GraphPad PRISM®

Version 4.0

Statistics Guide

Statistical analyses for laboratory and clinical researchers

Harvey Motulsky

Third printing February 2005

Citation: H.J. Motulsky, *Prism 4 Statistics Guide –Statistical analyses for laboratory and clinical researchers*. GraphPad Software Inc., San Diego CA, 2003, www.graphpad.com.

This book is a companion to the computer program GraphPad Prism version 4, which combines scientific graphics, basic biostatistics, and nonlinear regression (for both Windows and Macintosh). Prism comes with three volumes in addition to this one: *The Prism 4 User's Guide, Prism 4 Step-by-Step Examples,* and *Fitting Models to Biological Data using Linear and Nonlinear Regression*. You can learn more about Prism, and download all four volumes as Acrobat (pdf) files, from www.graphpad.com

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

Table of Contents

Part A:
Statistical Principles

1. Introduction to Statistical Thinking

Garbage in, garbage out

Computers are wonderful tools for analyzing data, but like any tool, data analysis programs can be misused. If you enter incorrect data or pick an inappropriate analysis, the results won't be helpful.

> Tip: Heed the first rule of computers: Garbage in, garbage out.

While this volume provides far more background information than most program manuals, it cannot entirely replace the need for statistics books and consultants.

GraphPad provides free technical support when you encounter problems with the program. Email us at support@graphpad.com. However, we can only provide limited free help with choosing statistics tests or interpreting the results (consulting and on-site teaching can sometimes be arranged for a fee).

When do you need statistical calculations?

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.
- ✓ Conversely, 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.

When you work with experimental systems with no biological variability and little experimental error, heed these aphorisms:

- ✓ If you need statistics to analyze your experiment, then you've done the wrong experiment.
- ✓ If your data speak for themselves, don't interrupt!

In many fields, however, scientists can't avoid large amounts of variability, yet care about relatively small differences. Statistical methods are necessary to draw valid conclusions from such data.

The key concept: Sampling from a population

What are samples and populations?

The basic idea of statistics is simple: you want to make inferences from the data you have collected to make general conclusions about the larger population from which the data sample was derived.

To do this, statisticians have developed methods based on a simple model: Assume that an infinitely large population of values exists and that 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 need for independent samples

It is not enough that your data are sampled from a population. Statistical tests are also 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.

How you can use statistics to extrapolate from sample to population

Statisticians have devised three basic approaches to make conclusions about populations from samples of data:

The first method is to assume that parameter values for populations follow a special distribution, known as the *Gaussian* (bell shaped) distribution. Once you assume that a population is distributed in that manner, statistical tests let you make inferences about the mean (and other properties) of the population. Most commonly used statistical tests assume that the population is Gaussian. These tests are sometimes called *parametric* tests.

The second method is to rank all values from low to high and then compare the distributions of ranks. This is the principle behind most commonly used *nonparametric* tests, which are used to analyze data from non-Gaussian distributions.

The third method is known as resampling. With this method, you create a population of sorts by repeatedly sampling values from your sample. This is best understood by an example. Assume you have a single sample of five values, and want to know how close that sample mean is likely to be from the true population mean. Write each value on a card and place the cards in a hat. Create many pseudo samples by drawing a card from the hat, writing down that number, and then returning the card to the hat. Generate many samples of N=5 this way. Since you can draw the same value more than once, the samples won't all be the same (but some might be). When randomly selecting cards gets tedious, use a computer program instead. The distribution of the means of these computer-generated samples gives you information about how accurately you know the mean of the entire population. The idea of resampling can be difficult to grasp. To learn about this approach to statistics, read the instructional material available at www.resample.com. Prism does not perform any tests based on resampling. Resampling methods are closely linked to bootstrapping methods.

Confidence intervals

The best way to use data from a sample to make inferences about the population is to compute a confidence interval (CI).

Let's consider the simplest example. You measure something (say weight or concentration) in a small sample, 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 with 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 confidence interval is a range of values.

The interpretation of a 95% CI is quite straightforward. If you accept certain assumptions (discussed later in this book for each kind of analyses), there is a 95% chance that the 95% CI of the mean you calculated contains the true population mean. In other words, if you generate many 95% CIs from many samples, you'll 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%. Since you don't know the population mean (unless you work with simulated data), you won'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 population mean lies within the 95% CI.

The concept is general. You can calculate the 95% CI for almost any value you compute when you analyze data, including the difference between the group means, a proportion, the ratio of two proportions, the best-fit slope of linear regression, and a best-fit value of an EC_{50} determined by nonlinear regression.

The concept is simple. 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. You want to know how precisely you have determined that value. A confidence interval tells that by expressing your results as a range of values. You can be 95% sure that the 95% CI contains the true (population) value.

While 95% CI are traditional, you can calculate a CI for any desired degree of confidence. If you want more confidence that an interval contains the true parameter, then the intervals have to be wider. So a 99% CI is wider than a 95% confidence interval, and a 90% CI is narrower than a 95% confidence interval.

Limitations of statistics

The statistical model is simple: Extrapolate from the sample you collected to a more general situation, assuming that each value in your sample was randomly and independently selected from a large population. 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 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. Thus, statistical logic is only part of data interpretation.

2. 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.

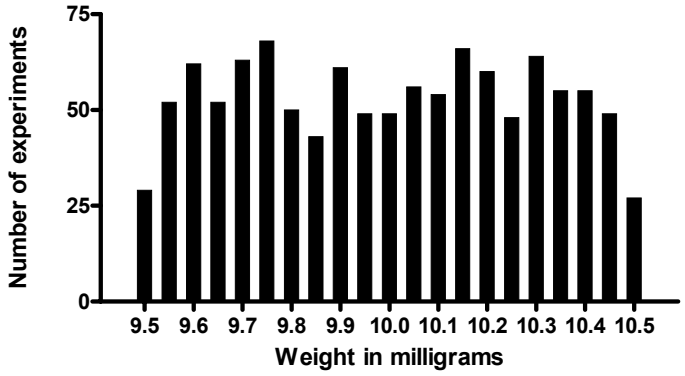
When reading in other books about the Gaussian distribution, two statistical terms might be confusing because they sound like ordinary words:

- ✓ 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.
- ✓ 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

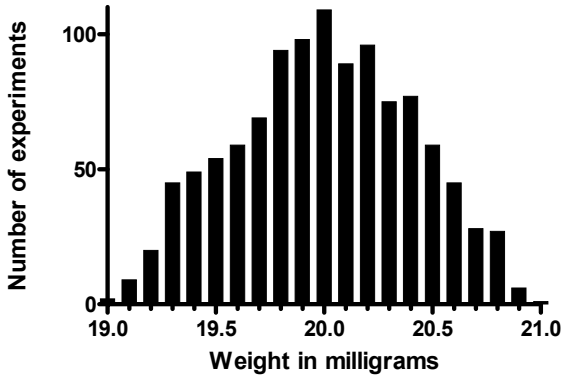
The Gaussian distribution emerges when many independent random factors act in an additive manner to create variability. This is best seen by an example.

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



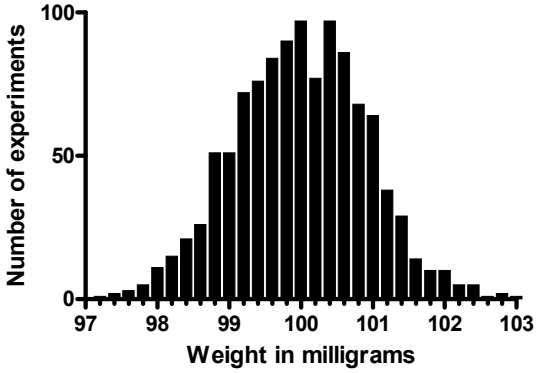
The average weight is 10 milligrams, the weight of 10 μL of water (at least on earth). The distribution is flat, with no hint of a Gaussian distribution.

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 .

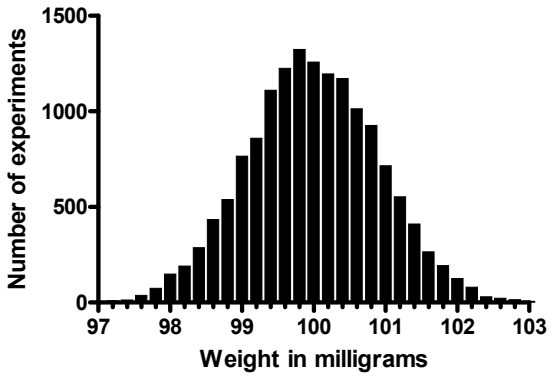


Each pipetting step has a flat random error. Add them up, and the distribution is not flat. For example, you'll get weights near 21 mg only if both pipetting steps err substantially in the same direction, and that is rare.

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.

3. P Values and Statistical Hypothesis Testing

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.

Statistical calculations can answer this question: 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? The answer to this question is called the *P value*.

The P value is a probability, with a value ranging from zero to one. If the P value is small enough, you'll conclude that the difference between sample means is unlikely to be due to chance. Instead, you'll conclude that the populations have different means.

What is a null hypothesis?

When statisticians discuss P values, they use the term *null hypothesis*. The null hypothesis simply states that there is no difference between the groups. Using that term, you can define the P value to be the probability of observing a difference as large as or larger than you observed if the null hypothesis were true.

Common misinterpretation of a P value

Many people misunderstand P values. If the P value is 0.03, that means that 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). It is tempting to conclude, therefore, that 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. However, this would be an incorrect conclusion. What you can say is that 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. This distinction may be clearer after you read *A Bayesian perspective on interpreting statistical significance* on page 21.

One-tail vs. two-tail P values

When comparing two groups, you must distinguish between one- and two-tail P values.

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 that compares the means of two groups. The same ideas can be applied to other statistical tests.

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?

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 ends up with the larger mean – even if it is quite a bit larger – you would have attributed that difference to chance.

It is usually best to use a two-tail P value for these reasons:

- ✓ 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.

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 α) 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 that the means come from different populations.

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.

Statistical significance in science

The term significant is *seductive* and easy to misinterpret. Using the conventional definition with alpha=0.05, a result is said to be statistically significant when that result would occur less than 5% of the time if the populations were, in fact, identical.

It is easy to read far too much into the word *significant* 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.

If a result is statistically significant, there are two possible explanations:

- ✓ The populations are identical, so there is, in fact, no difference. Owing simply to sampling variability, you obtained larger values in one sample group and smaller values in the other. In that case, it would be an error to infer a difference in populations based upon an observed difference in samples taken from those populations. Finding a statistically significant result when the populations are identical is called making a Type I error. If you define statistically significant to mean "P<0.05", then you'll make a Type I error in 5% of experiments where there really is no difference.
- ✓ The populations really are different, so your conclusion is correct. The difference may be large enough to be scientifically interesting, or it may be tiny and trivial.

"Extremely significant" results

Intuitively, you may think that P=0.0001 is more significant than P=0.04. Using strict definitions, this is not correct. Once you have set a threshold P value for statistical significance, every result is defined to be either statistically significant or is not statistically significant. Degrees of statistical significance are not distinguished. Some statisticians feel very strongly about this.

Many scientists are not so rigid and refer to results as being "almost significant", "very significant" or "extremely significant".

Prism summarizes the P value using the words in the middle column of this table. Many scientists label graphs with the symbols of the third column. 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

Report the actual P value

The concept of statistical hypothesis testing works well for quality control, where you must decide to accept or reject an item or batch based on the results of a single analysis. Experimental science is more complicated than that, because you often integrate many kinds of experiments before reaching conclusions. You don't need to make a significant/not significant decision for every P value. Instead, report exact P values, so you and your readers can interpret them as part of a bigger picture.

The tradition of reporting only "$P<0.05$" or "$P>0.05$" began in the days before computers were readily available. Statistical tables were used to determine whether the P value was less than or greater than 0.05, and it would have been very difficult to determine the P value exactly. Today, with most statistical tests, it is very easy to compute exact P values, and you shouldn't feel constrained to only report whether the P value is less than or greater than some threshold value.

4. Interpreting P Values and Statistical Significance

Statistical power

Type II errors and power

If you compare two treatments and your study concludes there is "no statistically significant difference", you should not necessarily conclude that the treatment was ineffective. It is possible that the study missed a real effect because you used a small sample or your data were quite variable. In this case you made a Type II error — obtaining a "not significant" result when, in fact, there is a difference.

When interpreting the results of an experiment that found no significant difference, you need to ask yourself how much power the study had to find various hypothetical differences (had they existed). The power depends on the sample size and amount of variation within the groups, where variation is quantified by the standard deviation (SD).

Here is a precise definition of power: Start with the assumption that two population means differ by a certain amount, but have the same SD. Now assume that you perform many experiments with the sample size you used, and calculate a P value for each experiment. Power is the fraction of these experiments that would lead to statistically significant results, i.e., would have a P value less than alpha (the largest P value you deem "significant", usually set to 0.05).

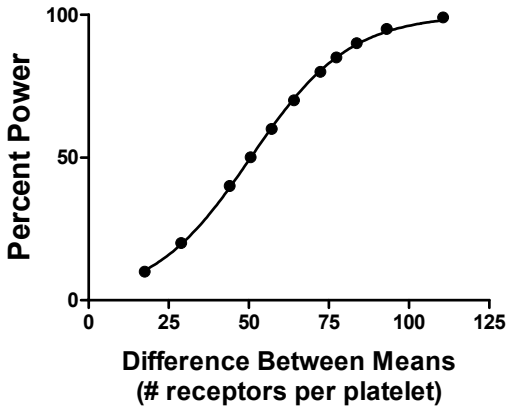
Example of power calculations

Motulsky et al. asked whether people with hypertension (high blood pressure) had altered numbers of α_2-adrenergic receptors on their platelets (Clinical Science 64:265-272, 1983). There are many reasons to think that autonomic receptor numbers may be altered in hypertensives. We studied platelets because they are easily accessible from a blood sample. The results are shown here:

Variable	Hypertensives	Controls
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 hypertensives do not have an altered number of α_2 receptors.

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. Prism does not compute power, but the companion program GraphPad StatMate does. Here are the results shown as a graph.



If the true difference between means was 50.58, then this study had only 50% power to find a statistically significant difference. In other words, if hypertensives really averaged 51 more receptors per cell, you'd find a statistically significant difference in about half of studies of this size, but you would not find a statistically significant difference in the other half of the studies. This is about a 20% change (51/257), large enough that it could possibly have a physiological impact.

If the true difference between means was 84 receptors/cell, then this study had 90% power to find a statistically significant difference. If hypertensives really had such a large difference, you'd find a statistically significant difference in 90% of studies this size and you would not find a significant difference in the other 10% of studies.

All studies have low power to find small differences and high power to find large differences. However, it is up to you to define "low" and "high" in the context of the experiment and to decide whether the power was high enough for you to believe the negative results. If the power is too low, you shouldn't reach a firm conclusion until the study has been repeated with more subjects. Most investigators aim for 80% or 90% power to detect a difference.

Since this study had only a 50% power to detect a difference of 20% in receptor number (50 sites per platelet, a large enough difference to possibly explain some aspects of hypertension physiology), the negative conclusion is not solid.

A Bayesian perspective on interpreting statistical significance

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 to detect the difference you are looking for with a P value less than 0.05. What happens when you repeat the experiment many times?

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 *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 = 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.

Your interpretation of a "statistically significant" result depends on the context of the experiment. 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.

Beware of multiple comparisons

Interpreting an individual P value is easy. Assuming the null hypothesis is true, the P value is the chance that the effects of 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. 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

However, many scientific studies generate more than one P value. Some studies in fact generate hundreds of P values. Interpreting multiple P values can be difficult.

If you test several independent null hypotheses and leave the threshold at 0.05 for each comparison, there is greater than a 5% chance of obtaining at least one "statistically significant" result by chance. The second column in the table below shows you how much greater.

Number of independent null hypotheses	Probability of obtaining one or more P values less than 0.05 by chance	Threshold to keep overall risk of type I error equal to 0.05
1	5%	0.0500
2	10%	0.0253
3	14%	0.0170
4	19%	0.0127

Number of independent null hypotheses	Probability of obtaining one or more P values less than 0.05 by chance	Threshold to keep overall risk of type I error equal to 0.05
5	23%	0.0102
6	26%	0.0085
7	30%	0.0073
8	34%	0.0064
9	37%	0.0057
10	40%	0.0051
20	64%	0.0026
50	92%	0.0010
100	99%	0.0005
N	$100(1.00 - 0.95^N)$	$1.00 - 0.95^{(1/N)}$.

To maintain the chance of randomly obtaining at least one statistically significant result at 5%, you need to set a stricter (lower) threshold for each individual comparison. This is tabulated in the third column of the table. If you only conclude that a difference is statistically significant when a P value is less than this value, then you'll have only a 5% chance of finding any "significant" difference by chance among all the comparisons.

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 didn't 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.

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 of P values, and scanning these can generate intriguing research hypotheses. You can't test hypotheses using the same data that prompted you to consider them. You need to test hypotheses with fresh data.

The examples above compared two groups, with multiple outcomes. If your experiment includes three or more groups you shouldn't do t tests at all (even if you correct for multiple comparisons). Instead analyze the data using one-way analysis of variance (ANOVA) followed by post tests. These methods account both for multiple comparisons and the fact that the comparisons are not independent. Read more about post tests in Chapter 9.

5. Outliers

What is an outlier?

When analyzing data, you'll sometimes find that one value is far from the others. Such a value is called an outlier, a term that is usually not defined rigorously. 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 those three questions, you have to decide what to do with the outlier. There are two possibilities.

One possibility is that 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 other possibility is that 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 and is misleading.

The problem, of course, is that you are rarely sure which of these possibilities is correct.

No mathematical calculation can tell you for sure whether the outlier came from the same or 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'd find one value as far from the others as you observed? If this probability is small, then you will conclude that the outlier is likely to be an erroneous value, and 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.

Prism does not perform any sort of outlier detection. If you want to perform an outlier test by hand, you can calculate Grubb's test, described below.

Detecting outliers with Grubbs' test

The 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 Grubb's test with the most extreme outlier, the value that leads to the largest value of Z.

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

Since 5% of the values in a Gaussian population are more than 1.96 standard deviations from the mean, your first thought might be to conclude that the outlier comes from a different population if Z is greater than 1.96. There are two problems with this approach:

When analyzing experimental data, you don't know the SD of the population. Instead, you calculate the SD from the data. The presence of an outlier increases the calculated SD. Since the presence of an outlier increases both the numerator (difference between the value and the mean) and denominator (SD of all values), Z does not get very large. In fact, no matter how the data are distributed, Z cannot get larger than $(N-1)/\sqrt{N}$ where N is the number of values. For example, if N=3, Z cannot be larger than 1.555 for any set of values. For small samples, therefore, the threshold must be less than 1.96.

If all your data comes from a Gaussian distribution (no true outliers), Grubb's test is designed so that there is a 5% chance of incorrectly identifying the value with the largest Z score as an outlier. If you used 1.96 as the threshold value, you'd identify 5% of the values as outliers, rather than identifying 5% of the samples as having an outlier. With large samples the threshold must be larger than 1.96.

Grubbs and others have tabulated critical values for Z, which are tabulated below for P=0.05 (two-tail). The critical value increases with sample size, as expected.

If your calculated value of Z is greater than the critical value in the table, then the P value is less than 0.05. This means that there is less than a 5% chance that you'd encounter an outlier so far from the others (in either direction) by chance alone, if all the data were really sampled from a single Gaussian distribution. Note that the method only works for testing the most extreme value in the sample (if in doubt, calculate Z for all values, but only calculate a P value for Grubbs' test from the largest value of Z).

N	Critical Z 5%	N	Critical Z 5%
3	1.15	27	2.86
4	1.48	28	2.88
5	1.71	29	2.89
6	1.89	30	2.91
7	2.02	31	2.92
8	2.13	32	2.94
9	2.21	33	2.95
10	2.29	34	2.97
11	2.34	35	2.98
12	2.41	36	2.99
13	2.46	37	3.00
14	2.51	38	3.01
15	2.55	39	3.03
16	2.59	40	3.04
17	2.62	50	3.13
18	2.65	60	3.20
19	2.68	70	3.26
20	2.71	80	3.31
21	2.73	90	3.35
22	2.76	100	3.38
23	2.78	110	3.42
24	2.80	120	3.44
25	2.82	130	3.47
26	2.84	140	3.49

Consult the references below for larger tables. You can also calculate an approximate P value as follows.

