Program design, manual and help screens:	Dr. Harvey J. Motulsky Paige Searle
Programming:	Mike Platt John Pilkington Harvey Motulsky

Macintosh conversion by Software MacKiev. www.mackiev.com

Project Manager:	Dennis Radushev
Programmers:	Alexander Bezkorovainy
	Dmitry Farina
Quality Assurance:	Lena Filimonihina
Help and Manual:	Pavel Noga
	Andrew Yeremenko

How to reach GraphPad:

Phone:	(US) 858-457-3909
Fax:	(US) 858-457-8141
Email:	support@graphpad.com or sales@graphpad.com
Web:	www.graphpad.com
Mail:	GraphPad Software, Inc. 5755 Oberlin Drive # 110 San Diego, CA 92121 USA

The entire text of this manual is available on-line at www.graphpad.com

Introduction to statistical principles

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 differences can be obscured by biological variability and experimental imprecision. This makes it hard to distinguish real differences from random variability.
- The human brain excels at finding patterns, even from 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 most useful when observed differences are small compared to experimental imprecision and biological variability. If you only care about large differences, 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!

But in many fields, scientists care about small differences and are faced with large amounts of variability. Statistical methods are necessary to draw valid conclusions from these data.

The key concept: Sampling from a population

Sampling from a population

The basic idea of statistics is simple: you want to extrapolate from the data you have collected to make general conclusions.

To do this, statisticians have developed methods based on this simple model: Assume that all your data are randomly sampled from an infinitely large population. Analyze this sample to make inferences about the population.

In some fields of science – for example, quality control – you really do collect random samples from a large (if not infinite) population. In other fields, you encounter two problems:

The first problem is that you don't really have a random sample. It is rare for a scientist to randomly select subjects from a population. More often you just did an experiment a few times and want to extrapolate to the more general situation. But you can define the population to be the results of a hypothetical experiment done many times (or a single experiment performed with an infinite sample size).

The second problem is that you generally want to make conclusions that extrapolate beyond the population. The statistical inferences only apply to the population your samples were obtained from. Let's say 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. Statistics can't help with this further extrapolation. You can use scientific judgment and common sense to make inferences that go beyond statistics. Statistical logic is only part of data interpretation.

Even though scientific research is not really based on drawing random samples from populations, the statistical tests based on this logic have proven to be very useful in analyzing scientific data. This table shows how the terms *sample* and *population* apply in various kinds of experiments.

Situation	Sample	Population
Quality control	The items you tested.	The entire batch of items produced.
Political polls	The voters you polled.	All voters.
Clinical studies	Subset of patients who attended Tuesday morning clinic in August.	All similar patients.
Laboratory research	The data you actually collected.	All the data you could have collected if you had repeated the experiment many times the same way.

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. The concept of independence is hard to grasp. Consider these 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, because the triplicate measurements in one animal are likely to be closer to each other than to measurements from the other animals. 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 laboratory 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 ten patients from an inner-city hospital and ten more patients from a suburban clinic. You have not independently sampled 20 subjects from one population. The data from the ten inner-city patients may be closer to each other than to the data from the suburban patients. You have sampled from two populations, and need to account for this in your analysis.

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 (that you didn't account for in the analysis of the data) can affect more than one, but not all, of the values, then the data are not independent.

How statistics can extrapolate from sample to population

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

The first method is to assume that the 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.

The second method is to convert all values to ranks, and look at the distribution of ranks. This is the principle behind most commonly used nonparametric tests.

The third method is known as resampling. This is best seen 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 them in a hat. Create many pseudo samples by drawing a card from the hat, then return it. You can 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. The distribution of the means of these pseudo samples gives you information about how well you know the population mean. The idea of resampling is hard to grasp. To learn about this approach to statistics, read the instructional material available at www.resampling.com. InStat does not perform any tests based on resampling.

Confidence intervals

Statistical calculations produce two kinds of results that help you make inferences about the population by analyzing the samples. Confidence intervals are explained here, and P values are explained in the next section.

Confidence interval of a mean

The mean you calculate from a sample is unlikely to equal the population mean. The size of the discrepancy depends on the size and variability of the sample. If your sample is small and variable, the sample mean may 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 confidence interval (CI) for the population mean. You can calculate intervals for any desired degree of confidence, but 95% confidence intervals are used most commonly. If you assume that your sample is randomly selected from some population (that follows a Gaussian distribution, see "What is the Gaussian distribution?" on page 19), you can be 95% sure that the confidence interval includes the population mean. More precisely, if you generate many 95% CI 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%. Since you don't know the population mean, you'll never know when this happens.

Confidence intervals in other situations

Statisticians have derived methods to generate confidence intervals for almost any situation. For example when comparing groups, you can calculate the 95% confidence interval for the difference between the population means. Interpretation is straightforward. If you accept the assumptions, there is a 95% chance that the interval you calculate includes the true difference between population means.

Similarly, methods exist to compute a 95% confidence interval for the relative risk, the best-fit slope of linear regression, and almost any other statistical parameter.

P values

What is a P value?

Assume that you've collected data from two samples, and the means are different. You want to know whether the data were sampled from populations with 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, and the difference you observed is a coincidence of random sampling. There is no way you can ever be sure whether the difference you observed reflects a true difference or a coincidence of random sampling. All you can do is calculate the probabilities.

The P value answers this question: If the populations really did have the same mean, what is the probability of observing such a large difference (or larger) between sample means in an experiment of this size?

The P value is a probability, with a value ranging from zero to one. If the P value is small, you'll conclude that the difference is quite unlikely to be caused by random sampling. You'll conclude instead that the populations have different means.

What is a null hypothesis?

When statisticians refer to 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 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. Wrong. 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.

The P value is a fraction, but what it is a fraction of? The P value is the fraction of all possible results obtained under the null hypothesis where the difference is as large or larger than you observed. That is NOT the same as the fraction of all experiments that yield a certain P value where the null hypothesis is true. To determine that fraction, you need to use Bayesian reasoning – beyond the scope of InStat.

One- vs. two-tail P values

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

Start with the null hypothesis that the two populations really are the same and that the 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 (or further) as 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 (or further) as 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 the same.

