

GraphPad

InStat®

Version 3.0

**The InStat
guide to
choosing and
interpreting
statistical tests**

Harvey Motulsky

© 1990-2003, GraphPad Software, Inc. All rights reserved.

Program design, manual and help screens: Dr. Harvey J. Motulsky
Paige Searle

Programming: Mike Platt
John Pilkington
Harvey Motulsky

InStat and GraphPad Prism are registered trademarks of GraphPad Software, Inc.

How to cite GraphPad InStat

When citing analyses performed by the program, include the name of the analysis, and InStat version number (including the second number after the decimal point). Use this example as a guide: "One-way ANOVA with Dunnett's post test was performed using GraphPad InStat version 3.05 for Windows 95, GraphPad Software, San Diego California USA, www.graphpad.com".

To find the full version number, pull down the Help menu (Windows), or InStat menu (Mac OS X), or a Apple menu (Mac OS 8-9, and choose About GraphPad InStat.)

If you want to cite this manual rather than the program itself, use this example as a guide: "GraphPad Software, InStat guide to choosing and interpreting statistical tests, 1998, GraphPad Software, Inc., San Diego California USA, www.graphpad.com". It makes sense to include the web address, since this manual is available on-line, but not in libraries.

System requirements

The Windows version requires Windows 95 or higher, 4Mb RAM, 3Mb free hard disk space. The Macintosh version requires OS 8.6 or higher (including OS X) and 4Mb free hard disk space. InStat 3.0 Mac is a native OS X (carbon) application. If you run InStat under Mac system 8.6-9.2, you must have CarbonLib (version 1.5 or later) installed on your system. You can find CarbonLib at www.graphpad.com.

Contents

Welcome to InStat	7
The InStat approach	7
What InStat does	7
How to start	8
Introduction to statistical principles	9
When do you need statistical calculations?	9
The key concept: Sampling from a population.....	9
The need for independent samples.....	10
How statistics can extrapolate from sample to population	11
Confidence intervals	11
P values	12
Hypothesis testing and statistical significance.....	14
The Gaussian distribution and testing for normality	17
What is the Gaussian distribution?	17
What's so special about the Gaussian distribution?	17
Nonparametric tests	17
Transforming data to create a Gaussian distribution	19
Testing for normality	19
Tutorial: The InStat approach	21
Step 1. Choose data format	21
Step 2. Enter data	23
Step 3. Summary statistics.....	24
Step 4. Select a statistical test.....	25
Step 5. View the results.....	26
Step 6. On your own.....	26
Descriptive statistics	27
Column statistics.....	27
Entering averaged data into InStat.....	28
One sample tests	30
Introduction to one sample tests.....	30
Choosing the one-sample t test or Wilcoxon test	30
The results of a one-sample t test	31
The results of a Wilcoxon test.....	35
Comparing two groups (t tests etc.)	38
Introduction to t tests	38
Entering t test data into InStat	38
Choosing a test to compare two columns	39
The results of an unpaired t test	41
How to think about results from an unpaired t test	43
The results of an unpaired t test, line by line.	45

The results of a paired t test	47
How to think about results of a paired t test	48
The results of a paired t test, line by line.....	50
The results of a Mann-Whitney test	52
How to think about the results of a Mann-Whitney test.....	53
How the Mann-Whitney test works.....	54
The results of a Wilcoxon test.....	54
How to think about the results of a Wilcoxon test	56
How the Wilcoxon matched pairs test works	56
Comparing three or more groups (one-way ANOVA, etc.)	58
Introduction to ANOVA.....	58
Entering ANOVA data into InStat	58
Choosing a one-way ANOVA analysis	59
The results of one-way ANOVA	62
How to think about results from one-way ANOVA.....	64
Results of one-way ANOVA. Line by line.	66
Post tests (one-way ANOVA).....	68
The results of repeated measures ANOVA	69
How to think about results from repeated measures one-way ANOVA.....	70
The results of repeated measures ANOVA, line by line.....	71
The results of a Kruskal-Wallis test.....	72
Approach to interpreting the results of a Kruskal-Wallis test	73
How the Kruskal-Wallis test works	74
Post tests following the Kruskal-Wallis test	74
The results of a Friedman test	75
Approach to interpreting the results of a Friedman test.....	76
How the Friedman test works.....	76
Post tests following the Friedman test	77
Contingency tables	78
Creating contingency tables.....	78
Analysis choices for contingency tables.....	79
Results of contingency table analyses	81
Linear regression and correlation	87
Introduction to linear regression and correlation.....	87
Entering data for correlation and linear regression.....	88
Choosing linear regression or correlation	88
Results of correlation	90
How to think about results of linear correlation	91
Correlation results line by line.....	91
Results of linear regression	92
How to think about the results of linear regression	93
Linear regression results line by line.....	93
Reading unknowns from standard curves.....	95
Multiple Regression and correlation	97
Introduction to multiple regression and correlation.....	97
Entering data for multiple regression and correlation.....	99
Analysis choices for multiple regression and correlation	100
Interpreting a correlation matrix	100