1. Calculate a t ratio from N (number of values in the sample) and Z (calculated for the suspected outlier as shown above).

$$t = \sqrt{\frac{N(N-2)Z^2}{(N-1)^2 - NZ^2}}$$

2. Determine the P value corresponding with that value of t with N-2 degrees of freedom. Use the Excel formula =TDIST(t,df,2), substituting values for t and df (the third parameter is 2, because you want a two-tailed P value).

3. Multiply the P value you obtain in step 2 by N. The result is an approximate P value for the outlier test. This P value is the chance of observing one point so far from the others if the data were all sampled from a Gaussian distribution. If Z is large, this P value will be very accurate. With smaller values of Z, the calculated P value may be too large.

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 use nonparametric tests, outliers will affect the results very little so do not need to be removed.

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. If you do this, you cannot use the table shown above. Rosner has extended the method to detecting several outliers in one sample. See the first reference below for details.

Here are two 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*.

Nonparametric tests are 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).

Excluding outliers in Prism

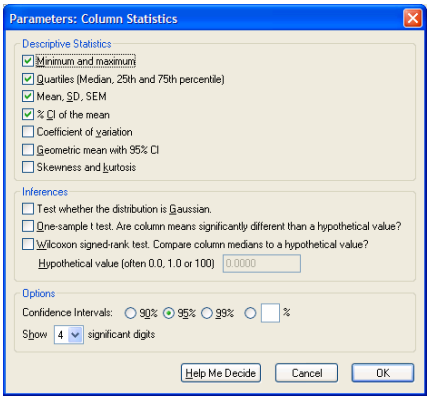
While Prism does not identify outliers automatically, it does let you manually exclude values you consider to be outliers. From a data table, select the outlier(s), drop the Edit menu, and choose **Exclude**. Prism will display excluded values in blue italics and append an asterisk. Such values will be ignored by all analyses and graphs. We suggest that you document your reasons for excluding values in the Prism Info section.

Part B:
Continuous Data

6. Descriptive Statistics

Choosing column statistics

From your data table, click **Analyze** and choose built-in analysis. Then choose **Column statistics** from the list of statistical analyses to bring up the **Parameters** dialog.



Choose the descriptive statistics you want to determine. Note: CI means confidence interval.

If you formatted the Y columns for entry of replicate values (for example, triplicates), Prism first averages the replicates in each row. It then calculates the column statistics on the means, without considering the SD or SEM of each row or the number of replicates you entered. If you enter 10 rows of triplicate data, the column statistics are calculated from the 10 row means, not from the 30 individual values. If you format the Y columns 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.

Note that this dialog also allows you to perform normality tests (details later in this chapter) and compare the mean or median with a hypothetical value using a one-sample t test or a Wilcoxon signed-rank test (details in the next chapter).

Interpreting descriptive statistics

Standard deviation (SD)

The ***standard deviation*** (SD) quantifies variability or scatter. If the data follow a bell-shaped Gaussian distribution, then 68% of the values lie within one SD of the mean (on either side) and 95% of the values lie within two SD of the mean. The SD is expressed in the same units as your data.

Prism calculates the SD using the equation below. (Each y_i is a value, y_{mean} is the average, and N is sample size).

$$SD = \sqrt{\frac{\sum (y_i - y_{mean})^2}{N-1}}$$

The standard deviation computed this way (with a denominator of N-1) is called the *sample SD*, in contrast to the *population SD* which would have a denominator of N. Why is the denominator N-1 rather than N? In the numerator, 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 population mean. This means that the numerator will be too small. So the denominator is reduced as well. It is reduced to N-1 because that is the number of *degrees of freedom* in your data.

Defining degrees of freedom rigorously is beyond the scope of this book. When computing the SD of a list of values, you can calculate the last value from N-1 of the values, so statisticians say there are N-1 degrees of freedom.

Standard error of the mean (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. The SEM is calculated by this equation:

$$SEM = \frac{SD}{\sqrt{N}}$$

With large samples, the SEM is always small. By itself, the SEM is difficult to interpret. It is easier to interpret the 95% confidence interval, which is calculated from the SEM.

The difference between the SD and SEM

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

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. 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.

The SD does not change predictably as you acquire more data. The SD quantifies the scatter of the data, and increasing the size of the sample does not change the scatter. The SD might go up, or it might go down; you can't predict. On average, the SD will stay the same as sample size gets larger.

If the scatter is caused by biological variability, you probably will want to show the variation. In this case, report the SD rather than the SEM. If you are using an *in vitro* system with no biological variability, the scatter can only result from experimental imprecision. In this case, you may not want to show the scatter, but instead show how well you have assessed the mean. Report the mean and SEM, or the mean with 95% confidence interval.

You should choose to show the SD or SEM based on the source of the variability and the point of the experiment. In fact, many scientists choose the SEM simply because it is smaller so creates shorter error bars.

95% confidence interval (CI)

Like the SEM, the ***confidence interval*** also quantifies the precision of the mean. The mean you calculate from your sample of data points depends on which values you happened to sample. Therefore, the mean you calculate is unlikely to equal the overall population mean exactly. The size of the likely discrepancy depends on the variability of the values (expressed as the SD) and the sample size. Combine those together to calculate a 95% confidence interval (95% CI), which is a range of values. You can be 95% sure that this interval contains the true population mean. More precisely, if you generate many 95% CIs from many data sets, you expect the CI to include the true population mean in 95% of the cases and not to include the true mean value in the other 5% of the cases. Since you don't know the population mean, you'll never know when this happens.

The confidence interval extends in each direction by a distance calculated from the standard error of the mean multiplied by a critical value from the t distribution. This value depends on the degree of confidence you want (traditionally 95%, but it is possible to calculate intervals for any degree of confidence) and on the number of degrees of freedom in this experiment (N-1). With large samples, this multiplier equals 1.96. With smaller samples, the multiplier is larger.

Coefficient of variation (CV)

The coefficient of variation (CV), also known as "relative variability", equals the standard deviation divided by the mean (expressed as a percent). Because it is a unitless ratio, you can compare the CV of variables expressed in different units. 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. In contrast, a temperature of "0.0" does not mean zero temperature (unless measured in degrees Kelvin). Don't calculate CV for variables, such as temperature, where the definition of zero is arbitrary.

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.

Quartiles and range

Quartiles divide the data into four groups, each containing an equal number of values. Quartiles are divided by the 25th percentile, 50th percentile, 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. The median is the 50th percentile.

Prism computes percentile values by first computing P*(N+1)/100, where P is 25, 50, or 75 and N is the number of values in the data set. The result is the rank that corresponds to the percentile value. If there are 68 values, the 25th percentile corresponds to a rank equal to 25*(68+1)/100 =17.25, so the 25th percentile lies between the value of the 17th and 18th value (when ranked from low to high). Prism computes the 25th percentile as the average of those two values.

While there is no ambiguity about how to calculate the median, there are several ways to compute the 25th and 75th percentiles. Prism uses what is probably the most commonly used method. With large data sets, all the methods give similar results. With small data sets, the results can vary quite a bit.

Tip: Because the various methods to compute 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.

Geometric mean

The ***geometric mean*** is the antilog of the mean of the logarithms of the values. This is the same as taking the Nth root (where N is the number of points) of the product of all N values. It is less affected by outliers than the mean. Prism also reports the 95% confidence interval of the geometric mean.

Skewness and kurtosis

Skewness quantifies the asymmetry of a distribution. A symmetrical distribution has a skewness of zero. An asymmetrical distribution with a long tail to the right (higher values) has a positive skew. An asymmetrical distribution with a long tail to the left (lower values) has a negative skew.

Kurtosis quantifies how closely the shape of a distribution follows the usual Gaussian shape. A Gaussian distribution, by definition, has a kurtosis of 0. A distribution with more values in the center, and less in the tails, has a negative kurtosis. A distribution with fewer values in the center and more in the tail has a positive kurtosis.

Skewness and kurtosis are computed by these equations.

$$skewness = \frac{\sum (Y_i - \mu)^3}{N\sigma^3}$$

$$kurtosis = \frac{\sum (Y_i - \mu)^4}{N\sigma^4} - 3$$

The results of normality tests

How the normality test works

Prism tests for deviations from Gaussian distribution using the Kolmogorov-Smirnov (KS) test. Since the Gaussian distribution is also called the "normal" distribution, the test is called a normality test. The KS statistic (which some other programs call D) quantifies the discrepancy between the distribution of your data and an ideal Gaussian distribution – larger values denoting larger discrepancies. It is not informative by itself, but is used to compute a P value.

Prism evaluates normality using the KS method, but the method as originally published cannot be used to calculate the P value because their method 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.

How to think about results from a normality test

The P value from the 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? More precisely, the P

value answers this question: If the population is really Gaussian, what is the chance that a randomly selected sample of this size would have a KS value as large as, or larger than, the observed value?

By looking at the distribution of a small sample of data, it is hard to tell whether or not the values came from a Gaussian distribution. Running a formal test does not make it easier. The tests simply have little power to discriminate between Gaussian and non-Gaussian populations with small sample sizes. How small? If you have fewer than five values, Prism doesn't even attempt to test for normality. In fact, the test doesn't really have much power to detect deviations from Gaussian distribution unless you have several dozen values.

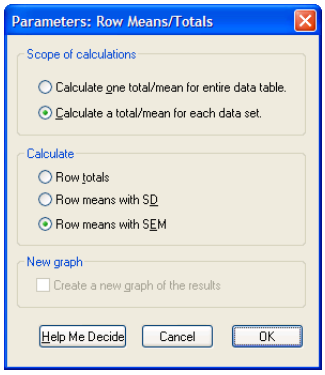
Your interpretation of the results of a normality test depends on the P value calculated by the test and on the sample size.

P value	Sample size	Conclusion
Small (<0.05)	Any	The data failed the normality test. You can conclude that the population is unlikely to be Gaussian.
Large	Large	The data passed the normality test. You can conclude that the population is likely to be Gaussian, or nearly so. How large does the sample have to be? There is no firm answer, but one rule-of-thumb is that the normality tests are only useful when your sample size is a few dozen or more.
Large	Small	You will be tempted to conclude that the population is Gaussian. But that conclusion may be incorrect. A large P value just means that the data are not inconsistent with a Gaussian population. That doesn't exclude the possibility of a non-Gaussian population. Small sample sizes simply don't provide enough data to discriminate between Gaussian and non-Gaussian distributions. You can't conclude much about the distribution of a population if your sample contains fewer than a dozen values.

Row means and totals

If you enter data with replicate Y values, Prism will automatically graph mean and SD (or SEM). You don't have to choose any analyses. Use settings on the Symbols dialog (double-click on any symbol to see it) to plot individual points or to choose SD, SEM, 95%CI or range error bars.

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** from the list of statistical analyses to bring up this dialog.



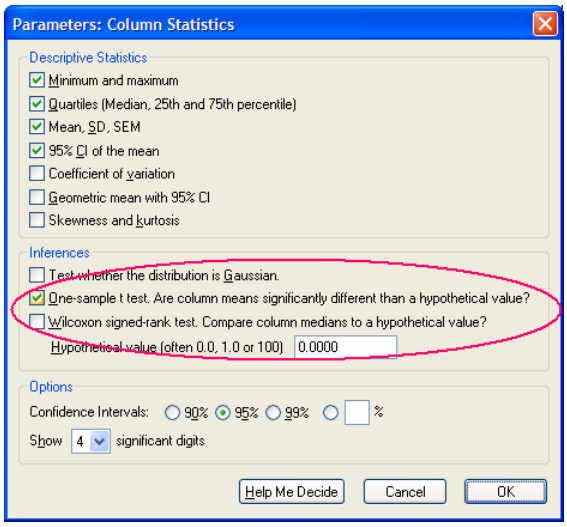
First, decide on the *scope of the calculations*. If you have entered more than one data set in the table, you have two choices. Most often, you'll calculate a row total/mean for each data set. The results table will have the same number of data sets as the input table. The other choice is to calculate one row total/mean for the entire table. The results table will then have a single data set containing the grand totals or means.

Then decide *what to calculate*: row totals, row means with SD, or row means with SEM. To review the difference between SD and SEM see *Interpreting descriptive statistics* on page 29.

7. Comparing a Mean or Median to a Hypothetical Value

Choosing a one-sample t test or Wilcoxon rank sum test

Prism can compare the mean of each column with a hypothetical value you enter using a one-sample t test. It can also compare the median of each column with a hypothetical value you enter using the Wilcoxon rank sum test. Choose either test from the Column Statistics analysis.



The results of a one-sample t test

How a one-sample t test works

A one-sample t test compares the mean of a single column of numbers against a hypothetical mean that you provide. Prism calculates the t ratio from this equation:

$$t = \frac{\text{Sample Mean} - \text{Hypothetical Mean}}{\text{Standard Error of Mean}}$$

A P value is computed from the t ratio and the numbers of degrees of freedom (which equals sample size minus 1).

How to think about results from the one-sample t test

Look first at the P value, which 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?

"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 seeking. How large a difference (between the population mean and the hypothetical mean) would you consider to be scientifically important? How small a difference would

you consider to be scientifically trivial? You need to use scientific judgment and common sense. Statistical calculations cannot answer these questions, because the answers depend on the context of the experiment.

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

If the P value is small (one-sample t test)

If the P value is small (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.

The true difference between population mean and hypothetical mean is probably not the same as the difference observed in this experiment. There is no way to know the true difference between the population mean and the hypothetical mean. Prism presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true difference between the overall (population) mean and the hypothetical value you entered.

To interpret the results in a scientific context, look at both ends of the confidence interval (the confidence limits) and ask whether they represent a discrepancy that is scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial	Trivial	Although the true difference is not zero (since the P value is low), the difference is tiny and uninteresting. The data have a mean distinct from the hypothetical value, but the discrepancy is too small to be scientifically interesting.
Trivial	Important	Since the confidence interval ranges from a difference that you think is biologically trivial to one you think would be important, you can't reach a strong conclusion from your data. You can conclude that the data has a mean distinct from the hypothetical value you entered, but don't know whether that difference is scientifically trivial or important. You'll need more data to obtain a clear conclusion.
Important	Important	Since even the low end of the confidence interval represents a difference large enough to be considered biologically important, you can conclude that the data have a mean distinct from the hypothetical value, and the discrepancy is large enough to be scientifically relevant.

If the P value is large (one-sample t test)

If the P value is large, 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.

How large could the true difference really be? Because of random variation, the difference between the hypothetical mean and the sample mean in this experiment is unlikely to be equal to the true difference between population mean and hypothetical mean. There is no way to know the true difference between the population mean and the hypothetical mean. Prism presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true difference between the true (population) mean of the data and the hypothetical mean you entered. When the P value is larger than 0.05, the 95% confidence interval will start with a negative number (the hypothetical mean is larger than the population mean) and go up to a positive number (the population mean is larger than the hypothetical mean).

To interpret the results in a scientific context, look at each end of the confidence interval and ask whether it represents a difference that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial	Trivial	You can reach a firm conclusion. Either the data has a mean equal to the hypothetical mean or they differ by a trivial amount.
Trivial	Important	You can't reach a strong conclusion. The data are consistent with a mean slightly smaller than the hypothetical mean, equal to the hypothetical mean, or larger than the hypothetical mean, perhaps large enough to be scientifically important. To reach a clear conclusion, you need to repeat the experiment with more subjects.
Important	Trivial	You can't reach a strong conclusion. The data are consistent with a mean smaller than the hypothetical mean (perhaps small enough to be scientifically important), equal to the hypothetical mean, or slightly larger than the hypothetical mean. You can't make a clear conclusion without repeating the experiment with more subjects.
Important	Important	You can't reach a strong conclusion. The data are consistent with a mean smaller than the hypothetical mean (perhaps small enough to be scientifically important), equal to the hypothetical mean, or larger than the hypothetical mean (perhaps large enough to be scientifically important). In other words, you can't draw any conclusion at all. You need to repeat the experiment with more subjects.

Checklist: Is a one-sample t test the right test for these data?

Before accepting the results of any statistical test, first think carefully about whether you chose an appropriate test. Before accepting results from a one-sample t test, ask yourself these questions:

Is the population distributed according to a Gaussian distribution?

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. Prism tests for violations of this assumption, but normality tests have limited utility (see page 32). 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 rank sum 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.

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. See *The need for independent samples* on page 10.

If you chose a one-tail P value, did you predict correctly?

If you chose a one-tail P value, you should have predicted whether the mean of your data would be larger than or smaller than the hypothetical mean. 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. See *One-tail vs. two-tail P values* on page 16.

The results of a Wilcoxon rank sum test

How the Wilcoxon rank sum test works

A Wilcoxon rank sum test compares the median of a single column of numbers against a hypothetical median you entered using these steps:

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'd expect W to be near zero. If W (the sum of signed ranks) is far from zero, the P value will be small. The P value answers this question: Assuming that you randomly sample N values from a population with the hypothetical median, what is the chance that W will be as far from zero (or further) than you observed?

Don't confuse the Wilcoxon rank sum test (compare one group with hypothetical median) with the Wilcoxon matched pairs test (compare medians of two paired groups). See *The results of a Wilcoxon matched pairs test* on page 54.

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

How to think about the results of a Wilcoxon rank sum test

The Wilcoxon signed rank test is a nonparametric test that compares the median of one column of numbers to a theoretical median.

Look first at the P value, which 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, 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, 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.

Checklist: Is the Wilcoxon test right for these data?

Before interpreting the results of any statistical test, first think carefully about whether you have chosen an appropriate test. Before accepting results from a Wilcoxon test, ask yourself these questions (Prism cannot help you answer them):

Are the "errors" independent?

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. Prism cannot test this assumption. See *The need for independent samples* on page 10.

Are the data clearly sampled from a non-Gaussian population?

By selecting a nonparametric test, you have avoided assuming that the data were sampled from a Gaussian distribution. 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 with logs or reciprocals) to create a Gaussian distribution and then using a one-sample t test.

Are the data distributed symmetrically?

The Wilcoxon test does not assume that the data are sampled from a Gaussian distribution. However it does assume that the data are distributed symmetrically around their median.

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 has the larger median before collecting data (see page 16). Prism does not ask you to record this prediction, but assumes it is correct. If your prediction was wrong, ignore the P value reported by Prism and state that P>0.50.

8. t Tests and Nonparametric Comparisons

Introduction to comparing two groups

Prism can compare two groups with a paired or unpaired t test, or with the nonparametric Mann-Whitney or Wilcoxon matched pairs test. These tests compare measurements (continuous variables) such as weight, enzyme activity, or receptor number. To compare two proportions, see *Contingency Tables* on page 99. To compare survival curves, see *Survival Curves* on page 107.

Entering data to compare two groups with a t test (or a nonparametric test)

From the Welcome (or New Table) dialog, choose any graph from the One grouping variable tab. Or choose to format the data table directly, and choose single columns of Y values and no X column (since X values are ignored by the t test analysis). The two groups do not have be the same size (it's OK to leave some cells empty).

	A	B
	Control	Treated
	Y	Y
1	34.0	45.0
2	43.0	47.0
3	39.0	52.0

If you have already averaged your data, format the data table for mean, SD (or SEM), and N. With this format, you can't pick nonparametric or paired tests, which require raw data. Enter data on only one row.

	A			B		
	Control			Treated		
	Y	SEM	N	Y	SEM	N
1	38.667000	2.6030	3	48.00000	2.0820	3

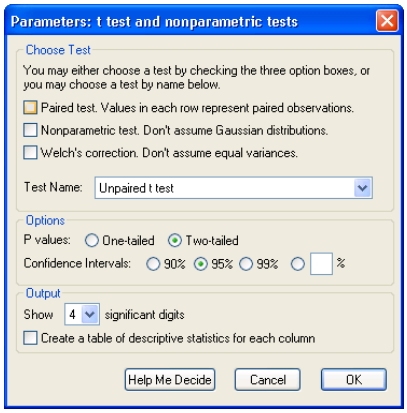
Indexed data

Many statistics programs expect you to enter data in an indexed format, as shown below. One column contains all the data, and the other column designates the group. Prism cannot analyze data entered in index format. If you have indexed data from another program, choose to "unstack" your data when you import it (an option on the Filter tab of the Import data dialog). This rearranges your data to a format Prism can analyze. Read the chapter on importing data in the Prism User's Guide. You can also insert indexed data into Prism using the Paste Special command.