You should only choose a one-tail P value when two things are true. First, you must have predicted which group will have the larger mean (or proportion) before you collected any data. That's easy, but the second criterion is harder. If the other group ends up with the larger mean – even if it is quite a bit larger – then you must attribute 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, no matter how large the difference is. You must say that the difference is not statistically significant. But most people would be tempted 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

The P value is a fraction. In many situations, the best thing to do is report that fraction to summarize your results ("P = 0.0234"). If you do this, you can totally avoid using the term "statistically significant", which is often misinterpreted.

In other situations, you'll want to make a decision based on a single comparison. In these situations, follow the steps of statistical hypothesis testing.

1. Set 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 fact, the threshold value (called α) is traditionally almost always set to 0.05.
2. Define the null hypothesis. If you are comparing two means, the null hypothesis is that the two populations have the same mean.
3. Do the appropriate statistical test to compute the P value.
4. Compare the P value to the preset threshold value.
5. If the P value is less than the threshold, state that you "reject the null hypothesis" and that the difference is "statistically significant".

6. 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

The term *significant* is seductive, and it is easy to misinterpret it. A result is said to be *statistically significant* when the result would be surprising if the populations were really 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 important or interesting. And 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 really is no difference. By chance, you obtained larger values in one group and smaller values in the other. 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.

Type II errors and statistical power

When a study reaches a conclusion of "no statistically significant difference", you should not necessarily conclude that the treatment was ineffective. It is possible that the study missed a small effect due to small sample size and/or large scatter. In this case you made a Type II error — concluding that there is no difference 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 if they existed. The power depends on the sample size and amount of variation within the groups, as quantified by the standard deviation (SD).

Here is a precise definition of power. Start with the assumption that the two population means differ by a certain amount and that the SD of the populations has a particular value. 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 have a P value less than α (the largest P value you deem ”significant“, usually set to 0.05). In other words, power equals the fraction of experiments that would lead to statistically significant results. InStat does not compute power.

"Extremely significant" results

Intuitively, you may think that P = 0.0001 is more statistically 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 either statistically significant or is not statistically significant. Some statisticians feel very strongly about this. Many scientists are not so rigid, and refer to results as being "very significant" or "extremely significant" when the P value is tiny.

InStat 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
>0.05	Not significant	ns
0.01 to 0.05	Significant	*
0.001 to 0.01	Very significant	**
< 0.001	Extremely significant	***

Beware of multiple comparisons

A result is said to be statistically significant when it would occur rarely under the null hypothesis. Therefore you conclude that the null hypothesis is unlikely to be true. But if you perform enough tests, statistically significant results will occur often (even if the null hypotheses are all true).

For example, assume you perform ten independent statistical tests and the null hypotheses are all true. The probability is 5% that any particular test will have a P value less then 0.05. But by performing ten tests, there is a very high chance that at least one of those comparisons will have a P value less than 0.05. The probability is about 40% (to calculate this, first calculate the probability of getting ten consecutive P values greater than 0.05, which is 0.95^{10}, or about 60%; so the chance that at least one of the P values is less than 0.05 is 100% - 60% or 40%).

The multiple comparison problem means that you cannot interpret a small P value without knowing how many comparisons were made. There are three practical implications:

- When comparing three or more groups, you should not perform a series of t tests. Instead, use one-way ANOVA followed by posttests (which take into account all the comparisons).
- Beware of data mining. If you look at many variables, in many subgroups, using many analyses, you are sure to find some small P values. But these are likely to occur by chance. Data exploration can be fun, and can lead to interesting ideas or hypotheses. But you'll need to test the hypotheses with a focussed experiment using new data.
- All analyses should be planned and all planned analyses should be reported. It is not fair to include in your papers the analyses that give small P values while excluding those that gave large P values.

The Gaussian distribution and testing for normality

What is the Gaussian distribution?

When many independent random factors act in an additive manner to create variability, data will follow a bell-shaped distribution called the Gaussian distribution. This distribution is also called a Normal distribution (don't confuse this use of the word "normal" with its usual meaning). The Gaussian distribution has some special mathematical properties that form the basis of many statistical tests. Although no data follows that mathematical ideal, many kinds of data follow a distribution that is approximately Gaussian.

What's so special about the Gaussian distribution?

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 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 about differences between means, the Central Limit Theorem lets these tests work well even when the populations are not Gaussian. The catch is that the samples have to be reasonably large. How large is that? It depends on how far the population distribution differs from a Gaussian distribution.

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

Nonparametric tests

The t test and ANOVA, as well as other statistical tests, assume 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 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 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. See "Type II errors and statistical power" on page 16.

- 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 fewer than a dozen or so categories (i.e. Apgar score). Clearly the population cannot be Gaussian in these cases.
- 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 values too low to measure an arbitrary low value, and values too high to measure an arbitrary high value. Since the nonparametric tests only consider the relative ranks of the values, it won't matter that you didn't know a few values exactly.
- You are sure that the population is far from Gaussian. Before choosing a nonparametric test, consider transforming the data (i.e. logarithms, reciprocals). Sometimes a simple transformation will convert nongaussian data to a Gaussian distribution. See "Transforming data to create a Gaussian distribution" on page 21.

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.

InStat 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 "Testing for normality" on page 22). 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 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 nongaussian populations. Small samples simply don't contain enough information to let you make inferences about the shape of the distribution in the entire population.

Transforming data to create a Gaussian distribution

If your data do not follow a Gaussian (normal) distribution, you may be able to transform the values to create a Gaussian distribution. If you know the distribution of your population, transforming the values to create a Gaussian dist0ribution is a good thing to do, as it lets you use statistical tests based on the Gaussian distribution.

This table shows some common normalizing transformations:

Type of data and distribution	Normalizing transformation
Count (C comes from Poisson distribution)	Square root of C
Proportion (P comes from binomial distribution)	Arcsine of square root of P
Measurement (M comes from lognormal distribution)	Log(M)
Time or duration (D)	1/D

Testing for normality

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

InStat uses the method of Kolmogorov and Smirnov to calculate KS. However, the method originally published by those investigators 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, InStat 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, InStat simply reports "P > 0.10" for large P values.

