

GraphPad Statistics Guide

**GraphPad Software Inc.
www.graphpad.com**

© 1995-2014 GraphPad Software, Inc.

This is one of three companion guides to GraphPad Prism 6.
All are available as web pages on graphpad.com.

Table of Contents

Foreword	0
Part I PRINCIPLES OF STATISTICS	9
1 The big picture.....	9
When do you need statistical calculations?	9
The essential concepts of statistics	10
Extrapolating from 'sample' to 'population'	13
Why statistics can be hard to learn	13
Ordinal, interval and ratio variables	14
The need for independent samples	16
Intuitive Biostatistics (the book)	17
2 The Gaussian distribution.....	18
Importance of the Gaussian distribution	18
Origin of the Gaussian distribution	19
The Central Limit Theorem of statistics	21
The lognormal distribution	21
3 Standard Deviation and Standard Error of the Mean.....	23
Key concepts: SD	23
Computing the SD	25
How accurately does a SD quantify scatter?	26
Key concepts: SEM	28
Computing the SEM	28
The SD and SEM are not the same	29
Advice: When to plot SD vs. SEM	30
Alternatives to showing the SD or SEM	31
4 Confidence intervals.....	31
Key concepts: Confidence interval of a mean	32
Interpreting a confidence interval of a mean	34
Other confidence intervals	36
Advice: Emphasize confidence intervals over P values	36
One sided confidence intervals	37
Compare confidence intervals, prediction intervals, and tolerance intervals	38
Confidence interval of a standard deviation	39
5 P Values.....	41
What is a P value?	41
The most common misinterpretation of a P value	41
More misunderstandings of P values	42
One-tail vs. two-tail P values	43
Advice: Use two-tailed P values	45
Advice: How to interpret a small P value	46
Advice: How to interpret a large P value	47
How Prism computes exact P values	48
6 Hypothesis testing and statistical significance.....	49
Statistical hypothesis testing	49
Extremely significant?	50
Advice: Avoid the concept of 'statistical significance' when possible	50
A Bayesian perspective on interpreting statistical significance	51

	A legal analogy: Guilty or not guilty?	53
	Advice: Don't keep adding subjects until you hit 'significance'	53
7	Statistical power.....	55
	Key concepts: Statistical Power	56
	An analogy to understand statistical power	57
	Type I, II (and III) errors	58
	Using power to evaluate 'not significant' results	59
	Why doesn't Prism compute the power of tests	61
	Advice: How to get more power	63
8	Choosing sample size.....	64
	Overview of sample size determination	64
	Why choose sample size in advance?	66
	Choosing alpha and beta for sample size calculations	68
	What's wrong with standard values for effect size?	69
	Sample size for nonparametric tests	71
9	The problem of multiple comparisons.....	72
	The multiple comparisons problem	72
	Approach 1: Don't correct for multiple comparisons	74
	Approach 2: Correct for multiple comparisons	75
	Approach 3: False Discovery Rate (FDR)	76
	Lingo: Multiple comparisons	77
	Multiple comparisons traps	78
	Planned comparisons	82
	Example: Planned comparisons	84
	The Bonferroni method	87
10	Testing for equivalence.....	88
	Key concepts: Equivalence	88
	Testing for equivalence with confidence intervals or P values	89
11	Nonparametric tests.....	92
	Key concepts: Nonparametric tests	92
	Advice: Don't automate the decision to use a nonparametric test	92
	The power of nonparametric tests	93
	Nonparametric tests with small and large samples	94
	Advice: When to choose a nonparametric test	95
	Lingo: The term "nonparametric"	96
12	Outliers.....	97
	An overview of outliers	98
	Advice: Beware of identifying outliers manually	99
	Advice: Beware of lognormal distributions	99
	How it works: Grubb's test	101
	How it works: ROUT method	102
	The problem of masking	104
	Simulations to compare the Grubbs' and ROUT methods	105
13	Analysis checklists.....	109
	Unpaired t test	109
	Paired t test	111
	Ratio t test	113
	Mann-Whitney test	114
	Wilcoxon matched pairs test	115
	One-way ANOVA	116
	Repeated measures one-way ANOVA	118
	Kruskal-Wallis test	120

Friedman's test	121
Two-way ANOVA	122
Repeated measures two-way ANOVA	124
Contingency tables	125
Survival analysis	126
Outliers	128

Part II STATISTICS WITH PRISM 6 130

1 Getting started with statistics with Prism.....	130
What's new in Prism 6 (statistics)?	130
Statistical analyses with Prism	131
Guided examples: Statistical analyses	132
2 Descriptive statistics and frequency distributions.....	133
Column statistics	134
How to: Column statistics.....	134
Analysis checklist: Column statistics.....	136
Interpreting results: Mean, geometric mean and median.....	138
Interpreting results: Quartiles and the interquartile range.....	139
Interpreting results: SD, SEM, variance and coefficient of variation (CV).....	141
Interpreting results: Skewness and kurtosis.....	142
Interpreting results: One-sample t test.....	143
Interpreting results: Wilcoxon signed rank test.....	144
Interpreting results: Normality tests.....	147
Frequency Distributions	148
Visualizing scatter and testing for normality without a frequency distribution.....	148
How to: Frequency distribution.....	149
Graphing tips: Frequency distributions.....	153
Fitting a Gaussian distribution to a frequency distribution.....	154
Describing curves	156
Smoothing, differentiating and integrating curves.....	156
Area under the curve.....	159
Row statistics	162
Overview: Side-by-side replicates.....	162
Row means and totals.....	163
3 Normality tests.....	163
How to: Normality test	164
How normality tests work	164
Interpreting results: Normality tests	165
Q&A: Normality tests	166
4 Identifying outliers.....	169
How to: Identify outliers	169
Analysis checklist: Outliers	172
5 One sample t test and Wilcoxon signed rank test.....	173
How to: One-sample t test and Wilcoxon signed rank test	174
Interpreting results: One-sample t test	174
Interpreting results: Wilcoxon signed rank test	175
6 t tests, Mann-Whitney and Wilcoxon matched pairs test.....	178
Paired or unpaired? Parametric or nonparametric?	178
Entering data for a t test.....	178
Choosing a test to compare two columns.....	179
Options for comparing two groups.....	181

What to do when the groups have different standard deviations?.....	182
Q&A: Choosing a test to compare two groups.....	185
The advantage of pairing.....	186
Unpaired t test	188
How to: Unpaired t test from raw data.....	188
How to: Unpaired t test from averaged data.....	189
Interpreting results: Unpaired t.....	191
The unequal variance Welch t test.....	193
Graphing tips: Unpaired t.....	195
Advice: Don't pay much attention to whether error bars overlap.....	196
Analysis checklist: Unpaired t test.....	198
Paired or ratio t test	200
How to: Paired t test.....	200
Testing if pairs follow a Gaussian distribution.....	201
Interpreting results: Paired t.....	202
Analysis checklist: Paired t test.....	203
Graphing tips: Paired t.....	205
Paired or ratio t test?.....	206
How to: Ratio t test.....	206
Interpreting results: Ratio t test.....	207
Analysis checklist: Ratio t test.....	208
Mann-Whitney or Kolmogorov-Smirnov test	209
Choosing between the Mann-Whitney and Kolmogorov-Smirnov tests.....	209
How to: MW or KS test.....	210
Interpreting results: Mann-Whitney test.....	213
The Mann-Whitney test doesn't really compare medians.....	216
Analysis checklist: Mann-Whitney test.....	218
Why the results of Mann-Whitney test can differ from prior versions of Prism	220
Interpreting results: Kolmogorov-Smirnov test.....	220
Analysis checklist: Kolmogorov-Smirnov test.....	222
Wilcoxon matched pairs test	223
"The Wilcoxon test" can refer to several statistical tests	223
How to: Wilcoxon matched pairs test.....	224
Results: Wilcoxon matched pairs test.....	226
Analysis checklist: Wilcoxon matched pairs test.....	228
How to handle rows where the before and after values are identical.....	230
Multiple t tests	230
How to: Multiple t tests.....	231
Options for multiple t tests.....	232
7 One-way ANOVA, Kruskal-Wallis and Friedman tests.....	234
How to: One-way ANOVA	234
Entering data for one-way ANOVA and related tests.....	234
Which multiple comparisons tests does Prism offer?.....	236
Experimental design tab: One-way ANOVA.....	237
Multiple comparisons tab: One-way ANOVA.....	239
Options tab: One-way ANOVA.....	241
Q&A: One-way ANOVA.....	244
One-way ANOVA	246
Interpreting results: One-way ANOVA.....	246
Analysis checklist: One-way ANOVA.....	249
Repeated-measures one-way ANOVA	251
What is repeated measures?.....	251
Sphericity and compound symmetry.....	251

Quantifying violations of sphericity with epsilon.....	255
Interpreting results: Repeated measures one-way ANOVA.....	256
Analysis checklist: Repeated-measures one way ANOVA.....	258
Kruskal-Wallis test	260
Interpreting results: Kruskal-Wallis test.....	260
Analysis checklist: Kruskal-Wallis test.....	262
Friedman's test	263
Interpreting results: Friedman test.....	263
Analysis checklist: Friedman's test.....	264
8 Multiple comparisons after ANOVA.....	265
The problem of multiple comparison	266
Bonferroni and Sidak methods	266
Tukey and Dunnett methods	269
The Holm-Sidak approach to multiple comparisons	270
Fisher's Least Significant Difference (LSD)	271
Testing for linear trend	272
Multiple comparisons after repeated measures ANOVA	274
Nonparametric multiple comparisons	274
Multiplicity adjusted P values	275
Beware of using multiple comparisons tests to compare dose-response curves or time courses	277
Q&A: Multiple comparisons tests	279
How Prism computes multiple comparison tests	283
9 Two-way ANOVA.....	284
How to: Two-way ANOVA	284
A note of caution for statistical novices.....	285
Deciding which factor defines rows and which defines columns?.....	285
Entering data for two-way ANOVA.....	286
Entering repeated measures data.....	287
Missing values and two-way ANOVA.....	289
Point of confusion: ANOVA with a quantitative factor.....	291
Experimental design tab: Two-way ANOVA.....	293
Multiple comparisons tab: Two-way ANOVA.....	296
Options tab: Two-way ANOVA.....	300
Summary of multiple comparisons available (two-way).....	302
Q&A: Two-way ANOVA.....	303
Ordinary (not repeated measures) two-way ANOVA	304
Interpreting results: Two-way ANOVA.....	304
Graphing tips: Two-way ANOVA.....	307
How Prism computes two-way ANOVA.....	307
Analysis checklist: Two-way ANOVA.....	309
Repeated measures two-way ANOVA	310
Interpreting results: Repeated measures two-way ANOVA.....	310
Graphing tips: Repeated measures two-way ANOVA.....	311
Analysis checklist: Repeated measures two-way ANOVA.....	313
Beware of three-way ANOVA	314
10 Categorical outcomes.....	317
Contingency tables	318
Key concepts: Contingency tables.....	318
How to: Contingency table analysis.....	319
Fisher's test or chi-square test?.....	322
Interpreting results: Relative risk and odds ratio.....	323
Interpreting results: Sensitivity and specificity.....	325
Interpreting results: P values from contingency tables.....	326

Analysis checklist: Contingency tables.....	327
Graphing tips: Contingency tables.....	328
The Confidence Interval of a proportion	328
How Prism can compute a confidence interval of a proportion.....	328
How to compute the 95% CI of a proportion.....	329
The meaning of “95% confidence” when the numerator is zero.....	332
A shortcut equation for a confidence interval when the numerator equals zero	333
Compare observed and expected distributions	333
How to: Compare observed and expected distributions.....	333
How the chi-square goodness of fit test works.....	335
The binomial test.....	336
McNemar's test.....	338
Don't confuse with related analyses.....	339
Analysis Checklist: Comparing observed and expected distributions.....	340
11 Survival analysis.....	340
Key concepts. Survival curves	341
How to: Survival analysis	342
Q & A: Entering survival data	343
Example of survival data from a clinical study	345
Example of survival data from an animal study	346
Analysis choices for survival analysis	347
Interpreting results: Kaplan-Meier curves	350
Interpreting results: P Value	351
Interpreting results: The hazard ratio	352
Interpreting results: Ratio of median survival times	355
Interpreting results: Comparing >2 survival curves	356
Multiple comparisons of survival curves	358
Analysis checklist: Survival analysis	359
Graphing tips: Survival curves	361
Q&A: Survival analysis	363
Determining the median follow up time	368
12 Correlation.....	369
Key concepts: Correlation	370
How to: Correlation	370
Interpreting results: Correlation	371
Analysis checklist: Correlation	373
The difference between correlation and regression	373
13 Diagnostic lab analyses.....	375
ROC Curves	375
Key concepts: Receiver-operator characteristic (ROC) curves.....	375
How to: ROC curve.....	376
Interpreting results: ROC curves.....	377
Analysis checklist: ROC curves.....	379
Calculation details for ROC curves.....	380
Computing predictive values from a ROC curve.....	381
Comparing ROC curves.....	382
Comparing Methods with a Bland-Altman Plot	383
How to: Bland-Altman plot.....	383
Interpreting results: Bland-Altman.....	385
Analysis checklist: Bland-Altman results.....	386
14 Simulating data and Monte Carlo simulations.....	387
Simulating a data table	388

How to: Monte Carlo analyses	389
Monte Carlo example: Power of unpaired t test	391

Index	397
--------------	------------

1 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](#)^[72] and the many ways you can [get trapped by multiple comparisons](#)^[78].
- [Testing for equivalence](#)^[88]
- The [danger of using outlier tests with lognormal distributions](#)^[99] and the problem of [masking](#)^[104] which can make it harder to find two outliers than to find one.
- [Why it doesn't make sense to automate the decision](#)^[92] to use a nonparametric test or not.
- The [distinction between SD and SEM](#)^[29], and [when to display each](#)^[30].
- The [advantages of reporting confidence intervals](#)^[36].
- The [most common misunderstanding about P values](#)^[41], and other [misunderstandings](#)^[42].
- [Why you can't peek at the results and add more subjects if the results are not quite significant yet.](#)^[53]
- A simple [analogy to understand statistical power](#)^[57].
- A set of [analysis checklists](#)^[109]. Each checklist lists questions you should ask yourself before accepting the results of a statistical analysis.

The [second half of the guide](#)^[130] 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.

1.1 The big picture

1.1.1 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.

1.1.2 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 Ouija board -- a method to produce preordained results, rather an objective method of analyzing data.

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.

"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%.

1.1.3 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.

1.1.4 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.

1.1.5 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 temperature'. However, temperature in degrees Kelvin is a ratio variable, as 0.0 degrees Kelvin really does mean 'no temperature'. Another counter example is pH. It is not a ratio variable, as $\text{pH}=0$ 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.

1.1.6 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.

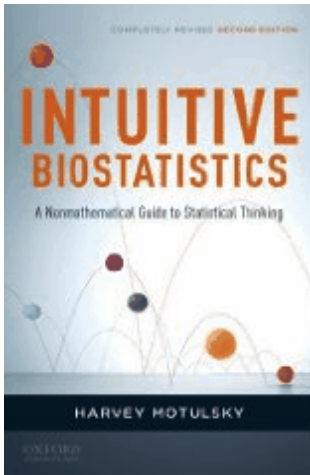
The concept of independence can be difficult to grasp. Consider the following three situations.

- 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.

1.1.7 Intuitive Biostatistics (the book)



H.J. Motulsky, [Intuitive Biostatistics](#), ISBN: 978-0199730063

[Table of contents](#)

[Excerpts](#)

[Reviews](#)

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).

1.2 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

1.2.1 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.

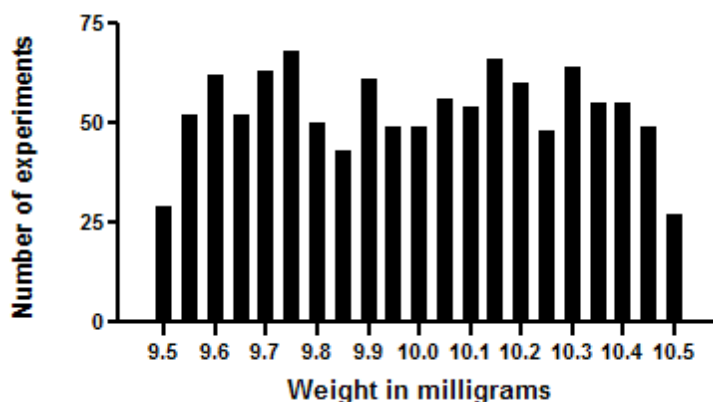
When reading in other books about the Gaussian distribution, two statistical terms might be confusing because they sound like ordinary words:

- 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”).

1.2.2 Origin of the Gaussian distribution

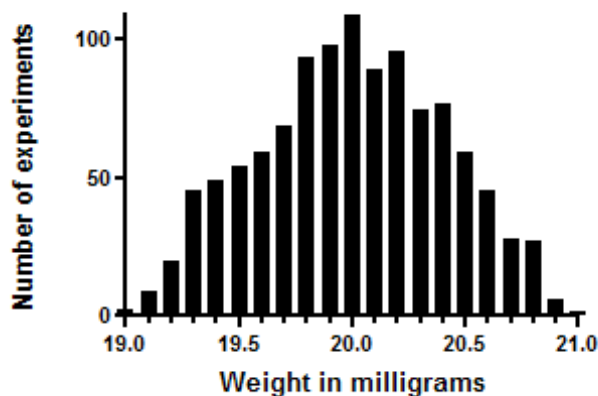
The Gaussian distribution emerges when many independent random factors act in an additive manner to create variability. This is best seen by an example.

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.

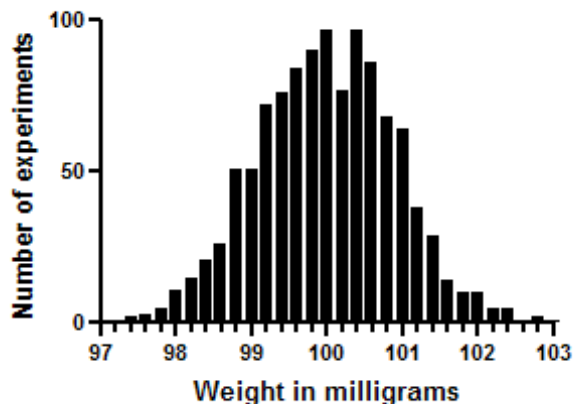
Now let's make the experiment more complicated. We pipette twice and weigh the result. On average, the weight will now be 20 milligrams. But you expect the errors to cancel out some of the time. The figure below is what you get.



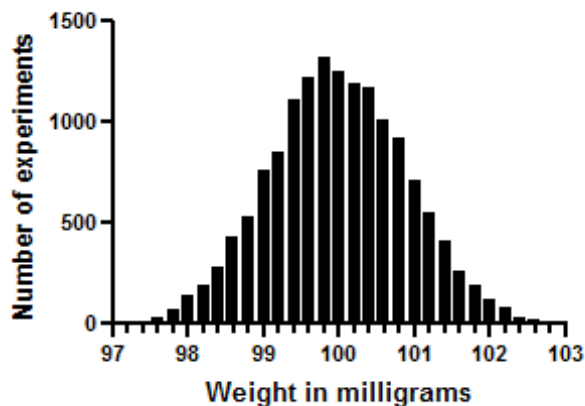
Each pipetting step has a flat random error. Add them up, and the distribution is not flat. For example, you'll get weights near 21 mg only if both pipetting steps err substantially in

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.

1.2.3 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.

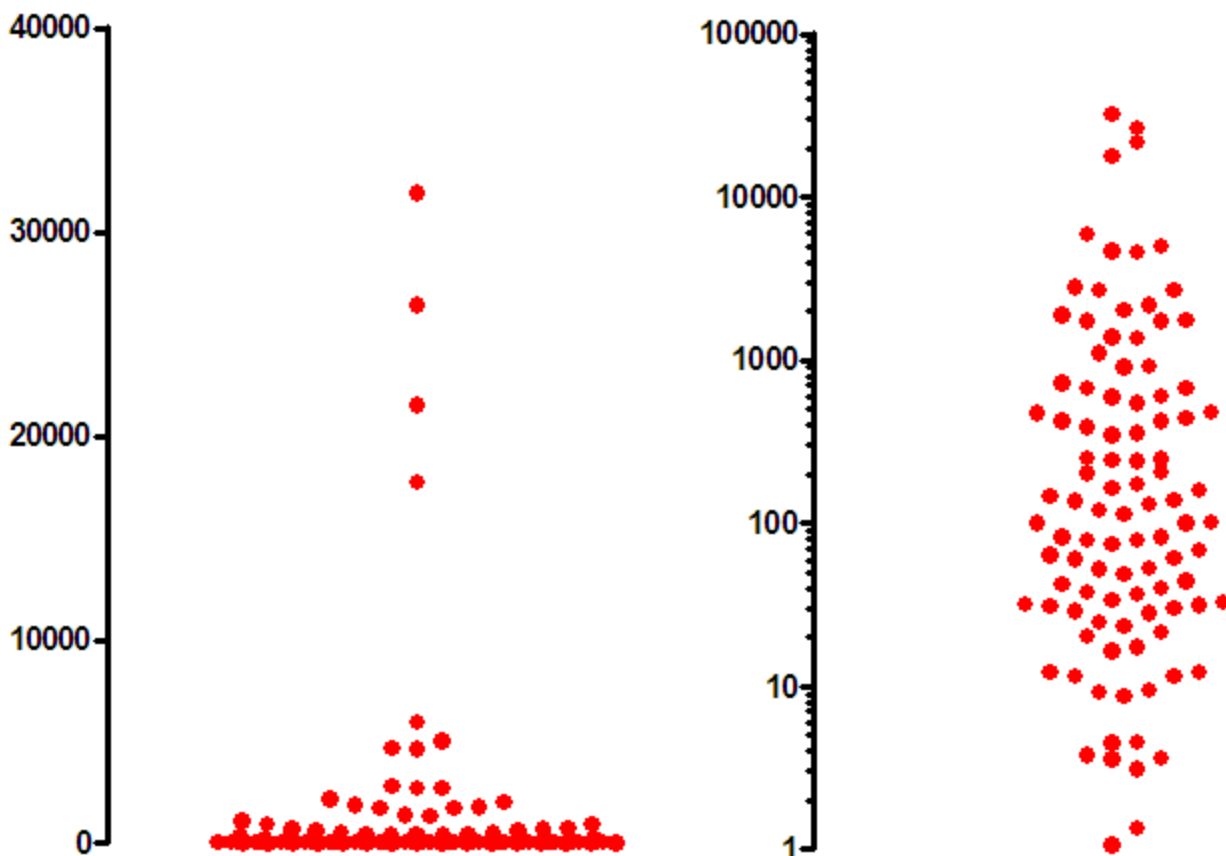
1.2.4 The lognormal distribution

The origin of the lognormal distribution

A Gaussian distribution emerges when variation is caused by [multiple sources of scatter which add together](#)^[19].

A lognormal distribution emerges when variation is caused by multiple sources of scatter which are *multiplicative*.

The two graphs below plot the same 50 values sampled from a lognormal distribution. The graph on the left has a linear (ordinary) Y axis, while the graph on the right has a logarithmic scale.



The graph on the left shows two problems:

- It is impossible to really get a sense of the distribution, since about half of the values are plotted in pile at the bottom of the graph.
- You might be tempted to remove the highest four values as outliers, since they seem to be so far from the others.

When plotted with a logarithmic axis (right), the distribution appears symmetrical, the highest points don't seem out of place, and you can see all the points. These data come from a lognormal distribution. That means that the logarithms of values follow a Gaussian distribution. Plotting such values on a logarithmic plot makes the distribution more symmetrical and easier to understand.

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.

1.3 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.

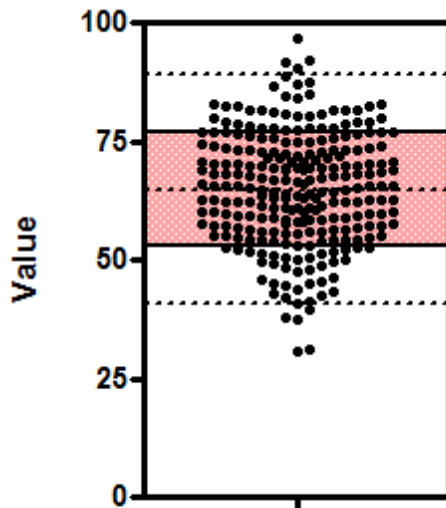
1.3.1 Key concepts: SD

What is the SD?

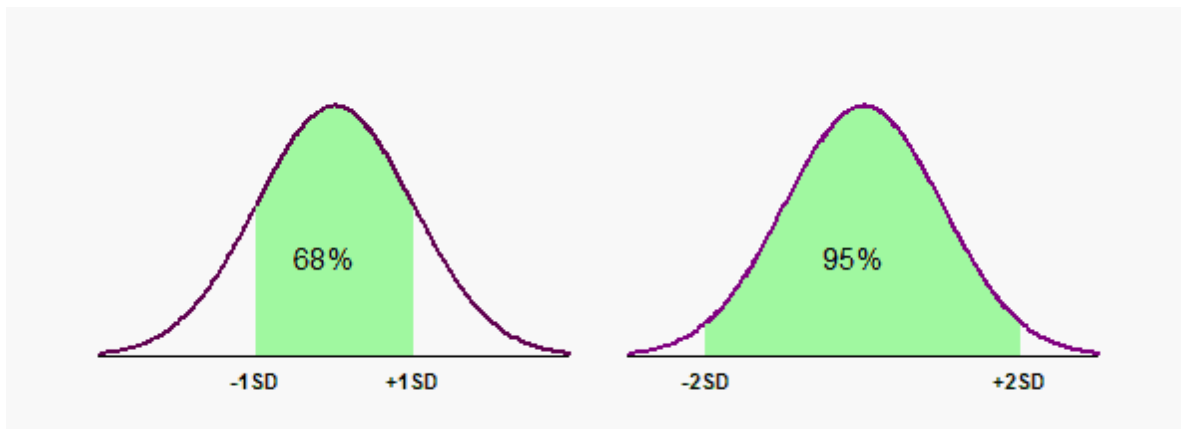
The standard deviation (SD) quantifies variability or scatter, and it is expressed in the same units as your data.

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.

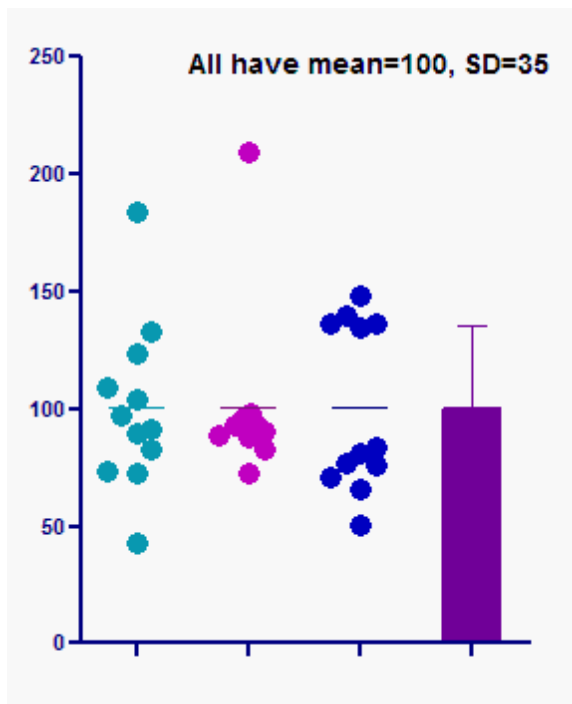


The graph that follows shows the relationship between the standard deviation and a Gaussian distribution. The area under a probability distribution represents the entire population, so the area under a portion of a probability distribution represents a fraction of the population. In the graph on the 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.



Beware, the data may not be 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 can be misleading if you assume the data are Gaussian, but that assumption isn't true.

1.3.2 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.

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](#)^[26].

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!](#)^[29]

1.3.3 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.

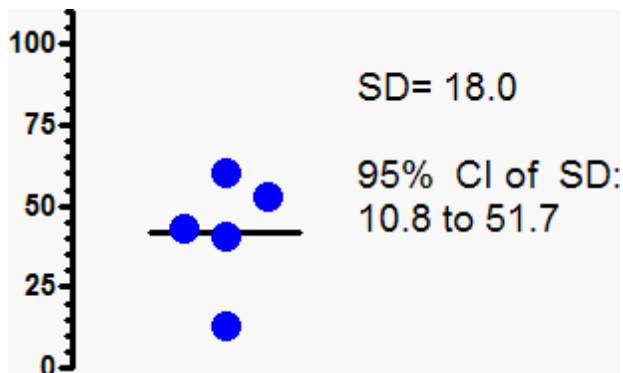
The 95% CI of the SD

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](#)^[32] 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](#)^[16] sampled from a [Gaussian](#)^[18] 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: $=SD * SQRT((N-1) / CHIINV(1-(alpha/2), N-1))$

1.3.4 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.

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!](#)²⁹

1.3.5 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:

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

1.3.6 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.

1.3.7 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.

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-whiskers 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.

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.

1.3.8 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-whiskers 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.](#)¹⁹⁶

1.4 Confidence intervals

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

1.4.1 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.

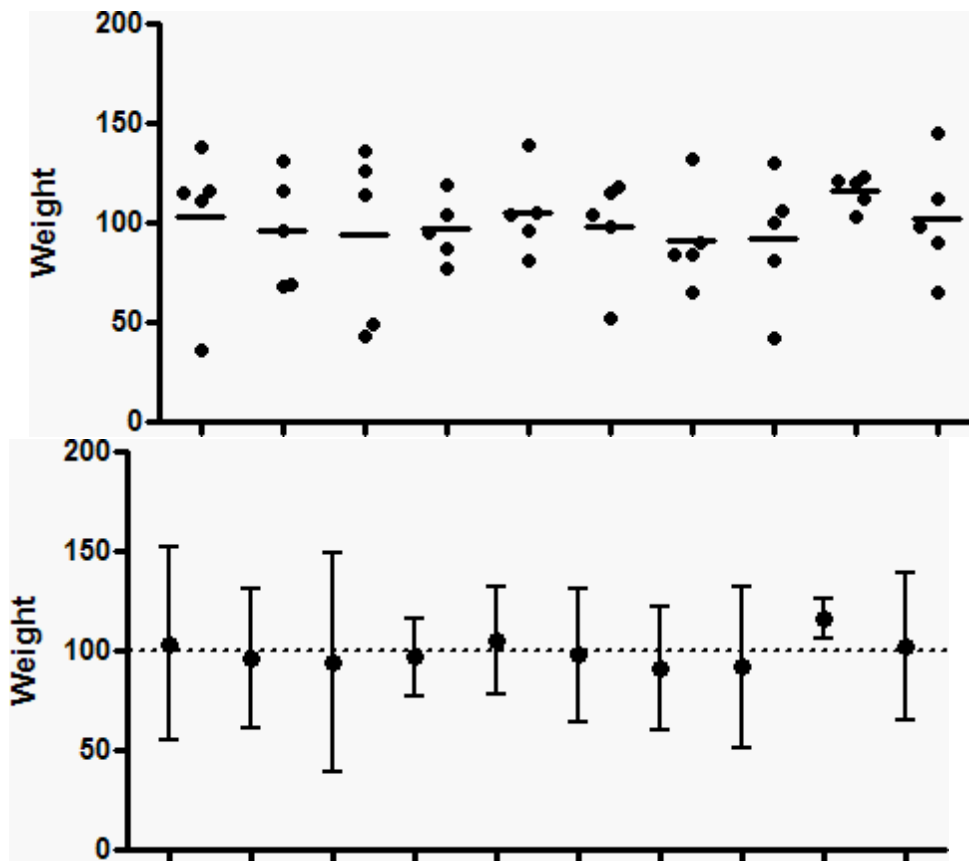
If your sample is small and variable, the sample mean is likely to be quite far from the population mean. If your sample is large and has little scatter, the sample mean will probably be very close to the population mean. Statistical calculations combine sample size and variability (standard deviation) 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](#)^[16] and randomly sampled from a population whose values are distributed according to a [Gaussian](#)^[18] 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

$$N = \text{TINV}(0.05, N-1)$$

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.

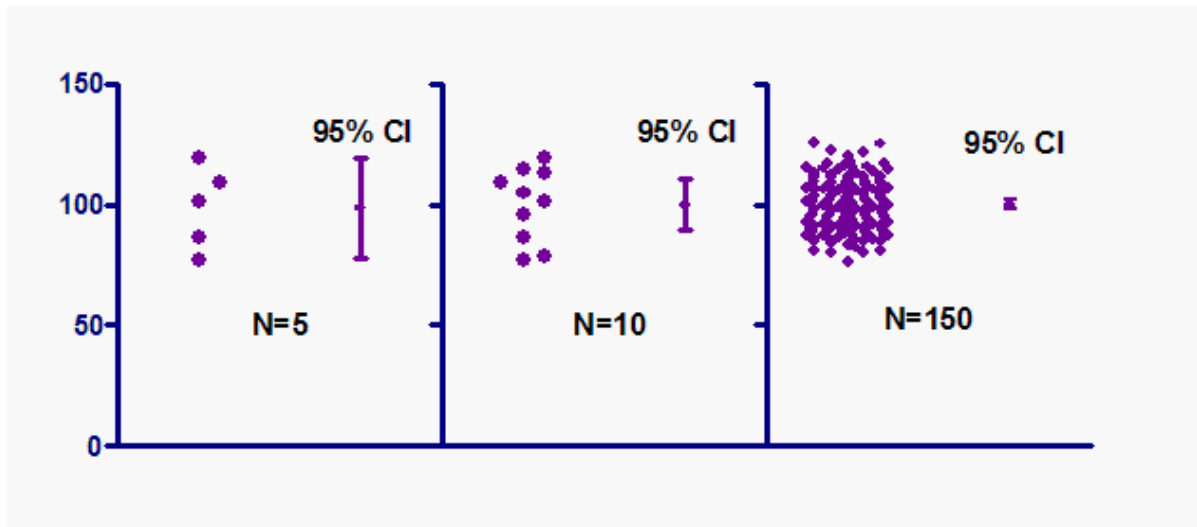
The last line in the table above shows you the equation to use to compute the multiplier in Excel.