The results of multiple regression	101
Using InStat	109
Online Help	109
Importing and exporting data	109
Working with the data table	112
Arranging data	116
Selecting columns to analyze	116
After you view the results	117
InStat files	119
GraphPad Software	120
Technical support	120
GraphPad Prism	121
Index	123

Welcome to InStat

The InStat approach

GraphPad InStat is designed to help the experimental or clinical scientist analyze small amounts of data. Although InStat can do some data manipulation and selection, it is not designed to manage a large database with many variables. InStat works best when you have a single question in mind. Enter the data to answer that question, and InStat guides you to pick the right test to answer it:

- The first step – even before entering or importing data – is to tell InStat what kind of data you wish to enter. InStat will then present you with a data table for that kind of data.
- The next step is to enter data. You can type in numbers, paste from the clipboard, or import a text file.
- InStat then asks you several questions to choose a test. On-line help (or this manual) can help you answer those questions.
- Finally InStat presents the results, avoiding jargon when possible. On-line help, and this manual, can help you interpret the values.

This manual provides a comprehensive explanation of all of InStat's features. In addition, this guide will help you review statistical principles, pick an appropriate test, and interpret the results.

What InStat does

InStat calculates these statistical tests:

Category	Tests that InStat performs
Column statistics	Mean, median, 95% CI, SD, SEM. Also tests whether the distribution conforms to a Gaussian distribution using the Kolmogorov-Smirnov test.
Group comparisons	Paired and unpaired t tests; Mann-Whitney and Wilcoxon nonparametric tests. Ordinary and repeated measures ANOVA followed by Bonferroni, Tukey, Student-Newman-Keuls or Dunnett post tests. Kruskal-Wallis or Friedman nonparametric tests followed by Dunn post test.
Contingency tables	Chi-square test (with or without Yates' correction). Fisher's exact test. Calculate 95% confidence interval for the difference of two proportions, relative risk, odds ratio, sensitivity or specificity. Chi-square test for trend.
Linear regression and correlation	Linear regression, optionally forcing the line through a defined point. Determine new points along the standard curve. Pearson linear correlation and Spearman nonparametric correlation.
Multiple regression	Determine the best linear equation that fits Y to two or more X variables.

InStat is not for everyone. It performs basic tests easily, but does not handle advanced statistical tests. For example, InStat does not perform two-way (or higher) ANOVA, logistic regression, the Mantel-Haenszel test (to analyze a stack of contingency tables), stepwise multiple regression, analyses of survival curves, analysis of covariance, factor or cluster analysis, polynomial regression or nonlinear regression.

Please note that our scientific graphics and analysis program, GraphPad Prism, can perform polynomial and nonlinear regression, analyze survival curves and perform two-way ANOVA. For more information, contact GraphPad Software or visit our web page at www.graphpad.com. Do you have the current version?

Like all software companies, GraphPad occasionally issues minor updates to Prism. If you are having trouble with InStat, check that you are running the current release.

The full version number is not on the manual cover or the CD label. You have to run the program and find out which version it is. Drop the Help menu (Windows), Apple menu (Mac OS8-9) or Prism menu (Mac OS X) and choose About InStat. Windows versions have two digits after the decimal point (i.e. 3.05). Mac versions have a single digit after the decimal followed by a letter (i.e. 3.0a).

Go to the Support page at www.graphpad.com to find out what version is most current. Download and install the updater if your version is not the most current. Updates (interim versions of GraphPad software containing bug fixes or minor improvements) are free to owners of the corresponding major releases. In contrast, upgrades (a new version with many new features) must be purchased.

How to start

There are three ways to proceed:

- Learn to use InStat systematically by carefully reading “Tutorial: The InStat approach” on page 21, and then browsing "Using InStat" on page 109.
- Review the principles of statistics before using the program by reading this manual from start to finish.
- Simply plunge in! Start using InStat, and consult the manual or help screens when you have questions. The InStat Guide (page 109) will help you learn the program quickly. You can complete your first analysis in just a few minutes.

