy as it sounds. The difficulty is that you rarely know the survival time for each subject.

- Some subjects may still be alive at the end of the study. You know how long they have survived so far, but don't know how long they will survive in the future.

- Others drop out of the study -- perhaps they moved to a different city or wanted to take a medication disallowed on the protocol. You know they survived a certain length of time on the protocol, but don't know how long they survived after that (or do know, but can't use the information because they weren't following the experimental protocol). In both cases, information about these patients is said to be censored.

You definitely don't want to eliminate these censored observations from your analyses -- you just need to account for them properly.The term "censored" seems to imply that the subject did something inappropriate. But that isn't the case. The term "censored" simply means that you don't know, or can't use, survival beyond a certain point. Prism automatically accounts for censored data when it creates and compares survival curves.

Not just for survival

The term *survival curve* is a bit restrictive as the outcome can be any well-defined end point that can only happen once per subject. Instead of death, the endpoint could be occlusion of a vascular graft, first metastasis of a tumor, or rejection of a transplanted kidney. The event does not have to be dire. The event could be restoration of renal function, discharge from a hospital, or graduation.

Analyzing other kinds of survival data

Some kinds of survival data are better analyzed with nonlinear regression. For example, don't use the methods described in this section to analyze cell survival curves plotting percent survival (Y) as a function of various doses of radiation (X). The survival methods described in this chapter are only useful if X is time, and you know the survival time for each subject.

Proportional hazards regression

The analyses built in to Prism can compare the survival curves of two or more groups. But these methods (logrank test, Gehan-Breslow-Wilcoxon test) cannot handle data where subjects in the groups are matched, or when you also want to adjust for age or gender or other variables. For this kind of analysis, you need to use proportional hazards regression, which Prism does not do.

5.17.1.2 How to: Survival analysis

1. Create a survival table

From the Welcome or New Table dialog, choose the Survival tab.

If you aren't ready to enter your own data yet, choose to use sample data, and choose one of the sample data sets.

2. Enter the survival times

Enter each subject on a separate row in the table, following these guidelines:

- Enter time until censoring or death (or whatever event you are tracking) in the X column. Use any convenient unit, such as days or months. Time zero does not have to be some specified calendar date; rather it is defined to be the date that each subject entered the study so may be a different calendar date for different subjects. In some clinical studies, time zero spans several calendar years as patients are enrolled. You have to enter duration as a number, and cannot enter dates directly.

- Optionally, enter row titles to identify each subject.

- Enter "1" into the Y column for rows where the subject died (or the event occurred) at the time shown in the X column. Enter "0" into the rows where the subject was censored[572] at that time. Every subject in a survival study either dies or is censored.

- Enter subjects for each treatment group into a different Y column. Place the X values for the subjects for the first group at the top of the table with the Y codes in the first Y column. Place the X values for the

second group of subjects beneath those for the first group (X values do not have to be sorted, and the X column may well contain the same value more than once). Place the corresponding Y codes in the second Y column, leaving the first column blank. In the example below, data for group A were entered in the first 14 rows, and data for group B started in row 15.

Table format: Survival		X Days after randomization	A Control	B Treated
		X	Y	Y
1	AB	34	1	
2	GT	66	1	
3	RF	64	0	
4	ED	89	1	
5	CD	98	1	
6	TT	111	1	
7	RV	123	1	
8	TV	145	1	
9	VD	134	1	
10	BM	145	0	
11	UJ	88		1
12	UV	143		1
13	IT	76		1
14	TO	111		0
15	AT	95		0
16	TU	134		1
17	XX	167		1
18	XY	198		1
19	XO	211		1
20	HO	234		1

- If the treatment groups are intrinsically ordered (perhaps increasing dose) maintain that order when entering data. Make sure that the progression from column A to column B to column C follows the natural order of the treatment groups. If the treatment groups don't have a natural order, it doesn't matter how you arrange them.

- Double check that the number of rows with data entered in a column matches the number of people/animals/whatever i that treatment group.

Entering data for survival studies can be tricky. See answers to common questions[576], an example of a clinical study[577], and an example of an animal study[579].

3. View the graph and results

After you are done entering your data, go to the new graph to see the completed survival curve. Go to the automatically created results sheet to see the results of the logrank test, which compares the curves (if you entered more than one data set).

Interpreting results: Kaplan-Meier curves[587]

Interpreting results: Comparing two survival curves[592]

Interpreting results: Comparing three or more survival curves[600]

Analysis checklist: Survival analysis[192]

Note that survival analysis works differently than other analyses in Prism. When you choose a survival table, Prism automatically analyzes your data. You don't need to click the Analyze button

5.17.1.3 Q & A: Entering survival data

How do I enter data for subjects still alive at the end of the study?

Those subjects are said to be censored. You know how long they survived so far, but don't know what will happen later. X is the # of days (or months…) they were followed. Y is the code for censored[572] observations, usually zero.

What if two or more subjects died at the same time?

Each subject must be entered on a separate row. Enter the same X value on two (or more) rows.

How do I enter data for a subject who died of an unrelated cause?

Different investigators handle this differently. Some treat a death as a death, no matter what the cause. Others treat death of an unrelated cause to be a censored observation. Ideally, this decision should be made in the study design. If the study design is ambiguous, you should decide how to handle these data before unblinding the study.

Do the X values have to be entered in order?

No. You can enter the rows of data in any order you want. It just matters that each Y value (code) be on the same row as the appropriate X value.

How does Prism distinguish between subjects who are alive at the end of the study and those who dropped out of the study?

It doesn't. In either case, the observation is censored. You know the patient was alive and on the protocol for a certain period of time. After

that you can't know (patient still alive), or can't use (patient stopped following the protocol) the information. Survival analysis calculations treat all censored subjects in the same way. Until the time of censoring, censored subjects contribute towards calculation of percent survival. After the time of censoring, they are essentially missing data.

I already have a life-table showing percent survival at various times. Can I enter this table into Prism?

No. Prism only can analyze survival data if you enter survival time for each subject. Prism cannot analyze data entered as a life table.

Can I enter a starting and ending date, rather than duration?

Yes. When you create a new survival table, you can choose to enter starting and ending dates, rather than number of days.

How do I handle data for subjects that were "enrolled" but never treated?

Most clinical studies follow the "intention to treat" rule. You analyze the data assuming the subject got the treatment they were assigned to receive, even if the treatment was never given. This decision, of course, should be made as part of the experimental design.

If the subject died right after enrollment, should I enter the patient with X=0?

No. The time must exceed zero for all subjects. If you enter X=0, Prism simply ignores that row. More on survival curves with X=0.

5.17.1.4 Example of survival data from a clinical study

Here is a portion of the data collected in a clinical trial:

Enrolled	Final date	What happened	Group
07-Feb-98	02-Mar-02	Died	Treated
19-May-98	30-Nov-98	Died	Treated
14-Nov-98	03-Apr-02	Died	Treated
01-Dec-98	04-Mar-01	Died	Control
04-Mar-99	04-May-01	Died	Control

Enrolled	Final date	What happened	Group
01-Apr-99	09-Sep-02	Still alive, study ended	Treated
01-Jun-99	03-Jun-01	Moved, off protocol	Control
03-Jul-99	09-Sep-02	Still alive, study ended	Control
03-Jan-00	09-Sep-02	Still alive, study ended	Control
04-Mar-00	05-Feb-02	Died in car crash	Treated

And here is how these data looked when entered in Prism.

Table format: Survival		Days X	Control Y	Treated Y
1	Title	1484		1
2	Title	195		1
3	Title	1236		1
4	Title	824	1	
5	Title	92	1	
6	Title	1257		0
7	Title	733	0	
8	Title	1164	0	
9	Title	980	0	
10	Title	703		0

Prism does not allow you to enter beginning and ending dates. You must enter elapsed time. You can calculate the elapsed time in Excel (by simply subtracting one date from the other; Excel automatically presents the results as number of days).

Unlike many programs, you don't enter a code for the treatment (control vs. treated, in this example) into a column in Prism. Instead you use separate columns for each treatment, and enter codes for survival or censored into that column.

There are three different reasons for the censored observations in this study.

- Three of the censored observations are subjects still alive at the end of the study. We don't know how long they will live.

- Subject 7 moved away from the area and thus left the study protocol. Even if we knew how much longer that subject lived, we couldn't use the information since he was no longer following the study protocol. We know that subject 7 lived 733 days on the protocol and either don't know, or know but can't use the information, after that.

- Subject 10 died in a car crash. Different investigators handle this differently. Some define a death to be a death, no matter what the cause. Others would define a death from a clearly unrelated cause (such as a car crash) to be a censored observation. We know the subject lived 703 days on the treatment. We don't know how much longer he would have lived on the treatment, since his life was cut short by a car accident.

Note that the order of the rows is entirely irrelevant to survival analysis. These data are entered in order of enrollment date, but you can enter in any order you want.

5.17.1.5 Example of survival data from an animal study

This example is an animal study that followed animals for 28 days after treatment. All five control animals survived the entire time. Three of the treated animals died, at days 15, 21 and 26. The other two treated animals were still alive at the end of the experiment on day 28. Here is the data entered for survival analysis.

Table format: Survival	X	A	B
	Days	Control	Treated
	X	Y	Y
1 Title	28	0	
2 Title	28	0	
3 Title	28	0	
4 Title	28	0	
5 Title	28	0	
6 Title	15		1
7 Title	21		1
8 Title	26		1
9 Title	28		0
10 Title	28		0

Note that the five control animals are each entered on a separate row, with the time entered as 28 (the number of days you observed the animals) and with Y entered as 0 to denote a censored observation. The

observations on these animals is said to be censored because we only
know that they lived for at least 28 days. We don't know how much longer
they will live because the study ended.

The five treated animals also are entered one per row, with Y=1 when
they died and Y=0 for the two animals still alive at the end of the study.

Data for each animal was entered on one row. So the number of rows
with data in a column equals the number of animals in that treatment
group.

5.17.1.6 Analysis choices for survival analysis

Prism analyzes survival curves without you having to choose any analysis

The survival analysis is unique in Prism. When you enter data on an
survival table, Prism automatically performs the analysis. You don't need
to click Analyze or make any choices on the parameters dialog.

From the results, you can click the analysis parameters button to bring up
the parameters dialog, if you want to make any changes.

Parameters: Survival Curve   ✕

Input

The X values are time. The Y values are coded as follows:

Death/Event: [1] ▲▼

Censored subject: [0] ▲▼

Note: All other Y values are ignored

Curve comparison

Calculations to compare two groups:

☑ Logrank (Mantel-Cox test)

☑ Gehan-Breslow-Wilcoxon test (extra weight for early time points)

Calculations to compare three or more groups:

☑ Logrank   Match SPSS and SAS (recommended) ⌄

☑ Logrank test for trend   Match SPSS and SAS (recommended) ⌄

☑ Gehan-Breslow-Wilcoxon test (extra weight for early time points)

Style

Tabulate: [Percent survival ⌄]

Express fraction survival error bars as:

◉ None   ○ SE   ○ 95% CI [Asymmetrical (more accurate; recommended) ⌄]

☑ Show censored subjects on graph.

Output

Show this many significant digits (for everything except P values): [4] ▲▼

P value style: [GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 ⌄] N = [5] ▲▼

☐ Use these settings as the default for future survival analyses

[Learn]   [Cancel]   [OK]

Input

The default choices are to use the code '1' for deaths and '0' for censored subjects, and these are almost universal. But some institutions use the opposite convention. The codes must be integers.

Curve comparison calculations: Comparing two survival curves

Prism can compare two survival curves using two methods. Choose either one, or both.

- **The logrank test**. There are two ways to compute this test. The two are almost equivalent, but can differ a bit in how they deal with multiple

deaths at exactly the same time point. Prism uses the Mantel-Haenszel approach but uses the name 'logrank' which is commonly used for both approaches. This method is also called the Mantel-Cox method.

- The **Gehan-Breslow-Wilcoxon test**. This method gives more weight to deaths at early time points, which makes lots of sense. But the results can be misleading when a large fraction of patients are censored at early time points. In contrast, the logrank test gives equal weight to all time points.

The logrank test is more standard. It is the more powerful of the two tests if the assumption of proportional hazards is true. Proportional hazards means that the ratio of hazard functions (deaths per time) is the same at all time points. One example of proportional hazards would be if the control group died at twice the rate as treated group at all time points.

The Gehan-Breslow-Wilcoxon test does not require a consistent hazard ratio, but does require that one group consistently have a higher risk than the other.

If the two survival curves cross, then one group has a higher risk at early time points and the other group has a higher risk at late time points. This could just be a coincidence of random sampling, and the assumption of proportional hazards could still be valid. But if the sample size is large, neither the logrank nor the Wilcoxon-Gehan test rests are helpful when the survival curves cross near the middle of the the time course.

If in doubt, report the logrank test (which is more standard). Choose the Gehan-Breslow-Wilcoxon test only if you have a strong reason to do so.

Curve comparison calculations: Comparing three or more survival curves

With three or more data sets, Prism offers three ways to compare survival curves. For the details on the first and third choices, look in the previous section.

- Logrank test. This is used most often.

- Logrank test for trend. The test for trend is only relevant when the order of groups (defined by data set columns in Prism) is logical. Examples would be if the groups are different age groups, different disease severities, or different doses of a drug. The left-to-right order of data sets in Prism must correspond to equally spaced ordered

categories. If the data sets are not ordered (or not equally spaced), it makes no sense to choose the logrank test for trend.

- The Gehan-Breslow-Wilcoxon test. This method gives more weight to the earlier time points. Choose it only if you have a strong reason to do so. Prism computes using equation 10.2 from Machin's text referenced below.

With three or more groups, Prism offers a choice of two methods for computing the P value

Match Prism 5 and earlier (conservative)

Prism 5 and earlier computed a P value to compare three or more groups using a conservative method shown in many text books. For each curve, this method computes a chi-square value by comparing the observed and expected number of deaths. It then sums those chi-square values to get an overall chi-square, from which the P value is determined. Here is it is as an equation, where O_i is the observed number of deaths in curve i, and E_i is the expected number of deaths:

$$\text{Chi square} = \sum_{\text{all curves}} \frac{(O_i - E_i)^2}{E_i}$$

This conservative method is documented in Machin (1), is easy to understand and works OK. The problem is that the P value is too high (that is what "conservative" means). Choose this method only if you want results to match results from prior versions of Prism.

Choose this method unless it is really important to you to match results from prior versions of Prism. Otherwise, choose the recommended method to match SPSS and SAS.

Match SPSS and SAS (recommended)

Prism can also compute the P value using a different method, explained in detail in the manuals for SPSS and NCSS. The method can only be understood in terms of matrix algebra. Like the conservative method, it also computes a chi-square value. For both methods, the number of degrees of freedom equals the number of groups minus 1. The difference is that the chi-square value is higher, so the P value is lower.

Style

The choices on how to tabulate the results (percents or fractions, death or survival), can also be made on the Format Graph dialog.

If you choose to plot 95% confidence intervals, Prism gives you two choices. The default is a transformation method, which plots asymmetrical confidence intervals. The alternative is to choose symmetrical Greenwood intervals. The asymmetrical intervals are more valid, and we recommend choosing them.

The only reason to choose symmetrical intervals is to be consistent with results computed by Prism version 4 and earlier. Note that the 'symmetrical' intervals won't always plot symmetrically. The intervals are computed by adding and subtracting a calculated value from the percent survival. At this point the intervals are always symmetrical, but may go below 0 or above 100. In these cases, Prism trims the intervals so the interval cannot go below 0 or above 100, resulting in an interval that appears asymmetrical.

A checkbox lets you decide to plot censored observations or not. The exception is when the largest X value (time) is censored. This is always shown, regardless of whether you check this option or not.

Output

Choose how many digits of precision to show and the format of P values.

Reference

David Machin, Yin Bun Cheung, Mahesh Parmar, [Survival Analysis: A Practical Approach](), 2nd edition, IBSN:0470870400.

5.17.2 Graphing survival data

Enter topic text here.

5.17.2.1 Customizing survival graphs

Prism automatically analyzes and graphs your survival data. When you first go look at the graph, Prism pops up a dialog where you can customize the graph. The choices are self-explanatory.
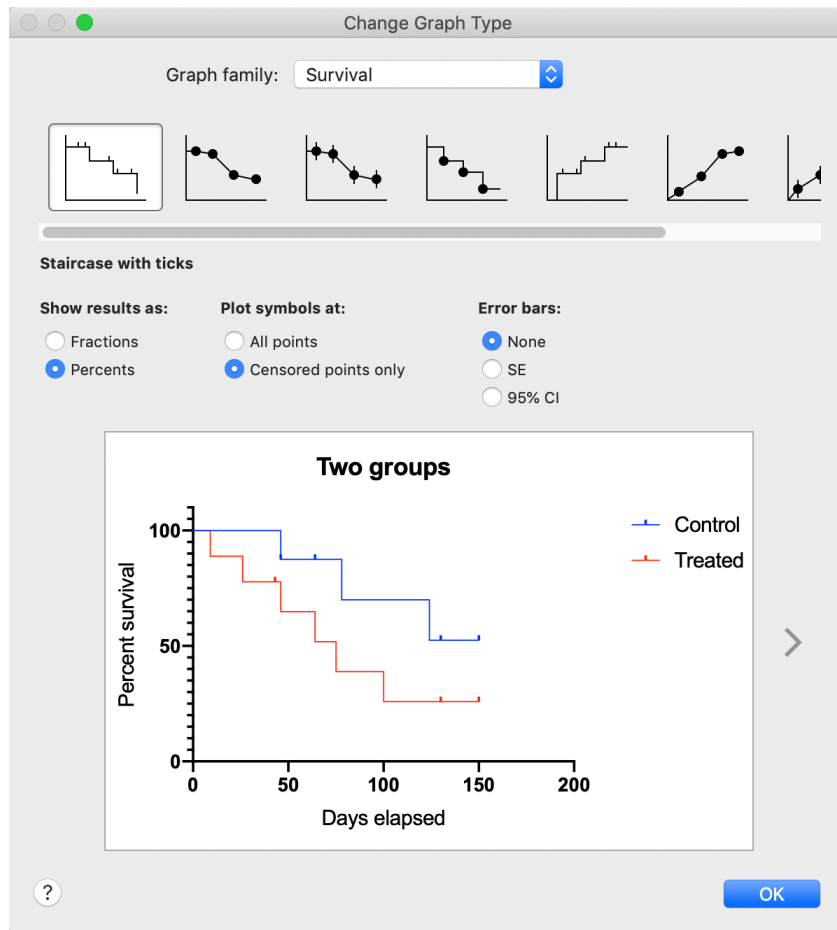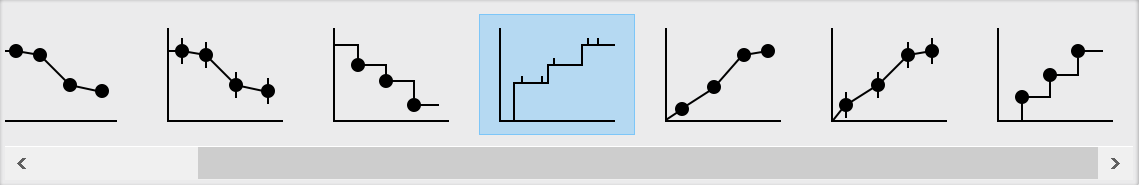
5.17.2.2 Cumulative incidence graphs

Survival graphs start at 100% (or1.0) and go down to zero.