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.

1.4.2 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.

Picky, picky, picky! 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.

Does it matter? It seems to me that it makes little difference, but some statisticians think this is a critical distinction.

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.

1.4.3 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](#)^[26]. Other confidence intervals 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 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.

1.4.4 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](#)^[41]. 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."

1.4.5 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 *Helicobacter 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).

1.4.6 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.

1.4.7 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: 1.50

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, ISBN:1584888148.

1.5 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.

1.5.1 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 disease 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?

1.5.2 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](#)^[51].

1.5.3 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 large 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*, ISBN:1591471184

1.5.4 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-sided 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-sided 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 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.

Recommendation

I recommend using only two-tailed P values for the following reasons:

- The relationship between P values and confidence intervals is more straightforward with

two-tailed P values.

- Two-tailed P values are larger (more conservative). Since many experiments do not completely comply with all the assumptions on which the statistical calculations are based, many P values are smaller than they ought to be. Using the larger two-tailed P value partially corrects for this.
- Some tests compare three or more groups, which makes the concept of tails inappropriate (more precisely, the P value has more than two tails). A two-tailed P value is more consistent with P values reported by these tests.
- Choosing one-tailed P values can put you in awkward situations. If you decided to calculate a one-tailed P value, what would you do if you observed a large difference in the opposite direction to the experimental hypothesis? To be honest, you should state that the P value is large and you found “no significant difference.” But most people would find this hard. Instead, they’d be tempted to switch to a two-tailed P value, or stick with a one-tailed P value, but change the direction of the hypothesis. You avoid this temptation by choosing two-tailed P values in the first place.

When interpreting published P values, note whether they are calculated for one or two tails. If the author didn’t say, the result is somewhat ambiguous.

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

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).

1.5.5 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 values.

1.5.6 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.](#)^[51]
- 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.

1.5.7 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.

1.5.8 How Prism computes exact P values

Calculations 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.

$$P_{\text{FromF}}(F_Ratio, DF_Numerator, DF_Denominator) = \text{BetaI}(DF_Denominator / 2, DF_Numerator / 2, DF_Denominator / (DF_Denominator + F_Ratio))$$

$$P_{\text{FromT}}(T_Ratio, DF) = \text{BetaI}(DF / 2, 1/2, DF / (DF + T_Ratio^2))$$

$$P_{\text{FromZ}}(Z_Ratio) = P_{\text{FromT}}(Z_Ratio, \text{Infinity})$$

$$P_{\text{FromR}}(R_Value) = P_{\text{FromT}}(|R_Value| / \text{SQRT}((1 - R_Value^2)/DF), DF)$$

$$P_{\text{FromChi2}}(\text{Chi2_Value}, DF) = \text{GammaQ}(DF / 2, \text{Chi2_Value} / 2)$$

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 Excel

If you want to compute P values using Excel, use these functions:

P value from F: $\text{FDIST}(F, DFn, DFd)$

P value from t (two tailed): $\text{TDIST}(t, df, 2)$ (The third argument, 2, specifies a two-tail P value.)

P value from ChiSquare: $\text{CHIDIST}(\text{ChiSquare}, DF)$

P value from z (two tailed): $\text{TDIST}(z, 10000, 2)$ (With a huge number of degrees of freedom, t and z are identical.)

or
 $2 * (1.0 - \text{NORMSDIST}(z))$

Reference

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

1.6 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.

1.6.1 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.

1.6.2 Extremely significant?

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 table. These definitions are not entirely standard. If you report the results in this way, you should define the symbols in your figure legend.

Here is the scheme that Prism uses:

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

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".

1.6.3 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 and maybe with clinical trials, it rarely occurs with basic research. 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'. Or consider simply reporting the confidence interval, without a P value.

1.6.4 A Bayesian perspective on interpreting statistical significance

Interpreting low (and high) P values is trickier than it looks.

Imagine that you are screening drugs to see if they lower blood pressure. 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](#)⁵⁶ 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.

It depends on the context of your experiment. Let's look at the same experiment performed in three alternative scenarios. In scenario A, you know a bit about the pharmacology of the drugs and expect 10% of the drugs to be active. In this case, the prior probability is 10%. In scenario B, you know a lot about the pharmacology of the drugs and expect 80% to be active. In scenario C, the drugs were selected at random, and you expect only 1% to be active in lowering blood pressure.

What happens when you perform 1000 experiments in each of these contexts? The details of the calculations are shown on pages 143-145 of *Intuitive Biostatistics*, by Harvey Motulsky (Oxford University Press, 1995). Since the power is 80%, you expect 80% of truly effective drugs to yield a P value less than 0.05 in your experiment. 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. Putting these calculations together creates these tables.

A. Prior probability=10%

	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

B. Prior probability=80%

	Drug really works	Drug really doesn't work	Total
P<0.05, "significant"	640	10	650
P>0.05, "not significant"	160	190	350
Total	800	200	1000

C. Prior probability=1%

	Drug really works	Drug really doesn't work	Total
P<0.05, "significant"	8	50	58
P>0.05, "not significant"	2	940	942
Total	10	990	1000

The totals at the bottom of each column are determined by the prior probability – the context of your experiment. The prior probability equals the fraction of the experiments that are in the leftmost column. To compute the number of experiments in each row, use the definition of power and alpha. Of the drugs that really work, you won't obtain a P value less than 0.05 in every case. You chose a sample size to obtain a power of 80%, so 80% of the truly effective drugs yield "significant" P values and 20% yield "not significant" P values. Of the drugs that really don't work (middle column), you won't get "not significant" results in every case. Since you defined statistical significance to be "P<0.05" (alpha=0.05), you will see a "statistically significant" result in 5% of experiments performed with drugs that are really inactive and a "not significant" result in the other 95% .

If the P value is less than 0.05, so the results are "statistically significant", what is the chance that the drug is, in fact, active? The answer is different for each experiment.

Prior probability	Experiments with P<0.05 and...		Fraction of experiments with P<0.05 where drug really works
	Drug really works	Drug really doesn't work	
A. Prior probability=10%	80	45	80/125 = 64%
B. Prior probability=80%	640	10	640/650 = 98%
C. Prior probability=1%	8	50	8/58 = The analysis checklists are part of Prism's help system, and have proven to be quite useful. We reprint them here, without the surrounding discussion of the tests. But the checklists alone might prove useful, even if only provoking you to read more about these tests. 14%

For experiment A, the chance that the drug is really active is 80/125 or 64%. If you observe a statistically significant result, there is a 64% chance that the difference is real and a 36% chance that the difference simply arose in the course of random sampling. For experiment B, there is a 98.5% chance that the difference is real. In contrast, if you observe a significant result in experiment C, there is only a 14% chance that the result is real and an 86% chance that it is due to random sampling. For experiment C, the vast majority of “significant” results are due to chance.

You can't interpret a P value in a vacuum. Your interpretation depends on the context of the experiment. Interpreting results requires common sense, intuition, and judgment.

1.6.5 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.

1.6.6 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:~~

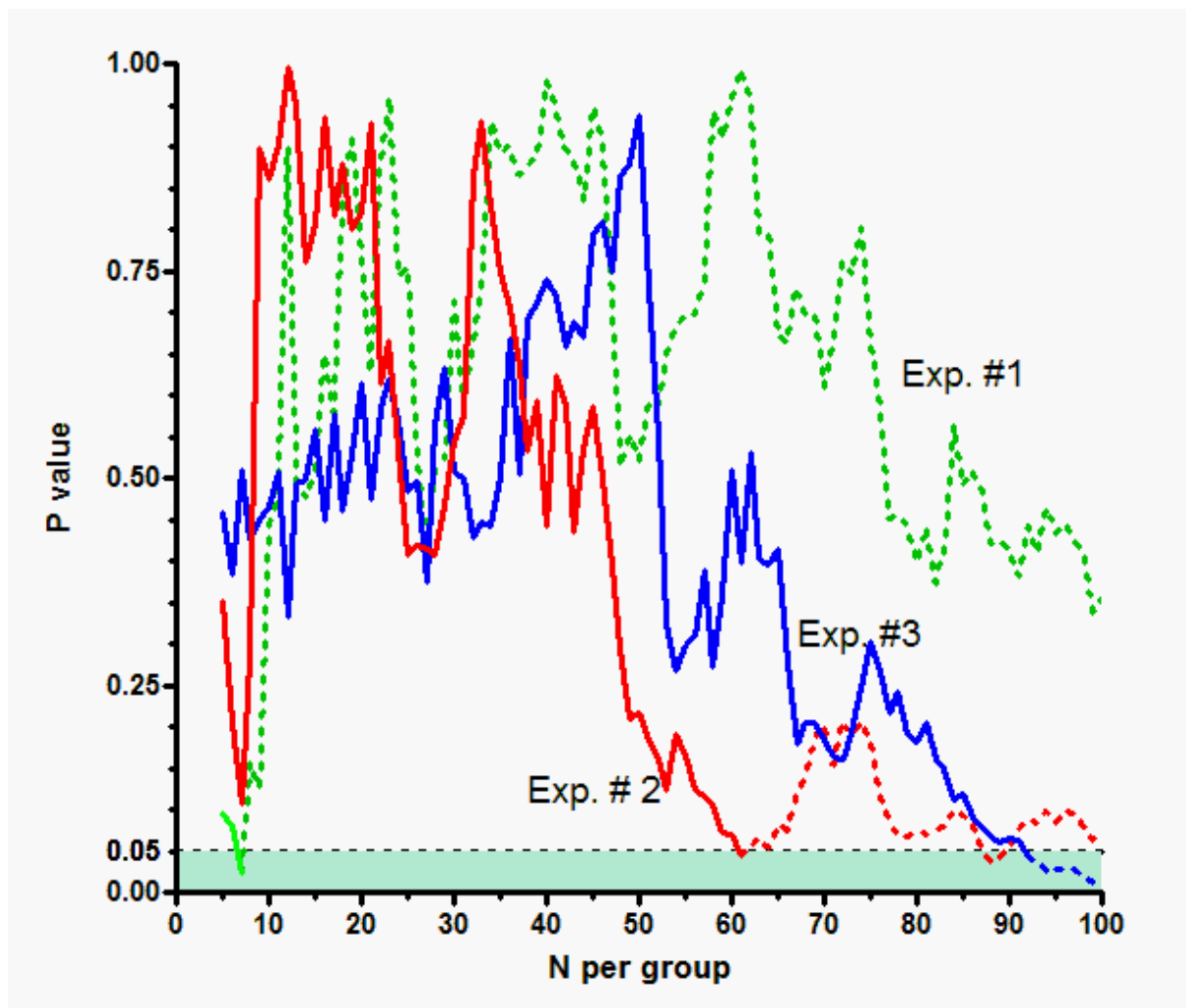
- ~~• If the result is not statistically significant, collect some more data, and reanalyze.~~
- ~~• If the result is statistically significant, stop the study.~~

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](#).^[58] 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.

1.7 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.

1.7.1 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](#)^[58].

GraphPad StatMate

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

1.7.2 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.

1.7.3 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.

1.7.4 Using power to evaluate 'not significant' results

Example data

Motulsky et al. asked whether people with hypertension (high blood pressure) had altered numbers of α_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 α_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.

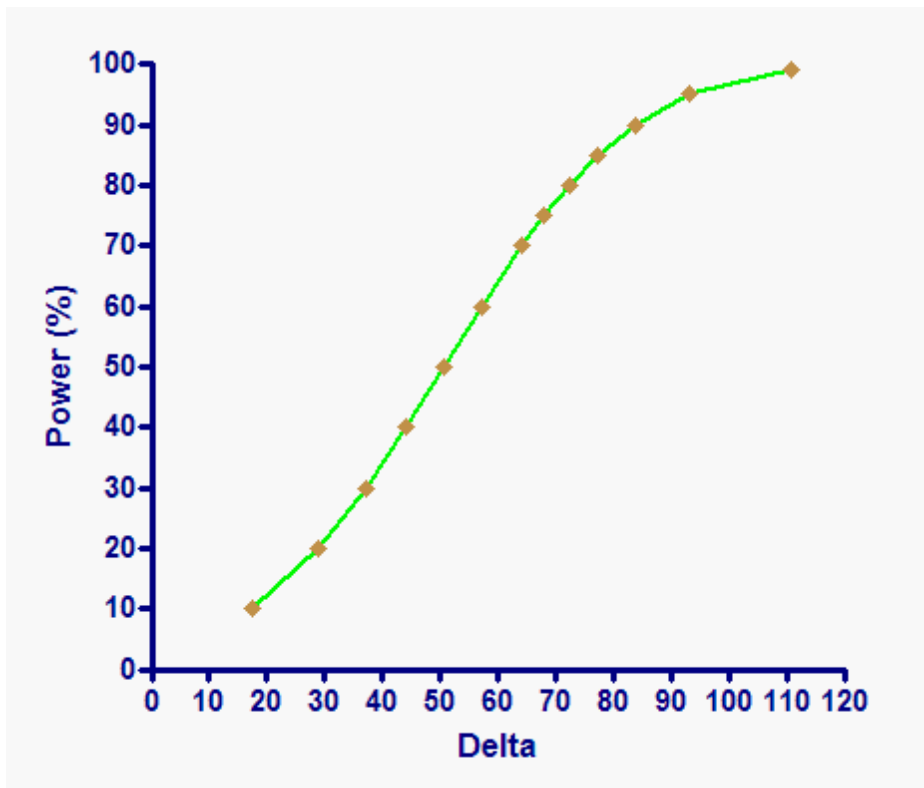
- The CI includes possibilities of a 20% change each way. A 20% change is huge. With such a wide CI, the data are inconclusive. Could be no change. Could be big decrease. Could be big increase.
- The CI tells us that the true difference is unlikely to be more than 20% in each direction. Since we are 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](#)^[61].

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.

1.7.5 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.

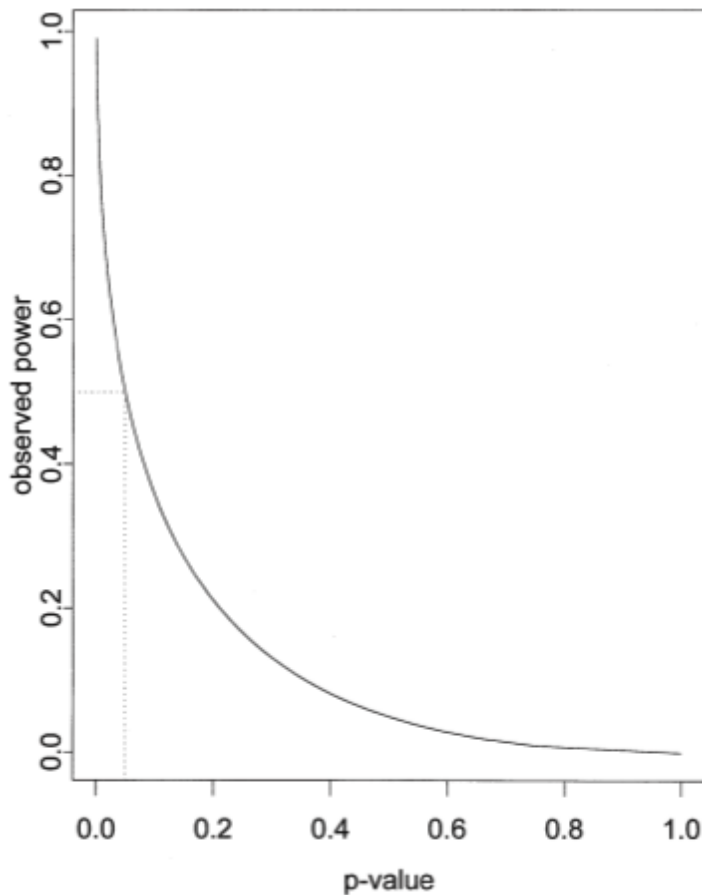
Which effect size should you calculate power for? How large a difference should you be looking for? These are not statistical questions, but rather scientific questions. It only makes sense to do a power analysis when you think about the data scientifically. It makes sense to compute the power of a study 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

Hoening 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.

Hoening 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

1.7.6 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*, ISBN:0521809169.

1.8 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.

1.8.1 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.

Reference

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

1.8.2 Why choose sample size in advance?

The appeal of choosing sample size as you go

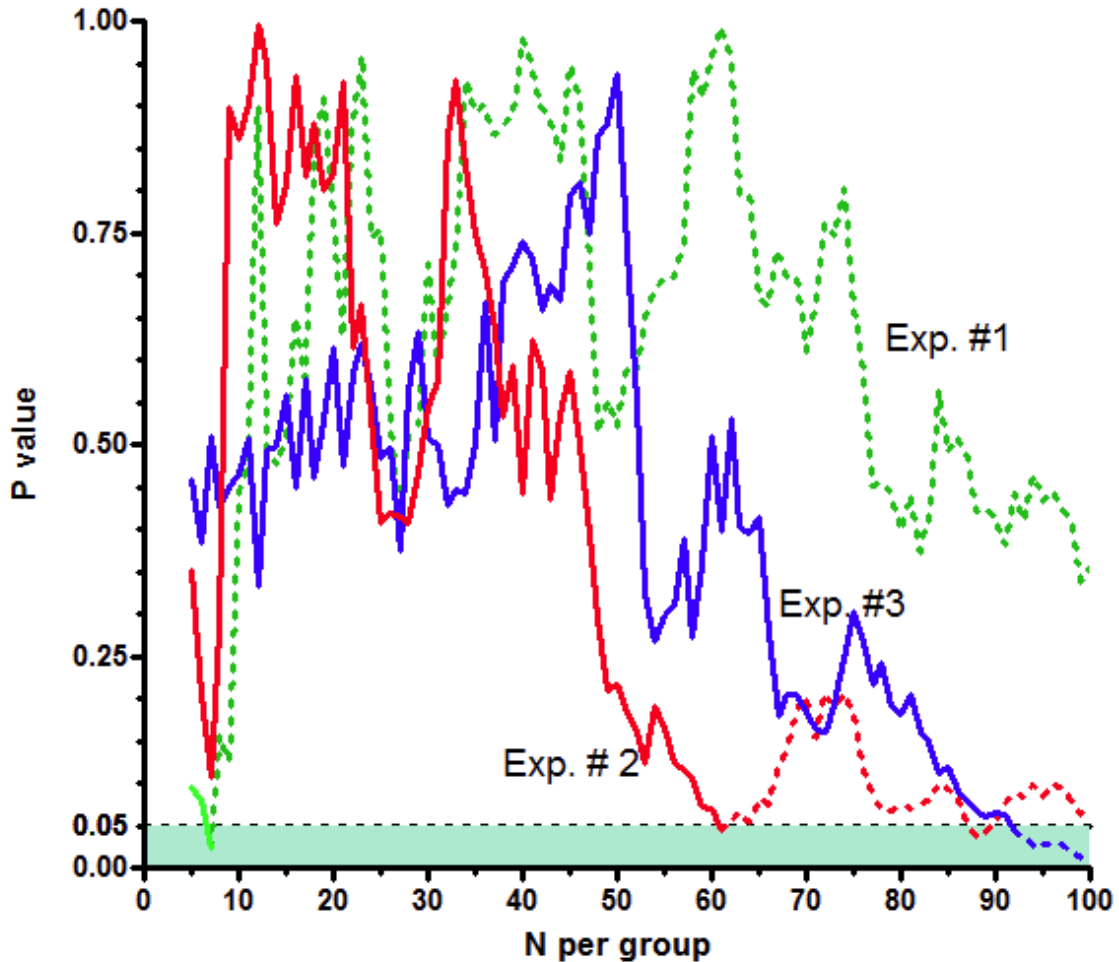
To many, calculating sample size before the study starts seems like a nuisance. Why not do the analyses as you collect data? If your results are not statistically significant, then collect some more data, and reanalyze. If your results are statistically significant result, then stop the study and don't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

1.8.3 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 in 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.

1.8.4 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](#).

1.8.5 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.

1.9 The problem of multiple comparisons

1.9.1 The multiple comparisons problem

Review of the meaning of P value and alpha

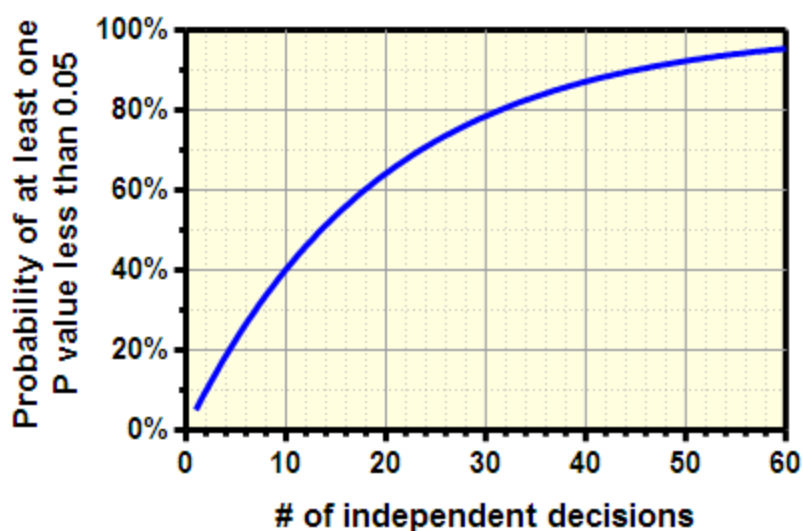
Interpreting an individual P value is straightforward. Consider the simple case of comparing two means. Assuming the null hypothesis is true, the P value is the probability that random subject selection alone would result in a difference in sample means (or a correlation or an association...) at least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

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](#), 2nd edition. Oxford University Press. ISBN=978-0-19-973006-3.

1.9.2 Approach 1: Don't correct for multiple comparisons

Multiple comparisons can be accounted for with [Bonferroni and other corrections](#)^[75], or by the approach of controlling the [False Discover Rate](#)^[76]. But these approaches are not always needed. Here are three situations where 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](#)^[271] follows this approach.

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

Other statisticians recommend not doing any formal corrections for multiple comparisons when the study focuses on only a few scientifically sensible comparisons, rather than every possible comparison. The term [planned comparison](#)^[82] is used to describe this situation. These comparisons must be designed into the experiment, and cannot be decided upon after inspecting the data.

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

1.9.3 Approach 2: Correct for multiple comparisons

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.

	"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 misled 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](#)^[266], [Tukey, Dunnett](#)^[269], [Dunn](#)^[274], [Holm](#)^[270] (and more) all use this approach.

1.9.4 Approach 3: False Discovery Rate (FDR)

Here again is the table from the [previous page](#)^[75] predicting the results from many comparisons.

	"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

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 row 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.

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 "statistically significant", what is the chance that the null hypothesis is true?

If you are only making a single comparison, you can't answer this without defining the prior odds and using [Bayesian reasoning](#)^[51]. But if you have many comparisons, simple methods let you answer that question approximately. In the table, above the False Discovery rate is the ratio $A/(A+C)$. This ratio is sometimes called Q . 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 90% of the discoveries to truly reflect actual differences, while 10% to be false positives. In other words, you expect $A/(A+C)$ to equal 0.10 (the value you set for Q).

Prism uses the concept of False Discovery Rate as part of our method to define outliers ([from a stack of values](#)^[102], or [during nonlinear regression](#)). Prism also can use the FDR method when calculating [many t tests at once](#)^[231].

1.9.5 Lingo: Multiple comparisons

The terminology is not always used consistently.

The term *multiple comparison test* applies whenever you make several comparisons at once. The term *post test* is often used interchangeably.

If you decide which comparisons you want to make *after* looking at the data, those comparisons are called *post hoc tests*.

If you focus on a few scientifically sensible comparisons chosen in advance, those are called "[planned comparisons](#)"^[84]. These choices must be based on the scientific questions you are asking, and must be chosen when you design the experiment. Some statisticians argue that you don't need to correct for multiple comparisons when you narrow down your goals to make only a few planned comparisons.

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.

[The difference between "planned comparisons", "post-hoc tests", "multiple comparison tests", "post tests", and "orthogonal comparisons"](#).

1.9.6 Multiple comparisons traps

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

These simple and sensible rules are commonly violated. When scientists don't get the results they want, they often resort to tactics such as:

- Change the definition of the outcome.
- Use a different time scale.
- Try different criteria for including or excluding a subject.
- Arbitrarily decide which points to remove as outliers.
- Try different ways to clump or separate subgroups.
- Try different ways to normalize the data.
- Try different algorithms for computing statistical tests.
- Try different statistical tests.
- If the results are still 'negative', then don't publish them.

If you try hard enough, eventually ‘statistically significant’ findings will emerge from any reasonably complicated data set. You 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.

Sequential Analyses

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, and the method or data source to compare it with, is unambiguous.

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).

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

Multiple comparisons can be interpreted correctly only when all comparisons are planned, and all planned comparisons are published. These simple ideas are violated in many ways in common statistical practice. "If you torture your data long enough, they will tell you whatever you want to hear." (Mills, 1993).

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.

1.9.7 Planned comparisons

What are 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.

The advantage of planned comparisons

By making only a limited number of comparisons, you increase the statistical power of each comparison.

Correct for multiple comparisons?

There are two approaches to analyzing planned comparisons:

- Use the Bonferroni correction for multiple comparisons, but only correct for the number of comparisons that were planned. Don't count other possible comparisons that were not planned, and so not performed. In this case, the significance level (often set to 5%) applies to the family of comparisons, rather than to each individual comparison.
- Set the significance level (or the meaning of the confidence interval) for each individual comparison. The 5% traditional significance level applies to each individual comparisons, rather than the whole family of comparisons as it does for multiple comparisons.

The second approach has more power to detect true differences, but also has a higher chance of falsely declaring a difference to be "significant". In other words, the second approach has a higher chance of making a Type I error but a lower chance of making a Type II error.

What is the logic of not correcting for multiple comparisons? It seems 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.

I don't really understand the logic of that second approach. It makes perfect sense that if you only plan to make two comparisons, the multiple comparisons should only correct for two comparisons and not the many others you could have made. I don't see how it makes sense to get rid of the whole idea of multiple comparisons just because they were preplanned. There is an inherent tradeoff between protecting against Type I errors (declaring differences "statistically significant" when they were in fact just due to a coincidence of random sampling) and Type II errors (declaring a difference "not statistically significant" even when there really is a difference). There is no way to avoid that tradeoff. Creating arbitrary rules just for preplanned comparisons does not seem justified to me.

Include all the groups when computing scatter?

Each comparison is made by dividing the difference between means by the standard error of that difference. Two alternative approaches can be used to compute that standard error:

- Do an ordinary t test, only using the data in the two groups being compared.
- Use the ANOVA results to account for the scatter in all the groups. ANOVA assumes that all the data are sampled from Gaussian populations and that the SD of each of those populations is identical. That latter assumption is called *homoscedasticity*. If that assumption is true, the scatter (variability) from all the groups can be pooled. The Mean Square Residual (also called Mean Square Error) of the ANOVA is used to compute the standard error of the difference.

If the assumption of homoscedasticity is valid, the second approach has more power. The calculation of the standard error is based on more data so is more precise. This shows up in the calculations as more degrees of freedom. But if that assumption is wrong, then pooling the scatter will give you an invalid measure of scatter.

Reference

Design and Analysis: A Researcher's Handbook (4th Edition), Geoffrey Keppel, Thomas D. Wickens, ISBN:0135159415.

1.9.8 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.

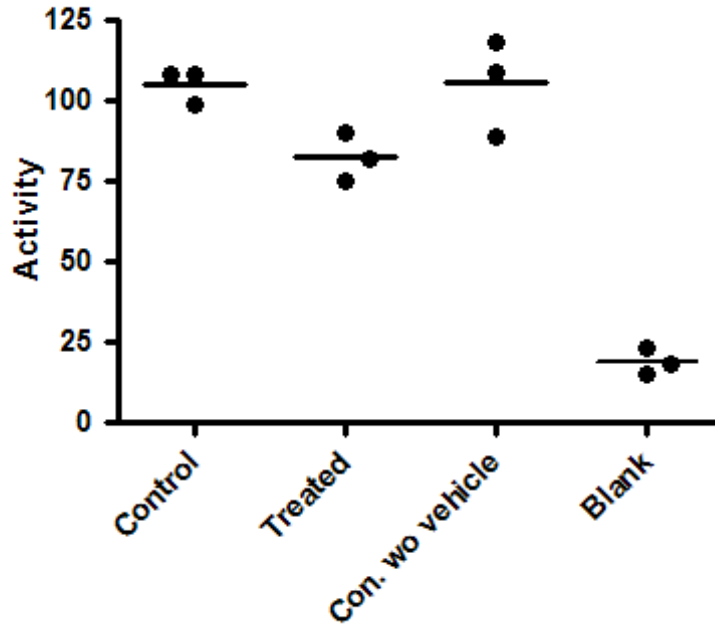
Some statisticians recommend not correcting for multiple comparisons when you make only a few *planned* comparisons. The idea is that you get some bonus power as a reward for having planned a focussed study.

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.

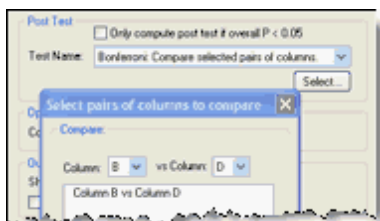
Tukey's multiple comparison tests were used to compare all pairs of means (table above). You only care about the first comparison -- control vs. treated -- which is not statistically significant ($P > 0.05$).

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):

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

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.

1.9.9 The Bonferroni method

How the Bonferroni test works

The Bonferroni method is a simple method for correcting for multiple comparisons. It can be used to correct any set of P values for multiple comparisons, and is not restricted to use as a followup test to ANOVA.

It works like this:

1. Compute a P value for each comparison. Do no corrections for multiple comparisons when you do this calculation.
2. Define the familywise significance threshold. Often this value is kept set to the traditional value of 0.05.
3. Divide the value you chose in step 2 by the number of comparisons you are making in this family of comparisons. If you use the traditional 0.05 definition of significance, and are making 20 comparisons, then the new threshold is $0.05/20$, or 0.0025.
4. Call each comparison "statistically significant" if the P value from step 1 is less than or equal to the value computed in step 3. Otherwise, declare that comparison to not be statistically significant.

The advantages of this method are that it is simple to understand and is very versatile. When you are making only a few comparisons at once, the method works pretty well. If you are making lots of comparisons, the power of this method is low.

Bonferroni tests in Prism

Prism can perform Bonferroni 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](#)^[269] if you want to compute confidence intervals for every comparison or the [Holm-Šidák test](#)^[270] 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-Šidák method, which has more power, but doesn't compute confidence intervals.
- As part of the new analysis to [perform many t tests at once](#)^[231].

1.10 Testing for equivalence

1.10.1 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](#)^[89].

1.10.2 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

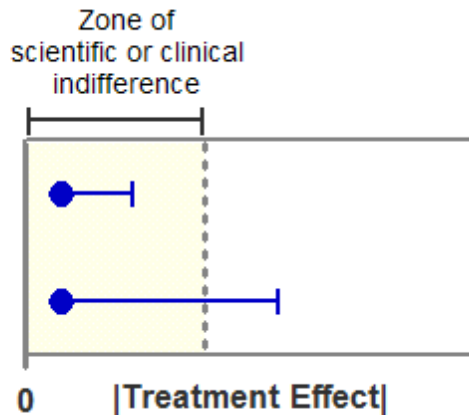
scientific or clinical judgment -- statistical analyses can't help.

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.

1.11 Nonparametric tests

1.11.1 Key concepts: Nonparametric tests

ANOVA, t tests, and many statistical tests assume that you have sampled data from populations that follow a [Gaussian](#)^[18] 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.

1.11.2 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.
- If data deviate substantially from a Gaussian distribution, using a nonparametric test

is not the only alternative. Consider transforming the data to create a Gaussian distribution. Transforming to reciprocals or logarithms are often helpful.

- Data can fail a normality test because of the presence of an outlier. 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, normality tests 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.

1.11.3 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 [power](#)^[56] 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](#)^[210] (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.

1.11.4 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.

1.11.5 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](#)^[318].
- **Small samples.** If you have tiny samples (a few subjects in each group), the nonparametric tests have [little or no power](#)^[93] to find a significant difference.
- **Normality tests [should not be used](#)**^[92] 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 misled.

1.11.6 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*, ISBN:0534919766.

Jean Dickinson Gibbons and Subhabrata Chakraborti, *Nonparametric Statistical Inference*, Fourth Edition: Revised and Expanded (Statistics: A Series of Textbooks and Monographs), ISBN:0824740521.

Myles Hollander and Douglas A. Wolfe, *Nonparametric Statistical Methods*, 2nd Edition, ISBN:0471190454.

