GraphPad

PRISM®

Version 5.0

Statistics Guide

Harvey Motulsky
President, GraphPad Software Inc.

This Statistics Guide is a companion to GraphPad Prism 5. Available for both Mac and Windows, Prism makes it very easy to graph and analyze scientific data. Download a free demo from www.graphpad.com

The focus of this Guide is on helping you understand the big ideas behind statistical tests, so you can choose a test, and interpret the results. Only about 10% of this Guide is specific to Prism, so you may find it very useful even if you use another statistics program.

The companion Regression Guide explains how to fit curves with Prism,

Both of these Guides contain exactly the same information as the Help system that comes with Prism 5, including the free demo versrion. You may also view the Prism Help on the web at:

http://graphpad.com/help/prism5/prism5help.html

Citation: H.J. Motulsky, Prism 5 Statistics Guide, 2007, GraphPad Software Inc., San Diego CA, www.graphpad.com.

To contact GraphPad Software, email support@graphpad.com or sales@graphpad.com.

Contents

I. Statistical principles

II. Descriptive statistics and normality tests

III. Comparing two groups (t tests ...)

IV. Comparing three or more groups (one-way ANOVA ...)

VII. Survival analysis

VIII. Diagnostic lab analyses

I. Statistical principles

Before you can choose statistical tests or make sense of the results, you need to know some of the basic principles of statistics. We try to make this as painless as possible, by avoiding equations and focussing on the practical issues that often impede effective data analysis.

The big picture

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.

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.

> *It is easy to lie with statistics, but it is easier to lie without them.*
>
> Frederick Mosteller

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 this book 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.

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.

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 in 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 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.

The Gaussian Distribution

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.

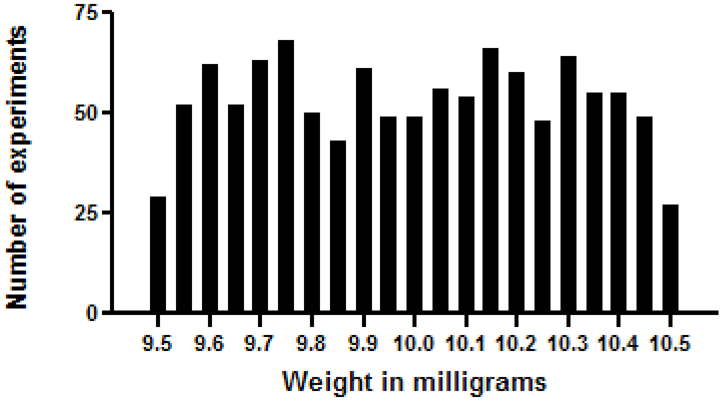
When reading in other books about the Gaussian distribution, two statistical terms might be confusing because they sound like ordinary words:

- 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”).

Origin of the Gaussian distribution

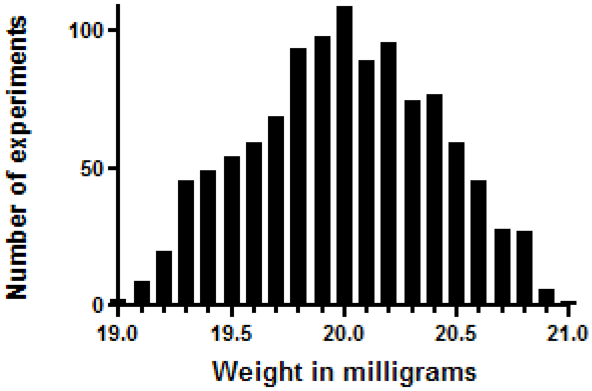
The Gaussian distribution emerges when many independent random factors act in an additive manner to create variability. This is best seen by an example.

Imagine a very simple "experiment". You pipette some water and weigh it. Your pipette is supposed to deliver 10 mL of water, but in fact delivers randomly between 9.5 and 10.5 mL. If you pipette one thousand times and create a frequency distribution histogram of the results, it will look like the figure below.



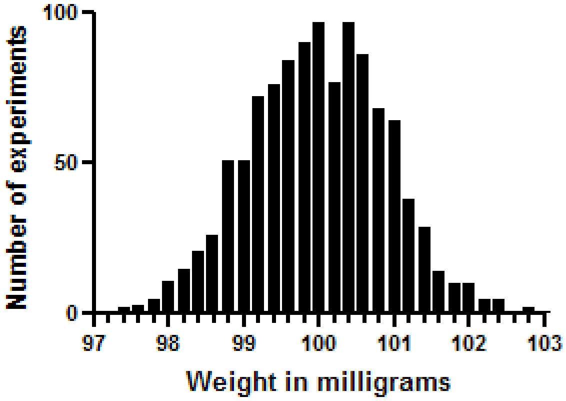
The average weight is 10 milligrams, the weight of 10 mL of water (at least on earth). The distribution is flat, with no hint of a Gaussian distribution.

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.

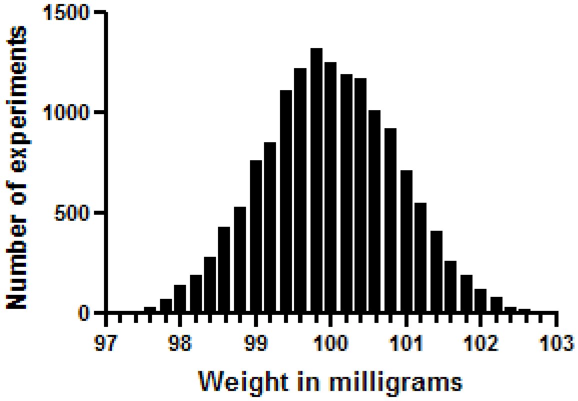


Each pipetting step has a flat random error. Add them up, and the distribution is not flat. For example, you'll get weights near 21 mg only if both pipetting steps err substantially in the same direction, and that is rare.

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.

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.

Standard Deviation and Standard Error of the Mean

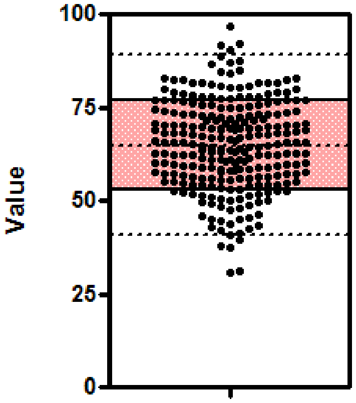
Key concepts: SD

What is the SD?

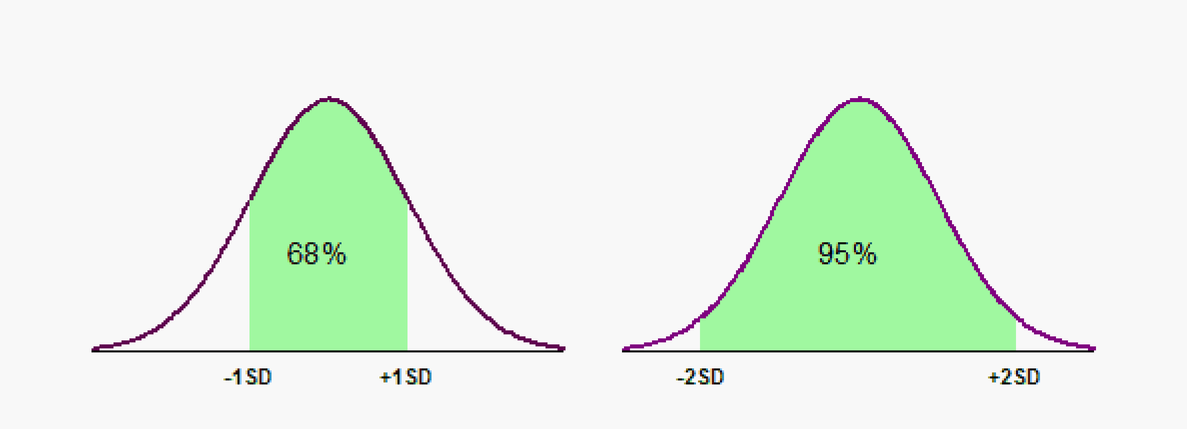
The standard deviation (SD) quantifies variability or scatter, and it is expressed in the same units as your data.

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.

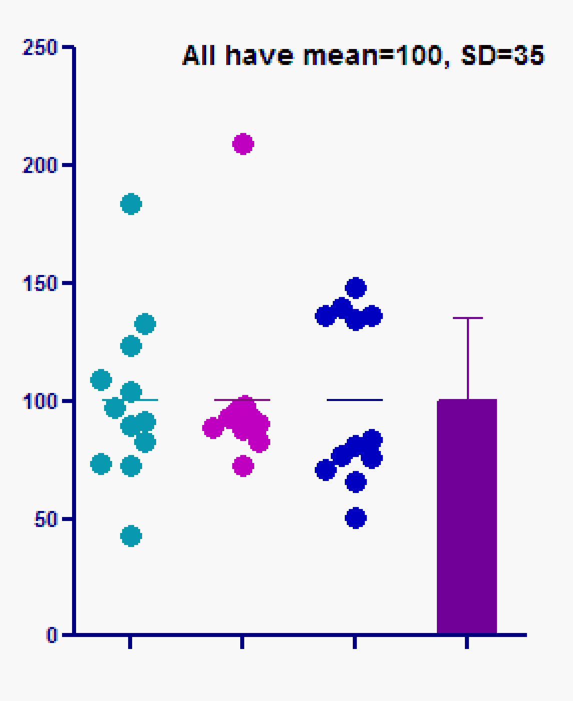


The graph that follows shows the relationship between the standard deviation and a Gaussian distribution. The area under a probability distribution represents the entire population, so the area under a portion of a probability distribution represents a fraction of the population. In the graph on the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

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.

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, then you can 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.

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 21.

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! 24

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.

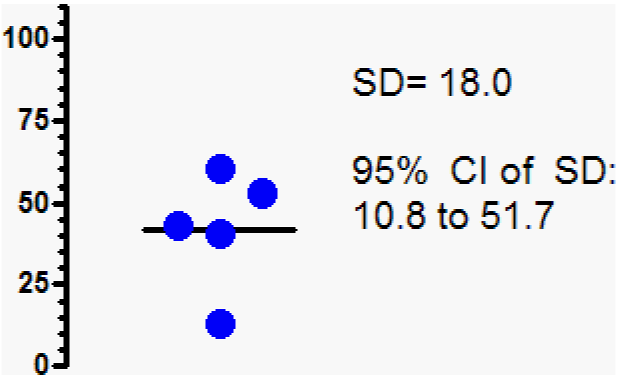
The 95% CI of the SD

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 26 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 12 sampled from a Gaussian 14 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))`

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! 24

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)

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 accurately you know the true mean of the population. It takes into account both the value of the SD and the sample size.
- 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.

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. Consider these points:

- Why show error bars? Consider showing the raw data in a scatter plot. This lets you show more information in the same amount of space.
- If the scatter is due to biological variation, show either raw data or the SD.
- If you don't want to show the scatter, but instead want to show how precisely you have determined the mean, then show the 95% confidence interval of the mean. If all the scatter is due to experimental imprecision (and not biological variation), you can justify putting the focus on the mean, and how precisely it has been determined, and not on the scatter among replicates.
- An alternative way to indicate the precision of a mean (or some other value) is to show the standard error instead of the confidence interval. I prefer confidence intervals, as they are much more straightforward to interpret. But showing the SE in graphs (as error bars) or tables is conventional in many fields of science.
- If in doubt, show raw data.
- If in doubt, but you really want to show an error bar, choose the SD.
- Always state whether your error bars are SD or SEM (or something else) in figure legends or in the methods section of a paper.

When you are trying to emphasize small and unimportant differences in your data, show your error bars as standard errors and hope that your readers think they are standard deviations.

When you are trying to cover-up large differences, show the error bars as standard deviations and hope that your readers think they are standard errors.

Steve Simon (in jest)

http://www.childrens-mercy.org/stats/weblog2005/standarderror.asp

Confidence intervals

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.

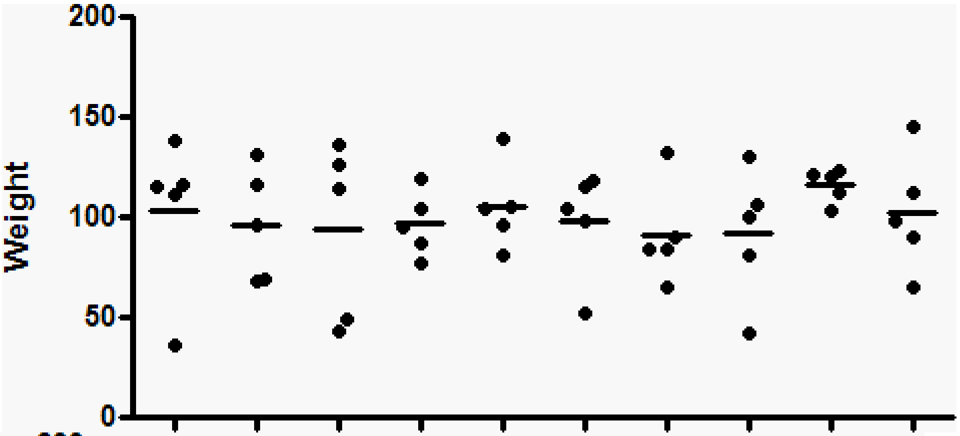
If your sample is small and variable, the sample mean is likely to be quite far from the population mean. If your sample is large and has little scatter, the sample mean will probably be very close to the population mean. Statistical calculations combine sample size and variability (standard deviation)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

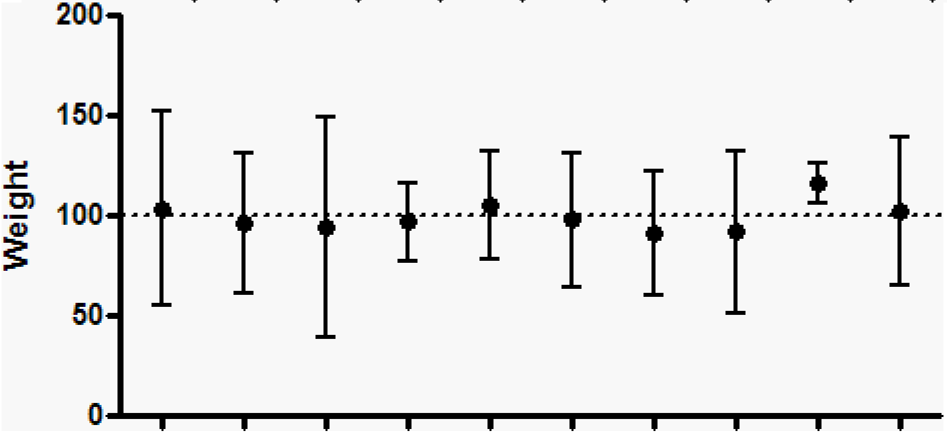
What assumptions are made in interpreting a CI of a mean?

To interpret the confidence interval of the mean, you must assume that all the values were independently 12 and randomly sampled from a population whose values are distributed according to a Gaussian 14 distribution. If you accept those assumptions, there is a 95% chance that the 95% CI contains the true population mean. In other words, if you generate many 95% CIs from many samples, you can expect the 95% CI to include the true population mean in 95% of the cases, and not to include the population mean value in the other 5%.

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

The upper panel below shows ten sets of data (N=5), randomly drawn from a Gaussian distribution with a mean of 100 and a standard deviation of 35. The lower panel shows the 95% CI of the mean for each sample.





Because these are simulated data, we know the exact value of the true population mean (100), so can ask whether or not each confidence interval includes that true population mean. In the data set second from the right in the graphs above, the 95% confidence interval does not include the true mean of 100 (dotted line).

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

How is the confidence interval of a mean computed?

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

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

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

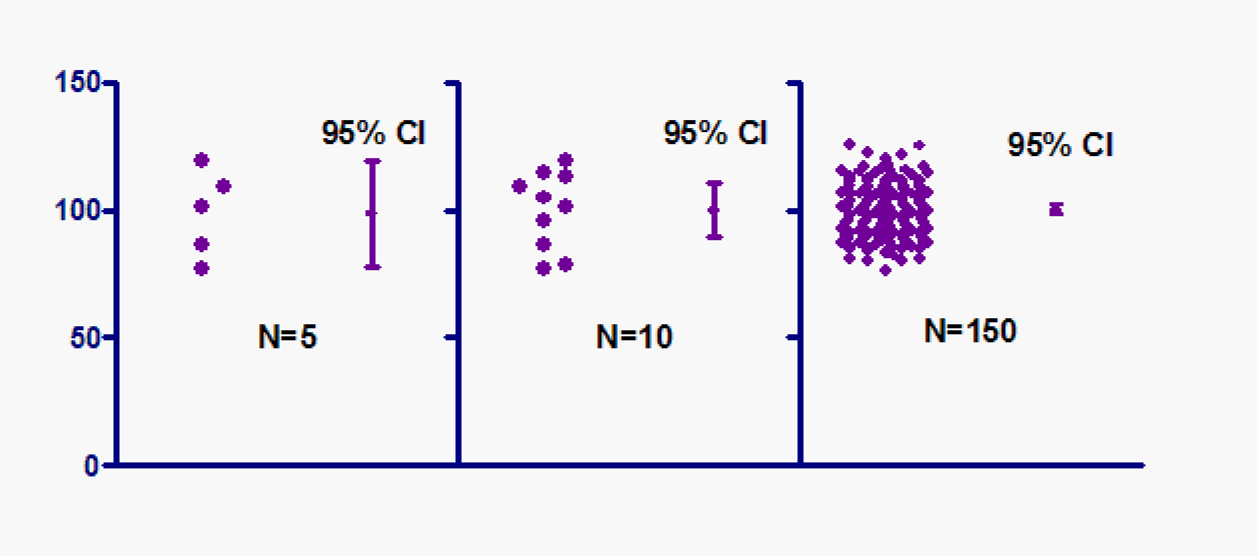
The last line in the table above shows you the equation to use to compute the multiplier in Excel.

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.

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.

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 21. 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.

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 33. 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."

One sided confidence intervals

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

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

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

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

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

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

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

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

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

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

Reference

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

P Values

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?

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 40.

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- and two-sided P values, which means the same thing.

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.

Note: This example is for an unpaired t test 98 that compares the means of two groups. The same ideas can be applied to other statistical tests.

Two-tail P value

The **two-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) you observed in this experiment with either group having the larger mean?

One-tail P value

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?

A one-tail P value is appropriate only when previous data, physical limitations or common sense tell you that a difference, if any, can only go in one direction. The issue is not whether you expect a difference to exist – that is what you are trying to find out with the experiment. The issue is whether you should interpret increases and decreases in the same manner.

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'.

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.

Advice: How to interpret a small 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 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. 40
- 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.

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.

Hypothesis testing and statistical significance

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.

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.

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

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.

A Bayesian perspective on interpreting statistical significance

Interpreting low (and high) P values is tricker 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 45 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.

	Experiments with P<0.05 and...		Fraction of experiments with P<0.05 where drug really works
Prior probability	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.

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.

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

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:~~

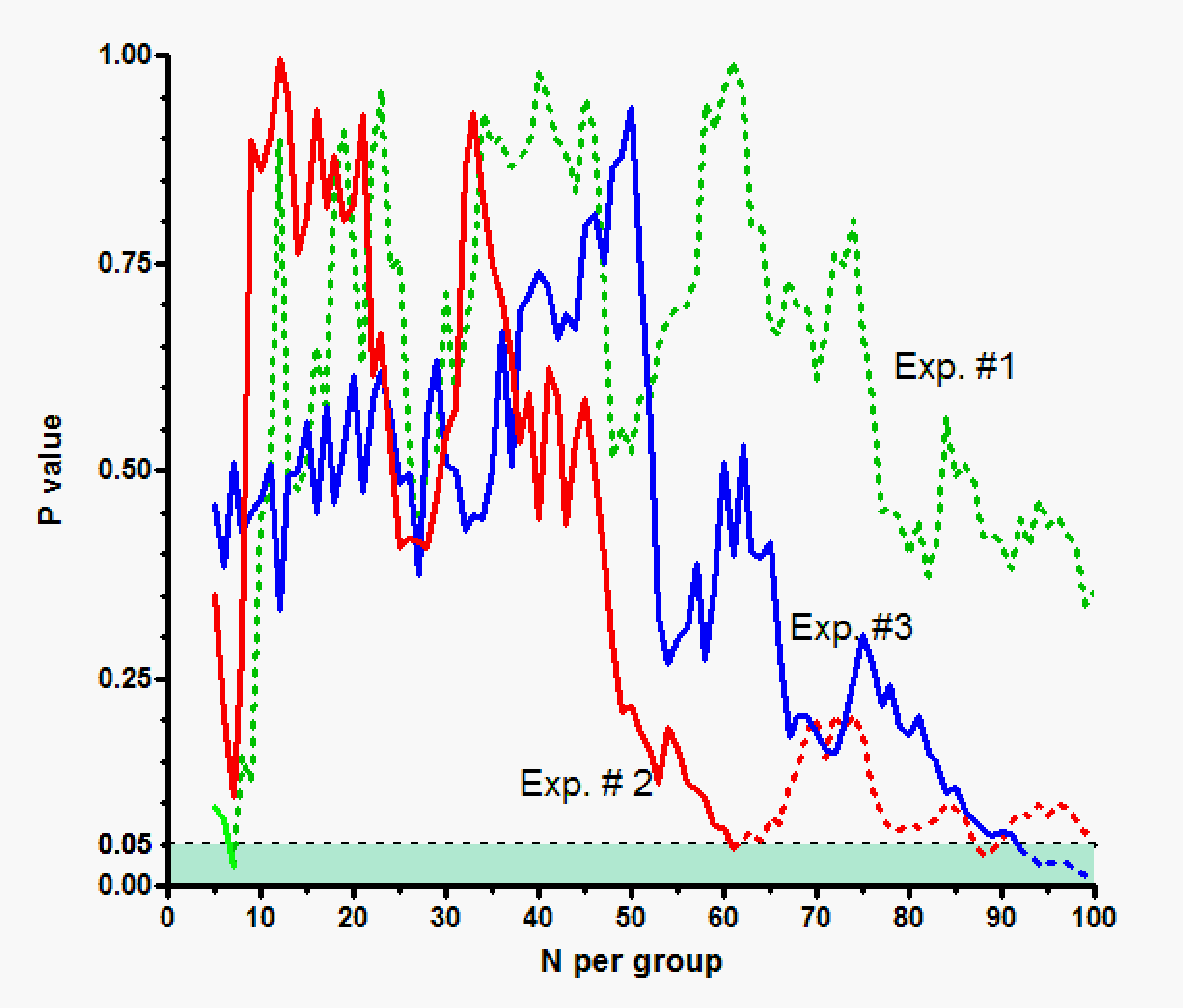
- ~~If the result is not statistically significant, collect some more data, and reanalyze.~~
- ~~If the result is statistically significant, stop the study.~~

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%.

> 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.

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. 47 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.

Note: 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' or 'adaptive' methods in advanced statistics books to learn more.

Statistical power

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 you use as your cutoff, alpha).

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 a 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.
- Ideally, your choice of acceptable power should depend on the consequence of making a Type II error 47.

GraphPad StatMate

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

An analogy to understand statistical power

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.

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.

Type I, II (and III) 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.

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.

Using power to evaluate 'not significant' results

Example data

Motulsky et al. asked whether people with hypertension (high blood pressure) had altered numbers of alpha$_2$-adrenergic receptors on their platelets (Clinical Science 64:265-272, 1983). There are many reasons to think that autonomic receptor numbers may be altered in hypertension. We studied platelets because they are easily accessible from a blood sample. The results are shown here:

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

The two means were almost identical, and a t test gave a very high P value. We concluded that the platelets of people with hypertension do not have an altered number of alpha_2 receptors.

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.

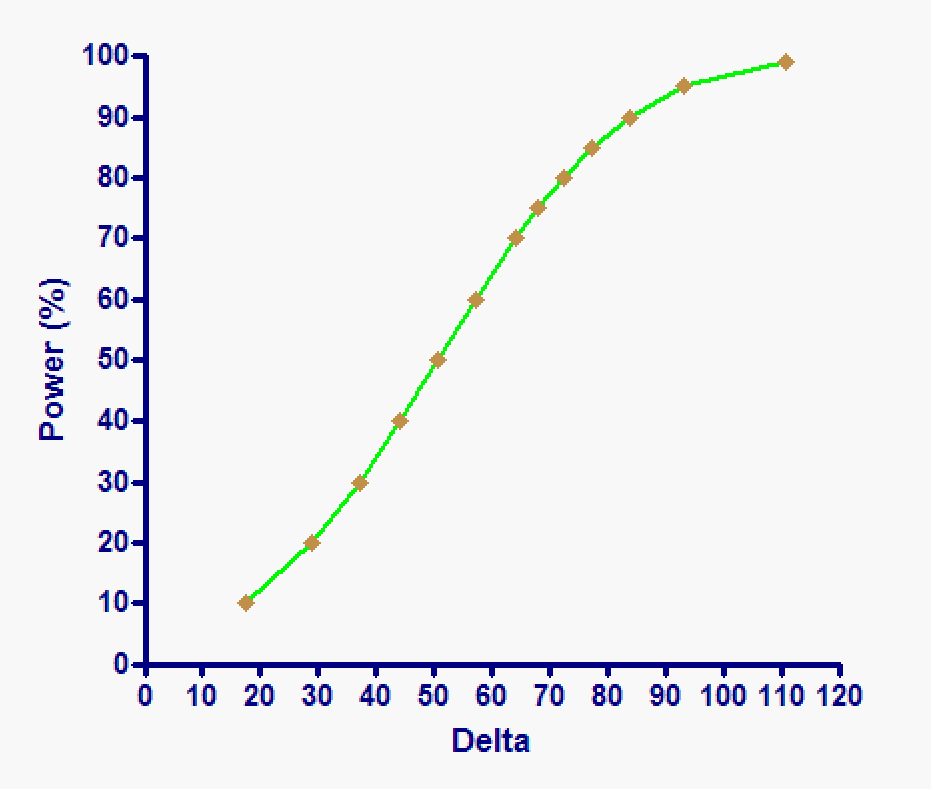
The interpretation of the confidence interval must be in a scientific context. Here are two very different approaches to interpreting this confidence interval.

- 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 50.

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.

Advice: Don't compute the power to detect the difference actually observed

It is never possible to just ask "what is the power of this experiment?". Rather, you must ask "what is the power of this experiment to detect an effect of some specified size?". Which effect size should you use? How large a difference should you be looking for? It only makes sense to do a power analysis when you think about the data scientifically. It isn't purely a statistical question, but rather a scientific one.

Some programs try to take the thinking out of the process by computing only a single value for power. These programs compute the power to detect the effect size (or difference, relative risk, etc.) actually observed in that experiment. The result is sometimes called **observed power**, and the procedure is sometimes called a **post-hoc power analysis** or retrospective power analysis.

Prism does not do this. Here is the reason:

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. You already know that the difference was not statistically significant, and now you know that the power of the study to detect that particular difference is low. Not helpful.

What would be helpful is to know the power of the study to detect some hypothetical difference that you think would have been scientifically or clinically worth detecting. GraphPad StatMate can help you make these calculations (for some kinds of experimental designs) but you must decide how large a difference you would have cared about. That requires a scientific judgment, that cannot be circumvented by statistical computations.

These articles listed below discuss the futility of post-hoc power analyses.

References

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

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.

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

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.

Multiple comparisons

Key concepts: multiple comparisons

Review of the meaning of P value and alpha

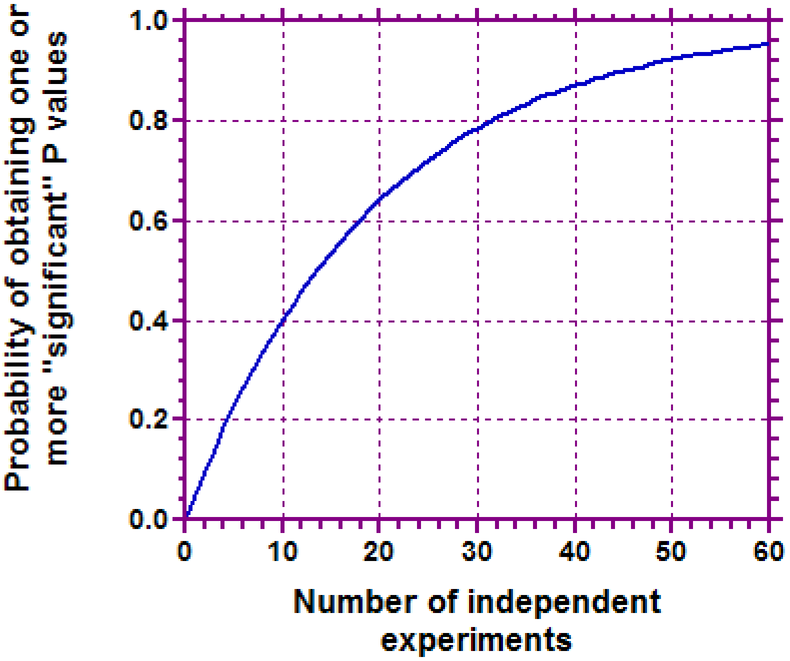
Interpreting an individual P value is easy. Assuming the null hypothesis is true, the P value is the probability that random subject selection alone would result in a difference in sample means (or a correlation or an association...) at least as large as that observed in your study.

Alpha is a threshold that you set. If the P value is less than alpha, you deem the comparison "statistically significant'. In other words, 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 generate more than one P value. Some studies in fact generate hundreds of P values.

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 on the Y axis is computed from N on the X axis using this equation: $100(1.00 - 0.95^N)$.



Remember the unlucky number 13. If you perform 13 independent experiments, your chances are about 50:50 of obtaining at least one 'significant' P value (<0.05) just by chance.

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”. To keep the overall chance of a false “significant” conclusion at 5%, you need to lower the threshold for each t test to 0.0170. 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%. Unless you correct for the multiple comparisons, it is easy to be fooled by the results. Now lets say you test ten different enzymes, at three time points, in two species, with four pre treatments. You can make lots of comparisons, and you are almost certain to find that some of them are 'significant', even if really all null hypotheses are true.

You can only account for multiple comparisons when you know about all the comparisons made by the investigators. If you report only “significant” differences, without reporting the total number of comparisons, others will not be able to properly evaluate your results. Ideally, you should plan all your analyses before collecting data, and then report all the results.

Distinguish between studies that test a hypothesis and studies that generate a hypothesis. Exploratory analyses of large databases can generate hundreds or thousands of P values, and scanning these can generate intriguing research hypotheses. But you can't test hypotheses using the same data that prompted you to consider them. You need to test your new hypotheses with fresh data.

Approaches to dealing with multiple comparisons

Perspective on dealing with multiple comparisons at once

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 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 the 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

Three approaches can be used to deal with multiple comparisons:

Approach 1: Don't correct for multiple comparisons

Use the standard definition of 'significance' so you expect the ratio of A/(A+B) to equal alpha, which is usually 5%. In other words, if the null hypothesis of no difference is in fact true, there is a 5% chance that you will mistakenly conclude that the difference is statistically significant. This 5% value applies to each comparison separately, so is per comparison error rate.

When using this approach, you have to beware of over interpreting a 'statistically' significant result. You expect 52 a significant result in 5% of comparisons where the null hypothesis is true. If you perform many comparisons, you would be surprised if none of the comparisons resulted in a 'statistically significant' conclusion.

This approach is sometimes called "Planned comparisons 160".

Approach 2: Correct for multiple comparisons

With this approach, you set a stricter threshold for significance, such that alpha is the chance of obtaining one or more 'significant' conclusions if the *all* the null hypotheses are true. In the table above, alpha is the probability that A will be greater than 0. If you set alpha to the usual value of 5%, this means you need 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 advantage of this approach is that you are far less likely to be mislead by false conclusions of 'statistical significance'. The disadvantage is that you need to use a stricter threshold of significance, so will have less power to detect true differences.

Approach 3: False Discovery Rate

The two approaches already discussed ask: If the null hypothesis is true what is the chance of getting "significant" results? The False Discovery Rate (FDR) answers a different question: If the comparison is "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 40. But if you have many comparisons, simple methods let you answer that question (at least approximately). In the table, above the False Discovery rate is the ratio A/(A+C). This ratio is sometimes called Q. If Q is set to 10%, that means the threshold for dividing 'significant' from not significant comparisons is established so we expect 90% of the 'significant' results to truly reflect actual differences, while 10% to be false positives.

Prism does not use the concept of False Discovery Rate, except indirectly as part of our method to define outliers in nonlinear regression.

Testing for equivalence

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 57.

Testing for equivalence with confidence intervals or P values

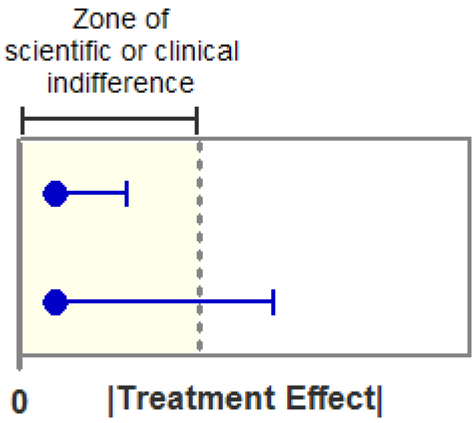
Before you can test for equivalence, you first have to define a range of treatment effects that you consider to be scientifically or clinically trivial. You must set this range based on scientific or clinical judgment -- statistical analyses can't help.

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*.

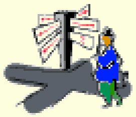
The two approaches are equivalent

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!

Outliers

What is an outlier?

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 have 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!

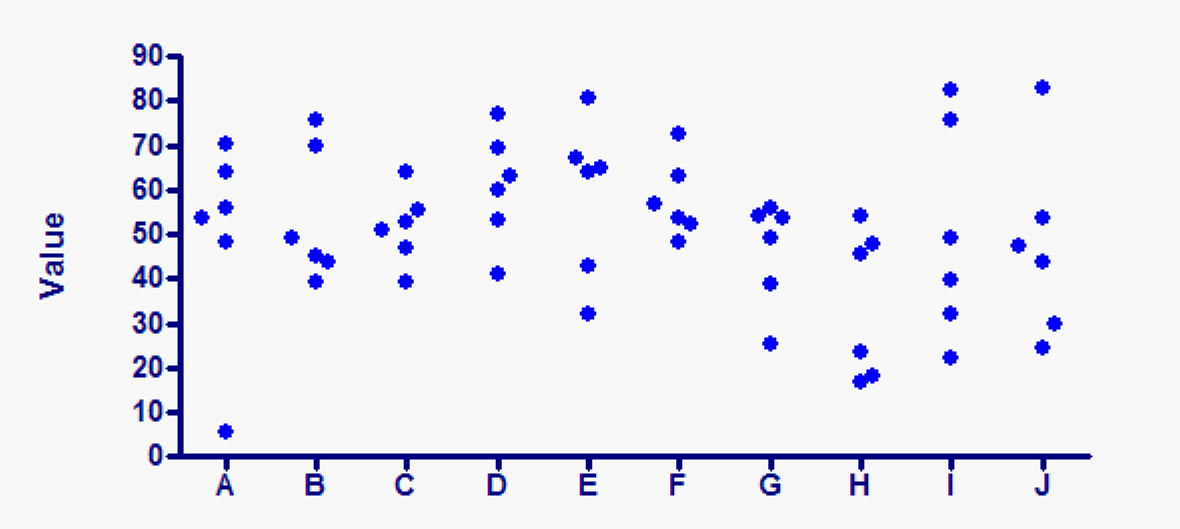
If you answered “no” to all three questions, you are left with two possibilities.