Group	Value
1	34
1	43
1	39
2	45
2	47
2	52

Choosing an analysis to compare two groups

Go to the data or results table you wish to analyze (see *Entering data to compare two groups with a t test (or a nonparametric test)* on page 40). Press the **Analyze** button, and choose **Built-in analysis**. Then select **t tests** from the Statistical analyses section. If your table has more than two columns, select the two columns you wish to compare. (If you want to compare three or more columns, see *One-way ANOVA and Nonparametric Comparisons* on page 57. Press OK from the Analyze dialog to bring up the parameters dialog for t tests and related nonparametric tests:



Paired or unpaired test?

When choosing a test, you need to decide whether to use a paired test. Choose a paired test when the two columns of data are matched. Here are some examples:

- ✓ You measure a variable (e.g., weight) before an intervention, and then measure it in the same subjects after the 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.

More generally, you should select a paired test whenever you expect a value in one group to be closer to a *particular* value in the other group than to a *randomly selected* value in the other group.

Ideally, the decision about paired analyses should be made before the data are collected. Certainly the matching 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 zip code, but it is not OK to match based on blood pressure.

t test or nonparametric test?

The t test, like many statistical tests, assumes that you have sampled data from populations that follow a Gaussian 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. But there is a drawback. Nonparametric tests are less powerful than the parametric tests that assume Gaussian distributions. This means that P values tend to be higher, making it harder to detect real differences. With large samples, the difference in power is minor. With small samples, nonparametric tests have little power to detect differences.

You may find it difficult to decide when to select nonparametric tests. You should definitely choose a nonparametric test in these situations:

- ✓ The outcome variable is a rank or score with only a few categories. Clearly the population is far from Gaussian in these cases.
- ✓ One, or a few, values are 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. Using a nonparametric test with these data is easy. 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 consider the relative ranks of the values, it won't matter that you didn't know one (or a few) of the values exactly.
- ✓ You are sure that the population is far from Gaussian. Before choosing a nonparametric test, consider transforming the data (e.g. to logarithms or reciprocals). Sometimes a simple transformation will convert non-Gaussian data to a Gaussian distribution.

In many situations, perhaps most, you will find it difficult to decide whether to select nonparametric tests. Remember that the Gaussian assumption is about the distribution of the overall population of values, not just the sample you have obtained in this particular experiment. Look at the scatter of data from previous experiments that measured the same variable. Also consider the source of the scatter. When variability is due to the sum of numerous independent sources, with no one source dominating, you expect a Gaussian distribution.

Prism can perform a normality test to attempt to determine whether data were sampled from a Gaussian distribution. Normality testing is part of the Column statistics analysis (see page 32). However, normality testing is less useful than you might hope, and not useful at all if you have less than a few dozen (or so) values. See page 32.

Your decision to choose a parametric or nonparametric test matters the most when samples are small for reasons summarized here:

	Large samples (> 100 or so)	Small samples (<12 or so)
Parametric tests	Robust. P value will be nearly correct even if population is fairly far from Gaussian.	Not robust. If the population is not Gaussian, the P value may be misleading.
Nonparametric test	Powerful. If the population is Gaussian, the P value will be nearly identical to the P value you would have obtained from a parametric test. With large sample sizes, nonparametric tests are almost as powerful as parametric tests.	Not powerful. If the population is Gaussian, the P value will be higher than the P value obtained from a t test. With very small samples, it may be impossible for the P value to ever be less than 0.05, no matter how the values differ.
Normality test	Useful. Use a normality test to determine whether the data are sampled from a Gaussian population.	Not very useful. Little power to discriminate between Gaussian and non-Gaussian populations. Small samples simply don't contain enough information to let you make inferences about the shape of the distribution in the entire population.

Assume equal variances?

The unpaired t test assumes that the two populations have the same variances (same standard deviations). A modification of the t test (developed by Welch) can be used when you are unwilling to make that assumption. Check the box for Welch's correction if you want this test.

This choice is only available for the unpaired t test. With Welch's t test, the degrees of freedom are calculated from a complicated equation and the number is not obviously related to sample size.

Welch's t test is used rarely. Don't select it without good reason.

One- or two-tail P value?

If you are comparing two groups, you need to decide whether you want Prism to calculate a one-tail or two-tail P value. To understand the difference, see *One-tail vs. two-tail P values* on page 16.

You should only choose a one-tail P value when:

- ✓ You predicted which group would have the larger mean (if the means are in fact different) before collecting any data.
- ✓ You will attribute a difference in the wrong direction (the other group ends up with the larger mean) to chance, no matter how large the difference.

Since those conditions are rarely met, two-tail P values are usually more appropriate.

Confirm test selection

Based on your choices, Prism will show you the name of the test you selected.

Test	Paired	Nonparametric	Equal variances
Unpaired t test	No	No	Yes
Welch's t test	No	No	No
Paired t test	Yes	No	N/A
Mann-Whitney test	No	Yes	N/A
Wilcoxon test	Yes	Yes	N/A

The results of an unpaired t test

How the unpaired t test works

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.

For the standard t test, the number of degrees of freedom (df) equals the total sample size minus 2. Welch's t test (a rarely used test which doesn't assume equal variances) calculates df from a complicated equation. Prism calculates the P value from t and df.

A standard t test assumes the two groups have equal variances. To test this assumption, Prism calculates the variance of each group (the variance equals the standard deviation squared) and then calculates F, which equals the larger variance divided by the smaller variance. The degrees of freedom for the numerator and denominator equal the sample sizes minus 1. From F and the two df values, Prism computes a P value that answers this question: If the two populations really have the same variance, what is the chance that you'd randomly select samples and end up with F as large (or larger) as observed in your experiment? If the P value is small, conclude that the variances (and thus the standard deviations) are significantly different.

Don't base your conclusion just on this one F test. Also consider data from other experiments in the series. If you conclude that the two populations really do have different variances, you have three choices:

- ✓ 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.
- ✓ Transform the data to equalize the variances, and then rerun the t test. You may find that converting values to their reciprocals or logarithms will equalize the variances and also make the distributions more Gaussian.
- ✓ Rerun the t test without assuming equal variances using Welch's modified t test.

How to think about results from an unpaired t test

The unpaired t test compares the means of two groups, assuming that data are sampled from Gaussian populations. The most important results are the P value and the confidence interval.

The P value answers this question: If the populations really have 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?

"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.

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

Interpreting a small P value from an unpaired t test

If the P value is small, then it is unlikely that the difference you observed is due to random sampling. You can conclude instead that the populations have different means. The difference is statistically significant, but is it scientifically important? The confidence interval helps you decide.

Because of random variation, the difference between the group means in this experiment is unlikely to equal the true difference between population means. There is no way to know what that true difference is. Prism presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true difference between the two means.

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference between means that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial difference	Trivial difference	Although the true difference is not zero, the true difference between means is tiny and uninteresting. The treatment had an effect, but a small one.
Trivial difference	Important difference	Since the confidence interval ranges from a difference that you think would be biologically trivial to one you think would be important, you can't reach a strong conclusion. You can conclude that the means are different, but you don't know whether the size of that difference is scientifically trivial or important. You'll need more data to obtain a clear conclusion.
Important difference	Important difference	Since even the low end of the confidence interval represents a difference large enough to be considered biologically important, you can conclude that there is a difference between treatment means and that the difference is large enough to be scientifically relevant.

Interpreting a large P value from an unpaired t test

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.

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. Prism presents the uncertainty 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 between means that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial decrease	Trivial increase	You can reach a crisp conclusion. Either the means really are the same or they differ by a trivial amount. At most, the true difference between means is tiny and uninteresting.
Trivial decrease	Large increase	You can't reach a strong conclusion. The data are consistent with the treatment causing a trivial decrease, no change, or an increase that might be large enough to be important. To reach a clear conclusion, you need to repeat the experiment with more subjects.
Large decrease	Trivial increase	You can't reach a strong conclusion. The data are consistent with a trivial increase, no change, or a decrease that may be large enough to be important. You can't make a clear conclusion without repeating the experiment with more subjects.
Large decrease	Large increase	You can't conclude anything until you repeat the experiment with more subjects.

Checklist: Is an unpaired t test the right test for these data?

Before accepting the results of any statistical test, first think carefully about whether you chose an appropriate test. Before accepting results from an unpaired t test, ask yourself the questions below. Prism can help you answer the first two questions. You'll have to answer the others based on experimental design.

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. While this assumption is not too important with large samples, it is important with small sample sizes (especially with unequal sample sizes). Prism tests for violations of this assumption, but normality tests have limited utility. See page 32. If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values (perhaps to logs or

reciprocals) to make the distributions more Gaussian. Another choice is to use the Mann-Whitney nonparametric test instead of the t test. A final option is to use the t test anyway, knowing that the t test is fairly robust to violations of a Gaussian distribution with large samples.

Do the two populations have the same variances?

The unpaired t test assumes that the two populations have the same variances (and thus the same standard deviation).

Prism tests for equality of variance with an F test. The P value from this test answers this question: If the two populations really have the same variance, what is the chance that you'd 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. See *The need for independent samples on page 10*.

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 followed by post tests. It is not appropriate to perform several t tests, comparing two groups at a time. Making multiple comparisons increases the chance of finding a statistically significant difference by chance and makes it difficult to interpret P values and statements of statistical significance.

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. See page 33.

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, 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. See *One-tail vs. two-tail P values on page 16.*

The results of a paired t test

How a paired t test works

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, therefore, 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. (See *Correlation coefficient* on page 93.) 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.

How to think about results of a paired t test

The paired t test compares two paired groups so you can make inferences about the size of the average treatment effect (average difference between the paired measurements). The most important results are the P value and the confidence interval.

The P value answers this question: If the treatment really had no effect, what is the chance that random sampling would result in an average effect as far from zero (or more so) as observed in this experiment?

"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 treatment effect 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.

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

Interpreting a small P value from a paired t test

If the P value is small, then it is unlikely that the treatment effect you observed is due to chance. You can reject the idea that the treatment does nothing, and conclude instead that the treatment had an effect. The treatment effect is statistically significant. But is it scientifically significant? The confidence interval helps you decide.

Random scatter affects your data, so the true average treatment effect is probably not the same as the average of the differences observed in this experiment. There is no way to know what that true effect is. Prism presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true treatment effect (the true mean of the differences between paired values).

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference between means that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial difference	Trivial difference	Although the true effect is not zero (since the P value is low), it is tiny and uninteresting. The treatment had an effect, but a small one.
Trivial difference	Important difference	Since the confidence interval ranges from a difference that you think are biologically trivial to one you think would be important, you can't reach a strong conclusion from your data. You can conclude that the treatment had an effect, but you don't know whether it is scientifically trivial or important. You'll need more data to obtain a clear conclusion.
Important difference	Important difference	Since even the low end of the confidence interval represents a treatment effect large enough to be considered biologically important, you can conclude that the treatment had an effect large enough to be scientifically relevant.

Interpreting a large P value from a paired t test

If the P value is large, the data do not give you any reason to conclude that the treatment had an effect. This is not the same as saying that the treatment had no effect. You just don't have evidence of an effect.

How large could the true treatment effect really be? The average difference between pairs in this experiment is unlikely to equal the true average difference between pairs (because of random variability). There is no way to know what that true difference is. Prism presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true treatment effect. 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 between means that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial decrease	Trivial increase	You can reach a crisp conclusion. Either the treatment has no effect or a tiny one.
Trivial decrease	Large increase	You can't reach a strong conclusion. The data are consistent with the treatment causing a trivial decrease, no change, or an increase that may be large enough to be important. To reach a clear conclusion, you need to repeat the experiment with more subjects.
Large decrease	Trivial increase	You can't reach a strong conclusion. The data are consistent with a trivial increase, no change, or a decrease that may be large enough to be important. You can't make a clear conclusion without repeating the experiment with more subjects.
Large decrease	Large increase	You can't reach any conclusion.

Checklist: Is the paired t test the right test for these data?

Before accepting the results of any statistical test, first think carefully about whether you chose an appropriate test. Before accepting results from a paired t test, ask yourself these questions. Prism can help you answer the first two questions listed below. You'll have to answer the others based on experimental design.

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. Prism tests for violations of this assumption, but normality tests have limited utility (see page 32). 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 Wilcoxon matched pairs nonparametric test instead of the t test.

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. See *Correlation coefficient* on page 93. If r is positive and P 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 you have seen in other similar experiments.

Are the pairs independent?

The results of a paired t test only make sense when the pairs are independent – 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. See *The need for independent samples* on page 10.

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 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 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. See *One-tail vs. two-tail P values* on page 16.

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 look at the ratio (treated/control) rather than the difference (treated – control). See below.

Ratio t tests for paired data

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.

Ratio t test with Prism

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 (see a detailed example below).

1. Transform both columns to logarithms.
2. Perform a paired t test on the transform results.
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?
4. Prism also reports the confidence interval of the difference between means. Since the data being analyzed are logs of the actual values, the difference between means is the same as the mean of the log(ratio). Take the antilog of each end of the interval (with a calculator) to compute the 95% confidence interval of the ratio.

> Note: Ratio t tests (like paired and unpaired t tests) are used to compare two groups when the outcome is a continuous variable like blood pressure or enzyme level. Don't confuse with the analysis of a contingency table, which is appropriate when there are only two possible outcomes (the outcome is a binary variable).

Example of ratio t test

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

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

If you perform a conventional paired t test, the P value is 0.07. The difference between control and treated is not substantial or 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.

It is always important to look at confidence intervals as well as P values. The difference between the means of the log(control) and log(treated) is -0.3042, with the 95% confidence interval extending from -0.3341 to -0.2745. This is equivalent to the confidence interval for the log of the ratio. Take the antilog (10 to the power) of both numbers to get the confidence interval of the ratio. The ratio of the means is 0.496 , with a 95% confidence interval extending from 0.463 to 0.531.

Analyzed with a paired t test, the results are very ambiguous. When the data are analyzed with a ratio t test, the results are very clear – the treatment cut the Km in half.

The results of a Mann-Whitney test

How the Mann-Whitney test works

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. If two values are the same, then they both get the average of the two ranks for which they tie. 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 populations really have the same median, 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).

How to think about the results of a Mann-Whitney test

The Mann-Whitney test is a nonparametric test to compare two unpaired groups. The key result is a P value that answers this question: If the populations really have the same median, what is the chance that random sampling would result in medians as far apart (or more so) as observed in this experiment?

If the P value is small, you can reject the idea that the difference is a due to random sampling, and you can 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. 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.

Checklist: Is the Mann-Whitney test the right test for these data?

Before interpreting the results of any statistical test, first think carefully about whether you have chosen an appropriate test. Before accepting results from a Mann-Whitney test, ask yourself these questions (Prism cannot help you answer them):

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 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. See "The need for independent samples" on page 10.

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 you really want to compare medians?

The Mann-Whitney test compares the medians of two groups. 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. See *One-tail vs. two-tail P values on page 16*.

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.

The results of a Wilcoxon matched pairs test

How the Wilcoxon matched pairs test works

The Wilcoxon test is a nonparametric test that compares two paired groups. It calculates the difference between each set of pairs and analyzes that list of differences. 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 as far from zero (or further) as observed in this experiment?

In calculating the Wilcoxon test, 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 treatment really had no effect overall, what is the chance that random sampling would lead to a sum of ranks as far apart (or more so) as observed here?

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.

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? Here, 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.

How to think about the results of a Wilcoxon matched pairs test

The Wilcoxon matched pairs test is a nonparametric test to compare two paired groups. It is also called the Wilcoxon matched pairs signed ranks test.

The Wilcoxon test analyzes only the differences between the paired measurements for each subject. The P value answers this question: If the median difference really is zero overall, what is the chance that random sampling would result in a median difference as far from zero (or more so) 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.

Checklist: Is the Wilcoxon test the right test for these data?

Before interpreting the results of any statistical test, first think carefully about whether you have chosen an appropriate test. Before accepting results from a Wilcoxon matched pairs test, ask yourself these questions:

Are the pairs independent?

The results of a Wilcoxon test only make sense when the pairs are independent – 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. See *The need for independent samples* on page 10.

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 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, 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. See *One-tail vs. two-tail P values* on page 16.

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.

9. One-way ANOVA and Nonparametric Comparisons

Introduction to comparisons of three or more groups

Prism can compare three or more groups with ordinary or repeated measures ANOVA or with the nonparametric Kruskal-Wallis or Friedman tests. Following ANOVA, Prism can perform the Bonferroni, Tukey, Newman-Keuls, or Dunnett's post test. Following nonparametric ANOVA, Prism can calculate the Dunn's post test.

These tests compare measurements (continuous variables) such as weight, enzyme activity, and receptor number. To compare proportions see *Contingency Tables* on page 99. To compare survival curves, see *Survival Curves* on page 107.

One-way ANOVA (and related nonparametric tests) compare three or more groups when the data are categorized in one way. For example, you might compare a control group with two treated groups. If your data are categorized in two ways (for example you want to compare control with two treated groups in both males and females) you'll want to use two-way ANOVA as explained beginning on page 76.

Entering data for ANOVA (and nonparametric tests)

From the Welcome (or New Table) dialog, choose any graph from the One grouping variable tab. Or choose to format the data table directly, and choose single columns of Y values and no X column (since X values are ignored by the ANOVA analysis).

The groups do not have be the same size (it's OK to leave some cells empty).

If you have already averaged your data, format the data table for mean, SD (or SEM), and N. With this format, you can't pick nonparametric or paired tests, which require raw data. Enter data on only the first row.

If you format the table for replicate Y values, Prism averages the replicates and bases the one-way ANOVA analysis only on the means.

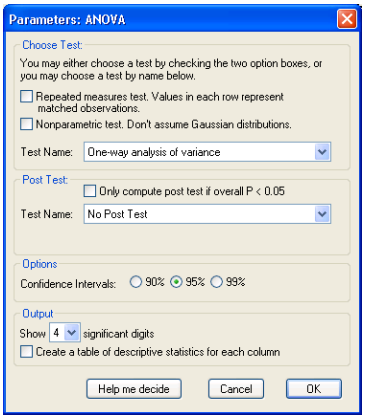
Indexed data

Many statistics programs expect you to enter data in an indexed format, as shown below. One column contains all the data, and the other column designates the group. Prism cannot analyze data entered in index format. If you have indexed data from another program, choose to "unstack" your data when you import it (an option on the Filter tab of the Import data dialog). This rearranges your data to a format Prism can analyze. Read the chapter on importing data in the Prism User's Guide.

Group	Value
1	34
1	43
1	39
2	45
2	47
2	52
3	76
3	99
3	82

Choosing one-way ANOVA and related analyses

Start from the data or results table you wish to analyze (see *Entering data for ANOVA (and nonparametric tests)* on page 57. Press **Analyze** and choose to do a built-in analysis. Then select **One-way ANOVA** from the list of statistical analyses. If you don't wish to analyze all columns in the table, select the columns you wish to compare. Press OK to bring up the Parameters dialog.



Repeated measures test?

You should choose repeated measures test when the experiment uses matched subjects. Here are some examples:

- ✓ You measure a variable in each subject before, during and after an intervention.
- ✓ You recruit subjects as matched sets. Each subject in the set has the same age, diagnosis, and other relevant variables. One of the set gets treatment A, another gets treatment B, another gets treatment C, etc.
- ✓ You run a laboratory experiment several times, each time with a control and several treated preparations handled in parallel.
- ✓ You measure a variable in triplets, or grandparent/parent/child groups.

More generally, you should select a repeated measures test whenever you expect a value in one group to be closer to a *particular* value in the other groups than to a *randomly selected* value in the other group.

Ideally, the decision about repeated measures analyses should be made before the data are collected. Certainly the matching 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 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 other 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.

ANOVA or nonparametric test?