Like any tool, data analysis programs can be misused. InStat won't be helpful if you designed the experiment badly, entered incorrect data or picked an inappropriate analysis. Heed the first rule of computers: *Garbage in, garbage out.*

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 17), 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.

- 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.

- Define the null hypothesis. If you are comparing two means, the null hypothesis is that the two populations have the same mean.
- Do the appropriate statistical test to compute the P value.
- Compare the P value to the preset threshold value.
- If the P value is less than the threshold, state that you "reject the null hypothesis" and that the difference is "statistically significant".
- If the P value is greater than the threshold, state that you "do not reject the null hypothesis" and that the difference is "not statistically significant". You cannot conclude that the null hypothesis is true. All you can do is conclude that you don't have sufficient evidence to reject the null hypothesis.

Statistical significance

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.

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 than 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 focused 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.

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 19.

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 19). 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	<u>Robust</u> . P value will be nearly correct even if population is fairly far from Gaussian.	<u>Not robust</u> . If the population is not Gaussian, the P value may be misleading.
Nonparametric test	<u>Powerful</u> . 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.	<u>Not powerful</u> . 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	<u>Useful</u> . Use a normality test to determine whether the data are sampled from a Gaussian population.	<u>Not very useful</u> . 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 distribution 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.

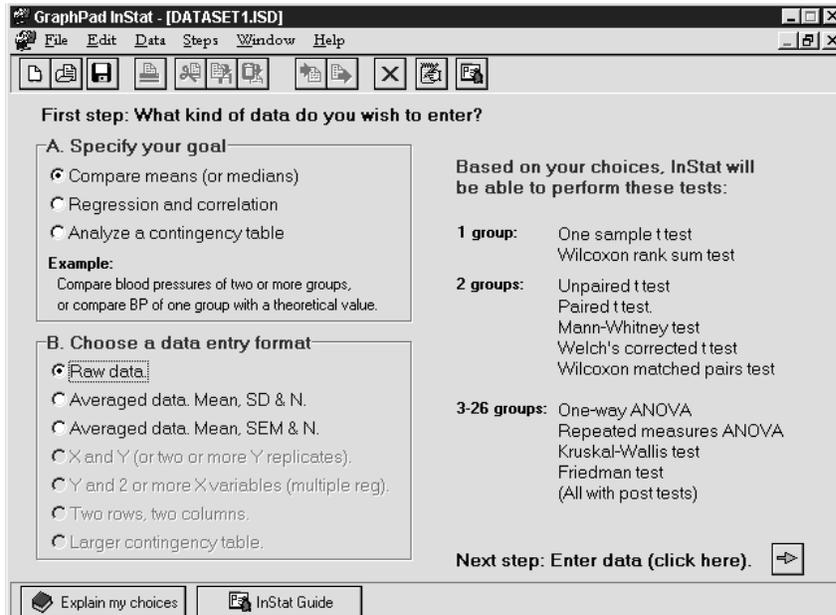
P value	Sample size	Conclusion
Small	Any	The data failed the normality test. You can conclude that the population is unlikely to be Gaussian.
Large	Large	The data passed the normality test. You can conclude that the population is likely to be Gaussian, or nearly so. How large does the sample have to 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.

Tutorial: The InStat approach

After installing InStat, the easiest way to learn InStat is to follow a simple example. In this example, we'll perform an unpaired t test. It will take just a few minutes. The screen shots are for Windows, but the Mac is very similar.

Step 1. Choose data format

Launch InStat by double clicking on its icon. You'll see this screen:

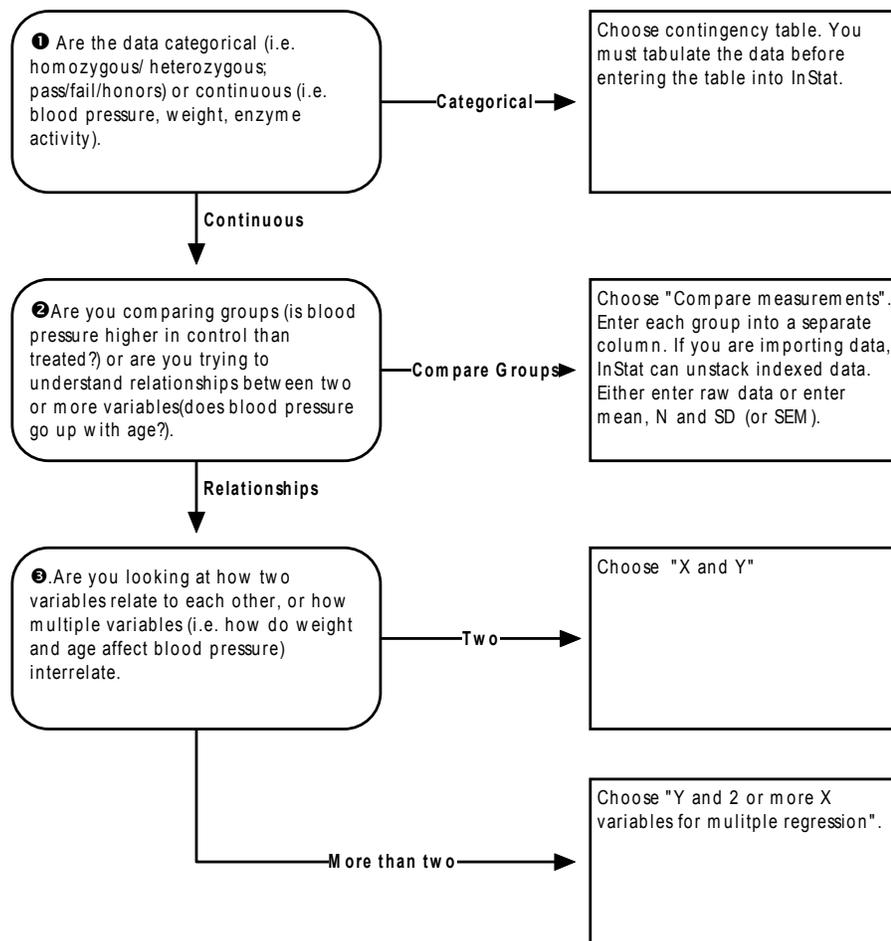


You may see the InStat Guide superimposed on this screen. While following this written tutorial, you may wish to turn off the Guide window (uncheck the box "Keep showing the InStat Guide", then press Close). The Guide is always on in the demo version. In the full version of the program, you decide when to show the Guide. Bring it back by selecting InStat Guide from the Help menu.

Before you can enter data, you first have to tell InStat what kind of data table you need. This important step makes InStat unique. Once you've chosen the right kind of data table for your experiment, InStat will be able to guide you to choose an appropriate statistical test.

InStat offers three goals on the top left of the screen, with more choices below. Based on your choices, you'll be able to perform different tests as shown on the right.

The three goals are distinct, and you shouldn't have a problem telling InStat what kind of data you have. Follow the logic of this flowchart:

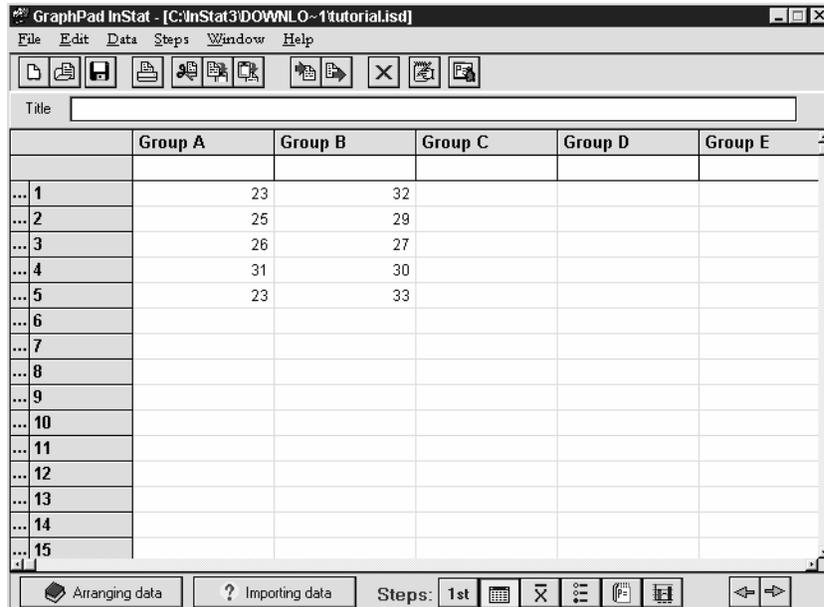


Choose “Compare means (or medians)” and “Raw data”. Then click the arrow button in the lower right of the screen to move to the next step.

Step 2. Enter data

Enter the values shown here.