- 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.

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.

Detecting outliers with Grubbs' test

How can outlier tests help?

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.

Grubbs' outlier test

Grubbs' method for assessing outliers is particularly 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. For this test, calculate the mean and SD from all the values, including the outlier. Calculate Z for all the values, but only perform the Grubbs' test with the most extreme outlier, the value that leads to the largest value of Z.

Since 5% of the values in a Gaussian population are more than 1.96 standard deviations from the mean, your first thought might be to conclude that the outlier comes from a different population if Z is greater than 1.96. But since the outlier affects the computation of the SD, this rule isn't right, so special tables are needed.

Grubbs test is not implemented in Prism, but we do make it available as a free web calculator (graphpad.com/quickcalcs).

The most that Grubbs' test (or any outlier test) can do is tell you that a value is unlikely to have come from the same Gaussian population as the other values in the group. You then need to decide what to do with that value. I would recommend removing significant outliers from your calculations in situations where experimental mistakes are common and biological variability is not a possibility. When removing outliers, be sure to document your decision. Others feel that you should never remove an outlier unless you noticed an experimental problem. Beware of a natural inclination to remove outliers that get in the way of the result you hope for, but to keep outliers that enhance the result you hope for.

If you decide to remove the outlier, you then may be tempted to run Grubbs' test again to see if there is a second outlier in your data. This is not valid unless you use an extension of the test designed to detect several outliers in one sample. See the first reference below for details.

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.

Statistical tests that are robust to the presence of outliers

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.

Most nonparametric tests compare the distribution of ranks. This makes the test robust because the largest value has a rank of 1, 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).

Dealing with outliers in Prism

Identifying and excluding outliers when fitting curves with nonlinear regression

Prism can automatically identify, and ignore, outliers when fitting curves with nonlinear regression. Read about how this is useful, when it should be avoided, and how it works.

Excluding data in Prism

While Prism does not identify outliers automatically, it does let you manually exclude values you consider to be outliers. From a data table, select the outlier(s), drop the Edit menu, and click the Exclude button ~~1.23~~.

Prism will display excluded values in blue italics and append an asterisk. Such values will be ignored by all analyses and graphs. We suggest that you document your reasons for excluding values in a floating note.

Note that sometimes outliers provide very useful information. Think about the source of variation before excluding an outlier.

The ROUT method for identifying outliers

When you use Prism to fit curves with nonlinear regression, you can choose automatic outlier rejection using a method we developed (reference below). This is a choice on the first tab of the nonlinear regression dialog. Because this method combines robust regression and outlier detection, we call it the ROUT method.

You can also use this method to detect one or more outliers in a stack of data in a Column table. To do this you have to convert your table to be an XY table, and then use nonlinear regression to fit a straight line model, constraining the slope to equal 0.0. In this case, Prism fits only one parameter which is labeled intercept, but is really the mean. Follow these steps:

1. Go to the Format Data Table dialog and change the table so it is an XY table.
2. In that dialog, choose to make X a series so you don't have to bother entering data. It doesn't matter what starting value and increment you use, as the X values will be ignored.
3. Click Analyze and choose Nonlinear regression from the list of XY analyses.
4. On the fit tab, choose the classic model "Polynomial: First order (straight line) and choose the

fitting method to be "automatic outlier elimination"

5. Go to the Constrain tab, and constrain slope to equal 0.
6. Go to the Weights tab and check the value of the ROUT coefficient. We recommend setting it to 1%. Increasing the value tells Prism to be more aggressive when defining a point to be an outlier, and decreasing the value makes Prism be more conservative about defining points to be outliers.
7. Click OK. Go to the results sheet.
8. At the bottom of the tabular results, Prism will report how many outliers it excluded. You can view a table of these values as a separate results page.
9. If you want to exclude these values from graphs or t tests or anovas, go to the data table, select the questionable values, right click and choose Exclude. The values will appear in blue italics, and will be ignored by all graphs and analyses.

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.

Nonparametric tests

Key concepts: Nonparametric tests

ANOVA, t tests, and many statistical tests assume that you have sampled data from populations that follow a Gaussian 14 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.

Advice: Don't automate the decision to use a nonparametric test

It sounds so simple. 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 follow 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 significantly from a Gaussian distribution, you should 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.
- 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.

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.

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 45 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 118 (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.

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.

Summary

	Large samples (> 100 or so)	**Small samples (<12 or so)**
Parametric tests on nongaussian data	OK. Tests are robust.	Results are misleading. Tests not robust.
Nonparametric test on Gaussian data.	OK. Almost as powerful as parametric tests.	Results are misleading. Tests are not powerful.
Normality test	Useful.	Not useful. Little power to discriminate between Gaussian and non-Gaussian populations.

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 206.

- **Small samples.** If you have tiny samples (a few subjects in each group), the nonparametric tests have little or no power 65 to find a significant difference.

- **Normality tests** should not be used 64 to automatically decide whether or not to use a nonparametric test. But they can help you make the decision.

- You really should choose your statistical test as part of the experimental design. If you try this test, then that test, until you get a result you like, you are likely to be mislead.

II. Descriptive statistics and normality tests

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. Assess the likelihood that the values were sampled from a Gaussian distribution. Test whether the mean (or median) differs significantly from a hypothetical value.

Column statistics

.

How to: Column statistics

The column statistics analysis computes descriptive statistics (and normality tests) for each data set. If your goal is simply to graph each column mean with SD or SEM, you don't need this analysis. Prism can graph error bars automatically when it creates a graph from a column table. Look elsewhere 92 if you want to compute the mean and SD or SEM of side by side replicates.

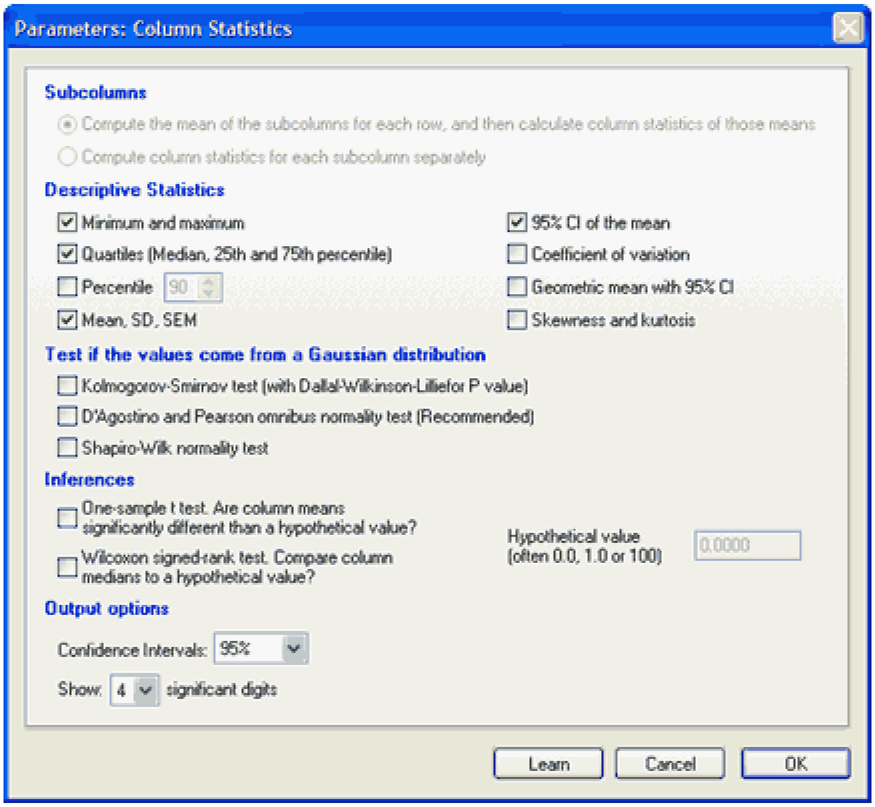
1. Entering data for column statistics

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: Making a column bar graph.

You can also choose the column statistics analysis from data entered onto XY or Grouped data tables.

2. Choose the column statistics analysis

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



Parameters: Column Statistics

Subcolumns

● Compute the mean of the subcolumns for each row, and then calculate column statistics of those means

○ Compute column statistics for each subcolumn separately

Descriptive Statistics

☑ Minimum and maximum

☑ Quartiles (Median, 25th and 75th percentile)

□ Percentile 90

☑ Mean, SD, SEM

☑ 95% CI of the mean

□ Coefficient of variation

□ Geometric mean with 95% CI

□ Skewness and kurtosis

Test if the values come from a Gaussian distribution

□ Kolmogorov-Smirnov test (with Dallal-Wilkinson-Lilliefor P value)

□ D'Agostino and Pearson omnibus normality test (Recommended)

□ Shapiro-Wilk normality test

Inferences

□ One-sample t test. Are column means significantly different than a hypothetical value?

□ Wilcoxon signed-rank test. Compare column medians to a hypothetical value?

Hypothetical value (often 0.0, 1.0 or 100) 0.0000

Output options

Confidence Intervals: 95%

Show: 4 significant digits

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 74, median 72, SD 18, SEM 22, confidence interval 26, coefficient of variation 75, geometric mean 72, skewness and kurtosis 76.

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. 76

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 77 to test that assumption. Or choose the nonparametric Wilcoxon signed-rank test 78.

Analysis checklist: Column statistics

Descriptive statistics

Value	Meaning
Minimum	The smallest value.
25th percentile 74	25% of values are lower than this.
Median 72	Half the values are lower; half are higher.
75th percentile 74	75% of values are higher than this.
Maximum	The largest value.
Mean 72	The average.
Standard Deviation 18	Quantifies variability or scatter.
Standard Error of Mean 22	Quantifies how precisely the mean is known.
95% confidence interval 26	Given some assumptions, there is a 95% chance that this range includes the true overall mean.
Coefficient of variation 75	The standard deviation divided by the mean.
Geometric mean 72	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).
Skewness 76	Quantifies how symmetrical the distribution is. A distribution that is symmetrical has a skewness of 0.
Kurtosis 76	Quantifies whether the shape of the data distribution matches the Gaussian distribution. A Gaussian distribution has a kurtosis of 0.

Normality tests

Normality tests 76 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 64.

Inferences

A one-sample t test 77 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 35 (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 78 is similar, but does not assume a Gaussian distribution. It asks whether the median of each column differs from a hypothetical median you entered.

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

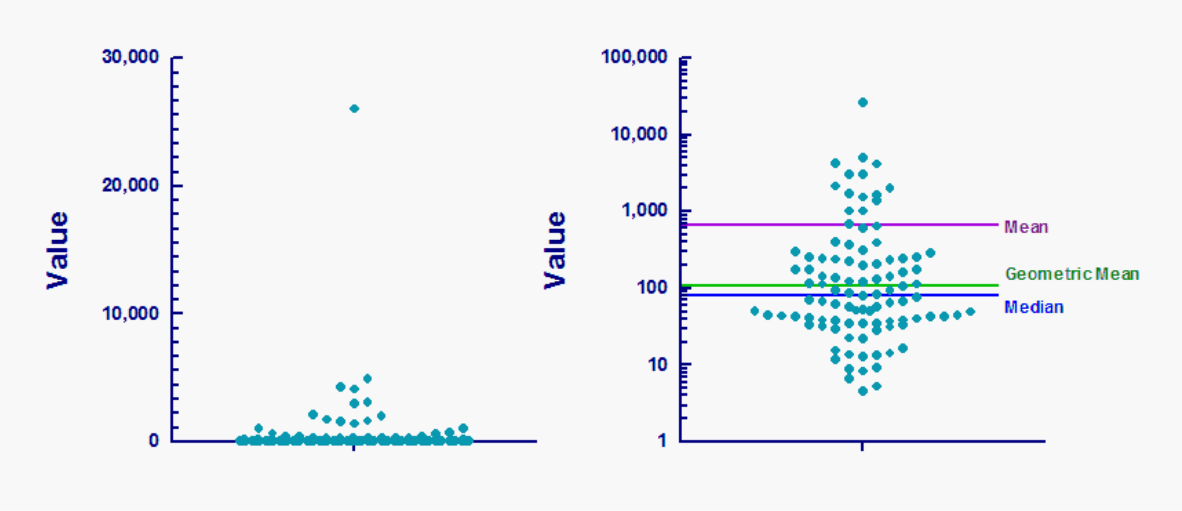
The median is the 50th percentile. Half the values are higher than the median, and half are lower.

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.

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

To compute a percentile value, first compute P*(N+1)/100, where P is the percentile value (i.e. 25, 50, or 75) and N is the number of values in the data set. The result is the rank that corresponds to that percentile value. If there are 68 values, the 25th percentile corresponds to a rank equal to 25*(68+1)/100 =17.25. Therefore, the 25th percentile lies between the value of the 17th and 18th value (when ranked from low to high). But where exactly? There is no clear answer, so not all programs compute the percentile the same way. Prism 5 computes the 25th percentile in this example as the value at 25% of the distance from the 17th to 18th value (earlier versions of Prism averaged the 17th and 18th values).

Because different methods for computing the 25th and 75th percentiles give different results with small data sets, we suggest that you only report the 25th and 75th percentiles for large data sets (N>100 is a reasonable cut off). For smaller data sets, we suggest showing a column scatter graph that shows every value.

Note that there is no ambiguity about how to compute the median. All programs do it the same way.

Interpreting results: Variance and coefficient of variation (CV)

Variance

The variance equals the SD squared, and therefore is expressed in the units of the data squared. Because these units are usually impossible to think about, most scientists avoid reporting the variance of data. Mathematicians like to think about variances because you can partition variances into different components -- the basis of ANOVA. In contrast, it is not correct to partition the SD into components.

Our advice: Leave variances to the mathematicians, and report and think only about 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.

Interpreting results: Skewness and kurtosis

Skewness quantifies how symmetrical the distribution is. A distribution that is symmetrical has a skewness of 0. If the skewness is positive, that means the right tail is 'heavier' than the left tail. If the skewness is negative, then the left tail of the distribution is dominant.

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, and a more peaked distribution has a positive kurtosis.

Interpreting results: Normality test

This section explains how normality test can assess whether your data are likely to have been sampled from a Gaussian distribution. Look elsewhere if you want to plot a frequency distribution 80 and for help on deciding when to use nonparametric tests 67.

Prism offers three normality tests as part of the Column Statistics analysis. These tests require seven or more values, and help you assess whether those values were sampled from a Gaussian 14 distribution.

Interpreting a normality test

The P value from a normality test 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).

How useful are normality tests?

Normality tests are less useful than some people guess. 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.

Normality tests can help you decide when to use nonparametric tests, but the decision should not be an automatic one 64.

How the normality tests work

We recommend relying on the **D'Agostino-Pearson** normality test. It first computes the skewness and kurtosis 76 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 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 statement[1]:*"The Kolmogorov-Smirnov test is only a historical curiosity. It should never be used."*[1]

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 (Am. Statistician, 40:294-296, 1986). Since that method is only accurate with small P values, Prism simply reports “P>0.10” for large P values.

Reference

[1] RB D'Agostino, "Tests for Normal Distribution" in *Goodness-Of-Fit Techniques* edited by RB D'Agostino and MA Stepenes, Macel Decker, 1986.

Interpreting results: One-sample t test

The one sample t test compares the mean of a column of numbers against a hypothetical mean. Don't confuse it with the unpaired t test 98 and paired t test 109 (which compare the means of two groups).

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 36, 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 35 (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 35.

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 12. 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).

Interpreting results: Wilcoxon signed rank test

> The Wilcoxon signed rank test is a nonparametric test that compares the median of a column of numbers against a hypothetical median. Don't confuse it with the Wilcoxon matched pairs test 126 which compares medians of two paired groups).

The nonparametric 64 Wilcoxon signed rank test compares the median of a single column of numbers against a hypothetical median.

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 35, 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 36, 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 12. 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 mean, 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 small samples, Prism computes an exact P value. With larger samples, Prism uses an approximation that is quite accurate.

Frequency Distributions

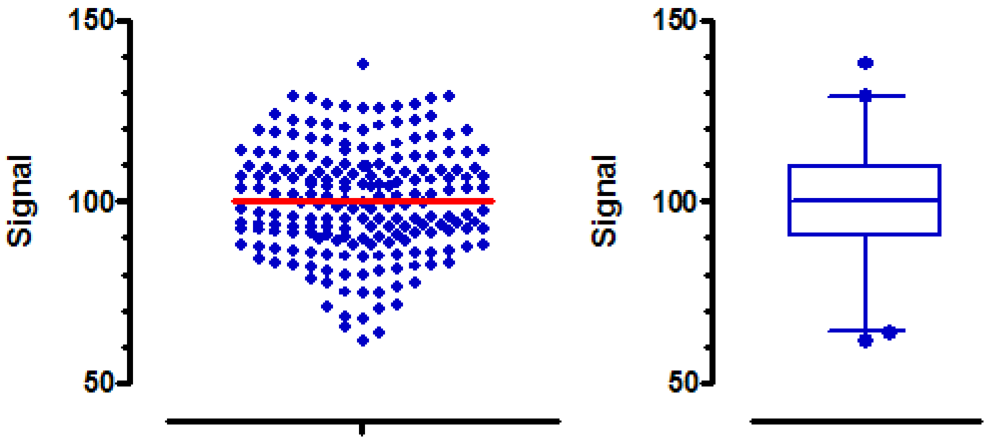
Visualizing scatter and testing for normality without a frequency distribution

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 76 as part of the column statistics analysis. You don't have to create a frequency distribution, and then fit a Gaussian distribution.

How to: Frequency distribution

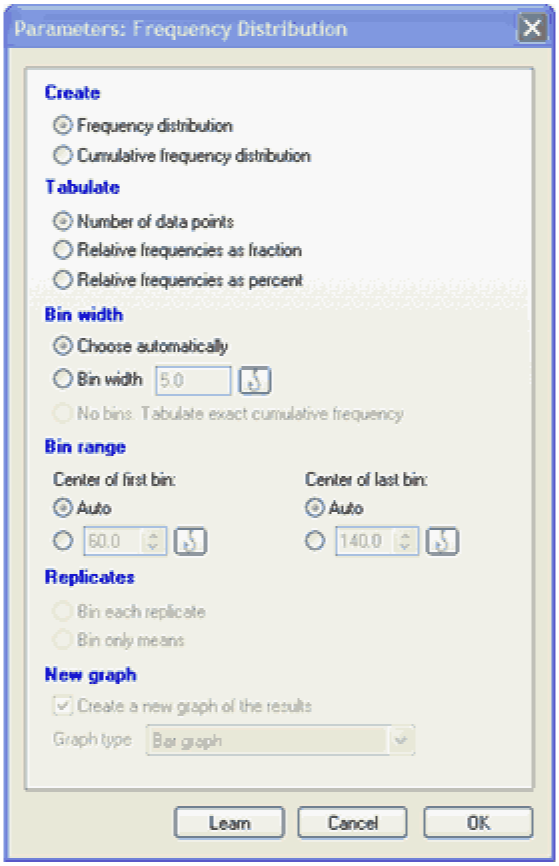
This section explains how to generate a frequency distribution (table and graph) from raw data. Before creating a frequency distribution, read about ways to visualize scatter and test for normality 80 without creating a frequency distribution.

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 for frequency distributions.

2. Choose the analysis

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


Parameters: Frequency Distribution

Create
- ● Frequency distribution
- ○ Cumulative frequency distribution

Tabulate
- ● Number of data points
- ○ Relative frequencies as fraction
- ○ Relative frequencies as percent

Bin width
- ● Choose automatically
- ○ Bin width 5.0
- ○ No bins. Tabulate exact cumulative frequency

Bin range

Center of first bin:
- ● Auto
- ○ 60.0

Center of last bin:
- ● Auto
- ○ 140.0

Replicates
- ○ Bin each replicate
- ○ Bin only means

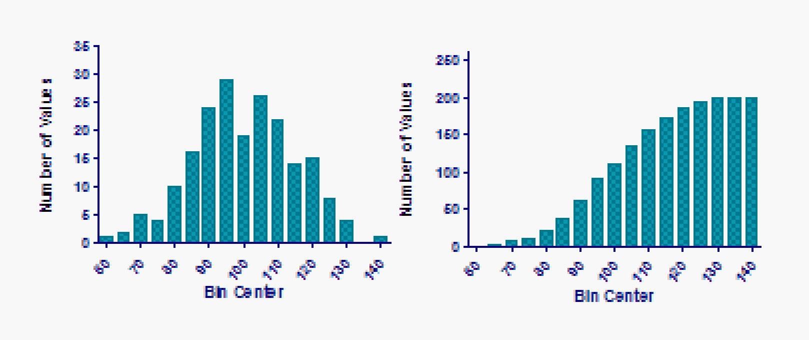
New graph
- ☑ Create a new graph of the results

Graph type: Bar graph

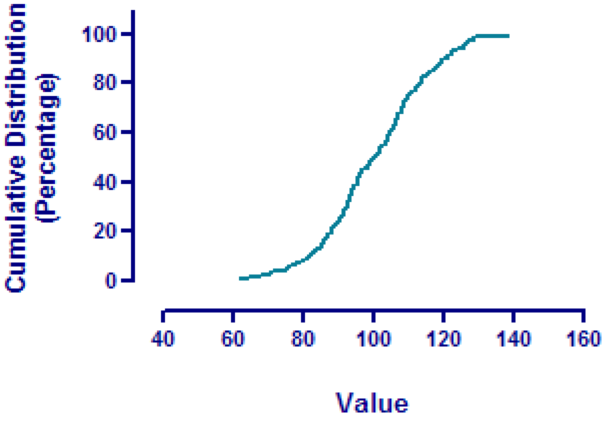
3. Choose analysis options

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.



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

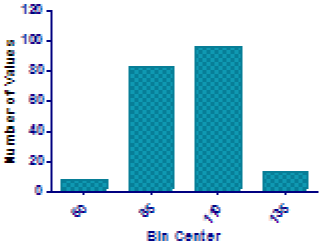
If you chose a cumulative frequency distributions, we suggest that you choose to create an exact distribution. In this case, you don't choose a bin width as each value is plotted individually.

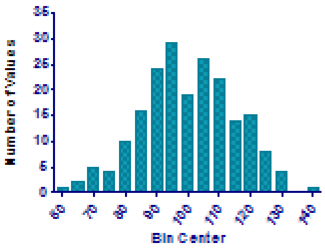
To create an ordinary frequency distribution, you must decide on a bin width. If the bin width is too large, there will only be a few bins, so you will not get a good sense of how the values distribute. If the bin width is too low, many bins might have only a few values (or none) and so the number of values in adjacent bins can randomly fluctuate so much that you will not get a sense of how the data are distributed.

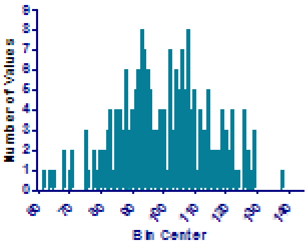
How many bins do you need? Partly it depends on your goals. And partly it depends on sample size.

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.







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 84.

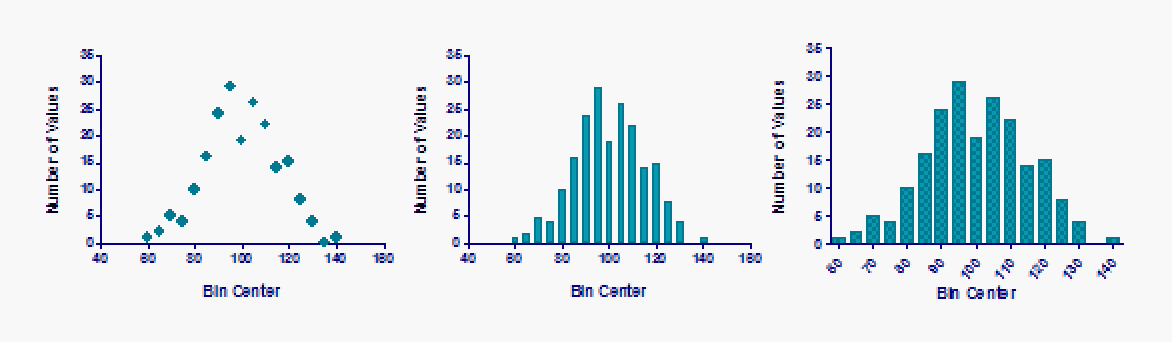
Prism can only make frequency distributions from numerical data. It can handle categorical data, but only if the categories are entered as values.

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

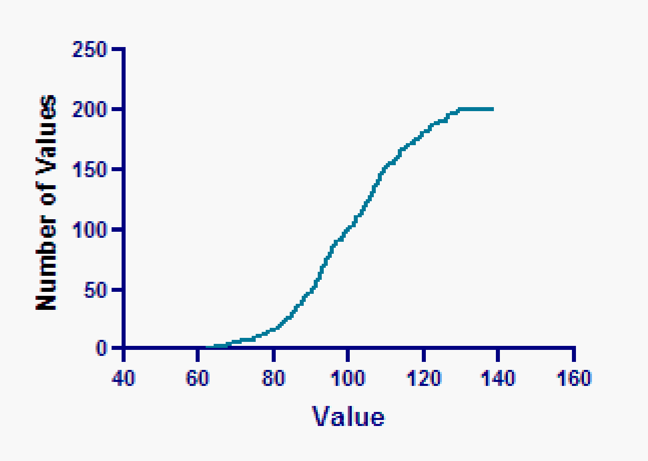
If you don't create a cumulative distribution, Prism gives you three choices illustrated below: XY graph with points, XY graph with spikes (bars). or a bar graph



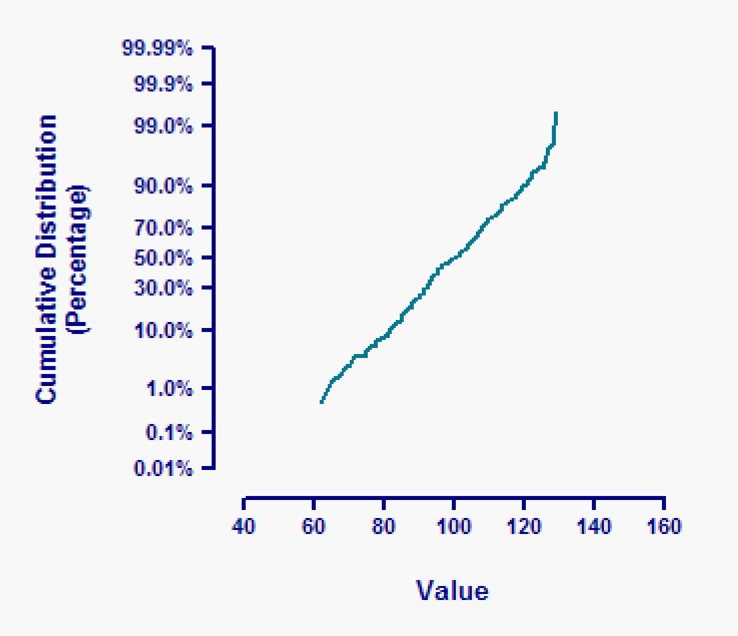
The last two graphs look very similar, but the graph on the right is a bar graph, while the one in the middle is an XY graph plotting bars or spikes instead of symbols. The graph in the middle has X values so you can fit a Gaussian distribution 86 to it. The graph on the right has no X values (just category 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.

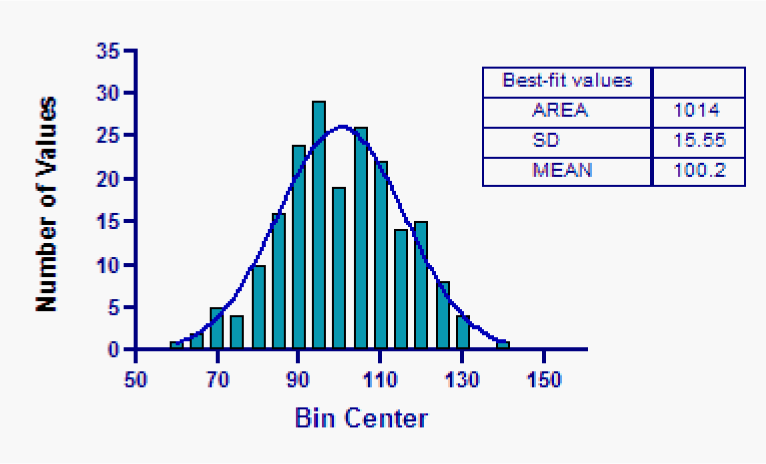
Fitting a Gaussian distribution to a frequency distribution

Why fit a Gaussian distribution to your data?

Does you data follow a Gaussian distribution? One way to answer that question is to perform a normality test 69 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.



Best-fit values	
AREA	1014
SD	15.55
MEAN	100.2

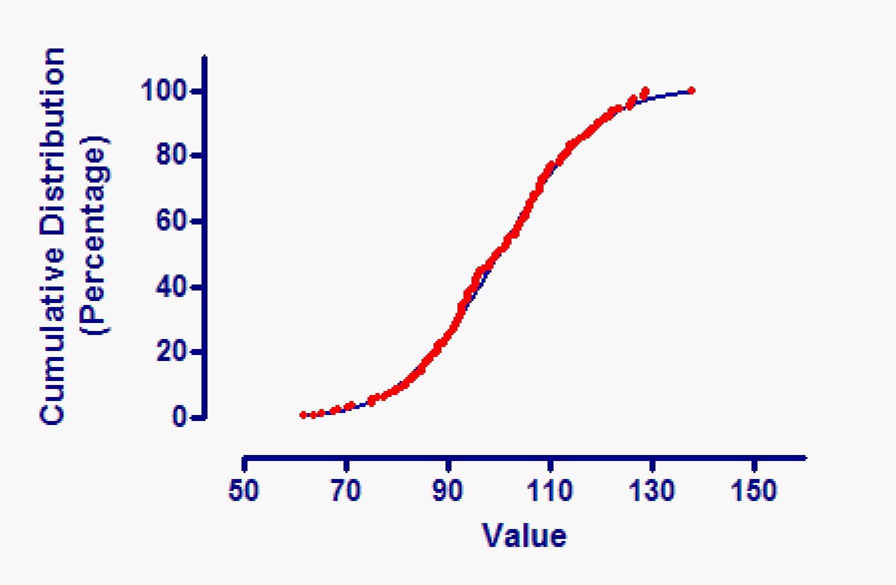
The results depend to some degree on which value you picked for bin width, so we recommend fitting the cumulative distribution as explained below.

Fitting a cumulative Gaussian distribution

The cumulative Gaussian distribution has a sigmoidal shape.

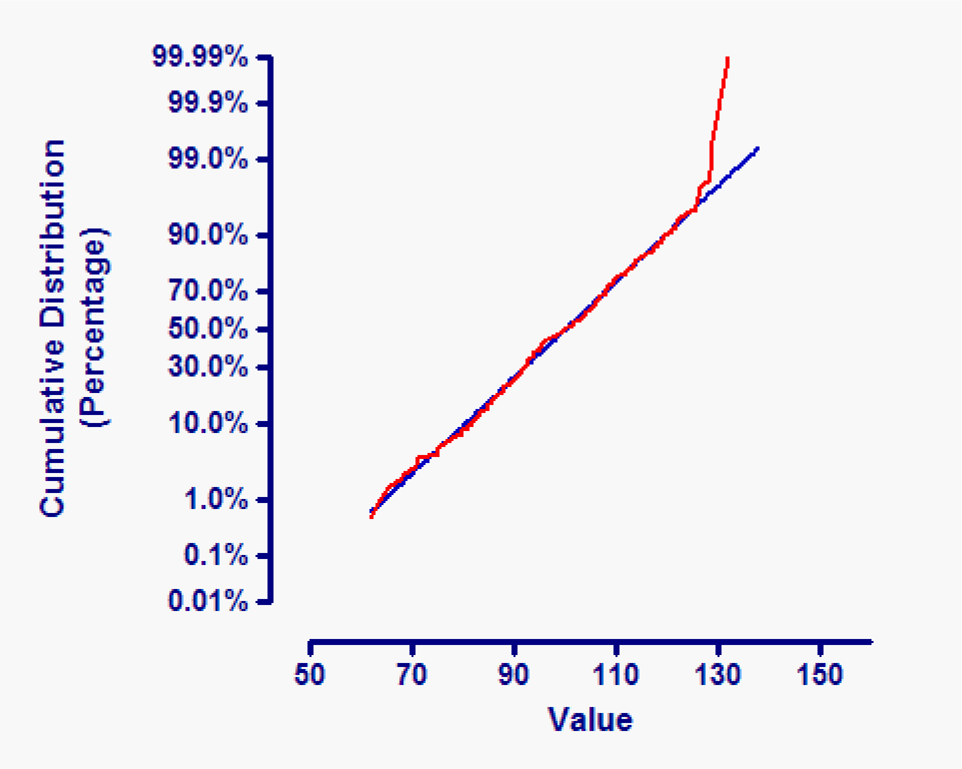
To fit the frequency distribution, you have to specify that the distribution be plotted as an XY plot, so the bin centers are X values (and not just row labels). Then click Analyze, choose nonlinear regression, and choose the one of the cumulative Gaussian models from the selection of Gaussian models.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

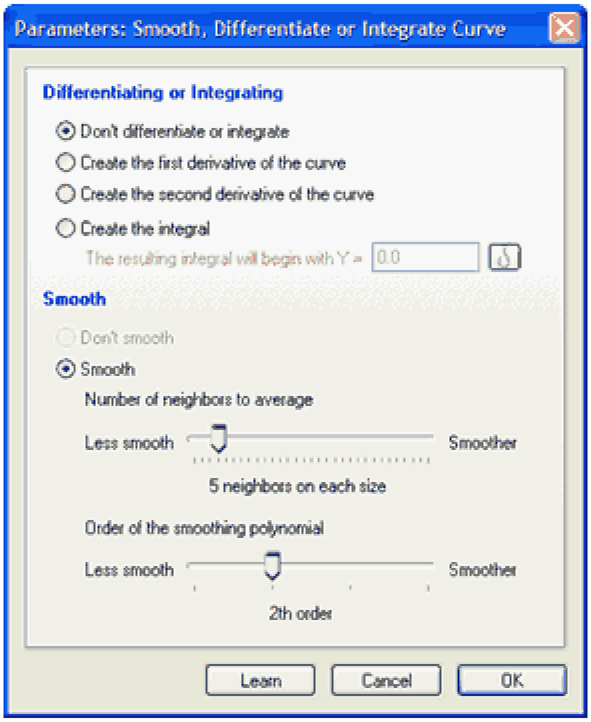
Below, the same graph is plotted using a probability Y axis. To do this, double-click on the Y axis to bring up the Format Axis dialog, drop down the choices for scale in the upper right corner, and choose "Probability (0..100%). The cumulative Gaussian distribution is linear when plotted on probability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



Describing curves

Smoothing, differentiating and integrating curves

A single Prism analysis smooths a curves and also (optionally) converts the resulting curve to its derivative or integral.