ANOVA, as well as other statistical tests, assumes that you have sampled data from populations that follow a Gaussian bell-shaped distribution. Biological data never follow a Gaussian distribution precisely, because a Gaussian distribution extends infinitely in both directions, so it includes both infinitely low negative numbers and infinitely high positive numbers! Many kinds of biological data, however, do follow a bell-shaped distribution that is approximately Gaussian. Because ANOVA works 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. But there is a drawback. Nonparametric tests are less powerful than the parametric tests that assume Gaussian distributions. This means that P values tend to be higher, making it harder to detect real differences as being statistically significant. If the samples are large the difference in power is minor. With small samples, nonparametric tests have little power to detect differences. With very small groups (just a few values in each groups), some nonparametric tests have zero power – the P value will always be greater than 0.05.

You may find it difficult to decide when to select nonparametric tests. You should definitely choose a nonparametric test in these situations:

- ✓ The outcome variable is a rank or score with only a few categories. Clearly the population is far from Gaussian in these cases.
- ✓ One, or a few, values are 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. Using a nonparametric test with these data is easy. Assign an arbitrary low value to values too low to measure, and an arbitrary high value to values too high to measure. Since the nonparametric tests only consider the relative ranks of the values, it won't matter that you didn't know one (or a few) of the values exactly.
- ✓ You are sure that the population is far from Gaussian. Before choosing a nonparametric test, consider transforming the data (perhaps to logarithms or reciprocals). Sometimes a simple transformation will convert non-Gaussian data to a Gaussian distribution.

In many situations, perhaps most, you will find it difficult to decide whether to select nonparametric tests. Remember that the Gaussian assumption is about the distribution of the overall population of values, not just the sample you have obtained in this particular

experiment. Look at the scatter of data from previous experiments that measured the same variable. Also consider the source of the scatter. When variability is due to the sum of numerous independent sources, with no one source dominating, you expect a Gaussian distribution.

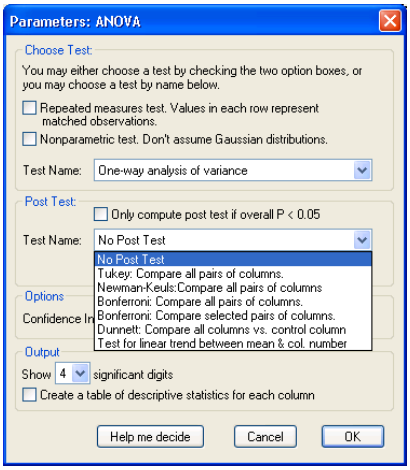
Prism performs normality testing in an attempt to determine whether data were sampled from a Gaussian distribution, but normality testing is less useful than you might hope (see page 32). Normality testing doesn't help if you have fewer than a few dozen (or so) values.

Your decision to choose a parametric or nonparametric test matters the most when samples are small for reasons summarized here:

	Large samples (> 100 or so)	**Small samples (<12 or so)**
Parametric tests	Robust. P value will be nearly correct even if population is fairly far from Gaussian.	Not robust. If the population is not Gaussian, the P value may be misleading.
Nonparametric test	Powerful. If the population is Gaussian, the P value will be nearly identical to the P value you would have obtained from a parametric test. With large sample sizes, nonparametric tests are almost as powerful as parametric tests.	Not powerful. If the population is Gaussian, the P value will be higher than the P value obtained from ANOVA. With very small samples, it may be impossible for the P value to ever be less than 0.05, no matter how the values differ.
Normality test	Useful. Use a normality test to determine whether the data are sampled from a Gaussian population.	Not very useful. Little power to discriminate between Gaussian and non-Gaussian populations. Small samples simply don't contain enough information to let you make inferences about the shape of the distribution in the entire population.

Which post test?

If you are comparing three or more groups, you may pick a post test to compare pairs of group means. Prism offers these choices:



Choosing an appropriate post test is not straightforward, and different statistics texts make different 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, you have three choices: the Bonferroni, Tukey, or Newman-Keuls (also known as the Student-Newman-Keuls or SNK) tests. The only advantage of the Bonferroni method is that it is easy to understand. 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. Choosing between the ***Tukey*** and ***Newman-Keuls*** test is not straightforward, and there appears to be no real consensus among statisticians. The two methods are related, and the rationale for the differences is subtle. The methods are identical when comparing the largest group mean with the smallest. For other comparisons, the Newman-Keuls test yields lower P values. The problem is that it is difficult to articulate exactly what null hypotheses the Newman-Keuls P values test. For that reason, and because the Newman-Keuls test does not generate confidence intervals, we suggest selecting Tukey's test. (If you select the Tukey test, you are actually selecting the Tukey-Kramer test, which includes the extension by Kramer to allow for unequal sample sizes.)

Confirm test selection

Based on the option boxes you selected, Prism will choose a test for you:

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

The results of one-way ANOVA

Interpreting the results of one-way ANOVA

P value

One-way ANOVA compares three or more unmatched groups, based on the assumption that the populations are Gaussian. 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?

R^2 value from one-way ANOVA

R^2 is the fraction of the overall variance (of all the data, pooling all the groups) attributable to differences among the group means. It compares the variability among group means with the variability within the groups. A large value means that a large fraction of the variation is due to the treatment that defines the groups. The R^2 value is calculated from the ANOVA table and equals the between group sum-of-squares divided by the total sum-of-squares. Some programs (and books) don't bother reporting this value. Others refer to it as η^2 (eta squared) rather than R^2. It is a descriptive statistic that quantifies the strength of the relationship between group membership and the variable you measured.

F ratio and ANOVA table (one-way ANOVA)

The F ratio and P value are computed from the ANOVA table. For completeness, Prism presents the complete table but you'll rarely need to look at more than the P value. If you want to understand the ANOVA table in details, you'll need to consult an advanced statistics book. The description below is not complete, and is intended to just give you a sense of what is going on.

ANOVA Table	SS	df	MS
Treatment (between column	13.61	2	6.804
Residual (within columns)	27.23	15	1.815
Total	40.84	17	

The key idea is that ANOVA partitions the variability among 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 answers this question: If the populations all have the same mean, what is the chance that randomly selected groups would lead to an F ratio as big (or bigger) as the one obtained in your experiment?

Bartlett's test for equal variances

ANOVA is based on the assumption that the populations all have the same variance. If your samples have four or more 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 variance, what is the chance that you'd randomly select samples whose variances are as different (or more different) as observed in your experiment? (Since the variance is the standard deviation squared, testing for equal variances is the same as testing for equal standard deviations).

Bartlett's test can be misleading, since it is very sensitive to deviations from a Gaussian distribution – more sensitive than are 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).

If the P value is small, you must decide whether you will conclude that the variances 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 variances is as important as the finding of different means. If the variances 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 variances, and then rerun the ANOVA. Often you'll find that converting values to their reciprocals or logarithms will equalize the variances and make the distributions more Gaussian.
- ✓ Use a modified ANOVA that does not assume equal variances. Prism does not provide such a test.

Post tests following one-way ANOVA

Post test for a linear 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 r^2, 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 r^2 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.

Other post tests

The Bonferroni, Tukey, Newman-Keuls and Dunnett's post tests are all modifications of t tests. They account for multiple comparisons, as well as for the fact that the comparisons are interrelated.

Recall that an unpaired t test computes the t ratio as the difference between two group means divided by the standard error of the difference (computed from the standard errors of the two group means and the two sample sizes). The P value is then derived from t. The post tests work in a similar way. Instead of dividing by the standard error of the difference, they divide by a value computed from the residual mean square (shown on the ANOVA table). Each test uses a different method to derive a P value from this ratio.

For the difference between each of a pair of means, Prism reports the P value as >0.05, <0.05, <0.01, or <0.001. These P values account for multiple comparisons. Prism reports a P value for the difference between each pair of means, but the probability values apply to the entire family of comparisons, not to each individual comparison. It is easier to think about this by asking about the chance that none of the comparisons are statistically significant. If the null hypothesis is true (all the values are sampled from populations with the same mean) there is a 95% chance that *every one of* the post tests will find no significant difference (P>0.05). This leaves a 5% that at least one of the post tests will find a significant difference at the 5% level (P<0.05).

Why does Prism only report the P value as less or greater than some standard value? Why not report the exact P value? There are two reasons:

- ✓ One reason is that the critical values for most post tests (Bonferroni is an exception) come from tables that are difficult to compute. Prism and InStat simply read 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 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. The P value is a probability that answers a question, and the question is about the family of comparisons, not about individual comparisons.

Prism also reports a 95% confidence interval for the difference between each pair of means (except for the Newman-Keuls post test, which cannot be used for confidence intervals). These intervals account for multiple comparisons. There is a 95% chance that *all* of these intervals contain the true differences between population, and only a 5% chance that any one or more of these intervals misses the true difference. A 95% confidence interval is computed for the difference between each pair of means, but the 95% probability applies to the entire family of comparisons, not to each individual comparison.

How to think about results from one-way ANOVA

One-way ANOVA compares the means of three or more groups, assuming that data are sampled from Gaussian populations. The most important results are the P value and the post tests.

The overall P value answers this question: If the populations really have the same mean, what is the chance that random sampling would result in means as far apart from one another (or more so) as you 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.

How to think about the results of post tests

If the columns are organized in a natural order, the post test for linear trend tells you whether the column means have a systematic trend, increasing or decreasing) as you go from left to right in the data table. See *Post test for a linear trend* on page 63.

With other post tests, look at which differences between column means are statistically significant. For each pair of means, Prism reports whether the P value is less than 0.05, 0.01 or 0.001.

"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.

As discussed below, you will interpret the post test results differently depending on whether the difference is statistically significant or not.

If the difference is statistically significant

If the P value for a post test is small, then it is unlikely that the difference you observed is due to random sampling. You can reject the idea that those two populations have identical means.

Because of random variation, the difference between the group means in this experiment is unlikely to equal the true difference between population means. There is no way to know what that true difference is. With most post tests (but not the Newman-Keuls test), Prism presents the uncertainty as a 95% confidence interval for the difference between all (or selected) pairs of means. You can be 95% sure that this interval contains the true difference between the two means.

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference between means that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial difference	Trivial difference	Although the true difference is not zero (since the P value is low) the true difference between means is tiny and uninteresting. The treatment had an effect, but a small one.
Trivial difference	Important difference	Since the confidence interval ranges from a difference that you think is biologically trivial to one you think would be important, you can't reach a strong conclusion from your data. You can conclude that the means are different, but you don't know whether the size of that difference is scientifically trivial or important. You'll need more data to reach a clear conclusion.
Important difference	Important difference	Since even the low end of the confidence interval represents a difference large enough to be considered biologically important, you can conclude that there is a difference between treatment means and that the difference is large enough to be scientifically relevant.

If the difference is not statistically significant

If the P value from a post test is large, the data do not give you any reason to conclude that the means of these two groups 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.

How large could the true difference really be? Because of random variation, the difference between the group means in this experiment is unlikely to equal the true difference between population means. There is no way to know what that true difference is. Prism presents the uncertainty as a 95% confidence interval (except with the Newman-Keuls test). 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 for each pair of means, and ask whether those differences would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial decrease	Trivial increase	You can reach a crisp conclusion. Either the means really are the same or they are different by a trivial amount. At most, the true difference between means is tiny and uninteresting.
Trivial decrease	Large increase	You can't reach a strong conclusion. The data are consistent with the treatment causing a trivial decrease, no change, or an increase that might be large enough to be important. To reach a clear conclusion, you need to repeat the experiment with more subjects.
Large decrease	Trivial increase	You can't reach a strong conclusion. The data are consistent with a trivial increase, no change, or a decrease that may be large enough to be important. You can't make a clear conclusion without repeating the experiment with more subjects.
Large decrease	Large increase	You can't reach any conclusion. Repeat the experiment with a much larger sample size.

Checklist: Is one-way ANOVA the right test for these data?

Before accepting the results of any statistical test, first think carefully about whether you chose an appropriate test. Before accepting results from a one-way ANOVA, ask yourself the questions below. Prism can help answer the first two questions. You'll need to answer the others based on experimental design.

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 32). 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.

See *Bartlett's test for equal variances* on page 62.

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. See *The need for independent samples* on page 10.

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, also called two factor ANOVA. See page 76.

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. If you need to perform

ANOVA with random effects variables, consider using the program NCSS from www.ncss.com.

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 (say glucose in column A, insulin concentration in column B, and glycosylated hemoglobin in column C), instead of different treatment groups, then one-way ANOVA is not an appropriate analysis.

The results of repeated-measures one-way ANOVA

How repeated-measures ANOVA works

Repeated-measures one-way ANOVA compares three or more matched groups, based on the assumption that the differences between matched values are Gaussian. For example, one-way ANOVA may compare measurements made before, during and after an intervention, when each subject was assessed three times. The P value answers this question: If the populations really have the same mean, what is the chance that random sampling would result in means as far apart (or more so) as observed in this experiment?

ANOVA table

The P value is 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.

The circularity assumption

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. 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. 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. See *Repeated measures test?* on page 58

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 is 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.

How to think about results from 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.

Your approach to interpreting repeated-measures ANOVA results will be the same as interpreting the results of ordinary one-way ANOVA. See *How to think about results from one-way ANOVA* on page 64.

Checklist: Is repeated-measures one way ANOVA the right test for these data?

Before accepting the results of any statistical test, first think carefully about whether you chose an appropriate test. Before accepting results from repeated-measures one-way ANOVA, ask yourself the questions listed below. Prism can help you answer the first question. You must answer the remaining questions based on experimental design.

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. See *The need for independent samples on page 10*.

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.

The results of a Kruskal-Wallis test

How the Kruskal-Wallis test works

The Kruskal-Wallis test is a nonparametric test that compares three or more unpaired groups. To perform the Kruskal-Wallis test, Prism first ranks all the values from low to high, without regard to which group each value belongs. If two values are the same, then they both get the average of the two ranks for which they tie. 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. Prism then sums the ranks in each group, and reports the sums. If the sums of the ranks are very different, the P value will be small.

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 populations really have the same median, what is the chance that random sampling would result in sums of ranks as far apart (or more so) as observed in this experiment? More precisely, if the null hypothesis is true, then what is

the chance of obtaining a Kruskal-Wallis statistic as high (or higher) as observed in this experiment?

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.

How Dunn's post test works

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.

How to think about the results of a Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric test to compare three or more unpaired groups. It is also called Kruskal-Wallis one-way analysis of variance by ranks. The key result is a P value that answers this question: If the populations really have the same median, what is the chance that random sampling would result in medians as far apart (or more so) as you observed in this experiment?

If the P value is small, you can reject the idea that the differences are all due to chance. This doesn't mean that every group differs from every other group, only that at least one group differs from one of the others. Look at the post test results to see which groups differ from which other groups.

If the overall Kruskal-Wallis 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 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 the groups differ.

How to think about post tests following the Kruskal-Wallis test

Dunn's post test calculates a P value for each pair of columns. These P values answer this question: If the data were sampled from populations with the same median, what is the chance that one or more pairs of columns would have medians as far apart as observed here? If the P value is low, you'll conclude that the difference is statistically significant. 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 separately to each individual comparison but rather to the *entire family* of comparisons.

Checklist: Is the Kruskal-Wallis test the right test for these data?

Before interpreting the results of any statistical test, first think carefully about whether you have chosen an appropriate test. Before accepting results from a Kruskal-Wallis test, ask yourself these questions about your experimental design:

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. See *The need for independent samples* on page 10.

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.

The results of a Friedman test

How the Friedman test works

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 your samples are small, 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.

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.

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.

How to think about the results of a Friedman test

The Friedman test is a nonparametric test to compare three or more matched groups. It is also called Friedman two-way analysis of variance by ranks (because repeated-measures one-way ANOVA is the same as two-way ANOVA without any replicates.)

The P value answers this question: If the median difference really is zero, what is the chance that random sampling would result in a median difference as far from zero (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.

How to think about post tests following the Friedman 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.

Checklist: Is the Friedman test the right test for these data?

Before interpreting the results of any statistical test, first think carefully about whether you have chosen an appropriate test. Before accepting results from a Friedman test, ask yourself these questions:

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.

10. Two-way ANOVA

Introduction 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.

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).

Prism can perform ordinary two-way ANOVA accounting for repeated measures when there is matching on one of the factors (but not both). Prism cannot perform any kind of nonparametric two-way ANOVA.

Entering data for two-way ANOVA

Arrange your data so the data sets (columns) represent different levels of one factor, and different rows represent different levels of the other factor. For example, to compare three time points in men and women, enter your data like this:

	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

The ANOVA calculations ignore any X values. You may wish to format the X column as text in order to label your rows. Or you may omit the X column altogether, or enter numbers.

You may leave some replicates blank and still perform ordinary two-way ANOVA (so long as you enter at least one value in each row for each data set). You cannot perform repeated-measures ANOVA if there are any missing values. Prism cannot perform repeated-measures two-way ANOVA, if any values are missing for any subject. However, Prism can perform repeated measures two-way ANOVA with different numbers of subjects in each group, so long as you have complete data (at each time point or dose) for each subject.

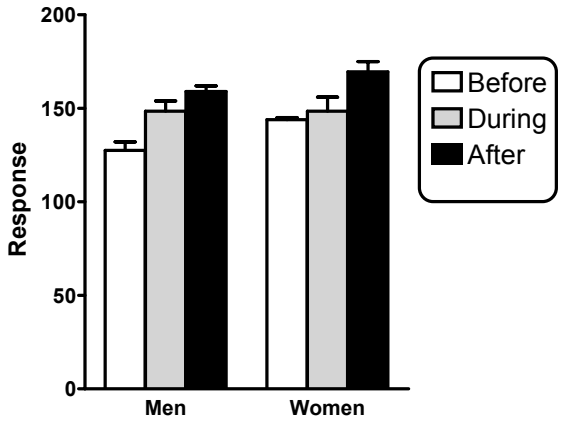
If you have averaged your data elsewhere, you may format the data table to enter mean, SD or SEM, and N. The N values do not have to all be the same, but you cannot leave N blank. You cannot perform repeated-measures two-way ANOVA if you enter averaged data.

Some programs expect you to enter two-way ANOVA data in an indexed format. All the data are in one column, and two other columns are used to denote different levels of the two factors. You cannot enter data in this way into Prism. Prism can import indexed data with a single index variable, but cannot import data with two index variables, as would be required for two-way ANOVA.

As with one-way ANOVA, it often makes sense to transform data by converting to logarithms or reciprocals to make the distributions more Gaussian.

When entering your data, you have to choose which grouping variable goes where. In the example above, we could have entered Men and Women into two columns (data sets) and Before, During and After into three rows. When deciding which grouping variable is denoted by rows and which by columns, keep two things in mind:

- ✓ When you create a bar graph of the data, each column will create bars which can have a different fill pattern and color. So in the example above, there will be one kind of bar for Before, another for During and another for After. Men and Women will appear as two bars of identical appearance within each of the three time points.

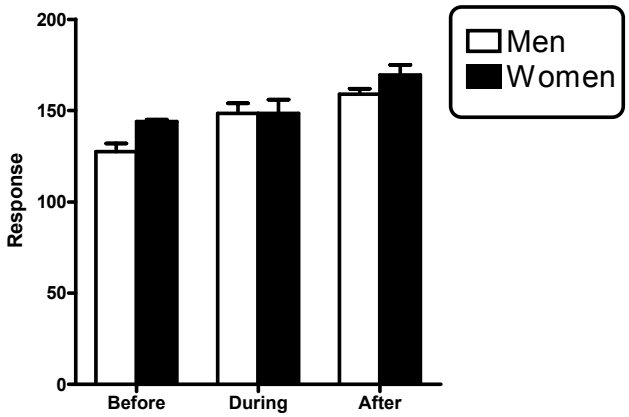


- ✓ Prism can only perform post tests within a row, comparing columns. When the data are entered as shown above, Prism can compute post tests comparing Before vs. During, Before vs. After, and During vs. After within each row. But Prism cannot compare Men vs. Women since that is comparing two rows.