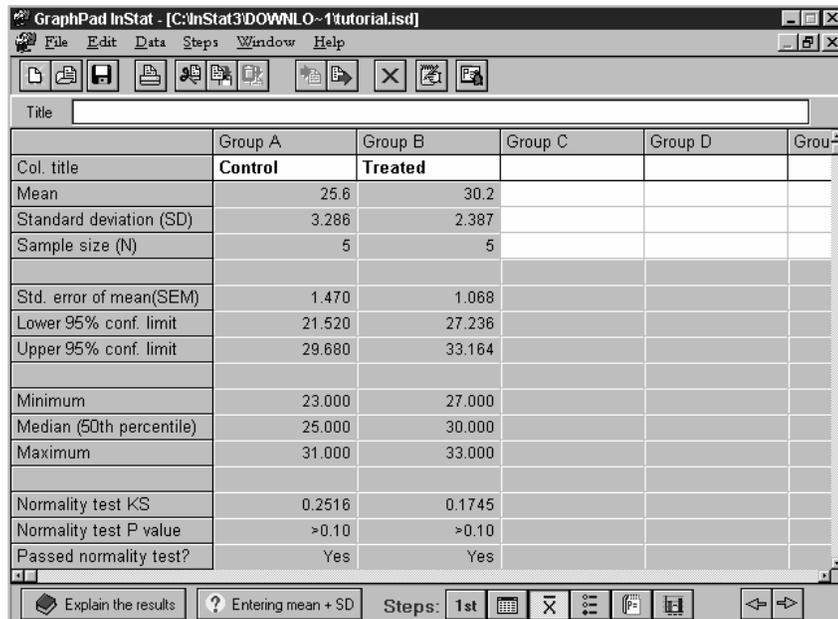
Group A	Group B
23	32
25	29
26	27
31	30
23	33



Press the buttons at the lower left of the window to learn how to arrange your data (important with InStat) or how to import data (including stacked or indexed data). As you move from step to step, these buttons will provide help on different topics.

Click the blue right arrow button (lower right of window) to go to the next step.

Step 3. Summary statistics



The screenshot shows the GraphPad InStat software interface. The main window displays a table of summary statistics for two groups: Control and Treated. The table includes columns for Col. title, Group A, Group B, Group C, Group D, and Group E. The data is as follows:

Col. title	Group A	Group B	Group C	Group D	Group E
Mean	25.6	30.2			
Standard deviation (SD)	3.286	2.387			
Sample size (N)	5	5			
Std. error of mean(SEM)	1.470	1.068			
Lower 95% conf. limit	21.520	27.236			
Upper 95% conf. limit	29.680	33.164			
Minimum	23.000	27.000			
Median (50th percentile)	25.000	30.000			
Maximum	31.000	33.000			
Normality test KS	0.2516	0.1745			
Normality test P value	>0.10	>0.10			
Passed normality test?	Yes	Yes			

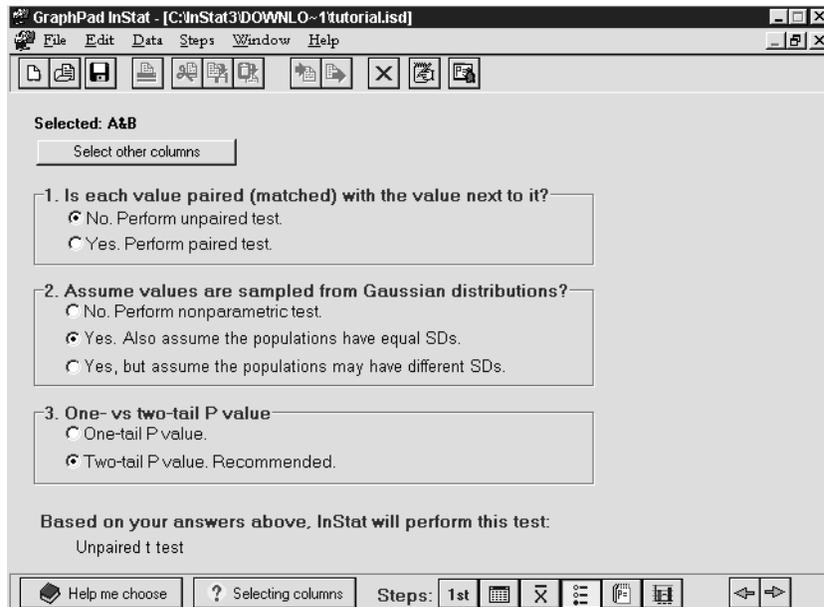
The software interface also includes a menu bar (File, Edit, Data, Steps, Window, Help), a toolbar with various icons, and a status bar at the bottom with buttons for "Explain the results", "? Entering mean + SD", and "Steps: 1st".