Erich L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, ISBN:013997735X.

Bryan F.J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Second Edition, ISBN:0412721309.

Larry Wasserman, *All of Nonparametric Statistics* (Springer Texts in Statistics), ISBN:1441920447.

Rand R. Wilcox, *Applying Contemporary Statistical Techniques*, ISBN:0127515410.

1.12 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](#)¹⁶⁹, or [while fitting a curve with nonlinear](#)

[regression.](#)

1.12.1 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

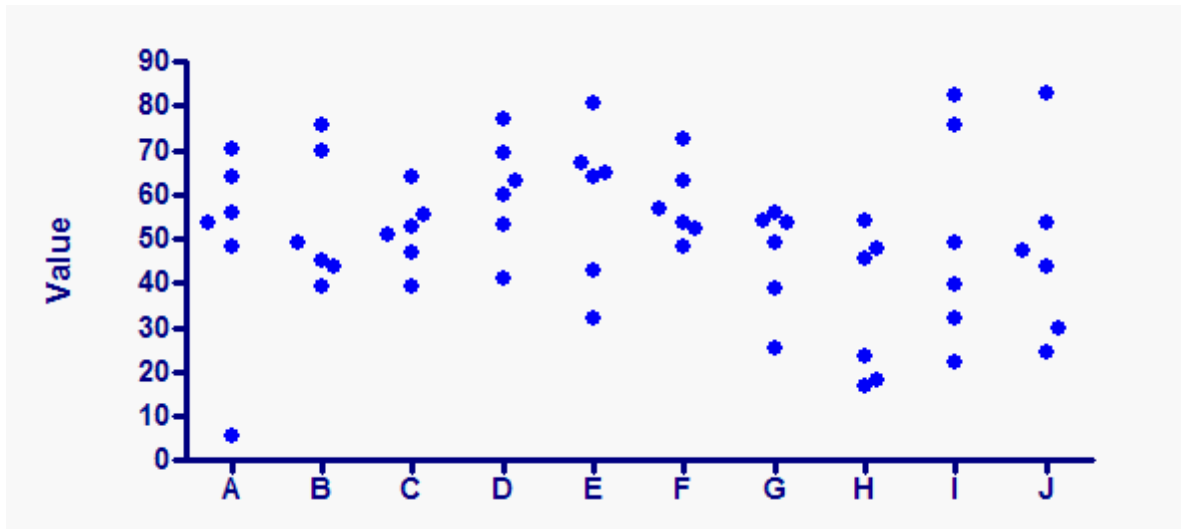
Some statistical tests are designed so that the results are not altered much by the presence of one or a few outliers. Such tests are said to be *robust*. When you use a robust method, there is less reason to want to exclude outliers.

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).

1.12.2 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.



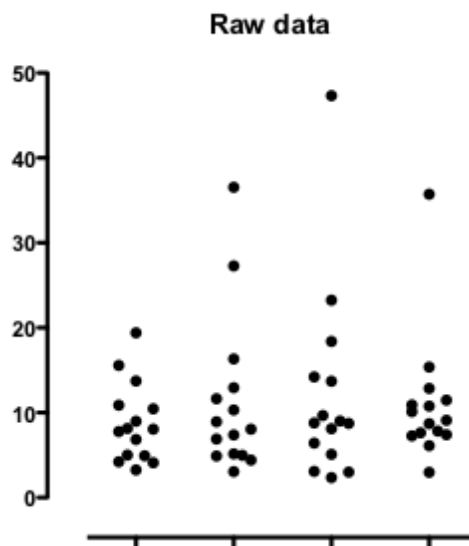
The graph above was created via simulation. The values in all ten data sets are randomly sampled from a Gaussian distribution with a mean of 50 and a SD of 15. But most people would conclude that the lowest value in data set A is an outlier. Maybe also the high value in data set J. Most people are unable to appreciate random variation, and tend to find 'outliers' too often.

1.12.3 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.

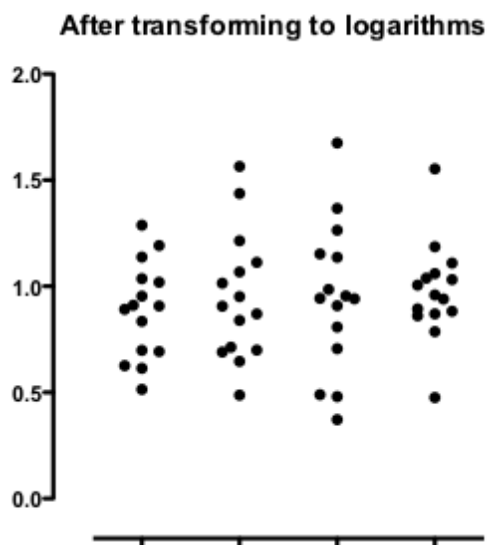
But what if the underlying distribution is not Gaussian? Then the outlier tests are misleading. A common situation is sampling from a lognormal distribution.

The graph below shows four data sets sampled from lognormal distributions.



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:



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.

1.12.4 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 analyses.

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.

$$Z = \frac{|\text{mean} - \text{value}|}{\text{SD}}$$

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.

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.

1.12.5 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 arbitrary X values to each row, and then fit to the model $Y = X \cdot o + M$.]

This method can detect any number of outliers (up to 30% of the sample size).

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](#)¹⁰⁵ 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 Grubbs' 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:

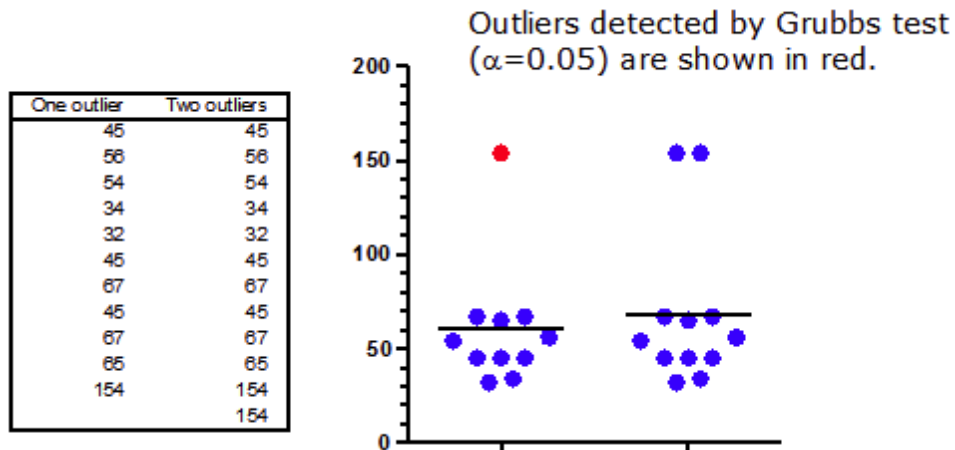
- Grubbs' is slightly better than the ROUT method for the task it was designed for: Detecting a single outlier from a Gaussian distribution.
- The ROUT method is much better than the iterative Grubbs' test at detecting two outliers in some situations.

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>.

1.12.6 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.



([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](#)^[101] 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

1.12.7 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				FDR
				0	1	2	>2	
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](#)^[104], 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 outliers) 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](#)

1.13 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](#)  109

[Paired t test](#)  111

[Mann-Whitney test](#)  114

[Wilcoxon matched pairs test](#)  115

[One-way ANOVA](#)  116

[Repeated measures one-way ANOVA](#)  118

[Kruskal-Wallis test](#)  120

[Friedman's test](#)  121

[Two-way ANOVA](#)  122



[Repeated measures two-way ANOVA](#)  124

[Contingency tables](#)  125

[Survival analysis](#)  126

[Outliers](#)  128

1.13.1 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](#)  179, and [interpreting the results](#)  191.

✓ **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. Prism can perform normality tests as part of the [Column Statistics](#)¹³⁴ analysis. [Learn more](#)¹⁶⁴.

✓ **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](#)²³⁴ 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](#)^[143].

✓ **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](#)^[43], 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 $P > 0.50$.

1.13.2 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](#)^[201].

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](#)^[16] – 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](#)^[72] 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](#)^[43] 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](#)^[206].

1.13.3 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](#)^[201].

✓ **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 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](#)^[16] – that whatever factor caused a ratio (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](#)^[72] 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](#)^[43] 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](#)^[200], not the ratio t test.

1.13.4 Mann-Whitney test

The [Mann-Whitney test](#)^[213] 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](#)^[216].

✓ **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](#)^[16] 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](#)^[213] 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](#)^[216]). 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.](#)^[43]

✓ **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.

1.13.5 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](#)^[179], and [interpreting the results](#)^[226].

✓ **Are the pairs independent?**

The results of a Wilcoxon test only make sense when the pairs are [independent](#)^[16] – 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](#)^[72] 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](#)^[43], 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.

1.13.6 One-way ANOVA

One-way ANOVA compares the means of three or more unmatched groups. Read elsewhere to learn about [choosing a test](#)^[237], and [interpreting the results](#)^[246].

✓ **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, 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.

Prism tests for equality of variance with Bartlett's test. The P value from this test answers 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 as those observed in your experiment. A small P value suggests that the variances are different.

Don't base your conclusion solely on Bartlett's 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 Bartlett's test (unless the P value is really tiny) and interpret the ANOVA results as usual. Some statisticians recommend ignoring Bartlett's test 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](#)^[284], 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.

1.13.7 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](#)^[237], and [interpreting the results](#)^[256].

✓ 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.

✓ 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 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, *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.

1.13.8 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](#)^[237], and [interpreting the results](#)^[260].

✓ 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.

1.13.9 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.

1.13.10 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](#)²⁸⁴, and [interpreting the results](#).³⁰⁴

✓ 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.

1.13.11 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](#)^[284], [graphing the data](#)^[311], and [interpreting the results](#)^[310]^[304]

✓ **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. 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.

✓ 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. You don't have to worry about circularity 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](#).^[291]

1.13.12 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](#)^[323], [sensitivity & specificity](#)^[325], and [interpreting P values](#)^[326].

✓ 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. Neither of these tests is offered by Prism.

✓ 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](#)^[338]. 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](#)^[333] 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.

1.13.13 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](#)^[350], and comparing [two](#)^[351] (or [more than two](#)^[356]) 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.

1.13.14 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](#)^[99]. 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](#)^[104] 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.

2 STATISTICS WITH PRISM 6

This second half of the GraphPad Statistics Guide 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.

If you are already familiar with Prism 5, you may be interested in these sections which are new to Prism 6:

- [How to identify outliers](#)^[169] in a stack of values.
- [Multiple t tests](#)^[231] (one per row)
- The [Holm-Šidák approach](#)^[270] to multiple comparisons
- [Fisher's Least Significance Difference \(LSD\) test](#)^[271]
- [Multiplicity adjusted P values](#)^[275] ("exact" P values for multiple comparisons tests)

2.1 Getting started with statistics with Prism

Enter topic text here.

2.1.1 What's new in Prism 6 (statistics)?

Most important statistical improvements in Prism 6

- **"Exact" P values after ANOVA.** Your #1 request has been for Prism to report "exact P values" for multiple comparisons made after one- or two-way ANOVA. Following Bonferroni, Tukey or Dunnett multiple comparisons testing, Prism 6 now can compute a *multiplicity adjusted P value*. For each comparison, this is the smallest significance level (applied to the entire family of comparisons) where this particular comparison would just barely be declared to be "statistically significant". An adjusted P value is an "exact P value" reported for each comparison, but its value depends on the number of comparisons.
- **Fisher's LSD.** Prism also offers a second approach to report "exact P values" following ANOVA, the unprotected [Fisher's unprotected LSD](#)^[271]. This method

does not correct for multiple comparisons.

- **Main and simple effects.** Prism 6 offers many more choices for multiple comparisons after two-way ANOVA, including testing for [main and simple effects](#).
- **Ratio t test.** The paired t test works by analyzing the difference between each pair of values, testing the null hypothesis that the average difference is zero. With some kinds of data, the difference between before and after is not a consistent measure of effect. The differences might be larger when the control values are larger. The ratio (after/before) may be a much more consistent way to quantify the effect of the treatment. Actually, it turns out that analyzing the logarithm of ratios works much better. Prism 6 makes it simple to do the [ratio t test](#) 206.
- **Kolmogorov-Smirnov test.** Like the Mann-Whitney (MW) test, the [Kolmogorov-Smirnov \(KS\) test](#), is a nonparametric method to compare two groups. The KS test works by comparing the two cumulative frequency distributions, and so has more power to detect subtle differences in the two distributions. In contrast, the MW test is better at detecting changes in the median. The use of the KS test has become standard in some scientific fields. Don't confuse this test with the the other version of the KS test used for normality testing.

Complete list of changes

View the complete list of changes in [multiple comparisons](#), and the other changes in [statistics](#).

2.1.2 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 sample data set

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](#)

Simulating data and Monte Carlo analyses

Prism can plot and analyze simulated data, as well as data you enter.

[Simulating a data table](#)

[Using a script to simulate many data sets](#)

[Key concepts: Monte Carlo analyses](#)

[Monte Carlo example: Accuracy of confidence intervals](#)

[Monte Carlo example: Power of unpaired t test](#)

2.1.3 Guided examples: Statistical analyses

Guided examples

These examples will guide you through most of Prism's statistical analyses.

Descriptive statistics

[Column statistics](#)  134

[Frequency distribution](#)  149

Compare two groups

[Unpaired t test from raw data](#)  188

[Paired t test](#)  200

[Mann-Whitney test](#)  210

[Wilcoxon matched pairs test](#)  224

Categorical outcomes

[Contingency table analysis](#)^[319]

[Survival analysis](#)^[342]

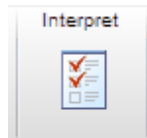
Diagnostic lab tests

[ROC curve](#)^[376]

[Bland-Altman plot](#)^[383]

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](#)^[109]

[Analysis checklist: Survival analysis](#)^[126]

[Analysis checklist: Repeated measures two-way ANOVA](#)^[124]

2.2 Descriptive statistics and frequency distributions

What can statistics help you say about a stack of numbers?

A lot! Quantify the center of the distribution and its scatter. Plot a frequency distribution. Test whether the mean (or median) differs significantly from a hypothetical value.

2.2.1 Column 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

[How to: Column statistics](#)¹³⁴

[Analysis checklist: Column statistics](#)¹³⁶

[Interpreting results: Mean, geometric mean and median](#)¹³⁸

[Interpreting results: Quartiles and the interquartile range](#)¹³⁹

[Interpreting results: SD, SEM, variance and coefficient of variation \(CV\)](#)¹⁴¹

[Interpreting results: Skewness and kurtosis](#)¹⁴²

[Interpreting results: One-sample t test](#)¹⁴³

[Interpreting results: Wilcoxon signed rank test](#)¹⁴⁴

[Interpreting results: Normality tests](#)¹⁴⁷

2.2.1.1 How to: Column statistics

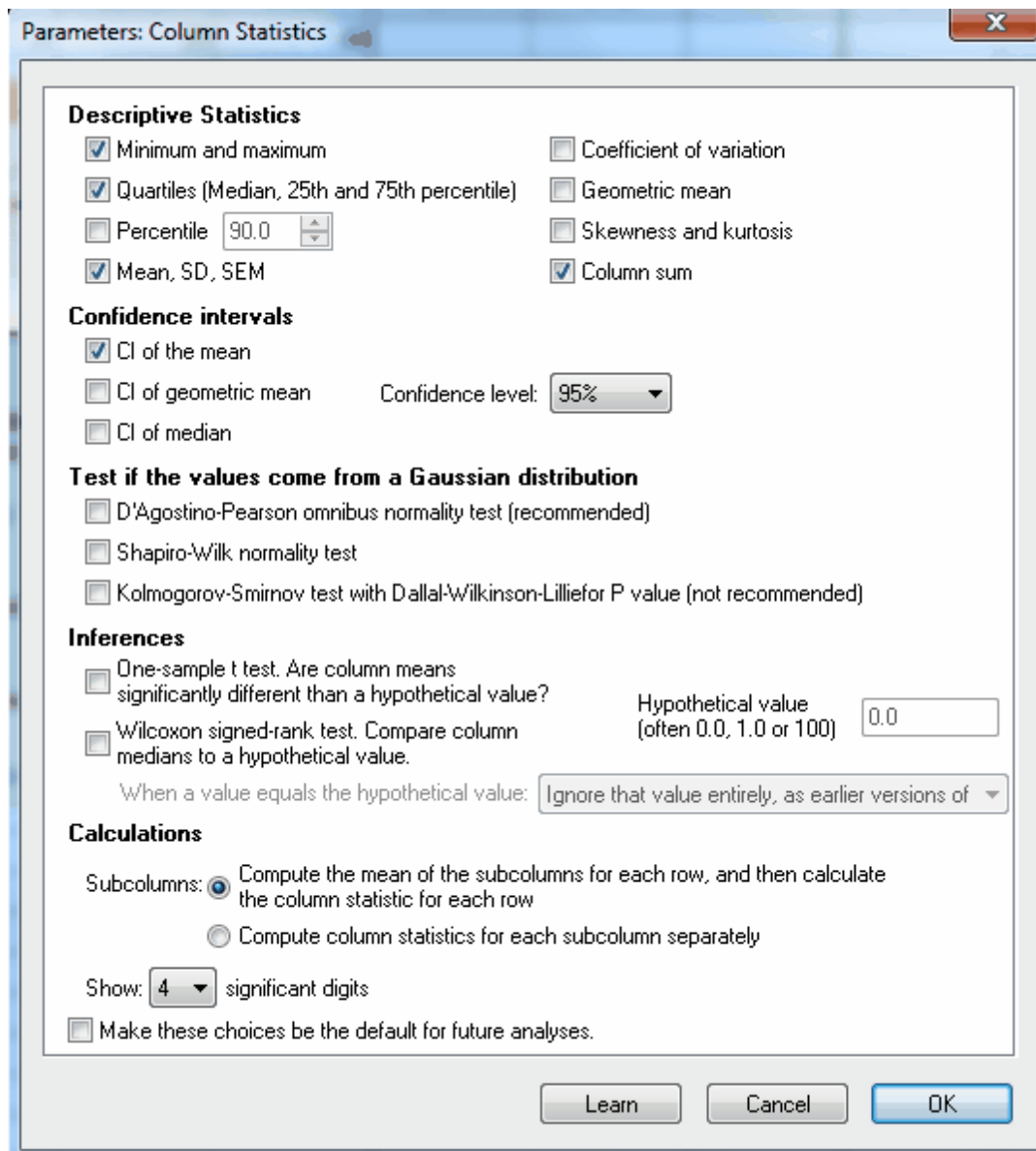
1. Entering data for column statistics

Column statistics 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 column statistics analysis from data entered onto XY or Grouped data tables.

2. Choose the column statistics analysis

Click  and choose Column statistics from the list of analyses for column data.



Prism's column statistics analysis computes descriptive statistics of each data set, tests for normality, and tests whether the mean of a column is different than a hypothetical value.

3. Choose analysis options

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.

Descriptive statistics

Learn more about [quartiles](#)^[139], [median](#)^[138], [SD](#)^[23], [SEM](#)^[28], [confidence interval](#)^[32], [coefficient of variation](#)^[141], [geometric mean](#)^[138], [skewness and kurtosis](#)^[142].

Test if the values come from a Gaussian distribution

One-way ANOVA and t tests depend on the assumption that your data are sampled from populations that follow a Gaussian distribution. Prism offers three tests for normality. We suggest using the D'Agostino and Pearson test. The Kolmogorov-Smirnov test is not recommended, and the Shapiro-Wilk test is only accurate when no two values have the same value. [Learn more about testing for normality.](#)^[164]

Inferences

If you have a theoretical reason for expecting the data to be sampled from a population with a specified mean, choose the [one-sample t test](#)^[143] to test that assumption. Or choose the nonparametric [Wilcoxon signed-rank test](#)^[144].

2.2.1.2 Analysis checklist: Column statistics

Descriptive statistics

Value	Meaning
Minimum	The smallest value.
25th percentile ^[139]	25% of values are lower than this.
Median ^[138]	Half the values are lower; half are higher.
75th percentile ^[139]	75% of values are higher than this.
Maximum	The largest value.
Mean ^[138]	The average.
Standard Deviation ^[23]	Quantifies variability or scatter.
Standard Error of Mean ^[28]	Quantifies how precisely the mean is known.
95% confidence interval ^[32]	Given some assumptions, there is a 95% chance that this range includes the true overall mean.
Coefficient of variation ^[141]	The standard deviation divided by the mean.
Geometric mean ^[138]	Compute the logarithm of all values, compute the mean of the logarithms, and then take the antilog. It is a better measure of central tendency when data follow a lognormal distribution (long tail).

Value[Skewness](#)^[142]**Meaning**

Quantifies how symmetrical the distribution is. A distribution that is symmetrical has a skewness of 0.

[Kurtosis](#)^[142]

Quantifies whether the shape of the data distribution matches the Gaussian distribution. A Gaussian distribution has a kurtosis of 0.

Normality tests

[Normality tests](#)^[166] are performed for each column of data. Each normality test reports a P value that answers this question:

If you randomly sample from a Gaussian population, what is the probability of obtaining a sample that deviates from a Gaussian distribution as much (or more so) as this sample does?

A small P value is evidence that your data was sampled from a nongaussian distribution. A large P value means that your data are consistent with a Gaussian distribution (but certainly does not prove that the distribution is Gaussian).

Normality tests are less useful than some people guess. With small samples, the normality tests don't have much power to detect nongaussian distributions. Prism won't even try to compute a normality test with fewer than seven values. 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.

Normality tests can help you decide when to use nonparametric tests, but the decision [should not be an automatic one](#)^[92].

Inferences

A [one-sample t test](#)^[143] compares the mean of a each 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 small](#)^[46] (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.

The [nonparametric Wilcoxon signed-rank test](#)^[144] is similar, but does not assume a Gaussian distribution. It asks whether the median of each column differs from a hypothetical median you entered.

2.2.1.3 Interpreting results: Mean, geometric mean and median

Mean and median

Mean

The mean is the average. Add up the values, and divide by the number of values.

Median

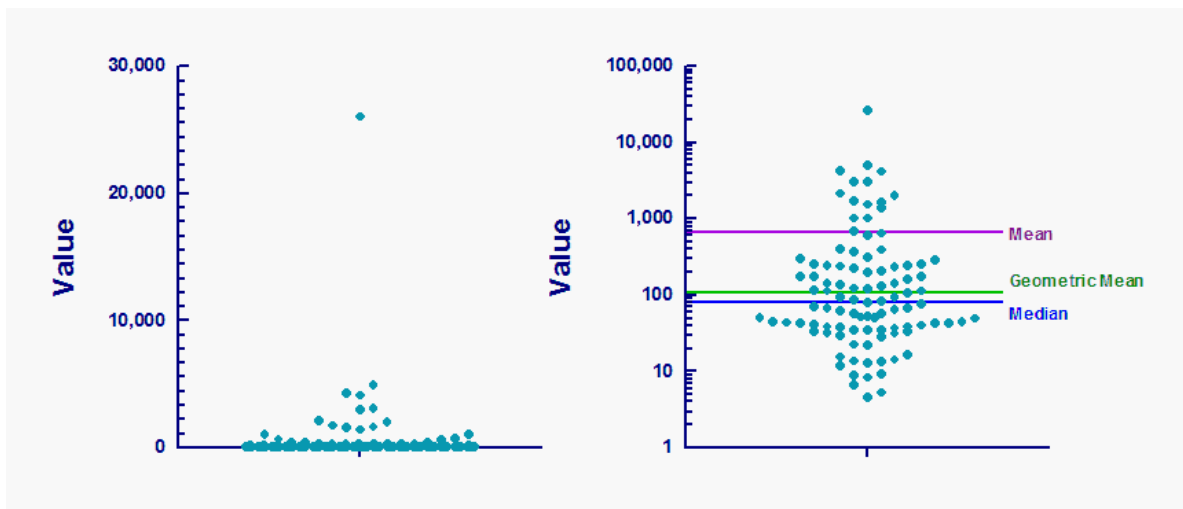
The median is the 50th percentile. Half the values are higher than the median, and half are lower.

Geometric mean

Compute the logarithm of all values, compute the mean of the logarithms, and then take the antilog. It is a better measure of central tendency when data follow a lognormal distribution (long tail).

Example

If your data are sampled from a Gaussian distribution, the mean, geometric mean and median all have similar values. But if the distribution is skewed, the values can differ a lot as this graph shows:



The graph shows one hundred values sampled from a population that follows a lognormal distribution. The left panel plots the data on a linear (ordinary) axis. Most of the data points are piled up at the bottom of the graph, where you can't really see them. The right panel plots the data with a logarithmic scale on the Y axis. On a log axis, the distribution appears symmetrical. The median and geometric mean are near the center of the data cluster (on a log scale) but the mean is much higher, being pulled up by some very large values.

Why is there no 'geometric median'? you would compute such a value by converting all the data to logarithms, find their median, and then take the antilog of that median. The result

would be identical to the median of the actual data, since the median works by finding percentiles (ranks) and not by manipulating the raw data.

Other ways to assess 'central tendency'

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.

Harmonic mean

To compute the harmonic mean, first transform all the values to their reciprocals. Then take the mean of those reciprocals. 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 not often used in biology, and is not computed by Prism.

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.

2.2.1.4 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. 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

2.2.1.5 Interpreting results: SD, SEM, variance and coefficient of variation (CV)

Standard Deviation

The [standard deviation](#)^[23] (SD) quantifies variability. It is expressed in the same units as the data.

Standard Error 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](#)^[29] and [when to use each](#)^[30].

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.

Coefficient of variation (CV)

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.

It only makes sense to report CV for a variable, 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. Therefore, it can make sense to express variation in 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 zero, 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.

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.

2.2.1.6 Interpreting results: Skewness and kurtosis

Interpreting skewness and kurtosis

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.

Kurtosis quantifies whether the shape of the data distribution matches the Gaussian distribution.

- A Gaussian distribution has a kurtosis of 0.
- A flatter distribution has a negative kurtosis,
- A distribution more peaked than a Gaussian distribution has a positive kurtosis.
- Kurtosis has no units.
- 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 skewness is computed

Skewness has been defined in multiple ways. The steps below explain the method used by Prism, called g_1 (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 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](#)

2.2.1.7 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](#)^[47], 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](#)^[46] (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](#)^[46].

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](#)^[16]. 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).

2.2.1.8 Interpreting results: Wilcoxon signed rank test

The [nonparametric](#)^[92] 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](#)^[223] 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](#)^[46], 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](#)^[47], 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](#)^[16]. 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 offers 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 can be different than from previous versions of Prism

Results from Prism 6 can differ from prior versions because Prism 6 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 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 uses 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 offers 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) [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
4. Pratt, J.W. and Gibbons, J.D. (1981), Concepts of Nonparametric Theory, New York: Springer Verlag.
5. 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

2.2.1.9 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 from it. The normality test tells you nothing about the alternative distributions.

If your 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](#)^[92].

2.2.2 Frequency Distributions

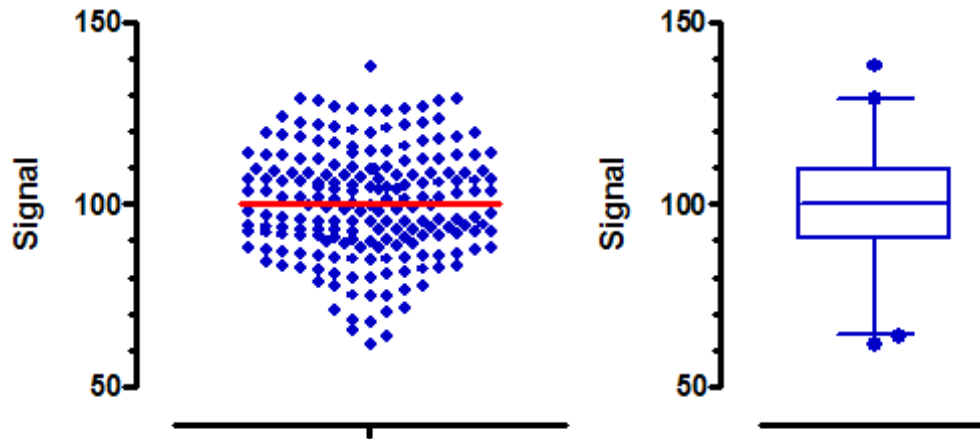
2.2.2.1 Visualizing scatter and testing for normality without a frequency distribution

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 on the right is a box-and-whiskers graph of the same data, showing the values lower than the first percentile and greater than the 99th percentile as circles. Note that Prism offers several choices for how to define the whiskers in this kind of plot.

Both graphs were created by Prism directly from the data table, with no analysis needed.



Testing for normality

Prism can [test for normality](#)^[164] as part of the column statistics analysis. You don't have to create a frequency distribution, and then fit a Gaussian distribution.

2.2.2.2 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

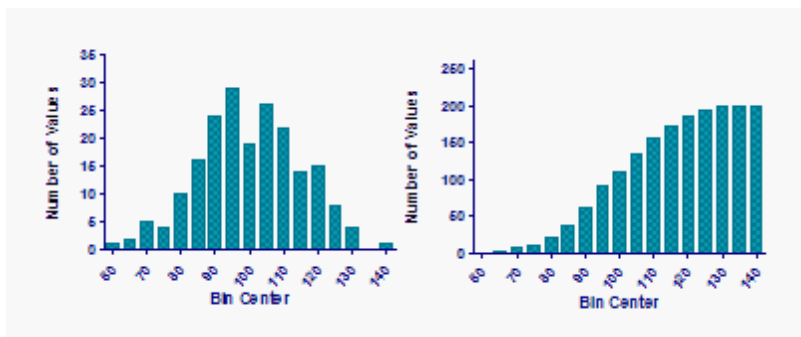
Click Analyze and then choose Frequency distribution from the list of analyses for Column data.



3. Choose analysis options

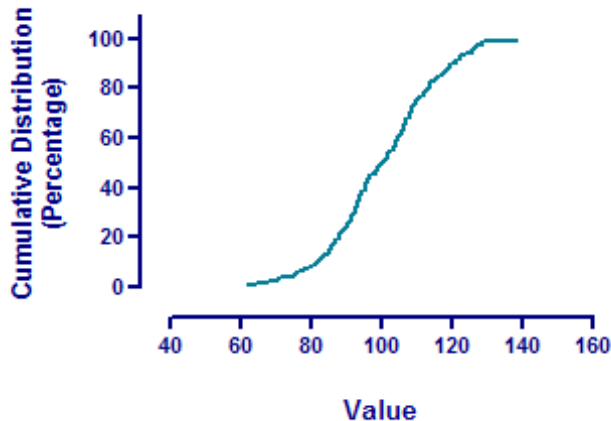
Cumulative?

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

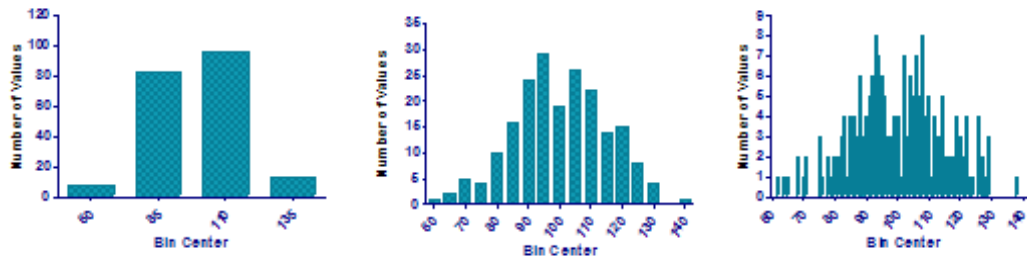
If you chose a cumulative frequency distributions, we suggest that you choose to create an exact distribution. In this case, you don't choose a bin width as each value is plotted individually.

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](#) ¹⁵³.



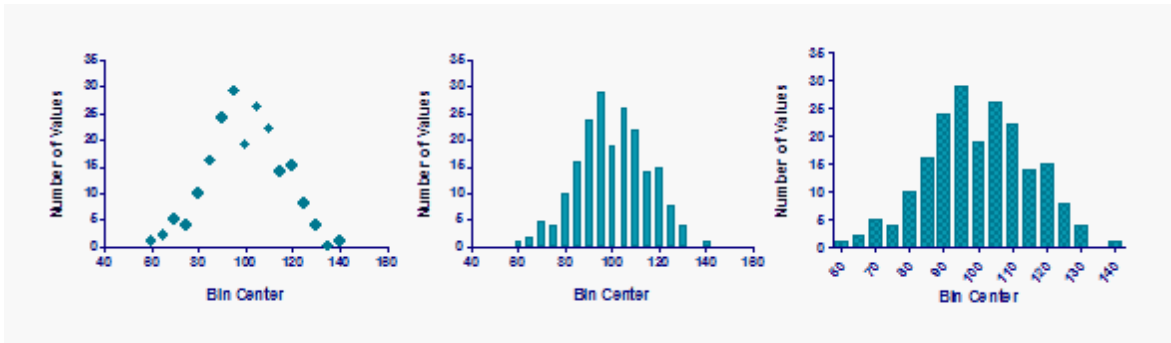
Prism can only make frequency distributions from numerical data. It can handle categorical data, but only if the categories are entered as values.