Parameters: Smooth, Differentiate or Integrate Curve

Differentiating or Integrating

● Don't differentiate or integrate
○ Create the first derivative of the curve
○ Create the second derivative of the curve
○ Create the integral
The resulting integral will begin with Y = 0.0

Smooth

○ Don't smooth
● Smooth
Number of neighbors to average
5 neighbors on each size
Order of the smoothing polynomial
2th order

Finding the derivative or integral of a curve

Note: 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 of that series of points. But if you give Prism an equation, it cannot compute a new equation that defines the derivative.

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 **integral** is the cumulative area under the curve. The integral at any value X equals the area of the curve for all values less than X.

This analysis integrates a curve, resulting in another curve showing cumulative area. Another Prism analysis computes a single value for the area under the curve 89.

Prism uses the trapezoid rule 89 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 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.

References

Savitsky and Golay (Analytical Chemistry, 36:1627-1639, 1964).

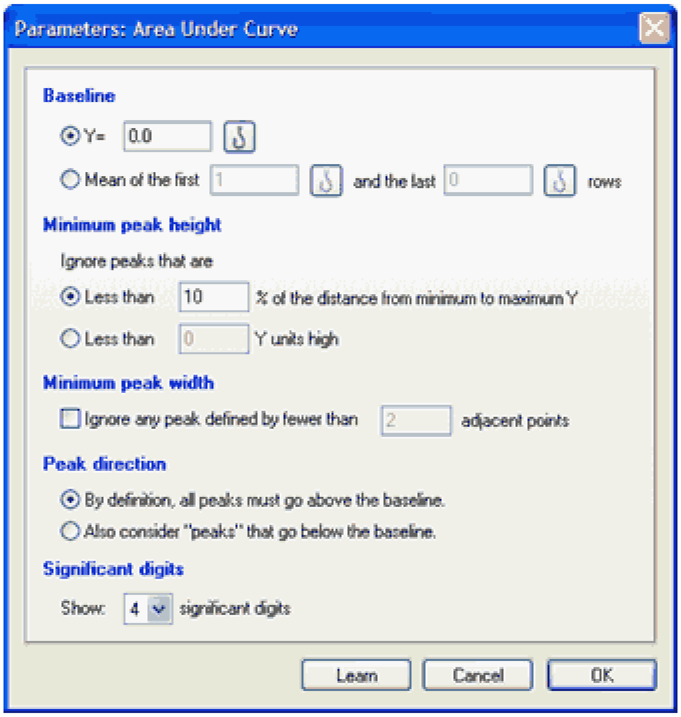
Area under the curve

This page explains how to compute the area under the curve. This analysis gives you one value for the area under the entire curve, as well as the area under well-defined peaks. A separate Prism analysis integrates a curve 88, resulting in another curve showing cumulative area.

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.

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.



Parameters: Area Under Curve

Baseline

● Y= 0.0

○ Mean of the first 1 and the last 0 rows

Minimum peak height

Ignore peaks that are

● Less than 10 % of the distance from minimum to maximum Y

○ Less than 0 Y units high

Minimum peak width

□ Ignore any peak defined by fewer than 2 adjacent points

Peak direction

● By definition, all peaks must go above the baseline.

○ Also consider "peaks" that go below the baseline.

Significant digits

Show: 4 significant digits

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.

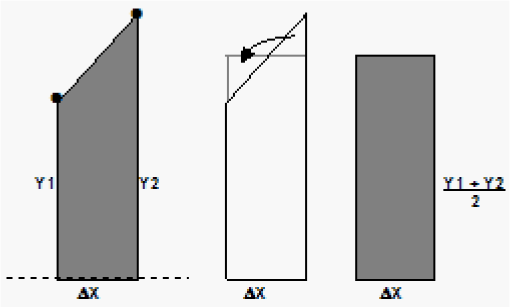
Note these limitations:

- 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”.

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 is simply a series of connected XY points, with equally spaced X values. 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 ΔX*(Y1+Y2)/2. Prism uses this formula repeatedly for each adjacent pair of points defining the curve.

Row statistics

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.

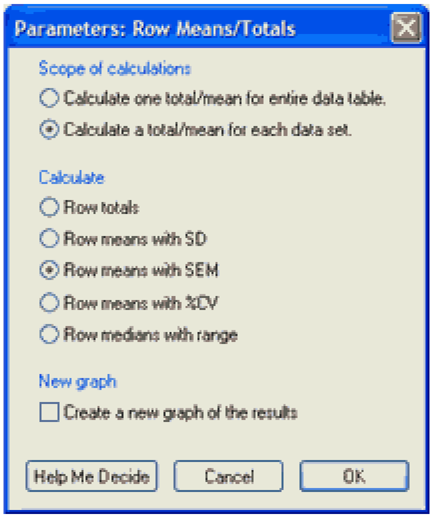
You can also format the table to enter mean, SD or SEM, and N. This is useful if you have already averaged the data in another program or if you have more than 52 replicates. Otherwise, it is best to enter the raw data into Prism, so you can plot every replicate.

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 92 analysis.

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.



Parameters: Row Means/Totals

Scope of calculations

○ Calculate one total/mean for entire data table.

● Calculate a total/mean for each data set.

Calculate

○ Row totals

○ Row means with SD

● Row means with SEM

○ Row means with %CV

○ Row medians with range

New graph

□ Create a new graph of the results

You can choose to compute the SD 18, SEM 22, or %CV 75 for each row each data set individually (what you'll usually want) or for the entire table.

III. Comparing two groups (t tests ...)

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?

Key concepts: t tests and related nonparametric tests

Q&A: Entering t test data

Is it possible to define the two groups with a grouping variable?

No. The two groups must be defined by 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 100 to see how. It is impossible to run a paired t test or a nonparametric test from data entered as mean, SD (or SEM) and N. You can only choose an unpaired t test.

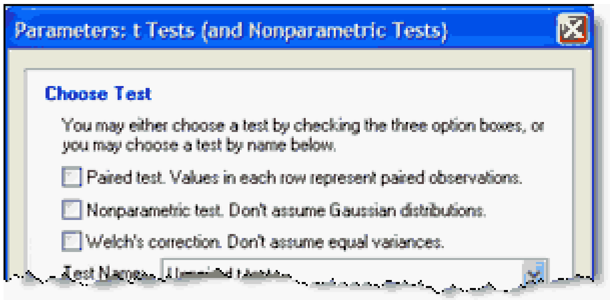
Can I enter data for many t tests on one table, and ask Prism to run them all at once?

No

Choosing a t test

The t test analysis compares the means (or medians) of two groups. If your goal is to compare the mean (or median) of one group with a hypothetical value, use the column statistics analysis 69 instead.

Prism offers five related tests that compare two groups. To choose among these tests, answer three questions:



Parameters: t Tests (and Nonparametric Tests)

Choose Test

You may either choose a test by checking the three option boxes, or you may choose a test by name below.

□ Paired test. Values in each row represent paired observations.

□ Nonparametric test. Don't assume Gaussian distributions.

□ Welch's correction. Don't assume equal variances.

Test Name: Unpaired t test

Are the data paired?

Choose a paired test when the columns of data are matched. 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.

Nonparametric test?

Nonparametric tests 64, unlike t tests, are not based on the assumption that the data are sampled from a Gaussian distribution 14. But nonparametric tests have less power 65, and report only P values but not confidence intervals. Deciding when to use a nonparametric test is not straightforward 67.

Equal variances?

If your data are not paired and you are not choosing a nonparametric test, you must 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, and you should accept it unless you have good reason not to. If you check the option for Welch's correction, the analysis will not assume equal variances (but will have less power).

Summary of tests

Test	Paired?	Nonparametric?	Welch correction?
Unpaired t 98	No	No	No
Welch's t 98	No	No	Yes
Paired t 109	Yes	No	N/A
Mann-Whitney 118	No	Yes	N/A
Wilcoxon matched pairs 126	Yes	Yes	N/A

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 133, and then follow up with multiple comparison post tests 158. The only exception 160 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 98 as mean, SD (or SEM) and N, and Prism can compute an unpaired t test. Prism cannot perform an paired test, as that requires analyzing each pair. It also cannot do any nonparametric tests, as these require ranking the data.

I only know the two group means, and don't have the raw data and don't know their SD or SEM. Can I run a t test?

No. The t test compares the difference between two means and compares that difference to the standard error of the difference, computed from the standard deviations and sample size. If you only know the two means, there is no possible way to do any statistical comparison.

Can I use a normality test to make the choice of when to use a nonparametric test?

It is not a good idea 64 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 206 and analyze with Fisher's 210 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 217. Don't run a t test on survival times.

Unpaired t test

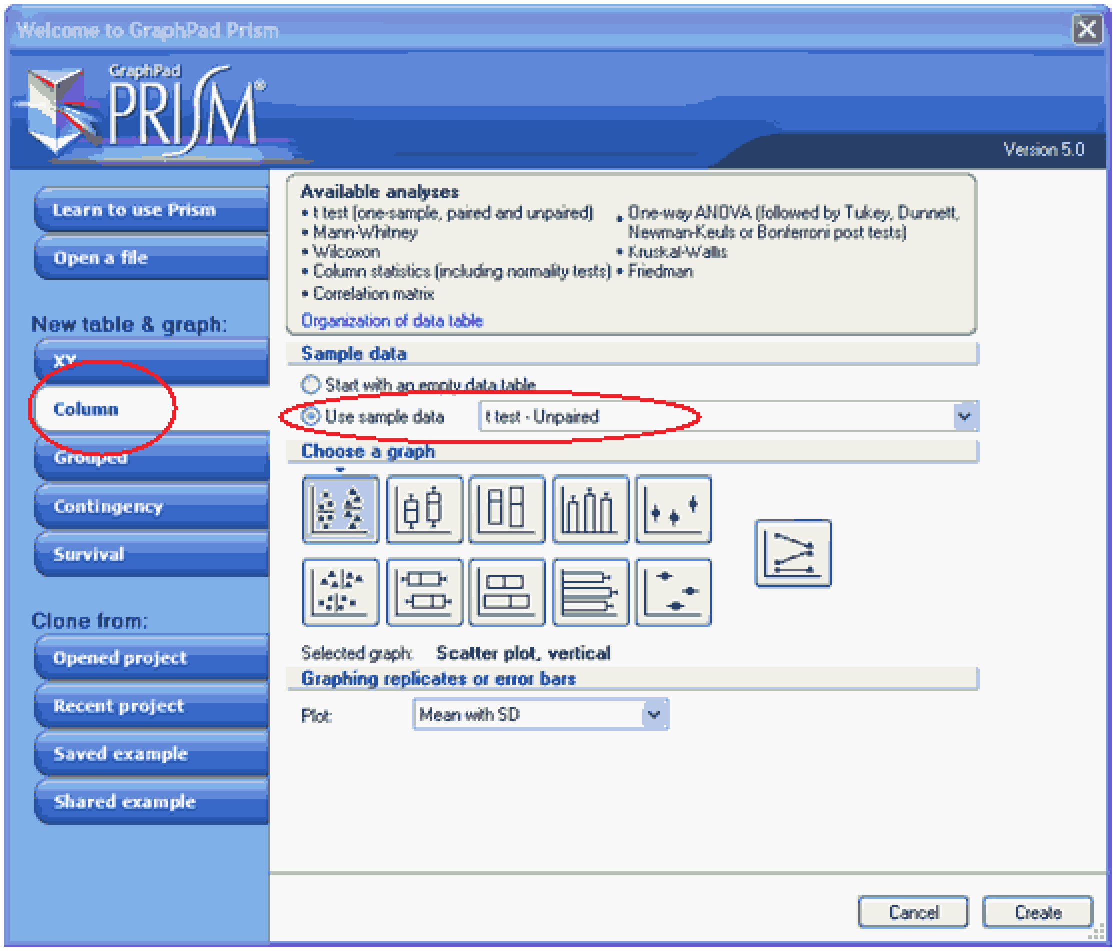
How to: Unpaired t test from raw data

This page explains how to enter and analyze raw data. Look elsewhere if you want to enter averaged data 100, if you want to perform the nonparametric Mann-Whitney test 118, or if your data are matched so you want to do a paired t test 109.

1. Create data table and enter data

From the Welcome (or New Table and graph) dialog, choose the Column tab, and then choose a scatter plot with a line at the mean.

If you are not ready to enter your own data, choose sample data and choose: t test - unpaired.


Welcome to GraphPad Prism

Available analyses
- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

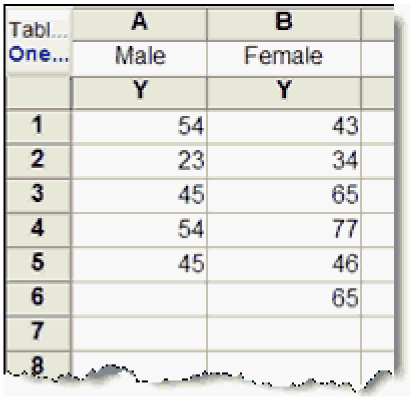
Sample data
○ Start with an empty data table
● Use sample data: t test - Unpaired

Selected graph: **Scatter plot, vertical**

Graphing replicates or error bars

Plot: Mean with SD

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.

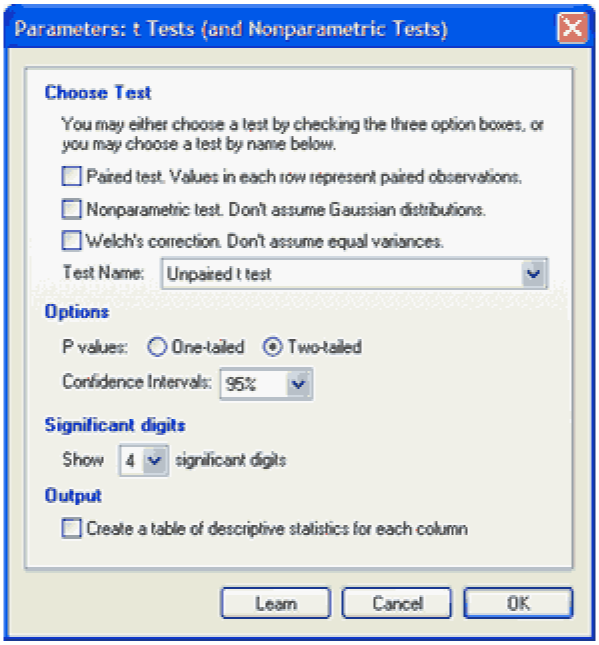


Tabl...	A	B
One...	Male	Female
	Y	Y
1	54	43
2	23	34
3	45	65
4	54	77
5	45	46
6		65
7		
8		

Stop. If you want to compare three or more groups, don't use t tests repeatedly. Instead, use one-way ANOVA 133 followed by multiple comparison 158 post tests.

2. Choose the unpaired t test

1. From the data table, click Analyze on the toolbar.
2. Choose t tests from the list of column analyses.
3. On the t test dialog, choose the 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.



Parameters: t Tests (and Nonparametric Tests)

Choose Test

You may either choose a test by checking the three option boxes, or you may choose a test by name below.

□ Paired test. Values in each row represent paired observations.

□ Nonparametric test. Don't assume Gaussian distributions.

□ Welch's correction. Don't assume equal variances.

Test Name: Unpaired t test

Options

P values: ○ One-tailed ● Two-tailed

Confidence Intervals: 95%

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

4. Choose a one- or two-tail P value 34. If in doubt, choose a two-tail P value

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 103 and graphing 105 the results.

Before accepting the results, review the analysis checklist 107.

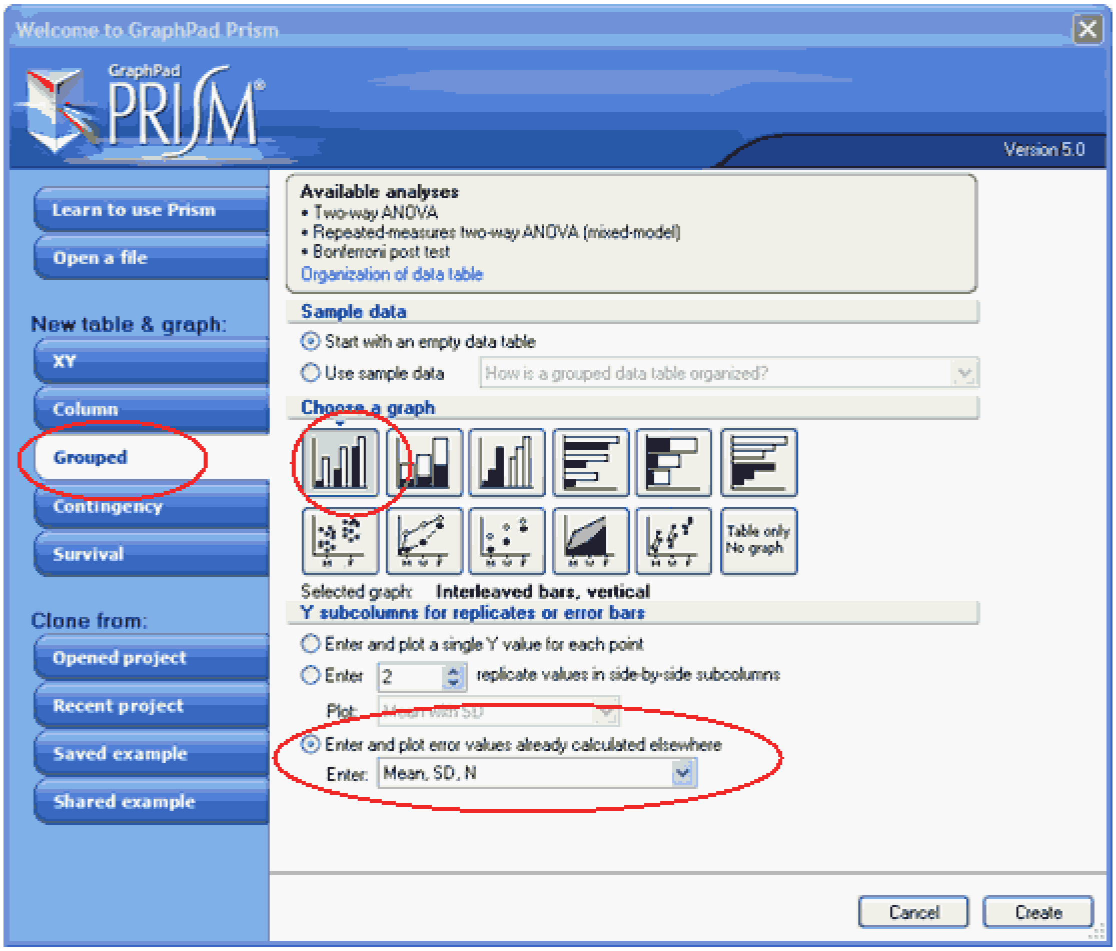
How to: Unpaired t test from averaged data

> The unpaired t test compares the means of two unmatched groups, assuming that the values follow a Gaussian distribution. This page gives detailed instructions for entering averaged data. Look elsewhere if you want to enter raw data 98.

1. Enter data

From the Welcome (or New Table and graph) dialog, choose the Grouped tab. (The t test is usually done from data tables formatted for column data, but Prism doesn't let you create column tables with subcolumns. Instead, create a Grouped table and enter data on one row).

Choose an interleaved bar graph, and choose to enter the data as Mean, SD and N (or as Mean, SEM and N).



Available analyses

- Two-way ANOVA
- Repeated-measures two-way ANOVA (mixed-model)
- Bonferroni post test

Organization of data table

Sample data

● Start with an empty data table

○ Use sample data How is a grouped data table organized?

Selected graph: Interleaved bars, vertical

Y subcolumns for replicates or error bars

○ Enter and plot a single Y value for each point

○ Enter 2 replicate values in side-by-side subcolumns

Plot: Mean with SD

● Enter and plot error values already calculated elsewhere

Enter: Mean, SD, N

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 table formatted for grouped data.

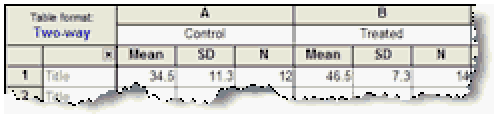
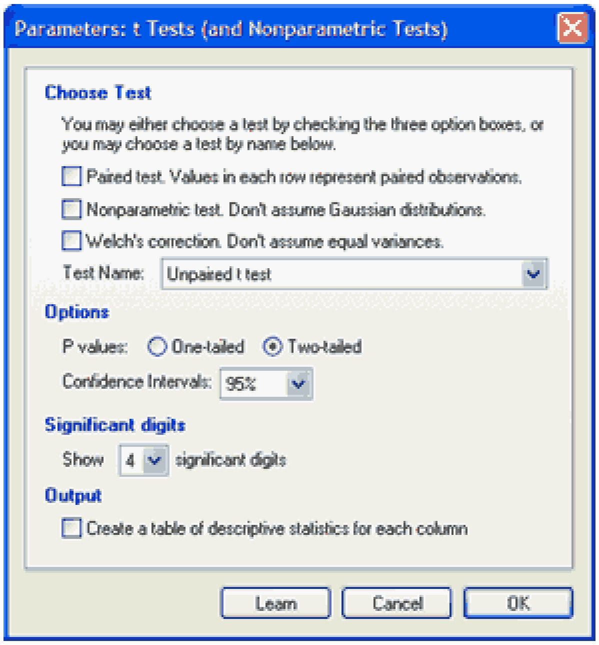


Table format: Two-way	A Control			B Treated		
×	Mean	SD	N	Mean	SD	N
1 Title	34.5	11.3	12	46.5	7.3	14
2 Title						

Stop. If you want to compare three or more groups, don't use t tests repeatedly. Instead, use one-way ANOVA 133 followed by multiple comparison 158 post tests.

2. Choose the unpaired t test

1. From the data table, click Analyze on the toolbar.
2. Choose t tests from the list of Column analyses.
3. On the t test dialog, choose the 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.



Choose Test

You may either choose a test by checking the three option boxes, or you may choose a test by name below.

□ Paired test. Values in each row represent paired observations.

□ Nonparametric test. Don't assume Gaussian distributions.

□ Welch's correction. Don't assume equal variances.

Test Name: Unpaired t test

Options

P values: ○ One-tailed ● Two-tailed

Confidence Intervals: 95%

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

4. Choose a one- or two-tail P value 34. If in doubt, choose a two-tail P value.

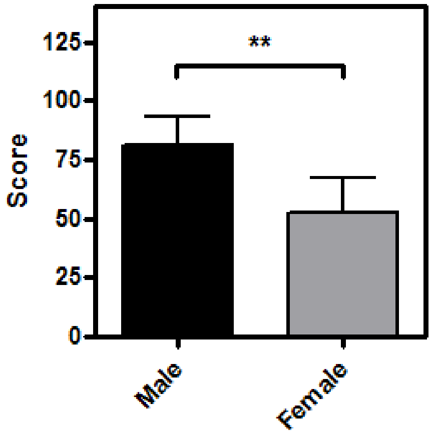
3. Review the results

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 103.

Before accepting the results, review the analysis checklist 107.

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? 22).
- To add the asterisks representing significance level 39 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.

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 107, 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.

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 38”.

You will interpret the results differently depending on whether the P value is small 35 or large 36.

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?

> **Note:** 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

If the P value is large (>0.05) you conclude that there is no evidence that the variances differ. If the P value is small, you conclude that the variances differ significantly. Then what? There are four answers.

- Ignore the result. With equal, or nearly equal, sample size (and moderately large samples), the assumption of equal variances is not a crucial assumption and the t test works pretty well even with unequal standard deviations. In other words, the t test is remarkably robust to violations of that assumption so long as the sample size isn't tiny and the sample sizes aren't far apart.
- Go back and rerun the t test, checking the option to do the modified Welch t test that allows for unequal variance. While this sounds sensible, Moser and Stevens (Amer. Statist. 46:19-21, 1992) have shown that it is not a good idea to first look at the F test to compare variances, and then switch to the modified (Welch modification to allow for different variances) t test when the P value is less than 0.05. If you think your groups might have different variances, you should always use this modified test (and thus lose some power).
- Transform your data (often to logs or reciprocals) in an attempt to equalize the variances, and then run the t test on the transformed results. Logs are especially useful, and are worth thinking about.
- Conclude that the two populations are different; that the treatment had an effect. In many experimental contexts, the finding of different variances is as important as the finding of different means. If the variances are truly different, then the populations are different regardless of what the t test concludes about differences between the means. This may be the most important conclusion from the experiment, so think about what it might mean before using one of the other approaches listed above.

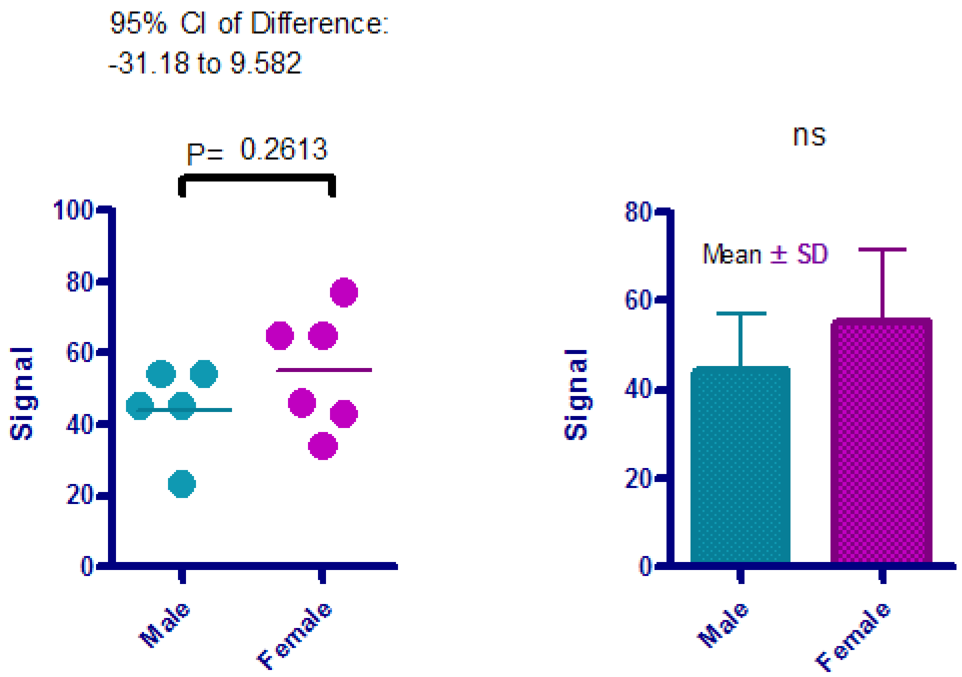
R squared from unpaired t test

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.

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.

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

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.

Analysis checklist: Unpaired t test

The unpaired t test compares the means of two unmatched groups, assuming that the values follow a Gaussian distribution.

✓ 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 help you test this assumption 111.

✓ 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 136 followed by multiple comparison tests 163. 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 160 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 77.

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 34, 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.

Paired t test

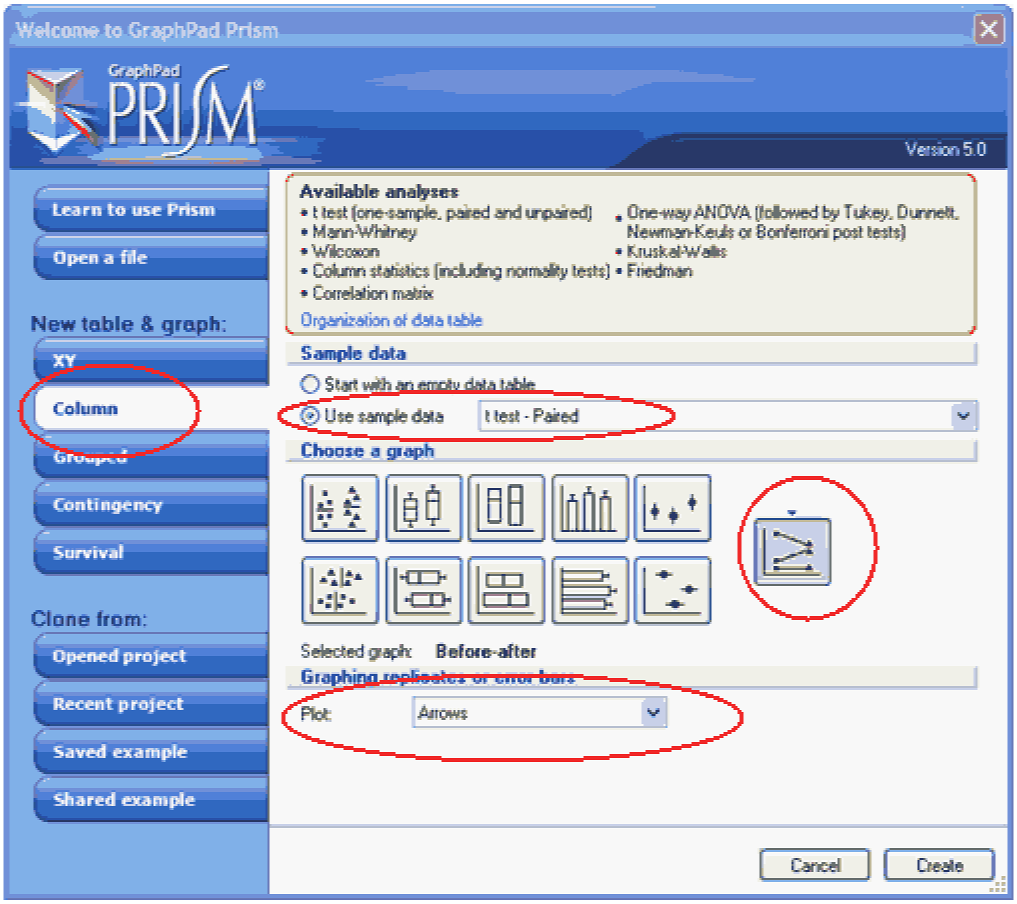
How to: Paired t test

The paired t test compares the means of two matched groups, assuming that the distribution of the before-after differences follows a Gaussian distribution. Look elsewhere if you want to perform the nonparametric Wilcoxon test 126.

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.



Available analyses
- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

Sample data
○ Start with an empty data table
● Use sample data: t test - Paired

Selected graph: Before-after

Plot: Arrows

Enter the data for each group into a separate column, with matched values on the same row. If you leave any missing values, that row will simply be ignored. Optionally, enter row labels to identify the source of the data for each row (i.e. subject's initials).

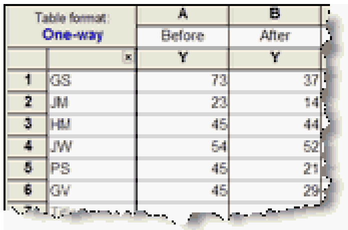
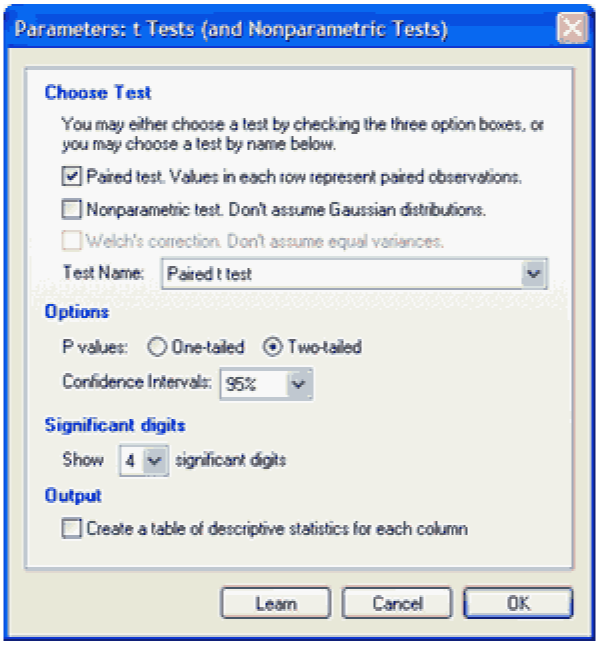


	Table format: One-way	A	B
		Before	After
	X	Y	Y
1	GS	73	37
2	JM	23	14
3	HM	45	44
4	JW	54	52
5	PS	45	21
6	GV	45	29

Stop. If you want to compare three or more groups, use repeated-measures one-way ANOVA 144 (not multiple t tests).

2. Choose the paired t test

1. From the data table, click Analyze on the toolbar.
2. Choose t tests from the list of column analyses.
3. On the t test dialog, choose the paired t test.



Parameters: t Tests (and Nonparametric Tests)

Choose Test

You may either choose a test by checking the three option boxes, or you may choose a test by name below.

☑ Paired test. Values in each row represent paired observations.

□ Nonparametric test. Don't assume Gaussian distributions.

□ Welch's correction. Don't assume equal variances.

Test Name: Paired t test

Options

P values: ○ One-tailed ● Two-tailed

Confidence Intervals: 95%

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

4. Choose a one- or two-tail P value 34. If in doubt, choose a two-tail P value.

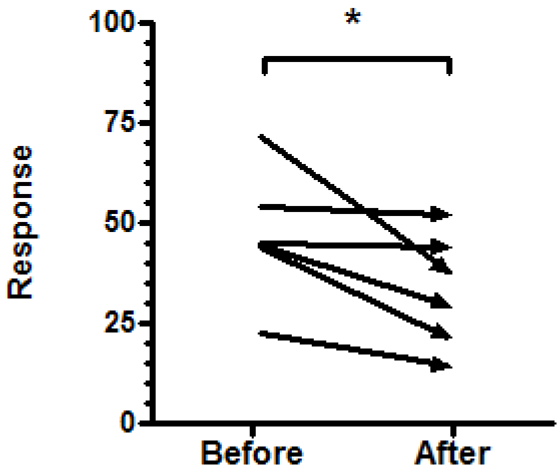
3. Review the results

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 112.

Before accepting the results, review the analysis checklist 116.

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 39 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 113.

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. From your data table, click Analyze and choose "Remove baseline...".
2. On the Remove Baseline dialog, define the baseline to be column B, and that you want to compute the difference.
3. View the results table showing the differences. Click Analyze and choose Column statistics. Note that you are chaining two analyses, first subtracting a baseline and then performing column statistics on the results.
4. 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 one.

5. 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 76.

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.

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 116, 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 38".

You will interpret the results differently depending on whether the P value is small 35 or large 36.

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. If the value in column B is larger, then the difference is positive. If the value in column A is larger, then the difference is negative. 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 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.