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 as much from a Gaussian distribution (or more so) as this sample does. More precisely, the P value answers this question: If the population was really Gaussian, what is the chance that a randomly selected sample of this size would have a KS distance as large, or larger, as observed?

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

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

P value	Sample size	Conclusion
Small	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. Don't do that. A large P value just means that the data are not inconsistent with a Gaussian population. That doesn't exclude the possibility of a nongaussian population. Small sample sizes simply don't provide enough data to discriminate between Gaussian and nongaussian distributions. You can't conclude much about the distribution of a population if your sample contains fewer than a dozen values.

Descriptive statistics

Column statistics

There are many ways to describe the distribution of a group of values. After you enter data for column comparisons, InStat next presents a table of descriptive statistics for each column.

Descriptive statistics

Statistic	Definition
Mean	The mean is the average of all the values in the column.
Standard deviation	The standard deviation (SD) quantifies variability or scatter among the values in a column. 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. InStat calculates the "sample SD" (which uses a denominator of N-1), not the "population SD" with a denominator of N. InStat does not report the variance. If you want to know the variance, simply square the standard deviation. Variance is expressed in the units of your data squared.
Standard error of the mean	The standard error of the mean (SEM) is a measure of the likely discrepancy between the mean calculated from your data and the true population mean (which you can't know without an infinite amount of data). The SEM is calculated as the SD divided by the square root of sample size. With large samples, therefore, 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.
Confidence interval	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 true 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. If the population is Gaussian (or nearly so), you can be 95% sure that this interval contains the true population mean. More precisely, if you generate many 95% CI 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%. Since you don't know the population mean, you'll never know when this happens.
Median	The median is the 50th percentile. Half the values are larger than the median, and half are lower. If there are an even number of values, the median is defined as the average of the two middle values.
Normality test	For each column, InStat reports the results of the normality test. If the P value is low, you can conclude that it is unlikely that the data were sampled from a Gaussian population. See "Testing for normality" on page 22.

SD vs. SEM

Many scientists are confused about the difference between the standard deviation (SD) and 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 population mean. The SEM gets smaller as your samples get larger, simply because the mean of a large sample is likely to be closer to the true 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 increase 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, your readers may want to see the variation. In this case, report the SD rather than the SEM. Better, show a graph of all data points, or perhaps report the largest and smallest value — there is no reason to only report the mean and SD.

If you are using an *in vitro* system with no biological variability, the scatter can only result from experimental imprecision. Since you don't really care about the scatter, the SD is less useful here. Instead, report the SEM to give your readers a sense of how well you have determined the mean.

Mean vs. median

The mean is the average. The median is the middle value. Half the values are higher than the median, and half are lower.

The median is a more robust measure of central tendency. Changing a single value won't change the median very much. In contrast, the value of the mean can be strongly affected by a single value that is very low or very high.

Entering averaged data into InStat

If you have already analyzed your data with another program, you may not need to enter every value into InStat. Instead, enter the mean, sample size (N) and either standard deviation (SD) or standard error of the mean (SEM) for each column. On the first step, choose that you want to enter mean with sample size and SD (or SEM). InStat won’t let you go to the data table. Enter the data directly on the column statistics page.

Paired, repeated measures, and nonparametric tests require raw data, and cannot be performed if you enter averaged data.

You can also enter raw data into some columns and averaged data into others. Format the data table for raw data. After entering raw data into some columns, go to the column statistics step. You'll see the mean, SD etc. for the data you have entered. In blank column(s) enter the mean, SD and N.

The one-sample t test and Wilcoxon rank sum test determine whether the values in a single column differ significantly from a hypothetical value.

You need to make three choices:

Choice	Explanation
Parametric or nonparametric?	InStat can compare the mean with the hypothetical value using a one-sample t test, or compare the median with the hypothetical value using the nonparametric Wilcoxon signed rank test. Choose the one-sample t test if it is reasonable to assume that the population follows a Gaussian distribution. Otherwise choose the Wilcoxon nonparametric test, realizing that the test has less power. See "Nonparametric tests" on page 20.
One- or two-tailed P value?	If in doubt, choose a two-tail P value. See ″One- vs. two-tail P values“ on page 14.
What is the hypothetical value?	Enter the hypothetical mean or median, often 0, 1, or 100. The hypothetical value comes from theory, from other kinds of experiments, or from common sense (for example, if data expressed as percent of control you may want to test whether the mean differs significantly from 100).

The results of a one-sample t test

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:

Question	Discussion
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. InStat tests for violations of this assumption, but normality tests have limited utility. See "Testing for normality" on page 22. 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 (see "Transforming data to create a Gaussian distribution" on page 21). 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 violations of 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. There is no way for InStat to test this assumption. See "The need for independent samples" on page 11.
Are you interested only in the means?	The one sample t test compares the *mean* of a group with a hypothetical mean. Even if the P value is tiny– clear evidence that the population mean differs from the hypothetical mean – the distribution of values may straddle the hypothetical mean with a substantial number of values on either side.
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. InStat 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 InStat and state that P > 0.50. See ″One- vs. two-tail P values“ on page 14.

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

The one-sample t test compares the mean of one column of numbers to a theoretical mean.

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 looking for. 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? 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

If the P value is small, then it is unlikely that the discrepancy you observed between sample mean and hypothetical mean is due to a coincidence of 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 it scientifically significant? The confidence interval helps you decide.

Your data are affected by random scatter, so 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 what that true difference is. InStat 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 and ask whether they represent a discrepancy that would be 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 true 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

If the P value is large, the data do not give you any reason to conclude that the overall 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 group mean in this experiment is unlikely to equal the true difference between population mean and hypothetical mean. There is no way to know what that true difference is. InStat presents the uncertainty as a 95% confidence interval. You can be 95% sure that this interval contains the true difference between overall (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 actual mean) and go up to a positive number (the actual mean is larger than the hypothetical mean).

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

Lower confidence limit	Upper confidence limit	Conclusion
Trivial	Trivial	You can reach a crisp conclusion. Either the data has a mean equal to the hypothetical mean or they differ by a trivial amount.
Trivial	Large	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.