Cumulative incidence graphs start at 0 and go up to a maximum of 100 (percent) or 1.0 (fraction). Choose a cumulative incidence graph at the top right of the Change graph type dialog.

5.17.3 Intepreting results: Survival analysis

5.17.3.1 Interpreting results: Survival proportions

The analysis tab for survival proportions or percentages is invisible by default. It is used to make a graph, but you won't see the analysis tab until you turn it on in the analysis tab menu.
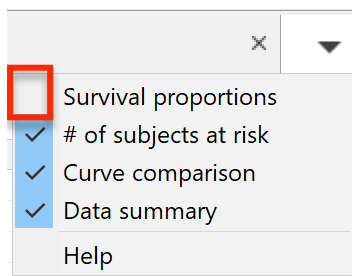
Survival proportions

Prism calculates survival proportions using the product limit (Kaplan-Meier) method. For each X value (time), Prism shows the fraction (or percentage) still alive (or the fraction or percentage already dead, if you chose to begin the curve at 0.0 rather than 1.0). This table contains the numbers used to graph survival vs. time.

The calculations take into account censored observations. Subjects whose data are censored[572]--either because they left the study, or because the study ended -- can't contribute any information beyond the time of censoring. This makes the computation of survival percentage somewhat tricky. While it seems intuitive that the curve ought to end at a survival fraction computed as the total number of subjects who died divided by the total number of subjects, this is only correct if there are no censored data. If some subjects were censored, then subjects were not all followed for the same duration, so computation of the survival fraction is not straightforward (and what the Kaplan-Meier method is for).

If the time of death of some subjects is identical to the time of censoring for others, Prism does the computations assuming the deaths come first.

The results and graphs are labeled "Percent survival" but a more accurate term would be "Probability of survival". If there are any censored observations, the graph and table doesn't show percent survival in the experiment but rather probability of survival at any time given the censoring.

Confidence intervals of survival percentages

Prism reports the uncertainty of the fractional survival as a standard error or 95% confidence intervals. Standard errors are calculated by the method of Greenwood.

You can choose between two methods of computing the 95% confidence intervals:

- Asymmetrical method (recommended). It is computed using the log-log transform method, which has also been called the exponential Greenwood formula. It is explained on page 42 and page 43 of Machin (reference below). You will get the same results from the survfit R function by setting error to *Greenwood* and conf.type to *log-log*. These intervals apply to each time point. The idea is that at each time point, there is a 95% chance that the interval includes the true population survival. We call the method asymmetrical because the distance that the interval extends above the survival time does not usually equal the distance it extends below. These are called *pointwise confidence limits.* It is also possible (but not by Prism) to compute *confidence bands* that have a 95% chance of containing the entire population survival curve. These confidence bands are wider than pointwise confidence limits.

- Symmetrical method. These intervals are computed as 1.96 times the standard error in each direction. In some cases the confidence interval calculated this way would start below 0.0 or end above 1.0 (or 100%). In these cases, the error bars are clipped to avoid impossible values. We provide this method only for compatibility with older versions of Prism, and don't recommend it.

How the Kaplan-Meier method works

The Kaplan-Meier method is logically simple. For each day (or week, however you scale the X values), it first computes the fraction of patients

who are alive at the start of that day who survived until the end of that day. To do this, it simply divides the number alive at the end of that day by the number alive at the beginning of that day (excluding any who were censored on that day from both the numerator and the denominator).

Then it computes the fraction of patients who survived from day 0 until the end of each particular day. To do this, it multiplies the fraction of patients who survived Day 1 by the fraction of patients who were alive at the beginning of day 2 that survived until the end of Day 2 (excluding any censored), and then by the fraction of patients who were alive at the beginning of Day 3 that survived until the end Day 3 (excluding any censored), and so on, until you eventually multiply by the fraction who survived until the of Day k. The result of multiplying all these probabilities is the fraction of all patients who survived until the end of Day k, and this is the survival fraction that gets tabulated and graphed by Prism. This method automatically accounts for censored patients, because both the numerator and the denominator are reduced on the day a patient is censored. Because it calculates the product of many survival fractions, this method is also called the *product-limit method*.

Reference

David Machin, Yin Bun Cheung, Mahesh Parmar, <u>Survival Analysis: A Practical Approach</u>, 2nd edition, IBSN:0470870400.

5.17.3.2 What determines how low a survival curve gets?

If there are no censored observations and all subjects die

If you follow each subject until the event occurs (the event is usually death, but survival curves can track time until any one-time event), then the curve will eventually reach 0. At the time (X value) when the last subject dies, the percent survival is zero.

If all subjects are followed for exactly the same amount of time

If all subjects are followed for the same amount of time, the situation is easy. If one third of the subjects are still alive at the end of the study, then the percent survival on the survival curve will be 33.3%.

If some subjects are censored along the way

If the data for any subjects are censored, the bottom point on the survival curve will *not* equal the fraction of subjects that survived.

Prior to censoring, a subject contributes to the fractional survival value. Afterward, she or he doesn't affect the calculations. At any given time, the fractional (or percent) survival value is the proportion of subjects *followed that long* who have survived.

Subjects whose data are censored --either because they left the study, or because the study ended--can't contribute any information beyond the time of censoring. You don't know whether or not they would have died after the time of censoring (or do know, but can't use the information because the experimental protocol was no longer being followed). So if any subjects are censored before the last time shown on the survival curve's X-axis, the final survival percentage shown on the survival graph will not correspond to the actual fraction of the subjects who survived. That simple survival percentage that you can easily compute by hand is not meaningful, because not all the subjects were not followed for the same amount of time.

When will the survival curve drop to zero?

If the survival curve goes all the way down to 0% survival, that does not mean that every subject in the study died. Some may have censored data at earlier time points (either because they left the study, or because the study ended while they were alive). The survival percentage will drop to zero when the event at the last time point is a death (or whatever outcome you track) and not a censoring. If your data are sorted by X value (which Prism can do using Edit..Sort), the curve will descend to 0% survival if the last Y value is 1 (death), and will end above 0% if the last Y value is 0 (censored).

In the example below, four of the ten subjects die. But the survival curve descends to zero, not to 60%. Why? Because six subjects were censored between 1 and 27 months. We have no idea what would have happened had they stayed in the study until month 28. Since we don't know if they would have lived or died, their data simply doesn't count after the time of censoring (but definitely counts before that). At time 27, only one subject is still being followed, and she or he died at month 28, dropping the percent survival down to zero.

Months elapsed	Status
1.00	0
4.00	0
13.00	0
14.00	1
16.00	0
19.00	1
20.00	0
26.00	1
27.00	0
28.00	1
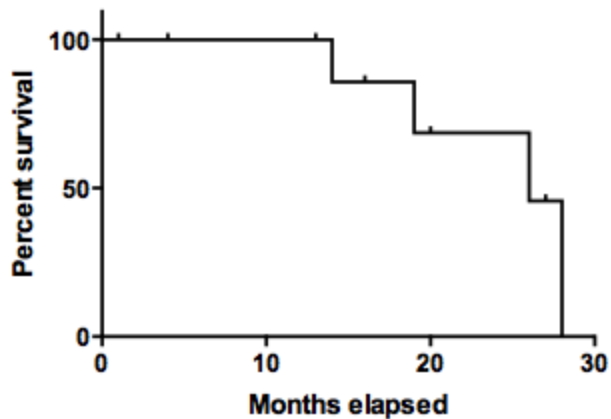
5.17.3.3 Interpreting results: Number at risk

Number of subjects at risk at various times

Since the number at risk applies to a range of days, and not to a single day, the table is a bit ambiguous. The values tabulated are the number of subjects at risk at the start of that day (which can be different than those at risk at the end of that day).

Here are values of that table for the control group of the sample data (comparing two groups) that you can choose from Prism's Welcome dialog.

Days	Standard
0	8
46	8
64	6
78	5
124	4
130	3
150	2

The experiment starts with 8 subjects receiving standard therapy. There are still 8 subjects at risk at the beginning of day 46, and this is shown on

the table. During day 46, one of the patients receiving standard therapy died and the data for another was censored. So 6 patients remain at risk between day 46 and the beginning of day 64, and this tabulated. On day 64, data on another patient is censored, so 5 patients are at risk until the beginning of day 78. The next death occurs on day 78, so 4 subjects are at risk until the beginning of day 124....

Prism does not graph this table automatically. If you want to create a graph of number of subjects at risk over time, follow these steps:

1. Go to the results tab for number of subjects at risk.

2. Click New, and then Graph of existing data.

3. Choose the XY tab and a graph with no error bars.

4. Change the Y-axis title to "Number of subjects at risk" and the X-axis title to "Days".

5.17.3.4 Interpreting results: P Value

Interpreting the P value

The P value tests the null hypothesis that the survival curves are identical in the overall populations. In other words, the null hypothesis is that the treatment did not change survival.

The P value answers this question:

If the null hypothesis is true, what is the probability of randomly selecting subjects whose survival curves are as different (or more so) than was actually observed?

Note that the P value is based on comparing entire survival curves, not on comparing only the median survival.

One-tail P value

Prism always reports a two-tail P value when comparing survival curves. If you wish to report a one-tail P value[75], you must have predicted which group would have the longer median survival before collecting any data. Computing the one-tail P value depends on whether your prediction was correct or not.

- If your prediction was correct, the one-tail P value is half the two-tail P value.

- If your prediction was wrong, the one-tail P value equals 1.0 minus half the two-tail P value. This value will be greater than 0.50, and you must conclude that the survival difference is not statistically significant.

5.17.3.5 Interpreting results: The hazard ratio

Key facts about the hazard ratio

- Hazard is defined as the slope of the survival curve — a measure of how rapidly subjects are dying.

- The hazard ratio compares two treatments. If the hazard ratio is 2.0, then the rate of deaths in one treatment group is twice the rate in the other group.

- The hazard ratio is not computed at any one time point, but is computed from all the data in the survival curve.

- Since there is only one hazard ratio reported, it can can only be interpreted if you assume that the population hazard ratio is consistent over time, and that any differences are due to random sampling. This is called the assumption of *proportional hazards*.

- If the hazard ratio is not consistent over time, the value that Prism reports for the hazard ratio will not be useful. If two survival curves cross, the hazard ratios are certainly not consistent (unless they cross at late time points, when there are few subjects still being followed so there is a lot of uncertainty in the true position of the survival curves).

- The hazard ratio is not directly related to the ratio of median survival times. A hazard ratio of 2.0 does not mean that the median survival time is doubled (or halved). A hazard ratio of 2.0 means a patient in one treatment group who has not died (or progressed, or whatever end point is tracked) at a certain time point has twice the probability of having died (or progressed...) by the next time point compared to a patient in the other treatment group.

- Prism computes the hazard ratio, and its confidence interval, using two methods, explained below. For each method it reports both the hazard ratio and its reciprocal. If people in group A die at twice the rate of people in group B (HR=2.0), then people in group B die at half the rate of people in group A (HR=0.5).

- For other cautions about interpreting hazard ratios, see two reviews by Hernan(1) and Spruance(2).

- Duerden (6) wrote a good easy-to-follow explanation of hazard ratios.

The two methods compared

Prism reports the hazard ratio computed by two methods: logrank and Mantel-Haenszel. The two usually give identical (or nearly identical) results. But the results can differ when several subjects die at the same time or when the hazard ratio is far from 1.0.

Bernstein and colleagues analyzed simulated data with both methods (3). In all their simulations, the assumption of proportional hazards was true. The two methods gave very similar values. The logrank method (which they refer to as the O/E method) reports values that are closer to 1.0 than the true Hazard Ratio, especially when the hazard ratio is large or the sample size is large.

When there are ties, both methods are less accurate. The logrank methods tend to report hazard ratios that are even closer to 1.0 (so the reported hazard ratio is too small when the hazard ratio is greater than 1.0, and too large when the hazard ratio is less than 1.0). The Mantel-Haenszel method, in contrast, reports hazard ratios that are further from 1.0 (so the reported hazard ratio is too large when the hazard ratio is greater than 1.0, and too small when the hazard ratio is less than 1.0).

What does it mean when the two hazard ratios are very different?

The simulations of reference 3 did not compare the two methods with data simulated where the assumption of proportional hazards is not true. I have seen one data set where the two estimate of HR were very different (by a factor of three), and the assumption of proportional hazards was dubious for those data. It seems that the Mantel-Haenszel method gives more weight to differences in the hazard at late time points, while the logrank method gives equal weight everywhere (but I have not explored this in detail).

If you see very different HR values with the two methods, think about whether the assumption of proportional hazards is reasonable. If that assumption is not reasonable, then of course the entire concept of a single hazard ratio describing the entire curve is not meaningful.

How the hazard ratio is computed

There are two very similar ways of doing survival calculations: logrank, and Mantel-Haenszel. Both are explained in chapter 3 of Machin, Cheung and Parmar, Survival Analysis (4).

The Mantel Haenszel approach:

1. Compute the total variance, V, as explained on page 38-40 of a handout by Michael Vaeth. Note that he calls the test "logrank" but in a note explains that this is the more accurate test, and also gives the equation for the simpler approximation that we call logrank.

2. Compute L = (O1 - E1) / V, where O1 is the total observed number of events in group1, and E1 is the total expected number of events in group1. You'd get the same value of L if you used the other group.

3. Note that L is the natural logarithm of the hazard ratio. So the hazard ratio equals exp(L).

4. The lower 95% confidence limit of the hazard ratio equals:

   exp(L - 1.96/sqrt(V))

5. The upper 95% confidence limit equals:

   exp(L + 1.96/sqrt(V))

The logrank approach:

1. As part of the Kaplan-Meier calculations, compute the number of observed events (deaths, usually) in each group (Oa, and Ob), and the number of expected events assuming a null hypothesis of no difference in survival (Ea and Eb).

2. The hazard ratio then is:

   HR= (Oa/Ea)/(Ob/Eb)

3. The standard error of the natural logarithm of the hazard ratio is S= sqrt(1/Ea + 1/Eb)

4. Calculate L = ln(HR). (Natural logarithm)

5. The lower 95% confidence limit of the hazard ratio equals:

exp(L - 1.96*S)

5. The upper 95% confidence limit equals:

exp(L + 1.96*S)

Prior versions of Prism

Prism 6 reported the hazard ratio twice, once computed with the Mantel-Haenszel method and again using the logrank method.

A bug in Prism 6. Note that both methods use the natural logarithm of the HR in their calculations. We define this value to be L above. The bug in Prism 6 is that the calculation for the logrank test actually calculated L using the Mantel-Haenszel approach when computing the confidence interval. Usually, the two HR values are nearly identical so this bug was mostly trivial. It only affects the calculations when the two HR values are very different. In this situation, one has to wonder if either definition is very helpful. I suspect this discrepancy happens when the data simply don't comply with the assumes of proportional hazards. The bug was fixed in 7.00 and 7.0a.

Prism 5 computed the hazard ratio and its confidence interval using the Mantel Haenszel approach. Prism 4 used the logrank method to compute the hazard ratio, but used the Mantel-Haenszel approach to calculate the confidence interval of the hazard ratio. The results can be inconsistent. In rare cases, the hazard ratio reported by Prism 4 could be outside the confidence interval of the hazard ratio reported by Prism 4.

References

1. M.A. Hernán. Hazards of Hazard Ratios, Epidemiology. 21:13-5, 2010.

2. S. L. Spruance et all, Hazard ratio in clinical trials, Antimicrobial Agents and Chemotherapy vol. 48 (8) pp. 2787, 2004.

3. L Bernstein, J. Anderson and MC Pike. Estimation of the proportional hazard in two-treatment-group clinical trials. Biometrics (1981) vol. 37 (3) pp. 513-519

4. David Machin, Yin Bun Cheung, Mahesh Parmar, <u>Survival Analysis: A Practical Approach</u>, 2nd edition, IBSN:0470870400.

5. Michael Vaeth, <u>Statistical analysis of survival data in clinical research</u> (2004).

6. Martin Duerden, <u>What are hazard ratios?</u> (2009)

5.17.3.6 Interpreting results: Median survival times

Median survival time

The median survival is the time at which fractional survival equals 50%. Notes:

• If survival exceeds 50% at the longest time point, then median survival cannot be computed. Prism reports that the median survival is "undefined". The logrank comparison of curves really does compare entire curves, and does not compare median survival times. So the P value computed by the logrank test is still valid even if one or both median survival times are undefined.