You could choose to enter the same data in an alternative manner like this:

	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

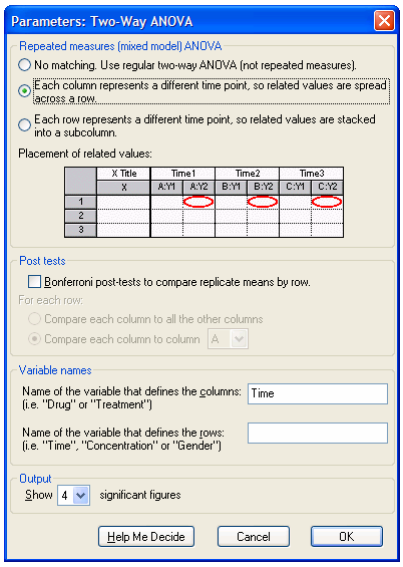
Now the bar graph will have different fill patterns (or colors) for men and women, with repeated bars for the three time points.



When entered this way, the post tests would compare men vs. women in each of the three time points. The rest of the two-way ANOVA results will be identical no matter how you enter your data.

Choosing the two-way ANOVA analysis

Start from the data or results table you wish to analyze (see *Entering data for two-way ANOVA* on page 76). Click **Analyze** and choose to do a built-in analysis. Then choose **Two-way ANOVA** from the list of statistical analyses



Variable names

Label the two factors to make the output more clear. If you don't enter names, Prism will use the generic names "Column factor" and "Row factor".

Repeated measures

You should choose a repeated-measures analysis when the experiment used paired or matched subjects. See *Repeated measures test?* on page 58. 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*.

	X Labels	A		B	
		Control		Treated	
	X	A:Y1	A:Y2	B:Y1	B:Y2
1	One	23	24	28	31
2	Two	34	41	41	54
3	Three	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.

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.

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 columns came from three other animals. The matching was by column.

The term *repeated measures* is appropriate for those examples, because you made repeated measurements from each animal. Some experiments involve matching but no 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 only uses the *term repeated measures* for any analysis where subjects were matched, regardless of whether measurements were actually repeated in those subjects.

It is also possible to design experiments with repeated measures in both directions. Here is an example: 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 columns (23, 34, 43 and 28, 41, 56). Prism cannot perform two-way ANOVA with repeated measures in both directions, and so cannot analyze this experiment.

Don't confuse replicates with repeated measures. Here is an example: 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.

Post tests following two-way ANOVA

Prism can perform post tests on each row. In the example above, data set A was control and data set B was treated. Each row represents a different time. 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).

Although other kinds of post tests are possible after two-way ANOVA, Prism only performs the post tests described above (which biologists use most frequently).

The results of two-way ANOVA

How two-way ANOVA works

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.

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.

With repeated-measures ANOVA, there is a fifth component: variation between subjects.

The ANOVA table shows how the sum of squares is partitioned into the four (or five) 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.

How Prism computes two-way ANOVA

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.

ANOVA from data entered as mean, SD (or SEM) and N

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.

Two-way ANOVA calculations with 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.

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.

The concept of repeated measures doesn't apply when your data are unreplicated.

Repeated-measures two-way ANOVA

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.

Post tests following 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.

For each row, Prism calculates:

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

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.

Confidence intervals at each row are computed using this equation:

$$\text{Span} = t^* \cdot \sqrt{\text{MS}_{\text{residual}}\left(\frac{1}{\text{N}_1} + \frac{1}{\text{N}_2}\right)}$$

$$95\%\ \text{CI: } \left[(\text{mean}_2 - \text{mean}_1) - \textit{Span}\ \right] \ \textit{to}\ \left[(\text{mean}_2 - \text{mean}_1) + \text{Span}\right]$$

The critical value of t is abbreviated t* in that equation (not a standard abbreviation). Its value does not depend on your data, only on your experimental design. It depends on the number of degrees of freedom and the number of rows (number of comparisons).

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, use $(SS_{subject} + SS_{residual})/(DF_{subject} + DF_{residual})$ instead of $MS_{residual}$ in the equation above, and set the number of degrees of freedom to the sum of $DF_{subject}$ and $DF_{residual}$.

How to think about results from two-way ANOVA

Two-way ANOVA partitions the overall variance of the outcome variable into three components, plus a residual (or error) term.

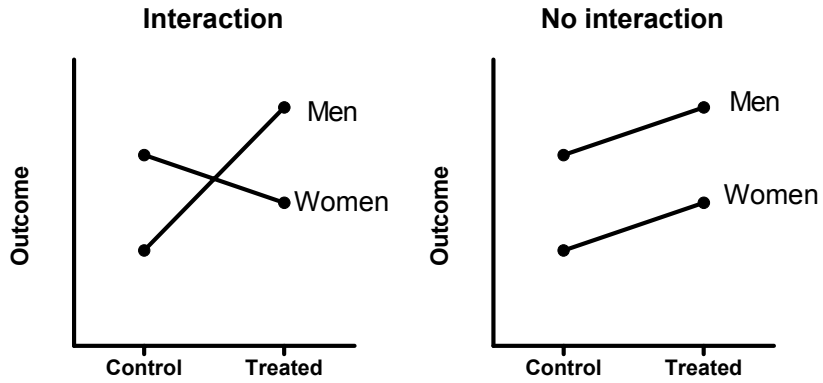
Interaction

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 col-

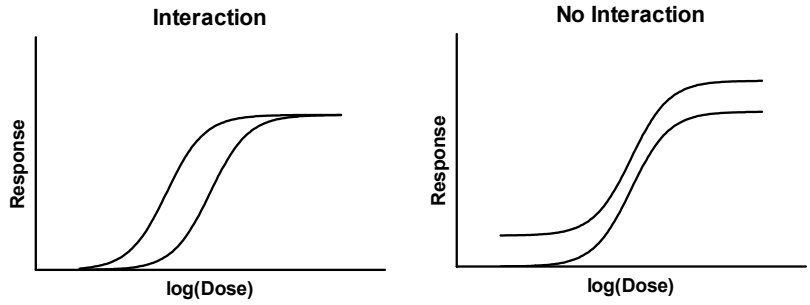
umns are the same for each row and that any systematic differences between rows are the same for each column. If columns represent drugs and rows represent gender, then the null hypothesis is that the differences between the drugs are consistent for men and women.

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? Often the test of interaction is the most important of the three tests.

If you graph the data, as shown below, there is no interaction when the curves are "parallel". In the graph below, the left panel shows interaction between treatment and gender, while the right panel shows no interaction.



But term "parallel" can be ambiguous. Pharmacologists consider two dose-response curves "parallel" when two drugs have similar effects at very low and very high concentrations, but different (and horizontally parallel) effects at moderate concentrations. Two-way ANOVA of such data would reject the null hypothesis of no interaction, because the difference between Y values in the middle of the curves is very different than the difference at the ends.



If you entered only a single value for each row/column pair, it is impossible to test for interaction between rows and columns. Instead, 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. The assumption cannot be tested without replicate values.

Note: If the interaction is statistically significant, it is difficult to interpret the row and column effects. Statisticians often recommend ignoring the tests of row and column effects when there is a significant interaction.

Column factor

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. If columns represent different drugs, the null hypothesis is that all the drugs produced the same effect. 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?

Row factor

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. If the rows represent gender, 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?

Subject (matching)

For repeated-measures ANOVA, Prism 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 to think about post tests following two-way ANOVA

If you have two data sets (columns), Prism can perform post tests to compare the two means from each row.

For each row, Prism reports the 95% confidence interval for the difference between the two means. These confidence intervals adjust for multiple comparisons, so you can be 95% certain that *all* the intervals contain the true difference between means.

For each row, Prism also reports the P value testing the null hypothesis that the two means are really identical. Again, the P value computations take into account multiple comparisons. If there really are no differences, there is a 5% chance that any one (or more) of the P values will be less than 0.05. The 5% probability applies to the entire family of comparisons, not to each individual P value.

If the difference is statistically significant

If the P value for a post test is small, then it is unlikely that the difference you observed is due to random sampling. You can reject the idea that those two populations have identical means.

Because of random variation, the difference between the group means in this experiment is unlikely to equal the true difference between population means. There is no way to know what that true difference is. With most post tests (but not the Newman-Keuls test), Prism presents the uncertainty as a 95% confidence interval for the difference between all (or selected) pairs of means. You can be 95% sure that this interval contains the true difference between the two means.

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent a difference between means that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial difference	Trivial difference	Although the true difference is not zero (since the P value is low) the true difference between means is tiny and uninteresting. The treatment had an effect, but a small one.
Trivial difference	Important difference	Since the confidence interval ranges from a difference that you think is biologically trivial to one you think would be important, you can't reach a strong conclusion from your data. You can conclude that the means are different, but you don't know whether the size of that difference is scientifically trivial or important. You'll need more data to draw a clear conclusion.
Important difference	Important difference	Since even the low end of the confidence interval represents a difference large enough to be considered biologically important, you can conclude that there is a difference between treatment means and that the difference is large enough to be scientifically relevant.

If the difference is not statistically significant

If the P value from a post test is large, the data do not give you any reason to conclude that the means of these two groups 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 evidence that they differ.

How large could the true difference really be? Because of random variation, the difference between the group means in this experiment is unlikely to equal the true difference between population means. There is no way to know what that true difference is. Prism presents the uncertainty as a 95% confidence interval (except with the Newman-Keuls test). 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 for each pair of means, and ask whether those differences would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial decrease	Trivial increase	You can reach a crisp conclusion. Either the means really are the same or they are different by a trivial amount. At most, the true difference between means is tiny and uninteresting.
Trivial decrease	Important increase	You can't reach a strong conclusion. The data are consistent with the treatment causing a trivial decrease, no change, or a large increase. To reach a clear conclusion, you need to repeat the experiment with more subjects.
Important decrease	Trivial increase	You can't reach a strong conclusion. The data are consistent with a trivial increase, no change, or a decrease that may be large enough to be important. You can't make a clear conclusion without repeating the experiment with more subjects.
Important decrease	Large increase	You can't reach any conclusion. Repeat the experiment with a much larger sample size.

Problems with post tests following two-way ANOVA

Post test are often used to compare dose-response curves or time course curves. Using two-way ANOVA in this way presents two problems. One problem is that ANOVA treats different doses (or time points) exactly as it deals with different species or different drugs. ANOVA ignores the fact that doses or time points come in order. You could jumble the doses in any order and get exactly the same ANOVA results. However, you did the experiment to observe a trend, so you should be cautious about interpreting results from an analysis method that doesn't recognize trends.

Another problem with the ANOVA approach is that it is hard to interpret the results. Knowing at which doses or time points the treatment had a statistically significant effect doesn't always help you understand the biology of the system and rarely helps you design new experiments. Some scientists like to ask which is the lowest dose (or time) at which the effect of the treatment is statistically significant. The post tests 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. 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. Finding the minimum dose that causes a "statistically significant" effect is rarely helpful, as the answer depends on sample size. With a large enough sample size (at each dose), you'll find that a tiny dose causes a statistically significant, but biologically trivial, effect.

Checklist: Is two-way ANOVA the right test for these data?

Before accepting the results of any statistical test, first think carefully about whether you chose an appropriate test. Before accepting results from a two-way ANOVA, ask yourself these questions:

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 matched?

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. See *The need for independent samples* on page 10.

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. See *Model I (fixed effects) vs. Model II (random effects) ANOVA* on page 81.

The circularity assumption in two-way repeated-measures ANOVA

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. 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.

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).

Calculating 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. 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 if you have more than two columns, Prism won't perform any post tests. And even with two columns, 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. 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).

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

to

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

Variable	Explanation
mean1, mean2	The mean of the two groups you want to compare.
N1, N2	The sample size of the two groups you want to compare.
t*	Critical value from the student t distribution. This variable is defined 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-tailed 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.

Variable	Explanation
$MS_{residual}$	The mean square for residuals calculated by Prism. The value of $MS_{residual}$ will be the same for each comparison. If you performed repeated measures two-way ANOVA, use the mean-square for subject instead.

To determine significance levels, calculate for each comparison:

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

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 at www.graphpad.com.

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.

11. Correlation

Introduction to correlation

When two variables vary together, statisticians say that there is a lot of *covariation* or *correlation*. The correlation coefficient, r, quantifies the direction and magnitude of correlation.

Correlation is not the same as linear regression, but the two are related. Linear regression finds the line that best predicts Y from X. Correlation quantifies how well X and Y vary together. In some situations, you might want to perform both calculations.

Correlation only makes sense when *both X and Y* variables are outcomes you measure. If you control X (often, you will have controlled variables such as time, dose, or concentration), don't use correlation, use linear regression.

> Tip: Linear and nonlinear regression are explained in the companion book, Fitting Biological Data to Models Using Linear and Nonlinear Regression.

Correlation calculations do not discriminate between X and Y, but rather quantify the relationship between the two variables. Linear regression does discriminate between X and Y. Linear regression finds the straight line that best predicts Y from X by minimizing the sum of the square of the vertical distances of the points from the regression line. The X and Y variables are not symmetrical in the regression calculations. Therefore only choose regression, rather than correlation, if you can clearly define which variable is X and which is Y.

Entering data for correlation

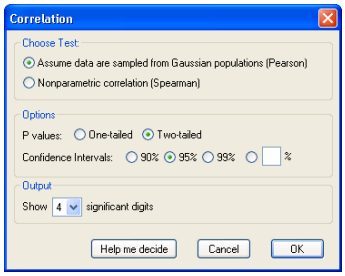
From the Welcome (or New Table) dialog, choose the XY graph tab and then choose the first graph type (no error bars). Or choose to format the data table directly, and choose single columns of Y values and numbers for the X column.

If you format the Y columns for replicates (for example, triplicates), Prism averages each set of replicates and considers this list of averages to be Y. If you format the Y columns to enter mean and SD (or SEM) on each row, Prism analyzes only the means and ignores the SD or SEM values.

If you enter Y values for several data sets (column A, B and C), Prism will report correlation results for X vs. YA, for X vs. YB, and for X vs. YC. However, Prism does not report the correlation of YA with YB, or YA with YC, etc.

Choosing a correlation analysis

To calculate correlation, start from your data table. Click the **Analyze** button, select **Built-in analysis**, and then select **Correlation** from the list of statistical analyses. If you have entered replicate values, Prism computes the average of each set of replicates and uses those averages for correlation calculations. If you entered mean and SD (or SEM) on each row, Prism performs calculations with the mean values, but ignores the SD or SEM values.



Choose one of two correlation tests. **Pearson correlation** calculations are based on the assumption that both X and Y values are sampled from populations that follow a Gaussian distribution, at least approximately. With large samples, this assumption is not too important. If you don't wish to make the Gaussian assumption, select **Nonparametric (Spearman) correlation** instead. Spearman correlation makes no assumption about the distribution of the values, as the calculations are based on ranks, not the actual values.

Prism can compute either a one-tailed or two-tailed P value. We suggest almost always choosing a two-tailed P value. You should only choose a one-tail P value when you have specified the anticipated sign of the correlation coefficient before collecting any data and are willing to attribute any correlation in the "wrong" direction to chance, no matter how large it is. See *One-tail vs. two-tail P values* on page 16.

For help interpreting the results, see *Results of correlation* on page 93.

Results of correlation

How correlation works

Correlation coefficient

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

Value of r (or r_s)	Interpretation
r = 0	The two variables do not vary together at all.
between 0 and 1	The two variables tend to increase or decrease together.
r = 1.0	Perfect correlation.
between 0 and -1	One variable increases as the other decreases.
r = -1.0	Perfect negative or inverse correlation.

If r or r_s is far from zero, there are four possible explanations:

- ✓ Changes in the X variable change the value of the Y variable.
- ✓ Changes in the Y variable change the value of the X variable.
- ✓ Changes in another variable influence both X and Y.
- ✓ X and Y don't really correlate at all, and you just happened to observe such a strong correlation by chance. The P value determines how often this could occur.

r^2 (from correlation)

Perhaps the best way to interpret the value of r is to square it to calculate r^2. Statisticians call this quantity the *coefficient of determination*, but scientists call it r *squared*. It is

a value that ranges from zero to one, and is the fraction of the variance in the two variables that is "shared". For example, if r^2=0.59, then 59% of the variance in X can be explained by variation in Y. Likewise, 59% of the variance in Y can be explained by variation in X. More simply, 59% of the variance is shared between X and Y.

Prism only calculates an r^2 value from the Pearson correlation coefficient. It is not appropriate to compute r^2 from the nonparametric Spearman correlation coefficient.

P value from correlation

The P value answers this question: If there really is no correlation between X and Y in the overall population, what is the chance that random sampling would result in a correlation coefficient as far from zero (or further) as observed in this experiment?

How to think about results of linear correlation

Look first at a graph of your data to see how X and Y vary together. Then look at the value of r (or r_s) which quantifies the correlation. Finally, look at the P value.

If the P value is small, you can reject the idea that the correlation is due to random sampling. Look at the confidence interval for r. You can be 95% sure that the true population r lies somewhere within that range.

If the P value is large, the data do not give you any reason to conclude that the correlation is real. This is not the same as saying that there is no correlation at all. You just have no compelling evidence that the correlation is real and not due to chance. Look at the confidence interval for r. It will extend from a negative correlation to a positive correlation. If the entire interval consists of values near zero that you would consider biologically trivial, then you have strong evidence that either there is no correlation in the population or that there is a weak (biologically trivial) association. On the other hand, if the confidence interval contains correlation coefficients that you would consider biologically important, then you couldn't make any strong conclusion from this experiment. To make a strong conclusion, you'll need data from a larger experiment.

Checklist. Is correlation the right analysis for these data?

To check that correlation is an appropriate analysis for these data, ask yourself these questions. Prism cannot help answer them.

Are the subjects independent?

Correlation assumes that any random factor affects only one subject, and not others. You would violate this assumption if you choose half the subjects from one group and half from another. A difference between groups would affect half the subjects and not the other half.

Are X and Y measured independently?

The calculations are not valid if X and Y are intertwined. You'd violate this assumption if you correlate midterm exam scores with overall course score, as the midterm score is one of the components of the overall score.

Were X values measured (not controlled)?

If you controlled X values (e.g., concentration, dose, or time) you should calculate linear regression rather than correlation.

Is the covariation linear?

A correlation analysis would not be helpful if Y increases as X increases up to a point, and then Y decreases as X increases further. You might obtain a low value of r, even

though the two variables are strongly related. The correlation coefficient quantifies linear covariation only.

Are X and Y distributed according to Gaussian distributions?

To accept the P value from standard (Pearson) correlation, the X and Y values must each be sampled from populations that follow Gaussian distributions. Spearman nonparametric correlation does not make this assumption.

Part C:
Categorical and Survival Data

12. 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 (Biometrika 26:404-413, 1934). The result is labeled an "exact" confidence interval (in contrast to the approximate intervals you can calculate conveniently by hand).

If you want to compute the 95% confidence interval by hand, most books present the Wald equation:

$$\left[p - \left(1.96\sqrt{\frac{p(1-p)}{N}}\right)\right] \quad to \quad \left[p + \left(1.96\sqrt{\frac{p(1-p)}{N}}\right)\right]$$

where

$$p = \frac{\#\ \text{of "}successes\text{"}}{\#\ of\ \ experiments} = \frac{S}{N}$$

N is the number of experiments (or subjects), and S is the number of those experiments or subjects with a particular outcome (termed "success"). This means the remaining N-S subjects or experiments have the alternative outcome. Expressed as a fraction, success

occurred in S/N of the experiments (or subjects). We'll define that proportion p. This use of the variable "p" is completely distinct from p values.

The Wald approximation (above) is known to work well only with large N and proportions not too close to 0.0 or 1.0. That is why most programs, including Prism, use the "exact" method of Clopper and Pearson. However, computer simulations by several investigators demonstrate that the so-called exact confidence intervals are also approximations. 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 (The American Statistician, 52:119-126, 1998) 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 as:

$$\left[p' - \left(1.96 \sqrt{\frac{p'(1-p')}{N+4}} \right) \right] \quad to \quad \left[p' + \left(1.96 \sqrt{\frac{p'(1-p')}{N+4}} \right) \right]$$

where

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

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 and z2, 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 at www.graphpad.com calculates the confidence interval of a proportion using the modified Wald method.

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 (J. Am. Med. Assoc., 249: 1743-1745, 1983) 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.

13. Contingency Tables

Introduction to contingency tables

Contingency tables summarize results where you compare two or more groups when 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 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.

Entering data into contingency tables

You must enter data in the form of a contingency table. Prism cannot cross-tabulate raw data to create a contingency table. Prism also cannot compare proportions directly. You need to enter the actual number of subjects in each category – you cannot enter fractions or percentages.