Large	Trivial	You can’t reach a strong conclusion. The data are consistent with a mean smaller than the hypothetical mean (perhaps enough smaller 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.

The results of a one-sample t test, line by line

Result	Explanation
P value	The P value that 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?
t ratio	InStat calculates the t ratio from this equation: t = (Sample Mean – Hypothetical Mean)/SEM
95% confidence interval	InStat calculates the 95% confidence interval for the difference between the mean calculated from your sample and the hypothetical (theoretical) mean you entered. You can be 95% sure that the interval contains the true difference.
Normality test	The one sample t test assumes that your data were sampled from a population that is distributed according to a Gaussian distribution. The normality test attempts to test this assumption. If the P value is low, conclude that the population is unlikely to be Gaussian. Either transform your data to make the distribution Gaussian, or choose the nonparametric Wilcoxon test. See "Testing for normality" on page 22.

The results of a Wilcoxon test

Checklist. Is the Wilcoxon rank sum 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 test, ask yourself these questions (InStat cannot help you answer them):

Question	Discussion
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. There is no way for InStat to test this assumption. See ″The need for independent samples“ on page 11.
Are the data clearly sampled from a nongaussian 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, InStat (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 logs or reciprocals) to create a Gaussian distribution and then using a 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 would have the larger median before collecting any data. InStat 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 InStat and state that P > 0.50. See ″One- vs. two-tail P values“ on page 14.

Approach to interpreting the results of a Wilcoxon signed rank 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 coincidence, 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 overall median differs from the hypothetical median. This is not the same as saying that the medians are the same. You just have no 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.

How the Wilcoxon rank sum test works

InStat follows these steps:

1. Calculate how far each value is from the hypothetical value.
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. InStat reports this value.
6. Sum the negative ranks. InStat also reports this value.
7. Add the two sums together. This is the sum of signed ranks, which InStat 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: Assume 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) as you observed?

Comparing two groups (t tests etc.)

Introduction to t tests

Use the t test, and corresponding nonparametric tests, to test whether the mean (or median) of a variable differs between two groups. For example, compare whether systolic blood pressure differs between a control and treated group, between men and women, or any other two groups.

Don't confuse t tests with correlation and regression. The t test compares one variable (perhaps blood pressure) between two groups. Use correlation and regression to see how two variables (perhaps blood pressure and heart rate) vary together.

Also don't confuse t tests with ANOVA. The t tests (and related nonparametric tests) compare exactly two groups. ANOVA (and related nonparametric tests) compare three or more groups.

Finally don't confuse a t test with analyses of a contingency table (Fishers or chi-square test). Use a t test to compare a continuous variable (i.e. blood pressure, weight or enzyme activity). Analyze a contingency table when comparing a categorical variable (i.e. pass vs. fail, viable vs. not viable).

Entering t test data into InStat

Enter each group into its own column. InStat compares the means (or medians) to ask whether the observed differences are likely to be due to coincidence.

Enter either raw data (enter each value) or averaged data (enter mean, N and SD or SEM). If you enter averaged data, InStat will not offer nonparametric or paired tests, which require raw data.

When entering raw data, simply leave a blank spot in the table to denote missing values. If you enter averaged data, you must enter the mean, N and SD (or SEM) for each column. It is okay if N differs among columns, but you must enter mean, N and SD (or SEM) for each column; you can't leave any of those values blank.

Ideally, you should decide about pairing before collecting data. Certainly the matching should not be based on the variable you are comparing. If you are comparing blood pressures in two groups, it is okay to match based on age or zip code, but it is not okay to match based on blood pressure.

Parametric or nonparametric test?

The t test, like many statistical tests, assumes that your data are sampled from a population that follows a Gaussian bell-shaped distribution. Alternative tests, known as nonparametric tests, make fewer assumptions about the distribution of the data, but are less powerful (especially with small samples). Choosing between parametric and nonparametric tests can be difficult. See ″Nonparametric tests“ on page 20. The results of a normality test can be helpful, but not always as helpful as you'd hope. See "Testing for normality" on page 22

Assume equal variances?

The unpaired t test assumes that the data are sampled from two populations with the same variance (and thus the same standard deviation). Use a modification of the t test (developed by Welch) when you are unwilling to make that assumption. This choice is only available for the unpaired t test. Use Welch's t test rarely, when you have a good reason. It is not commonly used.

One- or two-tail P value?

Choose a one-tailed P value only if:

- You predicted which group would have the larger mean before you collected any data.
- If the other group turned out to have the larger mean, you would have attributed that difference to coincidence, even if the means are very far apart.

Since those conditions are rarely met, two-tail P values are usually more appropriate. See ″One- vs. two-tail P values“ on page 14.

Summary of tests to compare two columns

Based on your answers…		…InStat chooses a test
Not paired	Gaussian distribution, equal SDs	Unpaired t test
Not paired	Gaussian distribution, different SDs	Welch's t test
Paired	Gaussian distribution of differences	Paired t test
Not paired	Not Gaussian	Mann-Whitney test
Paired	Not Gaussian	Wilcoxon test

The results of an unpaired t test

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 these questions:

Questions that InStat can help you answer

Question	Discussion
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). InStat tests for violations of this assumption, but normality tests have limited utility. See "Testing for normality" on page 22. If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values to make the distributions more Gaussian (see "Transforming data to create a Gaussian distribution" on page 21). 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 standard deviation?	The unpaired t test assumes that the two populations have the same standard deviation (and thus the same variance). InStat 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. See ”F test to compare variances“ on page 52.

Questions about experimental design

Question	Discussion
Are the data unpaired?	The unpaired t test works by comparing the difference between means with the pooled standard deviations of the two groups. If the data are paired or matched, then you should choose a paired t test. 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. There is no way for InStat to 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 11.
Are you comparing exactly two groups?	Use the t test only to compare two groups. To compare three or more groups, use one-way Analysis of Variance 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 a t test. Instead, use a one-sample t test. See "Choosing the one-sample t test or Wilcoxon test" on page 35.
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. InStat 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 InStat and state that P > 0.50. See "One- vs. two-tail P values" on page 14.