• If the survival curve is horizontal at 50% survival, then the median survival time is not really defined. In the survival curve below, the curve is horizontal at Y=50% between 9 and 17 months. It would be accurate to say that half the patients had died by 9 months, or that half were still alive at 17 months. Prism follows the suggestion of Machin and reports that the median survival is the average of those two values, 13 months.
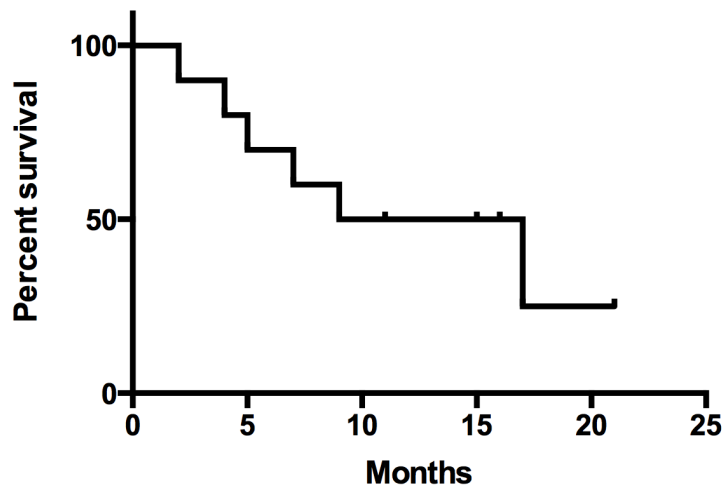
- Prism, like most programs, defines median survival as the time at which the staircase survival curve crosses 50% survival. Thus is is an accurate statement of median survival in the subjects or animals actually included in the data set. The graph on the left below, shows how Prism computes median survival (211 days for this example). If you connected the survival times with point-to-point lines rather than a staircase, you'd find that the line may intersect Y=50% at an earlier time, and thus you'd come up with a different value for median survival (193 days in the example on the right below) This would make sense if you were trying to predict median survival for future patients. Prism does not do this, as it is not standard.
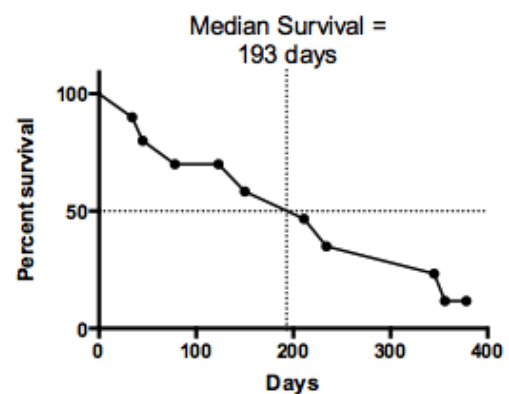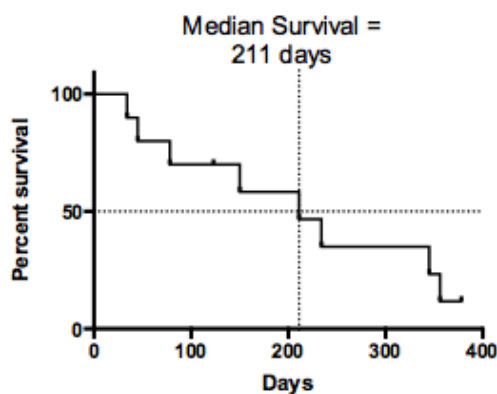
5.17.3.7 Intepreting results: Ratio of median survival times

Ratio of median survival times

If you compare two survival curves, Prism reports the ratio of the median survival times along with its 95% confidence interval of the ratio.

This calculation of the confidence interval of the ratio of survival times is based on an assumption that is not part of the rest of the survival comparison: that both survival curves follow an exponential decay. This means that the chance of dying in a small time interval is the same early in the study and late in the study. If your survival data follow a very different pattern, then the values that Prism reports for the 95% CI of the ratio of median survivals will not be meaningful.

Note that prior versions of Prism computed the confidence interval incorrectly (but computed the ratio just fine).

Why Prism doesn't compute the confidence interval of median survival time

While Prism computes the confidence interval for the ratio of median survivals (when you compare two groups), it does not compute the 95% confidence interval for the median survival time itself. The reason is that multiple methods for computing a confidence interval of median survival have been published and none seem to be standard, and the results don't match. To read more:

- One method is in Collett starting at page 35 .

- Brookmeer and Crowley, A confidence interval for the median survival time. Biometrics (1982) vol. 38 (1) pp. 29-41.

- Barker reviews several methods and points out how different their results can be. The Mean, Median, and Confidence Intervals of the Kaplan-Meier Survival Estimate—Computations and Applications. The American Statistician (2009) vol. 63 (1) pp. 78-80

5.17.3.8 Interpreting results: Comparing >2 survival curves

Logrank and Gehan-Breslow-Wilcoxon tests

The P value tests the null hypothesis that the survival curves are identical in the overall populations. In other words, the null hypothesis is that the treatment did not change survival.

The P value answers this question:

> If the null hypothesis is true, what is the probability of randomly selecting subjects whose survival curves are as different (or more so) than was actually observed?

The difference between the logrank and the Gehan-Breslow-Wilcoxon tests is that the latter places more weight on deaths at early time points.

Note that Prism lets you choose one of two algorithms[580] for computing the P value when comparing three or more groups. The results will show "(conservative)" or "(recommended)", to document your choice.

Logrank test for trend

If you compare three or more survival curves with Prism, it will show results for the overall logrank test, and also show results for the logrank test for trend.

When should you look at the results for the test for trend?

The test for trend is only relevant when the order of groups (defined by data set columns in Prism) is logical. Examples would be if the groups are different age groups, different disease severities, or different doses of a drug. The left-to-right order of data sets in Prism must correspond to equally spaced ordered categories.

If the data sets are not ordered (or not equally spaced), then you should ignore the results of the logrank test for trend.

Results of the logrank test for trend

The logrank test for trend reports a chi-square value, which is always associated with one degree of freedom (no matter how many data sets are being compared). It uses that chi-square value to compute a P value testing the null hypothesis that there is no linear trend between column

order and median survival. If the P value is low, you can conclude that there is a significant trend.

Prism assumes the groups are equally spaced

Computing the logrank test for trend requires assigning each group a code number. The test then looks at the trend between these group codes and survival. With some programs, you could assign these codes, and thus deal with ordered groups that are not equally spaced. Prism uses the column number as the code, so it can only perform the test for trend assuming equally spaced ordered groups. Even if you enter numbers as column titles, Prism does not use these when performing the test for trend.

How it works

The test looks at the linear trend between group code (column number in Prism) and survival. But it doesn't look at median survival, or five-year survival, or any other summary measure. It first computes expected survival assuming the null hypothesis that all the groups are sampled from population with the same survival experience. Then it quantifies the overall discrepancy between the observed survival and the expected survival for each group. Finally it looks at the trend between that discrepancy and group code. For details, see the text by Marchin.

Multiple comparison tests

After comparing three or more treatment groups, you may want to go back and compare two at a time. Prism does not do this automatically, but it is easy to duplicate the analysis, and change the copy to only compare two groups. Then repeat with a different two data sets. If you do this, you need to manually adjust the definition of 'significance' to account for multiple comparisons. [603] Or place all the P values into a new column table, and then analyze that stack of P values [642].

Reference

Survival Analysis: A Practical Approach, Second edition, by David Machin, Yin Bun Cheung, Mahesh Parmar, IBSN:0470870400.

5.17.3.9 The logrank test for trend

The logrank test for trend is used when you compare three or more survival curves when the columns are in a natural order (perhaps ages, or stage of cancer). It tests, essentially, whether there is a linear trend between column order and median survival.

Choosing the logrank test for trend

Prism will compute the logrank test for trend by default when you have three or more groups. You can turn off this test in the Parameters dialog. In this dialog, you also choose between two methods: an older method used in Prism 5 and a better method available since Prism 6 that matches SAS and SPSS.

When should you look at the results for the test for trend?

The test for trend is only relevant when the order of groups (defined by data set columns in Prism) is logical. Examples would be if the groups are different age groups, different disease severities, or different doses of a drug. The left-to-right order of data sets in Prism must correspond to equally spaced ordered categories.

 If the data sets are not ordered (or not equally spaced), then you should ignore the results of the logrank test for trend, or go to the parameters dialog and uncheck this option so the results are not shown.

Results of the logrank test for trend

The logrank test for trend reports a chi-square value, which is always associated with one degree of freedom (no matter how many data sets are being compared). It uses that chi-square value to compute a P value testing the null hypothesis that there is no linear trend between column order and median survival. If the P value is low, you can conclude that there is a significant trend.

Prism assumes the groups are equally spaced

Computing the logrank test for trend requires assigning each group a code number. The test then looks at the trend between these group codes and survival. With some programs, you could assign these codes, and thus deal with ordered groups that are not equally spaced. Prism uses the column number as the code, so it can only perform the test for trend assuming equally spaced ordered groups.

How it works

The test looks at the linear trend between group code (column number in Prism) and survival. But it doesn't look at median survival, or five-year survival, or any other summary measure. It first computes expected survival assuming the null hypothesis that all the groups are sampled from population with the same survival experience. Then it quantifies the overall discrepancy between the observed survival and the expected survival for each group. Finally it looks at the trend between that discrepancy and group code.

References

David Machin, Yin Bun Cheung, Mahesh Parmar, Survival Analysis: A Practical Approach, 2nd edition, IBSN:0470870400.

Douglas Altman, Practical Statistics for Medical Research, IBSN:0412276305

5.17.3.10 Multiple comparisons of survival curves

The need for multiple comparisons

When you compare three or more survival curves at once, you get a single P value testing the null hypothesis that all the samples come from populations with identical survival, and that all differences are due to chance. Often, you'll want to drill down and compare curves two at a time.

If you don't adjust for multiple comparisons, it is easy to fool yourself. If you compare many groups, the chances are high that one or more pair of groups will be 'significantly different' purely due to chance. To protect yourself from making this mistake, you probably should correct for multiple comparisons[117]. Probably? There certainly are arguments for not adjusting for multiple comparisons[128].

How multiple comparisons of survival curves work

Multiple comparison tests after ANOVA are complicated because they not only use a stricter threshold for significance, but also include data from all groups when computing scatter, and use this value with every comparison. By quantifying scatter from all groups, not just the two you are comparing, you gain some degrees of freedom and thus some power.

Multiple comparison tests for comparing survival curves are simpler. You simply have to adjust the definition of significance, and don't need to take into account any information about the groups not in the comparison (as that information would not be helpful).

Comparing survival curves two at a time with Prism

For each pair of groups you wish to compare, follow these steps:

1. Start from the results sheet that compares all groups.

2. Click New, and then Duplicate Current Sheet.

3. The Analyze dialog will pop up. On the right side, select the two groups you wish to compare and make sure all other data sets are unselected. Then click OK.

4. The parameters dialog for survival analysis pops up. Click OK without changing anything.

5. Note the P value (from the logrank or Gehan-Breslow-Wilcoxon test), but don't interpret it until you correct for multiple comparisons, as explained in the next section.

6. Repeat the steps for each comparison if you want each to be in its own results sheet. Or click Change.. data analyzed, and choose a different pair of data sets.

Which comparisons are 'statistically significant'?

When you are comparing multiple pairs of groups at once, you can't interpret the individual P in the usual way. Instead, you set a significance level, and ask which comparisons are 'statistically significant' using that threshold.

The simplest approach is to use the Bonferroni method. Note this is something you'll do manually by doing multiple survival analyzes. It does not involve ANOVA.

1. Define the significance level that you want to apply to the entire family of comparisons. This is conventionally set to 0.05.

2. Count the number of comparisons you are making, and call this value K. See the next section which discusses some ambiguities.

3. Compute the Bonferroni corrected threshold that you will use for each individual comparison. This equals the family-wise significance level (defined in step 1 above, usually .05) divided by K.

4. If a P value is less than this Bonferroni-corrected threshold, then the comparison can be said to be 'statistically significant' using a family-wise significance level of 5%.

How many comparisons are you making?

You must be honest about the number of comparisons you are making. Say there are four treatment groups (including control). You then go back and compare the group with the longest survival with the group with the shortest survival. It is not fair to say that you are only making one comparison, since you couldn't decide which comparison to make without looking at all the data. With four groups, there are six pairwise comparisons you could make. You have implicitly made all these comparisons, so you should define K in step 3 above to equal 6.

If you were only interested in comparing each of three treatments to the control, and weren't interested in comparing the treatments with each other, then you would be making three comparisons, so should set K equal to 3.

5.17.3.11 Analysis checklist: Survival analysis

Survival curves plot the results of experiments where the outcome is time until death. Usually you wish to compare the survival of two or more groups. Read elsewhere to learn about [interpreting survival curves](#)[587], and comparing [two](#)[592] (or [more than two](#)[600]) survival curves.

✓ Are the subjects independent?

Factors that influence survival should either affect all subjects in a group or just one subject. If the survival of several subjects is linked, then you don't have independent observations. For example, if the study pools data from two hospitals, the subjects are not independent, as it is possible that subjects from one hospital have different average survival

times than subjects from another. You could alter the median survival curve by choosing more subjects from one hospital and fewer from the other. To analyze these data, use Cox proportional hazards regression, which Prism cannot perform.

✓ Were the entry criteria consistent?

Typically, subjects are enrolled over a period of months or years. In these studies, it is important that the starting criteria don't change during the enrollment period. Imagine a cancer survival curve starting from the date that the first metastasis was detected. What would happen if improved diagnostic technology detected metastases earlier? Even with no change in therapy or in the natural history of the disease, survival time will apparently increase. Here's why: Patients die at the same age they otherwise would, but are diagnosed when they are younger, and so live longer with the diagnosis. (That is why airlines have improved their "on-time departure" rates. They used to close the doors at the scheduled departure time. Now they close the doors ten minutes before the "scheduled departure time". This means that the doors can close ten minutes later than planned, yet still be "on time". It's not surprising that "on-time departure" rates have improved.)

✓ Was the end point defined consistently?

If the curve is plotting time to death, then there can be ambiguity about which deaths to count. In a cancer trial, for example, what happens to subjects who die in a car accident? Some investigators count these as deaths; others count them as censored subjects. Both approaches can be justified, but the approach should be decided before the study begins. If there is any ambiguity about which deaths to count, the decision should be made by someone who doesn't know which patient is in which treatment group.

If the curve plots time to an event other than death, it is crucial that the event be assessed consistently throughout the study.

✓ Is time of censoring unrelated to survival?

The survival analysis is only valid when the survival times of censored patients are identical (on average) to the survival of subjects who stayed with the study. If a large fraction of subjects are censored, the validity of this assumption is critical to the integrity of the results. There is no

reason to doubt that assumption for patients still alive at the end of the study. When patients drop out of the study, you should ask whether the reason could affect survival. A survival curve would be misleading, for example, if many patients quit the study because they were too sick to come to clinic, or because they stopped taking medication because they felt well.

✓ Does average survival stay constant during the course of the study?

 Many survival studies enroll subjects over a period of several years. The analysis is only meaningful if you can assume that the average survival of the first few patients is not different than the average survival of the last few subjects. If the nature of the disease or the treatment changes during the study, the results will be difficult to interpret.

✓ Is the assumption of proportional hazards reasonable?

 The logrank test is only strictly valid when the survival curves have proportional hazards. This means that the rate of dying in one group is a constant fraction of the rate of dying in the other group. This assumption has proven to be reasonable for many situations. It would not be reasonable, for example, if you are comparing a medical therapy with a risky surgical therapy. At early times, the death rate might be much higher in the surgical group. At later times, the death rate might be greater in the medical group. Since the hazard ratio is not consistent over time (the assumption of proportional hazards is not reasonable), these data should not be analyzed with a logrank test.

✓ Were the treatment groups defined before data collection began?

 It is not valid to divide a single group of patients (all treated the same) into two groups based on whether or not they responded to treatment (tumor got smaller, lab tests got better). By definition, the responders must have lived long enough to see the response. And they may have lived longer anyway, regardless of treatment. When you compare groups, the groups must be defined before data collection begins.

5.17.3.12 Graphing tips: Survival curves

Prism offers lots of choices when graphing survival data. Most of the choices are present in both the Welcome dialog and the Format Graph

dialog, others are only present in the Format Graph dialog. Many are also present in the Parameters dialog for the survival analysis.

How to compute the data

These choices are straightforward matters of taste:

- Plot survival or deaths? The former, used more commonly, starts at 100% and goes down. The latter starts at 0% and goes up.

- Plot fractions or percents? This is simply a matter of preference. If in doubt, choose to plot percentages.

How to graph the data

You can change the appearance of the graph either in the Format Graph dialog or the Parameters dialog of nonlinear regression. If the graph only plots one survival analysis, the two are equivalent. If you plot two or more survival analyses on one graph, then the Format Graph dialog doesn't always work and it is best to make the changes on all the Parameters dialogs.