To create a data table for entering a contingency table, choose a graph on the Welcome (or New Table) dialog from the **Two grouping variable** tab and check the option for no error bars.

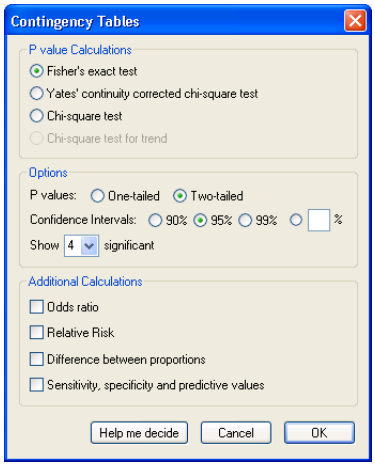
Each column represents an outcome, and each row represents a group. Each subject must contribute to one cell only – the groups and categories are mutually exclusive. In each cell, enter the number of subjects actually observed. Tables of averages, percentages or rates are not contingency tables, and cannot be analyzed by chi-square and related tests.

Most contingency tables have two rows (two groups) and two columns (two possible outcomes). 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.

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.

Choosing how to analyze a contingency table

To analyze a contingency table, start from your data table (see *Entering data into contingency tables* on page 100). Press **Analyze** and choose to do a built-in analysis. Then choose **Contingency table** analysis from the list of statistical tests.



Tables with exactly two rows and two columns

Prism offers two methods for calculating a P value from tables with two rows and two columns: **Fisher's exact test** and the **chi-square test**. We recommend always picking Fisher's test, as it calculates a P value that is exactly correct. The only advantage of

the chi-square test is that it is easier to calculate by hand and so is better known. We don't recommend using it to analyze contingency tables with two rows and two columns.

If you choose a chi-square test, also choose whether to apply Yates' continuity correction. This correction is designed to make the approximate results from a chi-square test more accurate with small samples. Statisticians disagree about whether to use it. If you always select Fisher's exact test (recommended), Yates' correction is of no concern.

If your table includes very large numbers (thousands), Prism will automatically perform the chi-square test even if you select Fisher's test. This is because the Fisher's test calculations are slow with large samples. With large samples, the chi-square test is very accurate and Yates' continuity correction has negligible effect.

Choose a two-sided P value, unless you have a good reason to pick a one-sided P value. (With contingency tables, Prism refers to "two-sided" P values rather than "two-tail P value").

Prism can summarize data on a two by two table with the relative risk, difference between proportions (P1-P2), and/or the odds ratio. If your data came from an experimental, cross-sectional or prospective study, summarize the data by choosing the relative risk or P1-P2. If your data came from a retrospective case-control study, pick only the odds ratio.

Table with more than two rows or two 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.

If your table has two columns and three or more rows, you may also select 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.

Interpreting analyses of contingency tables

How analyses of 2x2 contingency tables work

If your table has two rows and two columns, Prism computes relative risk, odds ratio and P1-P2 using the equations below:

	Outcome 1	**Outcome 2**
Group 1	A	B
Group 2	C	D

$$\text{Relative Risk} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}}$$

$$P1 - P2 = \frac{A}{A+B} - \frac{C}{C+D}$$

$$\text{Odds Ratio} = \frac{A/B}{C/D}$$

If any of the four values in the contingency table are zero, Prism adds 0.5 to all values before calculating the relative risk, odds ratio and P1-P2 (to avoid dividing by zero).

The word “risk” is appropriate when the first row is the exposed or treated group and the left column is the bad outcome. With other kinds of data, the term “risk” isn't appropriate, but you may still be interested in the ratio of proportions. Prism calculates the 95% confidence interval for the relative risk using the approximation of Katz. You can be 95% certain that this range includes the true relative risk.

If your data are from a case-control retrospective study, neither the relative risk nor P1-P2 is meaningful. Instead, Prism calculates an odds ratio and the confidence interval of the odds ratio using the approximation of Woolf. If the disease is rare, you can think of an odds ratio as an approximation of the relative risk.

Prism computes the P value using either the chi-square test or Fisher's exact test.

How analyses of larger contingency tables work

If your table has two columns and more than two rows (or two rows and more than two columns), Prism will perform both the chi-square test for independence and the chi-square test for trend.

The ***chi-square test for independence*** asks whether there is an association between the variable that defines the rows and the variable that defines the columns.

Prism first computes the expected values for each value. These expected values are calculated from the row and column totals, and are not displayed in the results. The discrepancies between the observed values and expected values are then pooled to compute chi-square, which is reported. A large value of chi-square tells you that there is a large discrepancy. The P value answers this question: If there is really no association between the variable that defines the rows and the variable that defines the columns, then what is the chance that random sampling would result in a chi-square value as large (or larger) as you obtained in this experiment?

The P value from the *chi-square test for trend* answers this question: If there is no linear trend between row (column) number and the fraction of subjects in the left column (top row), what is the chance that you would happen to observe such a strong trend as a consequence of random sampling? If the P value is small, you will conclude that there is a statistically significant trend.

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.

How to think about the relative risk, odds ratio and P1-P2

To understand the differences between the relative risk, odds ratio and P1-P2 consider this example. There are two groups of subjects, denoted by two rows. There are two outcomes denoted by columns:

	X Labels	A	B
	X Labels	Progress	No Progress
	X	Y	Y
1	AZT	76	399
2	Placebo	129	332

Method	Description
Difference between proportions	In the example, disease progressed in 28% of the placebo-treated patients and in 16% of the AZT-treated subjects. The difference is 28% - 16% = 12%.
Relative risk	The ratio 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.

How to think about sensitivity, specificity, and predictive values

Prism assesses the accuracy of a clinical test 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.

How to think about P values from a 2x2 contingency table

The P value 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? Equivalently, if there really is no association between rows and columns overall, what is the chance that random sampling would lead to a relative risk or odds ratio as far (or further) from 1.0 (or P1-P2 as far from 0.0) as observed in this experiment?

"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 relative risk, odds ratio or P1-P2 you are looking for. How large does the value need to be for you consider it to be scientifically important? How small a value 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.

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

If the P value is small (2x2 contingency table)

If the P value is small, then it is unlikely that the association you observed is due to random sampling. You can reject the idea that the association is a coincidence, and conclude instead that the population has a relative risk or odds ratio different than 1.0 (or P1-P2 different than zero). The association is statistically significant. But is it scientifically important? The confidence interval helps you decide.

Your data include the effects of random sampling, so the true relative risk (or odds ratio or P1-P2) is probably not the same as the value calculated from the data in this experiment. There is no way to know what that true value is. Prism presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true relative risk, odds ratio or P1-P2.

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent values that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial	Trivial	Although the true relative risk or odds ratio is not 1.0 (and the true P1-P2 is not 0.0), the association is tiny and uninteresting. The variables defining the rows is associated with the variable defining the columns, but weakly.
Trivial	Important	Since the confidence interval ranges from a relative risk (or odds ratio or P1-P2) that you think is biologically trivial to one you think would be important, you can't reach a strong conclusion from your data. You can conclude that the rows and columns are associated, but you don't know whether the association is scientifically trivial or important. You'll need more data to obtain a clear conclusion.
Important	Important	Since even the low end of the confidence interval represents an association large enough to be considered biologically important, you can conclude that the rows and columns are associated, and the association is strong enough to be scientifically relevant.

If the P value is large (2x2 contingency table)

If the P value is large, the data do not give you any reason to conclude that the relative risk or odds ratio differs from 1.0 (or P1-P2 differs from 0.0). This is not the same as saying

that the true relative risk or odds ratio equals 1.0 (or P1-P2 equals 0.0). You just don't have evidence that they differ.

How large could the true relative risk really be? Your data include the effects of random sampling, so the true relative risk (or odds ratio or P1-P2) is probably not the same as the value calculated from the data in this experiment. There is no way to know what that true value is. Prism presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true relative risk (or odds ratio or P1-P2). When the P value is larger than 0.05, the 95% confidence interval includes the null hypothesis (relative risk or odds ratio equal to 1.0 or P1-P2 equal to zero) and extends from a negative association (RR<1.0, OR<1.0, or P1-P2<0.0) to a positive association (RR>1.0, OR>1.0, or P1-P2>0.0)

To interpret the results in a scientific context, look at both ends of the confidence interval and ask whether they represent an association that would be scientifically important or scientifically trivial.

Lower confidence limit	Upper confidence limit	Conclusion
Trivial	Trivial	You can reach a crisp conclusion. Either there is no association between rows and columns, or it is trivial. At most, the true association between rows and columns is tiny and uninteresting.
Trivial	Important	You can't reach a strong conclusion. The data are consistent with the treatment causing a trivial negative association, no association, or a large positive association. To reach a clear conclusion, you need to repeat the experiment with more subjects.
Important	Trivial	You can't reach a strong conclusion. The data are consistent with a trivial positive association, no association, or a large negative association. You can't make a clear conclusion without repeating the experiment with more subjects.
Important	Important	You can't reach any conclusion at all. You need more data.

Checklist: Are contingency table analyses appropriate for your data?

Before interpreting the results of any statistical test, first think carefully about whether you have chosen an appropriate test. Before accepting results from a chi-square or Fisher's test, ask yourself these questions:

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.

14. Survival Curves

Introduction to survival curves

Survival curves plot the results of experiments where the outcome is time until death (or some other event). Usually you wish to compare the survival of two or more groups.

Prism creates survival curves, using the product limit method of Kaplan and Meier, and compares survival curves using the logrank test.

The name *survival curve* is a bit restrictive. The methods described in this chapter can analyze any kind of experiment where the result is expressed as the time to a well-defined end point. Instead of death, the endpoint could be occlusion of a vascular graft, first metastasis, 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.

The end point must be an event that can only occur one time per subject. Recurring events should not be analyzed with survival curves.

Some kinds of survival data are better analyzed with nonlinear regression. For example, don't use the methods in this chapter 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*.

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. Prism uses the method of Kaplan and Meier to create survival curves while accounting for censored data.

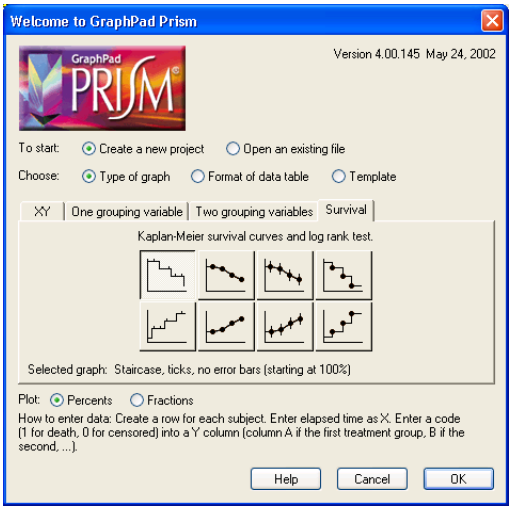
> Note. The term "censored" seems to imply that the subject did something inappropriate. But that isn't the case. The term "censored" simply means that you don't know, or can't use, survival beyond a certain point.

Entering survival data and creating a survival curve

Prism makes it very easy to create survival curves.

> Note to users of previous Prism versions. Prism 4 introduces a new, much easier, way to create survival curves.

When you create a new data table from the Welcome or New Table dialog, choose the fourth tab **Survival**. Then choose the basic look of your survival graph, which you can customize later.



Prism creates a data table formatted to enter X values as numbers, and Y values with no subcolumns. 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 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.
- ✓ 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 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.

10	46.0	1.0	
11	64.0	0.0	
12	78.0	0.0	
13	124.0	1.0	
14	127.0	0.0	
15	9.0		1.0
16	26.0		0.0
17	43.0		1.0
18	64.0		1.0

- ✓ 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.

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).

Examples of entering survival data

Example of survival data from an animal study

The first 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.

	X Values	A	B
	Days	Control	Treated
	X	Y	Y
1	28	0	
2	28	0	
3	28	0	
4	28	0	
5	28	0	
6	15		1
7	21		1
8	26		1
9	28		0
10	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.

Example of survival data from a clinical study

Here is a portion of the data collected in a clinical trial:

Enrolled	Final date	What happened	Group
07-Feb-98	02-Mar-02	Died	Treated
19-May-98	30-Nov-98	Died	Treated
14-Nov-98	03-Apr-02	Died	Treated
01-Dec-98	04-Mar-01	Died	Control
04-Mar-99	04-May-01	Died	Control

Enrolled	Final date	What happened	Group
01-Apr-99	09-Sep-02	Still alive, study ended	Treated
01-Jun-99	03-Jun-01	Moved, off protocol	Control
03-Jul-99	09-Sep-02	Still alive, study ended	Control
03-Jan-00	09-Sep-02	Still alive, study ended	Control
04-Mar-00	05-Feb-02	Died in car crash	Treated

And here is how these data looked when entered in Prism.

	X Values	A	B
	Days	Control	Treated
	X	Y	Y
1	1484		1
2	195		1
3	1236		1
4	824	1	
5	792	1	
6	1257		0
7	733	0	
8	1164	0	
9	980	0	
10	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.

Common questions about 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 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 patient died right after enrollment, should I enter the patient with X=0?

No. The time must exceed zero for all subjects.

Choosing a survival analysis

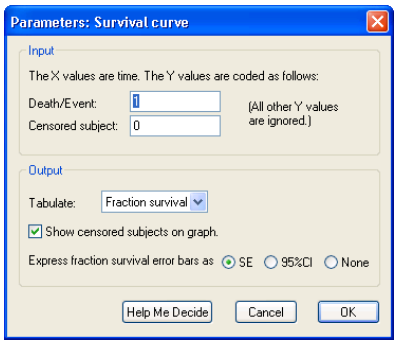
Automatic survival analysis

Survival analysis works differently than other analyses in Prism (and different than survival analysis in previous versions). When you create a new data table from the Welcome or New Table dialog, you can choose what kind of graph you want. If you choose a survival graph, Prism automatically analyzes your data and plots the survival curve. A survival analysis results sheet will appear in your Results folder. You don't need to click the Analyze button.

> Note: With automatic survival graphs, Prism assumes that deaths are coded with "1" and censored observations with "0". If you use a different coding scheme, create survival curves manually.

Manual survival analysis

You can manually analyze a table of data, even if you didn't choose a survival graph to begin with. Start from your table of survival data. Press **Analyze**, and choose to do a built-in analysis. Then choose **Survival curve** from the list of statistical analyses.

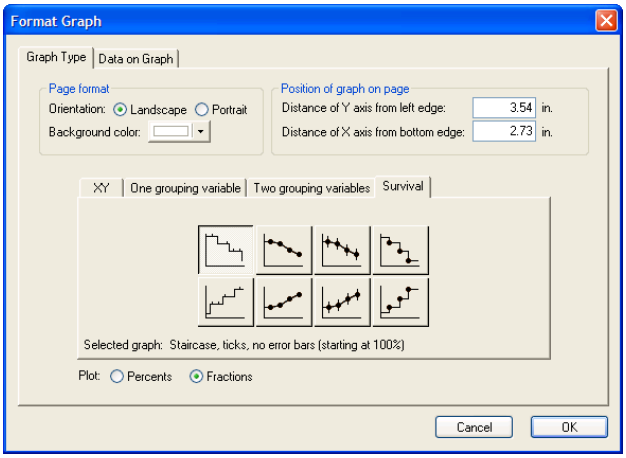


You can usually leave the choices on this dialog set to their default value. It is rare to use a code other than Y=0 for a censored subject and Y=1 for a death. The other choices on this dialog determine the initial look of the survival curve, and you can change these later from the graph (see the next section).

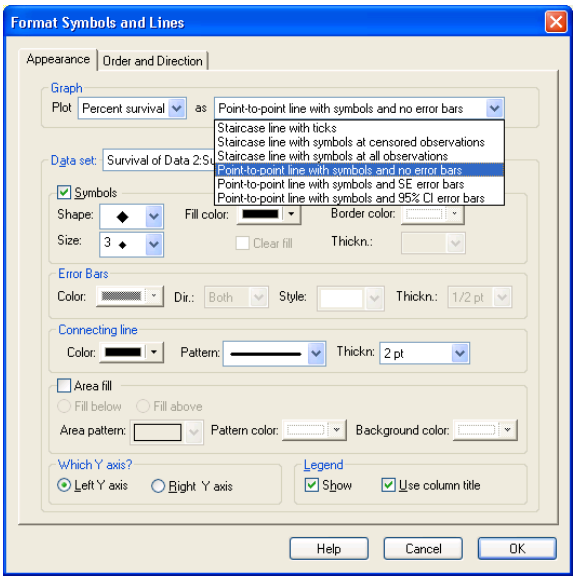
Modifying graphs of survival curves

Prism offers you two places to change the look of a survival graph.

To change the look of the graph entirely, click Change... graph type and choose from the Format Graph dialog.



To make smaller changes to the graph, double click on one of the symbols or choose Change Symbols and Lines. The top of that dialog offers the same choices as Format Graph, but as menu choices. The rest of the dialog lets you fine-tune the graph one data set at a time.



> Note: Prism offers three places (Analysis parameters, Format Graph, and Format Symbols & lines) to choose whether you want to tabulate and graph fraction death, fraction survival, percent death, or percent survival. If you make a change in any of these dialogs, it will also be made in the other. You cannot choose one format for tabulating and another for graphing (unless you repeat the analysis).

Results of survival analysis

The fraction (or percent) survival at each time

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.

Number of subjects at risk

Prism tabulates the number of patients still at risk at each time. The number of subjects still at risk decreases over time as subjects die or are 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”.

Curve comparison

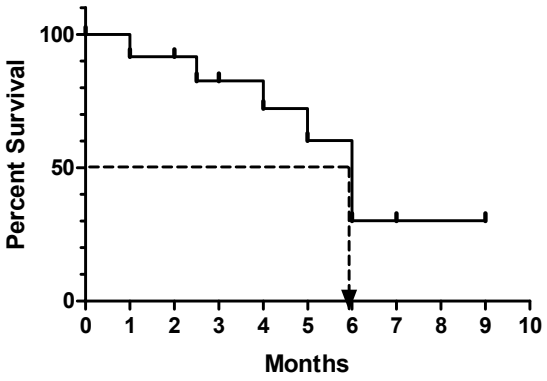
Prism compares survival curves using the logrank test. If you entered exactly two data sets, the logrank test is equivalent to the Mantel-Haenszel test. This test generates a P value testing the null hypothesis that the survival curves are identical in the overall populations. In other words, the null hypothesis is that the treatments 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-tailed P values. If you wish to report a one-tailed P value, you must have predicted which group would have the longer median survival before collecting any data. 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 is greater than 0.50, and you must conclude that the difference was due to chance, no matter how large it is.

If you entered three or more data sets, Prism also 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 number and median survival.

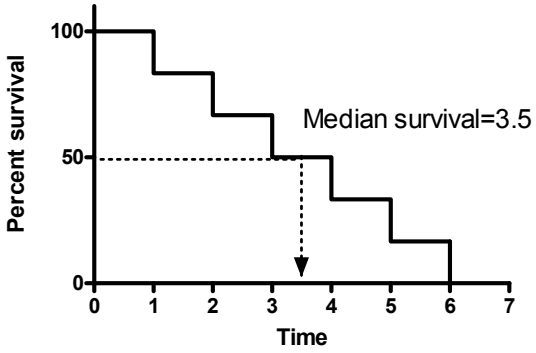
Median survival

The median survival is the time at which half the subjects have died. The example survival curve below shows 50% survival at 6 months, so median survival is 6 months.



If survival exceeds 50% at the longest time points, then median survival cannot be computed.

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.

Prism computes an approximate 95% confidence interval for the ratio of median survivals. 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. In other words, it assumes that the survival of patients or animals in your study follows the same model as radioactive decay. 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.

> Tip: Don't focus on the 95% CI of the ratio of median survivals unless the survival curves follow the same shape as an exponential decay.

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.

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.

Reference for survival calculations

Most of the calculations are quite standard. Prism follows the calculations as spelled out in detail in DG Altman, *Practical Statistics for Medical Research*, 1991, Chapman and Hall.