2.2.2.3 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

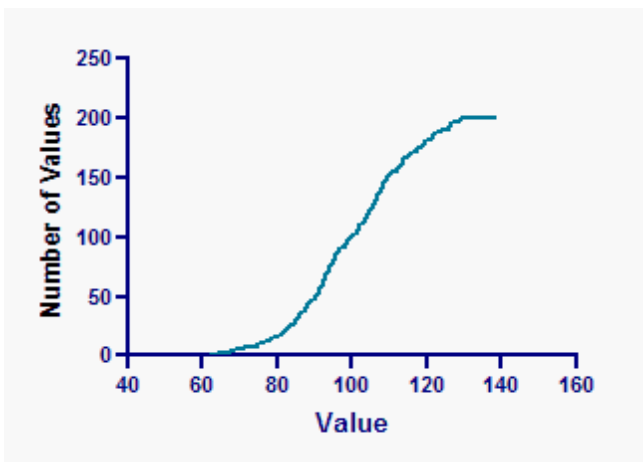
If you don't create a cumulative distribution, Prism gives you three choices illustrated below: XY graph with points, XY graph with spikes (bars). or a bar graph



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](#)^[154] 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.

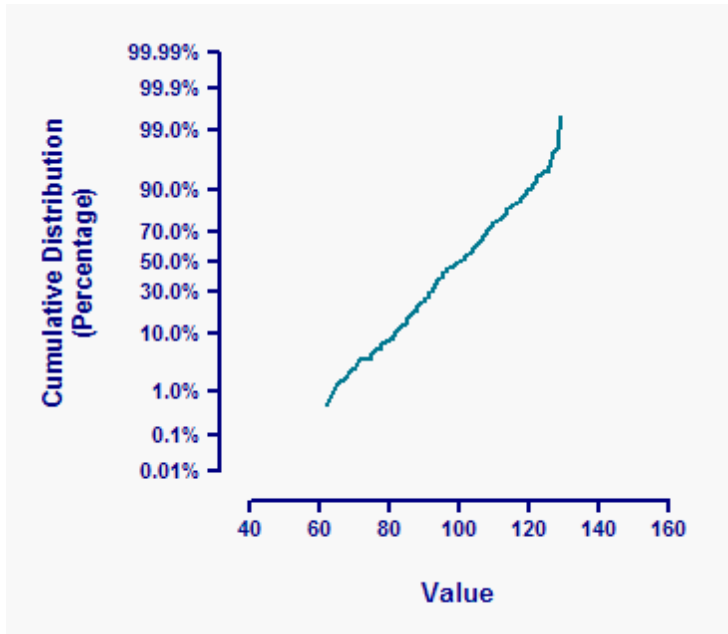
Graphs of cumulative frequency distributions

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.



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.

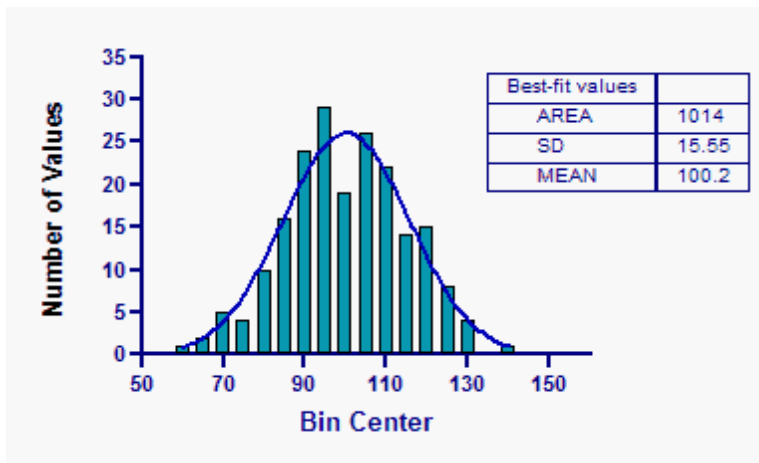
2.2.2.4 Fitting a Gaussian distribution to a frequency distribution

Why fit a Gaussian distribution to your data?

Does your data follow a Gaussian distribution? One way to answer that question is to [perform a normality test](#)^[134] on the raw data. Another approach is to examine the frequency distribution or the cumulative frequency distribution.

Fitting a Gaussian distribution

To fit the frequency distribution, you have to specify that the distribution be plotted as an XY plot, so the bin centers are X values (and not just row labels). Then click Analyze, choose nonlinear regression, and choose the Gaussian family of equations and then the Gaussian model.



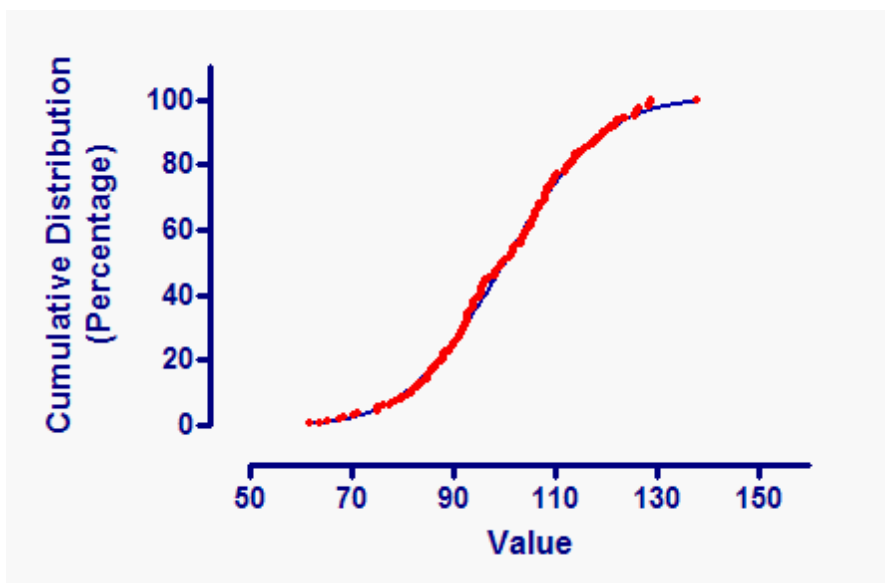
The results depend to some degree on which value you picked for bin width, so we recommend fitting the cumulative distribution as explained below.

Fitting a cumulative Gaussian distribution

The cumulative Gaussian distribution has a sigmoidal shape.

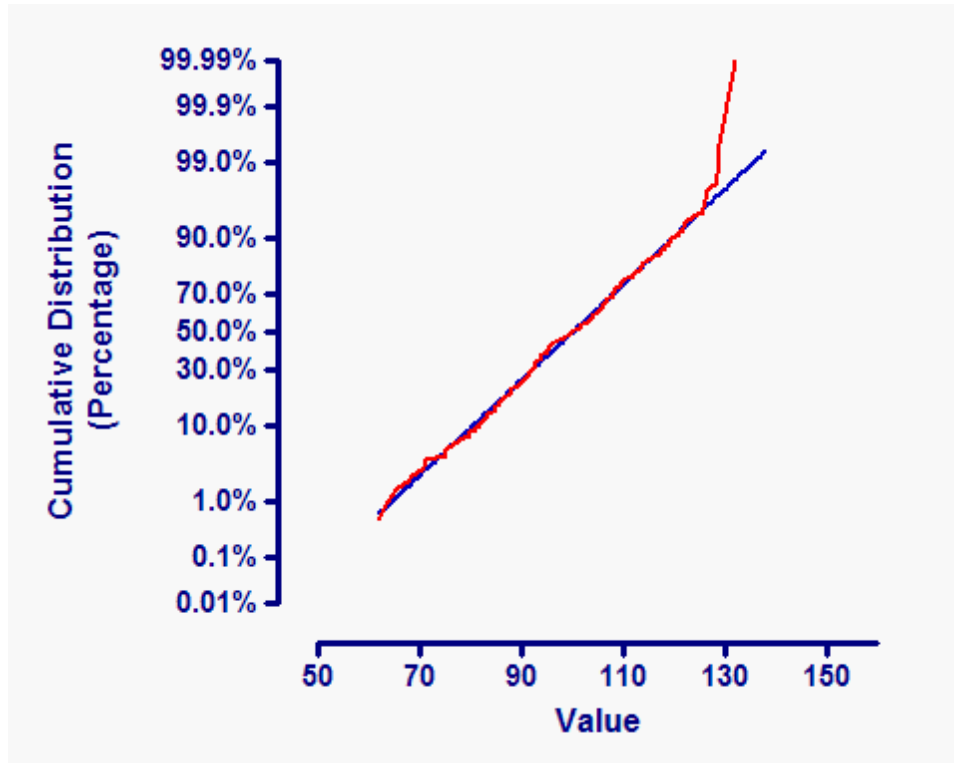
To fit the frequency distribution, you have to specify that the distribution be plotted as an XY plot, so the bin centers are X values (and not just row labels). Then click Analyze, choose nonlinear regression, and choose the one of the cumulative Gaussian models from the selection of Gaussian models. 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.

The graph below shows 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

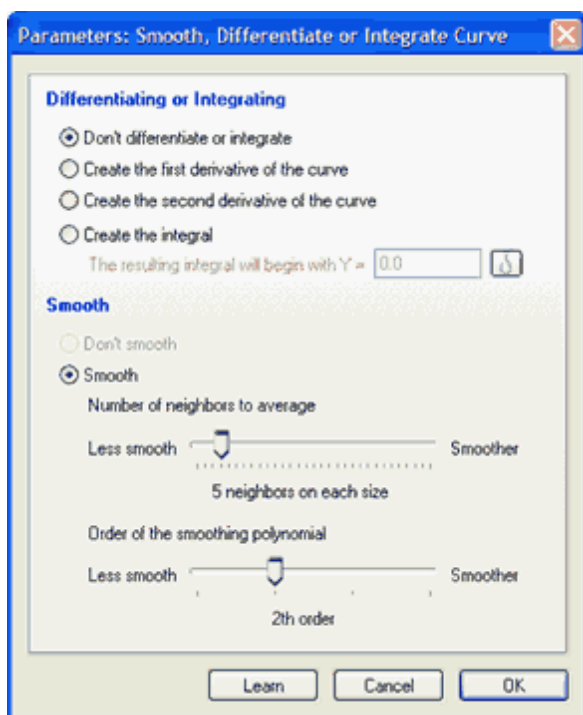
Below, the same graph is plotted using a probability Y axis. To do this, double-click on the Y axis to bring up the Format Axis dialog, drop down the choices for scale in the upper right corner, and choose "Probability (0..100%)". The cumulative Gaussian distribution is linear when plotted on probability 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.



2.2.3 Describing curves

2.2.3.1 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](#)¹⁵⁹.

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

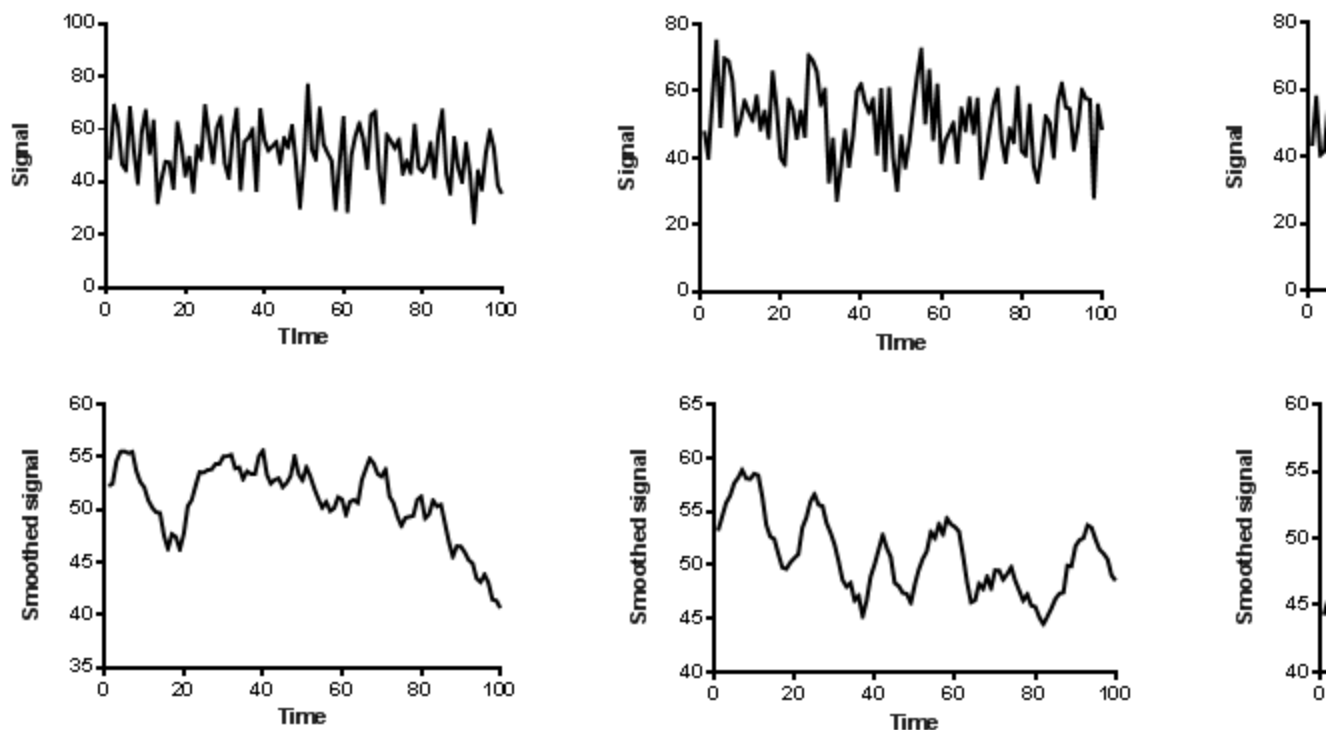
smooth a curve prior to nonlinear regression or other analyses. Smoothing is not a method of data analysis, but is purely a way to create a more attractive graph.

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).

$$x'[i] = (x[i+1] + x[i]) / 2$$

$$y' \text{ at } x'[i] = (y[i+1] - y[i]) / (x[i+1] - x[i])$$

- 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](#)^[159] 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 Savitsky and Golay (1).
- If you request that Prism 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

2.2.3.2 Area under the curve

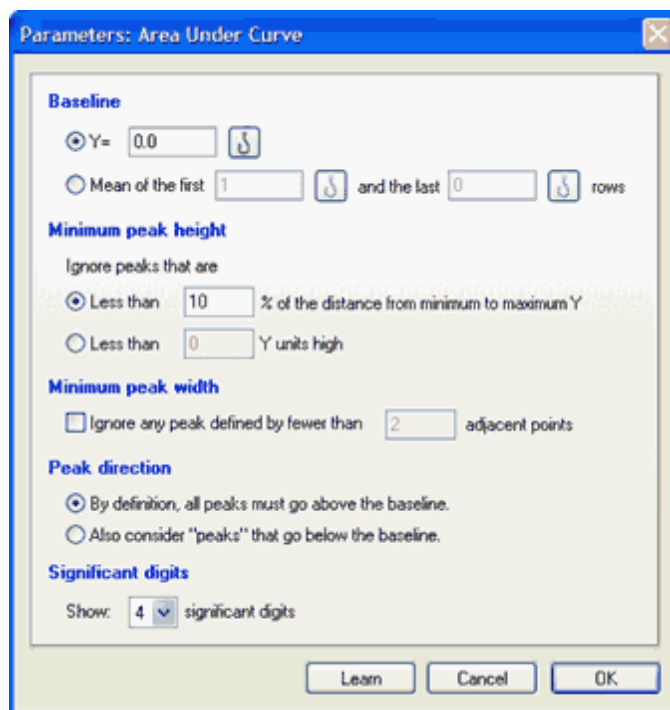
How to: Area under the curve

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](#)^[375].

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

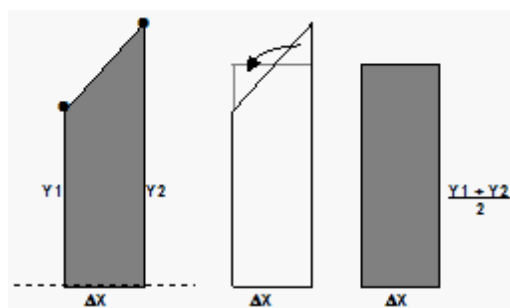
Note these limitations:

- 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.
- If you enter data with replicate Y values, or as Mean and SD or SEM, Prism only analyzes the mean values.
- Prism does not combine the SD values to come up with a confidence interval for the AUC or a SE for the AUC. These calculations have been described by Gagnon (1), but Prism does not yet do them.
- 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 $\Delta X \cdot (Y_1 + Y_2) / 2$. Prism uses this formula repeatedly for each adjacent pair of points defining the curve.

The area is computed between the baseline you specify and the curve, starting from the first X value in your data set and ending at the largest X value. Prism does not extend the

curve beyond your data.

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 always reports the *Total Area*, which includes: 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.

Prism also reports the *Total Peak Area*. Here Prism only includes the peaks you ask it 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. The total area is not a useful value to report, but it puts the results in context and might help you spot problems or better understanding what Prism is or is not including in the total area.

If you ask Prism to define peaks below the baseline as peaks, then Prism subtracts the area of peaks below the baseline from the area of peaks above the baseline, and reports this difference as the *Net Area*.

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.

2.2.4 Row statistics

2.2.4.1 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.

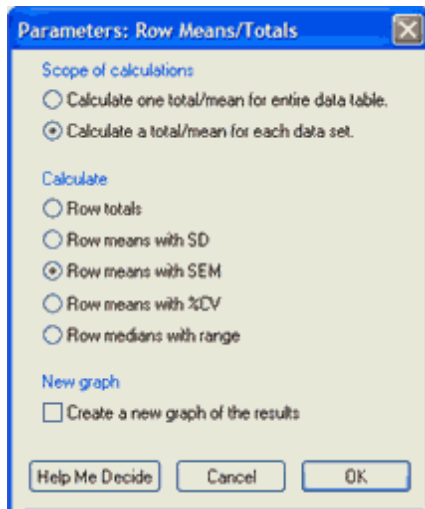
Prism can take your raw data, and create graphs with mean (or median) and error bars

(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](#)^[163] analysis.

2.2.4.2 Row means and totals

If you enter data onto XY or two-way tables with replicate Y values in subcolumns, Prism can automatically create graphs with the mean and SD (or SEM). You don't have to choose any analyses -- 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.

If you 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.



You can choose to compute the [SD](#)^[23], [SEM](#)^[28], or [%CV](#)^[141] for each row each data set individually (what you'll usually want) or for the entire table.

2.3 Normality 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.

2.3.1 How to: Normality test

Analyzing column data

1. Create a Column data table.
2. Enter each data set in a single Y column.
3. Click "Analyze... Statistical analyses... Column statistics"
4. Prism offers three options for testing for normality. Check one, or more than one, of these options.

Analyzing 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 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.

2.3.2 How normality tests work

How the normality tests work

We recommend relying on the **D'Agostino-Pearson** normality test. It first computes the [skewness and kurtosis](#)¹⁴² to quantify how far from Gaussian the distribution is 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 **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.

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 statement¹: *"The Kolmogorov-Smirnov test is only a historical*

curiosity. It should never be used."

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 (2). 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. RB D'Agostino, "Tests for Normal Distribution" in *Goodness-Of-Fit Techniques* edited by RB D'Agostino and MA Stephens, Macel Dekker, 1986.
2. 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.

2.3.3 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 from it. The normality test tells you nothing about the alternative distributions.

If your 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](#)^[92].

2.3.4 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 three 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-Lilliefors 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 7 or more values. The D'Agostino test requires 8 or more values.

▣ **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?**

Prism uses 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 superhigh 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 [should not be automated based on a normality test](#)^[92].

▣ **How should I interpret 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.

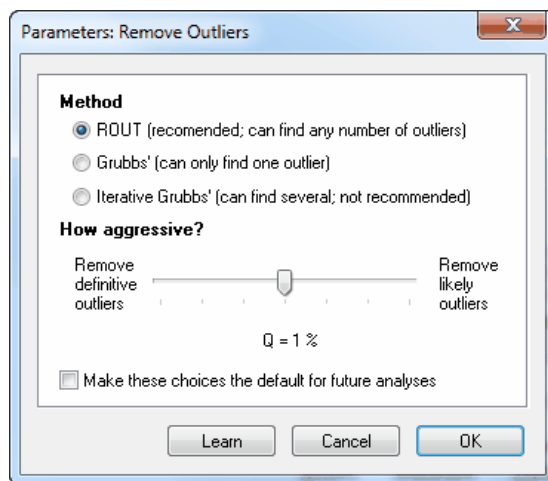
Parts of this page are excerpted from Chapter 24 of Motulsky, H.J. (2010). [Intuitive Biostatistics](#), 2nd edition. Oxford University Press. ISBN=978-0-19-973006-3.

2.4 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.

2.4.1 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.



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](#)^[102] 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](#)^[101] 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.

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](#)^[104].

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.

2.4.2 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](#)^[99]. 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^[104] 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.

2.5 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?

2.5.1 How to: One-sample t test and Wilcoxon signed rank test

The one-sample t test and the Wilcoxon rank sum tests are computed as part of Prism's Column Statistics analysis. Follow these steps

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... Statistical analyses... Column statistics"
4. At the bottom of the Column Statistics dialog, in the section labeled Inferences, check the option to perform either the one-sample t test or the Wilcoxon rank sum test.
5. To the right of that option, 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).

2.5.2 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](#)^[47], 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](#)^[46] (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](#)^[46].

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](#)^[16]. 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).

2.5.3 Interpreting results: Wilcoxon signed rank test

The [nonparametric](#)^[92] 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](#)^[223] 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](#)^[46], 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](#)^[47], 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](#)^[16]. 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 offers 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 can be different than from previous versions of Prism

Results from Prism 6 can differ from prior versions because Prism 6 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 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 uses 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 offers 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) [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
4. Pratt, J.W. and Gibbons, J.D. (1981), Concepts of Nonparametric Theory, New York: Springer Verlag.
5. 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

2.6 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?

2.6.1 Paired or unpaired? Parametric or nonparametric?

2.6.1.1 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](#)¹⁸⁸. The two groups do not have to 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](#)^[200] 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](#)^[189]. 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](#)^[189] 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!](#)^[231]

2.6.1.2 Choosing a test to compare two columns

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](#)^[92], unlike t tests, are not based on the assumption that the data are sampled from a [Gaussian distribution](#)^[18]. But nonparametric tests have [less power](#)^[93]. Deciding when to use a nonparametric test is [not straightforward](#)^[95].

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](#)^[193].

Parametric, paired

Choose the paired t test (which is standard in this situation) or the [ratio t test](#)^[206] (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 6 offers two choices: The Mann-Whitney test (which Prism has always offered) and the [Kolmogorov-Smirnov test](#)^[220] (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.

2.6.1.3 Options for comparing two groups

The second tab of the parameters dialog for t tests and nonparametric 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](#)^[43]. 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

All four options are new to Prism 6, and by default they are not selected. They can be useful to view the data with more depth, but none are essential to beginners.

- 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 ranks. 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 differences. 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 correlation. Graph one variable vs. the other to visually assess how correlated they are.

Additional results

These four choices are all new to Prism 6, and 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. [Calculation details.](#)
- 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 6 offers the option of [using the Pratt method instead](#)²³⁰. If your data has lots of ties, it is worth reading about the two methods and deciding which to use.

2.6.1.4 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 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 work around, 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 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 solution is to use the unequal variance (Welch) t test routinely. As mentioned above, it is not a good idea to first test for unequal standard deviations, and use that result 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 suggests that this is the best thing to do (3). 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 (4). 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. 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
5. P.I. Good and J.W. Hardin, [Common Errors in Statistics: \(and How to Avoid Them\)](#), 2003, ISBN:0471460680.

2.6.1.5 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 [one-way ANOVA](#)^[246], and then follow up with [multiple comparison tests](#)^[72]. 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 [enter data](#)^[188] as mean, SD (or SEM) and N, and Prism can compute an unpaired t test. Prism cannot perform a 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 [not a good idea](#)^[92] 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 [contingency table](#)^[318] and analyze with [Fisher's](#)^[322] 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](#)^[341]. 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?

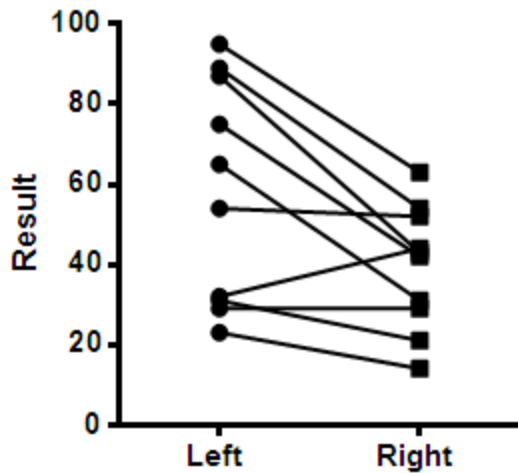
No. The results of any statistical test can only be interpreted at face value when the choice of analysis method was part of the experimental design.

2.6.1.6 The advantage of pairing

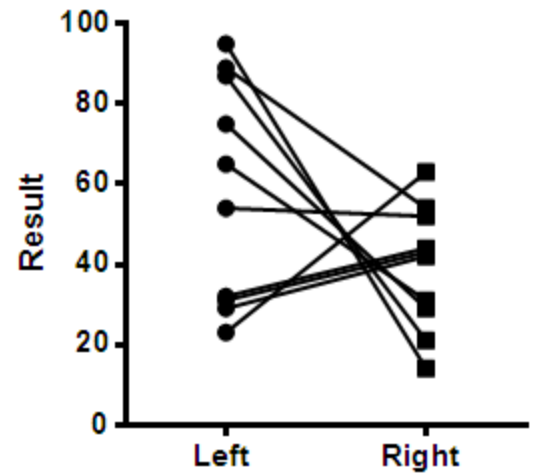
This set of graphs shows the importance of designing experiments where pairing or matching is part of the experimental design, and of accounting for that pairing when analyzing the data.

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 = 39.4
 SEM of differences = 12.2

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 for both analyses. 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.

[Download the Prism file.](#)

2.6.2 Unpaired t test

2.6.2.1 How to: Unpaired t test from raw data

1. Create data table and enter data

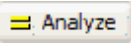
From the Welcome (or New Table and graph) dialog, choose the Column tab.

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  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](#)^[43]. 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](#)^[191] and [graphing](#)^[195] the results.

Before accepting the results, [review the analysis checklist](#)^[109].

2.6.2.2 How to: Unpaired t test from averaged data

1. Enter data

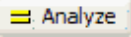
From the Welcome (or New Table and graph) dialog, choose the Column tab.

Choose to enter and plot error values computed elsewhere.

Enter the data all on one row.

Table format:		A			B		
Two-way		Control			Treated		
		Mean	SD	N	Mean	SD	N
1	Title	34.5	11.3	12	46.5	7.3	14

2. Choose the unpaired t test

- From the data table, click  on the toolbar.
- Choose t tests from the list of column analyses.
- 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.
- On the options tab, make these choices:
 - Choose a [one- or two-sided P value](#)⁴³. 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.

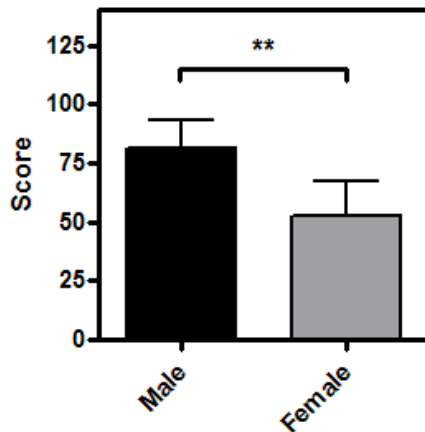
3. Review the results

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 t test](#)¹⁹¹.

Before accepting the results, [review the analysis checklist](#)¹⁰⁹.

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?](#)^[28]).
- To add the [asterisks representing significance level](#)^[50] 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.

2.6.2.3 Interpreting results: Unpaired t

Confidence Interval

The unpaired t test compares the means of two groups. The most useful result is the confidence interval for the difference between the means. If the [assumptions of the analysis are true](#)^[109], you can be 95% sure that the 95% confidence interval contains the true 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.

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.

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](#)”^[49].

You will interpret the results differently depending on whether the P value is [small](#)^[46] or [large](#)^[47].

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.

[What to do when the groups have different standard deviations?](#)^[182]

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.

2.6.2.4 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 assumes 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 have (or 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 6 reports and uses this fractional value for df. Earlier versions of Prism, as well as InStat and our QuickCalc all round the df down to next lower integer (which is common). For this reason, the P value reported by Prism 6 can be a bit smaller than the P values reported by prior versions of Prism.

When to chose the unequal variance (Welch) t test

Deciding when to use the unequal variance t test is not straightforward.

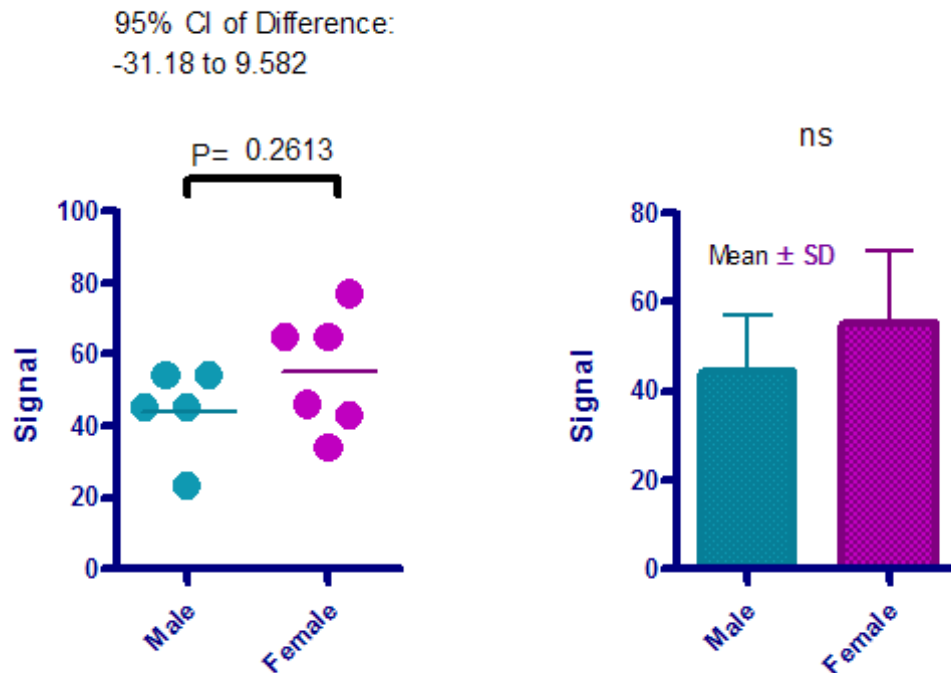
It seems sensible to first test whether the variances are different, and then choose the ordinary or Welch t test accordingly. In fact, [this is not a good plan](#). You should decide to use this test as part of the experimental planning.

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

2.6.2.5 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.

2.6.2.6 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, 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)

Unequal sample sizes

This page was updated 4/16/2010 to point out that the rules of thumb 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).

2.6.2.7 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](#)^[179], and [interpreting the results](#)^[191].

✓ 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. Prism can perform normality tests as part of the [Column Statistics](#)^[134] analysis. [Learn more](#)^[164].

✓ 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](#)^[234] 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](#)^[143].

✓ 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](#)^[43], 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 $P > 0.50$.

2.6.3 Paired or ratio t test

2.6.3.1 How to: Paired t test

1. Enter data

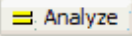
From the Welcome (or New Table and graph) dialog, choose the Column tab, and then a before-after graph.

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:		A	B
One-way		Before	After
	x	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 paired t test

- From the data table, click  Analyze on the toolbar.
- Choose t tests from the list of column analyses.
- 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
- On the options tab, make these choices:
 - Choose a [one- or two-sided P value](#)^[43]. 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.

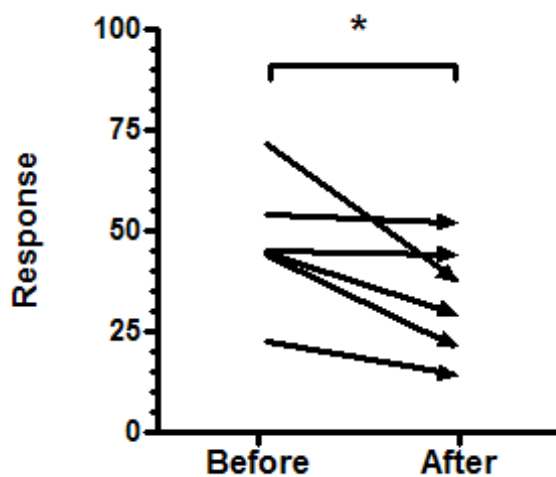
3. Review the results

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](#)^[202].

Before accepting the results, [review the analysis checklist](#)^[111].

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](#)^[50] 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](#)^[205].

2.6.3.2 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](#)^[147].

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.

2.6.3.3 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](#)^[111], 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](#)”^[49].

You will interpret the results differently depending on whether the P value is [small](#)^[46] or [large](#)^[47].

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.

2.6.3.4 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](#)^[201].

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](#)^[16] – 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](#)^[72] 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](#)^[43] 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 $P > 0.50$.