Graphs without error bars
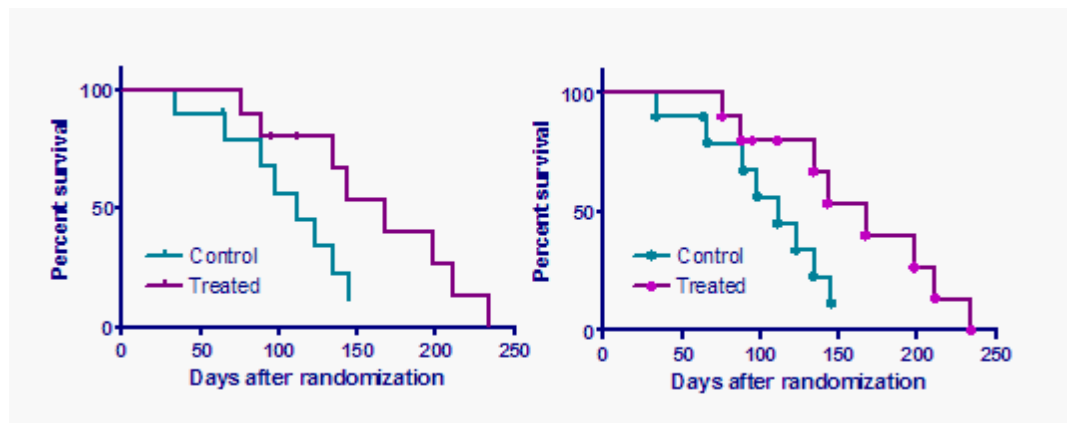
As shown above, survival curves are usually plotted as staircases. Each death is shown as a drop in survival.

In the left panel, the data are plotted as a tick symbol. These symbols at the time of death are lost within the vertical part of the staircase. You see the ticks clearly at the times when a subject's data was censored. The

example has two censored subjects in the treated group between 100 and 150 days.

The graph on the right plots the data as circles, so you see each subject plotted.

Graphs with error bars
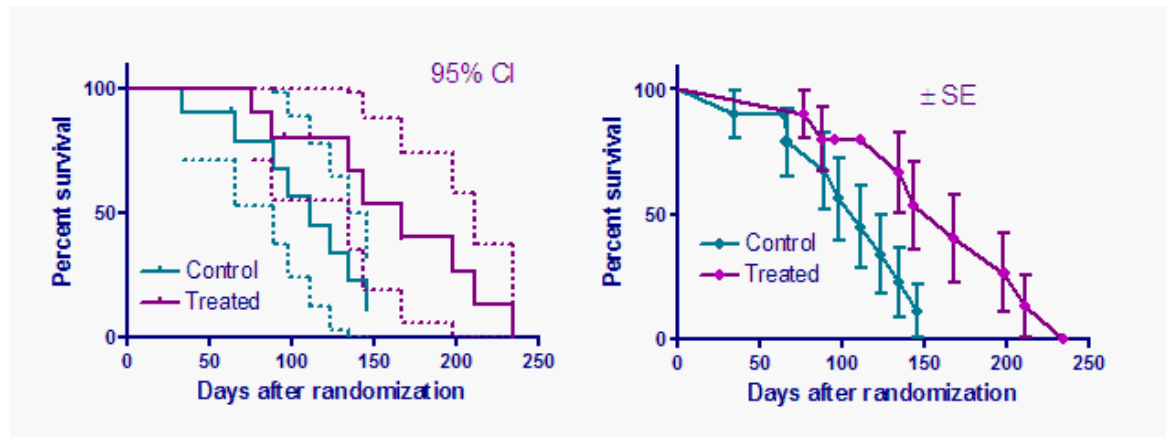
Showing error bars or error envelopes make survival graphs more informative, but also more cluttered. The graph on the left above shows staircase error envelopes that enclose the 95% confidence interval for survival. This shows the actual survival data very well, as a staircase, but it is cluttered. The graph on the right shows error bars that show the standard error of the percent survival. To prevent the error bars from being superimposed on the staircase curve, the points are connected by regular lines rather than by staircases.

How to add a grid line at Y=50% to show median survival

Add a grid line at Y=50 or Y=0.5. To do this, double clcik on the Y axis to bring up the Format Axis dialog, then enter the Y coordinate (50 or 0.5) and check the option to draw a grid line. Click the details button for more choices about line style, thickness and color.

5.17.3.13Q&A: Survival analysis

⊟ **How does Prism compute the confidence intervals of the survival percentages?**

Prism offers two choices.

- The symmetrical method was the only method offered in Prism 4 and earlier, and is offered now for compatibility. It uses the method of Greenwood. We don't recommend it.

- The asymmetrical method is more accurate and recommended. It is explained on page 42 and page 43 of Machin. That book does not give a name or reference for the method, The idea is that it first does a transform (square root and log) that makes the uncertainty of survival close to Gaussian. It then computes the SE and a symmetrical 95% CI on that transformed scale. Then it back transforms the confidence limits back to the original scale.

⊟ **Can Prism compute the mean (rather than median) survival time?**

Survival analysis computes the median survival with its confidence interval. The reason for this is that the median survival time is completely defined once the survival curve descends to 50%, even if many other subjects are still alive. And the median survival is defined, even if data from some subjects was censored.

In contrast, the mean survival is simply not defined until every subject dies, and only when you know the survival time for each subject (none were censored). These conditions occur in very very few studies, so Prism doesn't compute mean survival.

But there is an easy workaround: If you know he survival times for each subject, enter them into a column table, and ask Prism to do column statistics to calculate the mean with its confidence interval.

⊟ **Can Prism create a survival curve when you already know the percent survival at each time?**

Prism can create Kaplan-Meier survival curves, and compare these with the logrank test (or the Wlcoxon-Gehan-Breslow test). To do this, you

must enter data on a Prism table formatted as a survival table and you must enter one row of data per subject.

But what if you already know the percent survival at each time point, and just want to make a graph? In this case, do not enter data onto a survival data table. That table requires information about each subject. Instead, create an XY data table. If you only want to enter percent survival, format the data table to enter single Y values with no subcolumns. If you know the standard error of the survival at each time point (from calculations done elsewhere), then format the data table for entry of mean with SEM (in fact, the "mean" will be percent survival, and "SEM" will be SE of the survival percentage).

Enter time (as months or days or weeks) into X. You must enter this as a number, not a date.

Enter percent (or fraction) survival into Y. Just enter the values (don't append percent symbols).

Then polish your graph. If you want the graph to have a staircase look (which is traditional for survival curves), you can do that. This screen shot shows where to make this setting in the Format Graph dialog:
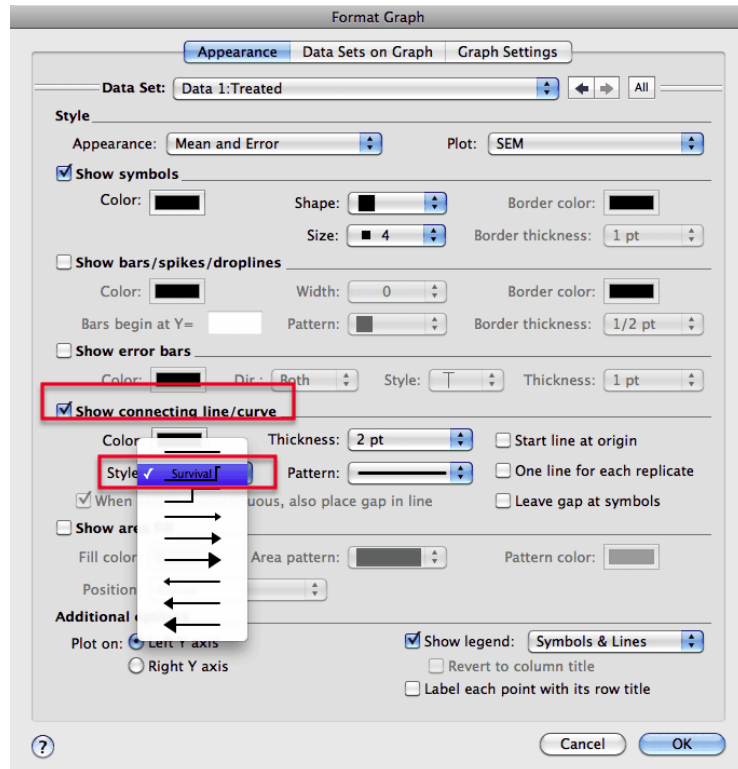
If you enter survival percentages on an XY table, it will not be possible to do any calculations. You won't be able to compute error bars or confidence bands, and wont' be able to compare survival curves under different treatments.

What determines how low a Kaplan-Meier survival curve ends up at late time points?

If there are no censored observations

If you follow each subject until the event occurs (the event is usually death, but survival curves can track time until any one-time event), then the curve will eventually reach 0. At the time (X value) when the last subject dies, the percent survival is zero.

If all subjects are followed for exactly the same amount of time

If all subjects are followed for the same amount of time, the situation is easy. If one third of the subjects are still alive at the end of the study, then the percent survival on the survival curve will be 33.3%.

If some subjects are censored along the way

If the data for any subjects are censored, the bottom point on the survival curve will not equal the fraction of subjects that survived.

Prior to censoring, a subject contributes to the fractional survival value. Afterward, she or he doesn't affect the calculations. At any given time, the fractional (or percent) survival value is the proportion of subjects followed that long who have survived.

Subjects whose data are censored --either because they left the study, or because the study ended--can't contribute any information beyond the time of censoring. So if any subjects are censored before the last time shown on the survival curve's X-axis, the final survival percentage shown on the survial graph will not correspond to the actual fraction of the subjects who survived. That simple survival percentage that you can easily compute by hand is not meaningful, because not all the subjects were not followed for the same amount of time.

When will the survival curve drop to zero?

If the survival curve goes all the way down to 0% survival, that does not mean that every subject in the study died. Some may have censored data at earlier time points (either because they left the study, or because the study ended while they were alive). The curve will drop to zero when a death happens after the last censoring. Make sure your data table is sorted by X value (which Prism can do using Edit..Sort). Look at the subject in the last row. If the Y value is 1 (death), the curve will descend to 0% survival. If the Y value is 0 (censored), the curve will end above 0%.

Why does Prism tell me that median survival is undefined?

Median survival is the time it takes to reach 50% survival. If more than 50% of the subjects are alive at the end of the study, then the median survival time is simply not defined.

The P value comes from the logrank test, which compares the entire curve, and works fine even if the percent survival is always greater than 50%. Two curves can be very different, even if they never dip down below 50%.

⊟ **Can Prism compute confidence bands as well as confidence intervals of survival curves?**

When Prism computes survival curves, it can also compute the 95% confidence interval at each time point (using two alternative methods). The methods are approximations, but can be interpreted like any confidence interval. You know the observed survival percentage at a certain time in your study, and can be 95% confident (given a set of assumptions) that the confidence interval contains the true population value (which you could only know for sure if you had an infinite amount of data).

When these confidence intervals are plotted as error bars (left graph below) there is no problem. Prism can also connect the ends of the error bars, and create a shaded region (right graph below). This survival curve plots the survival of a sample of only seven people, so the confidence intervals are very wide. Prism file.

The shaded region looks like the confidence bands computed by linear and nonlinear regression, so it is tempting to interpret these regions as confidence bands. But it is not correct to say that you can be 95% certain that these bands contain the entire survival curve. It is only correct to say that at any time point, there is a 95% chance that the interval contains the true percentage survival. The true survival curve (which you can't know) may be within the confidence intervals at some time points and outside the confidence intervals at other time points.
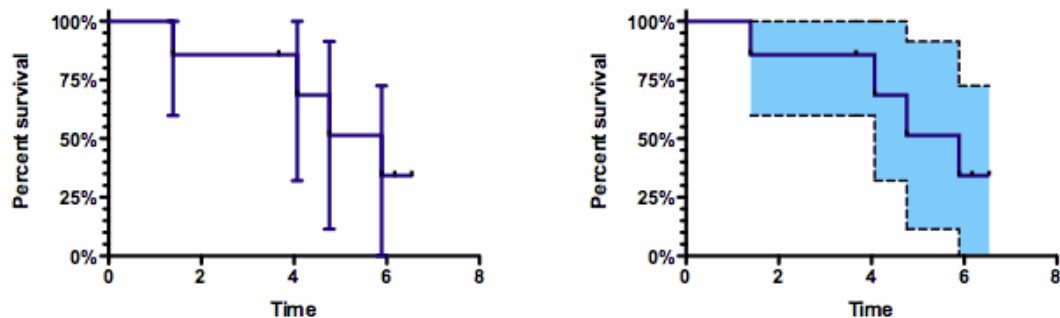
It is possible (but not with Prism) to compute true confidence bands for survival curves, and these are wider than the confidence intervals shown above. Confidence bands that are 95% certain to contain the

entire survival curve at all time points are wider than the confidence intervals for individual time points.

How does Prism deal with deaths at time zero?

When analyzing survival data, Prism simply ignores any rows with X=0. Our thinking is simple. If alternative treatments begin at time zero, then a death right at the moment treatment begins provides no information to help you decide which of two treatments is better. There is no requirement that X be an integer. If a death occurs half a day into treatment, and X values are tabulated in days, enter 0.5 for that subject.

Some fields (pediatric leukemia is one) do consider events at time zero to be valid. These studies to not simply track death, but track time until recurrence of the disease. But disease cannot recur until it first goes into remission. In the case of some pediatric leukemia trials, the treatment begins 30 days before time zero. Most of the patients are in remission at time zero. Then the patients are followed until death or recurrence of the disease. But what about the subjects who never go into remission? Some investigators consider these to be events at time zero. Some programs, we are told, take into account the events at time zero, so the Kaplan-Meier survival curve starts with survival (at time zero) of less than 100%. If 10% of the patients in one treatment group never went into remission, the survival curve would begin at Y=90% rather than 100%.

We have not changed Prism to account for deaths at time zero for these reasons:

- We have seen no scientific papers, and no text books, that explains what it means to analyze deaths at time zero. It seems far from standard.

- It seems wrong to combine the answers to two very different questions in one survival curve: What fraction of patients go into remission? How long do those in remission stay in remission?

- If we included data with X=0, we are not sure that the results of the survival analysis (median survival times, hazard ratios, P values, etc.) would be meaningful.

The fundamental problem is this: Survival analysis analyses data expressed as the time it takes until an event occurs. Often this event is death. Often it is some other well defined event that can only happen once. But usually the event is defined to be something that could possibly happen to every participant in the trial. With these pediatric leukemia trials, the event is defined to be recurrence of the disease. But, of course, the disease cannot recur unless it first went into remission. So the survival analysis is really being used to track time until the second of two distinct events. That leads to the problem of how to analyze the data from patients who never go into remission (the first event never happens).

We are willing to reconsider our decision to ignore, rather than analyze, survival data entered with X=0. If you think we made the wrong decision, please let us know. Provide references if possible.

There is a simple work around if you really want to analyze your data so deaths at time zero bring down the starting point below 100%, enter some tiny value other than zero. Enter these X values, say, as 0.000001. An alternative is to enter the data with X=0, and then use Prism's transform analysis with this user-defined transform:

```
X=IF(X=0, 0.000001, X)
```

In the results of this analysis, all the X=0 values will now be X=0.000001. From that results table, click Analyze and choose Survival analysis.

How is the percentage survival computed?

Prism uses the Kaplan-Meier method to compute percentage survival. This is a standard method. The only trick is in accounting for censored observations.

Consider a simple example. You start with 16 individuals. Two were censored before the first death at 15 months. So the survival curve drops at 15 months from 100% down to 13/14=92.86%. Note that the denominator is 14, not 16. Just before the death, only 14 people were being followed, not 16 (since data for two were censored before that).

Seven more individuals were censored before the next death at 93 months. So of those who survived more than 15 months, 5/6= 83.3%

were alive after 93 months. But this is a relative drop. To know the percent of people alive at 0 months who are still alive after 93 months, multiply 92.86% (previous paragraph) times 83.33% and you get 77.38%, which is the percent survival Prism reports at 93 months. Now you can see why these Kaplan-Meier calculations are sometimes called the product-limit method.

Reference

David Machin, Yin Bun Cheung, Mahesh Parmar, [Survival Analysis: A Practical Approach](#), 2nd edition, IBSN:0470870400.

5.17.3.14 Determining the median followup time

Survival analysis often deals with experimental designs where different subjects are followed for different durations. How can one quantify the median followup time? Survival analysis (in Prism and other programs) tells you the median survival time. But what about the median time of followup?

> Note the distinction between the median survival time and the median time that research subjects were followed (the topic of this page).

Prism presents you with a table of number of subjects at risk over time. One thought is to look at this table and see how long it takes for the number to drop to half the starting value. But there are two reasons why the number-at-risk drops over time: a subject can die or his data can be censored. Looking merely at the number-at-risk table treats those two situations identically. If someone dies, you don't know how long they would have been followed. From the point of view of tracking followup time, the roles of deaths and censoring are sort of reversed.

Schemper and Smith (1) followed that idea to its conclusion and devised a clever method to obtain the median followup time. Run the data through the Kaplan-Meier analysis again, but with the meaning of the status indicator reversed. The end point is loss-of-followup (which is usually considered censoring). If the patient died, you can't know how long they would have been followed. So death censors the true but unknown observation time of an individual. So create a Kaplan Meier curve where loss of followup is the event being followed, and a death is treated as censoring the data.

In Prism:

1. From the survival analysis results, click New, then Duplicate sheet.

2. OK the dialog that lists the data being analyzed.

3. On the parameters dialog, swap the two indicator variables. The default is for 1 to denote death and zero to denote censoring. Reverse this convention in the dialog (but leave the data alone).Tell Prism that 0 denotes "death" and 1 denotes "censoring".

4. OK from the dialog and look at the results page. Ignore the log rank test and its P value. These values cannot be interpreted. Instead, look at the "median survival". Since you swapped the meaning of survival and censored, this value is really the median followup time.

5. The Kaplan-Meier graph created from this analysis tracks the number of patients being followed over time. It is distinct from the Kaplan-Meier graph that tracks percent survival over time.