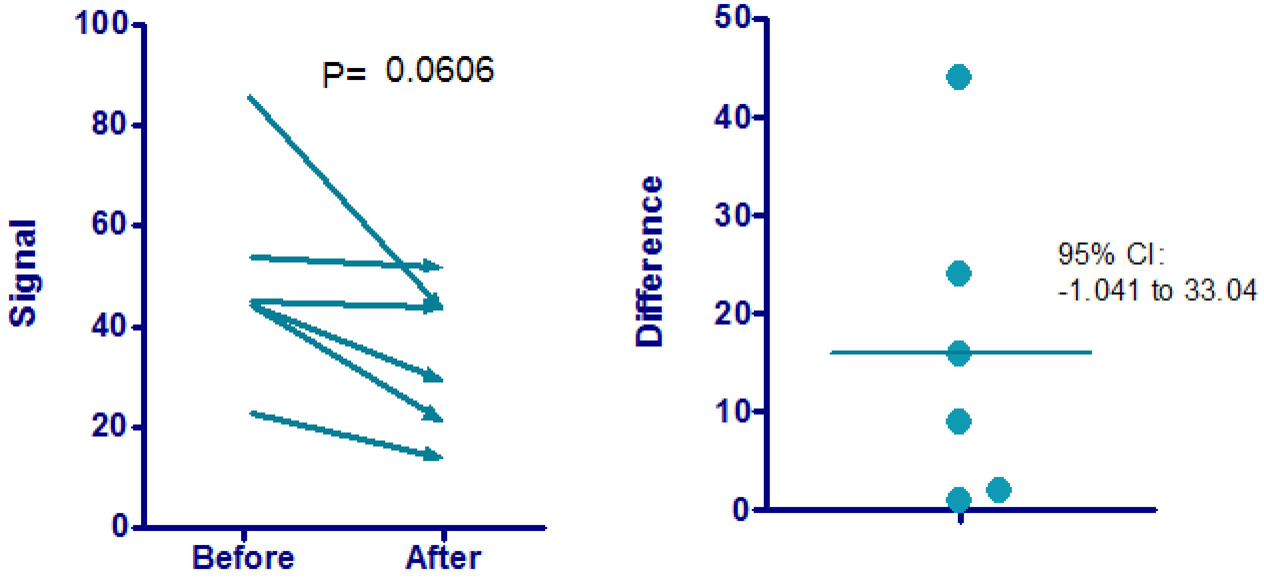
If the pairing was effective, r will be positive and the P value will be small. This means that the two groups are significantly correlated, so it made sense to choose a paired test.

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.

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 were computed using the Remove Baseline analysis.
- The confidence interval for the difference between means shown on the right graph was copy and pasted from the paired t test results.

An alternative to paired t test: Ratio t test

The paired t test analyzes the *differences* between pairs. For each pair, you calculate the difference. Then you calculate 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 mean is really zero. Prism does not perform a ratio t test directly, but you can do so indirectly by taking advantage of this simple mathematical fact.

$$\log\left(\frac{treated}{control}\right) = \log(treated) - \log(control)$$

To perform a ratio t test with Prism, follow these steps

1. Starting from your data table, click Analyze and choose Transform. On the Transform dialog, choose the transform Y=log(Y).

2. From the results of this transform, click Analyze and choose to do a t test. On the t test dialog, choose a paired t test. Notice that you are chaining the transform analysis with the t test analysis.

3. Interpret the P value: 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.

4. Manually compute the antilog of each end of the confidence interval of the difference between the means of the logarithms. The result is the 95% confidence interval of the ratio of the two means, which is much easier to interpret. (This will make more sense after you read the example below.)

Example

You measure the Km of a kidney enzyme (in nM) before and after a treatment. Each experiment was done with renal tissue from a different animal.

Control	Treated	Difference	Ratio
4.2	8.7	4.3	0.483
2.5	4.9	2.4	0.510
6.5	13.1	6.6	0.496

The P value from a conventional paired t test is 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 95% confidence interval for the difference between control and treated Km value is -0.72 to 9.72, which includes zero.

The ratios are much more consistent. It is not appropriate to analyze the ratios directly. Because ratios are inherently asymmetrical, you'll get a different answer depending on whether you analyze the ratio of treated/control or control/treated. You'll get different P values testing the null hypothesis that the ratio really equals 1.0.

Instead, we analyze the log of the ratio, which is the same as the difference between the log(treated) and log(control). Using Prism, click Analyze, pick Transform, and choose Y=log(Y). Then from the results table, click Analyze again and choose t test, then paired t test. The P value is 0.0005. Looked at this way, the treatment has an effect that is highly statistically significant.

The t test analysis reports the difference between the means, which is really the difference between means of log(control) and log(treated), or -0.3043. Take the antilog of this value (10 to that power) to obtain the ratio of the two means, which is 0.496. In other words, the control values are about half the treated values.

It is always important to look at confidence intervals as well as P values. Here the 95% confidence interval extends from -0.3341 to -0.2745. Take the antilog (10 to the power) of both numbers to get the confidence interval of the ratio, which extends from 0.463 to 0.531. We are 95% sure the control values are between 46% and 53% of the treated values.

You might want to look at the ratios the other way around. If you entered the treated values into column A and the control in column B, the results would have been the reciprocal of what we just computed. It is easy enough to compute the reciprocal of the ratio at both ends of its confidence interval. On average, the treated values are 2.02 times larger than the control values, and the 95% confidence interval extends from 1.88 to 2.16.

Analyzed with a paired t test, the results were very 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.

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 111.

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 12 – 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 52 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 34 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 114.

Mann-Whitney test

How to: Mann-Whitney test

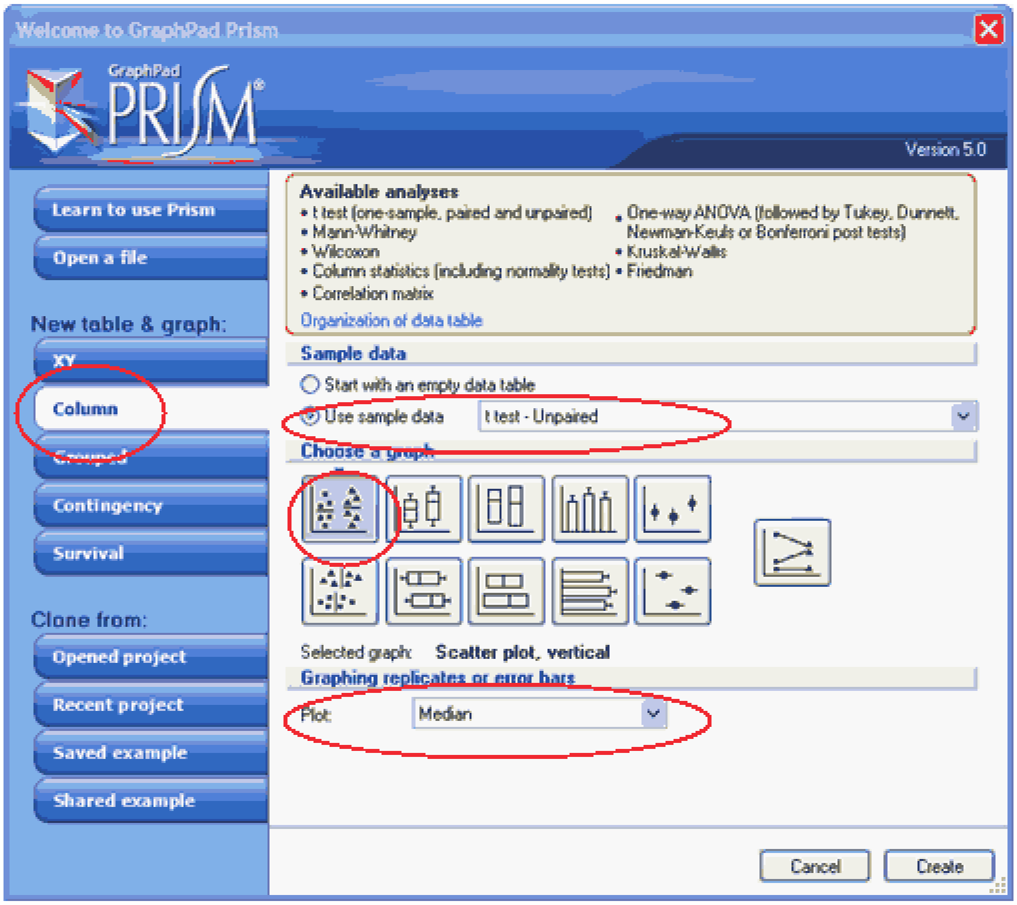
The Mann-Whitney test is a nonparametric test that compares the distributions of two unmatched groups. Look elsewhere if you want to compare three or more groups with the Kruskal-Wallis test 149, or perform the parametric unpaired t test 98.

This test is also called the Wilcoxon rank sum test. Don't confuse it with the Wilcoxon matched pairs test 126, which is used when the values are paired or the Wilcoxon signed-rank test 78 which compares a median with a hypothetical value.

1. Enter data

From the Welcome (or New Table and graph) dialog, choose the Column tab, and then choose a scatter plot with a line at the median.

If you are not ready to enter your own data, choose sample data and choose: t test - unpaired.



Welcome to GraphPad Prism

GraphPad PRISM® Version 5.0

Available analyses
- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

Organization of data table

Sample data

○ Start with an empty data table
● Use sample data: t test - Unpaired

Choose a graph

Selected graph: **Scatter plot, vertical**

Graphing replicates or error bars

Plot: Median

Since the Mann-Whitney tests compares the sum-of-ranks in the two groups, the test needs to have your raw data. It is not possible to perform a Mann-Whitney test if you entered your data as mean and SD (or SEM).

Enter the data for each group into a separate column. The two groups do not have to have the same number of values, and it's OK to leave some cells empty. Since the data are unmatched, it makes no sense to enter any row titles.

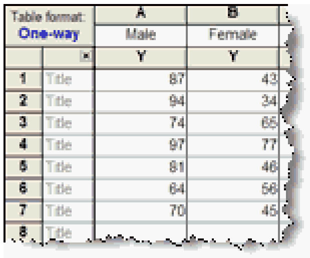
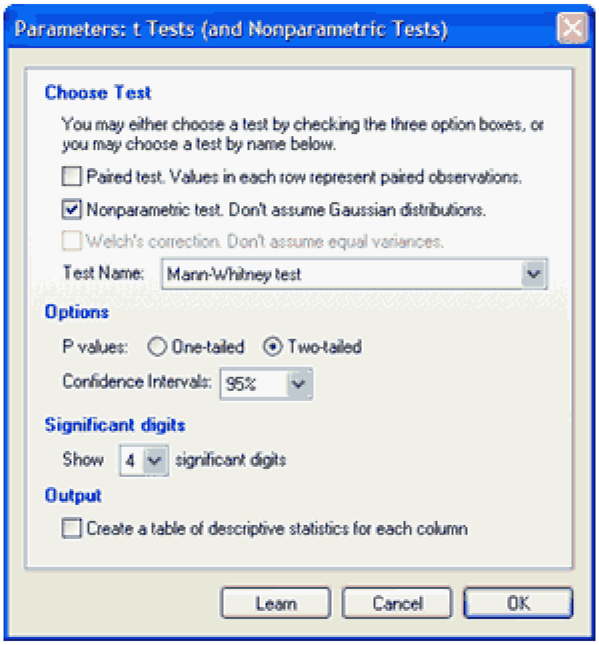


Table format: One-way		A	B
		Male	Female
	×	Y	Y
1	Title	87	43
2	Title	94	34
3	Title	74	65
4	Title	97	77
5	Title	81	46
6	Title	64	56
7	Title	70	45
8	Title		

Stop. If you want to compare three or more groups, use the Kruskal-Wallis test 149.

2. Choose the Mann-Whitney 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 Mann-Whitney test.



Parameters: t Tests (and Nonparametric Tests)

Choose Test

You may either choose a test by checking the three option boxes, or you may choose a test by name below.

□ Paired test. Values in each row represent paired observations.

☑ Nonparametric test. Don't assume Gaussian distributions.

□ Welch's correction. Don't assume equal variances.

Test Name: Mann-Whitney test

Options

P values: ○ One-tailed ● Two-tailed

Confidence Intervals: 95%

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

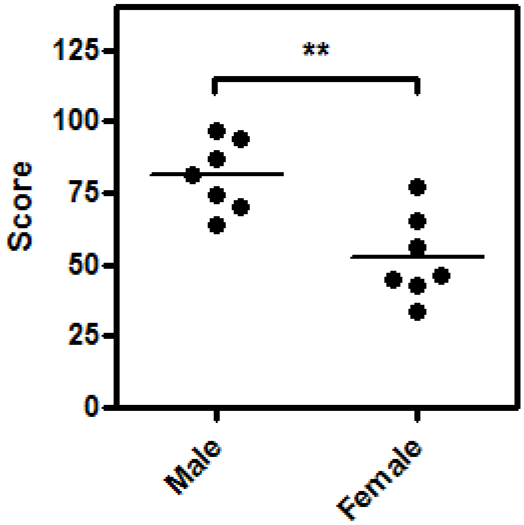
4. Choose a one- or two-tail P value 34. If in doubt, choose a two-tail P value.

3. Review the results

Learn more about interpreting the results of a Mann-Whitney test 122.

Before accepting the results, review the analysis checklist 124.

4. 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 39 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'.

Interpreting results: Mann-Whitney test

P value

The Mann-Whitney test, also called the 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 sums the ranks in each group, and reports the two sums. If the sums of the ranks are very different, the P value will be small.

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, and there are no ties, Prism calculates an exact P value. If your samples are large, or if there are ties, 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).

If the P value is small, you can reject the idea that the difference is due to random sampling, 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 Mann-Whitney test has little power 65. 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.

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.

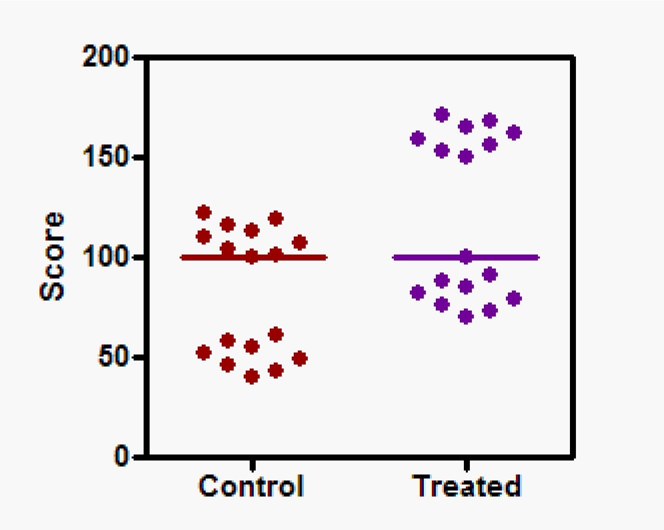
Prism uses a standard method to correct for ties when it computes U (or the sum of signed ranks; the two are equivalent).

Unfortunately, there isn't a standard method to get a P value from these statistics when there are ties. Prism always uses the approximate method, which converts U or sum-of-ranks to a Z value. It then looks up that value on a Gaussian distribution to get a P value. The exact test is only exact when there are no ties.

If you have large sample sizes and a few ties, no problem. But with small data sets or lots of ties, we're not sure how meaningful the P values are. One alternative: Divide your response into a few categories, such as low, medium and high. Then use a chi-square test to compare the two groups.

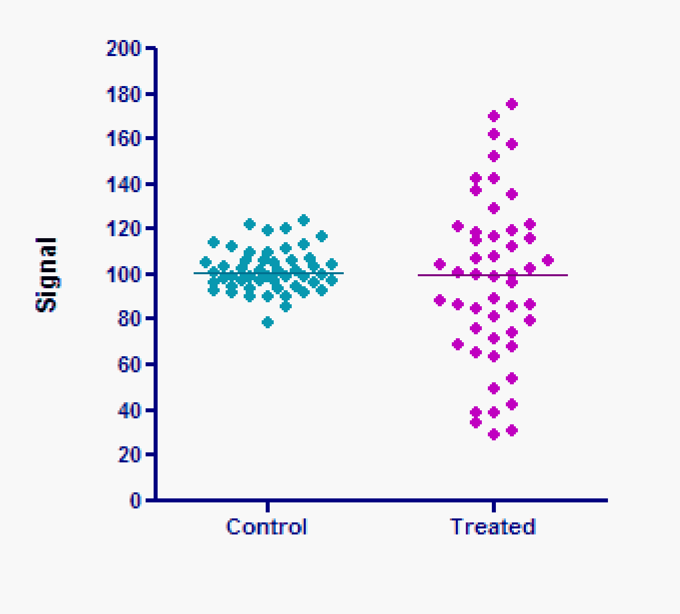
The Mann-Whitney test doesn't really compare medians

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 compared the distributions of ranks, which is quite different in the two groups.

It is not correct, however, 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 Mann-Whitney test compares sums of ranks -- it does not compare medians and does not compare distributions. To interpret the test as being a comparison of medians, you have to make an additional assumption -- that the distributions of the two populations have the same shape, even if they are shifted (have different medians). With this assumption, if you reject the Mann-Whitney test reports a small P value, you can conclude that the medians are different.

Analysis checklist: Mann-Whitney test

The Mann-Whitney test is a nonparametric test that compares the distributions of two unmatched groups.

✓ 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 12 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 122 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 123). 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. 34

✓ 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. 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 to create a Gaussian distribution and then using a t test.

Wilcoxon matched pairs test

How to: Wilcoxon matched pairs test

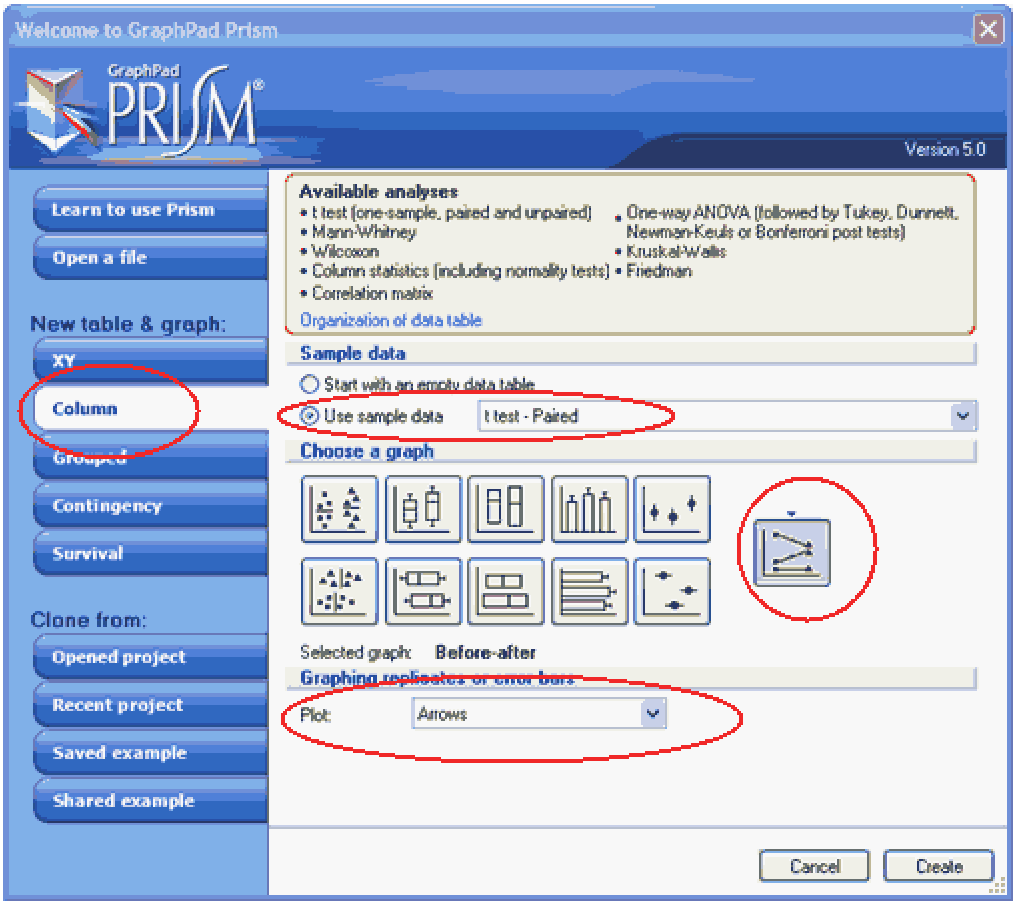
The Wilcoxon matched pairs compares two matched groups, without assuming that the distribution of the before-after differences follows a Gaussian distribution. Look elsewhere if you want to perform the paired t test 109.

Beware: Wilcoxon's name is used on two different tests. The test usually called the Mann-Whitney test 118 is also called the Wilcoxon rank-sum test. It compares two groups of unpaired data.

1. Enter data

From the Welcome (or New Table and graph) dialog, choose the one-way 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.



Available analyses
- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: t test - Paired

Choose a graph

Selected graph: **Before-after**

Graphing replicates or error bars

Plot: Arrows

Enter the data for each group into a separate column, with matched values on the same row. If you leave any missing values, that row will simply be ignored. Optionally, enter row labels to identify the source of the data for each row (i.e. subject's initials).

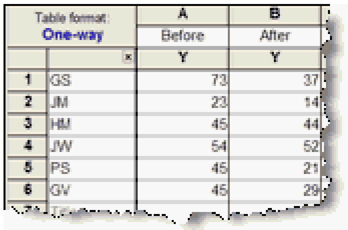
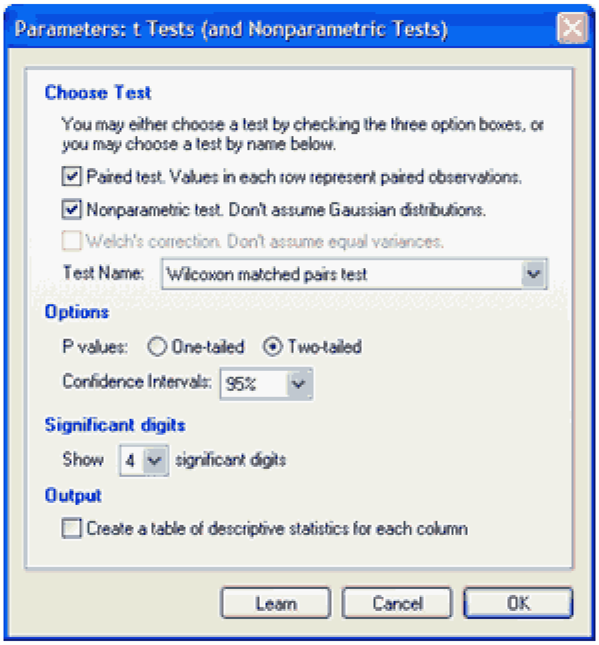


Table format: One-way		A	B
		Before	After
	×	Y	Y
1	GS	73	37
2	JM	23	14
3	HM	45	44
4	JW	54	52
5	PS	45	21
6	GV	45	29

Stop. If you want to compare three or more groups, use Friedman's test 154.

2. Choose the Wilcoxon matched pairs test

1. From the data table, click Analyze on the toolbar.
2. Choose t tests from the list of column analyses.
3. On the t test dialog, choose the Wilcoxon matched-pairs test.



Parameters: t Tests (and Nonparametric Tests)

Choose Test

You may either choose a test by checking the three option boxes, or you may choose a test by name below.

☑ Paired test. Values in each row represent paired observations.

☑ Nonparametric test. Don't assume Gaussian distributions.

□ Welch's correction. Don't assume equal variances.

Test Name: Wilcoxon matched pairs test

Options

P values: ○ One-tailed ● Two-tailed

Confidence Intervals: 95%

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

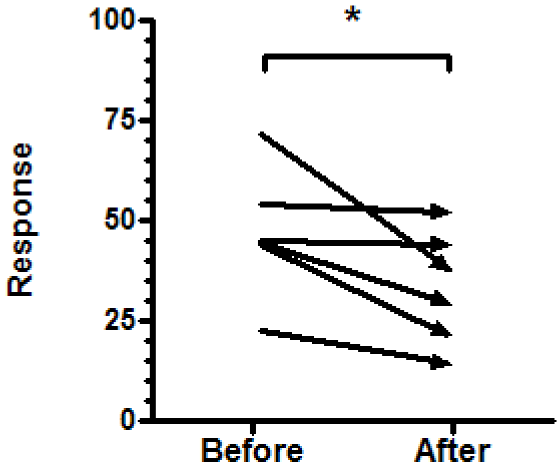
4. Choose a one- or two-tail P value 34. If in doubt, choose a two-tail P value.

3. Review the results

Learn more about interpreting the results of Wilcoxon's matched pairs test 129.

Before accepting the results, review the analysis checklist 130.

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 39 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'.

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 two sums of ranks are very different, 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 means 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 your samples are small and there are no tied ranks, Prism calculates an exact P value. If your samples are large or there are tied ranks, 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.

When some of the subjects have exactly the same value before and after the intervention (same value in both columns), there are two ways to compute the P value:

- Prism uses the method suggested by Wilcoxon and described in S Siegel and N Castellan, Nonparametric Statistics for the Behavioral Sciences and in WW Daniel, Applied Nonparametric Statistics (and many others). The subjects that show no change are simply eliminated from the analysis, reducing N. The argument is that since the outcome doesn't change at all in these subjects, they provide no information at all that will be helpful in comparing groups.

- Other books show a different method that still accounts for those subjects, and this alternative method gives a different P value. The argument is that the lack of change in these subjects brings down the average change altogether, so appropriately raises the P value.

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.

Analysis checklist: Wilcoxon matched pairs test

The Wilcoxon test is a nonparametric test that compares two paired groups.

✓ Are the pairs independent?

The results of a Wilcoxon test only make sense when the pairs are independent 12 – 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 52 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 34, 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.

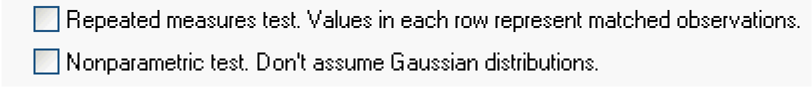
IV. Comparing three or more groups (one-way ANOVA ...)

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?

Key concepts: One-way ANOVA

Choosing a one-way ANOVA

Prism offers four related tests that compare three or more groups. Your choice of a test depends on two choices:



□ Repeated measures test. Values in each row represent matched observations.

□ Nonparametric test. Don't assume Gaussian distributions.

Repeated measures?

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*.

Nonparametric test?

Nonparametric tests 64, unlike ANOVA are not based on the assumption that the data are sampled from a Gaussian distribution 14. But nonparametric tests have less power 65, and report only P values but not confidence intervals. Deciding when to use a nonparametric test is not straightforward 67.

Test summary

Test	Matched	Nonparametric
Ordinary one-way ANOVA 136	No	No
Repeated measures one-way ANOVA 144	Yes	No
Kruskal-Wallis test 149	No	Yes
Friedman test 154	Yes	Yes

Q&A: Entering one-way ANOVA data

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 136 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.

Q&A: 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 133, and then follow up with multiple comparison post tests 158. The only exception 160 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 64. 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 206 and analyze with a chi-square test 208.

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 169 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.

One-way ANOVA

How to: One-way ANOVA

One-way ANOVA compares the means of three or more unmatched groups. If your data are matched, learn about repeated measures one-way ANOVA 144. If you want to perform a nonparametric test, read about the Kruskal-Wallis 149 and Friedman tests 154.

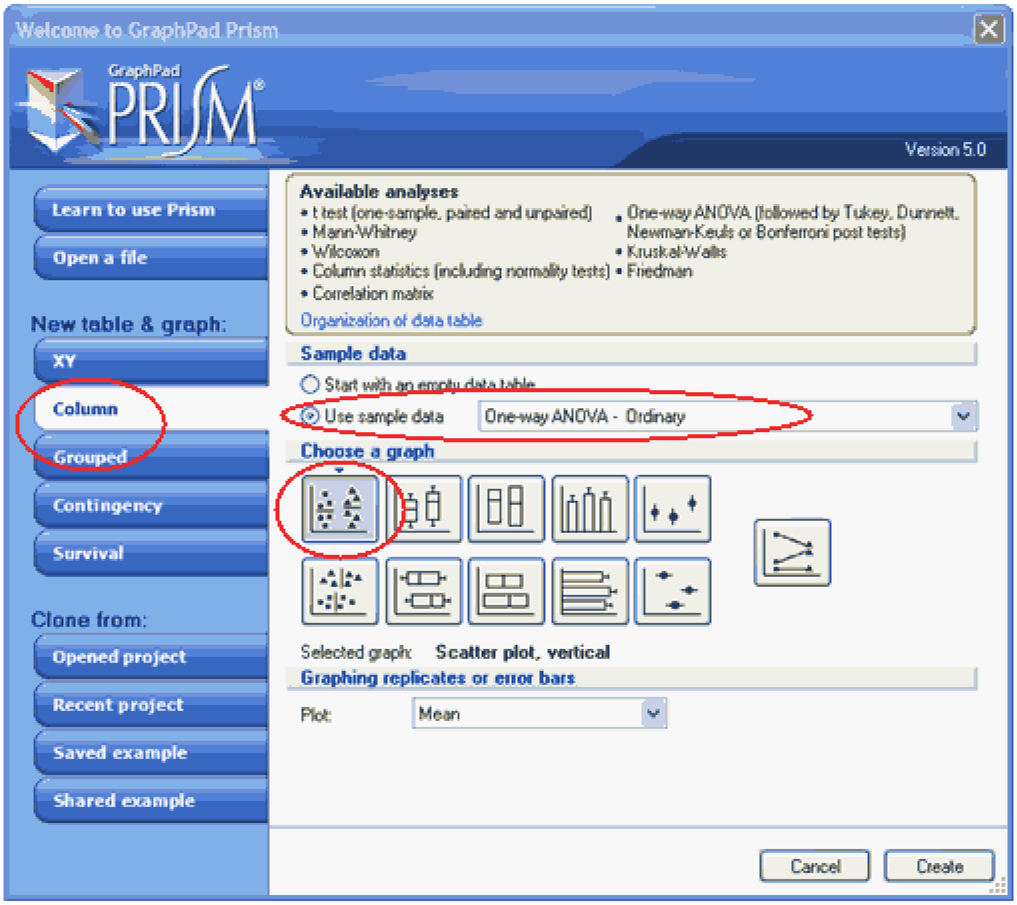
1. Enter data

You can enter the actual values you want to analyze (raw data) or the mean and SD (or SEM) of each group.

Enter raw data

From the Welcome (or New Table and graph) dialog, choose a vertical scatter plot from the Column tab.

If you aren't ready to enter your own data, choose to use sample data, and choose: One-way ANOVA - Ordinary.



Available analyses

- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: One-way ANOVA - Ordinary

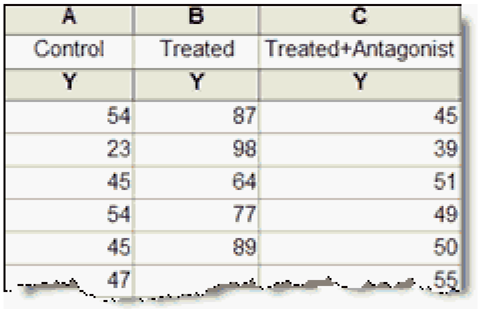
Choose a graph

Selected graph: **Scatter plot, vertical**

Graphing replicates or error bars

Plot: Mean

Enter the data for each group into a separate column. The two groups do not have be the same size (it's OK to leave some cells empty). Since the data are unmatched, it makes no sense to enter any row titles.

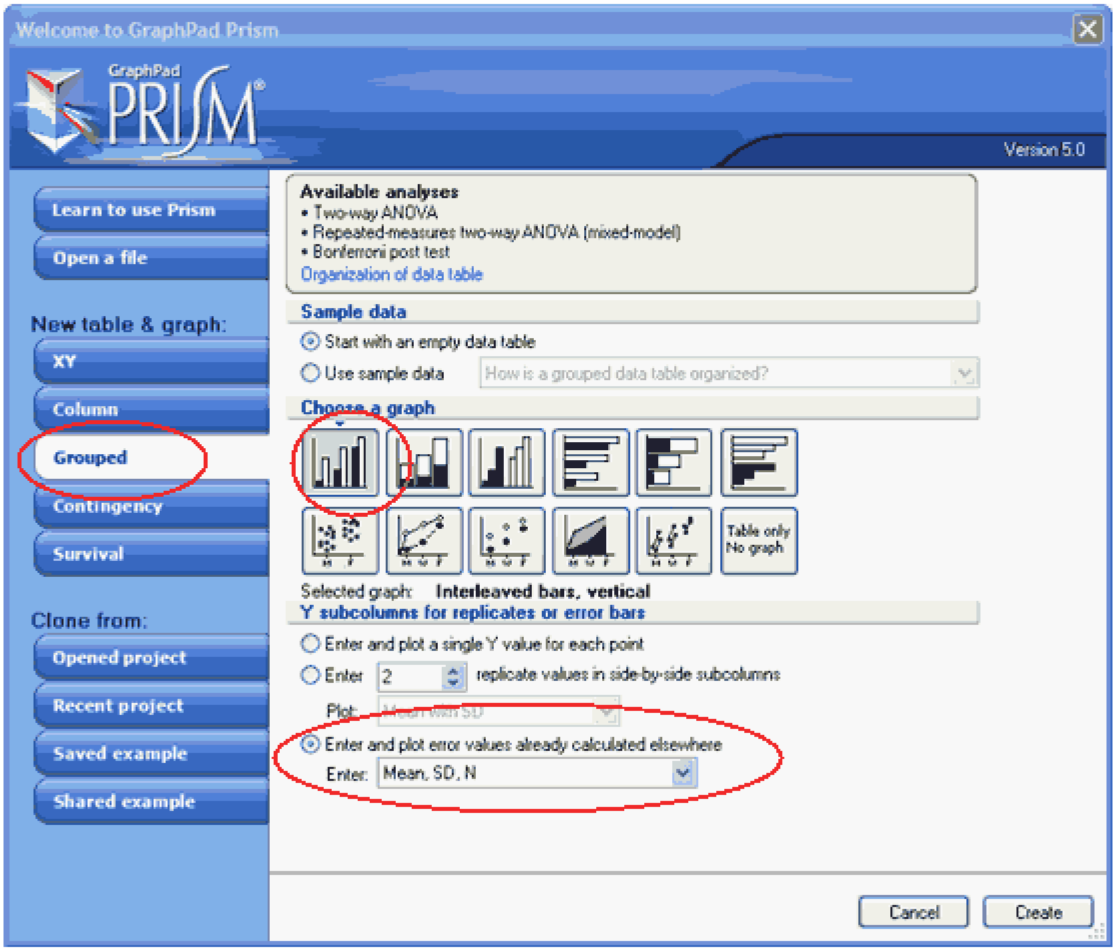


A	B	C
Control	Treated	Treated+Antagonist
Y	Y	Y
54	87	45
23	98	39
45	64	51
54	77	49
45	89	50
47		55

Enter averaged data

Prism also lets you perform one-way ANOVA with data entered as mean, SD (or SEM), and N. This can be useful if you are entering data from another program or publication.