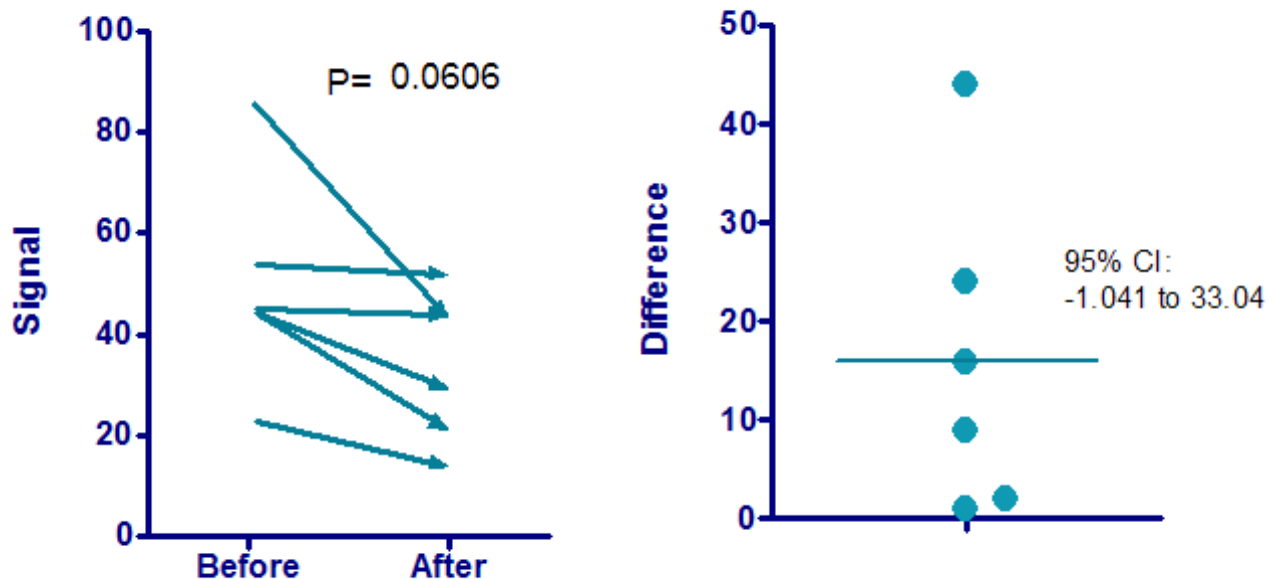
✓ Do you care about differences or ratios?

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](#)²⁰⁶.

2.6.3.5 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.

2.6.3.6 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.

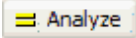
2.6.3.7 How to: Ratio t test

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  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.

2.6.3.8 Interpreting results: Ratio t test

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.

2.6.3.9 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](#)^[201].

✓ 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 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](#)^[16] – that whatever factor caused a ratio (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](#)^[72] 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](#)^[43] 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](#)^[200], not the ratio t test.

2.6.4 Mann-Whitney or Kolmogorov-Smirnov test

2.6.4.1 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](#)^[213] 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](#)^[220] 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.

2.6.4.2 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.

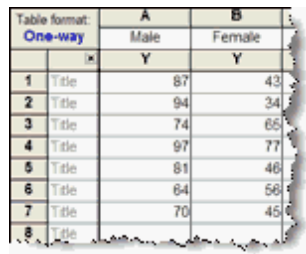
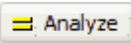
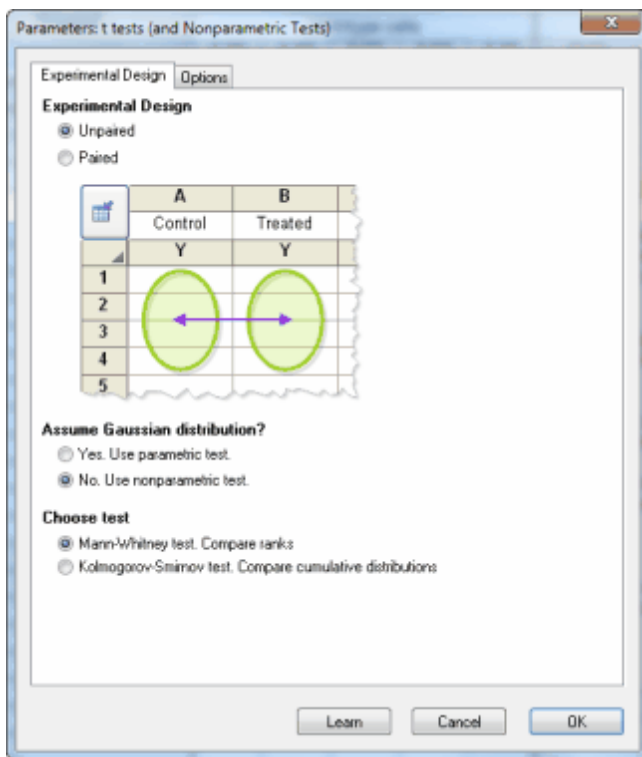


Table format	A	B
One-way	Male	Female
	Y	Y
1 Title	87	43
2 Title	94	34
3 Title	74	65
4 Title	97	77
5 Title	81	46
6 Title	64	56
7 Title	70	45
8 Title		

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](#)^[209].

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.

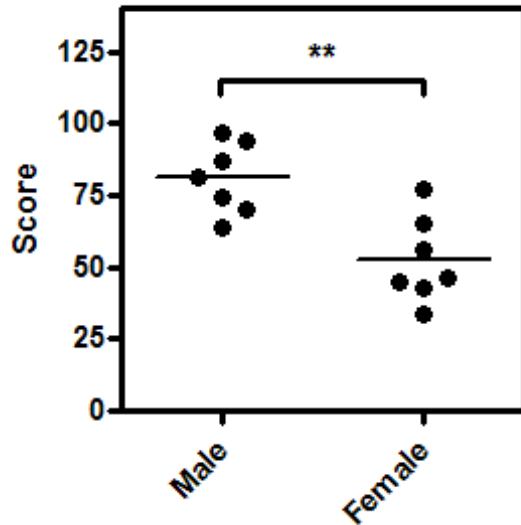
- Choose a [one- or two-tail P value](#)^[43]. If in doubt, choose a two-tail P value.
- If you chose the option below to compute the 95% CI of the difference between medians, specify how Prism will compute that difference (A-B or B-A).
- The MW test works by comparing ranks. Check an option to graph 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](#)^[213].

Before accepting the results, review the [analysis checklist](#)^[114].

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](#)^[50] 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".

2.6.4.3 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 6 computes the exact P value much faster than did prior 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](#)^[93]. 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

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.

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.

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 6 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 can report different results than prior versions

There are two reasons why Prism 6 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>

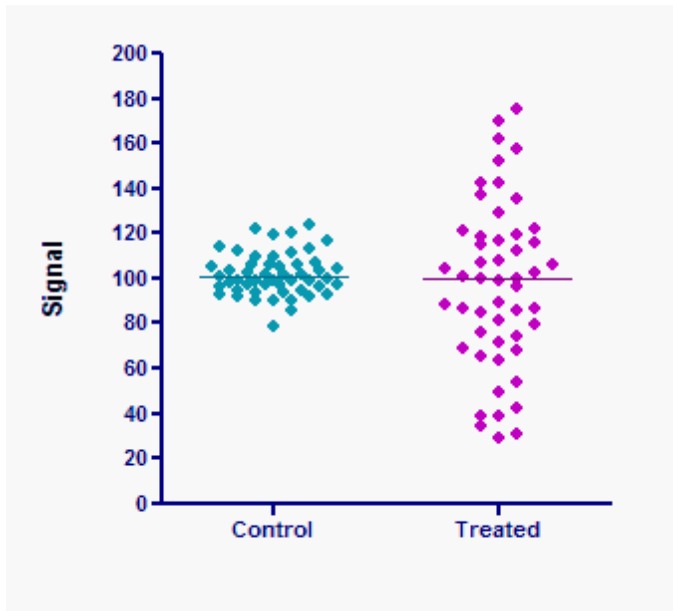
2.6.4.4 The Mann-Whitney test doesn't really compare medians

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-Whitney 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

2.6.4.5 Analysis checklist: Mann-Whitney test

The [Mann-Whitney test](#)^[213] 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](#)^[216].

✓ 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](#)^[16] 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](#)^[213] 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](#)^[216]). 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](#).^[43]

✓ 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.

2.6.4.6 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 previous version 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 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.](#)

2.6.4.7 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 Kolmogorov-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 dataset with the predictions of a Gaussian distribution. Prism [performs this normality test](#)^[164] 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 n_1 values from n_1+n_2 values is less than 60,000, where n_1 and n_2 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

2.6.4.8 Analysis checklist: Kolmogorov-Smirnov test

The [Kolmogorov-Smirnov test](#)^[220] 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](#) 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?

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.

2.6.5 Wilcoxon matched pairs test

2.6.5.1 "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](#)^[224] 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](#)^[144] 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](#)^[213].
- The [Gehan-Wilcoxon test](#)^[351] 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.

2.6.5.2 How to: Wilcoxon matched pairs test

1. Enter data

From the Welcome (or New Table and graph) dialog, choose the one-way 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:		A	B
One-way		Before	After
	x	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  on the toolbar.

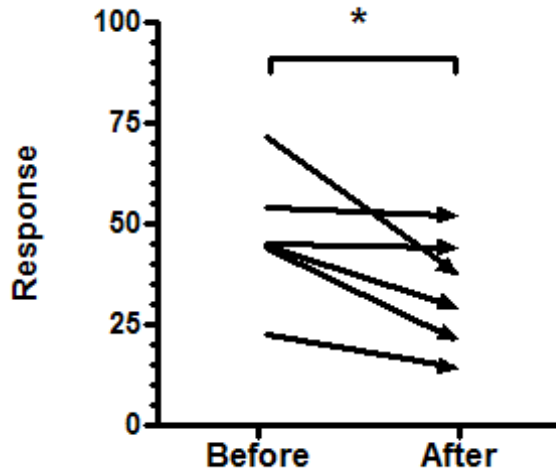
2. Choose t tests from the list of column analyses.
3. On the first (Experimental Design) tab of t test dialog, make these choices:
 - Experimental design: Paired
 - Assume Gaussian distribution: No
 - Choose test: Wilcoxon matched pairs test
4. On the options tab, make these choices:
 - Choose a [one- or two-sided P value](#)^[43]. 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?
 - [Choose how to handle](#)^[230] rows where both values are identical.

3. Review the results

Learn more about interpreting the results of [Wilcoxon's matched pairs test](#)^[226].

Before accepting the results, [review the analysis checklist](#)^[115].

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](#)^[50] 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".

2.6.5.3 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](#)^[144]. Prism 6 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 pair 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 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 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 can differ from prior versions because Prism 6 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 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.

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

2.6.5.4 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](#)^[179], and [interpreting the results](#)^[226].

✓ Are the pairs independent?

The results of a Wilcoxon test only make sense when the pairs are [independent](#)^[16] – 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](#)^[72] 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](#)^[43], 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.

2.6.5.5 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 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. Conover (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

2.6.6 Multiple t tests

2.6.6.1 How to: Multiple t tests

Distinguish the t test analysis from the multiple t test analysis

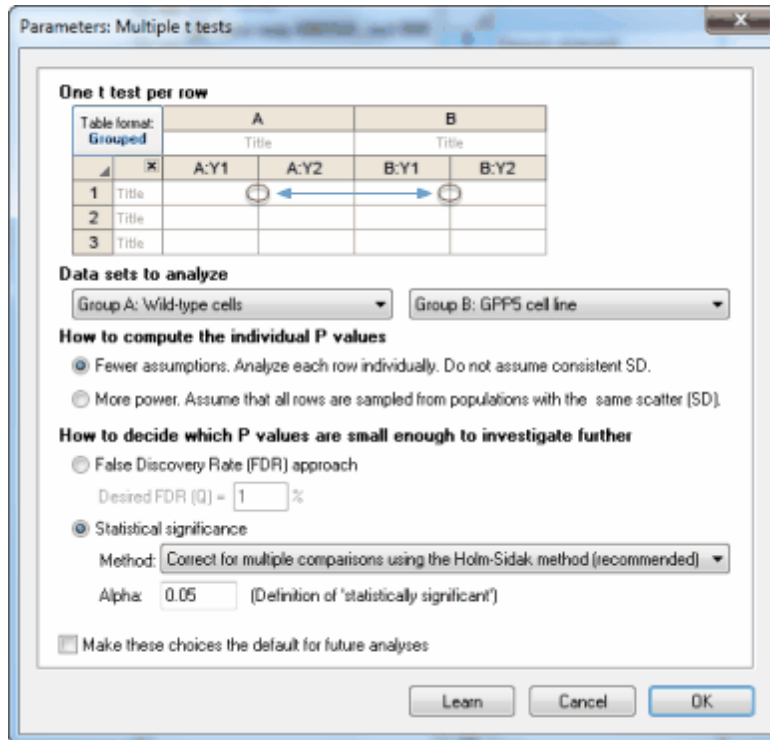
- The t test (and nonparametric) analysis compares two data set columns. Each set of replicate values are usually entered into a column, although Prism can also enter replicates entered into side-by-side subcolumns all on one row.
- The multiple t test analysis (new to Prism 6) performs many t tests at once -- one per row. Replicates are entered into side-by-side subcolumns.

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.](#)

232

2.6.6.2 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 analyzed. 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. 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.

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.

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.](#)^[76] Prism controls the FDR by the method of Benjamini and Hochberg (1).

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.

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-Šidák method](#)^[270] (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 Šidák-Bonferroni method (not recommended). The Bonferroni method is much simpler to understand and is better known than the Holm-Šidák method, but it has no other advantages. The Holm-Šidák method has more power, and we recommend it. Note that if you choose the Bonferroni approach, Prism always uses the [Šidák-Bonferroni method](#)^[266], often just called the Šidák 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.

Reference

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).

2.7 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?

2.7.1 How to: One-way ANOVA

Enter topic text here.

2.7.1.1 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

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 to 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

If the data are matched, so each row represents a different subject of experiment, then you may wish to use row titles to identify each row.

Table format: One-way		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	42	65
3	HM	45	65	99	105
4	JW	54	77	79	90
5	PS	45	46	75	86

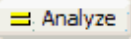
Enter and plot error values calculated elsewhere

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.

Enter the data all on one row. Because there is only one row, the data really only has one grouping variable even though entered on a grouped table.

Table format: Two-way		A			B			C		
		Control			Low dose			High dose		
		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

Run the ANOVA

1. From the data table, click  on the toolbar.
2. Choose one-way ANOVA from the list of column analyses.
3. [Choose the test](#)^[237] you want to perform on the first tab.
4. Choose the multiple comparisons tests on the [Multiple Comparisons](#)^[239] and [Options](#)^[241] tabs of the one-way ANOVA dialog.

2.7.1.2 Which multiple comparisons tests does Prism offer?

Tests that assume sampling from a Gaussian distribution

The choices for multiple comparisons that Prism makes available to you depends on two questions:

- Your goal. Which comparisons do you want to make? Answer this question, based on your experimental goals, on the [multiple comparisons tab](#)²³⁹ 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](#)²⁴¹ of the one-way ANOVA dialog.

Goal	Report CI as well as significance?	Method
Compare every mean to every other mean	Yes	Tukey (preferred) Bonferroni-Dunn Sidak-Bonferroni
	No	Holm-Sidak (preferred) Newman-Keuls
Compare every mean to a control mean	Yes	Dunnett
	No	Holm
Compare selected means	Yes	Bonferroni-Dunn Sidak-Bonferroni
	No	Holm-Sidak

Nonparametric tests

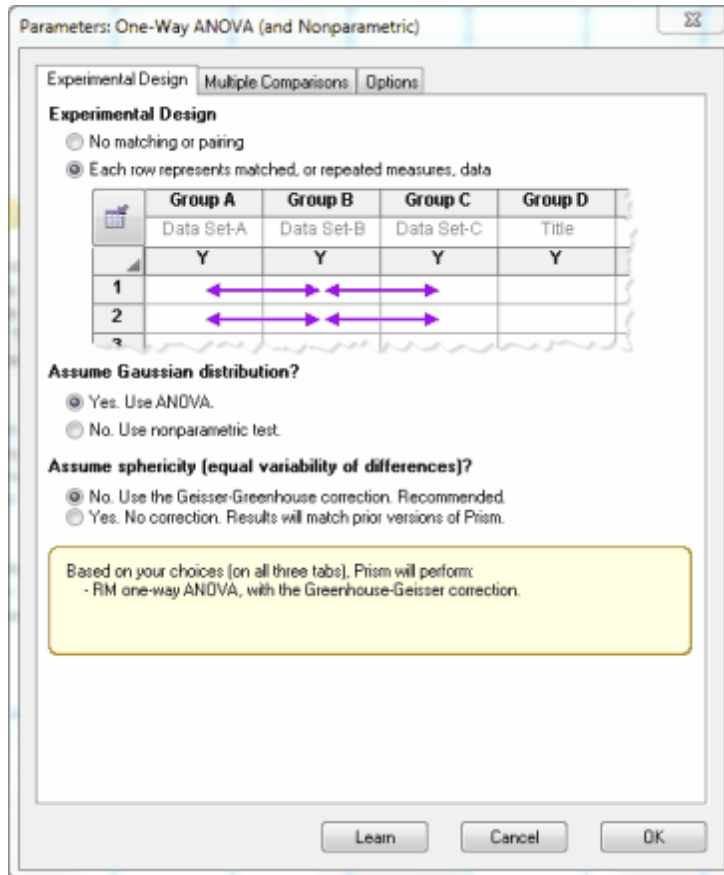
If you choose nonparametric ANOVA, Prism only offers Dunn's post test (1), either to compare all pairs of groups, or just selected pairs. Some books and programs don't use Dunn's name, but simply refer to this test as the post test following a Kruskal-Wallis test, and don't give it an exact name.

Reference

1. O.J. Dunn, *Technometrics*, 5:241-252, 1964

2.7.1.3 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*.

Assume Gaussian distribution?

[Nonparametric tests](#)^[92], unlike ANOVA are not based on the assumption that the data are sampled from a [Gaussian distribution](#)^[18]. But nonparametric tests have [less power](#)^[93], and report only P values but not confidence intervals. Deciding when to use a nonparametric test is [not straightforward](#)^[95].

Assume sphericity?

The concept of sphericity

The [concept of sphericity](#)^[251] 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](#)^[255].
- [Compute multiple comparisons tests differently.](#)^[274]

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
Ordinary one-way ANOVA ^[234]	No	No
Repeated measures one-way ANOVA ^[251]	Yes	No
Kruskal-Wallis test ^[260]	No	Yes
Friedman test ^[263]	Yes	Yes

2.7.1.4 Multiple comparisons tab: One-way ANOVA

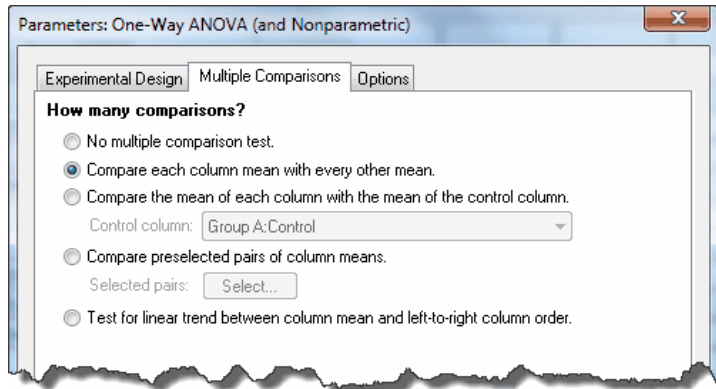
Overview of multiple comparisons choices

For both one- and two-way ANOVA, the 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 \(Options\)](#)^[241] 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?



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 than the other choices, it will have less power to detect differences.

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.

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](#)⁸⁴.

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](#)²⁷² 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, it cannot test for other trends (quadratic, etc.).

2.7.1.5 Options tab: One-way ANOVA

Choose a multiple comparisons test

This tab offers three choices for approaching multiple comparisons.

Correct for multiple comparisons: Confidence intervals and significance

We recommend this choice for two reasons:

- Confidence intervals are much easier for most to interpret than statements about statistical significance.
- Prism will be able to also compute [multiplicity adjusted P values](#)^[275]. It cannot do so for the methods that compute significance without confidence intervals.

If you choose a nonparametric test or choose to compare each mean against a control mean, you have no further choices.

If you chose to compare every column mean with every other column mean and did not choose a nonparametric comparison, Prism offers two methods. The Tukey test is designed for this purpose, and we recommend it. The alternative is to use the Bonferroni method. Its only advantage is that it is easier to understand, but it has less power.

Correct for multiple comparisons: Significance without confidence intervals.

The advantage of choosing a method that does not compute confidence intervals is that you can get a bit more power.

Comparing every column mean with every other column mean

Prism offers two choices. We recommend that you choose the [Holm-Šidák test](#)^[270] and avoid the Newman-Keuls test.

The Holm-Šidák method, which does not compute confidence intervals, is more powerful than the Tukey method, which does compute confidence intervals (1). That means that with some data sets, the Holm-Šidák method can find a statistically significant difference

where the Tukey method cannot.

The problem with the Newman-Keuls test is that 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. [Duncan's new multiple range test](#) is a modification of the Newman-Keuls test with more power. Prism does not offer it, because few, if any, statisticians recommend it.

Comparing every column mean with the mean of a control column

If you are comparing each column mean to a control mean, Prism offers no choices. It will calculate the Holm test. 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(1).

Don't correct for multiple comparisons. Each comparison stands alone.

An alternative to correcting for multiple comparisons is to perform each comparison individually. This is also called [Fisher's Least Significant Difference \(LSD\) test](#)^[271]. If you use this approach, you must correct for multiple comparisons informally while interpreting the results.

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.

We recommend avoiding the Fisher's LSD approach, unless you have a very good reason to use it.

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](#)^[275]. 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

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.

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.

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.

2.7.1.6 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](#)^[234] 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](#)^[246], and then follow up with multiple comparison post tests. An [exception](#)^[84] 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](#)^[92]. 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 outcome has two possibilities, and I know the fraction of each possible outcome in each group. How can I compare the groups?**

Not with ANOVA. Enter your data into a [contingency table](#)^[318] and analyze with a [chi-square test](#)^[319].

- ❑ **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](#)²⁵¹. 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.](#)

2.7.2 One-way ANOVA

2.7.2.1 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. 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" Bartlett'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, Bartlett'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 a modified ANOVA that does not assume that all standard deviations are equal. Prism does not provide such a test.
- 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 η^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.

2.7.2.2 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](#)^[237], and [interpreting the results](#)^[246].

✓ 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, 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.

Prism tests for equality of variance with Bartlett's test. The P value from this test answers 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 as those observed in your experiment. A small P value suggests that the variances are different.

Don't base your conclusion solely on Bartlett's 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 Bartlett's test (unless the P value is really tiny) and interpret the ANOVA results as usual. Some statisticians recommend ignoring Bartlett's

test 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](#)²⁸⁴, 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.

2.7.3 Repeated-measures one-way ANOVA

2.7.3.1 What is repeated measures?

The difference between ordinary and repeated measures ANOVA, is similar to the difference between unpaired and paired t tests. The term *repeated measures* means that you give treatments repeatedly to each subject.

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.

2.7.3.2 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 [matching rather than repeated measurements](#)^[251], then you can assume sphericity, as violations are essentially impossible.
- If your experiment design is repeated measures, 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.

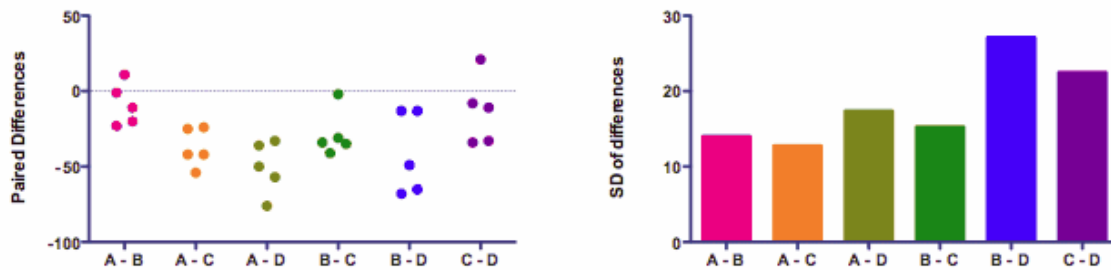
Here is the table of sample data from Prism (choose a Column table, and then choose sample data for repeated measures one-way ANOVA).

		A	B	C	D
		Control	Treatment 1	Treatment 2	Treatment 3
	Y	Y	Y	Y	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 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).

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](#) .

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. It lowers the values of degrees of freedom, and thus 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.

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. ISBN:0805837183.
2. Andy Field, [A Bluffer's Guide to ... Sphericity](#).

2.7.3.3 Quantifying violations of sphericity with epsilon

Deviations from [sphericity](#)^[251] 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. ISBN:0805837183.

2.7.3.4 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 consider using ordinary ANOVA rather than repeated-measures ANOVA.

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 ϵ_{255} , which is a measure of how badly the data violate the assumption of sphericity.

R²

Prism reports two different R² 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 [multiple comparisons tests after repeated measures ANOVA](#)^[274].

Before interpreting the results, [review the analysis checklist](#)^[118].

(1) G Keppel and TD Wickens, [Design and Analysis](#), Fourth Edition, 2004, ISBN: 0135159415

2.7.3.5 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](#)^[237], and [interpreting the results](#)^[256].

✓ 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.

✓ **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.

2.7.4 Kruskal-Wallis test

2.7.4.1 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. Prism labels the results 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).

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 handles ties differently than did prior versions. Prism 6 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 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

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, [review the analysis checklist](#)^[120].

2.7.4.2 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 [choosing a test](#)^[237], and [interpreting the results](#)^[260].

✓ 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.

2.7.5 Friedman's test

2.7.5.1 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.

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 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 6 will report different results than prior 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.

2.7.5.2 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.

2.8 Multiple comparisons after ANOVA

Interpreting multiple comparisons after ANOVA is tricky.

It is important to know exactly what statistical significance means in this situation.

2.8.1 The problem of multiple comparison

The problem of multiple comparisons was discussed in the Principles of Statistics section above.

[The multiple comparisons problem](#)^[72]

[Approach 1: Don't correct for multiple comparisons](#)^[74]

[Approach 2: Correct for multiple comparisons](#)^[75]

[Approach 3: False Discovery Rate \(FDR\)](#)^[76]

[Lingo: Multiple comparisons](#)^[77]

[Multiple comparisons traps](#)^[78]

[Planned comparisons](#)^[82]

[Example: Planned comparisons](#)^[84]

[The Bonferroni method](#)^[87]

2.8.2 Bonferroni and Sidak methods

Key facts about the Bonferroni and Šídák methods

- The Bonferroni and Šídák (the letter Š is pronounced "Sh") methods are very general so can be used whenever you are doing multiple comparisons.
- 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](#)^[269] when comparing every mean with every other mean, and use [Dunnett's method](#)^[269] 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.
- If you don't need or want a confidence interval, use the [Holm-Šídák method](#)^[270] instead, as it has more power (but cannot compute confidence intervals).

- The Šídák method has a bit more power than the Bonferroni method. So from a purely conceptual point of view, the Šídák method is always preferred.
- 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.
- Previous versions of Prism offered the Bonferroni method, but not the Šídák method.

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 $0.95 \times 0.95 \times 0.95$, 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:

$$\text{alphaPC} = 1.0 - (1.0 - \text{alphaFW})^{1/K}$$

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:

$$\text{alphaPC} = 1.0 - (1.0 - \text{alphaFW})^{1/K} = 1.0 - (1.0 - 0.05)^{1/3} = 0.0170$$

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.

What if the comparisons are not independent?

Both the Bonferroni and Šídák methods were derived assuming that each comparison is independent of the others. If this assumption is not correct, then the methods are said to be conservative. That means that the true familywise risk of a Type I error (declaring statistical significance, when in fact the null hypotheses are true) is less than the stated alphaFW. That means that when the assumption of independence is false, the Bonferroni and Šídák tests do a good job of protecting against false statements of statistical significance, but have less power to detect real differences.

Other names

Sheskin (and presumably some other authors) names the tests differently(2):

- He uses the name Šídák-Bonferroni method for the method we call the Šídák method.
- He uses the name Bonferroni-Dunn method for the method we call simply the Bonferroni method.

References

1. H Abdi. [The Bonferonni and Šídák 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

2.8.3 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 Dunnett 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.
- 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](#)^[275] 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, $q = \sqrt{2} * D / SED$. The only reason to look at these q ratios is to compare Prism's results with texts or other programs.
- 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](#).

2.8.4 The Holm-Sidak approach to multiple comparisons

The Holm-Šidák test in Prism

Prism 6 offers the Holm multiple comparison test (1) as a followup to one- or two-way ANOVA, or as part of the analysis to do [many t tests](#)^[231] at once. To choose the Holm-Šidák test following ANOVA, choose a multiple comparison test that computes significance but does not also report a confidence interval.

Key facts about the Holm test

- The Holm multiple comparison test only reports which comparisons are statistically significant, and does not also compute confidence intervals or [multiplicity adjusted P values](#)^[275].
- Holm's method has more power than the Bonferroni or Tukey methods (3). 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 (2).
- 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, the 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 Šidák modification of the Holm test makes it a bit more powerful, especially when there are many comparisons.

How the Holm-Šidák test works

Here is a brief description of how the Holm multiple comparison test works:

1. P values for each comparison are computed as they are for the Fisher's LSD test. 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 α/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 α/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 α/i ...

8. Continue until you find a comparison that is not statistically significant.

Prism actually uses the Šidák modification, so computes the Holm-Šidák test. At steps 5-7 above, the P value is not compared to α/i but rather to $1-(1-\alpha)^{(1/i)}$

The Holm's method is described in detail by SA Glantz (3). He recommends the Holm test as the best multiple comparison test to use after ANOVA. But note that it cannot compute confidence intervals, so is only "best" if you don't care about confidence intervals.

References:

1. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6 (2): 65–70.
2. MA Seaman, JR Levin and RC Serlin, [New Developments in pairwise multiple comparisons: Some powerful and practicable procedures](#), *Psychological Bulletin* 110:577-586, 1991.
3. SA Glantz, *Primer of Biostatistics*, 2005, ISBN=978-0071435093.

2.8.5 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.
- 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 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).
- 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 the Fisher's LSD test works

The Fisher's LSD test begins like the Bonferroni multiple comparison test. It takes the square root of the Residual Mean Square from the ANOVA and considers that to be the pooled SD. Taking into account the sample sizes of the two groups being compared, it computes a standard error of the difference between those two means. Then it computes a t ratio by dividing the difference between means by the standard error of that difference. To compute a P value and confidence interval, the Fisher's LSD test does not account for multiple comparisons.

2.8.6 Testing for linear trend

Overview

Prism can test for linear trend as part of the followup testing after one-way (but not two-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.

Alternative names are *testing for a linear contrast*, *post-test for trend*, and *test for linear trend*.

Interpreting the results

Slope

To compute the slope, Prism performs linear regression on the column numbers (A is 1, B is 2...) vs the column means.

R square

There are (at least) two different ways to define R^2 in the context of testing for linear trend after ANOVA.

Prism reports the *effect size* R^2 , which is the fraction of the total variance accounted for by the linear trend. It is the same value you'd get if you enter all the values into linear regression (using column order as X).

An alternative is the *alerting* R^2 . It 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 value

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.

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 groups 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.

When computing the slope, each column gets the same weight, even if you've entered more values into some columns than others (sample size is unequal). In contrast, when calculating the P value and R^2 , every value counts. Columns with more values get more weight. This apparent inconsistency is standard for this test. But note that while the P value and R^2 are useful values, the slope rarely is as it depends on the arbitrary assignment of code numbers to columns (A is 1, B is 2, etc.)

References

[Practical Statistics for Medical Research](#) by DG Altman, ISBN:0412276305

[Designing Experiments and Analyzing Data: A Model Comparison Perspective](#), Second Edition. by Scott E. Maxwell, Harold D. Delaney ISBN: 0805837183

2.8.7 Multiple comparisons after repeated measures 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).

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](#)²⁵¹.

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. ISBN:0805837183

2.8.8 Nonparametric multiple comparisons

Available choices

In the Multiple Comparisons tab you can choose to do 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, or to compare selected pairs of columns.

In the Options tab, you won't have any choice of method. Prism only computes one

choose the significance level alpha (usually 0.05).

How the tests are computed

Prism performs the Dunn's multiple comparison test, which is standard. One source is *Daniel's Applied nonparametric statistics*, second edition page 240-241.

First compute the value of alpha that accounts for multiple comparisons. Divide the family significance threshold, usually 0.05, by the number of comparisons you are making. If you compare each group with every other group, you are making $K*(K-1)/2$ comparisons, where K is the number of groups. If you are comparing each group to a control group, you are making K-1 comparisons. If you only comparing a set of preselected pairs of groups or treatments, K is the number of comparisons.

Next find the value of z from the normal distribution that corresponds to that two-tailed probability. [This free calculator](#) will help. For example, if there are 4 groups, you are making 6 comparisons, and the critical value of z (using the usual 0.05 significance level for the entire family of comparisons) is the z ratio that corresponds to a probability of 0.05/6 or 0.008333. That z ratio is 2.638.

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, divide 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, divide this difference in mean ranks by the square root of $[(N*(N+1) - \sum(T_i^3 - T_i) / (N - 1)) / 12 * (N_i + 1/N_j)]$, 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. Divide 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).