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) 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 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.

If the P value is small

If the P value is small, then it is unlikely that the difference you observed is due to a coincidence of random sampling. You can reject the idea that the difference is a coincidence, and conclude instead that the populations have different means. The difference is statistically significant. But is it scientifically significant? 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. InStat 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 (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 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 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.

If the P value is large

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 coincidence. 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. InStat 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 a large increase. 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.

The results of an unpaired t test, line by line.

P value

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? More precisely, the P value answers this question: If the populations really had the same mean, what is the chance of obtaining a t ratio as far from zero (or more so) than you obtained in this experiment.

If you chose a one-tail P value, you must have predicted which group would have the larger mean before collecting any data. InStat 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 InStat and state that $P > 0.50$.

See "P values" on page 13.

t ratio

The t ratio is an intermediate calculation of the t test. InStat first computes a t ratio, and then uses it to determine the P value.

InStat calculates the t ratio by dividing the difference between sample means by the standard error of the difference, calculated by pooling the SEMs of the two groups. If the difference is large compared to the SE of the difference, then the t ratio is also large (or is a large negative number), and the P value is small.

For the standard t test, the number of degrees of freedom (df) equals the total sample size minus 2. Welch's t test calculates df from a complicated equation. InStat calculates the P value from t and df.

CI for difference between means

Because of random variation, the difference between the group means in this experiment is unlikely to equal the true difference between population means. The size of the discrepancy depends on the scatter of your samples and the number of values in your sample. InStat reports the uncertainty as the 95% confidence interval of the mean. If you accept the assumptions of the analysis, you can be 95% sure that the confidence interval includes the true difference between group means.

The confidence interval is centered on the difference between the sample means. It extends in each direction by a distance calculated from the standard error of the difference (computed from the two SEM values) multiplied by a critical value from the t distribution for 95% confidence and corresponding to the number of degrees of freedom in this experiment. With large samples, this multiplier equals 1.96. With smaller samples, the multiplier is larger.

F test to compare variances

InStat tests whether the variances of the two groups are the same by calculating F, which equals the larger variance divided by the smaller variance. Remember that the variance equals the standard deviation squared. The degrees of freedom for the numerator and denominator equal the sample sizes minus 1. From F and the two df values, InStat 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 possible, don't base your conclusion just on this one F test. Also consider data from other experiments in the series, if possible. If you conclude that the two populations have different variances, you have three choices:

- Conclude that the two populations are different – 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, then rerun the t test. Often you'll find that converting values to their reciprocals or logarithms will equalize the variances and make the distributions more Gaussian. See "Transforming data to create a Gaussian distribution" on page 21.
- Rerun the t test without assuming equal variances using Welch's modified t test.

Normality test

The t test assumes that data are sampled from Gaussian populations. This assumption is tested with a normality test. See "Testing for normality" on page 22.

The results of a paired t test

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:

Questions that InStat can help you answer

Question	Discussion
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. InStat tests for violations of this assumption, but normality tests have limited utility. If your data do not come from Gaussian distributions, you have two options. Your best option is to transform the values to make the distributions more Gaussian (see "Transforming data to create a Gaussian distribution" on page 21. Another choice is to use the Wilcoxon 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. InStat tests the effectiveness of pairing by calculating the Pearson correlation coefficient, r, and a corresponding P value. See "Correlation coefficient" on page 102. 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 on this one P value, but also on the experimental design and the results you have seen in other similar experiments.

Questions about experimental design

Question	Discussion
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. There is no way for InStat to 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 11.
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 Analysis of Variance followed by post tests. It is not appropriate to perform several t tests, comparing two groups at a time.
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). It turns out that analyses of ratios are problematic. The problem is that the ratio is 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. If you have paired data and think that it makes more sense to look at ratios rather than differences, follow these steps. First transform both columns to logarithms. Then perform a paired t test. Note that the difference between logarithms (that InStat analyzes in this case) equals the log of the ratio.
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. InStat 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- vs. two-tail P values“ on page 14.

How to think about results of a paired t test

The paired t test compares two paired groups to 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.

If the P value is small

If the P value is small, then it is unlikely that the treatment effect you observed is due to a coincidence of random sampling. 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. InStat 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 there the treatment had an effect large enough to be scientifically relevant.

If the P value is large

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. InStat 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 a large increase. 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.

The results of a paired t test, line by line.

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.

P value

The P value answers this question: If the treatment is really ineffective so the mean difference is really zero in the overall population, what is the chance that random sampling would result in a mean difference as far from zero (or further) as observed in this experiment?

If you chose a one-tail P value, you must have predicted which group would have the larger mean before collecting any data. InStat 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 InStat and state that $P > 0.50$.

See "P values" on page 13.

t ratio

First InStat 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.

CI for difference between means

InStat reports the 95% confidence interval for the mean treatment effect. If you accept the assumptions of the analysis, you can be 95% sure that the confidence interval includes the true mean difference between pairs.

Test for adequate 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 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. InStat quantifies this by calculating the Pearson correlation coefficient, r. From r, InStat 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.

Normality test

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. See "Testing for normality" on page 22.

The results of a Mann-Whitney test

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 (InStat cannot help you answer them):

Question	Discussion
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. There is no way for InStat to 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 11.
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 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 Kruskall-Wallis test followed by post tests. It is not appropriate to perform several Mann-Whitney (or t) tests, comparing two groups at a time.
Are the shapes of the two distributions identical?	The Mann-Whitney test does not assume that the populations follow Gaussian distributions. But it does assume that the shape of the two distributions is identical. The medians may differ – that is what you are testing for – but the test assumes that the shape of the two distributions is identical. If two groups have very different distributions, transforming the data may make the distributions more similar.
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. InStat 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 InStat and state that $P > 0.50$. See ″One- vs. two-tail P values“ on page 14.

Are the data sampled from nongaussian 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, InStat (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 (see "Transforming data to create a Gaussian distribution" on page 21).

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 coincidence, and conclude instead that the populations have different medians.

If the P value is large, the data do not give you any reason to conclude that the overall medians differ. This is not the same as saying that the medians are the same. You just have no 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 the groups differ.