From the Welcome (or New Table and graph) dialog, choose the Grouped tab. That's right. Even though you want to do one-way ANOVA, you should pick the two-way tab (since the one-way table doesn't allow for entry of mean and SD or SEM). Then choose any graph from the top row, and choose to enter the data as Mean, SD and N or as Mean, SEM and N. Entering N is essential.



Available analyses
- Two-way ANOVA
- Repeated-measures two-way ANOVA (mixed-model)
- Bonferroni post test

Organization of data table

Sample data

● Start with an empty data table

○ Use sample data How is a grouped data table organized?

Choose a graph

Selected graph: **Interleaved bars, vertical**

Y subcolumns for replicates or error bars

○ Enter and plot a single Y value for each point

○ Enter 2 replicate values in side-by-side subcolumns

Plot: Mean with SD

● Enter and plot error values already calculated elsewhere

Enter: Mean, SD, N

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.

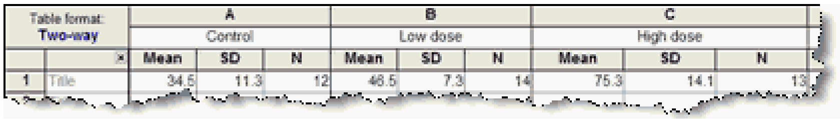
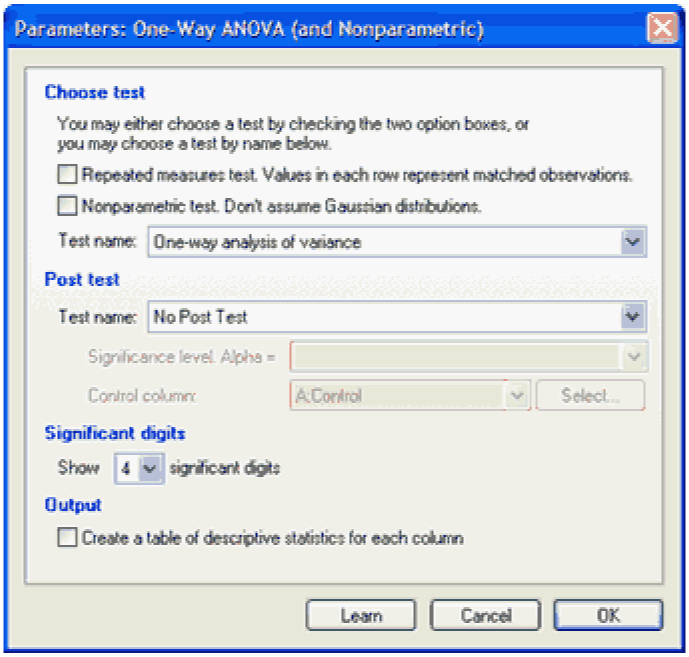


Table format: Two-way	A			B			C		
	Control			Low dose			High dose		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
1 Title	34.5	11.3	12	46.5	7.3	14	75.3	14.1	13

2. Choose one-way ANOVA

1. From the data table, click Analyze on the toolbar.
2. Choose one-way ANOVA from the list of column analyses.
3. On the ANOVA dialog, choose ordinary one-way ANOVA.



Parameters: One-Way ANOVA (and Nonparametric)

Choose test

You may either choose a test by checking the two option boxes, or you may choose a test by name below.

□ Repeated measures test. Values in each row represent matched observations.

□ Nonparametric test. Don't assume Gaussian distributions.

Test name: One-way analysis of variance

Post test

Test name: No Post Test

Significance level. Alpha =

Control column: A:Control

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

3. Choose a post test

How to choose. 163

4. Review the results and inspect the graph

ANOVA investigates the likelihood that the difference among the means could have been caused by chance. The most important results are the P value and the results of post tests.

Before accepting the results, review the analysis checklist 142.

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 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.

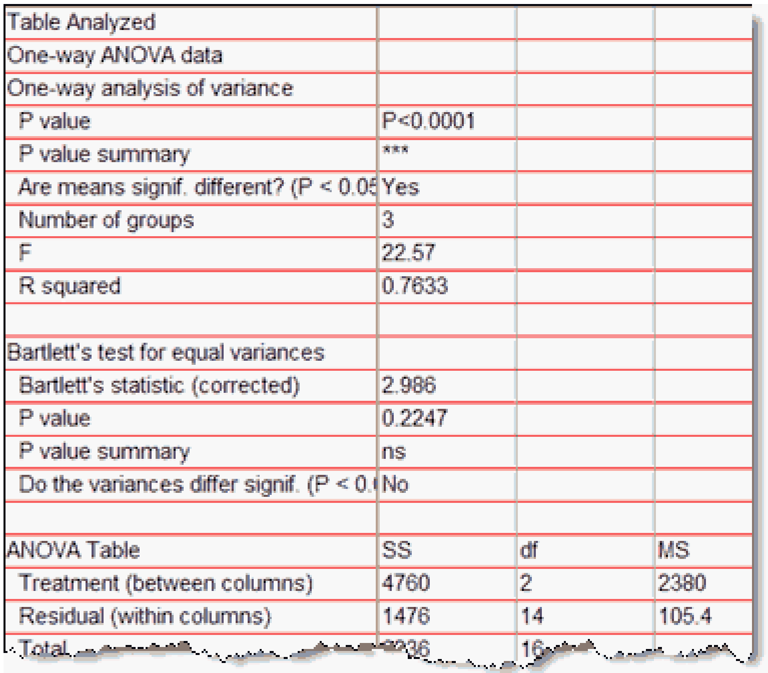


Table Analyzed			
One-way ANOVA data			
One-way analysis of variance			
P value	P<0.0001		
P value summary	***		
Are means signif. different? (P < 0.05	Yes		
Number of groups	3		
F	22.57		
R squared	0.7633		
Bartlett's test for equal variances			
Bartlett's statistic (corrected)	2.986		
P value	0.2247		
P value summary	ns		
Do the variances differ signif. (P < 0.(	No		
ANOVA Table	SS	df	MS
Treatment (between columns)	4760	2	2380
Residual (within columns)	1476	14	105.4
Total	6236	16	

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.

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.

Bartlett's test for equal variances

ANOVA is based on the assumption that the populations all have the same standard deviations. If ever group has at least five values, Prism tests this assumption using Bartlett's test. It reports the value of Bartlett's statistic along with a P value that answers 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?

If the P value is small, you must decide whether you will conclude that the standard deviations of the two populations are different. Obviously Bartlett's test is 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 three 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.
- Ignore Bartlett's test and interpret the ANOVA results as usual. Bartlett's test is very sensitive to deviations from a Gaussian distribution – more sensitive than the ANOVA calculations. A low P value from Bartlett's test may be due to data that are not Gaussian, rather than due to unequal variances. Since ANOVA is fairly robust to non-Gaussian data (at least when sample sizes are equal), some statisticians suggest ignoring the Bartlett's test, especially when the sample sizes are equal (or nearly so).

Why not switch to the nonparametric Kruskal-Wallis test?. While nonparametric tests do not assume Gaussian distributions, the Kruskal-Wallis test (and other nonparametric tests) does assume that the shape of the data distribution is the same in each group. So if your groups have very different standard deviations and so are not appropriate for one-way ANOVA, they should not be analyzed by the Kruskal-Wallis test either.

Some suggest using Levene's median test instead of Bartlett's test. Prism doesn't do this test (yet), but it isn't hard to do by Excel (combined with Prism). To do Levene's test, first create a new table where each value is defined as the absolute value of the difference between the actual value and median of its group. Then run a one-way ANOVA on this new table. The idea is that by subtracting each value from its group median, you've gotten rid of difference between group averages. (Why not subtract means rather than medians? In fact, that was Levene's idea, but others have shown the median works better.) So if this ANOVA comes up with a small P value, then it must be confused by different scatter (SD) in different groups. If the Levene P value is small then don't believe the results of the overall one-way ANOVA. See an example on pages 325-327 of Glantz.

Read more about the general topic of assumption checking after ANOVA in this article by Andy Karp.

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 η2 (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.

Analysis checklist: One-way ANOVA

One-way ANOVA compares the means of three or more unmatched groups.

✓ 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 (page 16). 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 168, 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.

Repeated-measures one-way ANOVA

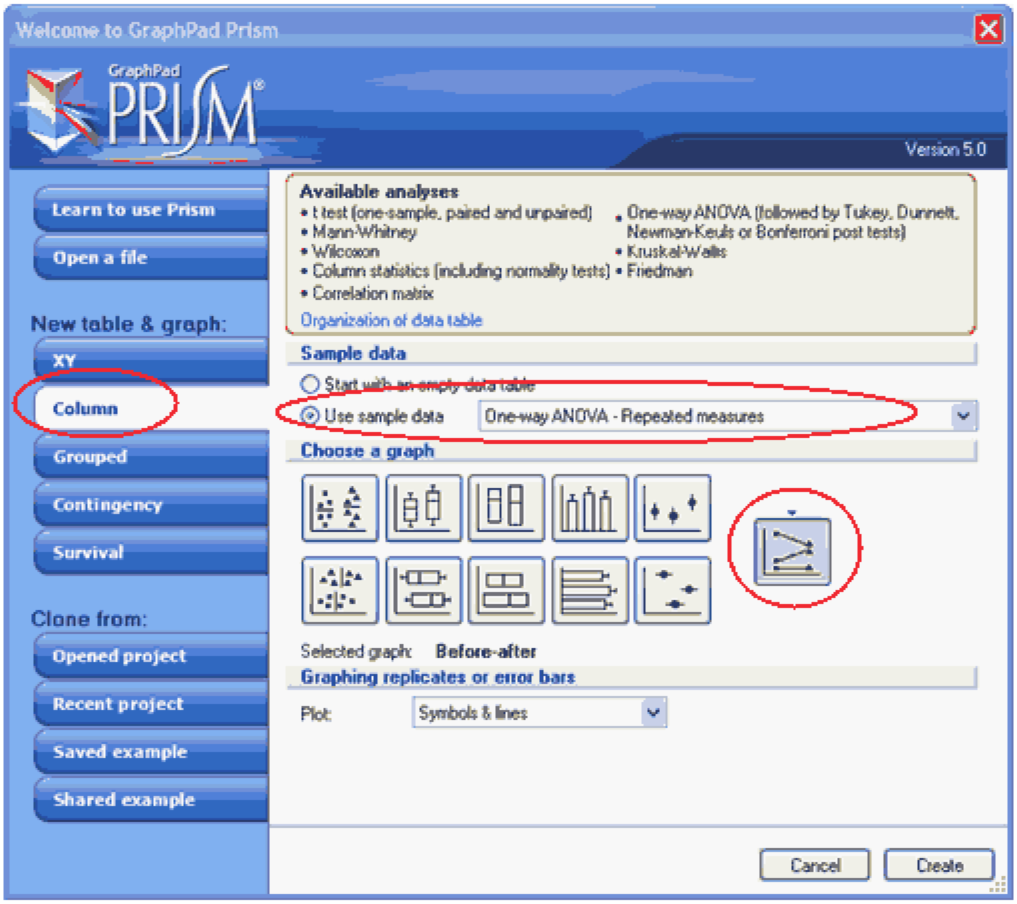
How to: Repeated measures one-way ANOVA

Repeated measures one-way ANOVA compares the means of three or more matched groups.

1. Enter data

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

If you aren't ready to enter your own data, choose to use sample data, and choose: One-way ANOVA - Repeated-measures.



Available analyses

- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: One-way ANOVA - Repeated measures

Choose a graph

Selected graph: **Before-after**

Graphing replicates or error bars

Plot: Symbols & lines

Enter the data for each group into a separate column. Enter data for each subject (or each matched set of results) on a separate row. Optionally, identify the subjects with row titles. If you leave any values blank, Prism will not be able to perform repeated measures ANOVA.

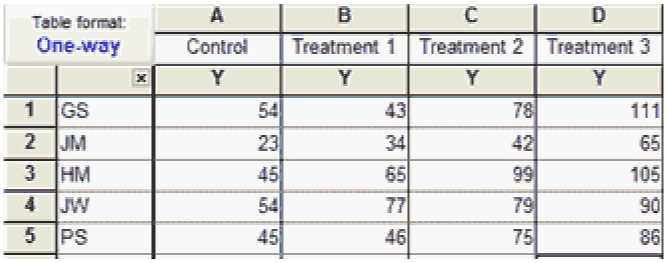


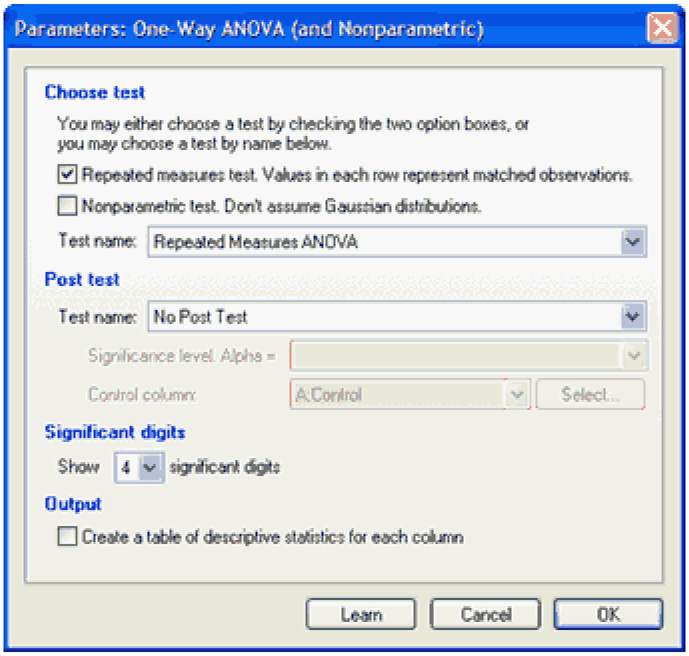
Table format: One-way	A	B	C	D
	Control	Treatment 1	Treatment 2	Treatment 3
☒	Y	Y	Y	Y
1 GS	54	43	78	111
2 JM	23	34	42	65
3 HM	45	65	99	105
4 JW	54	77	79	90
5 PS	45	46	75	86

2. Choose one-way ANOVA

1. From the data table, click Analyze on the toolbar.

2. Choose one-way ANOVA from the list of column analyses.

3. On the ANOVA dialog, choose repeated measures one-way ANOVA.



Parameters: One-Way ANOVA (and Nonparametric)

Choose test

You may either choose a test by checking the two option boxes, or you may choose a test by name below.

☑ Repeated measures test. Values in each row represent matched observations.

☐ Nonparametric test. Don't assume Gaussian distributions.

Test name: Repeated Measures ANOVA

Post test

Test name: No Post Test

Significance level. Alpha =

Control column: A:Control Select...

Significant digits

Show 4 significant digits

Output

☐ Create a table of descriptive statistics for each column

3. Choose a multiple comparison test

If you are comparing three or more groups, you may pick a post test to compare pairs of group means. Choosing an appropriate multiple comparison test 163 is not straightforward, and different statistics texts make different recommendations. With repeated-measures data, the ***test for linear trend*** often is most useful (assuming that the columns are arranged in a natural order e.g. dose or time).

The other post tests are less helpful with repeated-measures data, but you can choose them if you want.

4. Review the results and view the graph

ANOVA investigates the likelihood that the difference among the means could have been caused by chance. The most important results are the P value 139 and the results of post tests.

Before accepting the results, review the analysis checklist 142.

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.

Multiple comparisons tests and analysis checklist

Multiple comparisons tests after repeated measures ANOVA are not straightforward. Learn about interpreting the post test for linear trend 160 and post tests that compare group means 165.

Before interpreting the results, review the analysis checklist 147.

Analysis checklist: Repeated-measures one way ANOVA

Repeated measures one-way ANOVA compares the means of three or more matched 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 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.

Prism does not attempt to detect whether you violated this assumption, but some other programs do.

Kruskal-Wallis test

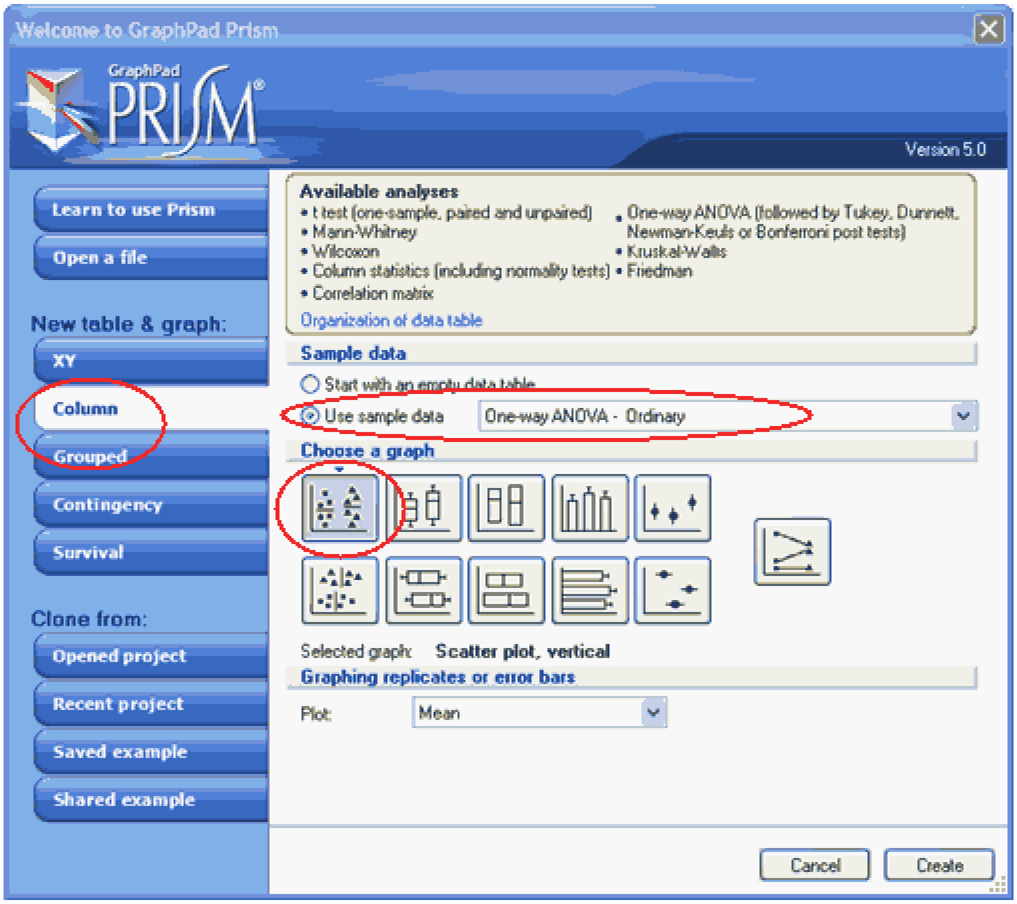
How to: Kruskal-Wallis test

The Kurskal-Wallis test is a nonparametric test that compares three or more unmatched groups.

1. Enter data

From the Welcome (or New Table and graph) dialog, choose any graph from the One-way tab. We suggest plotting a scatter graph showing every point, with a line at the median.

If you aren't ready to enter your own data, choose to use sample data, and choose: One-way ANOVA - Ordinary.



Available analyses

- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: One-way ANOVA - Ordinary

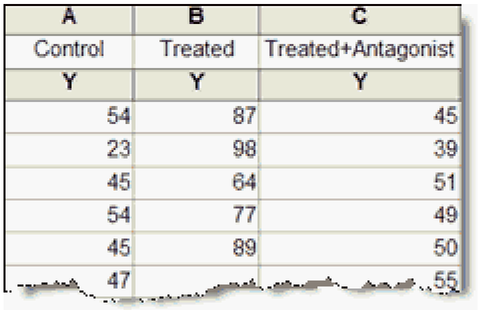
Choose a graph

Selected graph: Scatter plot, vertical

Graphing replicates or error bars

Plot: Mean

Enter the data for each group into a separate column. The two groups do not have be the same size (it's OK to leave some cells empty). Since the data are unmatched, it makes no sense to enter any row titles.



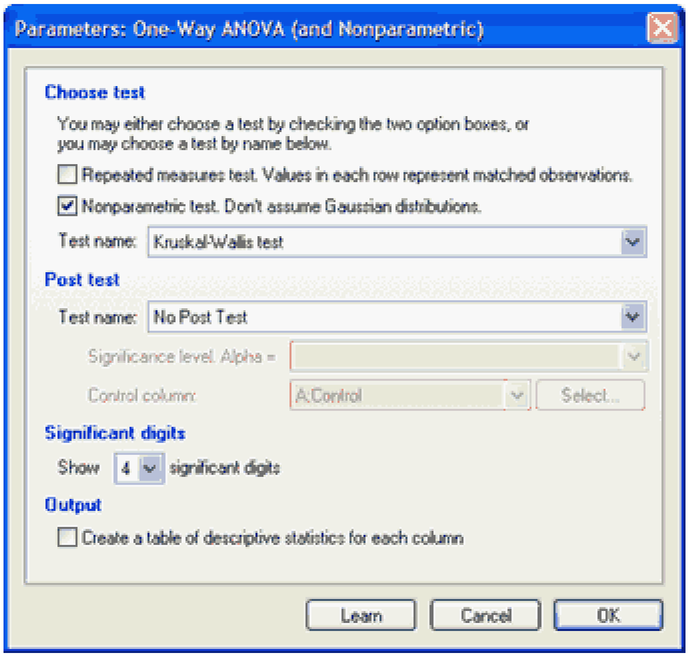
A	B	C
Control	Treated	Treated+Antagonist
Y	Y	Y
54	87	45
23	98	39
45	64	51
54	77	49
45	89	50
47		55

2. Choose the Kruskal-Wallis test

1. From the data table, click Analyze on the toolbar.

2. Choose one-way ANOVA from the list of column analyses.

3. On the ANOVA dialog, choose the Kruskal-Wallis test.



Parameters: One-Way ANOVA (and Nonparametric)

Choose test

You may either choose a test by checking the two option boxes, or you may choose a test by name below.

□ Repeated measures test. Values in each row represent matched observations.

☑ Nonparametric test. Don't assume Gaussian distributions.

Test name: Kruskal-Wallis test

Post test

Test name: No Post Test

Significance level. Alpha =

Control column: A:Control

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

3. Choose a post test

Prism only offers Dunn's post test, either to compare all pairs of groups, or just selected pairs.

4. Review the results and inspect the graph

ANOVA investigates the likelihood that the difference among the groups could have been caused by chance. The most important results are the P value 151 and the results of post tests.

Before accepting the results, review the analysis checklist 152.

Interpreting results: Kruskal-Wallis test

P value

The Kruskal-Wallis test is a nonparametric test that compares three or more unpaired 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, and there are no ties, Prism calculates an exact P value. If your samples are large, or if there are ties, 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).

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.

Unfortunately, there isn't a standard method to get a P value from these statistics when there are ties. Prism always uses the approximate method, which converts U or sum-of-ranks to a Z value. It then looks up that value on a Gaussian distribution to get a P value. The exact test is only exact when there are no ties.

If your samples are small and no two values are identical (no ties), Prism calculates an exact P value. If your samples are large or if there are ties, 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.

If you have large sample sizes and a few ties, no problem. But with small data sets or lots of ties, we're not sure how meaningful the P values are. One alternative: Divide your response into a few categories, such as low, medium and high. Then use a chi-square test to compare the groups.

Dunn's post test

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.

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 152.

Analysis checklist: Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric test that compares three or more paired or matched groups.

✓ 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.

Friedman's test

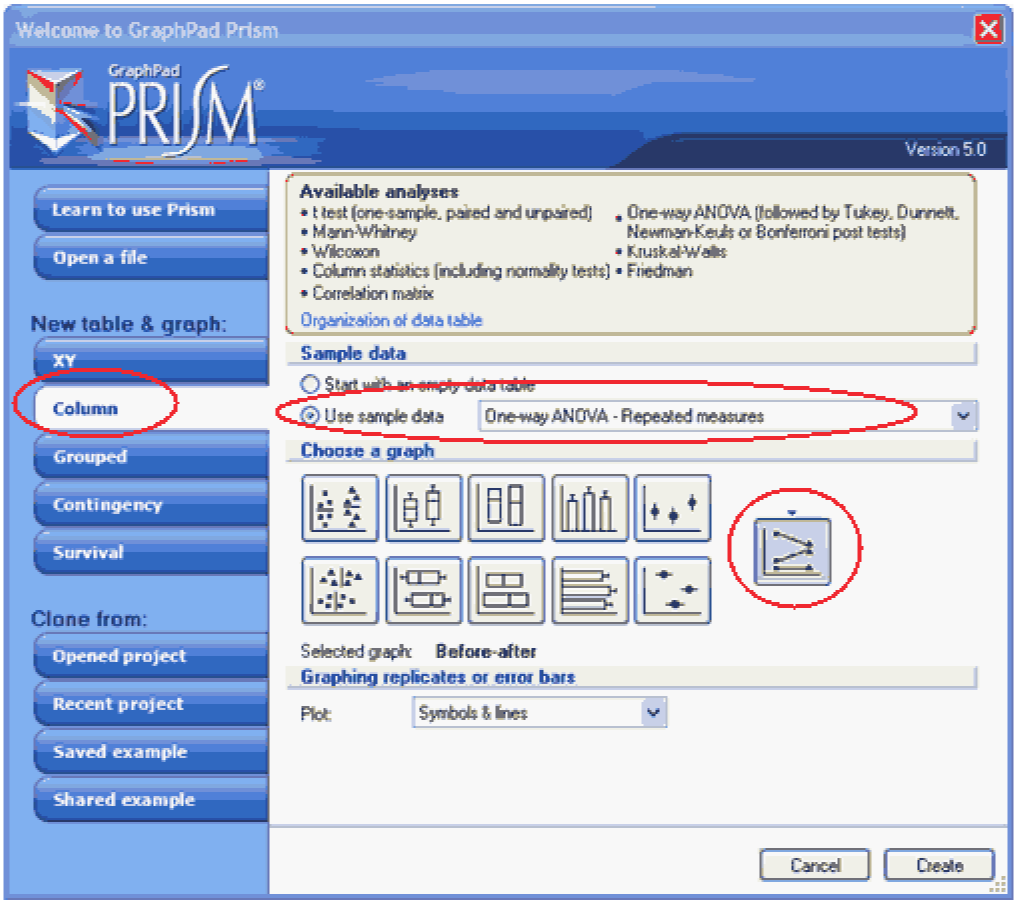
How to: Friedman test

Friedman's test is a nonparametric test to compare three or more matched groups.

1. Enter data

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

If you aren't ready to enter your own data, choose to use sample data, and choose: One-way ANOVA - Repeated-measures.



Welcome to GraphPad Prism

Available analyses

- t test (one-sample, paired and unpaired)
- Mann-Whitney
- Wilcoxon
- Column statistics (including normality tests)
- Correlation matrix
- One-way ANOVA (followed by Tukey, Dunnett, Newman-Keuls or Bonferroni post tests)
- Kruskal-Wallis
- Friedman

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: One-way ANOVA - Repeated measures

Choose a graph

Selected graph: **Before-after**

Graphing replicates or error bars

Plot: Symbols & lines

Enter the data for each group into a separate column. Enter data for each subject (or each matched set of results) on a separate row. Optionally, identify the subjects with row titles. If you leave any values blank, Prism will not be able to perform Friedman's test.

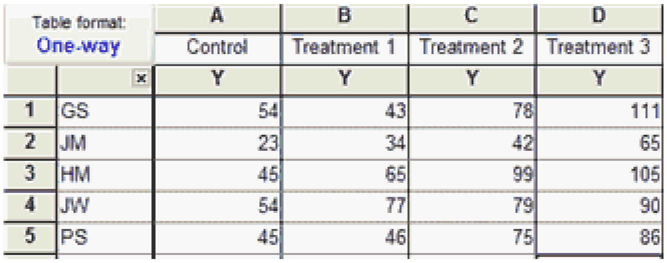
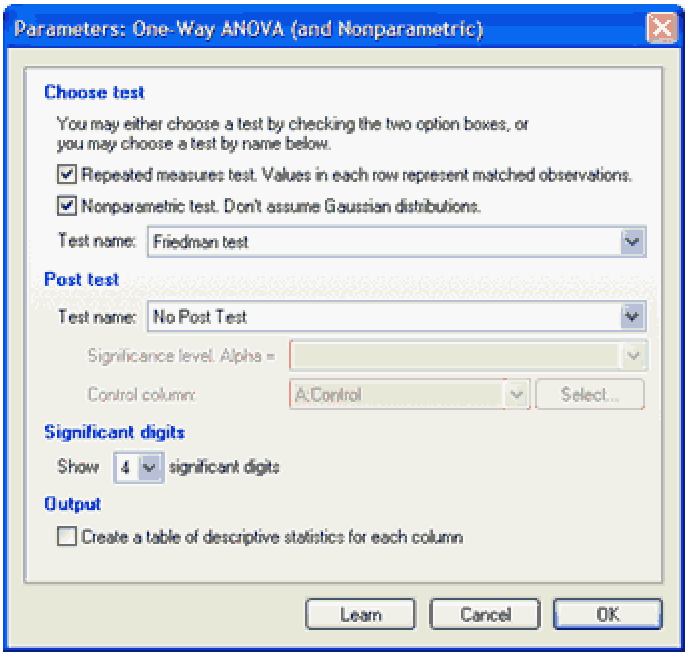


Table format: One-way	A Control	B Treatment 1	C Treatment 2	D Treatment 3
	Y	Y	Y	Y
1 GS	54	43	78	111
2 JM	23	34	42	65
3 HM	45	65	99	105
4 JW	54	77	79	90
5 PS	45	46	75	86

2. Choose Friedman's test

1. From the data table, click Analyze on the toolbar.
2. Choose one-way ANOVA from the list of column analyses.
3. On the ANOVA dialog, choose Friedman's test.



Parameters: One-Way ANOVA (and Nonparametric)

Choose test

You may either choose a test by checking the two option boxes, or you may choose a test by name below.

☑ Repeated measures test. Values in each row represent matched observations.

☑ Nonparametric test. Don't assume Gaussian distributions.

Test name: Friedman test

Post test

Test name: No Post Test

Significance level. Alpha =

Control column: A:Control

Significant digits

Show 4 significant digits

Output

□ Create a table of descriptive statistics for each column

3. Choose a post test

Prism only offers Dunn's post test, either to compare all pairs of groups, or just selected pairs.

4. Review the results and view the graph

ANOVA investigates the likelihood that the difference among the means could have been caused by chance. The most important results are the P value and the results of post tests.

Before accepting the results, review the analysis checklist 157.

Interpreting results: Friedman test

P value

The Friedman test is a nonparametric test that compares three or more 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.

Tied values

If two or more values (in the same row) have the same value, it is impossible to calculate the exact P value, so Prism computes the approximate P value.

If your samples are small and there are no ties, Prism calculates an exact P value. If your samples are large, it calculates the P value from 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 follow 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.

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.

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.

Multiple comparison tests

Key concepts: Multiple comparison tests

Terminology

The terminology is not always used consistently.

The term "multiple comparison test" applies whenever you make several comparisons at once. The term "post tests" 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 160". 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 do planned comparisons.

How multiple comparison tests work

All the multiple comparison tests (except planned comparisons) following one-way ANOVA do two things differently than a regular t test:

- First, the 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. With multiple comparison tests, scatter is 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 other aspect of multiple comparisons is an attempt to make the significance level apply to the entire family of comparisons, rather than to each comparison individually. That means if all the groups really have the same mean, there is a 5% chance that any one or more of the comparisons would reach a "statistically significant" conclusion by chance. To do this, the multiple comparison methods use a stricter definition of significance. This makes it less likely to make a Type I error (finding a 'significant' result by chance) but at the cost of decreasing the power to detect real differences. If you are only making a few comparisons, the correction is smaller than it would be if you made lots of comparisons, so the loss of power is smaller.

Note that these two aspects of multiple comparison tests have opposite effects on the power of the test. The increased power due to more degrees of freedom can be greater or smaller than the decreased power due to a stricter significance threshold. In most cases, however, the loss of power due to the stricter significance threshold is much larger than the gain due to increased numbers of degrees of freedom.

Advice: Choosing a multiple comparisons test

Multiple comparisons after ordinary one-way ANOVA

If you are comparing three or more groups with one-way ANOVA, you may pick a post test to compare pairs of group means. Choosing a multiple comparisons test is not 100% straightforward, so you may receive different recommendations depending on who you ask. Here are our recommendations:

Select ***Dunnett's*** test if one column represents control data and you wish to compare all other columns to that control column but not to each other.

Select the ***test for linear trend*** if the columns are arranged in a natural order (e.g. dose or time) and you want to test whether there is a trend such that values increase (or decrease) as you move from left to right across columns.

Select the ***Bonferroni test for selected pairs of columns*** when you only wish to compare certain column pairs. You must select those pairs based on experimental design and ideally should specify the pairs of interest before collecting any data. If you base your decision on the results (e.g., compare the smallest with the largest mean), then you have effectively compared all columns, and it is not appropriate to use the test for selected pairs.

If you want to compare all pairs of columns, choose the **Tukey** test. This is actually the Tukey-Kramer test, which includes the extension by Kramer to allow for unequal sample sizes. We recommend you don't use 163 the Newman-Keuls test or the Bonferroni test used to compare every pair of groups.

Multiple comparisons after repeated measures one-way ANOVA

The discussion above can also help you choose a multiple comparisons test after repeated measures one-way ANOVA.

All the multiple comparisons tests are based on the assumption that the values after each treatment were randomly drawn from populations with the same amount of scatter. But with some repeated measures designs, the scatter trends to increase with each sequential treatment. If the scatter systematically increases with each repeated treatment, then the multiple comparisons performed by Prism are not valid. There is an alternative approach to computing multiple comparisons, but this is not implemented in Prism. With this alternative test, each pair of groups is compared with a paired t test (without using any of the ANOVA results or degrees of freedom), with multiple comparisons being used to select the threshold for defining when a P value is low enough to be deemed "significant".

Multiple comparisons after nonparametric ANOVA

Prism only offers Dunn's post test, either to compare all pairs of groups, or just selected pairs.

For more information, 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.

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.

Post test for trend

If the columns represent ordered and equally spaced (or nearly so) groups, the post test for a linear trend determines whether the column means increase (or decrease) systematically as the columns go from left to right.

The post test for a linear trend works by calculating linear regression on group mean vs. column number. Prism reports the slope and r2, as well as the P value for the linear trend. This P value answers this 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? Equivalently, P is the chance of observing a value of r2 that high or higher, just as a consequence of random sampling.

Prism also reports a second P value testing for nonlinear variation. After correcting for the linear trend, this P value tests whether the remaining variability among column means is greater than that expected by chance. It is the chance of seeing that much variability due to random sampling.