For the sample data comparing two groups, the results (with some polishing) look like this:
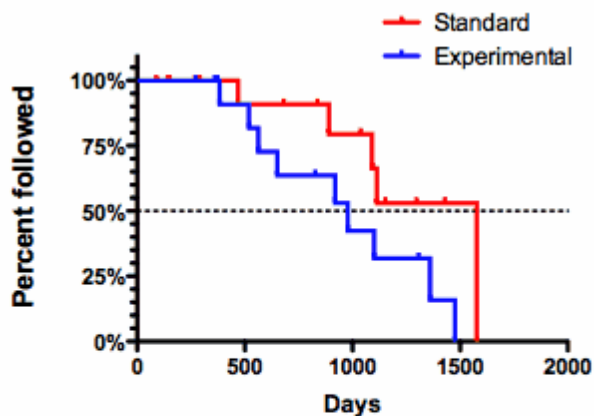
Median followup times:

Standard   1577
Experimental   978.0

1. M Schemper and TL Smith. A note on quantifying follow-up in studies of failure time. Controlled clinical trials (1996) vol. 17 (4) pp. 343-346

5.18 Correlation

When two variables vary together, statisticians
say that there is a lot of covariation or correlation.

STATISTICS WITH PRISM 8

The correlation coefficient, r, quantifies the

direction and magnitude of correlation.

5.18.1 Key concepts: Correlation

- When two variables vary together, statisticians say that there is a lot of covariation or correlation.

- The correlation coefficient, r, quantifies the direction and magnitude of correlation.

- Correlation is used when you measured both X and Y variables, and is not appropriate if X is a variable you manipulate.

- X and Y are almost always real numbers (not integers, not categories, not counts).

- The correlation analysis reports the value of the correlation coefficient. It does not create a regression line. If you want a best-fit line, choose linear regression.

- Note that correlation and linear regression are not the same. Review the differences. [625] In particular, note that the correlation analysis does not fit or plot a line.

- Correlation computes a correlation coefficient and its confidence interval. Its value ranges from -1 (perfect inverse relationship; ax X goes up, Y goes down) to 1 (perfect positive relationship; as X goes up so does Y). A value of zero means no correlation at all.

- Correlation also reports a P value testing the null hypothesis that the data were sampled from a population where there is no correlation between the two variables.

- The difference between Pearson and Spearman correlation, is that the confidence interval and P value from Pearson's can only be interpreted if you assume that both X and Y are sampled from populations with a Gaussian distribution. Spearman correction does not make this assumption.

© 1995-2020 GraphPad Software, LLC

- If either X or Y has only two possible values, the results of Pearson correlation are identical to point-biserial correlation.

5.18.2   How to: Correlation

Prism can perform correlation analyses from XY, Column, and Multiple Variables data tables. Click the Analyze button and choose correlation.

Compute correlation between which pairs of columns?

Compute the correlation between two specific columns, between all columns (correlation matrix), or between each column and a control data set (which is X, if you are analyzing an XY table).

Assume data are sampled from a Gaussian distribuiton?

Prism offers two ways to compute correlation coefficients:

- Pearson correlation calculations are based on the assumption that both X and Y values are sampled from populations that follow a Gaussian distribution, at least approximately. With large samples, this assumption is not too important.

- Spearman nonparametric correlation makes no assumption about the distribution of the values, as the calculations are based on ranks, not the actual values.

One- or two-tailed P values?

Prism can compute either a one-tailed or two-tailed P value. We suggest almost always choosing a two-tailed P value. You should only choose a one-tail P value when you have specified the anticipated sign of the correlation coefficient before collecting any data and are willing to attribute any correlation in the "wrong" direction to chance, no matter how striking that correlation is.

5.18.3 Interpreting results: Correlation

Correlation coefficient

The correlation coefficient, r, ranges from -1 to +1. The nonparametric Spearman correlation coefficient, abbreviated rs, has the same range. This latter value is sometimes denoted by the Greek letter ρ (rho).

Value of r (or rs)	Interpretation
1.0	Perfect correlation
0 to 1	The two variables tend to increase or decrease together.
0.0	The two variables do not vary together at all.
-1 to 0	One variable increases as the other decreases.
-1.0	Perfect negative or inverse correlation.

If r or rs is far from zero, there are four possible explanations:

- Changes in the X variable causes a change the value of the Y variable.

- Changes in the Y variable causes a change the value of the X variable.

- Changes in another variable influence both X and Y.

- X and Y don't really correlate at all, and you just happened to observe such a strong correlation by chance. The P value quantifies the likelihood that this could occur.

Notes on correlation coefficients:

- If you choose Spearman nonparametric correlation, Prism computes the confidence interval of the Spearman correlation coefficient by an approximation. According to Zar (Biostatistical Analysis) this approximation should only be used when n>10. So with smaller n,

Prism simply does not report the confidence interval of the Spearman correlation coefficient.

- If you ask Prism to compute a correlation matrix (compute the correlation coefficient for each pair of variables), it computes a simple correlation coefficient for each pair, without regard for the other variables. It does not compute multiple regression, or partial regression, coefficients.

- If all Y values are the same, it is not possible to compute a correlation coefficient (parametric or nonparametric), and Prism reports "horizontal line". Correlation asks how much X and Y vary together. If Y doesn't vary at all, that question is not meaningful and the correlation calculations can't be done (division by zero).

- If all the X values are the same, it is not possible to compute a correlation coefficient, and Prism reports "vertical line".

r^2

Perhaps the best way to interpret the value of r is to square it to calculate r^2. Statisticians call this quantity the coefficient of determination, but scientists call it "r squared". It is a value that ranges from zero to one, and is the fraction of the variance in the two variables that is "shared". For example, if r^2=0.59, then 59% of the variance in X can be explained by variation in Y. Likewise, 59% of the variance in Y can be explained by variation in X. More simply, 59% of the variance is shared between X and Y.

Prism only calculates an r^2 value from the Pearson correlation coefficient. It is not appropriate to compute r2 from the nonparametric Spearman correlation coefficient.

P value

The P value answers this question:

If there really is no correlation between X and Y overall, what is the chance that random sampling would result in a correlation coefficient as far from zero (or further) as observed in this experiment?

If the P value is small, you can reject the idea that the correlation is due to random sampling.

If the P value is large, the data do not give you any reason to conclude that the correlation is real. This is not the same as saying that there is no correlation at all. You just have no compelling evidence that the correlation is real and not due to chance. Look at the confidence interval for r. It will extend from a negative correlation to a positive correlation. If the entire interval consists of values near zero that you would consider biologically trivial, then you have strong evidence that either there is no correlation in the population or that there is a weak (biologically trivial) association. On the other hand, if the confidence interval contains correlation coefficients that you would consider biologically important, then you couldn't make any strong conclusion from this experiment. To make a strong conclusion, you'll need data from a larger experiment.

If you entered data onto a column table and requested a correlation matrix, Prism will report a P value for the correlation of each column with every other column. These P values do not include any correction for multiple comparisons.

Prism always reports two-tailed (two-sided) P values.

How Prism computes the P value for Spearman nonparametric correlation

With 17 or fewer XY pairs, Prism computes an exact P value for nonparametric (Spearman) correlation, looking at all possible permutations of the data. The exact calculations handle ties with no problem. With 18 or more pairs, Prism computes an approximate P value for nonparametric correlation). This approximation is standard. It first computes a t ratio from Rs, and then computes P from that.

Prism 5 used a cutoff of >13 pairs to do an approximate calculation in the absence of ties and always used the approximation in the presence of ties, while now Prism uses a cutoff of >17 pairs. Therefore Prism 5 will report different (less accurate) results for data sets with between 14 and 17 pairs or data sets with fewer than 17 pairs but with ties.

Prism 7 fixed a bug in Prism 6 (up to 6.05 and 6.0f, but not in earlier versions) that resulted in incorrect P values sometimes when Rs was negative, there were tied values, and the P value was computed exactly.

5.18.4 Analysis checklist: Correlation

✓ **Are the data points independent?**

Correlation assumes that any random factor affects only one data point, and not others. You would violate this assumption if you choose half the subjects from one group and half from another. A difference between groups would affect half the subjects and not the other half.

✓ **Are X and Y measured independently?**

The calculations are not valid if X and Y are intertwined. You'd violate this assumption if you correlate midterm exam scores with overall course score, as the midterm score is one of the components of the overall score.

✓ **Were X values measured (not controlled)?**

If you controlled X values (e.g., concentration, dose, or time) you should calculate linear regression rather than correlation.

✓ **Is the covariation linear?**

A correlation analysis would not be helpful if Y increases as X increases up to a point, and then Y decreases as X increases further. You might obtain a low value of r, even though the two variables are strongly related. The correlation coefficient quantifies linear covariation only.

✓ **Are X and Y distributed according to Gaussian distributions?**

To accept the P value from standard (Pearson) correlation, the X and Y values must each be sampled from populations that follow Gaussian distributions. Spearman nonparametric correlation does not make this assumption.

5.18.5 Correlation matrix

If you start with a data table with three or more Y columns, you can ask Prism to compute the correlation of each column with each other column, and thus generate a correlation matrix.

The results appear on three pages:

- The correlation coefficient r (or rs). This is computed for each pair of variables and doesn't not account for other variables. Prism does not compute a partial correlation coefficient.

- The P value (two-tail) testing the null hypothesis that the true population correlation coefficient for that pair of variables is zero.

- The number of XY pairs. This might not be the same for all pairs of variables if some data are missing.

Check an option (added in 8.1) to create a heat map of R^2 values. To make a heat map from P values or sample sizes: From the page, click New and choose Graph of existing data. Choose a Grouped graph, then choose the Heat Map tab.

5.18.6 The difference between correlation and regression

Correlation and linear regression are not the same.

What is the goal?

Correlation quantifies the degree to which two variables are related. Correlation does not fit a line through the data points. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does. When r is 0.0, there is no relationship. When r is positive, there is a trend that one variable goes up as the other one goes up. When r is negative, there is a trend that one variable goes up as the other one goes down.

Linear regression finds the best line that predicts Y from X.

What kind of data?

Correlation is almost always used when you measure both variables. It rarely is appropriate when one variable is something you experimentally manipulate.

Linear regression is usually used when X is a variably you manipulate (time, concentration, etc.)

Does it matter which variable is X and which is Y?

With correlation, you don't have to think about cause and effect. It doesn't matter which of the two variables you call "X" and which you call "Y". You'll get the same correlation coefficient if you swap the two.

The decision of which variable you call "X" and which you call "Y" matters in regression, as you'll get a different best-fit line if you swap the two. The line that best predicts Y from X is not the same as the line that predicts X from Y (however both those lines have the same value for R^2).

Assumptions

The correlation coefficient itself is simply a way to describe how two variables vary together, so it can be computed and interpreted for any two variables. Further inferences, however, require an additional assumption -- that both X and Y are measured (are interval or ratio variables), and both are sampled from Gaussian distributions. This is called a bivariate Gaussian distribution. If those assumptions are true, then you can interpret the confidence interval of r and the P value testing the null hypothesis that there really is no correlation between the two variables (and any correlation you observed is a consequence of random sampling).

With linear regression, the X values can be measured or can be a variable controlled by the experimenter. The X values are not assumed to be sampled from a Gaussian distribution. The distances of the points from the best-fit line is assumed to follow a Gaussian distribution, with the SD of the scatter not related to the X or Y values.

Relationship between results

Correlation computes the value of the Pearson correlation coefficient, r. Its value ranges from -1 to +1.

Linear regression quantifies goodness of fit with r^2, sometimes shown in uppercase as R^2. If you put the same data into correlation (which is rarely appropriate; see above), the square of r from correlation will equal r^2 from regression.

5.19 Diagnostic lab analyses

How do you decide where to draw the threshold between 'normal' and 'abnormal' test results? How do you compare two methods that assess the same outcome? Diagnostic labs have unique statistical needs, which we briefly discuss here.

5.19.1 ROC Curves

5.19.1.1 Key concepts: Receiver-operating characteristic (ROC) curves

- When evaluating a diagnostic test, it is often difficult to determine the threshold laboratory value that separates a clinical diagnosis of "normal" from one of "abnormal."

- If you set a high threshold value (with the assumption that the test value increases with disease severity), you may miss some individuals with low test values or mild forms of the disease. The *sensitivity*, the fraction of people who have the disease that will be correctly identified with a positive test, will be low. Few of the positive tests will be false positives, but many of the negative tests will be false negatives.

- If you set a low threshold, you will catch most individuals with the disease, but you may mistakenly diagnose many normal individuals as "abnormal." The *specificity*, the fraction of people who don't have the disease who are correctly identified with a negative test, will be low. Few of the negative tests will be false negatives, but many of the positive tests will be false positives.

- You can have higher sensitivity or higher specificity, but not both (unless you develop a better diagnostic test).

- A receiver operating characteristic (ROC) curve helps you visualize and understand the tradeoff between high sensitivity and high specificity when discriminating between clinically normal and clinically abnormal laboratory values.

- Which is the best combination of sensitivity and specificity? It depends on the circumstances. In some cases, you'll prefer more sensitivity at the expense of specificity. In other cases, just the opposite. Prism cannot help with those value judgments.

- Why the odd name? Receiver operating characteristic curves were developed during World War II, within the context of determining if a blip on a radar screen represented a ship or an extraneous noise. The radar-receiver operators used this method to set the threshold for military action.

- ROC curves can also be used as part of the presentation of the results of logistic regression. Prism does not do logistic regression so does not prepare this kind of ROC curve.

- The review by Berrar (1) is excellent both for understanding ROC curves and for appreciating some of their pitfalls.

- ROC curves can also be used as a way to display results from multiple logistic regression. This is not something Prism can do.

1. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform. Oxford University Press; 2011 Mar 21;13(1):bbr008–97.

5.19.1.2 How to: ROC curve

1. Enter ROC data
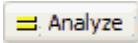
From the Welcome or New table dialog, choose the Column tab. If you are not ready to enter your own data, choose the sample ROC data.
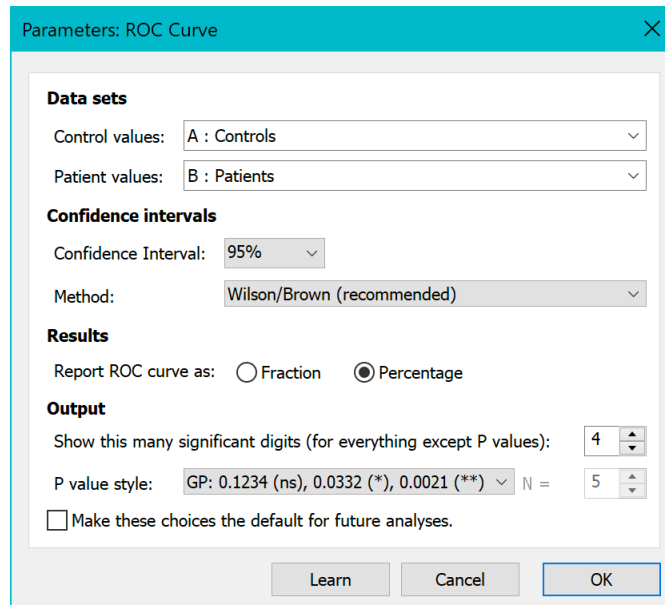
Enter diagnostic test results for controls into column A and patients in column B. Since the two groups are not paired in any way, the order in which you enter the data in the rows is arbitrary. The two groups may have different numbers of subjects.

Note that some other programs expect you to enter all the lab data into one column, and then differentiate patients from controls via a grouping

variable entered into another column. Prism cannot analyze data entered this way.

2. Create the ROC curve

From the data table, click ⇉ Analyze on the toolbar, and then choose Receiver-operator characteristic curve from the list of one-way analyses.

In the ROC dialog, designate which columns have the control and patient results, and choose to see the results (sensitivity and 1-specificity) expressed as fractions or percentages. Don't forget to check the option to create a new graph.

Note that Prism doesn't ask whether an increased or decrease test value is abnormal. Instead, you tell Prism which column of data is for controls and which is for patients, and it figures out automatically whether the patients tend to have higher or lower test results.

3. View the graph
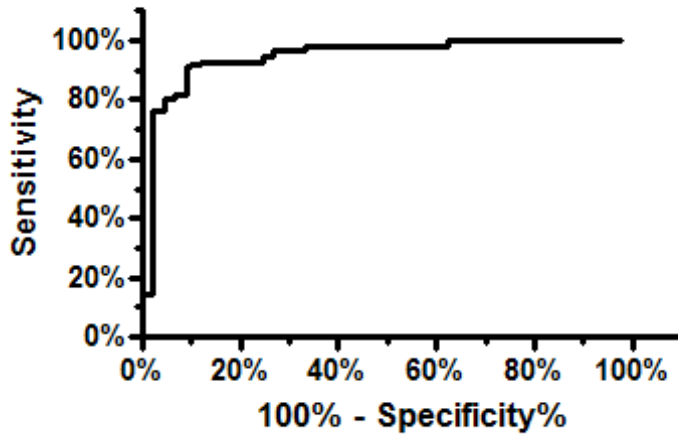
How to put several ROC curves on one graph

Each ROC analysis creates one ROC curve and graph. The XY points that define the graph are on a results page called "ROC curve". You can plot multiple ROC curves on one graph if you want to. The easiest way to do so is to go to a graph of one ROC curve, and drag the "ROC curve" results table from another one onto the graph. You can also change which data sets are plotted using the middle tab of the Format Graph dialog. The trick is realizing that the ROC curve is simply a data set created by an analysis, and it can be added to any graph.

5.19.1.3 Interpreting results: ROC curves

Sensitivity and specificity

The whole point of an ROC curve is to help you decide where to draw the line between 'normal' and 'not normal'. This will be an easy decision if all the control values are higher (or lower) than all the patient values. Usually, however, the two distributions overlap, making it not so easy. If you make the threshold high, you won't mistakenly diagnose the disease in many who don't have it, but you will miss some of the people who have the disease. If you make the threshold low, you'll correctly identify all (or almost all) of the people with the disease, but will also diagnose the disease in more people who don't have it.

To help you make this decision, Prism tabulates and plots the sensitivity and specificity of the test at various cut-off values.

Sensitivity: The fraction of people with the disease that the test correctly identifies as positive.