For calculating the logrank test, Prism uses the second of the two methods described in that reference (more difficult to calculate, but more accurate). Some books call this method the Mantel-Haenszel logrank test.

Checklist for interpreting survival analyses

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's 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". With an extra ten minutes preparation 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 is unrelated to survival?

The survival analysis only is valid when the survival times of censored patients are identical 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 didn't take 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.

Part D: Specialized Data

15. Comparing Methods with a Bland-Altman Plot

Introducing Bland-Altman plots

Bland and Altman devised a simple, but informative, way of graphing the comparison of two assay methods: First measure multiple samples using both methods. Then plot the difference between the two measurements as a function of the average of the two measurements of each sample (your best estimate of the true value). The resulting graph is called a *Bland-Altman plot*.

Creating a Bland Altman plot

To create a Bland-Altman plot, follow these steps:

- ✓ Create a new table. Choose the one-grouping variable tab and a before-after graph.
- ✓ 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.
- ✓ Click Analyze and then choose Bland-Altman from the list of Clinical lab analyses.
- ✓ Designate the columns with the data (usually A and B).
- ✓ Choose 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.

Bland-Altman results

The first page shows the difference and average values used to create the plot.

The second results page shows the 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 from this equation:

$$Bias - 1.96 \cdot SD \quad \text{to} \quad Bias + 1.96 \cdot SD$$

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.

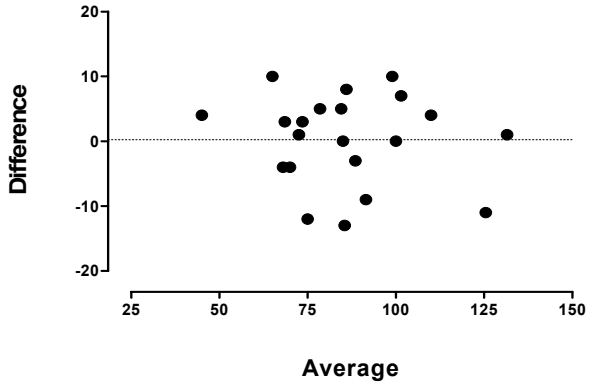
Example of Bland-Altman plot

Here is a cardiovascular example from Altman (1991), *Practical Statistics for Medical Researchers* (page 398) that will help to demonstrate the utility of the Bland-Altman plot.

To evaluate mitral valve disease, investigators measured blood flow through the mitral valve (MF), as well as left ventricular stroke volume (SV), in 21 normal subjects (shown here) as well as patients with mitral valve disease (not shown). In the absence of mitral valve disease (as is the case here), the two hemodynamic measurements should be equal. The investigator's data are shown here (in cm^3). The order of the rows does not affect the calculations or graphing, and in this example the data were sorted by the MF values in column A.

	A	B
	MF	SV
	Y	Y
1	47.0	43.0
2	66.0	70.0
3	68.0	72.0
4	69.0	81.0
5	70.0	60.0
6	70.0	67.0
7	73.0	72.0
8	75.0	72.0
9	79.0	92.0
10	81.0	76.0
11	85.0	85.0
12	87.0	82.0
13	87.0	90.0
14	87.0	96.0
15	90.0	82.0
16	100.0	100.0
17	104.0	94.0
18	105.0	98.0
19	112.0	108.0
20	120.0	131.0
21	132.0	131.0

The Bland-Altman plot graphs the average of the two values on each row on the X axis, and the difference between the measurements (A-B) on the Y axis.



Prism automatically graphs these Bland-Altman results. We modified this graph a bit using the Format Axes dialog box:

- ✓ Set the origin to be the lower left.
- ✓ Create a custom tick shown as a dotted line at the bias (Y=0.238, in this example).
- ✓ Offset the X and Y axes so that they do not touch

The bias is reported by Prism as:

Bias	0.24
SD of bias	6.96
95% Limit of agreement	
From	13.89
To	-13.41

As expected (among controls) the two methods had very similar results on average, and the bias (difference between the means) is only 0.24. In 95% of subjects the difference lies between -13.9 and +13.4.

The authors of this study used these results simply as a control, and then went on to investigate patients with mitral disease (not shown here).

Interpreting a Bland-Altman plot

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?

Checklist for interpreting 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.

16. Receiver-operator Curves

Introduction to 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 us to better 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.

Entering ROC data

- ✓ From the Welcome or New table dialog, select One grouping variable and then choose a scatter plot.
- ✓ Enter in one column (usually A) the diagnostic test results for a group of individuals who are known to not have that disease (controls).
- ✓ Enter in another column (usually B) the diagnostic test results for a group of individuals who are known to have the disease you are testing for (patients).

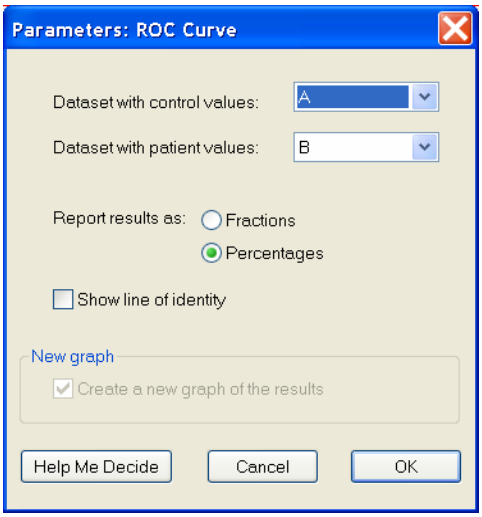
	A	B
	Controls	Patients
	Y	Y
1	97.9	112.7
2	94.9	104.0
3	98.6	126.7
4	77.3	123.3
5	97.9	120.5
6	99.7	130.3
7	83.0	129.6
8	102.5	140.2
9	104.5	119.7
10	108.9	139.9

The two groups are not paired in any way; therefore, the order in which you enter the data in the rows is arbitrary. The two groups may have different numbers of subjects.

Creating a ROC curve with Prism

From your data table, follow these steps to make a ROC curve.

1. Click the Analyze button and then choose Receiver-operator characteristic curve from the list of analyses for clinical labs.
2. In the ROC dialog box, designate the columns with the control and patient results (columns A and B, respectively, will be set as the default option).
3. Choose to see the results (sensitivity and 1-specificity) expressed as Fractions or Percentages.



4. Check the option to create a new graph.

Results of a ROC curve

Sensitivity and specificity

The calculation of sensitivity and specificity depends which value you use to separate normal from abnormal (the cutoff value). Unless your test discriminates perfectly (in which case an ROC curve isn't necessary), there will be a trade-off between high sensitivity and high specificity.

- ✓ If you make the threshold high, you increase the specificity of the test, but lose sensitivity.
- ✓ If you make the threshold low, you increase the test's sensitivity.

> Remember: Sensitivity is the fraction of people with the disease that the test correctly identifies as positive. Specificity is the fraction of people without the disease that the test correctly identifies as negative.

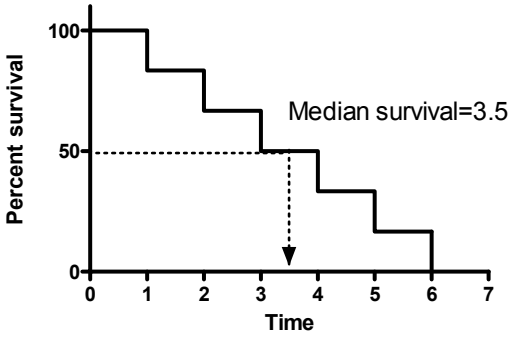
Prism calculates the sensitivity and specificity using each value in the data table as the cutoff value. This means that it calculates many pairs of sensitivity and specificity.

Prism tabulates sensitivity and 1-specifity, with 95% confidence intervals, for all possible cutoff values.

ROC graph

An ROC curve may help in analyzing this trade-off between sensitivity and specificity. Prism creates the ROC graph automatically. You'll only need to spend a few moments polishing it.

> Note: This graph is created from a data table, of which only the top portion was shown earlier.



When choosing which cutoff value you will use, don't think only about the tradeoff of sensitivity vs. specificity. Also consider the *clinical consequences* of false positive and false negative results. The ROC curve can't help with that.

Area under a ROC curve

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.

> Note: 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.

> Note: 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.

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.

In the example above, the area is 0.946 with a 95% confidence interval extending from 0.8996 to 0.9938. This means that a randomly selected patient has a 94.6% chance of having a higher test result than a randomly selected control.

Prism completes your ROC curve evaluation by reporting a P value and testing the null hypothesis that the area under the curve really equals 0.50. In other words, the null hypothesis is that the test diagnoses disease no better than flipping a coin. 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.

Comparing ROC curves

Prism does not compare ROC curves. It is, however, quite easy to 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:

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

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-tailed 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.

Checklist for 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?

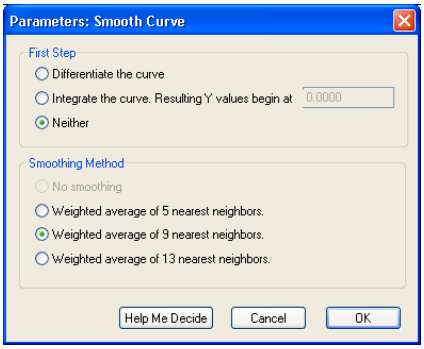
If some people are in the wrong group, the ROC curve won't be accurate. The method used to discriminate between patient and control must truly be a gold standard.

17. Smoothing, Differentiating and Integrating Curves

How to smooth, differentiate, or integrate a curve

Prism can compute numerical integrals or derivatives, and smooth the results. It can also smooth a curve directly, without computing its derivative or integral.

From a graph of the curve or a data table, click **Analyze** and choose to do a built-in analysis. Then select **Smooth, differentiate or integrate curve** to bring up this dialog.



Prism can only smooth data sets (and compute derivatives and integrals) where the X values are equally spaced. The X values in the table may be formatted either as individual numbers or as a sequence (you define the first value and the interval, and Prism fills in the rest).

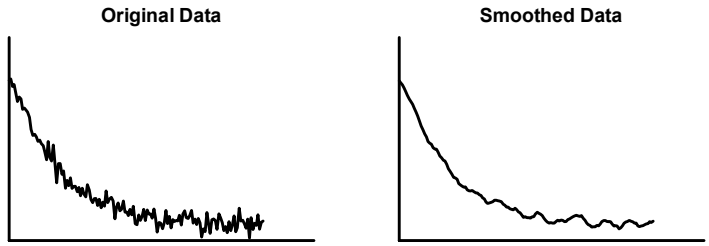
Smoothing a curve

If you import a curve from an instrument, smooth the data to improve the appearance of a graph. The purpose of smoothing is solely 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.

There is no point to smoothing curves created by nonlinear regression, since they are already smooth. It only makes sense to smooth curves collected from an instrument. It can make sense to compute the derivative or integral of a perfect (smooth) when X is time.

> Note: Smoothing is not a method of curve fitting.

Smoothing uses the method of Savitsky and Golay (Analytical Chemistry, 36:1627-1639, 1964) using a cubic equation. Each point in the curve is replaced by the weighted average of its nearest 5, 9, or 13 neighbors. The results table has a few less rows than the original data.



The derivative of a curve

The 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. Prism first calculates a simple numerical derivative. For every row i, the resulting X and Y values are:

$$Result(i) = \frac{X(i+1) + X(i)}{2}$$

$$ResultY(i) = \frac{\Delta Y}{\Delta X} = \frac{Y(i+1) - Y(i)}{X(i+1) - X(i)}$$

After calculating the numerical derivative, Prism can smooth the results, if you choose.

Prism does not do any symbolic algebra, and it cannot compute analytical derivatives. If you give Prism a series of XY points that define a curve, it can compute the derivative of that curve. If you give Prism an equation, it cannot compute a new equation that defines the derivative.

The integral of a 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.

Prism uses the trapezoid rule 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. Prism uses this equation (i refers to row number):

$$ResultY(i) = ResultY(i-1) + \frac{Y(i-1) + Y(i)}{2} \cdot \left[X(i) - X(i-1) \right]$$

After doing a simple integration, Prism can then smooth the results, if you choose.

Prism does not do any symbolic calculus and cannot compute analytical integrals.

18. Area Under the Curve

Usefulness of measuring the 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.

Before continuing, make sure you know the difference between integrating a curve and computing the area under the curve. When you integrate a curve, the result is another curve showing cumulative area. When you ask Prism to compute the area under the curve, it gives you one value for the area under the entire curve, as well as the area under well-defined peaks.

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.

Calculating area under curve using Prism

1. Start from a data table or graph.
2. Click Analyze and then choose Built-in analysis.
3. Select the category Curves and Regression, and then select Area under a curve.
4. Define the baseline by entering a Y value (usually Y=0); or, calculate the baseline as the mean of the first and last few values (you may choose how many values to include). If you enter 1 to compute the "mean of the first 1 rows" that means that row 1 defines the baseline.
5. Define the *minimum height of a peak* that you consider worth finding. Enter the height as measured in the units of the Y axis or enter the height as a percentage of the distance between the minimum and maximum Y values. If your data are "noisy", Prism may find too many peaks, some of which may be tiny and/or irrelevant. By defining your minimum height appropriately, you will avoid finding too many peaks in noisy data.
6. Define the *minimum width of a region* worth considering. Enter the width as the number of adjacent rows in the data table. Do not use the units of the X axis.
7. Choose whether values below the baseline should be treated as negative peaks or ignored.

Peaks that have negative Y values are called negative peaks. However, the area of these peaks is positive. So if you choose to include negative peaks, the total area will be larger.

> Note: If all values are above the baseline, then the preceding dialog choices (except for definition of the baseline) are irrelevant. Prism will find one peak, and report the area under the entire curve. Single-peak assessment is very useful in analysis of pharmacokinetic data.

Interpreting area under the curve

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.

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.

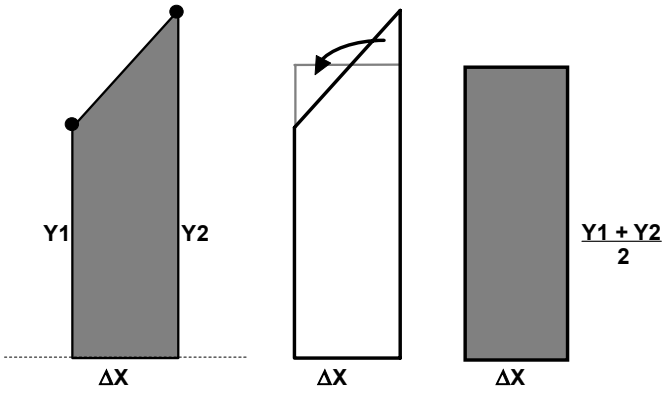
Prism may identify more regions than you are interested in. Simply go back to the *Parameters* dialog box and enter a larger value for the minimum width of a region and/or the minimum height of a peak.

Note the limitations of this Prism analysis:

- ✓ 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 calculates the area under a curve

Prism computes the area under the curve using the *trapezoid rule*, illustrated in the following figure.



Prism defines a curve as a series of connected X,Y points, with equally spaced X values. The left portion of the figure shows two of these points and the baseline (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 under the curve. Since the two triangles in the middle panel have the same area, the area of the trapezoid on the left (which we want to find out) is the same as the area of the rectangle on the right (which is easier to calculate).

The area under the curve, therefore, is calculated as:

$$\frac{\Delta X\left(Y1+Y2\right)}{2}$$

Prism repeatedly uses this formula for each adjacent pair of points defining the curve.

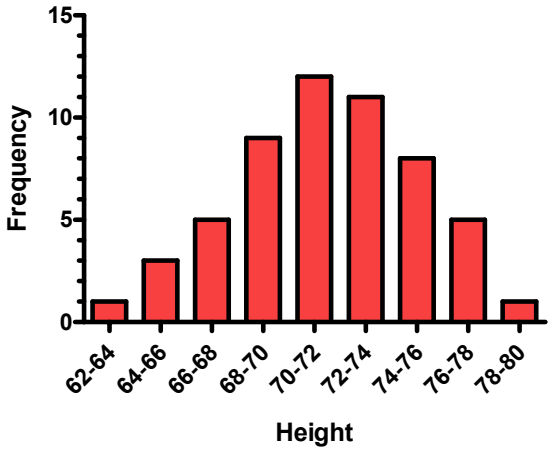
> Note: When Prism creates a curve for you, it generally defines the curve as 150 line segments. You can increase or decrease this in the parameter dialog for the analysis that created the curve.

19. Frequency Distributions

What is a frequency distribution?

A frequency distribution shows the distribution of Y values in each data set. The range of Y values is divided into bins, and Prism determines how many values fall into each bin.

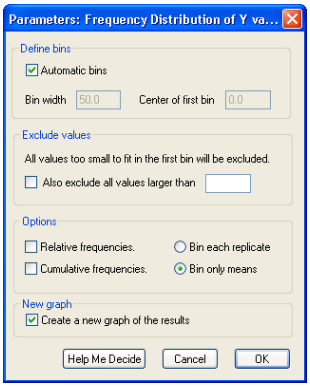
The term *histogram* is usually used (at least in the United States) to mean a graph of a frequency distribution. But this term is not used consistently, and some people use the term *histogram* to refer to any bar graph.



Creating a frequency distribution table with Prism

If you have already tabulated your data into a frequency distribution table, you can enter that directly and create a bar graph. In that case, there is no need to use any analysis.

If you have a column of raw data, Prism can create the frequency distribution table for you. Click **Analyze** and choose to do a built-in analysis. Then choose **Frequency distribution** from the list of statistical analyses to bring up the **Parameters** dialog.



Histograms generally look best when the bin width is a round number and there are 10-20 bins. Either use the default settings, or clear the **Automatic bins** option to enter the center of the first bin and the width of each.

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.

Select **Relative frequencies** to determine the fraction of values in each bin, rather than the number of values in each bin. For example, if you have 50 data points of which 15 fall into the third bin, the results for the third bin will be 0.30 (15/50) rather than 15.

Select **Cumulative frequencies** to see the 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.

Graphing frequency distribution histograms with Prism

We explain in detail how to graph frequency distribution histograms in the book of step by step examples.

Part E: Preprocessing Data

20. Transforming Data

Choosing a transform with Prism

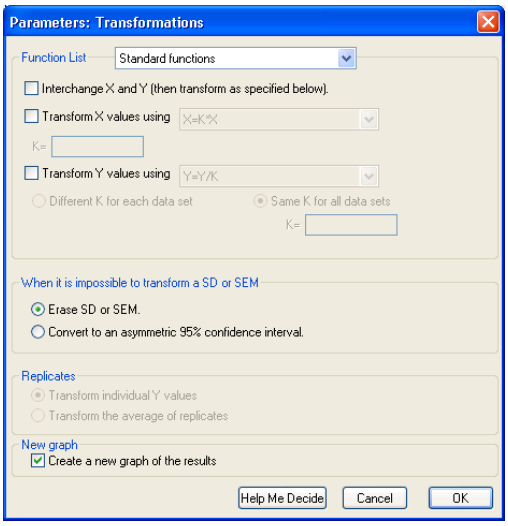
You'll often find it useful to transform your data before doing other analyses or making graphs. With Prism, transformations (and other kinds of data manipulations) are a kind of analysis. Prism never alters the data you enter, but instead creates a results table with the transformed values. You can then graph these results or analyze them further.

> Tip: Beware of using transforms to convert data that follows a curved relationship into a linear relationship. This can be useful for displaying data, but should usually be avoided when analyzing data.

Prism offers other kinds of data manipulations in addition to transforms. See *Subtracting (or dividing by) baseline values* on page 142, and *Normalizing data* on page 143.

To transform data with Prism:

1. Start from a data table, or a results table.
2. Click Analyze and choose Transform from the list of data manipulations.



3. Choose a built-in transform, a pharmacology/biochemistry transform, or enter your own.

4. Check the option box “Create new graph” to create a new graph of the processed data. The default selection (whether or not to create a new graph) is set in the Analysis options dialog (Tools menu).

Your results will appear in a new results table.

Interchanging X and Y

When you choose a standard function, you can choose to interchange X and Y values and also choose transforms of X or Y or both.