Finally, Prism shows an ANOVA table which partitions total variability into three components: linear variation, nonlinear variation, and random (residual) variation. It is used to compute the two F ratios, which lead to the two P values. The ANOVA table is included to be complete, but it will not be of use to most scientists.

For more information about the post test for a linear trend, see the excellent text, Practical Statistics for Medical Research by DG Altman, published in 1991 by Chapman and Hall.

Planned comparisons

What are planned comparisons?

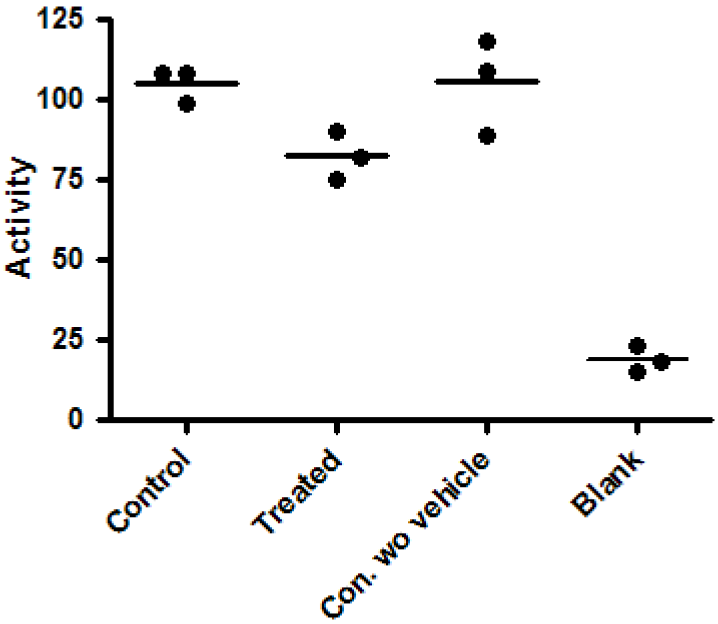
The term *planned comparison* is used when you focus in on a few scientifically sensible comparisons. You don't do every possible comparison. And you don't decide which comparisons to do after looking at the data. Instead, you decide -- as part of the experimental design -- to only make a few comparisons.

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.

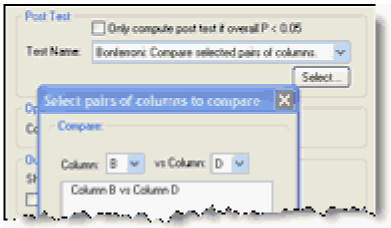
Tukey's multiple comparison tests were used to compare all pairs of means (table above). You only care about the first comparison -- control vs. treated -- which is not statistically significant (P>0.05).

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.



Post Test

□ Only compute post test if overall P < 0.05

Test Name: Bonferroni: Compare selected pairs of columns

Column: B vs Column: D

Column B vs Column D

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, you would need to do one at a time with Prism (so it doesn't do the Bonferroni correction for multiple comparisons). First do one comparison and note the results. Then change to compare a different pair of groups. When you report the results, be sure to explain that you are doing planned comparisons so have not corrected the P values or confidence intervals for multiple comparisons.

Example data analyzed by t test

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

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

How planned comparisons are calculated

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

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

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

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

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

Other multiple comparison tests

Tests that Prism offers, but we don't recommend

Bonferroni test to compare every pair of means

Prism offers the Bonferroni test for comparing every pair of means, but its only advantage over Tukey's test is that it is much easier to understand how it works. Its disadvantage is that it is too conservative, so you are more apt to miss real differences (also confidence intervals are too wide). This is a minor concern when you compare only a few columns, but is a major problem when you have many columns. Don't use the Bonferroni test with more than five groups.

Newman-Keuls

The Newman-Keuls (also called Student-Newman-Keuls test) compares all pairs of means following one-way ANOVA. The Newman-Keuls test is popular, but is no longer recommended for three reasons:

- The Newman-Keuls test does not maintain the family-wise error rate at the specified level. Most often, alpha is set to 5%. This is supposed to mean that the chance of making one or more Type I 47 errors is 5%. In fact the Newman-Keuls test doesn't do this(1). In some cases, the chance of a Type I error can be greater than 5%.
- You can't interpret a P value or significance level without stating a null hypothesis, but it is

difficult to articulate exactly what null hypotheses the Newman-Keuls test actually tests.

- Confidence intervals are more informative than significance levels, but the Newman-Keuls test cannot generate confidence intervals.

Although Prism still offers the Newman-Keuls test (for compatibility with prior versions), we recommend that you use the Tukey test instead. Unfortunately, the Tukey test has less power. This means that the Tukey test concludes that the difference between two groups is 'not statistically significant' in some cases where the Newman-Keuls test concludes that the difference is 'statistically significant'.

Tests Prism does not offer because many consider them obsolete

Fisher's LSD

While the Fisher's Least Significant Difference (LSD) test is of historical interest as the first post test ever developed, it is no longer recommended. The other tests are better. Prism does not offer the Fisher LSD test.

Fisher's LSD test does not correct for multiple comparisons as the other post tests do.

The other tests can be used even if the overall ANOVA yields a "not significant" conclusion. They set the 5% significance level for the entire family of comparisons -- so there is only a 5% chance than any one or more comparisons will be declared "significant" if the null hypothesis is true.

The Fishers LSD post test can only be used if the overall ANOVA has a P value less than 0.05. This first step sort of controls the false positive rate for the entire family of comparisons. But when doing each individual comparison, it sets the 5% significance level to apply to each individual comparison, rather than to the family of comparisons. This means it is easier to find statistical significance with the Fisher LSD test than with other post tests (it has more power), but that also means it is too easy to be mislead by false positives (you'll get bogus 'significant' results in more than 5% of experiments).

Duncan's test

This test is adapted from the Newman-Keuls method. Like the Newman-Keuls method, Duncan's test does not control family wise error rate at the specified alpha level. It has more power than the other post tests, but only because it doesn't control the error rate properly. Few statisticians, if any, recommend this test.

Multiple comparisons tests that Prism does not offer

Scheffe's test

Scheffe's test (not calculated by Prism) is used to do more all possible comparisons, including averages of groups. So you might compare the average of groups A and B with the average of groups C, D and E. Or compare group A, to the average of B-F. Because it is so versatile, Scheffe's test has less power to detect differences between pairs of groups, so should not be used when your goal is to compare one group mean with another.

Holm's test

Some statisticians highly recommend Holm's test. We don't offer it in Prism, because while it does a great job of deciding which group differences are statistically significant, it cannot compute confidence intervals for the differences between group means. (Let us know if you would like to see this in a future version of Prism.)

False Discovery Rate

The concept of the False Discovery Rate 54 is a major advance in statistics. But it is really only useful when you have calculated a large number of P values from independent comparisons, and now have to decide which P values are small enough to followup further. It is not used as a post test following one-way ANOVA.

References

1. MA Seaman, JR Levin and RC Serlin, Psychological Bulletin 110:577-586, 1991.

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 163 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 post test results valid?

You may find it surprising, but all the post 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 post 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, but these are for tests that Prism does not offer. Scheffe's test 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 Fisher's Least Significant Difference (LSD) test (which Prism does not offer). In its original form (called the restricted Fisher's LSD test) the post tests are performed only if the overall ANOVA finds a statistically significant difference among groups. But this LSD test is outmoded, and no longer recommended.

Are the results of the overall ANOVA useful at all? Or should I only look at post 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.

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 post test following ANOVA?

Post tests control for multiple comparisons. 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 post 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

Why don't post tests report exact P values?

Multiple comparison tests tell you the significance level of a comparison, but do not report P values. There are two reasons for this:

The critical values for most post tests (Bonferroni is an exception) come from tables that are difficult to compute. Prism simply reads the values from a table stored with the program, so they can only bracket the P value as less than, or greater than, a few key values (0.05, 0.01).

There is also a second, conceptual issue. The probabilities associated with post tests apply to the entire family of comparisons. So it makes sense to pick a threshold and ask which comparisons are "significant" at the proposed significance level. It makes less sense, perhaps no sense, to compute a P value for each individual comparison.

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 post 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 post 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.

V. Two-way ANOVA

You measured a response to three different drugs in both men and women. Is the response affected by drug? By gender? Are the two intertwined? These are the kinds of questions that two-way ANOVA answers.

Key concepts: Two-way ANOVA

Entering two-way ANOVA data

Groups are defined by rows and columns

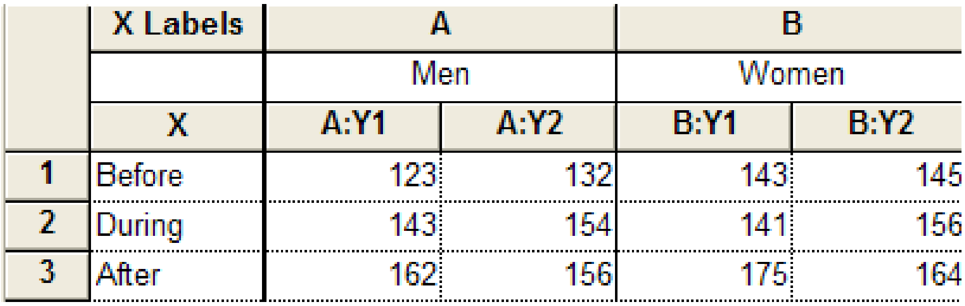
Prism organizes data for two-way ANOVA differently than do most other programs.

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.

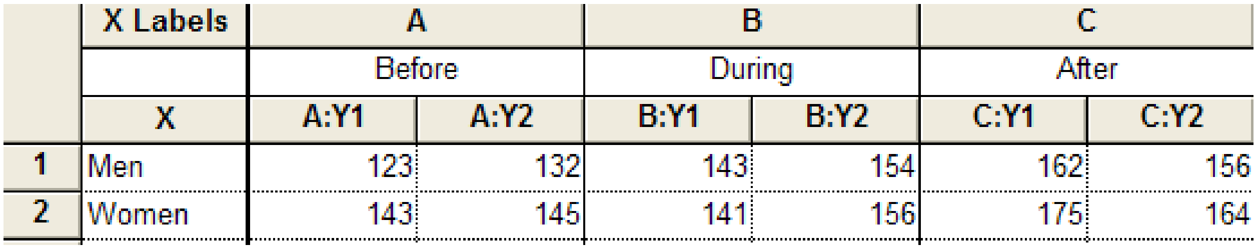
From the Welcome or New table dialog, choose the Grouped tab to set up the right kind of table.

Two ways to enter the data

Before entering data, choose which grouping variable is defined by the rows, and which is defined by the 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

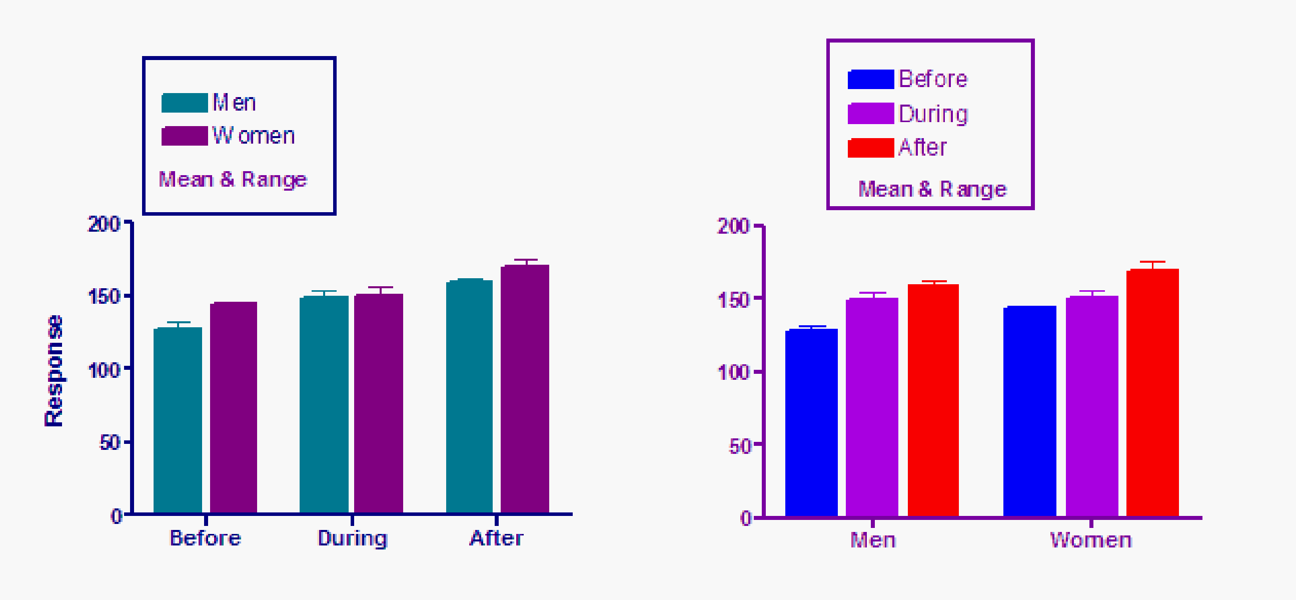
Of course, the ANOVA results will be identical no matter which way you enter the data. But the choice does matter, as it influences how the graph will appear and how Prism can do multiple comparison post tests.

Your choice affects how the graph will look

When you create a graph of the data, the points or bars from each column can have a different appearance and color.

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. This is illustrated in the left panel below.

The right panel below shows a graph of the second data table shown above. There is one bar color and fill for Before, another for During, and another for After. Men and Women appear as two bars of identical appearance.



Prism's multiple comparison post tests only compare within a row

Prism can only perform post tests within a row, comparing columns. Using the first approach, Prism compares Men vs. Women at each time point. Using the second approach, Prism compares Before vs. During, Before vs. After, and During vs. After for each gender. Arrange your data so these post tests address your scientific questions.

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 do two-way ANOVA on the results.

Choosing two-way ANOVA

What is two-way ANOVA used for?

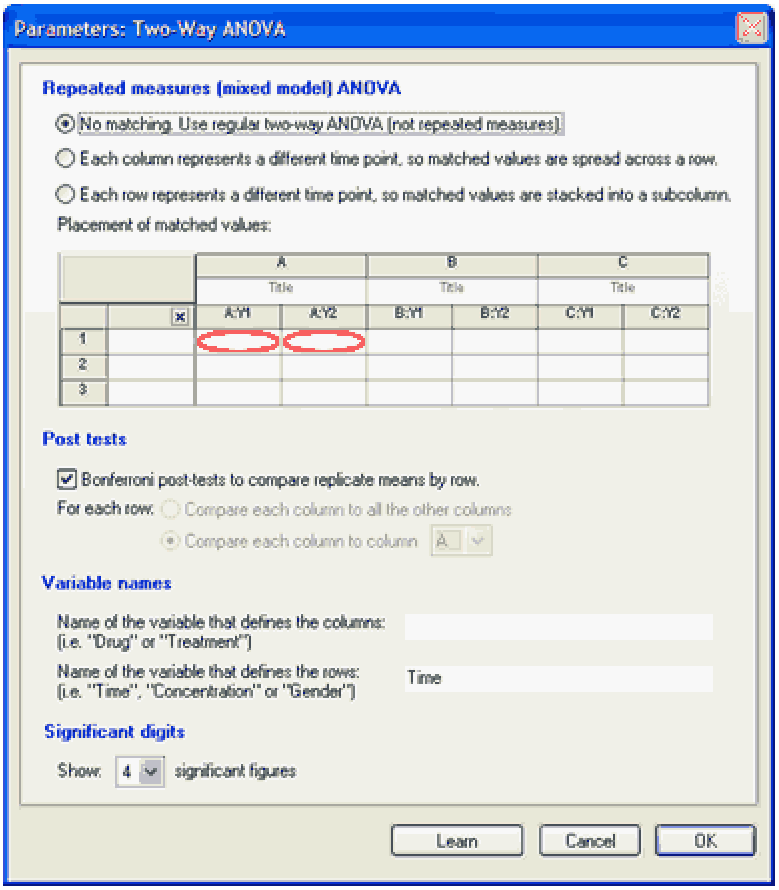
Two-way ANOVA, also called two-factor ANOVA, determines how a response is affected by two factors. For example, you might measure a response to three different drugs in both men and women. In this example, drug treatment is one factor and gender is the other.

Two-way ANOVA simultaneously asks three questions:

1. Does the first factor systematically affect the results? In our example: Are the mean responses the same for all three drugs?
2. Does the second factor systematically affect the results? In our example: Are the mean responses the same for men and women?
3. Do the two factors interact? In our example: Are the differences between drugs the same for men and women? Or equivalently, is the difference between men and women the same for all drugs?

Although the outcome measure (dependent variable) is a continuous variable, each factor must be categorical, for example: male or female; low, medium or high dose; or wild type or mutant. ANOVA is not an appropriate test for assessing the effects of a continuous variable, such as blood pressure or hormone level (use a regression technique instead).

Two-way ANOVA choices



Repeated measures (mixed model) ANOVA

● No matching. Use regular two-way ANOVA (not repeated measures).

○ Each column 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.

Placement of matched values:

		A		B		C	
		Title		Title		Title	
	☒	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1							
2							
3							

Post tests

☑ Bonferroni post-tests to compare replicate means by row.

For each row: ○ Compare each column to all the other columns

● Compare each column to column A

Variable names

Name of the variable that defines the columns: (i.e. "Drug" or "Treatment")

Name of the variable that defines the rows: (i.e. "Time", "Concentration" or "Gender") Time

Significant digits

Show: 4 significant figures

Repeated measures

Choose a repeated-measures analysis when the experiment used paired or matched subjects. Prism can calculate repeated-measures two-way ANOVA with matching by either row or column, but not both. Details. 184 This is sometimes called a **mixed model.**

You can only choose repeated measures when you have entered two or more side-by-side values into subcolumns for each row and dataset.

Post tests

Following two-way ANOVA, there are many possible multiple comparison tests that can help you focus in on what is really going on. However, Prism performs only the post tests 197 biologists use most frequently. At each row, Prism will compare each column with every column, or compare every column with a control column.

Variable names

If you enter names for the factors that define the rows and columns, the results will be easier to follow.

Point of confusion: 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.

GraphPad's advice: If one of your factors is quantitative (such as time or dose) think hard before choosing two-way ANOVA. Other analyses may make more sense.

Interpreting P values with a quantitative factor

Let's imagine you compare two treatments at six time points. The ANOVA analysis treats different time points exactly as it would treat different drugs or different species. The concept of trend is entirely ignored (except in some special post tests).

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 you probably expect no difference at early (or maybe late) 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.

ANOVA pays no attention to the order of your time points (or doses). If you randomly scramble the time points or doses, two-way ANOVA would report identical results. In other words, ANOVA ignores the entire point of the experiment, when one of the factors is quantitative.

Interpreting post tests with a quantitative factor

What about post tests?

Some scientists like to ask which is the lowest dose (or time) at which the change in response is statistically significant. Post 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), you will find that a tiny dose causes a statistically significant, but biologically trivial, effect. 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. For example, find the minimum dose that raises the pulse rate by more than 10 beats per minute.

If you look at all the post 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 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 an alternative to ANOVA. Will G Hopkins calls the alternative 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 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.

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?

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.

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 206 and analyze with a chi-square test 208.

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.

Ordinary (not repeated measures) two-way ANOVA

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. Drug treatment is one factor and gender is the other.

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 won't affect the ANOVA results, but the choice is important 169 as it affects the appearance of graphs and the kinds of post tests Prism can compare.

This page explains how to do ordinary two-way ANOVA. If you want to do repeated measures two-way ANOVA see separate examples for repeated measures by row 186 and repeated measures by column 188.

1. Create a 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, chose to use sample data and choose: Two-way ANOVA -- Ordinary.

If you plan to enter your own data, it is important that you choose the subcolumn format correctly, for the maximum number of replicates you have.

Leave the graph set to its default -- interleaved bars, vertical.

Entering averaged data

If you have already averaged your replicates in another program, you can choose (at the bottom of the dialog) to enter and plot the mean and SD (or SEM) and N. If your data has more than 52 replicates, this is the only way to enter data into Prism for two-way ANOVA.

2. Enter data

Here are the sample data:

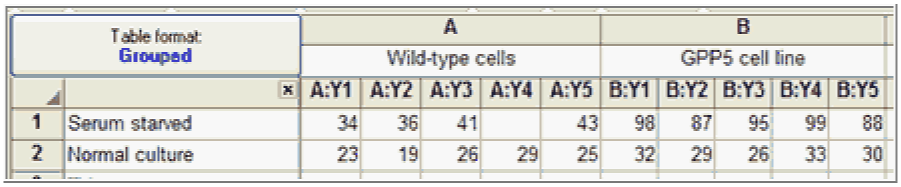


Table format: Grouped	A					B				
	Wild-type cells					GPP5 cell line				
	A:Y1	A:Y2	A:Y3	A:Y4	A:Y5	B:Y1	B:Y2	B:Y3	B:Y4	B:Y5
1 Serum starved	34	36	41		43	98	87	95	99	88
2 Normal culture	23	19	26	29	25	32	29	26	33	30

Note that one value is blank. It is fine to have some missing values, but you must have at least one value in each row for each data set. The following table cannot be analyzed by two-way ANOVA because there are no data for treated women. But it doesn't matter much that there are only two (not three) replicates for control men and treated men.

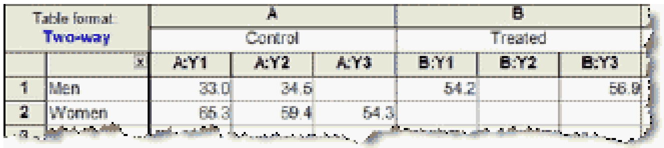


Table format: Two-way	A			B		
	Control			Treated		
	A:Y1	A:Y2	A:Y3	B:Y1	B:Y2	B:Y3
1 Men	33.0	34.5		54.2		56.9
2 Women	65.3	59.4	54.3			

If you are entering mean, SD (or SEM) and N, You must enter N (number of replicates) but it is ok if N is not always the same.

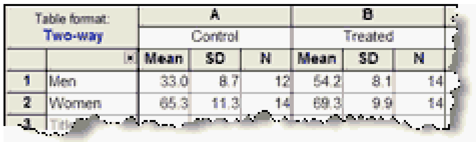
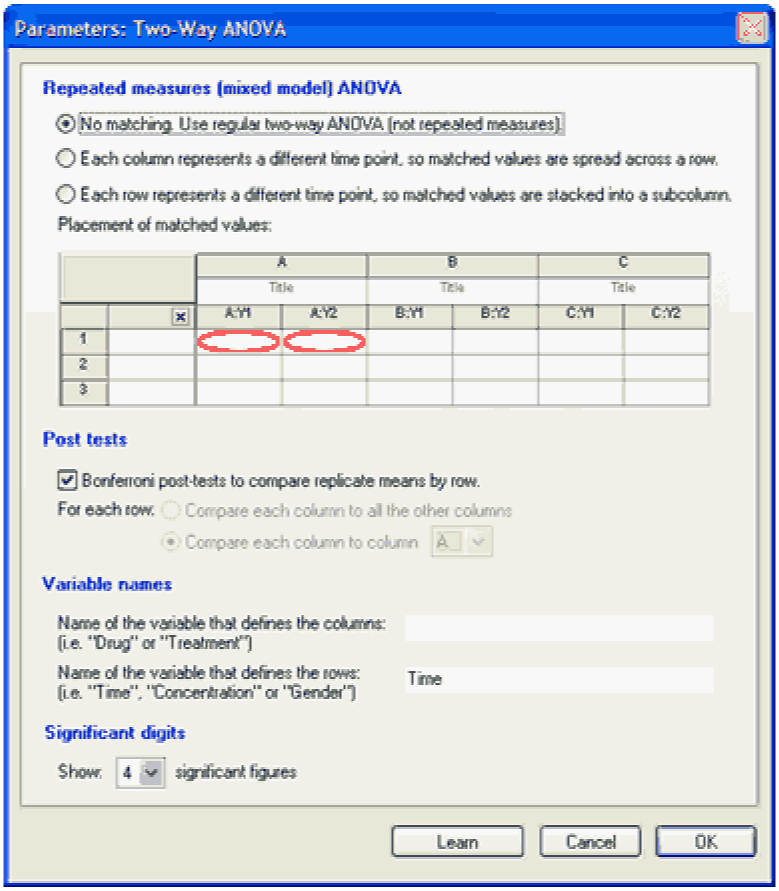


Table format: Two-way	A			B		
	Control			Treated		
	Mean	SD	N	Mean	SD	N
1 Men	33.0	8.7	12	54.2	8.1	14
2 Women	65.3	11.3	14	69.3	9.9	14

3. Choose two-way ANOVA

1. From the data table, click Analyze on the toolbar.
2. Choose **Two-way ANOVA** from the list of grouped analyses.
3. Select 'no matching' because your data doesn't involve repeated measure 184. Choose post tests 197 if they will help you interpret your results, and enter the name of the factors that define columns and rows.



Repeated measures (mixed model) ANOVA

● No matching. Use regular two-way ANOVA (not repeated measures).

○ Each column 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.

Placement of matched values:

		A		B		C	
		Title		Title		Title	
	☒	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1							
2							
3							

Post tests

☑ Bonferroni post-tests to compare replicate means by row.

For each row: ○ Compare each column to all the other columns

● Compare each column to column A

Variable names

Name of the variable that defines the columns: (i.e. "Drug" or "Treatment")

Name of the variable that defines the rows: (i.e. "Time", "Concentration" or "Gender") Time

Significant digits

Show: 4 significant figures

5. Interpret the results

Interpreting results: Two-way ANOVA 177

Graphing tips: Two-way ANOVA data 180

How Prism computes two-way ANOVA 181

Checklist: Two-way ANOVA 182

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.

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).

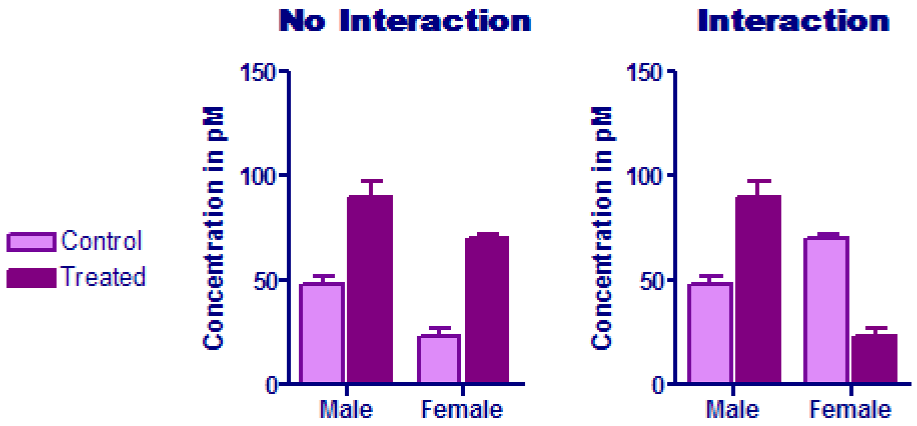
Interaction P value

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.

Tip: If the interaction is statistically significant, you won't learn much from the other two P values. Instead focus on the multiple comparison post tests 197.

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

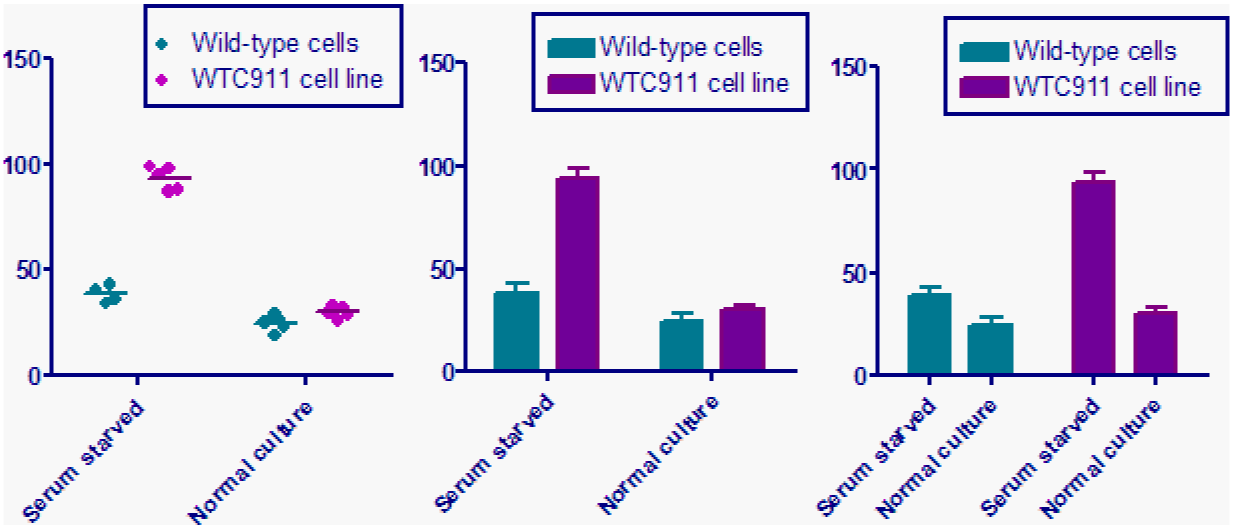
The null hypothesis is that the mean of each row (totally ignoring the columns) is the same in the overall population, and that all differences we see between row means are due to chance. In the example above, the rows represent gender, so the null hypothesis is that the mean response is the same for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

Post tests

Prism can compare columns at each row using multiple comparison post tests 197.

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.

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 in SA Glantz and BK Slinker, Primer of Applied Regression and Analysis of Variance, McGraw-Hill, 1990. 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.

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.

✓ 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.

Repeated measures two-way ANOVA

What is repeated measures?

You should choose a repeated-measures analysis when the experiment used paired or matched subjects. Prism can calculate repeated-measures two-way ANOVA with matching by either row or column, but not both. This is sometimes called a **mixed model.**

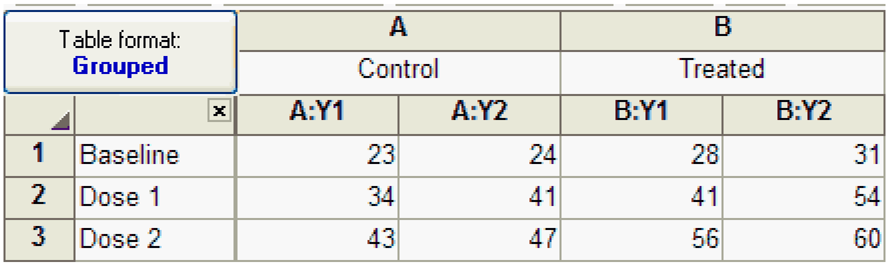


Table format: Grouped		A		B	
		Control		Treated	
		A:Y1	A:Y2	B:Y1	B:Y2
1	Baseline	23	24	28	31
2	Dose 1	34	41	41	54
3	Dose 2	43	47	56	60

The table above shows example data testing the effects of three doses of drugs in control and treated animals. The decision to use repeated measures depends on the experimental design.

Repeated measures by row

Here is an experimental design that would require analysis using repeated measures by row:

> The experiment was done with six animals, two for each dose. The control values were measured first in all six animals. Then you applied a treatment to all the animals and made the measurement again. In the table above, the value at row 1, column A, Y1 (23) came from the same animal as the value at row 1, column B, Y1 (28). The matching is by row.

Repeated measures by column

Here is an experimental design that would require analysis using repeated measures by column:

> 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

Here is an experimental design where both factors are repeated measures:

> 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).

Prism cannot perform two-way ANOVA with repeated measures in both directions, and so cannot analyze the data from this experiment.

"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.

Example without repeated measures

Finally, here is an example of an experiment done with replicates but no 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.

How to: Two-way ANOVA. Repeated measures by row

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.

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 169 as it affects the appearance of graphs and the kinds of post tests Prism can compare.

This page shows you how to enter and analyze data with repeated measurements placed on a single row. Use a different 'how to' page 188 if you enter repeated measurements in a subcolumn.

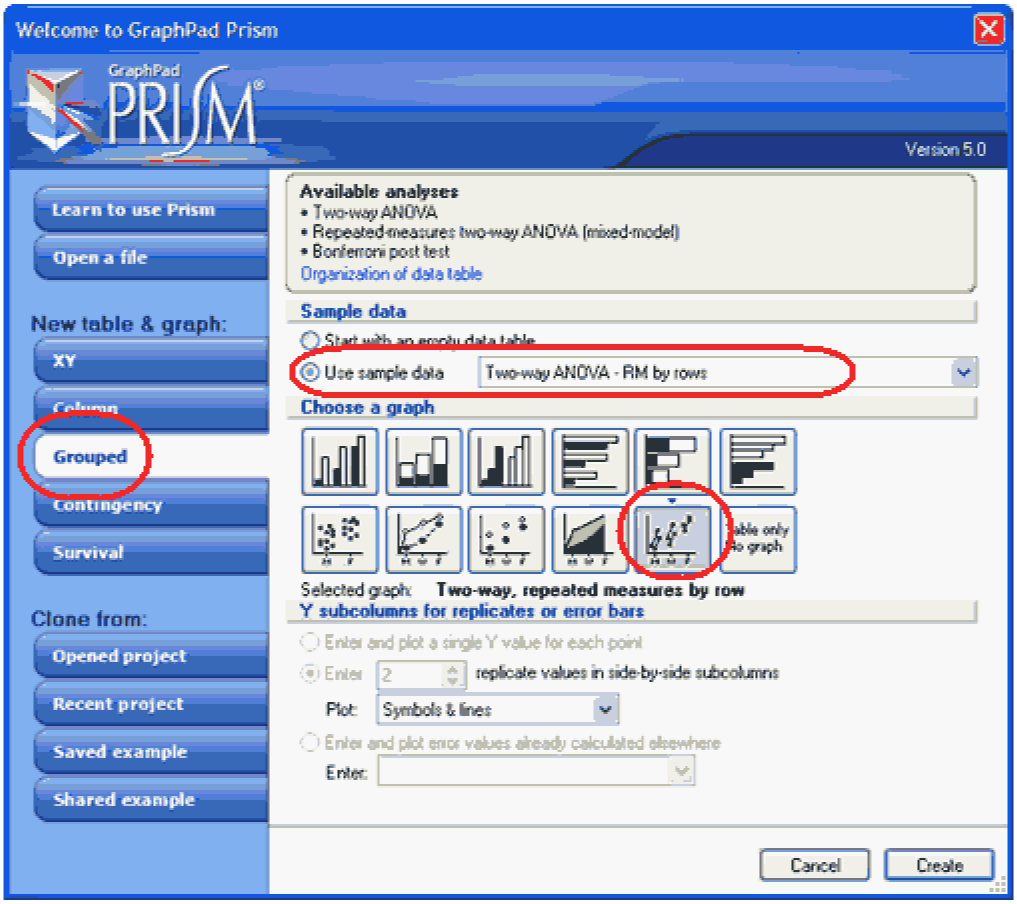
1. Create a data table and enter data

From the Welcome (or New Data Table and Graph) dialog, choose the Grouped tab.