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, InStat 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. InStat 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, InStat calculates an exact P value. If your samples are large, it approximates 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. The approximation is quite accurate with large samples.

The results of a Wilcoxon test

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:

Question	Discussion
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. There is no way for InStat to 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 they are not independent. See "The need for independent samples" on page 11.
Is the pairing effective?	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 will not affect the difference between before and after. By analyzing only the differences, therefore, a paired test controls for some of the sources of scatter. The pairing should be part of the experimental design and not something you do after collecting data. InStat tests the effectiveness of pairing by calculating the Spearman correlation

	coefficient, r_s, and a corresponding P value. See ”Results of correlation“ on page 101. If r_s is positive and P is small, the two groups are significantly correlated. This justifies the use of 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 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. InStat 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 InStat and state that $P > 0.50$. See ”One- vs. two-tail P values“ on page 14.
Are the data clearly sampled from nongaussian 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, InStat (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 logs or reciprocals) to create a Gaussian distribution and then using a t test. See "Transforming data to create a Gaussian distribution" on page 21.
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.

How to think about the results of a Wilcoxon test

The Wilcoxon 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 a coincidence, 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 evidence that they differ. If you have small samples, the Wilcoxon test has little power to detect small differences.

How the Wilcoxon matched pairs test works

P value

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, InStat first computes the differences between each set of pairs. Then it ranks the absolute values of the differences from low to high. Finally, it sums the ranks of the differences where column A was higher (positive ranks) and the sum of the ranks where column B was higher (it calls these negative ranks), and reports these 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 you chose a one-tail P value, you must have predicted which group would have the larger median before collecting any data. InStat 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 InStat and state that $P > 0.50$.

If your samples are small, InStat 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.

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 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. InStat quantifies this by calculating the nonparametric Spearman correlation coefficient, r_s. From r_s, InStat calculates a P value that answers this question: If the two groups really are not correlated at all, what is the chance that randomly selected subjects would have a correlation coefficient as large (or larger) as observed in your experiment (the P value is one-tail, as you are not interested in the possibility of observing a strong negative correlation).

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

If the P value is large (say larger than 0.05), you should question whether it made sense to use a paired test. Your choice of whether to use a paired test or not should not be based on this one P value, but also on the experimental design and the results you have seen in other similar experiments (assuming you have repeated the experiments several times).

If r_s is negative, it means that the pairing was counter productive! 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.

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

Introduction to ANOVA

Use one-way analysis of variance (ANOVA), and corresponding nonparametric tests, to test whether the mean (or median) of a variable differs among three or more groups. For example, compare whether systolic blood pressure differs between a control group and two treatment groups, or among three (or more) age groups.

Rather than using one-way ANOVA, you might be tempted to use a series of t tests, comparing two groups each time. Don't do it. If you have three or more groups, use one-way ANOVA (perhaps followed by post tests) – don't use a series of t tests.

Don't confuse ANOVA with multiple regression. ANOVA test whether the mean (or median) of a single variable (perhaps blood pressure) differs among three or more groups. Multiple regression is used to find out how three or more variables (perhaps blood pressure, age and heart rate) vary together.

One way ANOVA compares three or more groups defined by a single factor. For example, you might compare control, with drug treatment with drug treatment plus antagonist. Or you might compare control with five different drug treatments.

Some experiments involve more than one factor. For example, you might compare the effects of three different drugs administered at two times. There are two factors in that experiment: drug treatment and time. These data need to be analyzed by two-way ANOVA, also called two factor ANOVA. InStat does not perform two-way ANOVA.

Entering ANOVA data into InStat

Enter each group into its own column. InStat compares the means (or medians) to ask whether the observed differences are likely to be due to coincidence.

Enter either raw data (enter each value) or averaged data (enter mean, N and SD or SEM). If you enter averaged data, InStat will not offer nonparametric or paired tests, which require raw data.

The results of one-way ANOVA

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 these questions:

Questions that InStat can help you answer

Question	Discussion
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). InStat tests for violations of this assumption, but normality tests have limited utility. See "Testing for normality" on page 22. If your data do not come from Gaussian distributions, you have three options. Your best option is to transform the values (perhaps logs or reciprocals) to make the distributions more Gaussian (see "Transforming data to create a Gaussian distribution" on page 21. Another choice is to use the Kruskal-Wallis nonparametric test instead of ANOVA. A final option is to use ANOVA anyway, knowing that it is fairly robust to violations of a Gaussian distribution with large samples.
Do the populations have the same standard deviation?	One-way ANOVA assumes that all the populations have the same standard deviation (and thus the same variance). This assumption is not very important when all the groups have the same (or almost the same) number of subjects, but is very important when sample sizes differ. InStat tests for equality of variance with 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 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 75.

Questions about experimental design

Question	Discussion
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. There is no way for InStat to 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 11.
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. InStat does not perform two-way ANOVA.

Is the factor ″fixed“ rather than ″random“?	InStat 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 in biology, and InStat does not perform it.

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) than 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 coincidence. This is not the same as saying that the true means are the same. You just don’t have evidence that they differ.

If the overall P value is small, then it is unlikely that the differences you observed are due to a coincidence of 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 understand where the differences are.

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 linear trend" on page 76.

With other post tests, look at which differences between column means are statistically significant. For each pair of means, InStat 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.

You will interpret the post test results differently depending on whether the difference is statistically significant or not.

If the difference is statistically significant – the P value is small

If the P value for a post test is small, then it is unlikely that the difference you observed is due to a coincidence of 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), InStat 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 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 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.

If the difference is not statistically significant – the P value is large

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 coincidence. 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. InStat 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 a large increase. 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.

Results of one-way ANOVA. Line by line.

P value

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

See "P values" on page 13.

R^2 value

This is the fraction of the overall variance (of all the data, pooling all the groups) attributable to the difference mea 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 (for a definition of sum-of-squares see "ANOVA table " on page 76). 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.