If the ratio calculated in the preceding paragraph is larger than the critical value of z computed in the paragraph before that, then conclude that the difference is statistically significant.

Note that this method accounts for ties when computing the ranks, and thus when computing the mean ranks which are compared.

2.8.9 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.

Key facts about multiplicity adjusted P values

- A separate adjusted P value 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 Fisher's Least Significant Difference test.
- Choosing to 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 calculation (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, ISBN: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.

2.8.10 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.

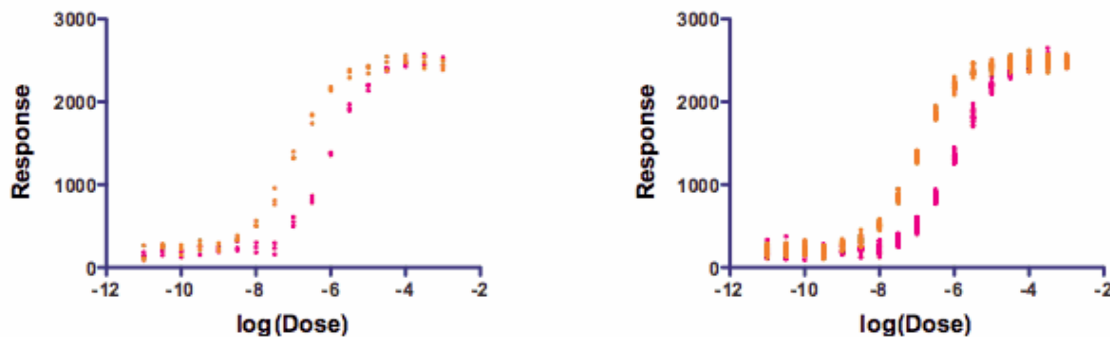
It is tempting to then run multiple comparison tests at each dose (or each time point) asking whether the difference between treatment groups is statistically significant. I don't see how these multiple comparison tests provide useful information. If you have two distinct dose-response curves, you expect 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.

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 (Prism file). 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.

2.8.11 Q&A: 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 post tests will find a significant difference between pairs of means. But 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 post 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 post test 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 the post tests of Bonferroni, Tukey, Dunnett, or Newman-Keuls can find significant differences even when the overall ANOVA showed no significant differences among groups. These tests are more focussed, 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)

There are two exceptions. 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. Another exception is the restricted Fisher's Least Significant Difference (LSD) test. In this form (the *restricted* Fisher's 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 is unrestricted -- the results

don't depend on the overall ANOVA P value.

▼ **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 focussed and answered by multiple comparison tests (post tests). In these cases, you can safely ignore the overall ANOVA results and jump right to the post test results.

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

▼ **The q or t ratio**

Each Bonferroni comparison is reported with its t ratio. Each comparison with the Tukey, Dunnett, or Newman-Keuls post test is reported with a q ratio. We include it so people can check our results against text books or other programs. The value of q won't help you interpret the results.

For a historical reason (but no logical reason), the q ratio reported by the Tukey (and Newman-Keuls) test and the one reported by Dunnett's test differ by a factor of the square root of 2, so cannot be directly compared.

▼ **How to interpret statistical significance**

“Statistically significant” is not the same as “scientifically important”. Before interpreting the P value or confidence interval, you should think about the size of the difference you are looking for. How large a difference would you 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 of the experiment.

▼ **Compared to comparing two groups with a t test, is it always harder to find a 'significant' difference when I use a multiple comparisons test following ANOVA?**

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.

But multiple comparisons tests do more than set a stricter threshold of significance. They also use the information from all of the groups, even when comparing just two. It uses the information in the other groups to get a better measure of variation. Since the scatter is determined from more data, there are more degrees of freedom in the calculations, and this usually offsets some of the increased strictness mentioned above.

In some cases, the effect of increasing the df overcomes the effect of controlling for multiple comparisons. In these cases, you may find a 'significant' difference in a post test where you wouldn't find it doing a simple t test. In the example below, comparing groups 1 and 2 by unpaired t test yields a two-tail P value equals 0.0122. If we set our threshold of 'significance' for this example to 0.01, the results are not 'statistically significant'. But if you compare all three groups with one-way ANOVA, and follow with a Tukey post test, the difference between groups 1 and 2 is statistically significant at the 0.01 significance level.

Group 1	Group 2	Group 3
34	43	48
38	45	49
29	56	47

▼ How can I get "exact" P values from multiple comparisons tests?

This is a common question, but it has three alternative answers:

Answer 1: Don't bother

The whole idea of statistical hypothesis testing is to make a crisp decision from comparison. That is the only reason to use the term "statistical significance".

In many situations in science, this is not a useful way of thinking about the data. You don't need to make a decision from each comparison, so don't need each comparison to be reported as "statistically significant" or not. In these cases, you can ignore the significance conclusions altogether, and (perhaps) also ignore the P value.

Instead focus on the confidence intervals. Many find it much simpler to think about confidence intervals rather than P values. Note that the confidence intervals reported with multiple comparisons tests (except for Fisher's LSD) adjust for multiple comparisons. Given the usual assumptions, you can be 95% confident that all the intervals contain the true population value, which leaves a 5% chance that one or more of the intervals do not include the population value.

Answer 2: Multiplicity adjusted P values

A [multiplicity adjusted P value](#)²⁷⁵ 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. Prism does not yet compute adjusted P values, but this is high on our priority list. Think twice before reporting adjusted P values. They are a bit hard to understand. And since they are not commonly reported, they may be misunderstood by others.

Prism can compute a *multiplicity adjusted P value* for each comparison. Since adjusted P values are not reported by most programs, and are not widely reported in scientific papers, be sure [you fully understand what they mean](#)^[275] before reporting these values.

Answer 3: Fisher's Least Significant Differences. P values that don't correct for multiple comparisons

An alternative to adjusted P values is to compute a P value (and confidence interval) for each comparison, without adjusting for multiple comparisons. This is sometimes called the unprotected [Fisher's Least Significant Difference \(LSD\) test](#)^[271]. 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. Since the Fishers LSD test does not correct for multiple comparisons, it is easy to be misled by the results. Since the analysis method does not account for multiple comparisons, the reader must do so when evaluating the results.

Adjusted P values are very different than Fisher's LSD P values.

Note the huge difference:

Approach	Correct for multiple comparisons?
Adjusted P values	Yes
Fishers LSD test	No

The "exact" P values computed by the two methods, therefore, will give very different results and must be interpreted very differently. If you report either, be sure to be very explicit about exactly what P value you are reporting.

▼ Is it enough to notice whether or not two sets of error bars overlap?

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.

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.

▼ How can Prism summarize a comparison after ANOVA with *, but answer "Statistically significant?" with "No"?

The summary (with asterisks) always uses [the same scale](#) regardless of your choice of alpha. In contrast, the answer to "Statistically significant" depends on the value of alpha you selected in the Options tab of the ANOVA dialog. If you set alpha to 0.01 for example, all P values between 0.01 and 0.05 will be labeled with * since a single asterisk is used to label P values in that range. But the answer to the question "Significant?" is No because the P value is greater than 0.01.

[Details.](#)

2.8.12 How Prism computes multiple comparison tests

Nonparametric multiple comparisons are explained [here](#)²⁷⁴.

The details of how Prism does its calculations are in [this eight-page document](#). All calculations use standard methods detailed in Maxwell and Delaney (1).

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.

1. Scott E. Maxwell, Harold D. Delaney, [Designing Experiments and Analyzing Data: A Model Comparison Perspective](#), Second Edition, ISBN: 0805837183
2. Margaret Diponizio, 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.

2.9 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.

2.9.1 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.

[A note of caution for statistical novices](#)²⁸⁵

[Deciding which factor defines rows and which defines columns?](#)²⁸⁵

[Entering data for two-way ANOVA](#)²⁸⁶

[Entering repeated measures data](#)²⁸⁷

[Missing values and two-way ANOVA](#)²⁸⁹

[Point of confusion: ANOVA with a quantitative factor](#)²⁹¹

[Experimental design tab: Two-way ANOVA](#)²⁹³

[Multiple comparisons tab: Two-way ANOVA](#)²⁹⁶

[Options tab: Two-way ANOVA](#)³⁰⁰

[Summary of multiple comparisons available \(two-way\)](#)³⁰²

[Q&A: Two-way ANOVA](#)³⁰³

2.9.1.1 A note 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.

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.

2.9.1.2 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:

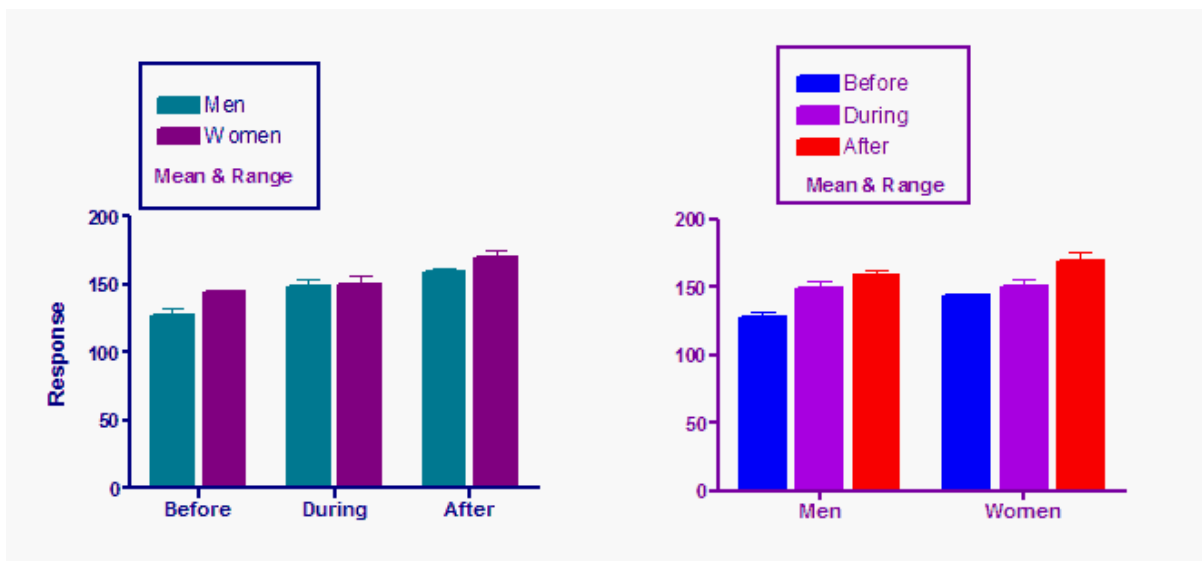
	X Labels	A		B	
		Men		Women	
	X	A:Y1	A:Y2	B:Y1	B:Y2
1	Before	123	132	143	145
2	During	143	154	141	156
3	After	162	156	175	164

	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.

2.9.1.3 Entering data for two-way ANOVA

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

If you are not ready to enter your own data, choose to use sample data and choose of of the two-way ANOVA sample data sets.

If you plan to enter your own data, create a table with enough subcolumns to hold the maximum number of replicates you have.

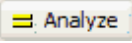
Entering averaged data

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. 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  on the toolbar.
2. Choose Two-way ANOVA from the list of grouped analyses.
3. On the first tab (Experimental Design), define whether or not your experimental design used [repeated measure](#)^[287]. Optionally name the grouping variables that define the rows and columns.
4. On the second ([Multiple Comparisons](#)^[296]) and third ([Options](#)^[300]) tab, choose multiple comparisons.

2.9.1.4 Entering repeated measures data

The term *repeated-measures* refers to an experiment that collects multiple measurements from each subject. The analysis of repeated measures data is identical to the analysis of randomized block experiments that use paired or matched subjects. 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.

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](#)²⁸⁵ as it affects the appearance of graphs.

Table format: Grouped		A		B	
		Control		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

The experiment was done with six animals. Each animal was given one of two treatments at one of three doses. The measurement was then made in duplicate. The value at row 1, column A, Y1 (23) came from the same animal as the value at row 1, column A, Y2 (24). Since the matching is within a treatment group, it is a replicate, not a repeated measure. Analyze these data with ordinary two-way ANOVA, not repeated-measures ANOVA.

Matched values are spread across a rows

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.

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 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](#), 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.

2.9.1.5 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				
	x	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 for ordinary (but not repeated measures) ANOVA, 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. Since this example will be analyzed by ordinary two-way ANOVA, it doesn't matter much that there are only two (not three) replicates for control men and treated men. When you want to use repeated-measures ANOVA, missing values are not allowed.

Table format: Two-way		A			B		
		Control			Treated		
	x	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, You must never leave n blank or enter zero, but it is ok if n is not always the same.

Missing values with repeated measures ANOVA

The rules regarding missing values and repeated measures two-way ANOVA are:

- Prism can compute repeated measures two-way ANOVA fine if alternative treatments were given to different numbers of subjects.
- Prism cannot compute repeated measures two-way ANOVA if values at some time points are missing.

For the examples shown below, assume that the matched values are stacked into a subcolumn, which is the most common way to enter data.

OK. Every subject has data at every time point.

Table format: Grouped		Group A			Group B		
		Control			Treated		
		A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Baseline	34	65	54	39	65	
2	Injection	35	67	58	41	54	
3	1 hour later	78	111	98	167	211	
4	6 hours later	54	98	89	143	178	
5	12 hours later	42	89	87	136	146	

Not OK. Two subjects have a missing value.

Table format: Grouped		Group A			Group B		
		Control			Treated		
		A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1	Baseline	34	65	54	39	65	73
2	Injection	35	67	58	41		69
3	1 hour later	78		98	167	211	198
4	6 hours later	54	98	89	143	178	167
5	12 hours later	42	89	87	136	146	143

In the top table, column A has data in three subcolumns, while column B had data in only two subcolumns. This is not a problem, and repeated measures ANOVA will work fine. Note that it is not possible to format a table, in this case, with three subcolumns for data set A, and two for data set B. Prism always creates tables with the same number of subcolumns in each data set, but it is ok to simply leave subcolumns blank.

Prism cannot perform repeated measures ANOVA from the second data table below. The problem is that the value in row 2 of subcolumn 2 of data set B is missing. Computing repeated measures ANOVA with missing values is far from straightforward, and Prism doesn't attempt it.

2.9.1.6 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.

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 first-step in comparing curves.

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. It rarely makes sense to test this hypothesis. Of course time affects the outcome! That's why you did a time course.
- Another P value tests the null hypothesis that the treatment makes no difference, on average. This can be somewhat useful. But in most cases, you expect no difference at early time points, and only care about differences at late time points. So it may not be useful to ask if, on average, the treatments differ.
- 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.

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. This kind of analysis does not ask a fundamental question, and so the results are rarely 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.

2.9.1.7 Experimental design tab: Two-way ANOVA

Parameters: Two-Way ANOVA

Experimental Design Multiple Comparisons Options

Experimental Design

No matching. Use regular two-way ANOVA (not repeated measures)
 Each subcolumn represents a different time point, 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

Table format:		A		B		C	
Grouped		Title		Title		Title	
	☒	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Time1						
2	Time2						
3	Time3						

Assume sphericity (equal variability of differences)?

Yes. Results will match prior versions of Prism.
 No. Recommended.

Factor names

Name the factor that defines the columns:

Name the factor that defines the rows:

Based on your choices (on all three tabs), Prism will perform:
 - RM two-way ANOVA, matched values are both stacked and spread across a row.

Learn Cancel OK

Experimental Design

Repeated measures defined

Repeated measures means that the 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*.

Which factor is matched?

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 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).

Table format: Grouped		A		B		C	
		Title		Title		Title	
	⊗	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Title						
2	Title						
3	Title						
4	Title						

Each subcolumn represents a different time point, so matched values are spread across a row.

Table format: Grouped		A		B		C	
		Time1		Time2		Time3	
	⊗	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Title						
2	Title						
3	Title						
4	Title						

Each row represents a different time point, so matched values are stacked into a subcolumn.

Table format: Grouped		A		B		C	
		Title		Title		Title	
	⊗	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Time1						
2	Time2						
3	Time3						
4	Title						

Repeated measures by both factors.

Table format: Grouped		A		B		C	
		Title		Title		Title	
	x	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Time1						
2	Time2						
3	Time3						
4	Time4						

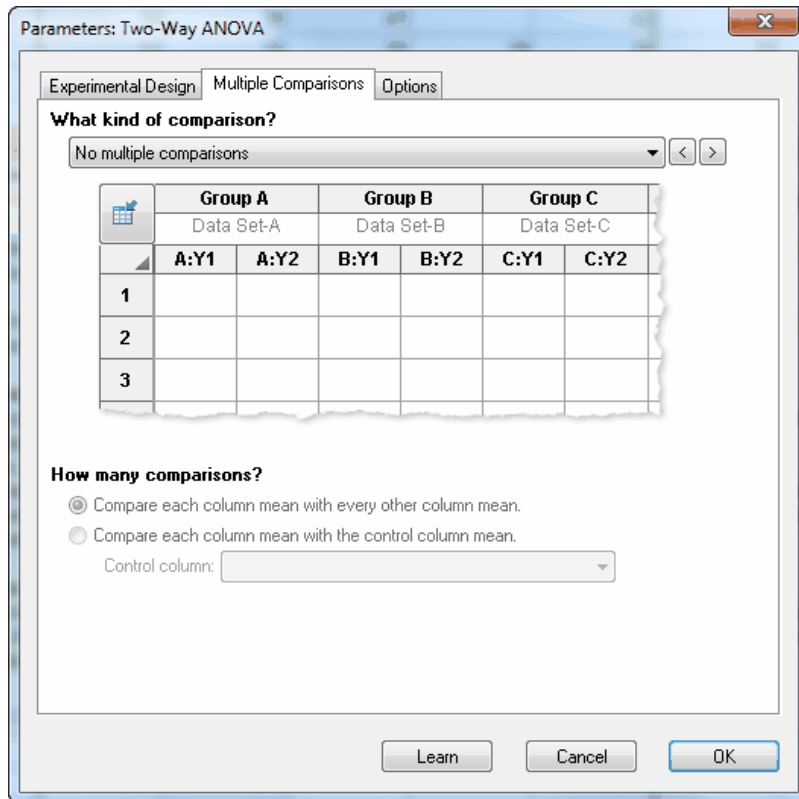
Factor names

Entering descriptive names for the two factors will make it easier to interpret your results.

Sphericity

With two-way repeated measures ANOVA, Prism always assumes [sphericity](#)^[251]. It cannot do the Greenhouse-Geisser correction for two-way ANOVA, as it does for one way, nor can it calculate [epsilon](#)^[255]. 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).

2.9.1.8 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.

	Group A		Group B	
	Data Set-A		Data Set-B	
	A:Y1	A:Y2	B:Y1	B:Y2
1	Mean		Mean	
2	Mean		Mean	
3	Mean		Mean	

- Compare each cell mean with the other mean in that column

	Group A		Group B		Group C	
	Data Set-A		Data Set-B		Data Set-C	
	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Mean		Mean		Mean	
2	Mean		Mean		Mean	

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.

	Group A		Group B		Group C	
	Data Set-A		Data Set-B		Data Set-C	
	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Mean		Mean		Mean	
2	Mean		Mean		Mean	
3	Mean		Mean		Mean	

- 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.

	Group A		Group B		Group C	
	Data Set-A		Data Set-B		Data Set-C	
	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1	Mean		Mean		Mean	
2	Mean		Mean		Mean	
3	Mean		Mean		Mean	

▣ 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.

	Group A		Group B		Group C	
	Data Set-A		Data Set-B		Data Set-C	
	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1						
2	Mean		Mean		Mean	
3						

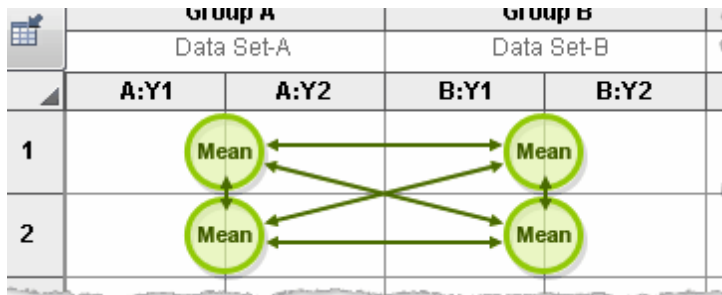
▣ Main row effects

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.

	Group A		Group B	
	Data Set-A		Data Set-B	
	A:Y1	A:Y2	B:Y1	B:Y2
1	Mean			
2	Mean			
3	Mean			

- ☐ 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 either or 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 prior versions of Prism, [note this bug](#) in versions prior to 5.04 (Windows) and 5.0d (Mac).

2.9.1.9 Options tab: Two-way ANOVA

Choose a multiple comparisons test

On the top of the options tab, refine your choice of multiple comparisons test

Multiple comparisons test

Correct for multiple comparisons: Confidence intervals and significance. Recommended.
Test: Tukey (recommended) ▾

Correct for multiple comparisons: Significance without confidence intervals. More power.
Test: Holm-Sidak (recommended) ▾

Don't correct for multiple comparisons. Each comparison stands alone.
Test: Fisher's LSD test

Your choice depends on the answer to two questions:

Do you want to compute confidence intervals as well as determining statistical significance?

We recommend that you choose a method that reports confidence intervals as well as significance for two reasons:

- Confidence intervals are easier for most to interpret than statements about statistical significance.
- If you choose a method that also computes confidence intervals (Tukey, Dunnett, Bonferroni), Prism can also compute [multiplicity adjusted P values](#)^[275]. It cannot do so for the other methods (Holm, Newman-Keuls).

But note that there is an advantage to picking a method that does not report confidence intervals. These methods tend to be more powerful. The [Holm-Šidák method](#)^[270] (which cannot compute confidence intervals) is more powerful than the Tukey or Dunnett method (1). That means that with some data sets, the Holm-Šidák method can find a statistically significant difference where the Tukey method cannot.

Do you want to correct for multiple comparisons at all?

An alternative to correcting for multiple comparisons is to perform each comparison individually. This is also called [Fisher's Least Significant Difference \(LSD\) test](#)^[271]. If you use this approach, you must correct for multiple comparisons informally while interpreting the results.

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.

We recommend avoiding the Fisher's LSD approach, unless you have a very good reason to

use it.

Which multiple comparisons test?

In most cases, choosing the experimental design and the goals of multiple comparisons will also choose the multiple comparisons test. In a few cases, you can choose between two tests.

Tukey vs. Bonferroni?

If you choose to compare every column mean with every other column mean using a method that produces confidence intervals as well as decisions about statistical significance, you'll have two choices. The Tukey test is designed for this purpose, and we recommend it. The alternative is Bonferroni. Its only advantage is that it is easier to understand, but it has less power.

Holm-Šidák vs. Newman-Keuls?

If you choose to compare every column mean with every other column mean using a method that reaches decisions about statistical significance without creating a confidence interval, you'll have two choices. We recommend that you choose the [Holm-Šidák test](#) and [avoid the Newman-Keuls test](#). The problem with the Newman-Keuls test is that 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.

[Duncan's new multiple range test](#) is a modification of the Newman-Keuls test with more power. Prism does not offer it, because it does a worse job of controlling the error rate than does the Newman-Keuls test. Few statisticians, if any, recommend Duncan's test.

Multiplicity adjusted P values

If you choose the Bonferroni, Tukey or Dunnett multiple comparisons test, Prism can also report [multiplicity adjusted P values](#)^[275]. 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".

Multiplicity adjusted P values are not commonly reported. If you choose to compute these values, take the time to be sure you understand what they mean. If you include them in publications, be sure to explain clearly what values you are reporting.

Choosing a value for alpha and the confidence level of confidence intervals

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.

Other options

Prism gives you options to create some extra graphs, and an extra page of results showing the results as narrative paragraphs rather than tables.

References

1. SA Glantz states that the Holm test is less conservative than Tukey on page 111 of *Primer of Biostatistics*, sixth edition, ISBN= 978-0071435093. I have also created some sample data sets that shows the same thing. However, I have not seen a proof that the Holm method is always less conservative than Tukey for all possible data sets.
2. MA Seaman, JR Levin and RC Serlin, *Psychological Bulletin* 110:577-586, 1991.

2.9.1.10 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](#)²⁹⁶ 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	Dunnett
	No	Holm

2.9.1.11 Q&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 outcome has two possibilities, and I know the fraction of each possible outcome in each group. How can I compare the groups?

Not with ANOVA. Enter your data into a [contingency table](#)³¹⁸ and analyze with a [chi-square test](#)³¹⁹.

What does 'two-way' mean?

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 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. Prism does not perform three-way ANOVA, but other programs do.

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 6 can analyze data where neither factor is repeated measures, one of the two factors is repeated measures, or when factors are repeated measures. Earlier versions of Prism could not analyze data where both factors are repeated

measures.

2.9.2 Ordinary (not repeated measures) two-way ANOVA

2.9.2.1 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).

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).

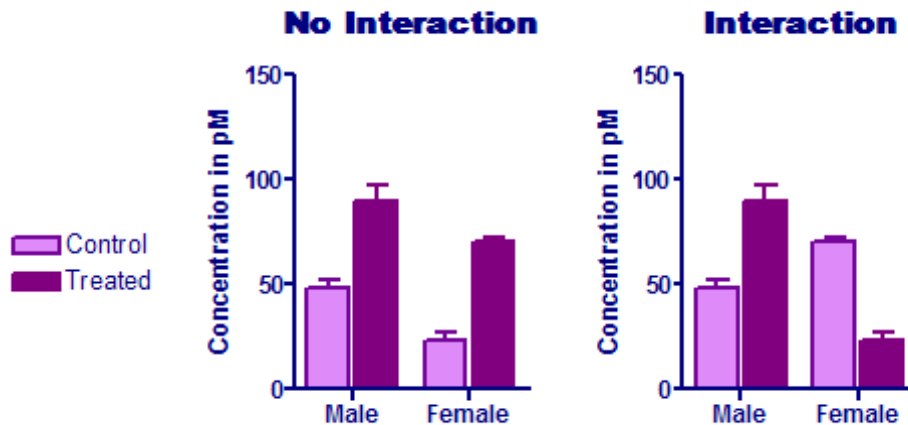
Interaction P value

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

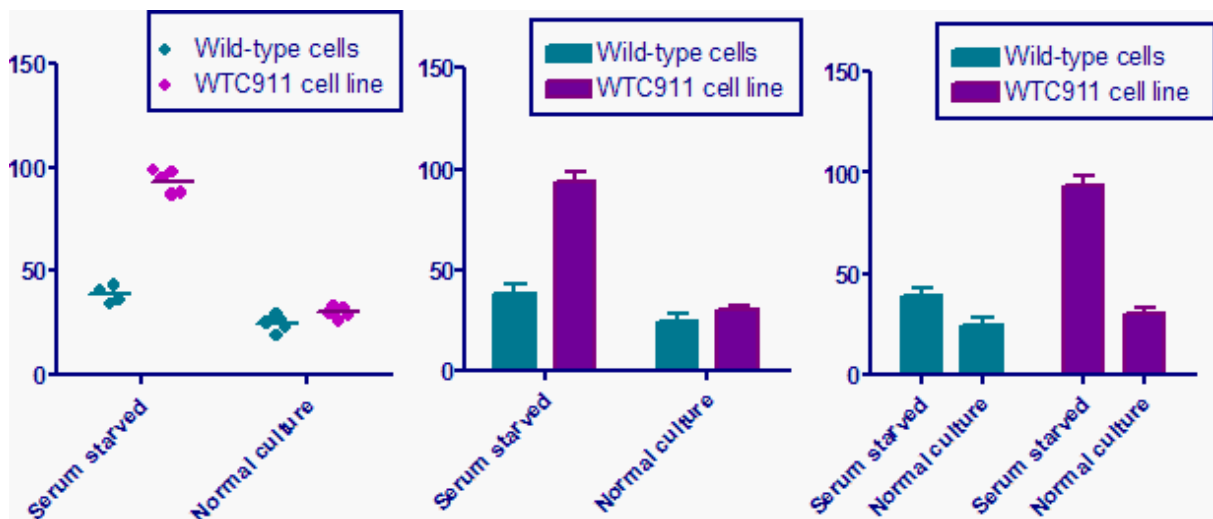
The null hypothesis is that the mean of each row (totally ignoring the columns) is the same in the overall population, and that all differences we see between row means are due to chance. In the example above, the rows represent gender, so the null hypothesis is that the mean response is the same for 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.

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](#)⁷².

2.9.2.2 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. You can 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.

Don't forget to include in the figure legend whether the error bars are SD or SEM or something different.

2.9.2.3 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. 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. 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.

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. In other cases, however, the 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.

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.

2.9.2.4 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](#)^[284], and [interpreting the results](#).^[304]

✓ 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.

2.9.3 Repeated measures two-way ANOVA

2.9.3.1 Interpreting results: Repeated measures two-way ANOVA

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](#)^[291].

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](#)^[304] first. Also read the general page on the assumption of [sphericity](#)^[251], and [assessing violations of that assumption with epsilon](#)^[255].

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

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](#)⁷².

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.

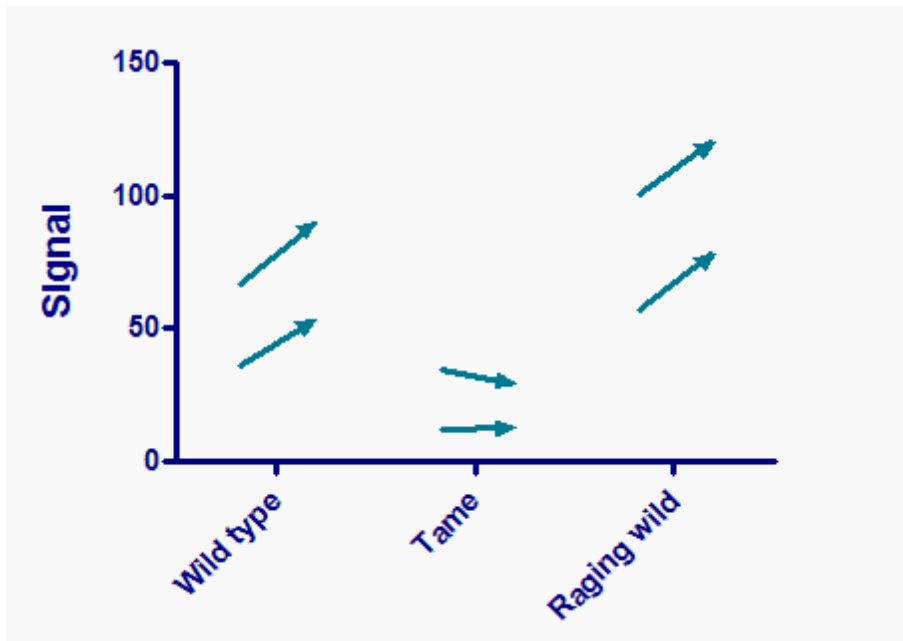
2.9.3.2 Graphing tips: Repeated measures two-way ANOVA

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.

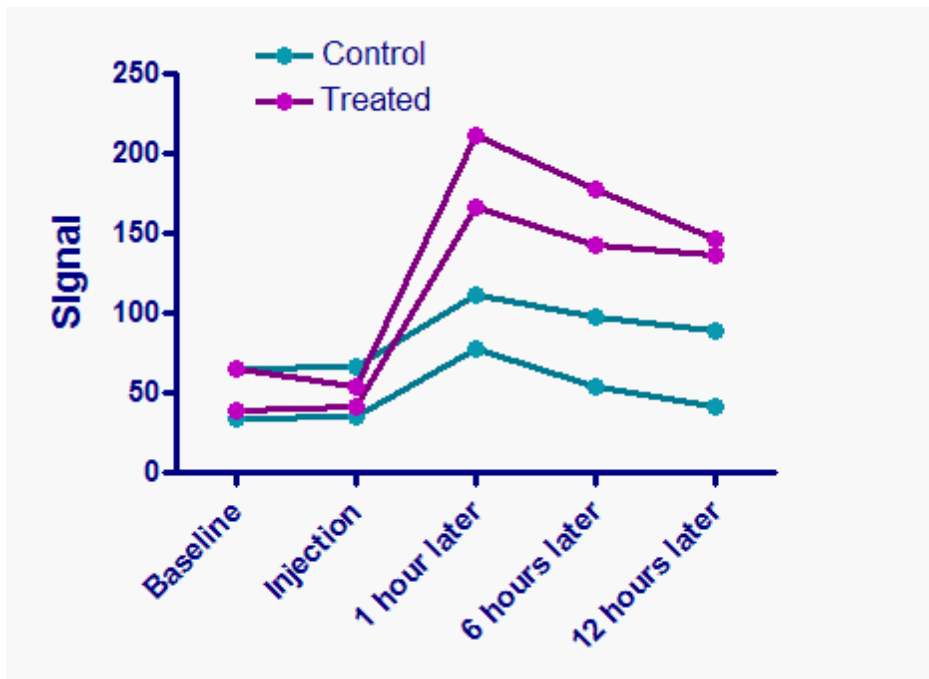


Graphing two-way ANOVA with repeated measures by column

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'.



2.9.3.3 Analysis checklist: 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](#)^[284], [graphing the data](#)^[311], and [interpreting the results](#)^{[310], [304]}

✓ 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. 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.