Specificity: The fraction of people without the disease that the test correctly identifies as negative.

Prism calculates the sensitivity and specificity using each value in the data table as the cutoff value. This means that it calculates many pairs of sensitivity and specificity. If you select a high threshold, you increase the specificity of the test, but lose sensitivity. If you make the threshold low, you increase the test's sensitivity but lose specificity.

Prism displays these results in two forms. The table labeled "ROC" curve is used to create the graph of 100%-Specificity% vs. Sensitivity%. The table labeled "Sensitivity and Specifity" tabulates those values along with their 95% confidence interval for each possible cutoff between normal and abnormal.

Area (C statistic)

The area under a ROC curve is called the *C statistic,* the *concordance* statistic or *the C-index.* It quantifies the overall ability of the test to discriminate between those individuals with the disease and those without the disease. A truly useless test (one no better at identifying true positives than flipping a coin) has an area of 0.5. A perfect test (one that has zero false positives and zero false negatives) has an area of 1.00. Your test will have an area between those two values. Even if you choose to plot the results as percentages, Prism reports the area as a fraction.

Prism computes the area under the entire AUC curve, starting at 0,0 and ending at 100, 100. Note that whether or not you ask Prism to plot the ROC curve out to these extremes, it computes the area for that entire curve.

While it is clear that the area under the curve is related to the overall ability of a test to correctly identify normal versus abnormal, it is not so obvious how one interprets the area itself. There is, however, a very intuitive interpretation.

If patients have higher test values than controls, then:

The area represents the probability that a randomly selected patient will have a higher test result than a randomly selected control.

If patients tend to have lower test results than controls:

The area represents the probability that a randomly selected patient will have a lower test result than a randomly selected control.

For example: If the area equals 0.80, on average, a patient will have a more abnormal test result than 80% of the controls. If the test were perfect, every patient would have a more abnormal test result than every control and the area would equal 1.00.

If the test were worthless, no better at identifying normal versus abnormal than chance, then one would expect that half of the controls would have a higher test value than a patient known to have the disease and half would have a lower test value. Therefore, the area under the curve would be 0.5.

The area under a ROC curve can never be less than 0.50. If the area is first calculated as less than 0.50, Prism will reverse the definition of abnormal from a higher test value to a lower test value. This adjustment will result in an area under the curve that is greater than 0.50.

Berrar points out that ROC curves must be interpreted with care, and there is more to interpretation than looking at the AUC (1).

SE and Confidence Interval of Area

Prism also reports the standard error of the area under the ROC curve, as well as the 95% confidence interval. These results are computed by a nonparametric method that does not make any assumptions about the distributions of test results in the patient and control groups.

Interpreting the confidence interval is straightforward. If the patient and control groups represent a random sampling of a larger population, you can be 95% sure that the confidence interval contains the true area.

P Value

Prism completes your ROC curve evaluation by reporting a P value that tests the null hypothesis that the area under the curve really equals 0.50. In other words, the P value answers this question:

If the test diagnosed disease no better flipping a coin, what is the chance that the area under the ROC curve would be as high (or higher) than what you observed?

If your P value is small, as it usually will be, you may conclude that your test actually does discriminate between abnormal patients and normal controls.

If the P value is large, it means your diagnostic test is no better than flipping a coin to diagnose patients. Presumably, you wouldn't collect enough data to create an ROC curve until you are sure your test actually can diagnose the disease, so high P values should occur very rarely.

1. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform. Oxford University Press; 2011 Mar 21;13(1):bbr008–97.

5.19.1.4 Analysis checklist: ROC curves

✔ **Were the diagnoses made independent of the results being analyzed?**

The ROC curve shows you the sensitivity and specificity of the lab results you entered. It does this by comparing the results in a group of patients with a group of controls. The diagnosis of patient or control must be made independently, not as a result of the lab test you are assessing.

✔ **Are the values entered into the two columns actual results of lab results?**

Prism computes the ROC curve from raw data. Don't enter sensitivity and specificity directly and then run the ROC analysis.

✔ **Are the diagnoses of patients and controls accurate?**

If some people are in the wrong group, the ROC curve won't be accurate. The method used to discriminate between patient and control must truly be a gold standard.

5.19.1.5 Calculation details for ROC curves

Sensitivity and specificity at various thresholds

The list of thresholds is taken by sorting all the values in both groups (patients and controls) and averaging adjacent values in that sorted list. So each threshold value is midway between two values in the data.

Each sensitivity is the fraction of values in the patient group that are above the threshold. The specificity is the fraction of values in the control group that are below the threshold. Each confidence intervals is computed from the observed proportion by the Clopper method (1), without any correction for multiple comparisons.

Area under the ROC curve

Prism uses the same method it uses for the Area Under Curve[233] analysis.

SE of the area

Prism uses the method of Hanley (1), which uses the equation below where A is the area, na and nn are the number of abnormals (patients) and normals (controls).

$$SE = \sqrt{\frac{A(1-A) + (na-1)(Q1-A^2) + (nn-1)(Q2-A^2)}{na \cdot nn}}$$

Q1 is the probability that two randomly chosen patients will both get a more positive test result than a randomly chosen control which is approximately A/(2-A).

Q2 is the probability that one randomly chosen patient will get a more positive test result than two randomly chosen controls which is approximately 2A*A/(1+A).

Prism actually computes Q1 and Q1 using a more complicated equation.

P value

When computing the P value, Prism computes the SE differently, assuming that the area is really 0.5 (the null hypothesis). This simplifies the equation to

$$SE = \sqrt{\frac{0.25 + (na + nn - 2)}{na \cdot nn \cdot 12}}$$

It then computes a z ratio using the equation below, and determines the P value from the normal distribution (two-tail).

$$z = \frac{A - 0.5}{SE}$$

Reference

1. C. J. Clopper and E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika 1934 26: 404-413.

2. Hanley JA, McNeil BJ. The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve, Radiology 1982 143 29-36

5.19.1.6 Computing predictive values from a ROC curve

The Positive and Negative Predictive Values

If you enter test values from patients and controls, Prism can create a ROC curve. This plots the tradeoff of sensitivity vs specificity for various possible cutoff values to define the borderline between "normal" and "abnormal" test results.

The *sensitivity* is the proportion of patients who will have an abnormal test result.

The *specificity* is the proportion of controls who will have a negative test result.

But those two values may not answer the questions you really want the answer to:

• If the result is "abnormal", what is the chance that the person really has the disease. This is the Positive Predictive Value (PPV).

• If the results is "normal", what is the chance that the person really does not have the disease. This is the Negative Predictive Value (NPV).

It is possible to compute the PPV and NPV from the sensitivity and specificity, but only if you know the prevalence of the disease in the population you are testing.

Example

You examined the ROC curve, and chose a test value to use as the cutoff between "normal" and "abnormal". For this cutoff, the sensitivity is 90% and the specficity is 95%. In the population you are testing, the prevalence of the disease is 10%. What are the PPV and NPV? You can figure it out by filling in a table.

1. Assume a value for the total number of patients examined. In the end, everything will be a ratio, so this value doesn't matter much. I chose 10,000 and put that into the bottom right of the table.

2. The prevalence is 10%, so 1,000 patients will have the disease and 9,000 will not. These values form the bottom (total) row of the table.

3. The sensitivity is 90%, so 0.9*1,000=900 people with the disease (left column) will have a positive test, and 100 will not . Thse values go into the left column.

4. The specificity is 95%, so 0.95*9000= 8550 people without the disease will have a negative test. That leaves 450 with a positive test. These values go into the second (disease absent) column.

5. Fill in the last (total) column.

6. The positive predictive value is the fraction of people with a positive test who have the disease: 900/1350 = 66.7%

7. The negative predictive value is the fraction of those with a negative test who do not have the disease: 8550/8650= 98.8%

	Disease present	Disease Absent	Total
Positive test	900	450	1,350
Negative test	100	8,550	8,650
Total	1,000	9,000	10,000

If you want to automate these calculations (perhaps in Excel), the bottom of this page (from MedCalc) gives the necessary equations.

5.19.1.7 Comparing ROC curves

Prism does not compare ROC curves. It is, however, quite easy to manually compare two ROC curves created with data from two different (unpaired) sets of patients and controls.

1. Separately use Prism to create two ROC curves by separately analyzing your two data sets.

2. For each ROC curve, note the area under the curve and standard error (SE) of the area.

3. Combine these results using this equation:

$$z = \frac{|Area_1 - Area_2|}{\sqrt{SE^2_{Area1} + SE^2_{Area2}}}$$

4. If you investigated many pairs of methods with indistinguishable ROC curves, you would expect the distribution of z to be centered at zero with a standard deviation of 1.0. To calculate a two-tail P value, therefore, use the following Microsoft Excel function:

   =2*(1-NORMSDIST(z))

The method described above is appropriate when you compare two ROC curves with data collected from different subjects. A different method is needed to compare ROC curves when both laboratory tests were evaluated in the same group of patients and controls. To account for the correlation between areas under your two curves, use the method described by Hanley and McNeil (1).

1. Hanley, J.A., and McNeil, B. J. (1983). Radiology 148:839-843. Accounting for the correlation leads to a larger z value and, thus, a smaller P value.

5.19.2 Comparing Methods with a Bland-Altman Plot

5.19.2.1 How to: Bland-Altman plot
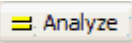
A Bland-Altman plot compares two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis.

1. Enter the data

Create a new table. Choose the Column tab. If you don't have data yet, choose the sample data: Bland-Altman method comparison.

Enter the measurements from the first method into column A and for the other method into column B. Each row represents one sample or one subject.

2. Choose the Bland-Altman analysis

From the data table, click ⇄ Analyze on the toolbar, and then choose Bland-Altman from the list of one-way analyses.
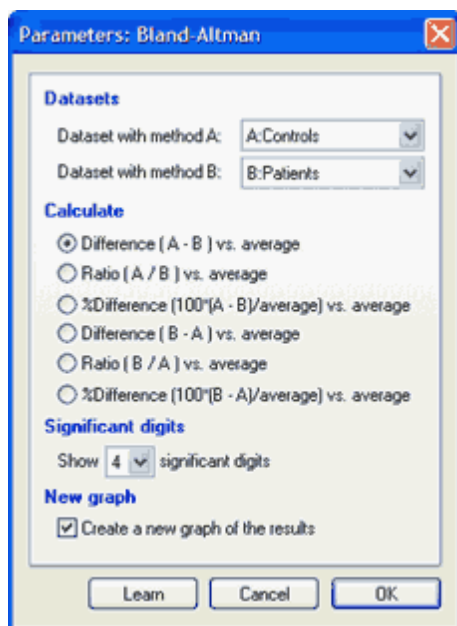
Designate the columns with the data (usually A and B), and choose how to plot the data. You can plot the difference, the ratio, or the percent difference. If the difference between methods is consistent, regardless of the average value, you'll probably want to plot the difference. If the difference gets larger as the average gets larger, it can make more sense to plot the ratio or the percent difference.

3. Inspect the results

The Bland-Altman analysis creates two pages of results. The first page shows the difference and average values, and is used to create the plot. The second results page shows the bias[640], which is the average of the differences, and the 95% limits of agreement[640].

If you used the sample data, the two methods have very similar results on average, and the bias (difference between the means) is only 0.24. The 95% limits of agreement are between -13.4 and 13.9.

Bland-Altman method comp...		
Bias & agreement		Value
1	Bias	0.238095
2	SD of bias	6.96351
3	95% Limit of agreement	
4	From	-13.4104
5	To	13.8866

4. Plot the Bland-Altman graph
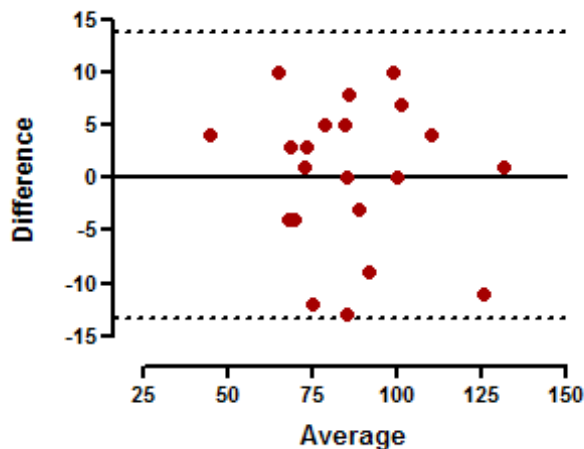
The 95% limits of agreement are shown as two dotted lines. To create these, double click on the Y axis to bring up Format Axis. At the bottom of that dialog, click the "..." button (Windows, shown below) or the gear icon (Mac) to bring up the Format Additional Ticks and Grids dialog. Then click the fish hook icon to 'hook' the Y location of the grid line to an

analysis constant created by the Bland-Altman analysis. Repeat for the other grid line.
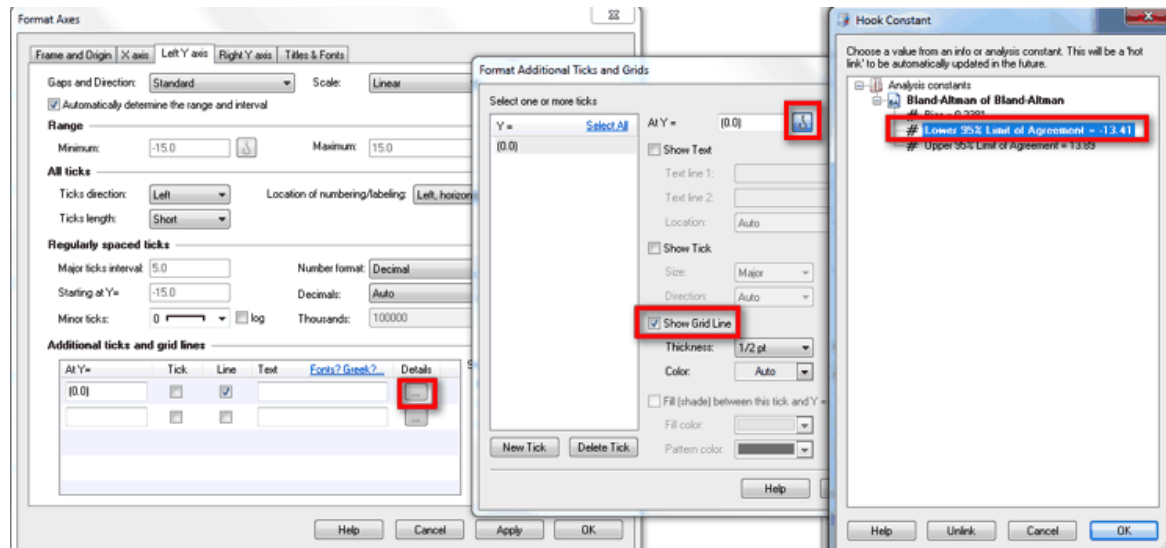
The origin of the graph was moved to the lower left (and offset) on the first tab of the Format Axes dialog.

5.19.2.2 Interpreting results: Bland-Altman

Difference vs. average

The first page of Bland-Altman results shows the difference and average values and is used to create the plot.

Bias and 95% limits of agreement

The second results page shows the average bias, or the average of the differences. The bias is computed as the value determined by one method minus the value determined by the other method. If one method is sometimes higher, and sometimes the other method is higher, the average of the differences will be close to zero. If it is not close to zero, this indicates that the two assay methods are systematically producing different results.

This page also shows the standard deviation (SD) of the differences between the two assay methods (labeled as the SD of bias) and the 95% *limits of agreement*, computed as the mean difference (bias) plus or minus 1.96 times its SD. If we assume a Normal distribution for the differences, and a large enough sample size that the sample mean and sample SD are quite close to their population values, then 95% of the

differences between the two assay methods would be expected to be within the range of values described by the 95% limits of agreement.

Especially with small sample sizes, the sample mean and sample SD may not have values close to the true population mean and SD. To account for this possible discrepancy, it is possible to compute *95% prediction bands* for the difference between the two assay methods. These 95% prediction bands are wider than the 95% limits of agreement (especially for small sample sizes), and so provide a more accurate prediction of where to expect future differences between the two assay methods to be found. Prism does not compute the prediction bands but they can easily be computed by hand using a formula on page 146 of a review by Giavarina (1).

Interpreting the Bland-Altman results

Bland-Altman plots are generally interpreted informally, without further analyses. Ask yourself these questions:

- How big is the average discrepancy between methods (the bias)? You must interpret this clinically. Is the discrepancy large enough to be important? This is a clinical question, not a statistical one.

- How wide are the limits of agreement? If it is wide (as defined clinically), the results are ambiguous. If the limits are narrow (and the bias is tiny), then the two methods are essentially equivalent.

- Is there a trend? Does the difference between methods tend to get larger (or smaller) as the average increases?

- Is the variability consistent across the graph? Does the scatter around the bias line get larger as the average gets higher?

1. Giavarina, D. (2015). Understanding Bland Altman analysis. Biochem Med *25* : 141–151.

5.19.2.3 Analysis checklist: Bland-Altman results

✔ **Are the data paired?**

The two values on each row must be from the same subject.

✔️ **Are the values entered into the two columns actual results of lab results?**

Prism computes the Bland-Altman plot from raw data. Don't enter the differences and means, and then run the Bland-Altman analysis. Prism computes the differences and means.

✔️ **Are the two values determined independently?**

Each column must have a value determined separately (in the same subject). If the value in one column is used as part of the determination of the other column, the Bland-Altman plot won't be helpful.

5.20 Analyzing a stack of P values

Prism lets you input a stack of P values computed

elsewhere and decide which ones are low enough

to be worthy of further study.

5.20.1 Key concepts: Analyzing a stack P values

- Our goal with Prism has always been to make basic biostatistics very accessible and easy. Analyzing a stack of P values is really is beyond "basic biostatistics". Nonetheless, the concepts are not all that complicated.