Bartlett's test for equal variances

ANOVA is based on the assumption that the populations all have the same variance. If your samples have five or more values, InStat tests this assumption with Bartlett's test. It reports the value of Bartlett's statistic and the 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 is very sensitive to deviations from a Gaussian distribution – more sensitive than the ANOVA calculations are. 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 nongaussian data (at least when sample sizes are equal), the Bartlett's test can be misleading. 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 have to decide whether you wish to 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 – the treatments 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 ANOVA concludes about differences among the means. This may be the most important conclusion from the experiment.

- Transform the data to equalize the variances, 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. See "Transforming data to create a Gaussian distribution" on page 21.
- Use a modified ANOVA that does not assume equal variances. InStat does not provide such a test.

ANOVA table

The P value is calculated from the ANOVA table. The key idea is that variability among the values can be partitioned into variability among group means and variability within the groups. Variability within groups is quantified as the sum of the squares of the differences between each value and its group mean. This is the residual sum-of-squares. Total variability is quantified as the sum of the squares of the differences between each value and the grand mean (the mean of all values in all groups). This is the total sum-of-squares. The variability between group means is calculated as the total sum-of-squares minus the residual sum-of-squares. This is called the between-groups sum-of-squares.

Even if the null hypothesis is true, you expect values to be closer (on average) to their group means than to the grand mean. The calculation of the degrees of freedom and mean square account for this. See a statistics book for detail. The end result is the F ratio. If the null hypothesis is true, you expect F to have a value close to 1.0. If F is large, the P value will be small. 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?

Post tests (one-way ANOVA)

Post test for linear trend

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

Result	Discussion
Slope	The slope of the best-fit line where the X values are column number (1, 2, 3…) and the Y values are the column means. It is the average increase (decrease, if negative) in column mean as you go from one column to the next column to the right.
R squared	A measure of goodness-of-fit for that best-fit line. See "r2" on page 102.

P value for 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, it is the chance of observing a value of r^2 that high or higher, just by coincidence of random sampling.
P value for nonlinear variation	After correcting for the linear trend, this P value tests whether the remaining variability among column means is greater than expected by chance. It is the chance of seeing that much variability due to random sampling.
ANOVA table	This ANOVA table partitions total variability into three components: linear variation, nonlinear variation, and random or 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 will not be of use to most scientists.

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

Other post tests

For each pair of columns, InStat reports the P value as >0.05, <0.05, <0.01 or <0.001. These P values account for multiple comparisons. If the null hypothesis is true (all the values are sampled from populations with the same mean), then there is only a 5% chance that any one or more comparisons will have a P value less than 0.05. The probability is for the entire family of comparisons, not for each individual comparison.

InStat also reports the 95% confidence intervals for the difference between each pair of means. These intervals account for multiple comparisons. There is a 95% chance that all of these intervals contain the true differences between population means, and only a 5% chance that any one or more of these intervals misses the true population difference.

The results of repeated measures ANOVA

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 these questions. InStat can help you answer the first; you must answer the rest based on experimental design.

Question	Discussion
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. InStat tests the effectiveness of matching with an F test (distinct from the main F test of differences between columns). If this P value is large (say larger than 0.05), you should question whether it made sense to use a repeated measures test. Your choice of whether to use a repeated measures test 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 subjects independent?	The results of repeated measures ANOVA only make sense when the subjects are independent. There is no way for InStat to test this assumption. You must think about the experimental design. For example, the errors are not independent if you have six rows of data of values, 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 11.
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 same for each individual), an individual effect (the same for each treatment) 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. InStat 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. These data need to be analyzed by two-way ANOVA, also called two factor ANOVA. InStat does not perform two-way ANOVA.

Is the factor ″fixed“ rather than ″random“?	InStat 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 in biology, and InStat does not perform it.

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 block of matched subjects. The analyses are identical for repeated measures and randomized block experiments, and InStat 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 72.

The results of repeated measures ANOVA, line by line

P value

Repeated measures one-way ANOVA compares three or more matched groups, based on the assumption that the differences between matched values are Gaussian. 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?

Interpreting the P value from repeated measures ANOVA requires thinking about one of the assumptions of the analysis. Repeated measures ANOVA assumes that the random error truly is truly 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 (equivalently) *sphericity*. It is closely related to another term you may encounter, *compound symmetry*.

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.

Repeated measures ANOVA is quite sensitive to violations of the assumption of circularity. InStat does not attempt to test for violations of the assumption of circularity. When the assumption is violated, the P value from repeated measures ANOVA will be too low. InStat also reports a second P value calculated using the method of Geisser and Greenhouse. This P value is computed from the same F ratio but uses different numbers of degrees of freedom (the numerator df equals one; the denominator df equals one less than the number of subjects). This P value is conservative (too high). No matter how badly the assumption of circularity is violated, the true P value will be between the two P values that InStat presents. If these two P values are very different and you think your experiment may have violated the circularity assumption, use a more advanced program that can apply complicated methods (Huynh&Feldt or Box) that correct for violations of circularity more precisely.

You only have to worry about the assumption of circularity and the Geisser and Greenhouse corrected P value when you perform a repeated measures experiment, where each row of data represents repeated measurements from a single subject. If you performed a randomized block experiment, where each row of data represents data from a matched set of subjects, use the standard ANOVA P value and ignore the corrected P value.

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 tests the null hypothesis that the row means are identical (this is the test for effective matching). In both cases, the F ratio is expected to be near 1.0 if the null hypotheses are true. If F is large, the P value will be small.

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 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.

InStat 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.

Post tests

Interpret post tests following repeated measures ANOVA the same as regular ANOVA. See "Post test for linear trend" on page 76, and "Other post tests" on page 77.

The results of a Kruskal-Wallis test

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 (InStat cannot help you answer them):

Question	Discussion
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. There is no way for InStat to 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 triplicates from one animal to be high or low. See ″The need for independent samples“ on page 11.
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 nongaussian 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, InStat (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 logs or reciprocals) to create a Gaussian distribution and then using ANOVA. See "Transforming data to create a Gaussian distribution" on page 21.
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.

Approach to interpreting 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 a coincidence. This doesn't mean that every group differs from every other group, only that at least one group differs from the others. Then look at post tests to see which group(s) differ from which other group(s).

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 to EACH comparison but rather to the ENTIRE family of comparisons.