If you are not ready to enter your own data, chose to use sample data and choose: Two-way ANOVA data -- RM by rows. ("RM" means Repeated Measures).

If you plan to enter your own data, it is important that you choose the subcolumn format correctly, for the maximum number of replicates you have.

Since your data are repeated measures, you want to make a graph that shows that. Choose the fifth choice on the second row of graph types, so values are connected properly on the graph. If you choose sample data, Prism automatically chooses the appropriate graph.



Available analyses
- Two-way ANOVA
- Repeated-measures two-way ANOVA (mixed-model)
- Bonferroni post test

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: Two-way ANOVA - RM by rows

Choose a graph

Selected graph: **Two-way, repeated measures by row**

Y subcolumns for replicates or error bars

○ Enter and plot a single Y value for each point

● Enter 2 replicate values in side-by-side subcolumns

Plot: Symbols & lines

○ Enter and plot error values already calculated elsewhere

Enter:

Cancel | Create

Enter the data putting matched values in the same row, in corresponding columns. In the sample data shown below, the two circled values represent repeated measurements in one animal.

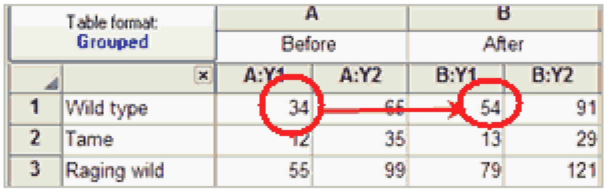
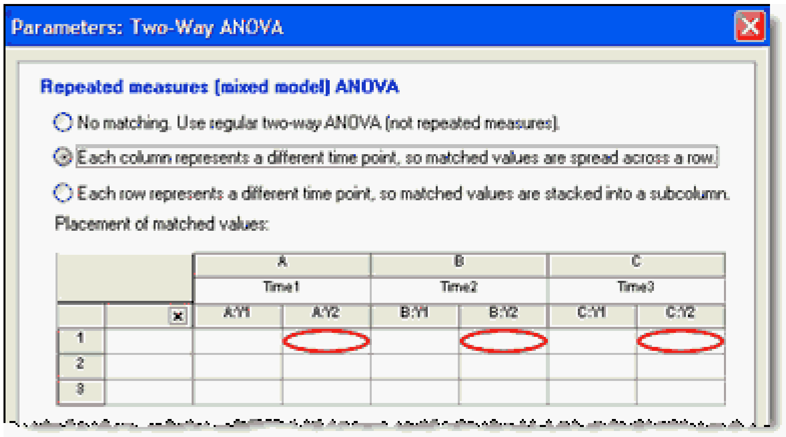


Table format: Grouped	A		B	
	Before		After	
	A:Y1	A:Y2	B:Y1	B:Y2
1 Wild type	34	65	54	91
2 Tame	12	35	13	29
3 Raging wild	55	99	79	121

2. Choose two-way ANOVA

1. From the data table, click Analyze on the toolbar.
2. Choose Two-way ANOVA from the list of grouped analyses.
3. Choose the option that specifies that matched values are spread across a row.



Parameters: Two-Way ANOVA

Repeated measures (mixed model) ANOVA

○ No matching. Use regular two-way ANOVA (not repeated measures).

● Each column 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.

Placement of matched values:

		A		B		C	
		Time1		Time2		Time3	
	☒	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1							
2							
3							

3. Interpret the results

Interpreting results: Repeated measures two-way ANOVA 194

Graphing tips: Repeated measures two-way ANOVA 191

Checklist: Repeated measures two-way ANOVA 195

How to: Two-way ANOVA. Repeated measures by column

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.

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 169 as it affects the appearance of graphs and the kinds of post tests Prism can compare.

This page shows you how to enter and analyze data with repeated measurements placed in a subcolumn. Use a different 'how to' page 186 if you enter repeated measurements in a row.

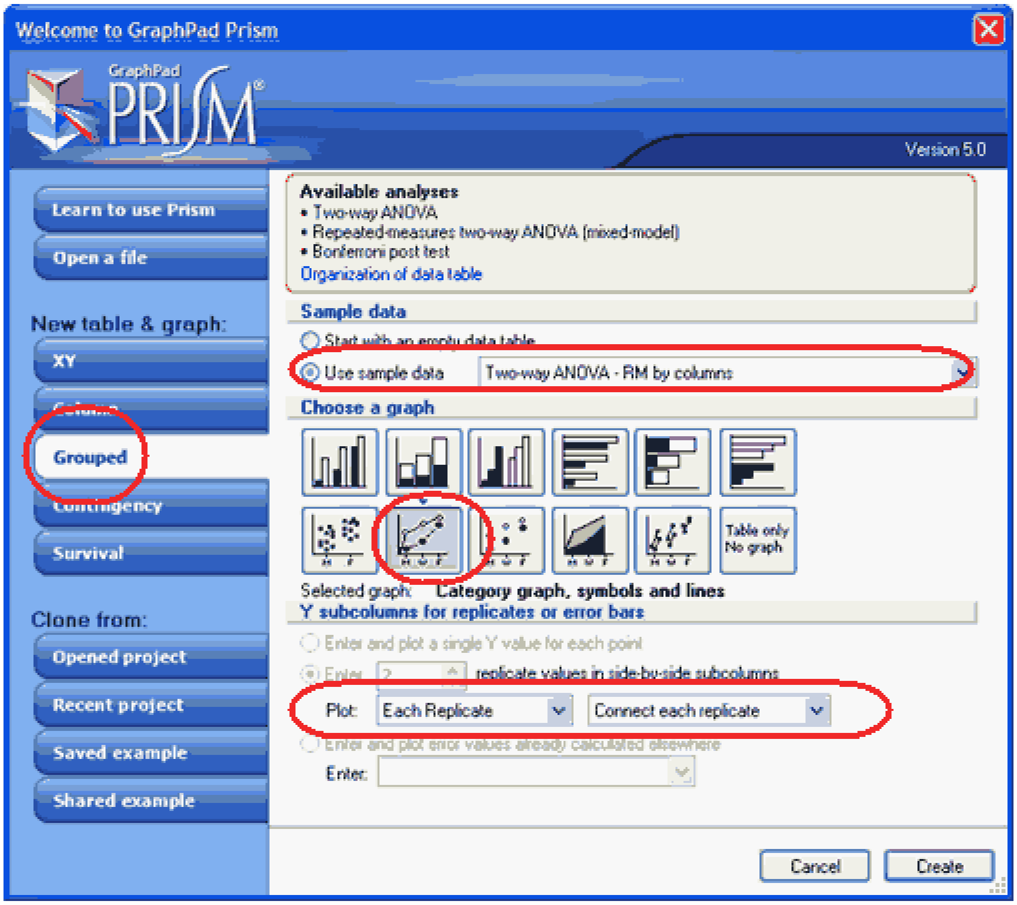
1. Create a data table and enter data

From the Welcome (or New Data Table and Graph) dialog, choose the Grouped tab.

If you are not ready to enter your own data, chose to use sample data and choose: Two-way ANOVA data -- RM by columns. ("RM" means Repeated Measures).

If you plan to enter your own data, it is important that you choose the subcolumn format correctly, for the maximum number of subjects you have with any treatment.

Since your data are repeated measures, you want to make a graph that shows that. Choose the second choice on the second row of graph types, so values are connected properly on the graph. Choose to plot each replicate, connecting each subcolumn. If you choose the sample data, Prism will automatically choose the appropriate graph.



Available analyses
- Two-way ANOVA
- Repeated-measures two-way ANOVA (mixed-model)
- Bonferroni post test

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: Two-way ANOVA - RM by columns

Choose a graph

Selected graph: **Category graph, symbols and lines**

Y subcolumns for replicates or error bars

○ Enter and plot a single Y value for each point

● Enter 2 replicate values in side-by-side subcolumns

Plot: Each Replicate; Connect each replicate

○ Enter and plot error values already calculated elsewhere

Enter:

Arrange your data so the data sets (columns) represent different levels of one factor, and different rows represent different levels of the other factor. The sample data set compares five time points in two subjects under control conditions and two after treatment.

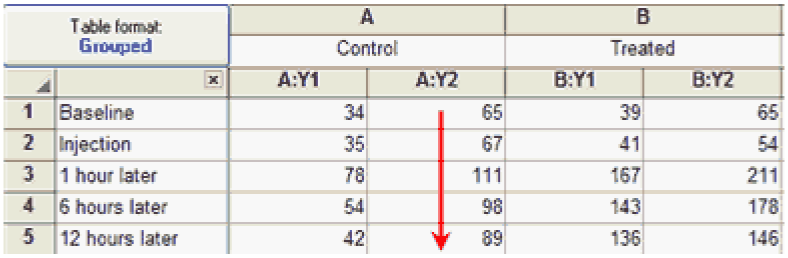


Table format: Grouped	A		B	
	Control		Treated	
	A:Y1	A:Y2	B:Y1	B:Y2
1 Baseline	34	65	39	65
2 Injection	35	67	41	54
3 1 hour later	78	111	167	211
4 6 hours later	54	98	143	178
5 12 hours later	42	89	136	146

Each subcolumn (one is marked with a red arrow above) represents repeated measurements on a single subject.

Missing values

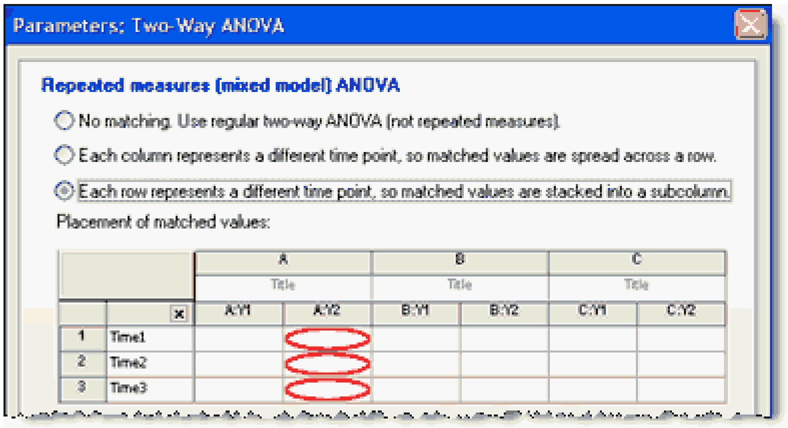
It is OK if some treatments have more subjects than others. In this case, some subcolumns will be entirely blank. But you must enter a measurement for each row for each subject. Prism cannot handle missing measurements with repeated measures ANOVA.

2. Choose two-way ANOVA

1. From the data table, click Analyze on the toolbar.

2. Choose Two-way ANOVA from the list of grouped analyses.

3. Choose the option that specifies that matched values are stacked in a subcolumn.



Parameters: Two-Way ANOVA

Repeated measures (mixed model) ANOVA

○ No matching. Use regular two-way ANOVA (not repeated measures).

○ Each column 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.

Placement of matched values:

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

Also, choose post tests 197 if they will help you interpret your results, and enter the name of the factors that define columns and rows.

3. Interpret the results

Interpreting results: Repeated measures two-way ANOVA 194

Graphing tips: Repeated measures two-way ANOVA 191

Checklist: Repeated measures two-way ANOVA 195

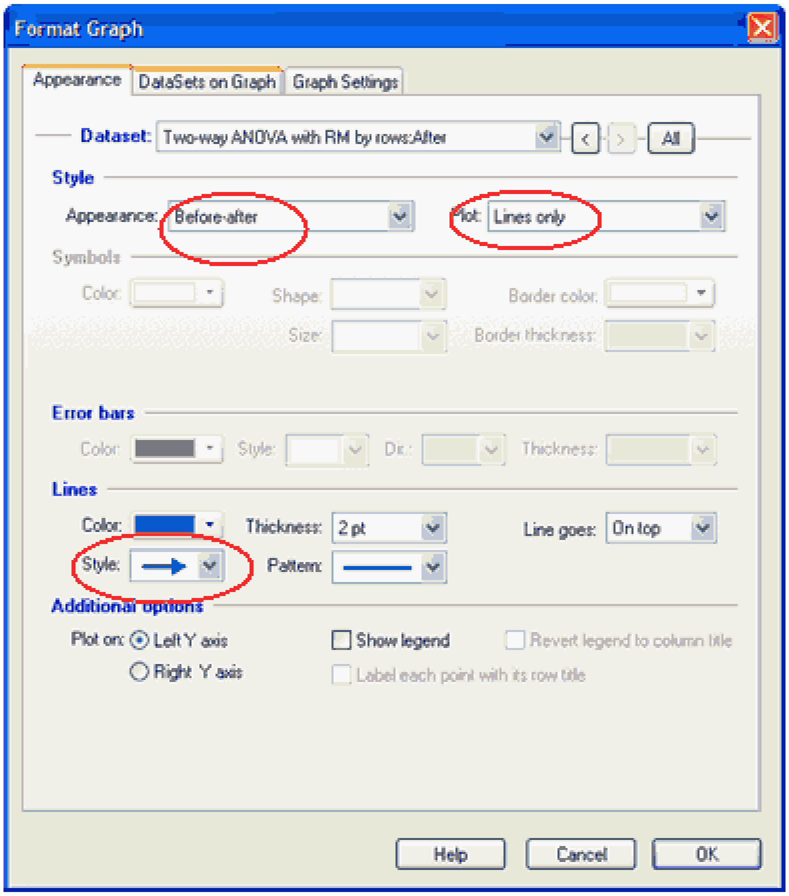
Graphing tips: Repeated measures two-way ANOVA

Graphing two-way ANOVA with repeated measures by row

From the welcome dialog, you can choose a graph designed for repeated measures by rows. This is the fifth 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 '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.



Format Graph

Appearance | DataSets on Graph | Graph Settings

Dataset: Two-way ANOVA with RM by rows:After

Style

Appearance: Before-after

Plot: Lines only

Symbols

Color:

Shape:

Border color:

Size:

Border thickness:

Error bars

Color:

Style:

Dir.:

Thickness:

Lines

Color:

Thickness: 2 pt

Line goes: On top

Style:

Pattern:

Additional options

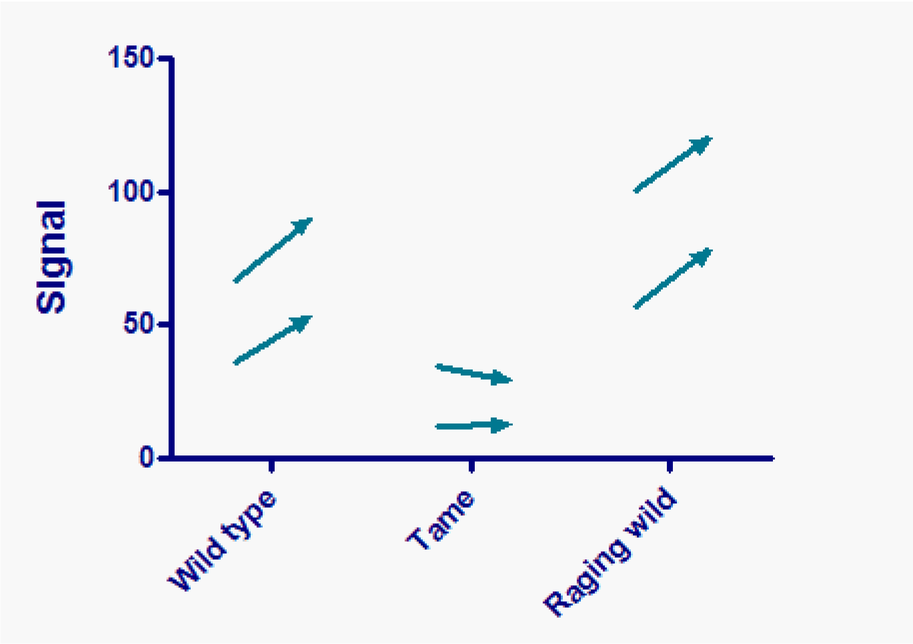
Plot on: ● Left Y axis

○ Right Y axis

□ Show legend

□ Revert legend to column title

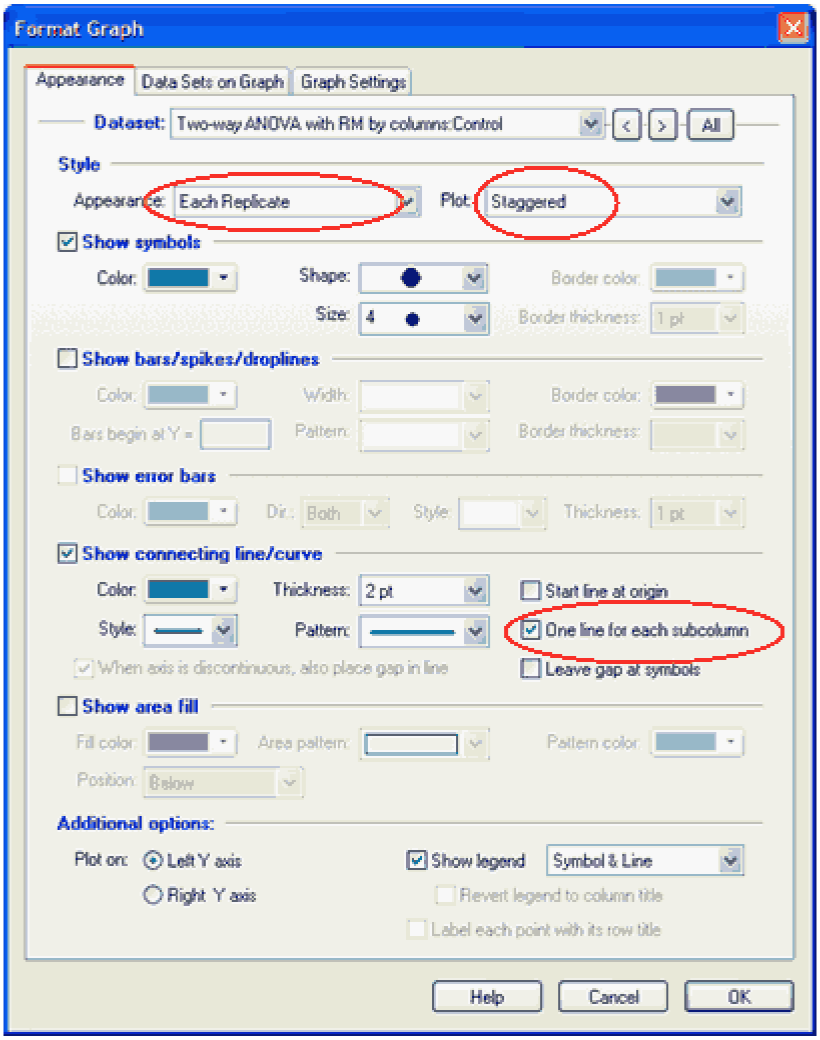
□ Label each point with its row title



Graphing two-way ANOVA with repeated measures by column

From the welcome 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:



Dataset: Two-way ANOVA with RM by columns:Control

Appearance: Each Replicate

Plot: Staggered

☑ Show symbols

Color: 🟦 Shape: ● Border color:

Size: 4 Border thickness: 1 pt

□ Show bars/spikes/droplines

Color: Width: Border color:

Bars begin at Y = Pattern: Border thickness:

□ Show error bars

Color: Dir.: Both Style: Thickness: 1 pt

☑ Show connecting line/curve

Color: 🟦 Thickness: 2 pt □ Start line at origin

Style: Pattern: ☑ One line for each subcolumn

☑ When axis is discontinuous, also place gap in line □ Leave gap at symbols

□ Show area fill

Fill color: Area pattern: Pattern color:

Position: Below

Additional options:

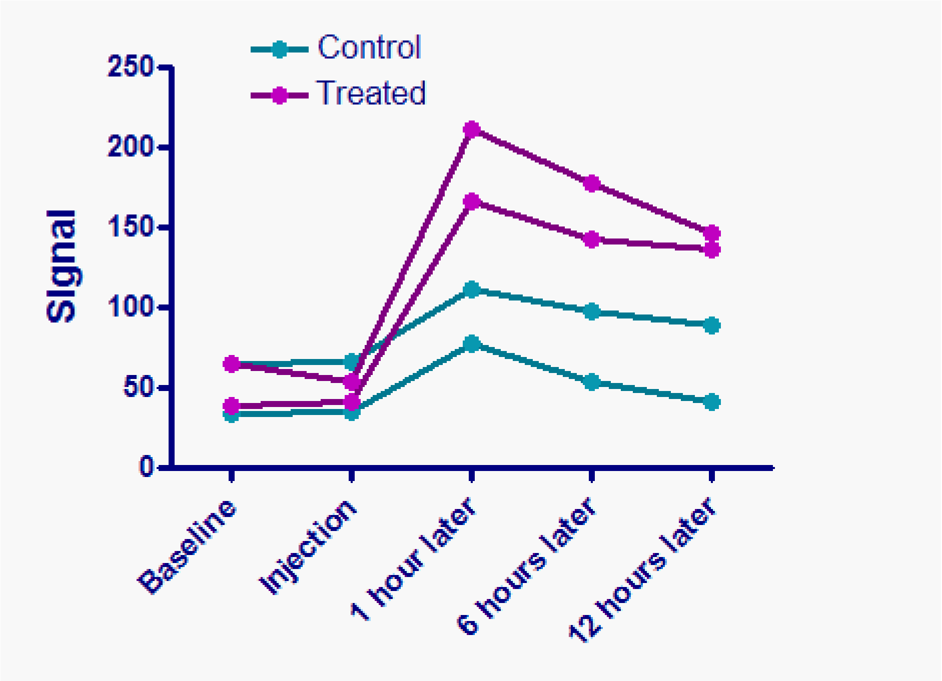
Plot on: ● Left Y axis ○ Right Y axis

☑ Show legend Symbol & Line

□ Revert legend to column title

□ Label each point with its row title

- 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'.



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 172.

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 177 first.

Repeated measures ANOVA reports an additional P value: the P value for subject (matching) This tests the null hypothesis that the matching was not effective. You expect a low P value if the repeated-measures design was effective in controlling for variability between subjects. If the P value was high, reconsider your decision to use repeated-measures ANOVA

How the repeated measures ANOVA is calculated

Prism computes repeated-measures two-way ANOVA calculations using the standard method explained especially well in SA Glantz and BK Slinker, Primer of Applied Regression and Analysis of Variance, McGraw-Hill, 1990.

Interpreting post tests after repeated measures ANOVA

The use of post tests after repeated measures ANOVA is somewhat controversial.

The post 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. Prism does this automatically.

But with repeated measures, it is common that the scatter increases with time, so later treatments give a more variable response than earlier treatments. In this case, some argue that the post tests are misleading.

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).

✓ Are the data matched?

If the matching is effective in controlling for experimental variability, repeated-measures ANOVA will be more powerful than regular ANOVA.

✓ 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 'circularity'

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.

Some advanced programs do special calculations to detect, and correct for, violations of the circularity or sphericity assumption. Prism does not.

✓ Consider alternatives to repeated measures two-way ANOVA.

Two-way ANOVA may not answer the questions your experiment was designed to address. Consider alternatives. 172

Post tests after two-way ANOVA

What post tests can Prism perform following two-way ANOVA?

Following two-way ANOVA, there are many possible multiple comparison tests that can help you focus in on what is really going on. However, Prism performs only the post tests described below (which biologists use most frequently).

- Prism can only compare one value with another value within the same row.
- Prism does not compare values within the same column.
- Prism does not compute row means and then compare one row mean with another.
- Prism does not compute column means and then compare one column mean with another.

Post tests are most often done when there are two columns of data (representing control and treated), and multiple rows (perhaps representing different time points). Prism can perform post tests to compare the control value and the treated value at each time.

If you have three columns, Prism can also perform post tests. Say that data set A is control, data set B is one treatment, and data set C is another treatment. Each row represents a different time point. Prism can do two kinds of post tests. It can do all possible comparisons at each time point (row). In this example, there are three columns, and prism can compare A with B, A with C, and B with C at each row. Or you can specify that one data set is the control (A in this case) and Prism will compare each other data set to the control. In this example, Prism would compare A with B, and A with C at each time point (row).

How Prism calculates post tests after two-way ANOVA

Prism performs post tests following two-way ANOVA using the Bonferroni method as detailed in pages 741-744 and 771 in J Neter, W Wasserman, and MH Kutner, Applied Linear Statistical Models, 3rd edition, Irwin, 1990.

The numerator is the difference between the mean response in the two data sets (usually control and treated) at a particular row (usually dose or time point). The denominator combines the number of replicates in the two groups at that dose with the mean square of the residuals (sometimes called the mean square of the error), which is a pooled measure of variability at all doses.

Statistical significance is determined by comparing the t ratio with the t distribution for the number of df shown in the ANOVA table for $MS_{residual}$, applying the Bonferroni correction for multiple comparisons. The Bonferroni correction lowers the P value that you consider to be significant to 0.05 divided by the number of comparisons. This means that if you have five rows of data with two columns, the P value has to be less than 0.05/5, or 0.01, for any particular row in order to be considered significant with P<0.05. This correction ensures that the 5% probability applies to the entire family of comparisons, and not separately to each individual comparison.

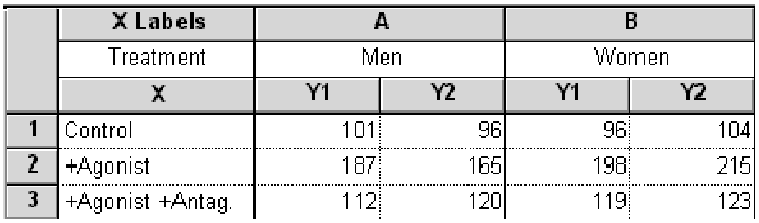
Post tests following repeated-measures two-way ANOVA use exactly the same equation if the repeated measures are by row. If the repeated measures are by column, the term (SSsubject + SSresidual)/(DFsubject + DFresidual) is used instead of MSresidual in the equation above, and the number of degrees of freedom equals the sum of DFsubject and DFresidual.

Calculating more general post tests

Need for more general post tests

Prism only performs one kind of post test following two-way ANOVA. If your experimental situation requires different post tests, you can calculate them by hand without too much trouble. Or use the free web QuickCalcs provided at graphpad.com.

Consider this example where you measure a response to a drug after treatment with vehicle, agonist, or agonist+antagonist, in both men and women.



	X Labels	A		B	
	Treatment	Men		Women	
	X	Y1	Y2	Y1	Y2
1	Control	101	96	96	104
2	+Agonist	187	165	198	215
3	+Agonist +Antag.	112	120	119	123

Prism will compare the two columns at each row. For this example, Prism's built-in post tests compare the two columns at each row, thus asking:

- Do the control responses differ between men and women?
- Do the agonist-stimulated responses differ between men and women?
- Do the response in the presence of both agonist and antagonist differ between men and

women?

If these questions match your experimental aims, Prism's built-in post tests will suffice. Many biological experiments compare two responses at several time points or doses, and Prism built-in post tests are just what you need for these experiments. But you may wish to perform different post tests. In this example, based on the experimental design above, you might want to ask these questions:

- For men, is the agonist-stimulated response different than control? (Did the agonist work?)
- For women, is the agonist-stimulated response different than control?
- For men, is the agonist response different than the response in the presence of agonist plus antagonist? (Did the antagonist work?)
- For women, is the agonist response different than the response in the presence of agonist plus antagonist?
- For men, is the response in the presence of agonist plus antagonist different than control? (Does the antagonist completely block agonist response?)
- For women, is the response in the presence of agonist plus antagonist different than control?

One could imagine making many more comparisons, but we'll make just these six. The fewer comparisons you make, the more power you'll have to find differences, so it is important to focus on the comparisons that make the most sense. But you must choose the comparisons based on experimental design and the questions you care about. Ideally you should pick the comparisons before you see the data. It is not appropriate to choose the comparisons you are interested in after seeing the data.

How to do the calculations

For each comparison (post test) you want to know:

- What is the 95% confidence interval for the difference?
- Is the difference statistically significant (P<0.05)?

Although Prism won't calculate these values for you, you can easily do the calculations yourself, starting from Prism's ANOVA table. For each comparison, calculate the confidence interval for the difference between means using this equation (from pages 741-744 and 771, J Neter, W Wasserman, and MH Kutner, Applied Linear Statistical Models, 3rd edition, Irwin, 1990).

$$(mean_1 - mean_2) - t^* \sqrt{MS_{residual}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$to$$

$$(mean_1 - mean_2) + t^* \sqrt{MS_{residual}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

In this equation, mean1 and mean 1 are the means of the two groups you are comparing, and N1 and N2 are their sample size. MSresidual is reported in the ANOVA results, and is the same for all post tests.

The variable t* is the critical value from the student t distribution, using the Bonferroni correction for multiple corrections. When making a single confidence interval, t* is the value of the t ratio that corresponds to a two-tail P value of 0.05 (or whatever significance level you chose). If you are making six comparisons, t* is the t ratio that corresponds to a P value of 0.05/6, or 0.00833. Find the value using this Excel formula =TINV(0.00833,6), which equals 3.863. The first parameter is the significance level corrected for multiple comparisons; the second is the number of degrees of freedom for the ANOVA (residuals for regular two-way ANOVA, 'subject' for repeated measures). The value of t* will be the same for each comparison. Its value depends on the degree of confidence you desire, the number of degrees of freedom in the ANOVA, and the number of comparisons you made.

To determine significance levels, calculate for each comparison:

$$t = \frac{\text{mean}_1 - \text{mean}_2}{\sqrt{\text{MS}_{\text{residual}}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

The variables are the same as those used in the confidence interval calculations, but notice the key difference. Here, you calculate a t ratio for each comparison, and then use it to determine the significance level (as explained in the next paragraph). When computing a confidence interval, you choose a confidence level (95% is standard) and use that to determine a fixed value from the t distribution, which we call t*. Note that the numerator is the absolute value of the difference between means, so the t ratio will always be positive.

To determine the significance level, compare the values of the t ratio computed for each comparison against the standard values, which we abbreviate t*. For example, to determine whether the comparison is significant at the 5% level (P<0.05), compare the t ratios computed for each comparison to the t* value calculated for a confidence interval of 95% (equivalent to a significance level of 5%, or a P value of 0.05) corrected for the number of comparisons and taking into account the number of degrees of freedom. As shown above, this value is 3.863. If a t ratio is greater than t*, then that comparison is significant at the 5% significance level. To determine whether a comparison is significant at the stricter 1% level, calculate the t ratio corresponding to a confidence interval of 99% (P value of 0.01) with six comparisons and six degrees of freedom. First divide 0.01 by 6 (number of comparisons), which is 0.001667. Then use the Excel formula =TINV(0.001667,6) to find the critical t ratio of 5.398. Each comparison that has a t ratio greater than 5.398 is significant at the 1% level.

Tip: All these calculations can be performed using a free QuickCalcs web calculator.

Example

For this example, here are the values you need to do the calculations (or enter into the web calculator).

Comparison	**Mean1**	**Mean2**	**N1**	**N2**
1: Men. Agonist vs. control	176.0	98.5	2	2
2: Women. Agonist vs. control	206.5	100.0	2	2
3: Men. Agonist vs. Ag+Ant	176.0	116.0	2	2

4: Women. Agonist vs. Ag+Ant	206.5	121.0	2	2
5: Men Control vs. Ag+Ant	98.5	116.0	2	2
6: Women. Control vs. Ag+Ant	100.0	121.0	2	2

And here are the results:

Comparison	**Significant? (P < 0.05?)**	**t**
1: Men. Agonist vs. control	Yes	8.747
2: Women. Agonist vs. control	Yes	12.020
3: Men. Agonist vs. Ag+Ant	Yes	6.772
4: Women. Agonist vs. Ag+Ant	Yes	9.650
5: Men Control vs. Ag+Ant	No	1.975
6: Women. Control vs. Ag+Ant	No	2.370

Comparison	**Mean1 - Mean2**	**95% CI of difference**
1: Men. Agonist vs. control	+ 77.5	+ 43.3 to + 111.7
2: Women. Agonist vs. control	+ 106.5	+ 72.3 to + 140.7
3: Men. Agonist vs. Ag+Ant	+ 60.0	+ 25.8 to + 94.2
4: Women. Agonist vs. Ag+Ant	+ 85.5	+ 51.3 to + 119.7
5: Men Control vs. Ag+Ant	-17.5	-51.7 to + 16.7
6: Women Control vs. Ag+Ant	-21.0	-55.2 to + 13.2

The calculations account for multiple comparisons. This means that the 95% confidence level applies to all the confidence intervals. You can be 95% sure that all the intervals include the true value. The 95% probability applies to the entire family of confidence intervals, not to each individual interval. Similarly, if the null hypothesis were true (that all groups really have the same mean, and all observed differences are due to chance) there will be a 95% chance that all comparisons will be not significant, and a 5% chance that any one or more of the comparisons will be deemed statistically significant with P< 0.05.

For the sample data, we conclude that the agonist increases the response in both men and women. The combination of antagonist plus agonist decreases the response down to a level that is indistinguishable from the control response.

VI. 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?

The Confidence Interval of a proportion

An example of the confidence interval of a proportion

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 13.73 percent.

How to compute the 95% CI of a proportion

Prism does not compute the confidence interval of a single proportion, but does compute the confidence interval of each of two proportions when analyzing a 2x2 contingency table. 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 also 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. The 95% CI is calculated using the following equation (note that the variable “p” as used here is completely distinct from p values) :

$$p' = \frac{\#\text{ "successes"} + 2}{\#\text{of experiment s} + 4} = \frac{S+2}{N+4}$$

$$p' - 1.96 \cdot \sqrt{\frac{p'(1-p')}{N+4}} \quad \text{to} \quad \text{p}' + 1.96 \cdot \sqrt{\frac{p'(1-p)'}{N+4}}$$

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.

This method works very well. For any values of S and N, there is close to a 95% chance that it contains the true proportion. With some values of S and N, the degree of confidence can be a bit less than 95%, but it is never less than 92%.

Where did the numbers 2 and 4 in the equation come from? Those values are actually $z^2/2$ and z^2, where z is a critical value from the Gaussian distribution. Since 95% of all values of a normal distribution lie within 1.96 standard deviations of the mean, z = 1.96 (which we round to 2.0) for 95% confidence intervals.

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.

One of the GraphPad QuickCalcs free web calculators computes the confidence interval of a proportion using both methods.

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.

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.

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.