- Enter a bunch of P values that were computed elsewhere into a column A of a Prism Column table.

- This analysis then determines which of those P values are small enough for the corresponding comparison to be worth investigating further.

5.20.2 How to: Analyzing a stack of P values

1. Create a Column table

From the Welcome or New Table dialog, choose the Column tab.

2. Enter the P values

Enter P values, calculated elsewhere, into column A.

Every value must be a P value (between 0.0 and 1.0). You cannot enter "<0.0001" or "ns" or "**". Each P value you enter must be a decimal fraction.

Optionally, enter a column title adjacent to each P value to identify the corresponding comparison. This will make it easier to understand the results. If you don't enter row titles, Prism will label the results using the original row number instead.

This analysis is usually used with hundreds or thousands of P values, but can be done with any number.

3. Choose the analysis

Click Analyze, look in the list of Column analyses, and choose Analyze a stack of P values.
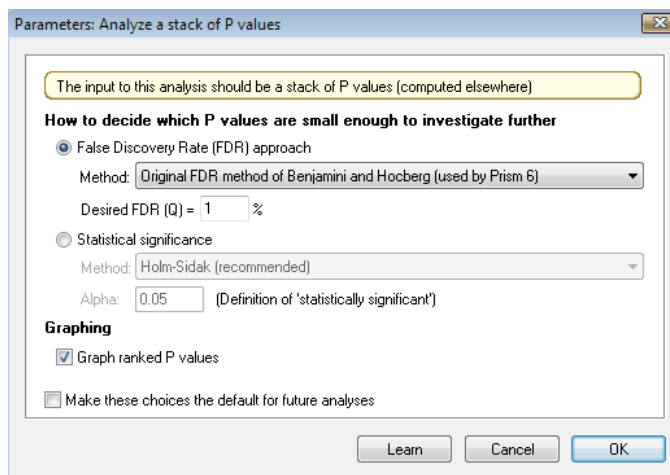
Choose one of two general approaches: Control the False Discovery Rate[138] (FDR) or control the Type I error rate for the family of comparisons[128]. Then choose the actual method you want Prism to use.

Control the False Discovery Rate (FDR)

Choose one of three approaches[386] to decide which P values are small enough to flag as "discoveries". We recommend the adaptive method of

Benjamini, Krieger and Yekutieli (1), as it has more power. The original FDR method of Benjamini and Hochberg (2) is better known, but has less power. The method of Benjamini and Yekutieli makes fewer assumptions, but has much less power.

Also set the value of Q, with the goal that no more than Q% of the comparisons flagged as discoveries, are false discoveries (false positives). Enter a percentage, not a fraction. If you want at most 5% of discoveries to be false, enter "5" not "0.05" and not "5%". There really is no standard value. If you enter a larger value, more comparisons will be flagged as discoveries, but more of these will be false discoveries.

Statistical significance (control Type I error rate for the family of comparisons)

Choose how Prism will correct for multiple comparisons. We recommend using the [Holm-Šídák method][133], which has the most power. The alternatives are the methods of [Bonferroni-Dunn][131] or [Bonferroni-Šídák][131]. The method we call *Bonferroni-Dunn* is often referred to simply as *Bonferroni*. The method we call Bonferroni-Šídák is often referred to simply as Šídák. The two are very similar. The Bonferroni-Šídák method has a bit more power but is less widely understood.

Set the value of alpha that applies to the entire family of P values. Prism then decides which P values are small enough for the related comparison to be designated "statistically significant" after correcting for multiple comparisons. Enter a fraction, not a percentage. If you want 5% of comparisons under the null hypothesis to be falsely flagged as "significant", enter 0.05 not 5. If you enter a larger value, more comparisons will be flagged as "significant", but more of these will be false positives. For statistical significance, alpha is often set to 0.05 when making one or a few comparisons. But when making many comparisons, you'll probably want to enter a higher value.

Graphing

Check the option if you want to see a graph of P value rank vs. P value. This is a common way to visualize the distribution of P values.
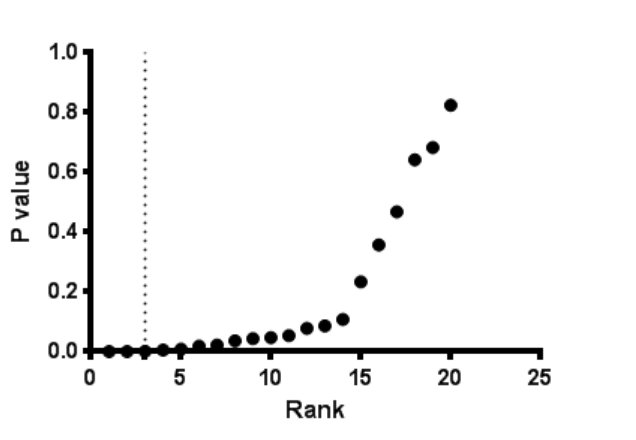
References

1. Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93, 491–507 (2006).

2. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289–300 (1995).

3. Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165–1188.

5.20.3   Interpreting results: Analyzing a stack of P values

Statistical significance approach

Adjusted P values

You entered each P value on a different row, and the results table has the same number of rows, each with three values:

- The first column states whether the comparison associated with that P values is statistically significant or not, stated simply as yes or no.

- The second column shows the P value you entered.

- The third column shows the adjusted P value. If the alpha value you entered exactly equaled this adjusted P value, then this comparison would be right at the border of being defined as "statistically significant" or not.

Significant results

This table shows only the comparisons flagged as "statistically significant". The table is sorted, with the smallest P value on top. Each row shows the row title (or row number in the data table, if you didn't enter titles), the P value you entered and the adjusted P value.

Ranked P value

This table is only created if you checked the option to graph the P value distribution. The X values are the ranks of the P values (1 is smallest) and the Y values are the P values themselves. The table is not all that useful by itself, but is used to automatically create the graph.

Threshold P value

A floating note on the results page tells you the threshold P value. If a P value is less than this threshold, the result of the associated comparison is considered to be "significant". This is not the same as the value of alpha you entered. With the Bonferroni method, for example, the threshold equals alpha/K where K is the number of P values you entered.

False Discovery Rate method

q values

You enter each P value on a different row, and the results table has the same number of rows, each with three values:

- The first column states whether the comparison associated with that P values is defined to be a "discovery" or not, stated simply as yes or no.

- The second column shows the P value you entered.

- The third column shows the q value. If the Q value you entered exactly equaled this q value, then this comparison would be right at the border of being defined a discovery.

Discoveries

This table shows only the comparisons flagged as "discoveries". The table is sorted, with the smallest P value on top. Each row shows the row title (or row number in the data table, if you didn't enter titles), the P value you entered and the q value

Ranked P value

This table is only created if you checked the option to graph the P value distribution. The X values are the ranks of the P values (1 is smallest) and the Y values are the P values themselves. The table is not all that useful by itself, but is used to automatically create the graph.

Threshold P value

A floating note on the results page tells you the threshold P value. If a P value is less than this threshold, the result of the associated comparison is considered to be a "discovery". This is not the same as the value of alpha you entered.

Estimated number of true null hypotheses

If you chose the adaptive method of Benjamini, Krieger and Yekutieli, Prism will report in a floating note the estimated number of true null hypotheses. The adaptive method works by first estimating this value, then using it when deciding which P values are small enough to be called "discoveries".

5.21 Simulating data and Monte Carlo simulations

Simulating data is a powerful way to understand statistical analyses and plan experiments. Monte Carlo analysis lets you simulate many data sets, analyze each, and then look at the distribution of parameters (results) of those analyses. This can let

you "experiment" with alternative experimental designs via computer before you collect any data. Prism makes it easy, without requiring any programming or scripting.

5.21.1 Simulating a XY data table

To simulate a family of XY data sets with random error, start from any data table or graph, click Analyze, open the Simulate data category, and then select Simulate XY Data.

X values tab. Generate a regular series (arithmetic or geometric) of X values or use the X values from the data table you are analyzing.

Equation tab. You can choose to use Y values from the data table you are analyzing, and then add random scatter. More often, you will choose an equation on this tab.

Parameter values tab. On top of the tab, choose how many data sets you wish to simulate, and how many replicates each data set will have. The main part of the tab is where you enter the values of each parameter. If you choose to simulate more than one data set, then you can choose to enter a parameter value just for one data set, or to enter a parameter that applies to several, or all, curves. Choose the data sets on the top part of the dialog, and enter the parameter values for that data set (or that group of data sets) below.

Random error tab. Choose among several methods for generating random scatter and also adding outliers.

5.21.2 Simulating a Column data table

To simulate a set of Column data sets with random error, start from any data table or graph, click Analyze, open the Simulate data category, and then select Simulate Column Data.

Experimental design tab

Choose the number of data sets. In most cases you will want to specify the population mean of each column. But you can instead enter the mean and SD of a set of means, and have each column mean chosen randomly.

For each column you wish to simulate, specify how many rows it will have, its column mean, and its title. Note you can hook the column mean to a results computed in another analysis.

Random error tab

Choose among several methods for generating random scatter and also adding outliers.

5.21.3 Simulating a Contingency data table

Experimental design

Choose the total sample size, the grand total of all four cells in the contingency table.

Also choose the experimental design:

- Cross-sectional. You sample subjects without regard to their exposure or disease.

- Prospective. You choose subjects based on exposure defined by the rows, and then inquire about outcome defined by the columns.

- Experimental. You assign treatments that define the row, and then tabulate outcome that defines the columns.

- Case-control-retrospective. You choose cases (with disease) and controls and these define the columns. You then look back at exposure and tabulate these results into rows.

Rows and columns

Enter titles for both rows and both columns.

Enter various probabilities as asked. The details depend on the experimental method you chose.

5.21.4 How to: Monte Carlo analyses

How to begin a Monte Carlo analysis

Simulate a data table using[650] one of Prism's simulation analyses.. Note that these simulations include random scatter, so will produce new results when they are updated.

1. Analyze that simulated data set as appropriate.

2. From that results page, click Analyze and choose Monte-Carlo analysis. This analysis will repeat the simulations many times, and tabulate selected results. The Monte Carlo analysis will only be available for analyses that create analysis constants. Note that linear regression does not, but you can fit a straight line with the nonlinear regression analysis.

The explanations below explain the basic ideas of the Monte Carlo analysis. Follow the example[657]to learn the details.

Simulations tab

How many simulations?

How many simulations should you run? If you make only a few simulations, the results will be affected too much by chance. Running more simulations will give you more precise results, but you'll have to wait longer for the calculations to repeat. When just playing around, it might make sense to use as few as 100 simulations so you can see the results instantly. When trying to polish simulation results, it can make sense to use as many as 10,000 or 100,000 simulations. A good compromise is 1000.

Append?

If you go back to run more simulations, check an option box to append the new simulations to the existing results, rather than start fresh.

Random seed

The choice of random numbers used in a series of simulations depends on the random number seed used to generate the first set of results. By default, Prism picks this seed automatically (based on the time of day),

and presents this seed in a floating note superimposed on Monte Carlo results.

If you want two or more Monte Carlo analyses to use precisely the same data sets (so you can compare two ways of analyzing those data), enter that random seed on the Simulation tab.

Parameters to tabulate tab

Prism lists all of the analysis constants generated by the analysis. Check the ones whose values you want to tabulate.

You cannot change the set of values included in this list. Let us know if there are parameters missing, and we can add them in a future version.

Hits tab (optional)

If you skip this Hits tab, Prism will tabulate the selected parameters (different columns) for each simulation (rows).

Prism can also reduce the results down to a single number -- the fraction of the simulations that are "hits". Define a hit to be when a value tabulated by the analysis equals a certain value or is within a specified range. Click New...Graph of existing data from this table, and choose a parts-of-whole graph to create a pie graph of the fraction of hits vs. not hits.

Prism can also tabulate the selected parameters only for simulations that are hits, and/or for only the simulations that are not hits. Choose any or all of these options (Hits, Not hits, All simulations) at the bottom of the Hits tab. Each option you check will create its own results table.

5.21.5 Monte Carlo example: Accuracy of confidence intervals

Overview

When you fit a curve with nonlinear regression, one of the most important set of results are the 95% confidence intervals of the parameters. These intervals are computed from the standard errors which are based on some mathematical simplifications. They are called "asymptotic" or "approximate" standard errors. They are calculated assuming that the

equation is linear, but are applied to nonlinear equations. This simplification means that the intervals can be too optimistic, too narrow, so your true confidence level may be less than 95%.

How can you know whether the intervals really do have 95% confidence? There is no general way to answer this. But for any particular situation, you can get an answer using simulations. This page explains how to do this with Prism using the Monte Carlo feature new to Prism 6. We will simulate a dose-response curve and ask how accurate the 95% confidence intervals are for the Hill Slope. Christopolous suggested that the distribution of HillSlope can be asymmetrical, and suggested fitting the logarithm of the HillSlope instead (1).

Step 1. Simulating the first experiment

From anywhere, click New..Analysis and choose Simulate XY Data. To follow this example exactly, make these choices:

- X values tab. Start at X=-9, increment each X value by 0.5, and stop when X equals or exceeds -3.0.

- Equation tab. Choose the folder "Dose-response - Stimulation", and choose the equation: log(agonist) vs. response --Variable slope.

- Parameter values tab. Choose to simulate one data set, with 3 replicate values. Set Bottom=250, Top=5000, logEC50=-6, and HillSlope=0.5.

- Random error tab. Random error is Gaussian (absolute) with SD=200.

Step 2. Fit the first experiment

1. From the graph, click Analyze and choose Nonlinear regression. Or click the nonlinear regression shortcut button in the Analysis part of the toolbar.

2. On the first (Fit) tab, choose the folder: Dose-response - Stimulation. Then choose the equation: log(agonist) vs. response --Variable slope. You can accept all the defaults for the other tabs.

3. Click OK, and Prism will fit the model to the data and graph the curve on the graphs.

Step 3. Simulate a few more data sets

Note the Hill Slope, the parameter we wish to investigate in this simulation. To simulate new data with different random numbers, click the red die icon, or drop the Change menu and choose Simulate Again. Note that the Hill slope changes.

Step 4. Simulate many data sets

Start from the nonlinear regression results, click Analyze and choose Monte Carlo simulation.

On the first (Simulations) tab, choose how many simulations you want Prism to perform. For this example, enter 1000.

On the second (Parameters to tabulate) tab, choose which parameters you want to tabulate. The choice is the list of analysis constants that Prism creates when it analyzes the data. For this example, we only want to tabulate the two confidence limits for the Hill Slope.

On the third (Hits) tab, define a criterion which makes a given simulated results a "hit". For this example, we'll define a hit to mean that the confidence interval brackets the true value of 0.5 (set in the simulation). So a hit is defined when the lower limit is less than or equal to 0.5 and the upper limit is greater than or equal to 0.5.
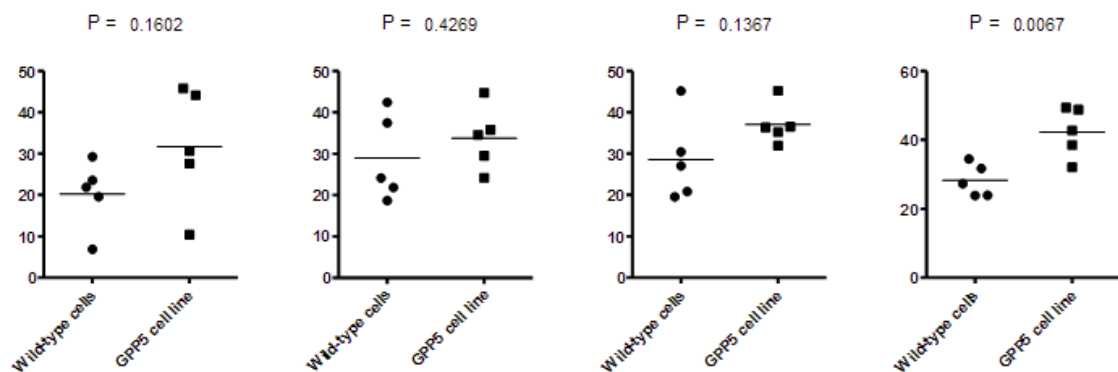
Click OK and Prism will run the simulations. Depending on the speed of your computer, it will take a few or a few dozen seconds.

Step 5. Monte-Carlo results

The results of the simulations are shown on only one page (since we unchecked all the options for reporting individual simulations in the Hits tab above.
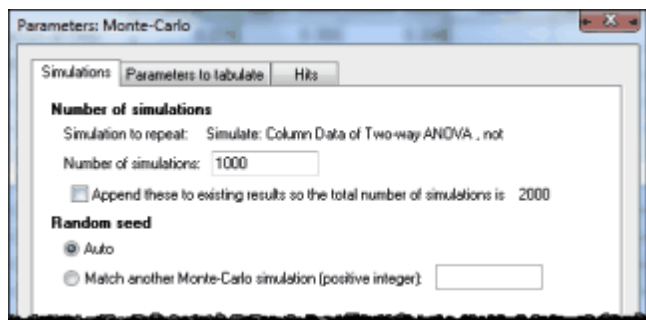
The fraction of hits is 0954. For this example, we defined "hit" to mean that the confidence interval included the true parameter value. In other words, the confidence interval for HillSlope included the true value (used in the simulation) in 95.4% of the 1000 simulated data sets. (Since these results depend on which random numbers Prism generates, you'll get somewhat different results when you try to follow this example). This is what you'd expect for 95% confidence intervals, and means that you can trust the confidence intervals in experiment of this design. There is no need to consider fitting the logarithm of the Hill Slope instead, for this experimental design

Step 6. Try variations

If the confidence interval for this result is too wide for your tastes (it ranges from 93.9% to 96.6%), go back and run this Monte Carlo simulation with many more iterations (perhaps 10,000).