✓ **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 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 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. You don't have to worry about circularity 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.](#)^[297]

2.9.4 Beware of three-way ANOVA

GraphPad Prism does not perform three-way ANOVA, but many have suggested that we add three way ANOVA to a future version. One reason we have been reluctant to add three-way ANOVA to our programs is 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 three time points (1, 2 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, 2 and 3 weeks). So it seems logical to think that three-way ANOVA is the appropriate analysis. Put the data into a three-way ANOVA program, and out come 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 three 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 time 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 three 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, doesn't mean 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.

2.10 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?

2.10.1 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).

2.10.1.1 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 Prism does not offer. Logistic regression is used when the outcome is categorical, but can be used when there are multiple independent variables, which can be categorical or numerical. 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 would use logistic regression.

2.10.1.2 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.

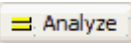
For calculation of P values, the order of rows and columns does not matter. But it does matter for calculations of relative risk, odds ratio, etc. Use the sample data to see how the data should be organized.

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](#)³³⁸.

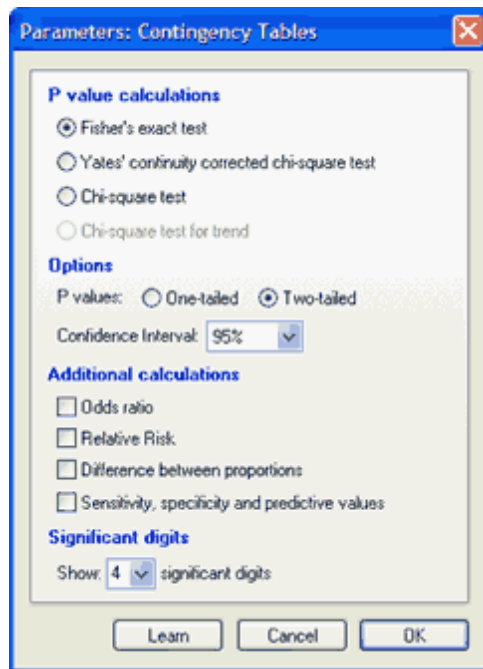
If you want to compare an observed distribution of values with a distribution expected by theory, do not use a contingency table. Prism offers [another analysis](#)³³³ for that purpose.

3. Analyze

From the data table, click  on the toolbar, and choose **Chi-square (and Fisher's exact) test**.

If your table has exactly two rows and two columns:

Prism will offer you several choices:



We [suggest you always choose Fisher's exact test](#)^[322] with a [two-tail P value](#)^[43].

Your choice of additional calculations 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.

If your table has more than two rows or two columns

If your table has two columns and three or more rows, 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.

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.

4. Review the results

[Interpreting results: relative risk and odds ratio](#)^[323]

[Interpreting results: sensitivity and specificity](#)^[325]

[Interpreting results: P values \(from contingency tables\)](#)^[326]

[Analysis checklist: Contingency tables](#)^[125]

2.10.1.3 Fisher's test or chi-square test?

If you entered data with two rows and two columns, you must choose the chi-square test 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.

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). <http://www.cytel.com/Papers/twobinomials.pdf>

2.10.1.4 Interpreting results: Relative risk and odds ratio

Relative risk and difference between proportions

The most important results from analysis of a 2x2 contingency table is the relative risk, odds ratio and difference between proportions. Prism reports these with confidence intervals.

	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 ($P_1 - P_2$) is $28\% - 16\% = 12\%$.

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.

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.

When any value equals zero

If any cell has a zero, Prism adds 0.5 to all cells before calculating the relative risk, odds ratio, or P1-P2 (to prevent division by zero).

How Prism computes the confidence interval of the odds ratio and relative risk

Prism computes the confidence interval of the odds ratio and relative risk using approximations detailed by Altman (1), which work well so long as the four values in the table are not tiny.

The confidence interval of the natural logarithm of the Odds Ratio (LOR) is computed using this equation (where A, B, C and D are the four entries in the contingency table):

$$\text{LOR} = \ln(\text{Odds ratio})$$

$$\text{SE of LOR} = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

$$95\% \text{ CI of LOR} = \text{LOR} - 1.96 \cdot (\text{SE of LOR}) \text{ to } \text{LOR} + 1.96 \cdot (\text{SE of LOR})$$

The 95% CI of the Odds Ratio is computed as the antilogarithm (exp function) of the two confidence limits computed above.

The confidence interval of the natural logarithm of the Relative Risk (LRR) is computed using this equation:

$$\text{SE of LRR} = \sqrt{\frac{1}{A} + \frac{1}{A+C} + \frac{1}{B} + \frac{1}{B+D}}$$

$$95\% \text{ CI of LRR} = \text{LRR} - 1.96 \cdot (\text{SE of LRR}) \text{ to } \text{LRR} + 1.96 \cdot (\text{SE of LRR})$$

The 95% CI of the Relative Risk is computed as the antilogarithm (exp function) of the two confidence limits computed above.

References

1. D.G. Altman, 1991, *Practical Statistics for Medical Research*, Chapman and Hall, ISBN=0-412-27630-5

2.10.1.5 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.

2.10.1.6 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. 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”](#)^[50].

You will interpret the results differently depending on whether the P value is [small](#)^[46] or [large](#)^[47].

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.

2.10.1.7 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](#)^[323], [sensitivity & specificity](#)^[325], and [interpreting P values](#)^[326].

✓ 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. Neither of these tests is offered by Prism.

✓ 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](#)^[338]. 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](#)^[333] 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.

2.10.1.8 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.

2.10.2 The Confidence Interval of a proportion

2.10.2.1 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.

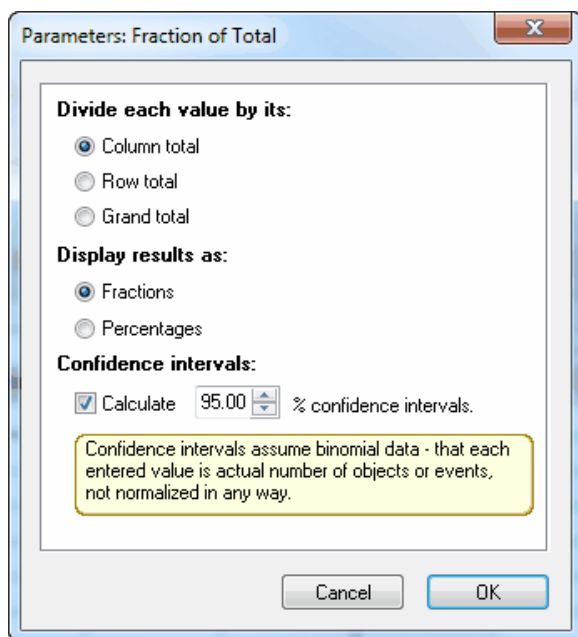
For example, you look in a microscope at cells stained so that live cells are white and dead cells are blue. Out of 85 cells you looked at, 6 were dead. The fraction of dead cells is $6/85 = 0.0706$.

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.

There are two ways to get Prism to compute the confidence interval of a proportion.

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.



2.10.2.2 How to compute the 95% CI of a proportion

Prism (like most other programs) computes the confidence interval of a proportion using a method developed by Clopper and Pearson (1). The result is labeled an “exact” confidence interval (in contrast to the approximate intervals you can calculate conveniently by hand). Computer simulations demonstrate that the so-called exact confidence intervals are really approximations(2). The discrepancy varies depending on the values of S and N. The so-called “exact” confidence intervals are not, in fact, exactly correct. These intervals may be wider than they need to be and so generally give you more than 95% confidence.

Agresti and Coull (3) recommend a method they term the **modified Wald** method. It is

easy to compute by hand and is more accurate than the so-called “exact” method. Define S to be the number of “successes” (the numerator) and n to be the total number of trials (the denominator). The CI is calculated using the following equation. The general equation is shown first, followed by an approximation for 95% confidence interval (for 95%, $=1.96$, which is close to 2).

$$p' = \frac{S + 0.5z^2}{n + z^2} \approx \frac{S + 2}{n + 4}$$

Compute W , the margin of error (or half-width) of the CI.

$$W = z \sqrt{\frac{p'(1-p')}{n + z^2}} \approx 2 \sqrt{\frac{p'(1-p')}{n + 4}}$$

The CI ranges from $p' - W$ to $p' + W$.

In some cases, the lower limit calculated using that equation is less than zero. If so, set the lower limit to 0.0. Similarly, if the calculated upper limit is greater than 1.0, set the upper limit to 1.0.

Note that the confidence interval is centered on p' , which is not the same as p , the proportion of experiments that were “successful”. If p is less than 0.5, p' is higher than p . If p is greater than 0.5, p' is less than p . This makes sense, as the confidence interval can never extend below zero or above one. So the center of the interval is between p and 0.5.

Agresti and Coull (3) showed that this method works very well, as it comes quite close to actually having 95% confidence of containing the true proportion, for any values of S and N . With some values of S and N , the degree of confidence can be less than 95%, but it is never less than 92% confidence.

One of the GraphPad QuickCalcs [free web calculators](#) computes the confidence interval of a proportion using both methods. Prism always uses the so called “exact” method.

References

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. RG Newcombe, Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* 17: 857-872, 1998.
3. Agresti, A., and Coull, B. A. (1998), Approximate is Better than “exact” for interval estimation of binomial proportions, *The American Statistician*, 52: 119-126.

2.10.2.3 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.

2.10.2.4 A shortcut equation for a confidence interval when the numerator equals zero

JA Hanley and A Lippman-Hand (1) devised a simple shortcut equation for estimating the 95% confidence interval of a proportion when the numerator is zero. If you observe zero events in N trials, you can be 95% sure that the true rate is less than 3/N. To compute the usual “95% confidence interval” (which really gives you 97.5% confidence), estimate the upper limit as 3.5/N. This equation is so simple, you can do it by hand in a few seconds.

Here is an example. You observe 0 dead cells in 10 cells you examined. What is the 95% confidence interval for the true proportion of dead cells? The exact 95% CI is 0.00% to 30.83. The adjusted Wald equation gives a 95% confidence interval of 0.0 to 32.61%. The shortcut equation computes upper confidence limits of 3.5/10, or 35%. With such small N, the shortcut equation overestimates the confidence limit, but it is useful as an estimate you can calculate instantly.

Another example: You have observed no adverse drug reactions in the first 250 patients treated with a new antibiotic. What is the confidence interval for the true rate of drug reactions? The exact confidence interval extends from 0% to 1.46% (95% CI). The shortcut equation computes the upper limits as 3.5/250, or 1.40%. With large N, the shortcut equation is quite accurate.

Reference

1. Hanley JA, Lippman-Hand A: "If nothing goes wrong is everything all right? Interpreting zero numerators". Journal of the American Medical Association 249(13): 1743-1745, 1983.

2.10.3 Compare observed and expected distributions

2.10.3.1 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:		A
Parts of whole		# 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 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.

Parameters: Compare observed distribution with expected

Data set to analyze:
A: # of seeds

Enter expected values as:

- Actual numbers of objects or events
- Percentages

With two rows, perform:

- Binomial test (recommended)
- Chi-square test for goodness of fit

Expected distribution

Row	Outcome	Observed #	Expected #
1	Round and yellow	315	312.75
2	Round and green	108	104.25
3	Angular and yellow	101	104.25
4	Angular and green	32	34.75

This analysis expects that each value in the data table is an actual number of events or items, and is not normalized in any way.

Learn Cancel OK

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 discrepancy 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.

2.10.3.2 How the chi-square goodness of fit tes 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](#)³³⁶, which gives an exact result instead of either form of the chi-square calculation, which is only an approximation.

2.10.3.3 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 at 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*.

2.10.3.4 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:

$$\chi^2 = \frac{(|R - S| - 1)^2}{R + S}$$

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.

2.10.3.5 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](#)^[318]. 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](#)^[134].

Relationship to the Kolmogorov-Smirnov test

The [Kolmogorov-Smirnov test](#)^[220] 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.

2.10.3.6 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.

2.11 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.

2.11.1 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.

2.11.2 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](#)³⁴¹ 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		X	A	B
Survival		Days after randomization	Control	Treated
	X	X	Y	Y
1	AB	34	1	
2	GT	66	1	
3	RF	64	0	
4	ED	88	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	78		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.

Entering data for survival studies can be tricky. See [answers to common questions](#)³⁴³, an [example of a clinical study](#)³⁴⁵, and an [example of an animal study](#)³⁴⁶.

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](#)³⁵⁰

[Interpreting results: Comparing two survival curves](#)³⁵¹

[Interpreting results: Comparing three or more survival curves](#)³⁵⁶

[Analysis checklist: Survival analysis](#)¹²⁶

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

2.11.3 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](#)³⁴¹ 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?

No. You must enter the number of days (or months, or some other unit of time). Use a spreadsheet to subtract dates to calculate duration.

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.](#)

2.11.4 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
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:		X	A	B
Survival		Days	Control	Treated
	x	X	Y	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.

2.11.5 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	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

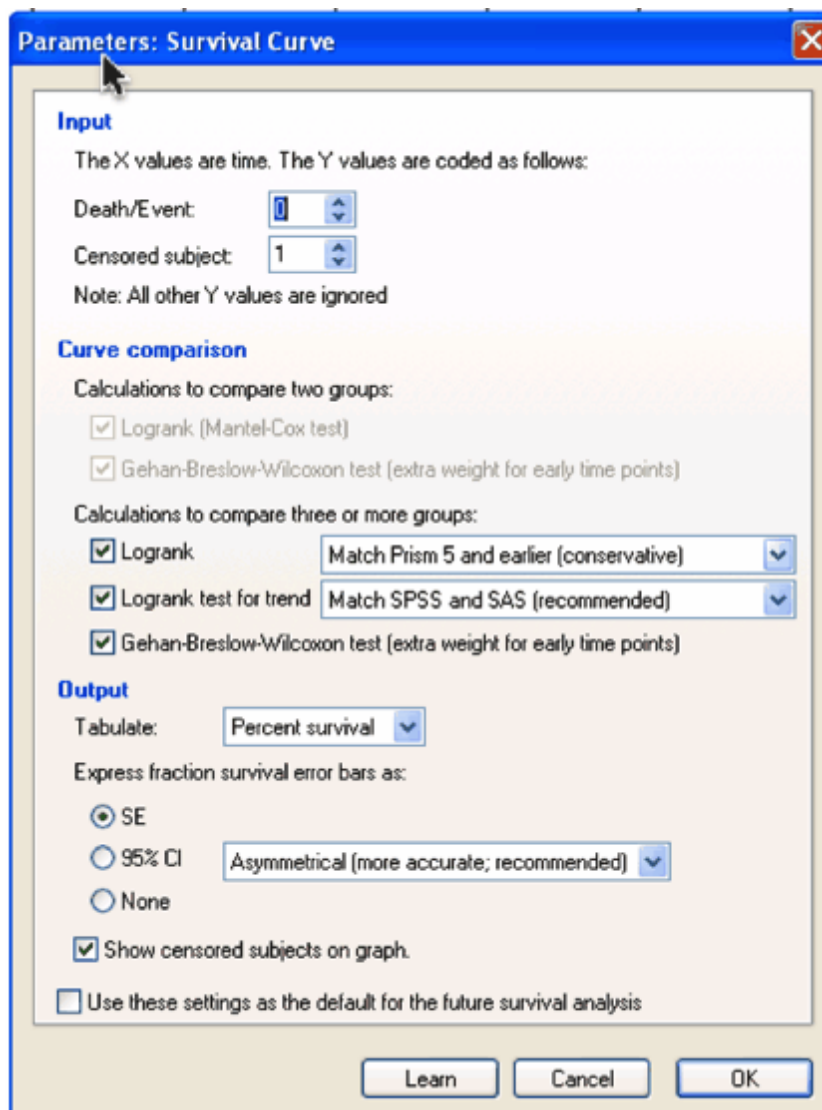
for the two animals still alive at the end of the study.

2.11.6 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.



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.** This is equivalent to the Mantel-Haenszel method, except the two methods 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.

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 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 6 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.

Output

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.

Reference

David Machin, Yin Bun Cheung, Mahesh Parmar, [Survival Analysis: A Practical Approach](#), 2nd edition, ISBN:0470870400.

2.11.7 Interpreting results: Kaplan-Meier curves

Kaplan-Meier survival fractions

Prism calculates survival fractions using the product limit (Kaplan-Meier) method. For each X value (time), Prism shows the fraction still alive (or the fraction 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](#) ³⁴¹--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.

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.

Number of subjects at risk at various times

One of the pages (or 'views') in the survival analysis page is "# of subjects at risk". Since the number at risk applies to a range of days, and not to a single day, the table is a bit ambiguous.

Here are the top six rows of that table for the sample data (comparing two groups) that you can choose from Prism's Welcome dialog.

Days	Standard	Experimental
0	16	14
90	16	
142	15	
150	14	
369	13	
272		14

The experiment starts with 16 subjects receiving standard therapy and 14 receiving experimental therapy. On day 90, one of the patients receiving standard therapy died. So the value next to 90 tells you that there were 16 subjects alive up until day 90, and 15 at risk between day 90 and 142. At day 142, the next patient dies, also on standard therapy. So between days 142 and 150 (the next death), 14 subjects are at risk in the standard group.

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 subpage of 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".

Reference

David Machin, Yin Bun Cheung, Mahesh Parmar, [Survival Analysis: A Practical Approach](#), 2nd edition, ISBN:0470870400.

2.11.8 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](#)^[43], 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.

2.11.9 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 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).

The two methods compared

Prism 6 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).

They did not test 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 $K = (O_1 - E_1) / V$, where O_1 - is the total observed number of events in group 1, E_1 - is the total expected number of events in group 1. You'd get the same value of K if you used the other group.
3. The hazard ratio equals $\exp(K)$
4. The lower 95% confidence limit of the hazard ratio equals:

$$\exp(K - 1.96/\sqrt{V})$$
5. The upper 95% confidence limit equals:

$$\exp(K + 1.96/\sqrt{V})$$

The logrank approach:

1. As part of the Kaplan-Meier calculations, compute the number of observed events (deaths, usually) in each group (O_a , and O_b), and the number of expected events assuming a null hypothesis of no difference in survival (E_a and E_b).
2. The hazard ratio then is:

$$HR = (O_a/E_a)/(O_b/E_b)$$
3. The standard error of the natural logarithm of the hazard ratio is: $\sqrt{1/E_a + 1/E_b}$
4. The lower 95% confidence limit of the hazard ratio equals:

$$\exp((O_a - E_a)/V - 1.96 * \sqrt{1/E_a + 1/E_b})$$
5. The upper 95% confidence limit equals:

$$\exp((O_a - E_a)/V + 1.96 * \sqrt{1/E_a + 1/E_b})$$

Prior versions of Prism

Prism 6 reports the hazard ratio twice, once computed with the Mantel-Haenszel method and again using the logrank method.

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

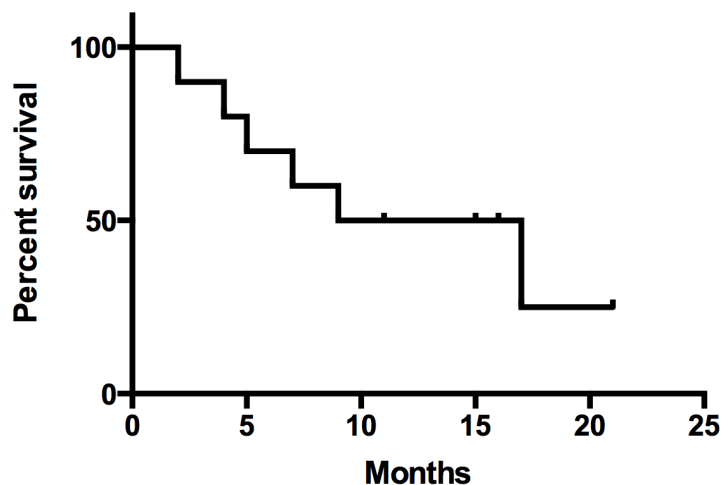
1. M.A. Hernán. [Hazards of Hazard Ratios](#), Epidemiology. 21:13-5, 2010.
2. S. L. Spruance et al, [Hazard ratio in clinical trials](#), Antimicrobial Agents and Chemotherapy vol. 48 (8) pp. 2787, 2004.
3. L Bernstein, J. Anderson and MC Pike. Estimation of the proportional hazard in two-treatment-group clinical trials. Biometrics (1981) vol. 37 (3) pp. 513-519
4. David Machin, Yin Bun Cheung, Mahesh Parmar, [Survival Analysis: A Practical Approach](#), 2nd edition, ISBN:0470870400.

2.11.10 Interpreting results: Ratio of median survival times

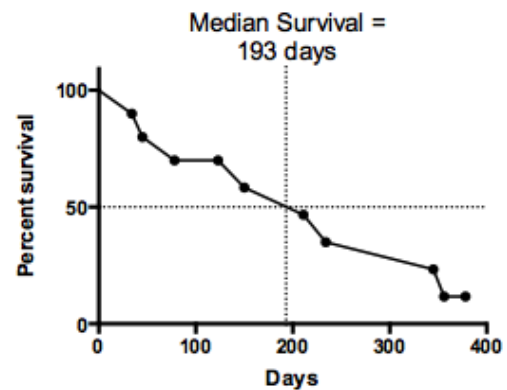
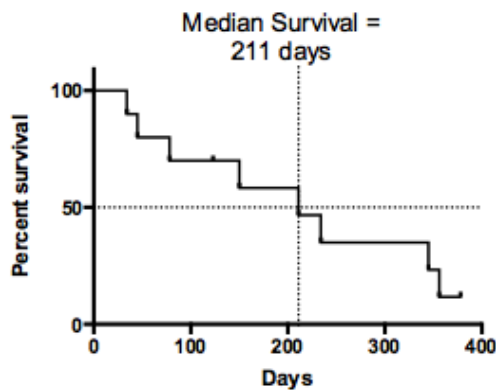
Median survival time

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.



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).

2.11.11 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](#)^[347] 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. But if you do this, you need to adjust the definition of 'significance' to [account for multiple comparisons](#).^[358]

Reference

[Survival Analysis: A Practical Approach](#), Second edition, by David Machin, Yin Bun Cheung, Mahesh Parmar, ISBN:0470870400.

2.11.12 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](#).^[72] Probably? There certainly are arguments for [not adjusting for multiple comparisons](#).^[75]

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:

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'.

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.

2.11.13 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](#)^[350], and comparing [two](#)^[351] (or [more than two](#)^[356]) 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.

2.11.14 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.

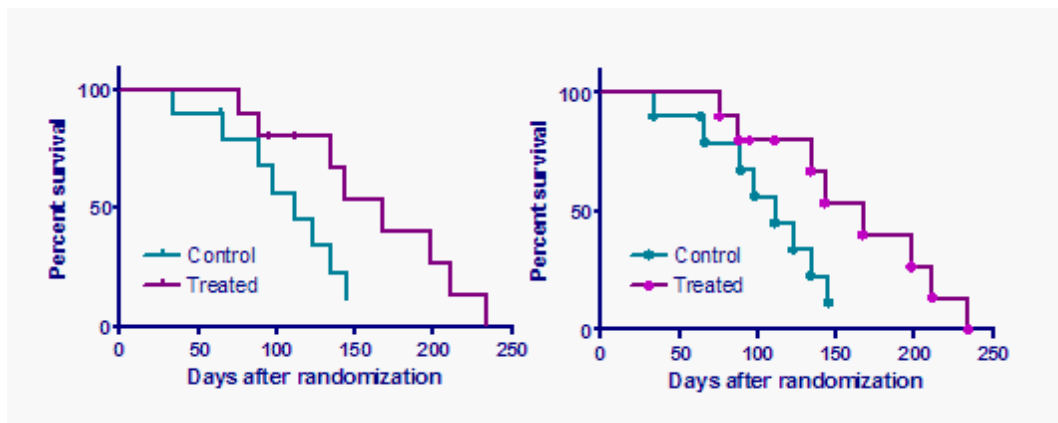
How to compute the data

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

Graphs without error bars

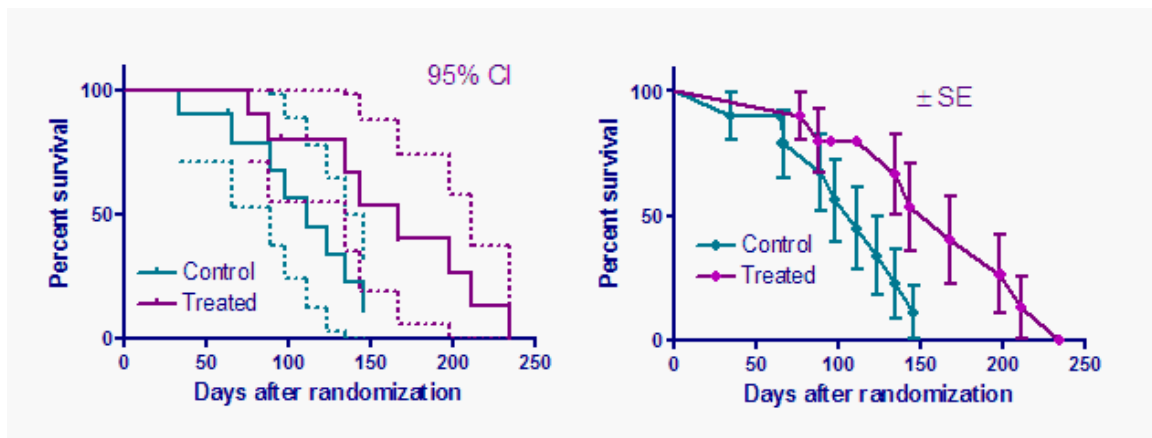


As shown above, survival curves are usually plotted as staircases. Each death is shown as a drop in survival.

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.

2.11.15 Q&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 the 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 Wilcoxon-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.

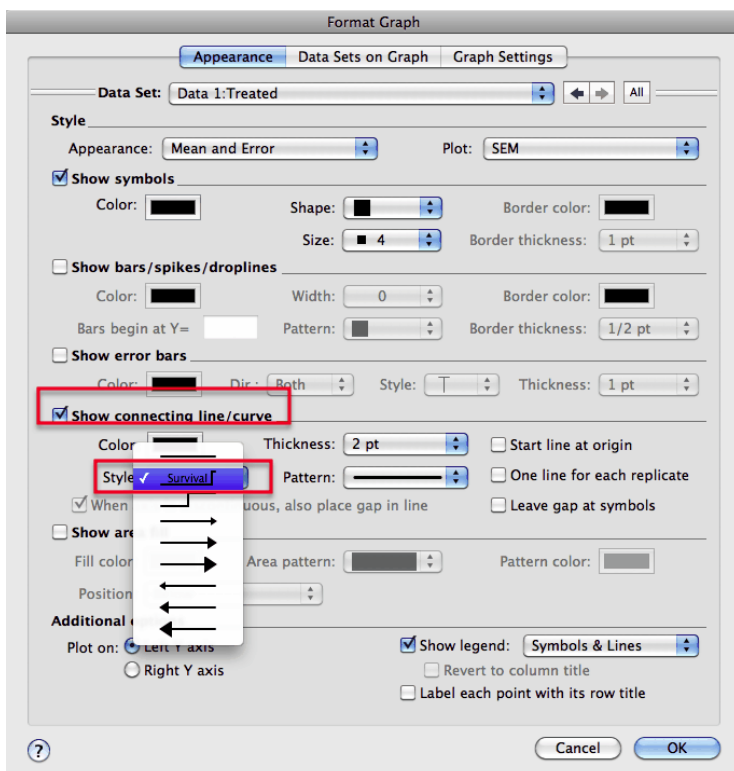
But what if you already know the percent survival at each time point, and just want to

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 won't 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 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 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.

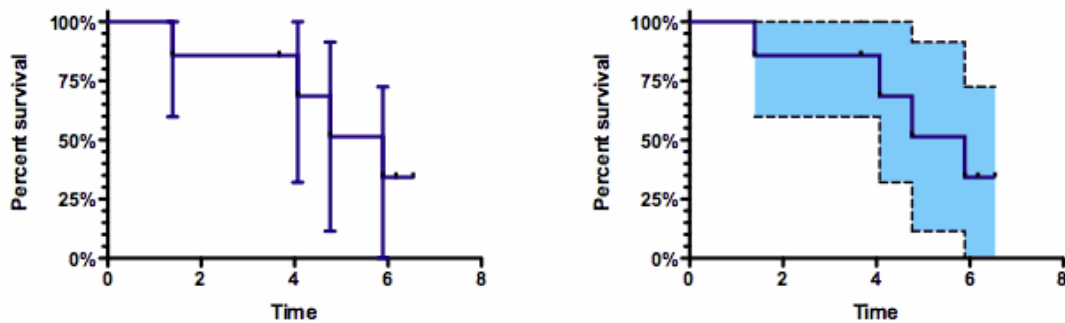
The P value comes from the logrank test, which compares the entire curve, and works fine even if the percent survival is always greater than 50%. Two curves can be very different, even if they never dip down below 50%.

▣ 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 do 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, ISBN:0470870400.

2.11.16 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.

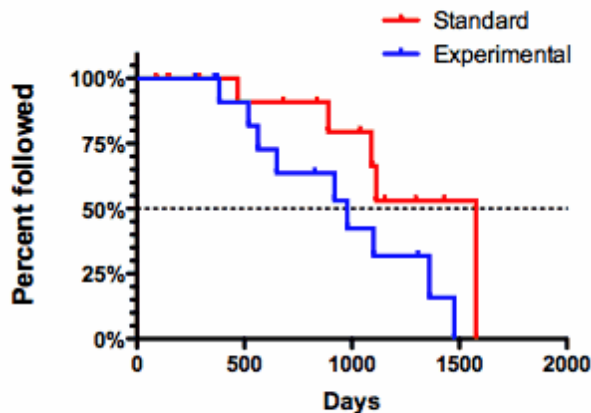
Schemper and Smith (1) followed that idea to its conclusion and devised a clever method to obtain the median followup time. Run the data through the Kaplan-Meier analysis again, but with the meaning of the status indicator reversed. The end point is loss-of-followup (which is usually considered censoring). 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

2.12 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.

2.12.1 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.

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.](#)³⁷³ In particular, note that the correlation analysis does not fit or plot a line.

2.12.2 How to: Correlation

Prism can perform correlation analyses either from an XY or Column table. 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 distribution?

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.

2.12.3 Interpreting results: Correlation

Correlation coefficient

The correlation coefficient, r , ranges from -1 to $+1$. The nonparametric Spearman correlation coefficient, abbreviated r_s , has the same range.

Value of r (or r_s) 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 r_s 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.

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 r^2 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.

How Prism computes the P value for Spearman nonparametric correlation

With 16 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 17 or more pairs, Prism computes an approximate P value for nonparametric correlation).

Prism 6 does the exact Spearman calculations hundreds of times faster than prior versions, so the cutoff for performing approximate calculations was moved up from >13 pairs to >17 pairs. Therefore Prism 6 will report different (more accurate) results for data sets with between 14 and 17 pairs than did prior versions.

2.12.4 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.

2.12.5 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 variable 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.

2.13 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.

2.13.1 ROC Curves

2.13.1.1 Key concepts: Receiver-operator 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-operator 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-operator 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.

2.13.1.2 How to: ROC curve

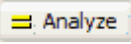
1. Enter ROC data

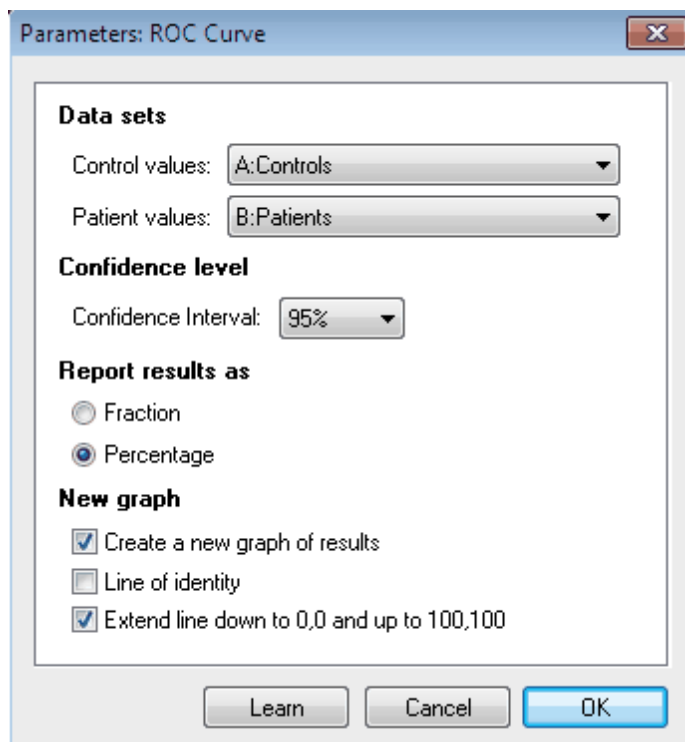
From the Welcome or New table dialog, choose the Column tab. If you are not ready to enter your own data, choose the sample ROC data.

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

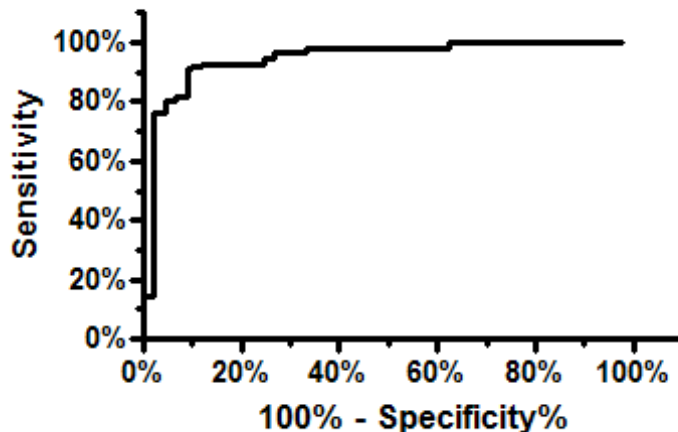
From the data table, click  on the toolbar, and then choose Receiver-operator characteristic curve from the list of one-way analyses.