Some notes on interchanging X and Y values:

- ✓ Prism can interchange data on tables with more than one data set (more than one Y column), even though the results sheet has only a single X column. It does this by creating additional rows. The results will be staggered down the page with only one data set in any particular row.
- ✓ If you entered replicate Y values (or mean with SD or SEM) Prism interchanges X and Y by putting the mean Y value into the X column. Information about the scatter of Y is ignored.
- ✓ If you selected X or Y transforms (in addition to interchanging), Prism applies the transform to the data after interchanging X and Y. This means that the X transform is applied to data that were originally in the Y column, and the Y transform is applied to data originally in the X column.

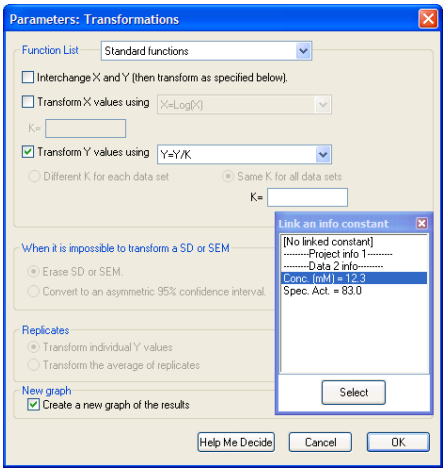
Standard functions

Choose from one of these functions for transforming Y values (analogous functions are available for X):

Function	Comments
Y= Y * K	Enter K in the box provided.
Y = Y + K	"
Y = Y – K	"
Y = Y / K	"
Y = Y squared	
Y = Y ^ K	Enter K in the box provided.
Y = log(Y)	Log base 10
Y = -1*log(Y)	
Y = ln(Y)	Natural logarithm (base e)
Y = 10 ^ Y	Ten to the Yth power (inverse of log).
Y = exp(Y)	eY (inverse of ln)
Y = 1/Y	
Y = sqrt(Y)	Square root.
Y = logit(y)	ln(Y/1-Y)
Y = probit(Y)	Y must be between 0.0 and 1.0
Y = rank(Y)	Column rank. Smallest Y value gets rank of 1.
Y = zscore(Y)	Number of standard deviations from the col. mean.
Y = sin(Y)	Y is in radians.
Y = cos(Y)	"
Y = tan(Y)	"
Y = arcsin(Y)	Result is in radians.
Y = ABS(Y)	Absolute value. If Y is negative, multiply by -1.
Y = Y + Random	Gaussian. Mean=0. SD=K (you enter).
Y = X / Y	
Y = Y / X	
Y = Y – X	
Y = Y + X	
Y = Y * X	
Y = X – Y	

Many of the functions include the variable "K". Enter a value for K on the dialog. When transforming Y values, you can enter one value of K for all data sets or a separate value of K for *each* data set. To enter different K values for each data set, choose a data set, enter K, choose another data set, enter its K, and so on.

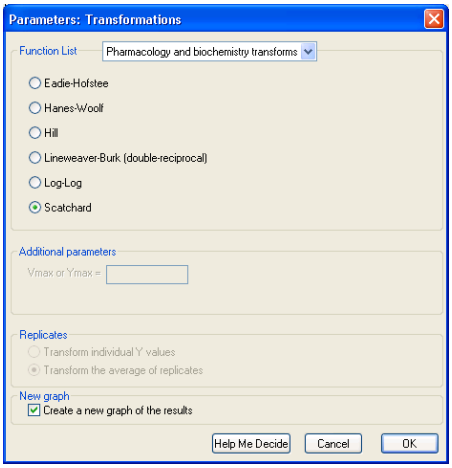
Rather than enter the value of K on this table, you can choose a value from a linked info table. If you have created a linked info table and it contains numbers, Prism will popup a list of values you can choose from.



Use this option to transform different data sets with different values of K. If you don't want to transform some data sets at all, don't select them for the analysis. You can choose which data sets to analyze right on the Analyze dialog.

Special functions for pharmacology and biochemistry

On top of the transform dialog choose *Pharmacology and biochemistry functions* from the drop down and choose a transform by name. Eadie-Hofstee, Hanes-Woolf, and Lineweaver-Burk transforms are used to plot enzyme-kinetic results. Scatchard transforms are used to display radioligand binding and Hill plots are used to plot dose-response data.



These transforms are very useful as a way to display data. They are much less useful as a method to analyze data. You'll get better results by using nonlinear regression on the actual data. See the companion book on nonlinear curve fitting for details.

Here is the mathematical definition of each transform:

Function	X becomes	Y becomes
Eadie-Hofstee	Y/X	No change

Hanes-Woolf	No change	X/Y
Hill	No change if you entered your data as log(conc). Log10(X) if you entered your data as concentration.	log10(Y/(Ymax-Y)) (Prism prompts for Ymax)
Lineweaver-Burk	1/X	1/Y
Log-log	Log10(X)	Log10(Y)
Scatchard	Y	Y/X

> Tip. Prism can also create Bland-Altman plots, which require a simple transform of the data. However, this is not done via a transform, but rather via a separate analysis. See Comparing Methods with a Bland-Altman Plot on page 118.

User-defined transforms

At the top of the Parameters dialog for transforms, switch between built-in transforms and user-defined transforms of X or Y. Select a user-defined transform you have used before or enter a new one.

You can write equations so different data sets get different transforms. Put <B> in front of a line in your transform that only applies to data set B. Put <~A> in front of a line that applies to all data sets except data set A.

If you are transforming X values, you may use Y in the function. If the data table contains several data sets (so has several Y values for a single X value), Prism will stagger the results down the page, repeating X values as needed. The results for column A will appear on top of the results table. Below that Prism will place the results for column B. For these rows, column A will be empty.

> Note to users of previous Prism versions. Prism 4 is much more flexible than earlier versions, which did not let you use X in Y transforms or Y in X transforms.

Available functions for user-defined transformations

When you enter your transforms, you can use any of the functions listed below.

Function	Explanation	Excel equivalent
abs(k)	Absolute value. If k is negative, multiply by -1.	abs(k)
arccos(k)	Arccosine. Result is in radians.	acos(k)
arccosh(k)	Hyperbolic arc cosine.	acosh(k)
arcsin(k)	Arcsine. Result is in radians.	asin(k)
arcsinh(k)	Hyperbolic arcsin. Result in radians.	asinh(k)
arctan(k)	Arctangent. Result is in radians.	atan(k)
arctanh(k)	Hyperbolic tangent. K is in radians.	atanh(k)

Function	Explanation	Excel equivalent
artctan2(x,y)	Arctangent of y/x. Result is in radians.	atan2(x,y)
besselj(n,x)	Integer Order J Bessel, N=0,1,2...	besselj(x,n)
bessely(n,x)	Integer Order Y Bessel, N=0,1,2...	bessely(x,n)
besseli(n,x)	Integer Order I Modified Bessel, N=0,1,2...	besseli(x,n)
besselk(n,x)	Integer Order K Modified Bessel, N=0,1,2...	besselk(x,n)
beta(j,k)	Beta function.	exp(gammaln(j) +gammaln(k) -gammaln(j+k))
binomial(k,n,p)	Binomial. Probability of k or more "successes" in n trials, when each trial has a probability p of "success".	1 - binomdist(k,n,p,true) + binomdist(k,n,p,false)
chidist(x2,v)	P value for chi square equals x2 with v degrees of freedom.	chidist(x2,v)
ceil(k)	Nearest integer not smaller than k. Ceil (2.5)=3.0. Ceil(-2.5)=2.0	(no equivalent)
cos(k)	Cosine. K is in radians.	cos(k)
cosh(k)	Hyperbolic cosine. K is in radians.	cosh(k)
deg(k)	Converts k radians to degrees.	degrees(k)
erf(k)	Error function.	2*normsdist(k*sqrt(2))-1
erfc(k)	Error function, complement.	2-2*normsdist(k*sqrt(2))
exp(k)	e to the kth power.	exp(k)
floor(k)	Next integer below k. Floor(2.5)=2.0. Floor(-2.5)=-3.0.	(no equivalent)
fdist(f,v1,v2)	P value for F distribution with v1 degrees of freedom in the numerator and v2 in the denominator.	fdist(f,v1,v2)
gamma(k)	Gamma function.	exp(gammaln(k))
gammaln(k)	Natural log of gamma function.	gammaln(k)
hypgeometricm (a,b,x)	Hypergeometric M.	(no equivalent)
hypgeometricu(a,b,x)	Hypergeometric U.	(no equivalent)
hypgeometricf(a,b,c,x)	Hypergeometric F.	(no equivalent)
ibeta(j,k,m)	Incomplete beta.	(no equivalent)

Function	Explanation	Excel equivalent
if(condition, j, k)	If the condition is true, then the result is j. Otherwise the result is k. See details below.	(similar in excel)
igamma(j,k)	Incomplete gamma.	(no equivalent)
igammac(j,k)	Incomplete gamma, complement.	(no equivalent)
int(k)	Truncate fraction. INT(3.5)=3 INT(-2.3) = -2	trunc()
ln(k)	Natural logarithm.	ln(k)
log(k)	Log base 10.	log10(k)
max(j,k)	Maximum of two values.	max(j,k)
min(j,k)	Minimum of two values.	min(j,k)
j mod k	The remainder (modulus) after dividing j by k.	mod(j,k)
psi(k)	Psi (digamma) function. Derivative of the gamma function.	(no equivalent)
rad(k)	Converts k degrees to radians.	radians(k)
sgn(k)	Sign of k. If k>0, sgn(k)=1. If k<0, sgn(k)= -1. If k=0, sgn(k)=0.	sign(k)
sin(k)	Sine. K is in radians.	sin(k)
sinh(k)	Hyperbolic sine. K is in radians.	sinh(k)
sqr(k)	Square.	k*k
sqrt(k)	Square root.	sqrt(k)
tan(k)	Tangent. K is in radians.	tan(k)
tanh(k)	Hyperbolic tangent. K is n radians.	tanh(k)
tdist(t,v)	P value (one-tailed) corresponding to specified value of t with v degrees of freedom. T distribution.	tdist(t,v,1)
zdist(z)	P value (one-tailed) corresponding to specified value of z. Gaussian distribution.	normsdist(z)

Using the IF function

Prism allows you to introduce some branching logic through use of the IF function. The syntax is:

```
IF (conditional expression, value if true, value if false)
```

You can precede a conditional expression with NOT, and can connect two conditional expressions with AND or OR. Examples of conditional expressions:

```
MAX>100
Ymax=Constraint
(A<B or A<C)
NOT(A<B AND A<C)
FRACTION<>1.0
X<=A and X>=B
```

Note: "<>" means not equal to, "<=" means less than or equal to, and ">=" means greater than or equal to.

Here is an example.

```
 Y= IF (Y<Y0, Y, Y*Y)
```

If Y is less than Y0, then Y is unchanged. Otherwise Y is transformed to Y squared.

Here is a function that returns Y if Y is positive or zero, but otherwise leaves the results blank. In other words, it removes all negative values. The way to leave a result blank is to do an impossible mathematical transform such as dividing by zero.

```
Y = IF (Y<0, Y/0, Y)
```

Transferring transforms with data files

Prism maintains a list of user-defined transformations. Whenever you transform data, you can choose from transformations you used before.

What happens when you want to transfer a file to another computer? There are no explicit commands to import or export transforms. Prism handles the situation automatically by including the transform in the project file. When you open the file, Prism first reads the name of the transform from the file.

1. If a transform with exactly the same name already exists in the equation list on that computer, the equation is read from the list even if it is different than the one saved on the file. Prism will not use the transform saved with the file, but will instead use the transform with the same name already stored on the computer. This allows you to use a revised function with stored data.
2. If the list does not contain a transform with the same name, then Prism reads the transform from the file and adds it to the list stored on the computer.

Transforming replicates and error bars

If you entered replicate Y values, Prism can transform each replicate or the mean of the replicates.

If you entered data as mean, SD (or SEM), and N, Prism tries to transform the error bar as well as the mean. When a transform is intrinsically asymmetrical (i.e. logarithms), it is mathematically impossible to transform a SD and end up with a SD. You have two choices. You may either transform the mean only or erase the error bars. Or you may convert the error bars to 95% confidence intervals, and then transform both ends of the confidence interval. The resulting 95% CI will be asymmetrical.

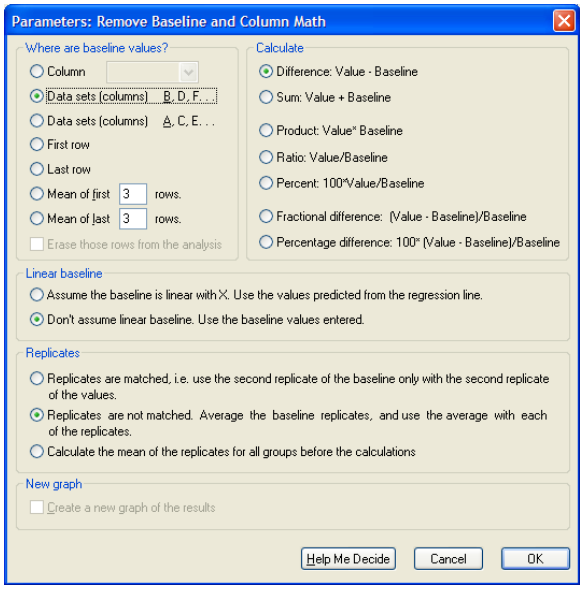
21. Removing Baselines, Normalizing, Pruning, and Transposing

Subtracting (or dividing by) baseline values

Many kinds of data combine a measurement (signal) you care about with a baseline or background (noise) you don't care about. You can analyze these data using two approaches. One approach is to perform analysis on the total signal. The other approach is to subtract or divide by a baseline or nonspecific value and then analyze and graph the results.

> Tip: Although the dialog is called Subtract or Divide a baseline, you can also add or multiply two columns.

Click **Analyze** and choose **built-in analyses**. Then choose **Remove Baseline** from the list of data manipulations to bring up this dialog. The choices are self-explanatory:



Where are the baseline values?

Pick one of the first three choices when you have measured a baseline value at every value of X, and have placed baseline or nonspecific values in datasets adjacent to the main data. Pick one of the last four choices when you have collected data over time, and the first or last few points define the baseline.

Calculate

Choose if you want to compute the difference, ratio, percent difference, etc. You can also choose sum or product.

Linear baseline

If you choose the option to assume a linear baseline, Prism performs linear regression with the background (nonspecific) values and then subtracts (or divides) the Y value predicted from that line. This method is appropriate when you know that the nonspecific or background values must be linear with the X values (for example nonspecific binding is often linear with ligand concentration), and is particularly useful when you have not collected baseline or nonspecific measurements at every value of X (Prism will fill in the missing nonspecific values from linear regression). When Prism fits linear regression, it does not assume that the line goes through the origin and does not display the regression results.

Replicates

If you state that the replicates are matched, Prism computes the difference (or ratio, percent difference, etc.) for the first replicate of total and the first replicate of nonspecific. Then repeats for the second replicate. Otherwise, Prism computes the mean of the total replicates and the mean of the nonspecific and computes the difference (or ratio, percent difference, etc) of the means.

Create a new graph

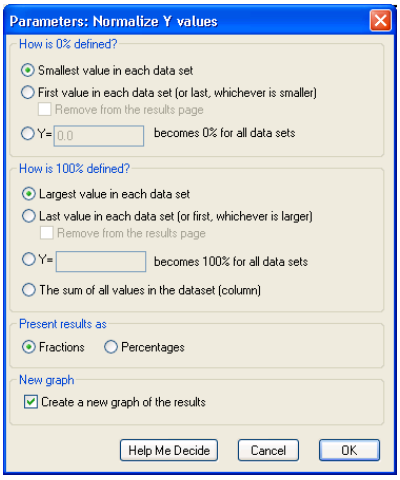
You'll almost always want to check this option, so Prism makes a new graph of the results.

Normalizing data

Normalize the data to convert Y values from different data sets to a common scale. This is useful when the want to compare the shape or position (EC_{50}) of two or more curves, and don't want to be distracted by different maximum and minimum values.

Investigators who analyze dose-response curves commonly normalize the data so all curves begin at 0% and plateau at 100%. If you then fit a sigmoid dose-response curve to the normalized data, be sure to set the top and bottom plateaus to constant values. If you've defined the top and bottom of the curves by normalizing, you shouldn't ask Prism to fit those parameters.

To normalize, click **Analyze** and choose **Built-in analyses**. Then select **Normalize** from the list of data manipulations to bring up this dialog.



To normalize between 0 and 100%, you must define these baselines. Define zero as the smallest value in each data set, the value in the first row in each data set, or to a value you

enter. Define one hundred as the largest value in each data set, the value in the last row in each data set, a value you enter, or the sum of all values in the column. Prism can express the results as fractions or percentages.

Notes:

- ✓ If you have entered replicate values, zero and one hundred percent are defined by the mean of the replicates. It is not possible to normalize each sub column separately.
- ✓ The X values are copied to the results table. They are not normalized.
- ✓ Each SD or SEM is normalized appropriately.
- ✓ If you normalize to the smallest and largest value in the data set, you can remove those values (which would become 0.000 and 1.000) from the results.

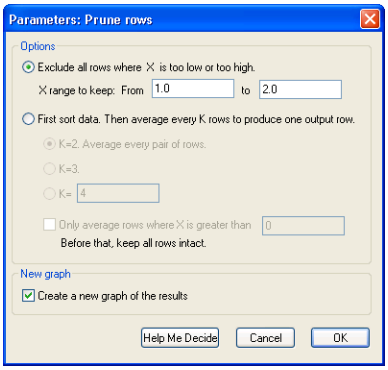
> Tip: The Remove Baseline analysis lets you subtract (or divide) all values by the mean of the first few rows. With some experimental designs, this is the best way to normalize.

Pruning rows

This analysis reduces the size of large data sets to speed curve fitting and graphing. Use it to preprocess large data sets imported from an instrument. The pruning analysis starts with a large data table and generates a shorter results table. Another way to deal with large data sets is to decimate data while importing, so Prism only reads every tenth (or some other number) row. See the chapter in importing data in the Prism User's Guide.

> Note: After pruning, the project contains both the original data and the pruned data. Therefore pruning increases the size of the project file. If you don't want the original data any more, you should go to that data table and use the Delete Sheet command (on the Edit menu) to remove it.

To prune, click **Analyze** and choose **Built-in analyses**. Then choose **Prune** from the list of data manipulations to bring up this dialog.



One choice is to remove all rows where X is too low or too high, and keep only rows where X is between limits you enter.

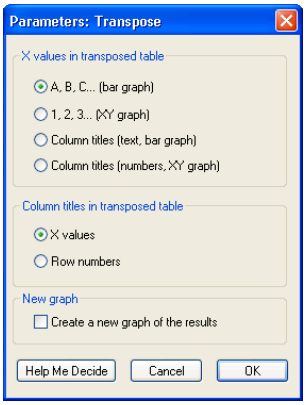
The other choice is to average every K rows to produce one output row (you enter K). First Prism sorts the table by X (if not already sorted), then it averages every K rows to produce one output row. For example, if K=3, the first X value in the results table is the average of the X values of the first three rows. The first Y value of each data set is the average of the Y values in the first three rows. The second row in the results table is the average of rows 4

to 6, and so on. If your data have more than one replicate, the pruning is done by replicate. So if K=2, the value of Y1 (the first replicate) in the results will be the average of Y1 of row 1 of the original data and Y1 of row 2 of the original data.

Optionally average only rows after a threshold X value. When X is lower than that threshold, keep all data intact. After that threshold, reduce the number of rows by a factor of K. This is useful if your experiment reaches a plateau value, and you only want to prune the values near the plateau.

Transposing rows and columns

Click **Analyze** and choose **Transpose** from the list of data manipulations to bring up this dialog.



Each row of Y values becomes one column (data set) in the results table. The first row becomes the first data set, the second row becomes the second data set, etc. You may not transpose a data table with more than 104 rows, because Prism cannot create a table with more than 104 columns.

The column and row titles in the results table are determined by your choices in the dialog.

Note: You can also transpose via copy and place, without using the Transpose analysis. To do this, select a portion, or all, of a table, and copy to the clipboard. Position the insertion point to a different part of the same table or to a new table. Drop the Edit menu and choose Paste Special. Finally, choose Transpose on the Placement tab.

Index

H

I

K

L

M

N

O

P

Q

R

S

T

U

W

Y