The summary statistics screen shows the mean, SD, SEM, confidence interval, etc. for each column. You can also enter data here if you have calculated the mean and SD (or SEM) in another program.

Click "Explain the results" for a definition of the statistical terms and a discussion of how they are used.

Go to the next step.

Step 4. Select a statistical test



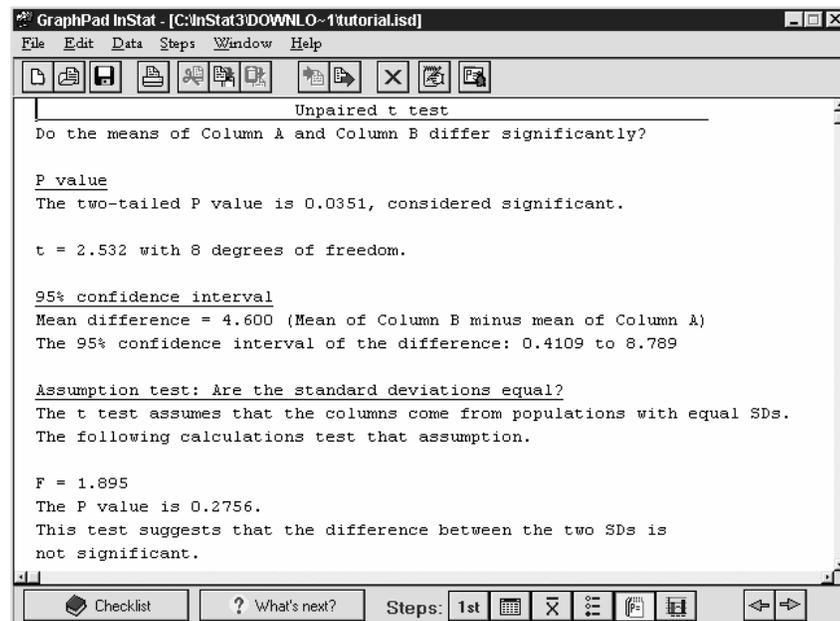
For this example, InStat presents questions necessary to choose an analysis to compare two columns of data. You'd see different choices if you entered a different number of columns of data, or if you created a different kind of data table.

Select an unpaired test, assuming the populations are Gaussian with equal standard deviations, and a two-tail P value.

If you are unsure of the choices, InStat can help you understand these questions. Press "Help me choose" for guidance.

Go to the next step.

Step 5. View the results



InStat presents the results using as little statistical jargon as possible. Of course you can print the results or export them to another program.

Press the "Checklist" button to confirm that you picked an appropriate test and to understand the results.

You don't have to follow InStat's steps in order. Use one of the six step buttons (lower right) to jump from step to step.



Step 6. On your own

Press the last step button to see a notebook quality graph, suitable for getting the sense of your data and to spot errors in data entry. You cannot change or edit the graph, but can print it or copy it to the clipboard.

You don't have to follow the steps in order. Click the data table button to go back to the data and change one or more of the values. Now click the results step button ("P=") to view the results. Note that InStat instantly recomputed the results to correspond with the new data.

An InStat file consists of one data table and one analysis. Click "What's next" to learn how to manage multiple analyses.

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	<p>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.</p> <p>InStat calculates the "sample SD" (which uses a denominator of N-1), not the "population SD" with a denominator of N.</p> <p>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.</p>
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 19.

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.

One sample tests

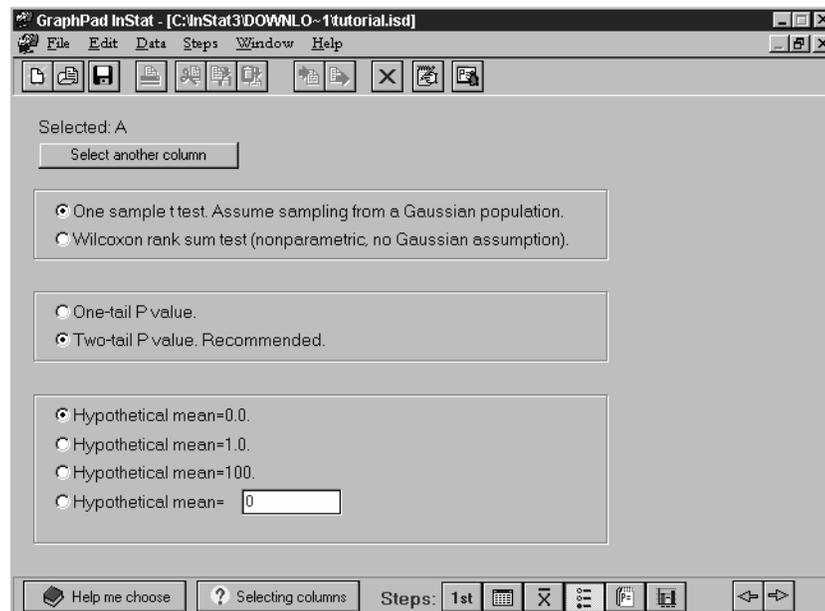
Introduction to one sample tests

The one sample t test (and nonparametric Wilcoxon test) tests whether the mean (median) of a sample differs significantly from a value set by theory.