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



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.

2.13.1.3 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 Specificity" tabulates those values along with their 95% confidence interval for each possible cutoff between normal and abnormal.

Area

The area under a ROC curve 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.

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.

2.13.1.4 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.

2.13.1.5 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](#)¹⁵⁹ 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).

$$SE = \sqrt{\frac{A(1-A) + (na-1)(Q1-A^2) + (nn-1)(Q2-A^2)}{na \cdot nn}}$$

where Q1 is defined as $A/(2-A)$ and Q2 as $2A^2/(1+A)$.

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

$$SE = \sqrt{\frac{0.25 + (na + nn - 2)0.083333}{na \cdot nn}}$$

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

2.13.1.6 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 specificity 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 $0.9 \times 1,000 = 900$ people with the disease (left column) will have a positive test, and 100 will not. These values go into the left column.
4. The specificity is 95%, so $0.95 \times 9000 = 8550$ 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.

- The positive predictive value is the fraction of people with a positive test who have the disease: $900/1350 = 66.7\%$
- 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.

2.13.1.7 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.

- Separately use Prism to create two ROC curves by separately analyzing your two data sets.
- For each ROC curve, note the area under the curve and standard error (SE) of the area.
- Combine these results using this equation:

$$z = \frac{|Area_1 - Area_2|}{\sqrt{SE_{Area1}^2 + SE_{Area2}^2}}$$

- 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).

- 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.

2.13.2 Comparing Methods with a Bland-Altman Plot

2.13.2.1 How to: Bland-Altman plot

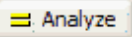
A Bland-Altman plot compares two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis.

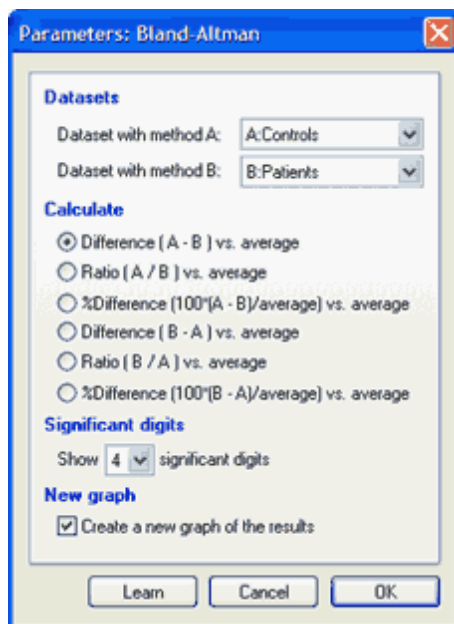
1. Enter the data

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  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

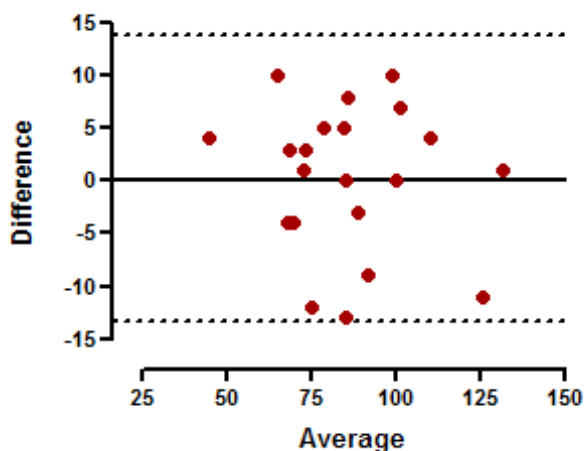
The Bland-Altman analysis creates two pages of results. The first page shows the difference

and average values, and is used to create the plot. The second results page shows the [bias](#) ³⁸⁵, which is the average of the differences, and the [95% limits of agreement](#) ³⁸⁵.

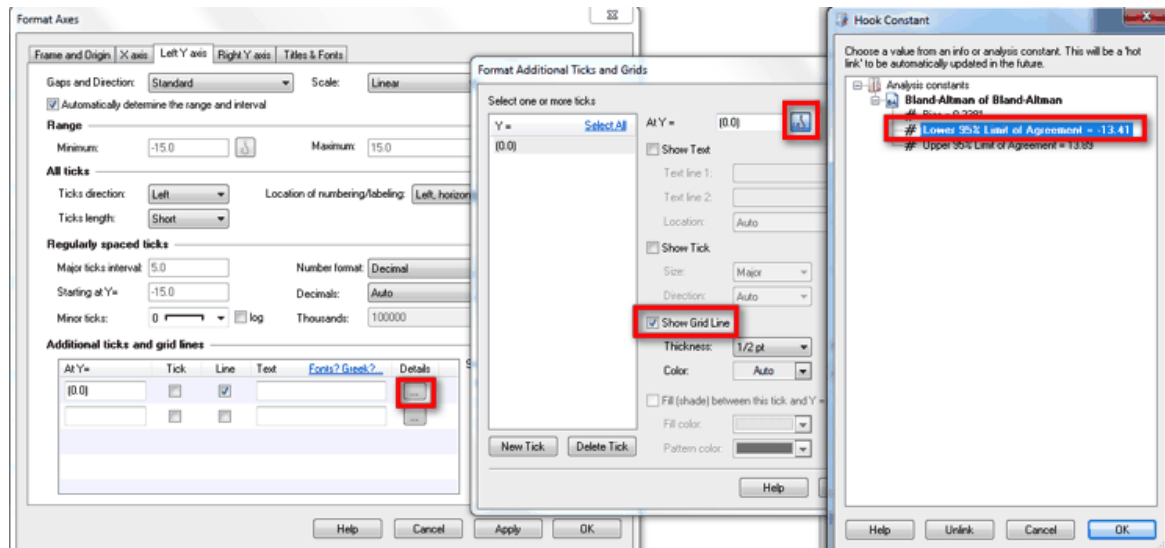
If you used the sample data, the two methods have very similar results on average, and the bias (difference between the means) is only 0.24. The 95% limits of agreement are between -13.4 and 13.9.

Bland-Altman method comp...		Value
Bias & agreement		
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.

2.13.2.2 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). The SD value is not very useful by itself, but is used to calculate the limits of agreement, computed as the mean bias plus or minus 1.96 times its SD.

For any future sample, the difference between measurements using these two assay methods should lie within the limits of agreement approximately 95% of the time.

Actually, the limits of agreement are a description of the data. It is possible to compute 95% prediction bands for the difference, and these limits would be further from the bias in each direction than do the limits of agreement (especially when the sample is small).

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?

2.13.2.3 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.

2.14 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.

2.14.1 Simulating a data table

Simulate a Column data table

To simulate a family 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.

On the Experimental design tab, choose the number of data sets, and the mean of each data set. For each data set, enter the number of values you wish to simulate for that data set (number of rows of data).

On the Random error tab, choose among several methods for generating random scatter and also adding outliers. You must choose one setting for the random values for all the data sets. For example, if you choose Gaussian error (the most common), you can only choose one standard deviation, which applies to all the data sets.

Simulate a 2x2 contingency table

To simulate a contingency table, start from any data table or graph, click Analyze, open the Simulate data category, and then select Simulate Contingency Table.

On the Experimental design tab, choose the sample size (total number of subjects for both rows and both columns). Also specify which of [four experimental designs](#)³¹⁸ you wish to simulate.

On the Rows and columns tab, name the two rows and two columns, and specify (on average) how many subjects go in each.

Prism will use the binomial random values to decide how many subjects go into each cell, maintaining the total you entered.

Note that this analysis only can simulate a 2x2 contingency table.

How Prism generates random numbers

Prism generates pseudo random numbers from the binomial or Poisson distribution, using ideas adapted from pages 372-377 of Numerical Recipes, third edition, by WH Press and colleagues.

2.14.2 How to: Monte Carlo analyses

How to begin a Monte Carlo analysis

[Simulate a data table using](#)³⁸⁸ 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](#)³⁹¹ 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.

2.14.3 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.

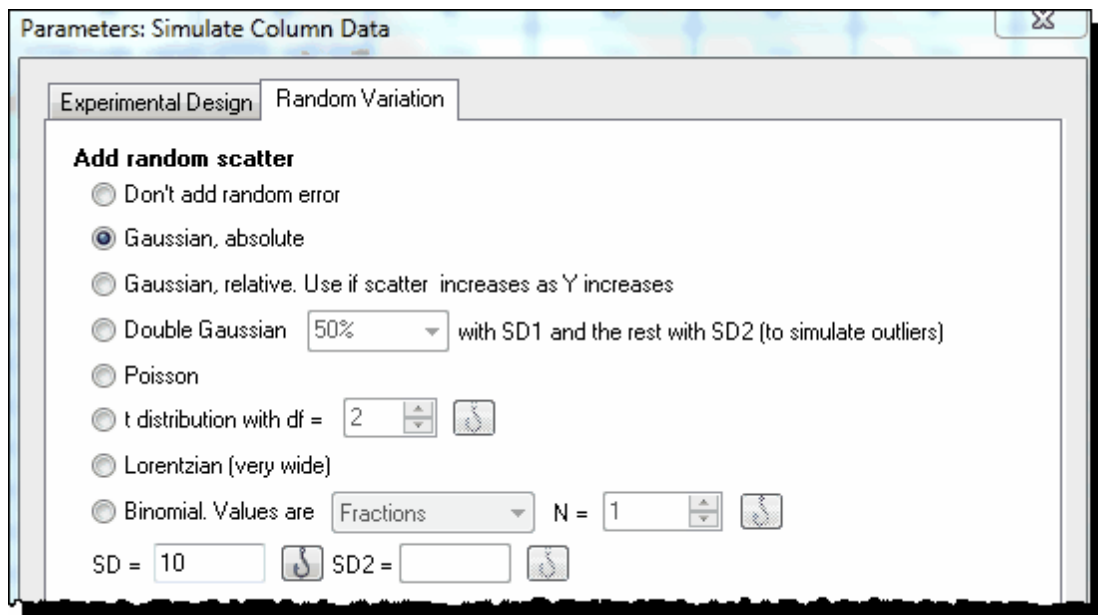
Parameters: Simulate Column Data

Experimental Design Random Variation

Number of data sets
Number of data sets: 2

Population column means
 Choose randomly from a gaussian distribution. Mean: SD:
 Enter column means individually

Index	Number of Rows	Column Mean	Hook	Column Title
A	5	25		Wild-type cells
B	5	35		GPP5 cell line



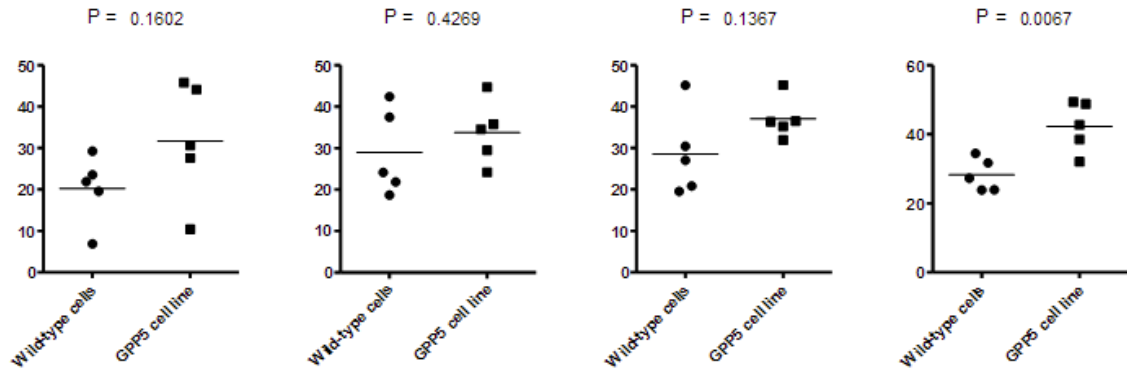
Step 2. Analyze the data with a t test

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

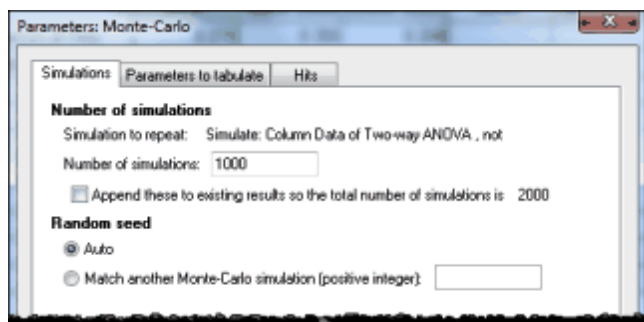
The layout below shows four such graphs placed on the layout as unlinked pictures that do not update when the graph changes. Even though there is only one graph in the project, this made it possible to put four different versions of it (with different random data) onto the layout. You can see that with random variation of the data, the P value varies a lot.



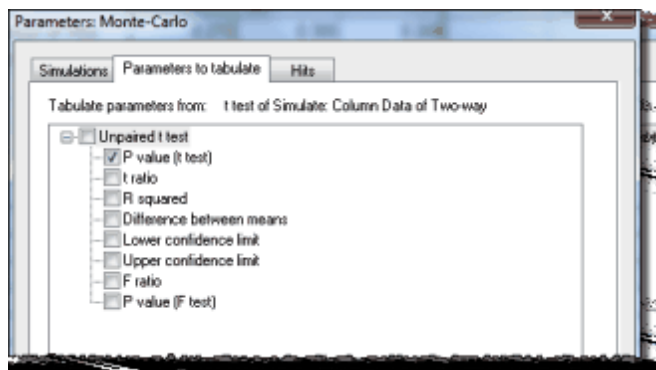
Step 4. Monte Carlo simulation

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.

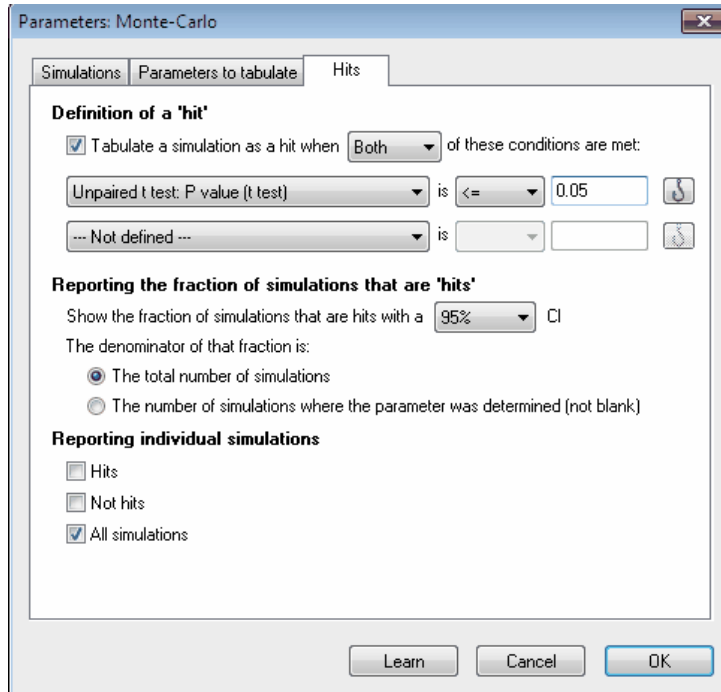


On the second (Parameters to tabulate) tab, choose which parameters you want to tabulate. The choice is the list of analysis constants that Prism creates when it analyzes the data. For this example, we only want to tabulate the P value (from the t test which compares means; don't mix it up with the P 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 $P < 0.05$.



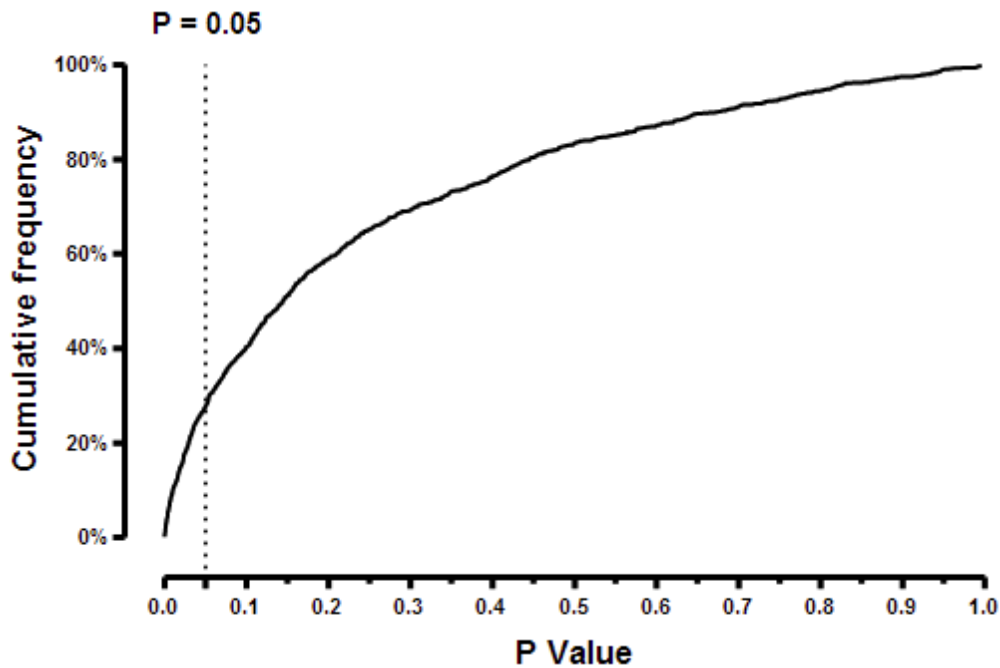
Click OK and Prism will run the simulations. Depending on the speed of your computer, it will take a few seconds or a few dozen seconds.

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

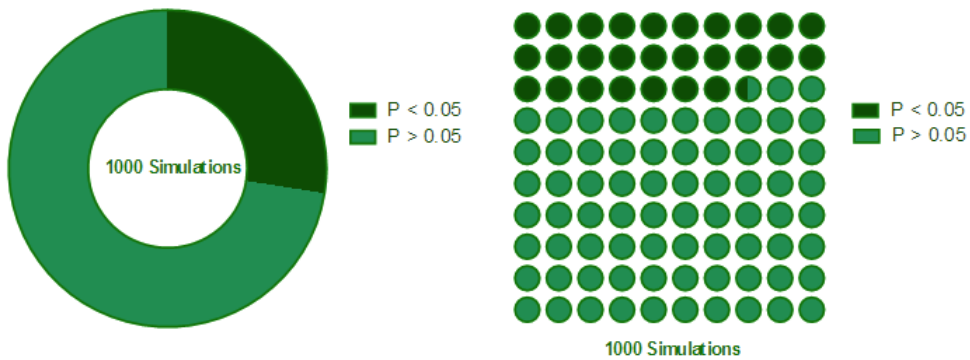
The other results table summarizes the fraction of hits. For this set of simulations, 27.5% of the simulations were hits (P value less than 0.05), with a 95% confidence interval ranging from 24.8% to 30.4%. Another way of stating these results is that the power of our experimental design is 27.5%.

 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?

Index

- * -

** Asterisks to denote significance 50

- 5 -

5-number summary 139

- A -

Adjusted P values (for multiple comparisons) 275
 Algorithms, multiple comparisons 283
 Alpha, defined 49
 Altman (Bland-Altman plot) 383
 Analysis checklist: Bland-Altman results 386
 Analysis checklist: Column statistics 136
 Analysis checklist: Contingency tables 125, 327
 Analysis checklist: Friedman's test 121, 264
 Analysis checklist: Kolmogorov-Smirnov test 222
 Analysis checklist: Kruskal-Wallis test 120, 262
 Analysis checklist: Mann-Whitney test 114, 218
 Analysis checklist: One-way ANOVA 116, 249
 Analysis checklist: Outliers 128, 172
 Analysis checklist: Paired t test 111, 203
 Analysis checklist: Repeated measures two-way ANOVA 124, 313
 Analysis checklist: Repeated-measures one way ANOVA 118, 258
 Analysis checklist: ROC curves 379
 Analysis checklist: Survival analysis 126, 359
 Analysis checklist: Two-way ANOVA 122, 309
 Analysis checklist: Unpaired t test 109, 198
 Analysis checklist: Wilcoxon matched pairs test 115, 228
 Analysis of variance, two-way 304
 ANOVA table, two-way 304
 ANOVA, one-way, instructions 234
 ANOVA, one-way, interpreting results 246
 ANOVA, repeated measures one-way 251
 ANOVA, two-way 304
 ANOVA, two-way repeated measures 310
 ANOVA, two-way, instructions 286
 Area under the curve 159

Asterisks, to denote significance 50
 AUC, area under the curve 159

- B -

Barnard test 322
 Bartlett's test 246
 Basement, analogy about power 57
 Bayesian perspective on statistical significance 51
 Berger's test 322
 Beta, defined 56
 Bias, in Bland-Altman results 385
 Bin width in frequency distributions 149
 Binomial test 336
 Biostatistics, Intuitive 17
 Bland-Altman plot 383
 Bonferroni method 87
 Bonferroni method, one-way ANOVA 241
 Bonferroni-Dunn multiple comparisons 266
 Breslow test 351
 Brown-Forsythe test 246

- C -

Case-control study 318
 Categorical variables 14
 Censored data, defined 341
 Central Limit Theorem 21
 Chi-square test or Fisher's test? 322
 Choosing a t test 179
 CI of mean 32
 Circularity 124, 313
 Coefficient of variation 141
 Column statistics, how-to 134
 Compound symmetry 124, 251, 313
 Confidence interval of a mean 32
 Confidence interval of a proportion 328
 Confidence interval of proportion 328
 Confidence intervals, advice 36
 Confidence intervals, one sided 37
 Confidence intervals, principles 34
 Contingency tables 318
 Continuity correction 322
 Correlation vs. regression 373
 Correlation, key concepts 370
 Correlation: How to 370
 Correlation: Interpreting results 371

Cross-sectional study 318
 Cumulative Gaussian distribution, fitting 154
 CV, coefficient of variation 141

- D -

D'Agostino normality test 134
 Derivative of curve 156
 Diagnostic test, accuracy of 325
 Dunnett multiple comparisons test 269
 Dunnett's test 241
 Dunn's multiple comparison test 236
 Dunn's post test 260

- E -

Epsilon, to quantify specificity 255
 Equivalence 88
 Error bars, on survival curves 361
 Error bars, overlapping 196
 Error, defined 18
 Exact P values, multiple comparisons 275
 Extremely significant, defined 50

- F -

F test for unequal variance 191
 False Discovery Rate 75
 Family-wise significance level 75
 FDR 75
 Fisher's LSD test 241, 271
 Fisher's test or chi-square test? 322
 Five-number summary 139
 Fixed vs. random factors 124, 313
 Frequency distribution, how to 149
 Friedman test, interpreting results: 263
 Friedman's test 263
 Friedman's test, analysis checklist 121, 264

- G -

Gaussian distribution, defined 18
 Gaussian distribution, fitting to frequency distribution 154
 Gaussian distribution, origin of 19
 Gehan-Breslow-Wilcoxon test to compare survival curves 351

Geisser-Greenhouse correction 237
 Geometric mean 134, 138
 Graphing hints: Contingency tables 328
 Graphing hints: Survival curves 361
 Graphing tips: Frequency distributions 153
 Graphing tips: Paired t 205
 Graphing tips: Repeated measures two-way ANOVA 311
 Graphing tips: Two-way ANOVA 307
 Graphing tips: Unpaired t 195
 Greenhouse-Geisser correction 251
 Grubbs' method 169
 Grubbs' outlier test 101
 Guilty or not guilty? Analogy to understand hypothesis testing. 53

- H -

Harmonic mean 138
 Hazard ratio 351
 Hodges-Lehmann 213
 Holm's test 241
 Holm-Sidak method 270
 Holm-Sidak test 241
 How many comparisons? 239
 How to: Bland-Altman plot 383
 How to: Column statistics 134
 How to: Contingency table analysis 319
 How to: Frequency distribution 149
 How to: Friedman test 234
 How to: Kruskal-Wallis test 234
 How to: Mann-Whitney test 210
 How to: Multiple t tests 231
 How to: One-way ANOVA 234
 How to: Paired t test 200
 How to: ROC curve 376
 How to: Survival analysis 342
 How to: Two-way ANOVA 286
 How to: Unpaired t test from averaged data 189
 How to: Unpaired t test from raw data 188
 How to: Wilcoxon matched pairs test 224
 Hypothesis testing, defined 49

- I -

Independent samples, need for 16
 Integral of a curve 156
 Interaction, in two way ANOVA 304

Interpreting results: Bland-Altman 385
 Interpreting results: Chi-square 326
 Interpreting results: Comparing >2 survival curves 356
 Interpreting results: Comparing two survival curves 351
 Interpreting results: Correlation 371
 Interpreting results: Fisher's exact test 326
 Interpreting results: Friedman test 263
 Interpreting results: Kaplan-Meier curves 350
 Interpreting results: Kolmogorov-Smirnov test 220
 Interpreting results: Kruskal-Wallis test 260
 Interpreting results: Mann-Whitney test 213
 Interpreting results: One-sample t test 143, 174
 Interpreting results: One-way ANOVA 246
 Interpreting results: Paired t 202
 Interpreting results: Relative risk and odds ratio 323
 Interpreting results: Repeated measures one-way ANOVA 256
 Interpreting results: Repeated measures two-way ANOVA 310
 Interpreting results: ROC curves 377
 Interpreting results: Sensitivity and specificity 325
 Interpreting results: Two-way ANOVA 304
 Interpreting results: Unpaired t 191
 Interpreting results: Wilcoxon matched pairs test 226
 Interpreting results: Wilcoxon signed rank test 144, 175
 Interquartile range 139
 Interval variables 14
 Intuitive Biostatistics 17
 Iterative Grubbs' method 169

- K -

Kaplan-Meier curves, interpreting 350
 Key concepts. Survival curves 341
 Key concepts: Contingency tables 318
 Key concepts: Multiple comparison tests 77
 Key concepts: Receiver-operator characteristic (ROC) curves 375
 Key concepts: t tests 179
 Kolmogorov-Smirnov normality test 134
 Kolmogorov-Smirnov test, analysis checklist 222
 Kolmogorov-Smirnov test, interpreting results 220
 Kruskal-Wallis test 260
 Kruskal-Wallis, interpreting results 260
 Kurtosis 142

- L -

Least Significant Difference 271
 Likelihood ratio 325
 Lognormal 21
 Lognormal distribution and outliers 99
 Log-rank (Mantel-Cox) test 351
 Logrank test 356
 Logrank test for trend 356
 LSD, Fisher's 271

- M -

Mann-Whitney test 209
 Mann-Whitney test, analysis checklist 114, 218
 Mann-Whitney test, step by step 210
 Mantel-Cox comparison of survival curves 351
 Masking outliers 104
 McNemar's test 336, 338
 Mean, geometric 138
 Mean, harmonic 138
 Mean, trimmed 138
 Mean, Winsorized 138
 Median 138
 Median survival 355
 Median survival ratio 351
 Medians, does the Mann-Whitney test compare? 216
 Method comparison plot 383
 Mixed model ANOVA 287, 293
 Mode 138
 Monte Carlo simulations 388
 Motulsky 17
 Multiple comparisons 77
 Multiple comparisons 72
 Multiple comparisons of survival curves 358
 Multiple comparisons test, list of 236
 Multiple comparisons, after one-way ANOVA 77
 Multiple comparisons, algorithms 283
 Multiple t test options 232
 Multiple t tests, step by step 231
 Multiplicity adjusted P values 275

- N -

Newman-Keuls test 241

Nominal variables 14
 Nonparametric tests, choosing 95
 Nonparametric tests, power of 93
 Nonparametric tests, why Prism doesn't choose automatically 92
 nonparametric, defined 92
 Normal distribution, defined 18
 Normality tests, choosing 134
 Not guilty or guilty? Analogy to understand hypothesis testing. 53
 Null hypothesis, defined 41

- O -

Odds ratio 323
 One sample t test, choosing 134
 One-sample t test, interpreting 143, 174
 One-sided confidence intervals 37
 One-tail vs. two-tail P value 43
 One-way ANOVA, analysis check list 116, 249
 One-way ANOVA, instructions 234
 One-way ANOVA, interpreting results 246
 One-way ANOVA, options tab 241
 Options tab, one-way ANOVA 241
 Ordinal variables 14
 Orthogonal comparisons 77
 Outlier, defined 98
 Outliers and lognormal distributions 99
 Outliers, danger of identifying manually 99
 Outliers, generating 388
 Outliers, identifying 169
 Outliers, masking 104
 Overlapping error bars 196

- P -

P value, common misinterpretation 41
 P value, defined 41
 P value, interpreting when large 47
 P value, interpreting when small 46
 P values, multiplicity adjusted 275
 P values, one- vs. two-tail 43
 Paired data, defined 179
 Paired t test 200
 Paired t test, analysis checklist 111, 203
 Paired t test, interpreting results 202
 Paired t test, step by step 200
 Parameters: Area under the curve 159

Parameters: Bland-Altman plot 383
 Parameters: Column statistics 134
 Parameters: Contingency tables 319
 Parameters: Frequency distribution 149
 Parameters: One-way ANOVA 237
 Parameters: ROC curve 376
 Parameters: Smooth, differentiate or integrate a curve 156
 Parameters: Survival analysis 342, 347
 Parameters: t tests (and nonparametric tests) 179
 Pearson 370
 Percentile plot 149
 Percentiles 139
 Planned comparisons 77, 84, 239
 Population, defined 13
 Post test for trend 239, 272
 Post tests 77
 Post tests, after comparing survival curves 358
 Post tests, after one-way ANOVA 77
 Post-hoc tests 77
 Power of nonparametric tests 93
 Power, defined 56
 Pratt 226
 Predictive value 325
 Probability vs. statistics 13
 Probability Y axis 154
 Proportion, confidence interval of 328
 Proportional hazards 126, 359
 Proportional hazards regression 341
 Prospective 318

- Q -

Q-Q plot 153
 Quartiles 139

- R -

r 371
 r, interpreting 371
 Random numbers 388
 Random vs. fixed factors 124, 313
 Randomized block 251
 Randomized block, defined 287
 Randomized blocks 237
 Ratio of median survival times 355
 Ratio t test 206

Ratio variables 14
 Receiver-operator characteristic (ROC) curves 375
 Regression vs. correlation 373
 Relative risk 323
 Repeated measures 251
 Repeated measures ANOVA, multiple comparisons 274
 Repeated measures two-way ANOVA 310
 Repeated measures vs. "randomized block" 287
 Repeated measures, defined 287
 Repeated measures, one-way ANOVA 237
 Repeated-measures one-way ANOVA 251
 Results: Wilcoxon matched pairs test 226
 Retrospective case-control study 318
 ROC curves 375
 Row statistics 163

- S -

Sample, defined 13
 Savitsky, method to smooth curves 156
 Scheffe's multiple comparisons 239
 SD vs. SEM 29, 30
 SD, interpreting 23
 SEM vs. SD 29, 30
 SEM, computing 28
 Sensitivity 325
 Sequential approach to P values 53
 Shapiro-Wilk normality test 134
 Šidák-Bonferroni multiple comparisons 266
 Sidak-Holm method 270
 Sign test 336
 Signed rank test, interpreting results 144, 175
 Significant, defined 49
 Simulating data with random error 388
 Simulation, to demonstrate Gaussian distribution
 Skewness 142
 Smoothing a curve 156
 Spearman 370
 Spearman correlation calculations 371
 Specificity 325
 Sphericity 124, 237, 251, 313
 Sphericity and multiple comparisons 274
 Standard deviation, computing 25
 Standard deviation, confidence interval of 26
 Standard deviation, interpreting 23
 Stars, used to denote significance 50
 Statistics vs. probability 13

STDEV function of Excel 25
 Stevens' categories of variables 14
 Survival analysis 340

- T -

t test, one sample 134
 t test, paired 200, 202
 t test, unpaired 188, 191
 t test, use logs to compare ratios 206
 Test for trend 239
 Trend, post test for 272
 Trimmed mean 138
 Tukey multiple comparisons test 269
 Two outliers, masking 104
 Two-tail vs. one-tail P value 43
 Two-way ANOVA 284, 304
 Two-way ANOVA, experimental design 293
 Two-way ANOVA, instructions 286
 Two-way ANOVA, interaction 304
 Two-way ANOVA, multiple comparisons 296
 Two-way ANOVA, options 300
 Two-way ANOVA, repeated measures 310
 Two-way ANOVA. How to enter data. 285
 Type I, II (and III) errors 58

- U -

Unpaired t test 188
 Unpaired t test, interpreting results 191
 Unpaired t test, step by step 188

- V -

Variance, defined 141
 Very significant, defined 50

- W -

Welch t test 179
 Wilcoxon (used to compare survival curves) 351
 Wilcoxon matched pairs test 223
 Wilcoxon matched pairs test, analysis checklist 115, 228
 Wilcoxon matched pairs test, step by step 224
 Wilcoxon signed rank test, choosing 134

Wilcoxon signed rank test, Interpreting results 144,
175

Winsorized mean 138

- Y -

Yates correction 322