The conclusion is specific to the experimental design and parameter values you chose. Go back to the simulation, and change the HillSlope to 4.0. Then change the Monte Carlo dialog to redefine a "hit" to mean that the confidence interval brackets the true value of 4.0. A hit is defined when the lower limit is less than or equal to 4.0 and the upper limit is greater than or equal to 4.0. With such a steep HillSlope, the confidence intervals really are not accurate at all. The "95%" confidence intervals only include the true value (4.0) in 83.1% of the simulations. Doubling the number of concentrations (changing the increment of X values from 0.5 down to 0.25 in the X values tab of the Simulate XY data dialog)

solves this problem. With so many data points, the 95% confidence intervals contain the true value in 95.3% of the simulations.

By varying the choices in the simulation dialog, and seeing the effect via Monte Carlo simulations, you can design better experiments.

Another example

Here is another detailed example, using simulations to find the power of a t test.

Reference

1. Arthur Christopoulos, Assessing the distribution of parameters in models of ligand-receptor interaction: to log or not to log, Trends in Pharmacological Sciences, Volume 19, Issue 9, 1 September 1998, Pages 351-357

Prism file for this example

5.21.6 Monte Carlo example: Power of unpaired t test

Overview

This example will compute the power of an unpaired t test. The goal of this example, however, is broader -- to show how easy it is to perform Monte Carlo analyses with Prism and to show you how useful they can be.

The question here is this: Given a certain experimental design and assumptions about random scatter, what is the chance (power) that an unpaired t test will give a P value less than 0.05 and thus be declared statistically significant?

Step 1. Simulate the first experiment

From anywhere, click New..Analysis and choose Simulate Column Data. Choose to simulate two groups, with five values per group, sampled from populations with means of 25 and 35 distributed according to a Gaussian distribution with a SD of 10.

Step 2. Analyze the data with a t test

From the simulated data table, click Analyze and choose t test from the list of Column analyses. Accept all the default choices to perform an unpaired t test, reporting a two-tail P value.

Step 3. View a few simulated results

Copy the P value from the results and paste onto a graph of the data. It will paste with a live link, so the P value will change if the values change. To simulate new data with different random numbers, click the red die icon, or drop the Change menu and choose Simulate Again

The layout below shows four such graphs placed on the layout as unlinked pictures that do not update when the graph changes. Even though there is only one graph in the project, this made it possible to put four different versions of it (with different random data) onto the layout. You can see that with random variation of the data, the P value varies a lot.
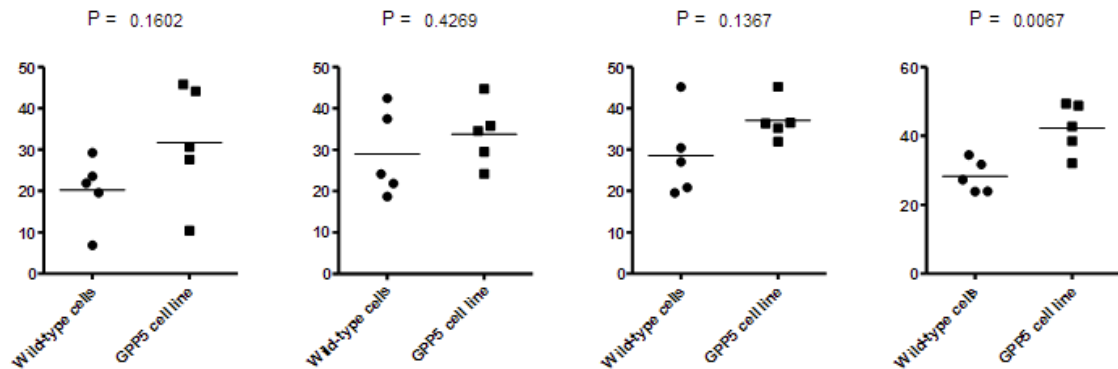
Step 4. Monte Carlo simulation

Start from the t test result, click Analyze and choose Monte Carlo simulation.

On the first (Simulations) tab, choose how many simulations you want Prism to perform. This example used 1000 simulations.
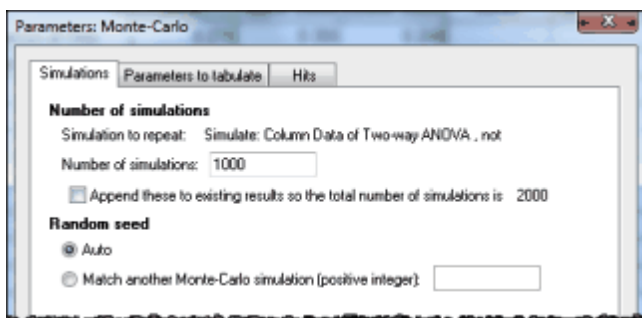
On the second (Parameters to tabulate) tab, choose which parameters you want to tabulate. The choice is the list of analysis constants that Prism creates when it analyzes the data. For this example, we only want to tabulate the P value (from the t test which compares means; don't mix it up with the P value from the F test which compares variances).
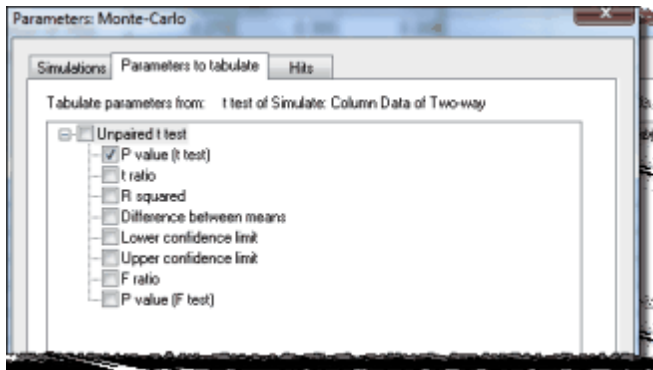
On the third (Hits) tab, define a criterion which makes a given simulated result a "hit". For this example, we'll define a hit to mean statistical significance with P<0.05.
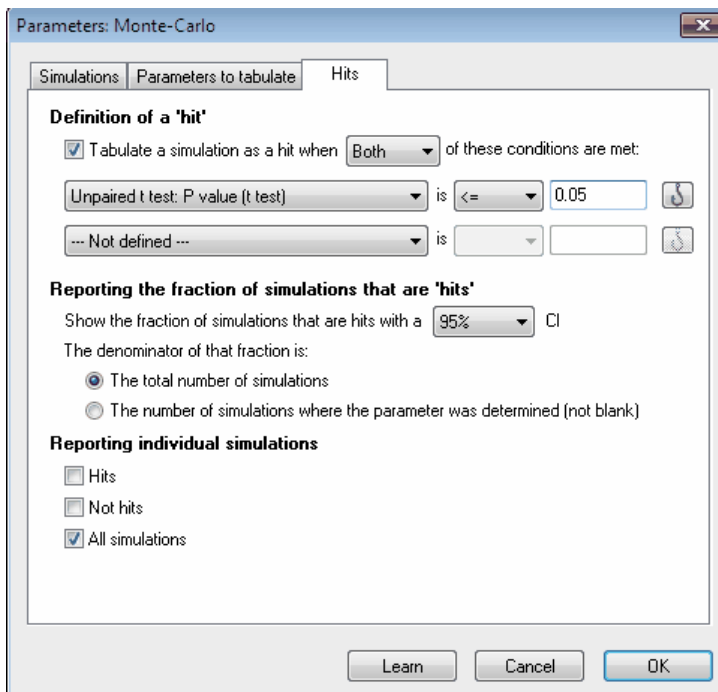
Click OK and Prism will run the simulations. Depending on the speed of your computer, it will take a few seonds or a few dozen seconds.

Step 5. Monte-Carlo results

Distribution of P values

The results of the simulations are shown in two pages.

One shows the tabulated parameters for all simulations. In this example, we only asked to tabulate the P value, so this table is a list of 1000 (the number of simulations requested) P values. To create a frequency distribution from this table, click Analyze, and choose Frequency Distribution. Choose a cumulative frequency distribution. You can see that about a quarter of the P values are less than 0.05.
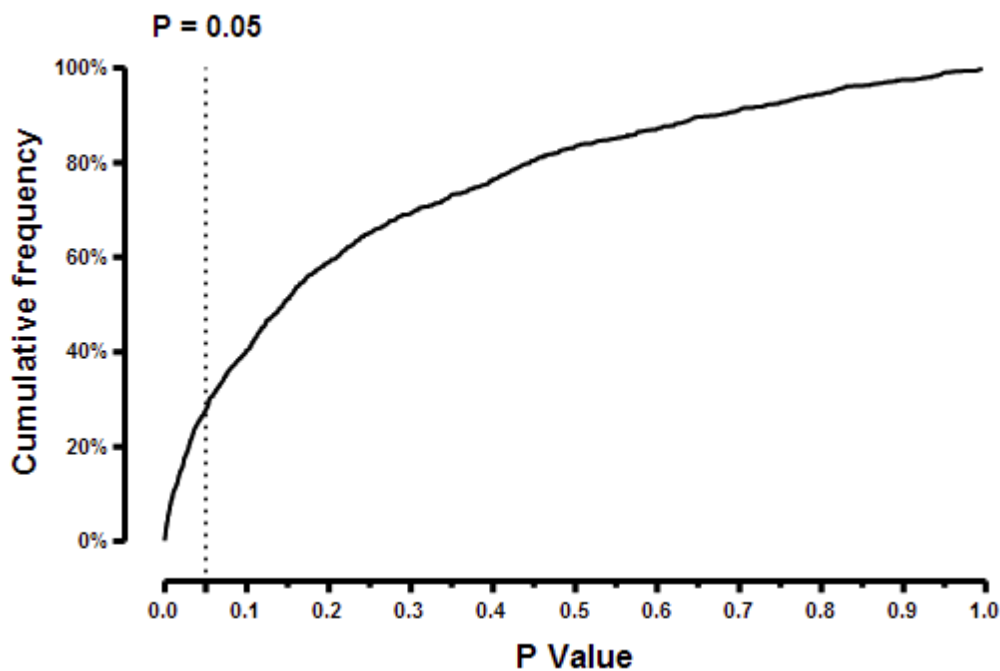
Fraction of hits

The other results table summarizes the fraction of hits. For this set of simulations, 27.5% of the simulations were hits (P value less than 0.05), with a 95% confidence interval ranging from 24.8% to 30.4%. Another way of stating these results is that the power of our experimental design is 27.5%.

Monte Carlo Fraction of hits		A		
		Data Set-A		
		Mean	Upper Limit	Lower Limit
1	Fraction of 'Hits'	0.275	0.304	0.248
2	Fraction of 'Not hits'	0.725	0.752	0.696

Note that the simulations depend on random number generation, which is initialized by the time of day you begin. So if your results might not be identical to what is shown above.

If we had run more simulations, of course that confidence interval would be narrower.

From this table, click New...Graph of existing data to create a pie or percentage plot.
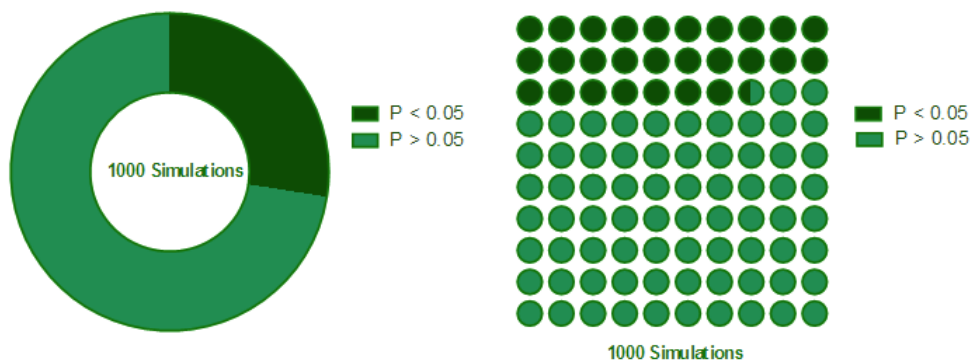
Step 6. Further explorations

Go back to step 1 and simulate a larger experiment, say with 10 values in each group. Or 20 or 100. How much will that increase the power?

Try reducing the definition of hit to be a P value less than 0.01 rather than 0.05. How does that affect power?

5.21.7 Using a script to simulate many data sets

This page will help people who set up these scripts in earlier versions of Prism, or people for whom the Monte-Carlo analysis doesn't provide

enough versatility. But for most people, the [Monte Carlo analysis](#)[652] will be much easier.

Follow these steps:

1. Create a Prism project that [simulates](#)[650] a data set with random scatter (the first results sheet), and then fits a curve through these data with nonlinear regression (second results sheet).

2. Make sure that the first data table is empty and formatted for column data (no X column; no Y subcolumns). The script will write the results into this table. You can click in the upper left corner of the data table to change its format, if needed.

3. Click the Prism button at the left of the toolbar, and choose Run Script. Then choose New Script, and start with a blank page.

4. Enter this script:

```
Table Prism 1 Clear
Foreach 100
 Goto R 1
 Regenerate
 Goto R 2
 WTable "logEC50",5,1
Next
```

5. Click the Run button on the Script dialog.

6. After the script is complete, go to the first data table and view the results.

7. Click analyze, and choose to do Column statistics (with normality test) or perhaps create a Frequency Distribution.

The first line of the script specifies which data table will hold the results. Then the script loops 100 times. With each loop, it goes to the first results page (with the simulation) and regenerates with new random scatter. It then goes to the second results page (with curve fit results) and writes the value in the fifth row of the first column into the data table, and labels that column "logEC50". You will want to adjust the row number, and perhaps add additional lines to output additional results. You may also want to loop more than 100 times.

5.21.8　How Prism generates random numbers

Random or pseudo-random?

The only way to generate truly random numbers is through a random physical process, such as tossing dice or measuring intervals between radioactive decays. Prism, like all computer programs, generates "random" numbers from defined calculations. Since the sequence of numbers is reproducible, mathematicians say that the numbers are "pseudo-random".

The difference between truly random and pseudo-random numbers rarely creates a problem. For most purposes, computer-generated random numbers are random enough to simulate data and test analytical methods.

Prism uses the time of day when calculating the first random number, so you will get a different series of random numbers every time you run the program.

Random numbers from a Gaussian distribution

Prism generates random values from a Gaussian distribution using routines adapted from ideas presented in Numerical Recipes in C, (W. H. Press et al, second edition, Cambridge Press, 1992). The function RAN3 (defined in Numerical Recipes) generates uniformly distributed random numbers and the function GASDEV transforms them to a Gaussian distribution with a mean of zero and a standard deviation you enter.

If you choose relative error, Prism first calculates a random number from a Gaussian distribution with a mean of zero and with a SD equal to the percent error you enter. It then multiplies that percentage times the ideal Y value to yield the actual random value that is added to the Y value.

Random numbers from a t distribution

Prism also can generate random numbers from a t distribution with any number of degrees of freedom (df). This lets you simulate wider scatter than Gaussian. If df is low, this distribution is very wide. If df is high (more than 20 or so), it is almost indistinguishable from a Gaussian distribution. If df=1, the distribution is extremely wide (lots of outliers)

and is identical to a Lorentzian distribution, also known as the Cauchy distribution. Prism uses this equation to generate random numbers from the t distribution with df degrees of freedom:

$$\frac{Rand}{\sqrt{\sum_{i=1}^{df} \frac{Rand^2}{df}}}$$

In this equation, Rand is a random number drawn from a Gaussian distribution with mean=0 and SD=1. To compute a random number from a t distribution with df degrees of freedom, Prism generates df+1 different random numbers drawn from a Gaussian distribution.

Random numbers from a Poisson or Binomial distribution

We adapted ideas from pages 372-377 of Numerical Recipes, third edition, by WH Press and colleagues.

Note an important difference between Poisson and binomial random values and Gaussian.

- With Gaussian random values, the simulated value equals the calculated ideal value *plus* a random error drawn from the Gaussian distribution (with a mean of zero and a SD you specify). We say "plus" the random value, but this random value will be negative half the time (on average), so the simulated value will be less than the idea value half the time.

- With Poisson, the simulated value *is* a random value drawn from a Poisson distribution, using the ideal value computed in the simulation as the population value of the Poisson distribution. With Gaussian errors, you specify a standard deviation in the simulate dialog. Not so with Poisson error. The simulated Poisson value is always an integer.

- With binomial, the simulated value *is* a random variable drawn from a binomial distribution with a population proportion (or percentage) computed by the simulation, and using a sample size you entered.

Random numbers with Excel

If you want to do simulations not possible with Prism, it might be possible with Excel. The methods listed below work well with Excel 2003 and later, but should not be used with earlier versions of Excel.

Excel offers two random number functions.

- The RANDBETWEEN(low, high) formula chooses a random integer randomly from the range between (and including) the low and high values entered. The result is always an integer, so this function is not very useful.

- The RAND() function generates a random value between 0.0 and 1.0. It is an Excel function, at least in name, so it needs to be followed by parentheses, but the function has no inputs so there is nothing between the parentheses.

If you want to generate random numbers from a Gaussian distribution, you'll need to transform the values generated with the RAND() function. This Excel formula computes a random number from a Gaussian distribution with a mean of 0.0 and a SD of 1.0.

```
=NORMSINV(RAND())
```

The RAND() function calculates a random number from 0 to 1. the NORMSINV() function takes a fraction between 0 and 1 and tells you how many standard deviations you need to go above or below the mean for a cumulative Gaussian distribution to contain that fraction of the entire population.

Multiple by the standard deviation and add a mean, and you'll have random numbers drawn from a Gaussian distribution with that mean and SD. For example, use this formula to sample from a Gaussian distribution with a mean of 100 and a SD of 15:

```
=(NORMSINV(RAND())*15)+100
```

Index

- * -

- 5 -

- A -

- B -

- C -

- N -

- O -

- P -

- T -

- U -

- V -

- W -

- Y -