Choosing the one-sample t test or Wilcoxon test

If you entered a single column of data, InStat will present the choices for comparing one column of data to a hypothetical mean or median.

If you entered more columns, InStat will present the choices for comparing two columns or comparing three or more columns. If this happens, click the button “select other columns” that appears over all the choices. Uncheck all but one of the columns, and InStat will present the choices for analyzing one column of data.



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:

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 17.

One- or two-tailed P value?

If in doubt, choose a two-tail P value. See "One- vs. two-tail P values" on page 13.

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:

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 19. 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 19). 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 10.

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 13.

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: $t = (\text{Sample Mean} - \text{Hypothetical Mean}) / \text{SEM}$

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 19.

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):

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 10.

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 13.

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.

Do not enter indexed data

InStat expects you to enter data in a format that is natural to many scientists. For example, to compare the blood pressure of a group of men and a group of women with InStat, enter the men's blood pressure in one column and the women's blood pressure in another.

Some other statistics programs expect you to arrange data differently, putting all of the data into one column and using another column to define group. For the blood pressure example, you would enter all the blood pressure values (for both groups) in one column. In another column you

would enter a code or index (perhaps 1 for men and 2 for women). Don't arrange data like this when using InStat. If you have data files arranged like this (sometimes called indexed or stacked), InStat can import them, automatically rearranging the values. See "Importing indexed data" on page 110.

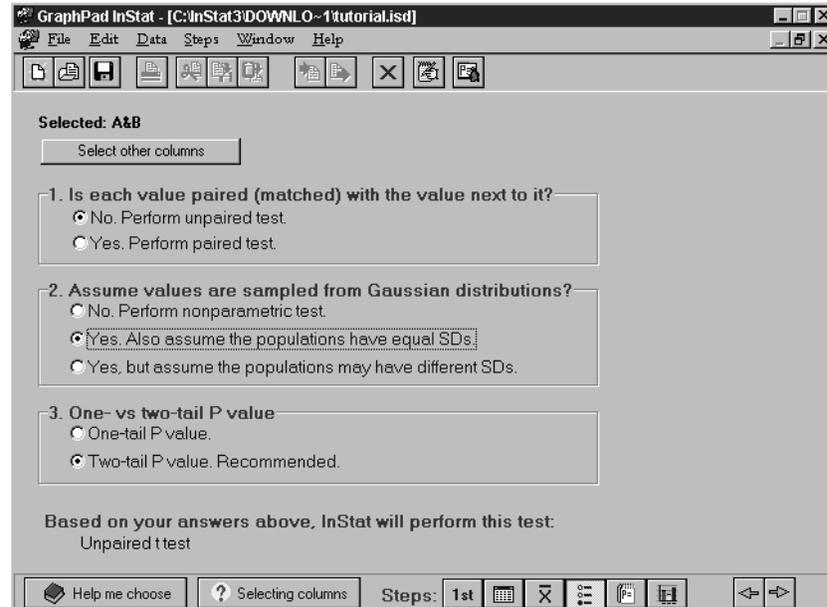
Consider transforming the data

Before comparing columns, consider whether you should first transform the values. The t test assumes that your data are sampled from a population that follows 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. See "Transforming data to create a Gaussian distribution" on page 19.

If you know the distribution of your population, transforming the values to create a Gaussian distribution is a good thing to do, as it lets you use a t test, which has more power than a nonparametric test.

If you plan to use a nonparametric test, transforming the data will make no difference.

Choosing a test to compare two columns



InStat can perform paired and unpaired t tests, and the nonparametric Mann-Whitney and Wilcoxon tests. To choose between these tests, you must answer four questions:

Are the data paired?

Choose a paired test when the experiment follows one of these designs:

- You measure a variable before and after an intervention in each subject.
- You recruit subjects as pairs, matched for variables such as age, ethnic group or disease severity. One of the pair gets one treatment; the other gets an alternative treatment.
- You run a laboratory experiment several times, each time with a control and treated preparation handled in parallel.
- You measure a variable in twins, or child/parent pairs.