Contingency tables

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

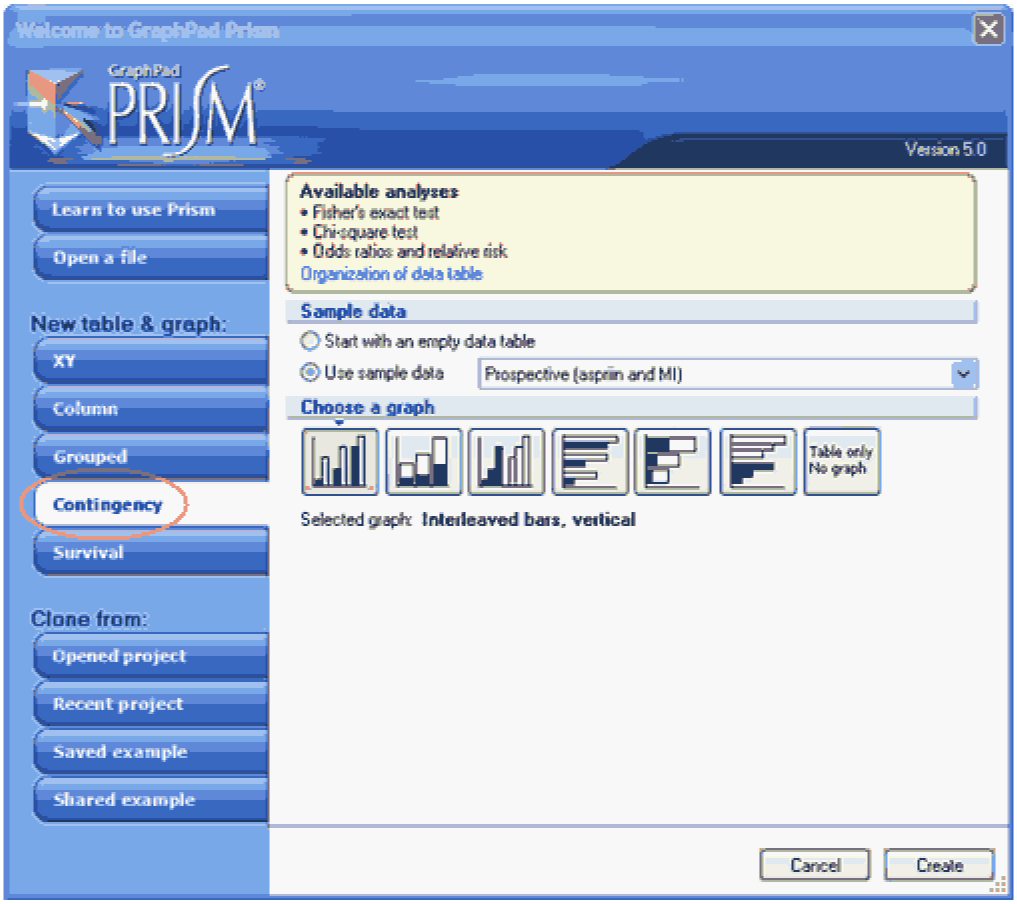
Contingency tables analyze data where the outcome is categorical, and where there is one independent (grouping) variable that is also categorical. If your experimental design is more complicated, you need to use logistic regression which Prism does not offer. Logistic regression is used when the outcome 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.

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 to use sample data, and choose any of the provided data sets.



Welcome to GraphPad Prism

Available analyses
- Fisher's exact test
- Chi-square test
- Odds ratios and relative risk

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: Prospective (asprin and MI)

Choose a graph

Selected graph: Interleaved bars, vertical

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. Don't enter averages, percentages or rates.

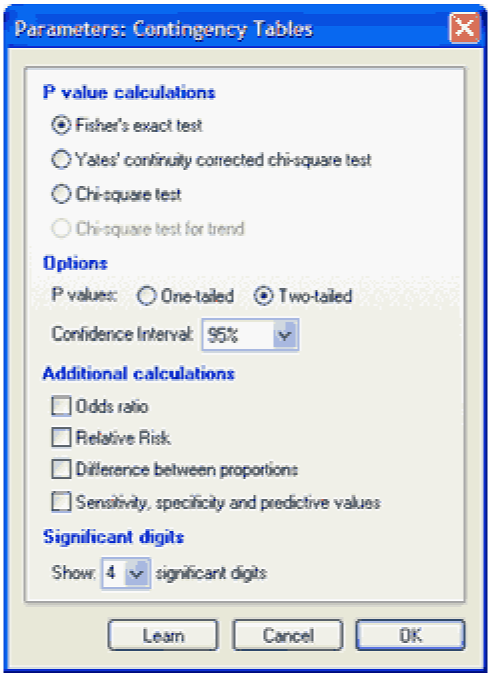
If your experimental design matched patients and controls, you should **not** analyze your data with contingency tables. Instead you should use McNemar's test. This test is not offered by Prism, but it is calculated by the free QuickCalcs web calculators available on www.graphpad.com.

3. Analyze

From the data table, click Analyze 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:



Parameters: Contingency Tables

P value calculations

● Fisher's exact test

○ Yates' continuity corrected chi-square test

○ Chi-square test

○ Chi-square test for trend

Options

P values: ○ One-tailed ● Two-tailed

Confidence Interval: 95%

Additional calculations

□ Odds ratio

□ Relative Risk

□ Difference between proportions

□ Sensitivity, specificity and predictive values

Significant digits

Show: 4 significant digits

We suggest you always choose Fisher's exact test 210 with a two-tail P value 34.

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 211

Interpreting results: sensitivity and specificity 212

Interpreting results: P values (from contingency tables) 213

Analysis checklist: Contingency tables 214

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.

GraphPad's advice: Choose Fishers tests, unless someone requires you to use chi-square test. If your values are huge, Prism will override your choice and compute the chi-square test, which is very accurate with large values.

Conventional advice

In the days before computers were readily available, people analyzed contingency tables by hand, or using a calculator, using chi-square tests. But the chi-square test is only an approximation. The **Yates' continuity correction** is designed to make the chi-square approximation better, but it overcorrects 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.

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.

Fisher's test. Exactly correct answer to wrong question?

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. Since the constraints of your study design don't 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. Berger modified this test to one that is easier to calculate yet more powerful. He also provides an online web calculator so you can try it yourself.

It is worth noting that Fisher convinced Barnard to repudiate his test!

At this time, we do not plan to implement Bernard's or Berger's test in Prism. There certainly does not seem to be any consensus among statisticians that these tests are preferred. But let us know if you disagree.

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 (P1-P2) 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

If your data represent results of a case-control retrospective study, choose to report the results as an 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. With case-control data, direct calculations of the relative risk or the difference between proportions should not be performed, as the results are not meaningful.

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).

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.

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” 39.

You will interpret the results differently depending on whether the P value is small 35 or large 36.

Why isn't the P value 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.

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).

✓ 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 (which Prism does not do, but can be performed by GraphPad's QuickCalcs web page at www.graphpad.com). 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.

✓ 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.

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.

VII. 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.

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

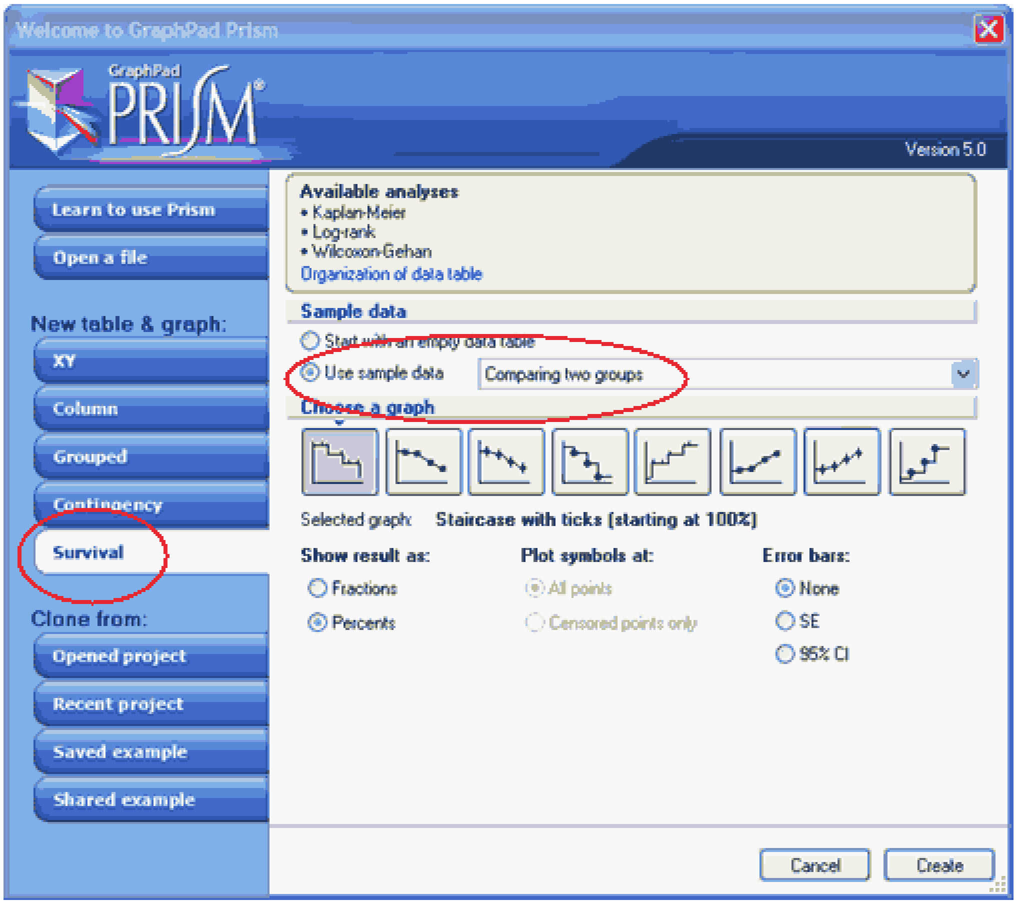
The analyses built in to Prism can compare the survival curves of two or more groups. But these methods (logrank test, Gehan-Breslow-Wilcoxon test) cannot handle data where subjects in the groups are matched, or when you also want to adjust for age or gender or other variables. For this kind of analysis, you need to use proportional hazards regression, which Prism does not do.

How to: Survival analysis

1. Create a survival table

From the Welcome or New Table dialog, choose the Survival tab, and choose the kind of survival graph you want.

If you aren't ready to enter your own data yet, choose to use sample data, and choose one of the sample data sets.



Available analyses
- Kaplan-Meier
- Log-rank
- Wilcoxon-Gehan

Organization of data table

Sample data

○ Start with an empty data table

● Use sample data: Comparing two groups

Choose a graph

Selected graph: Staircase with ticks (starting at 100%)

Show result as: ○ Fractions ● Percents

Plot symbols at: ● All points ○ Censored points only

Error bars: ● None ○ SE ○ 95% CI

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 217 at that time. Every subject in a survival study either dies or is censored.

- Enter subjects for each treatment group into a different Y column. Place the X values for the subjects for the first group at the top of the table with the Y codes in the first Y column. Place the X values for the second group of subjects beneath those for the first group (X values do not have to be sorted, and the X column may well contain the same value more than once). Place the corresponding Y codes in the second Y column, leaving the first column blank. In the example below, data for group A were entered in the first 14 rows, and data for group B started in row 15.

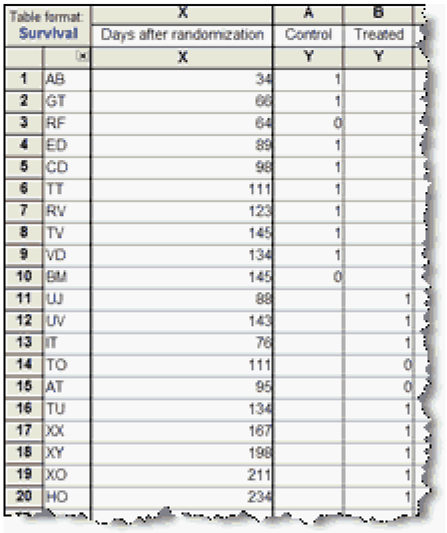


Table format: Survival		X	A	B
		Days after randomization	Control	Treated
		X	Y	Y
1	AB	34	1	
2	GT	66	1	
3	RF	64	0	
4	ED	89	1	
5	CD	98	1	
6	TT	111	1	
7	RV	123	1	
8	TV	145	1	
9	VD	134	1	
10	BM	145	0	
11	UJ	88		1
12	UV	143		1
13	IT	76		1
14	TO	111		0
15	AT	95		0
16	TU	134		1
17	XX	167		1
18	XY	198		1
19	XO	211		1
20	HO	234		1

- If the treatment groups are intrinsically ordered (perhaps increasing dose) maintain that order when entering data. Make sure that the progression from column A to column B to column C follows the natural order of the treatment groups. If the treatment groups don't have a natural order, it doesn't matter how you arrange them.

Entering data for survival studies can be tricky. See answers to common questions 221, an example of a clinical study 223, and an example of an animal study 225.

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 228

Interpreting results: Comparing two survival curves 229

Interpreting results: Comparing three or more survival curves 231

Analysis checklist: Survival analysis 234

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.

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 217 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 data in any order you want.

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.

If the subject died right after enrollment, should I enter the patient with X=0?

No. The time must exceed zero for all subjects.

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.

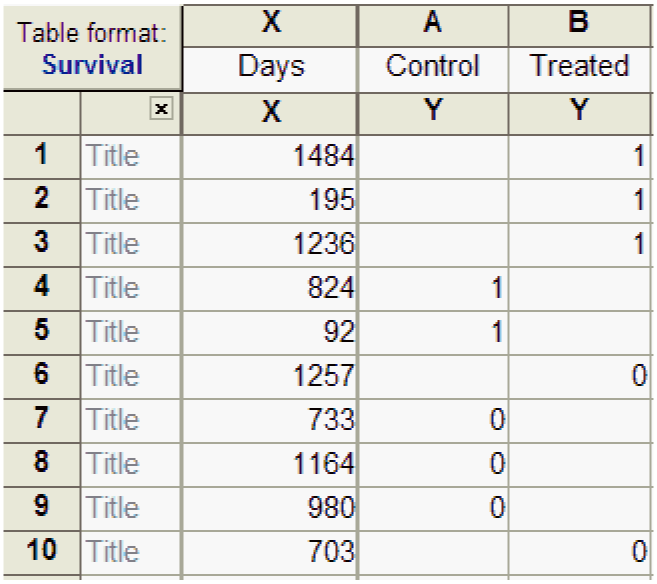


Table format: Survival		X	A	B
		Days	Control	Treated
	☒	X	Y	Y
1	Title	1484		1
2	Title	195		1
3	Title	1236		1
4	Title	824	1	
5	Title	92	1	
6	Title	1257		0
7	Title	733	0	
8	Title	1164	0	
9	Title	980	0	
10	Title	703		0

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.

Example of survival data from an animal study

This example is an animal study that followed animals for 28 days after treatment. All five control animals survived the entire time. Three of the treated animals died, at days 15, 21 and 26. The other two treated animals were still alive at the end of the experiment on day 28. Here is the data entered for survival analysis.

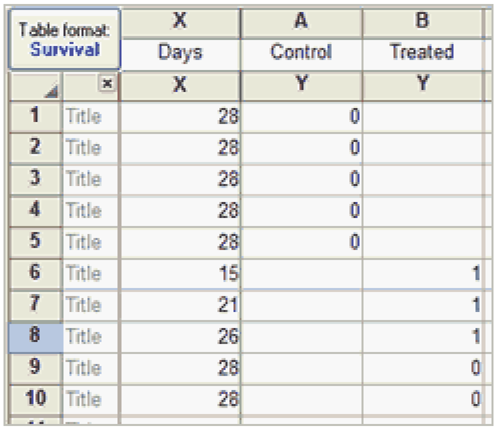


Table format: Survival		X	A	B
		Days	Control	Treated
		X	Y	Y
1	Title	28	0	
2	Title	28	0	
3	Title	28	0	
4	Title	28	0	
5	Title	28	0	
6	Title	15		1
7	Title	21		1
8	Title	26		1
9	Title	28		0
10	Title	28		0

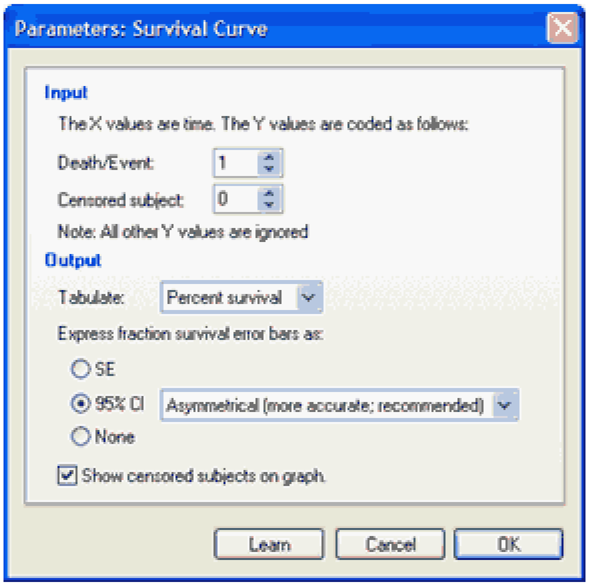
Note that the five control animals are each entered on a separate row, with the time entered as 28 (the number of days you observed the animals) and with Y entered as 0 to denote a censored observation. The observations on these animals is said to be censored because we only know that they lived for at least 28 days. We don't know how much longer they will live because the study ended.

The five treated animals also are entered one per row, with Y=1 when they died and Y=0 for the two animals still alive at the end of the study.

Analysis choices for survival 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, but there is little reason to ever do this as the default input choices are totally standard and the output choices can also be made on the graph.


Parameters: Survival Curve

Input

The X values are time. The Y values are coded as follows:

Death/Event: 1

Censored subject: 0

Note: All other Y values are ignored

Output

Tabulate: Percent survival

Express fraction survival error bars as:

○ SE

● 95% CI Asymmetrical (more accurate; recommended)

○ None

☑ Show censored subjects on graph

Input

The default choices are to use the code '1' for deaths and '0' for censored subjects, and these are almost universal. But you can choose different codes by entering values here. The codes must be integers.

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 5 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 prior versions of Prism. 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.

Prism always compares survival curves by performing both the **log-rank (Mantel-Cox)** test and the **Gehan-Breslow-Wilcoxon** test. P values from both tests are reported. You can't choose which test is computed (Prism always does both), but you should choose which test you want to report.

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.

Prism also reports the uncertainty of the fractional survival as a standard error or 95% confidence intervals. Standard errors are calculated by the method of Greenwood. The 95% confidence 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.

The calculations take into account censored observations. Subjects whose data are censored 217-- 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).

Number of subjects at risk at various times

Prism tabulates the number of patients still at risk at each time. The number of subjects still at risk decreases each time a subject dies or is censored.

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”.

Interpreting results: Comparing two survival curves

Two methods to compute P value

Prism compares two survival curves by two methods: the log-rank test (also called the Mantel-Cox test) and the Gehan-Breslow-Wilcoxon test. It doesn't ask for your preference, but always reports both.

- The **log-rank (Mantel-Cox) test** 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. Prism actually computes the Mantel-Haenszel method, which is nearly identical to the log-rank method (they differ only in how they deal with two subjects with the same time of death).
- The **Gehan-Breslow-Wilcoxon** method gives more weight to deaths at early time points. This often 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 log-rank test gives equal weight to all time points. The Gehan-Wilcoxon test does not require a consistent hazard ratio, but does require that one group consistently have a higher risk than the other.

You need to choose which P value to report. Ideally, this choice should be made before you collect and analyze your data.

If in doubt, report the log-rank test (which is more standard) and report the Gehan-Wilcoxon results only if you have a strong reason.

Note that Prism cannot perform Cox proportional hazards regression.

If 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. In this case, neither the log-rank nor the Wilcoxon-Gehan test rests will be very helpful.

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?

Prism always calculates two-tail P values. If you wish to report a one-tail P value 34, 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.

Ratio of median survivals

The median survival is the time at which fractional survival equals 50%.

If survival exceeds 50% at the longest time point, then median survival cannot be computed and Prism leaves it blank. Even so, the P values and hazard ratio are still valid.

If the survival curve is horizontal at 50% survival, the median survival is ambiguous, and different programs report median survival differently. Prism reports the average of the first and last times at which survival is 50%.

When comparing two survival curves, Prism also reports the ratio of the median survival times along with its 95% confidence interval. You can be 95% sure that the true ratio of median survival times lies within that range.

This calculation is based on an assumption that is not part of the rest of the survival comparison. The calculation of the 95% CI of ratio of median survivals assumes that the survival curve follows 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 correct.

Hazard ratio

If you compare two survival curves, Prism reports the hazard ratio and its 95% confidence interval.

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 computation of the hazard ratio assumes that the ratio is consistent over time, and that any differences are due to random sampling. So Prism reports a single hazard ratio, not a different hazard ratio for each time interval. Prism 5 computes the hazard ratio and its confidence interval using the Mantel-Haenszel method. Prism 4 computed the hazard ratio itself (but not the confidence interval) by the log-rank method. The two are usually quite similar.

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.

Interpreting results: Comparing >2 survival curves

Logrank test

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?

Logrank test for trend

If you entered three or more data sets, Prism automatically calculates the **logrank test for trend**. This test is only meaningful if the data sets were entered in a logical order, perhaps corresponding to dose or age. If the data sets are not ordered (or not equally spaced), then you should ignore the results of the logrank test for trend.

The logrank test for trend calculates 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.

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. 232

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 should correct for multiple comparisons 52.

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.

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.

✓ 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.

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.

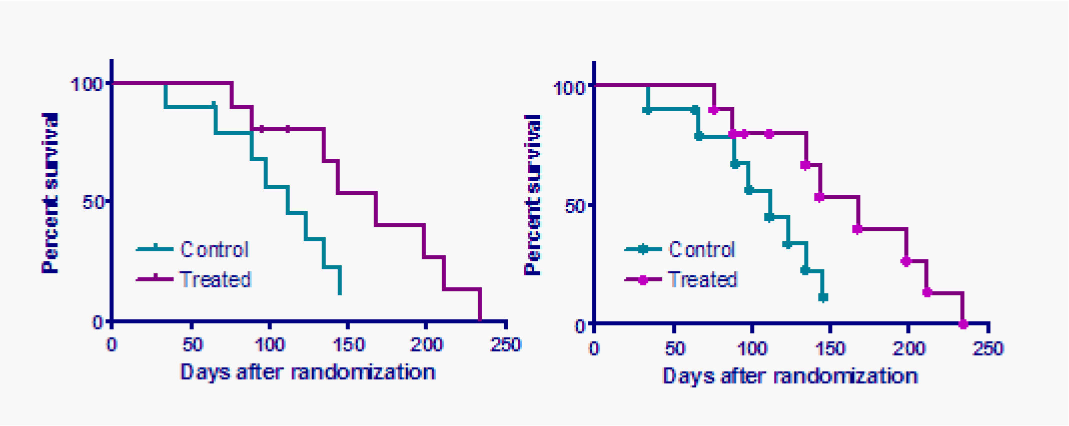
How to compute the data

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

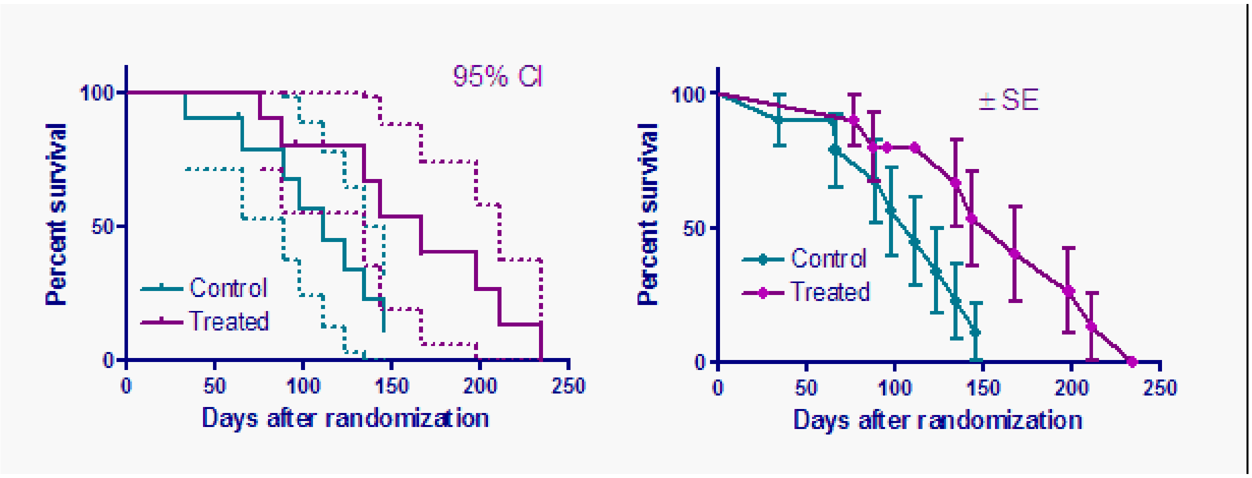


As shown above, survival curves are usually plotted as staircases. Each death is shown as a drop in survival.

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 left 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.

VIII. 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.

ROC Curves

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.

On the other hand, 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.

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.

How to: ROC curve

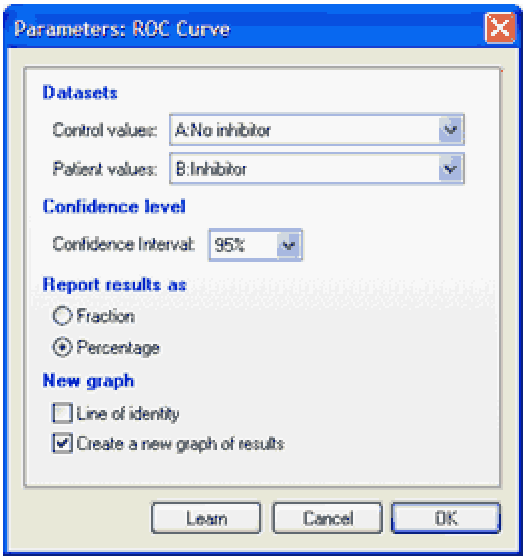
1. Enter ROC data

From the Welcome or New table dialog, choose the one-way tab, and then choose a scatter plot. 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.

2. Create the ROC curve

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



Parameters: ROC Curve

Datasets

Control values: A:No inhibitor

Patient values: B:Inhibitor

Confidence level

Confidence Interval: 95%

Report results as

○ Fraction

● Percentage

New graph

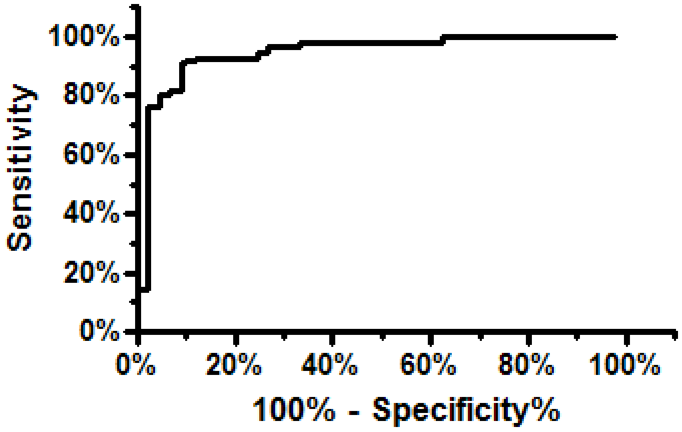
□ Line of identity

☑ Create a new graph of results

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



Interpreting results: ROC curves

Sensitivity and specificity

The whole point of an ROC curve is to help you decide where to draw the line between 'normal' and 'not normal'. This will be an easy decision if all the control values are higher (or lower) than all the patient values. Usually, however, the two distributions overlap, making it not so easy. If you make the threshold high, you won't mistakenly diagnose the disease in many who don't have it, but you will miss some of the people who have the disease. If you make the threshold low, you'll correctly identify all (or almost all) of the people with the disease, but will also diagnose the disease in more people who don't have it.

To help you make this decision, Prism tabulates and plots the sensitivity and specificity of the test at various cut-off values.

Sensitivity: The fraction of people with the disease that the test correctly identifies as positive.

Specificity: The fraction of people without the disease that the test correctly identifies as negative.

Prism calculates the sensitivity and specificity using each value in the data table as the cutoff value. This means that it calculates many pairs of sensitivity and specificity. If you select a high threshold, you increase the specificity of the test, but lose sensitivity. If you make the threshold low, you increase the test's sensitivity but lose specificity.

Prism displays these results in two forms. The table labeled "ROC" curve is used to create the graph of 100%-Specificity% vs. Sensitivity%. The table labeled "Sensitivity and Specifity" tabulates those values along with their 95% confidence interval for each possible cutoff between normal and abnormal.

Area

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.

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. This method is described by Hanley, J.A., and McNeil, B. J. (1982). Radiology 143:29-36.

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.

Prism calculates z= (AUC - 0.5)/SEarea and then determines P from the z ratio (normal distribution). In the numerator, we subtract 0.5, because that is the area predicted by the null hypothesis. The denominator is the SE of the area, which Prism reports.

Comparing ROC curves

Prism does not compare ROC curves. It is, however, quite easy to manually compare two ROC curves created with data from two different (unpaired) sets of patients and controls.

1. Calculate the two ROC curves using separate analyses of your two data sets.
2. For each data set, calculate separate values for the area under the curve and standard error (SE) of the area.
3. Combine these results using this equation:

$$z = \frac{\left|Area_1 - Area_2\right|}{\sqrt{SE^2_{Area1} + SE^2_{Area2}}}$$

4. If you investigated many pairs of methods with indistinguishable ROC curves, you would expect the distribution of z to be centered at zero with a standard deviation of 1.0. To calculate a two-tail P value, therefore, use the following Microsoft Excel function:

   =2*(1-NORMSDIST(z))

The method described above is appropriate when you compare two ROC curves with data collected from different subjects. A different method is needed to compare ROC curves when both laboratory tests were evaluated in the same group of patients and controls.

Prism does not compare paired-ROC curves. To account for the correlation between areas under your two curves, use the method described by 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.

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.

Comparing Methods with a Bland-Altman Plot

How to: Bland-Altman plot

A Bland-Altman plots compare 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.

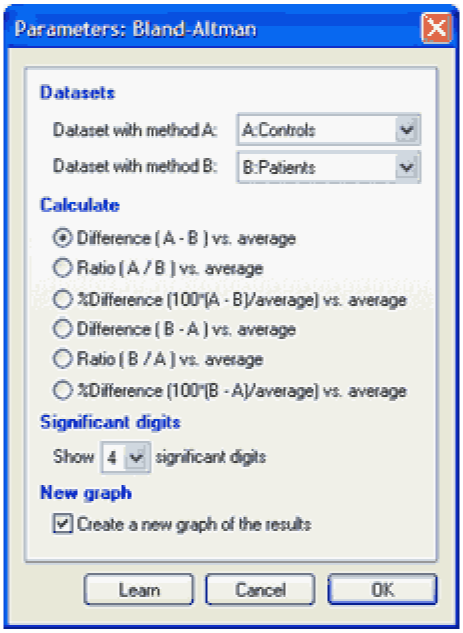
1. Enter the data

Create a new table. Choose the one-way tab and a before-after graph. If you don't have data yet, choose the Bland-Altman sample data.

Enter the measurements from the first method into column A and for the other method into column B. Each row represents one sample or one subject.

2. Choose the Bland-Altman analysis

From the data table, click Analyze on the toolbar, and then choose Bland-Altman from the list of one-way analyses.


Parameters: Bland-Altman

Datasets

Dataset with method A: A:Controls

Dataset with method B: B:Patients

Calculate

● Difference (A - B) vs. average

○ Ratio (A / B) vs. average

○ %Difference (100*(A - B)/average) vs. average

○ Difference (B - A) vs. average

○ Ratio (B / A) vs. average

○ %Difference (100*(B - A)/average) vs. average

Significant digits

Show 4 significant digits

New graph

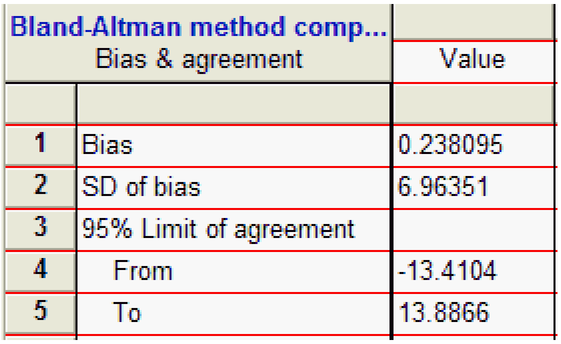
☑ Create a new graph of the results

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

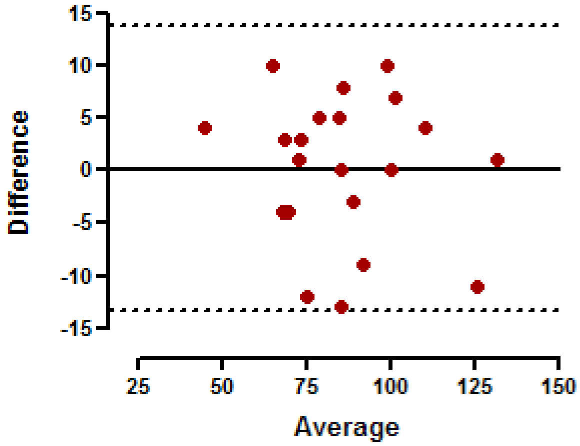
The Bland-Altman analysis creates two pages of results. The first page shows the difference and average values, and is used to create the plot. The second results page shows the bias 246, or the average of the differences, and the 95% limits of agreement 246.

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

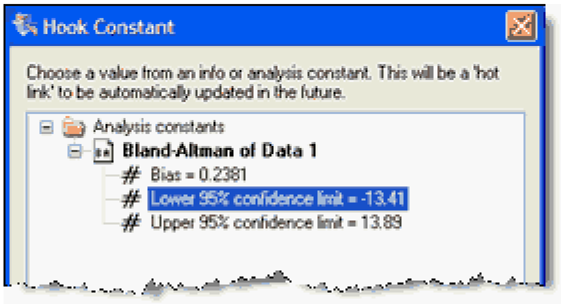


	Bland-Altman method comp... Bias & agreement	Value
1	Bias	0.238095
2	SD of bias	6.96351
3	95% Limit of agreement	
4	From	-13.4104
5	To	13.8866

4. Plot the Bland-Altman graph



The 95% confidence limits of the bias are shown as two dotted lines. They are created using Additional ticks and grid lines on the Format axes dialog. The position of each line was 'hooked' to an analysis constant created by the Bland-Altman analysis.



Choose a value from an info or analysis constant. This will be a 'hot link' to be automatically updated in the future.

- Analysis constants
  - Bland-Altman of Data 1
    - Bias = 0.2381
    - Lower 95% confidence limit = -13.41
    - Upper 95% confidence limit = 13.89

The origin of the graph was moved to the lower left (and offset) on the first tab of the Format Axes dialog.

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% limit 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 producing different results.

This page also shows the standard deviation (SD) of the differences between the two assay methods. The SD value 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% confidence limits for the difference, and these limits would go further in each direction than do the limits of agreement.

Interpreting the Bland-Altman results

Bland-Altman plots are generally interpreted informally, without further analyses. Ask yourself these three 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.
- 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?

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.

Index