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 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 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, InStat 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 all the groups. InStat 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 larger value of the Kruskal-Wallis statistic corresponds to a larger 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 value of the Kruskal-Wallis statistic as high (or higher) as observed in this experiment.

If your samples are small, InStat calculates an exact P value. If your samples are large, it approximates the P value from the chi-square distribution. The approximation is quite accurate with large samples. With medium size samples, InStat can take a long time to calculate the exact P value. You can interrupt the calculations if an approximate P value is good enough for your purposes.

Post tests following the Kruskal-Wallis 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, InStat 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 NJ Castellan, 1988. The original reference is O.J. Dunn, Technometrics, 5:241-252, 1964.

InStat 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.

The results of a Friedman test

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:

Question	Discussion
Was the matching effective?	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 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 pairing should be part of the experimental design and not something you do after collecting data. InStat 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 effect can affect values in more than one row. There is no way for InStat to 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 nongaussian 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, InStat (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 logs or reciprocals) to create a Gaussian distribution and then using repeated measures ANOVA.

Approach to interpreting 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. (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 coincidences of random sampling, and conclude instead that at least one of the treatments (columns) differs from the rest. Then look at post tests to see which group(s) differ from which other group(s).

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 evidence that they differ. If you have small samples, Friedman’s test has little power.

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. InStat 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. 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, InStat 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, InStat can take a long time to calculate the exact P value. You can interrupt the calculations if an approximate P value is close enough.

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, InStat 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 NJ Castellan, 1988. The original reference is O.J. Dunn, Technometrics, 5:241-252, 1964.

InStat 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 Friedman test, and don't give it an exact name.

Contingency tables

Creating contingency tables

Use contingency tables to display the results of five kinds of experiments.

Term	Design of experiment and arrangement of data
Cross-sectional study	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 who has 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.
Prospective study	Use two samples 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 and tabulate the numbers that get leukemia. Subjects that get leukemia are tabulated in one column; the rest are tabulated in the other column.
Retrospective case-control study	Use two samples of subjects selected based on the outcome variable. 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.

Experiment	Use 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.
Assess accuracy of diagnostic test	Select two samples of subjects. One sample has the disease or condition you are testing for, the other does not. Then perform the test on all subjects and tabulate positive test results in one column and negative test results in the other.

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

Here is an example contingency table. Subjects with HIV infection were divided into two groups and given placebo or AZT. The result was recorded as disease progression or no progression (from New Eng. J. Med. 329:297-303, 1993).

	Disease progression	No progression	Total
AZT	76	399	475
Placebo	129	332	461
Total	205	731	936

The values in a contingency table represent the number of subjects actually observed in this experiment. Tables of averages, percentages or rates are not contingency tables. Note also that the columns are mutually exclusive. A subject can be in one or the other, but not both. The rows are also mutually exclusive.

Results of contingency table analyses

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:

Question	Discussion
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. There is no way for InStat to 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 (not offered by InStat).
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. Data like this should be analyzed by special methods such as McNemar's test. Paired data should not be analyzed by chi-square or Fisher's test.
Is your table really a contingency table?	To be a contingency table, the values 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. InStat 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, InStat 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), InStat 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.

Interpreting relative risk, odds ratio, P1-P2, etc.

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

Relative risk

The relative risk is the proportion of subjects in the top row who are in the left column divided by the proportion of subjects in the bottom row who are in the left column. For the AZT example, the relative risk is 16%/28% =0.57. A subject treated with AZT has 57% the chance of disease progression as a subject treated with placebo. The word "risk" is appropriate in some studies, but not others. Think of the relative risk as being simply the ratio of proportions. InStat also reports the 95% confidence interval for the relative risk, calculated by the approximation of Katz. For the example, the 95% confidence interval ranges from 0.4440 to 0.7363. You can be 95% certain that this range includes the true population relative risk.

P1-P2

You can also summarize the results by taking the difference of the two proportions. In the example, the disease progressed in 28% of the placebo-treated patients and in 16% of the AZT-treated subjects. The difference is 28% - 16% = 12%. InStat also reports an approximate 95% confidence interval (unless the sample sizes are very small). For the example, the confidence interval ranges from 6.68% to 17.28%.

Odds ratio

When analyzing case-control retrospective studies, you cannot meaningfully calculate the difference between proportions or the relative risk. The best way to summarize the data is via an odds ratio. In most cases, you can think of an odds ratio as an approximate relative risk. So if the odds ratio equals 4, the disease occurs four times as often in people exposed to the risk factor as in people not exposed.

Sensitivity, specificity, and predictive values

Term	Meaning
Sensitivity	The fraction of those with the disease correctly identified as positive by the test.
Specificity	The fraction of those without the disease correctly identified as negative by the test.
Positive predictive value	The fraction of people with positive tests who actually have the condition.
Negative predictive value	The fraction of people with negative tests who actually don't have the condition.
Likelihood ratio	If you have a positive test, how many times more likely are you to have the disease? If the likelihood ratio equals 6.0, then someone with a positive test is six times more likely to have the disease than someone with a negative test. The likelihood ratio equals sensitivity/(1.0-specificity).

The sensitivity, specificity and likelihood ratios are properties of the test. The positive and negative predictive values are properties of both the test and the population you test. If you use a test in two populations with different disease prevalence, the predictive values will be different. A test that is very useful in a clinical setting (high predictive values) may be almost worthless as a screening test. In a screening test, the prevalence of the disease is much lower so the predictive value of a positive test will also be lower.

Interpreting P values from analyses of a 2x2 contingency table

If you set up the contingency table to evaluate the accuracy of a diagnostic test, the most important results will be the sensitivity, specificity and predictive power (see page 93), and you'll probably ignore the P value. In other situations, you'll be interested both in the P value and the confidence interval for the relative risk, odds ratio, or P1-P2.

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

If the P value is small, then it is unlikely that the association you observed is due to a coincidence of 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. InStat 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 rows and columns are associated, 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

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. InStat 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	Large	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.

Large	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.

Interpreting analyses of larger contingency tables

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

Chi-square test for independence

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.

InStat 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-squared 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.

Chi-square test for trend

The P value from the 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 coincidence 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.