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

Ideally, 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 17. The results of a normality test can be helpful, but not always as helpful as you’d hope. See “Testing for normality” on page 19

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 13.

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

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 19. 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 19). 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 46.

Questions about experimental design

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 10.

Are you comparing exactly two groups?

Use the t test only to compare two groups. To compare three or more groups, use one-way 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 30.

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 13.

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 12.

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 19.
- 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 19.

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

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 19). 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 91. 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

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 10.

Are you comparing exactly two groups?

Use the t test only to compare two groups. To compare three or more matched groups, use repeated measures one-way 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 13.

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 12.

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 19.

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):

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 10.

Are the data unpaired?

The Mann-Whitney test works by ranking all the values from low to high, and comparing the mean rank in the two groups. If the data are paired or matched, then you should choose a Wilcoxon test instead.

Are you comparing exactly two groups?

Use the Mann-Whitney test only to compare two groups. To compare three or more groups, use the Kruskal-Wallis test followed by post tests. It is not appropriate to perform several Mann-Whitney (or t) tests, comparing two groups at a time.

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 13.

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 19).

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:

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 10. Are the pairs independent?

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 90. 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 13.

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 19.

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.

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.

Do not enter indexed data

InStat expects you to enter data in a format that is natural to many scientists. For example, to compare the blood pressure of three groups with InStat, enter the men's blood pressure in one column and the women's blood pressure in another.

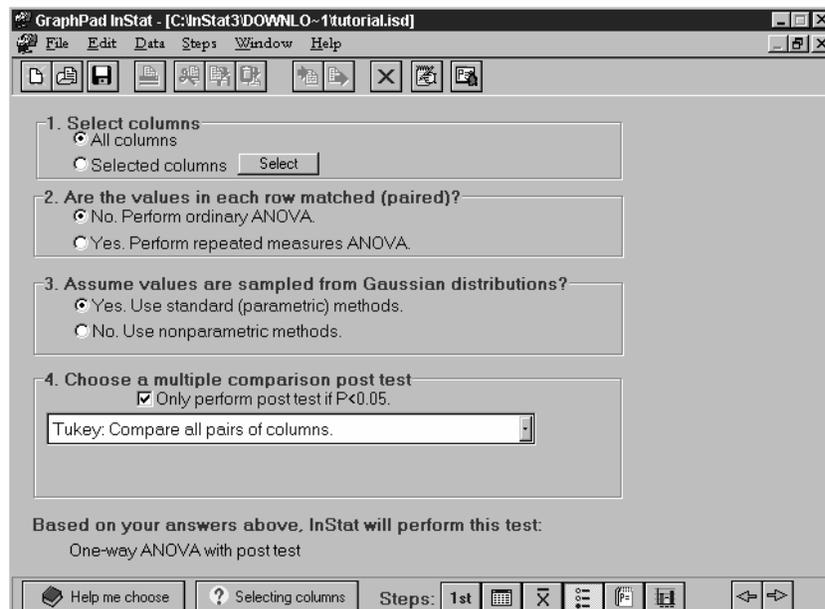
Some other statistics programs expect you to arrange data differently, putting all of the data into one column and using another column to define group. Don't arrange data like this when using InStat. If you have data files arranged like this (called indexed or stacked format), InStat can import them, automatically rearranging the values. See "Importing indexed data" on page 110.

Consider transforming the data

Before comparing columns, consider whether you should first transform the values. ANOVA assumes that your data are sampled from populations that follow Gaussian distributions. If your data do not follow a Gaussian (normal) distribution, you may be able to transform the values to create a Gaussian distribution. See "Transforming data to create a Gaussian distribution" on page 19.

If you know the distribution of your population, transforming the values to create a Gaussian distribution is a good thing to do, as it lets you use ANOVA, which has more power than a nonparametric test.

Choosing a one-way ANOVA analysis



InStat can perform ordinary one-way ANOVA, repeated measures ANOVA and the nonparametric tests of Kruskal-Wallis and Freidman. To choose among these tests, you must answer three questions:

Are the data matched?

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

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

The term *repeated measures* applies strictly only to the first example – you are giving treatments repeatedly to one subject. The other two examples are called *randomized block* experiments (each set of subjects is called a *block* and you randomly assign treatments within each block). The analyses are identical for repeated measures and randomized block experiments, and InStat always uses the term repeated measures.

Ideally, you should decide about matching 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 postal code, but it is not okay to match based on blood pressure.

Assume sampling from a Gaussian distribution?

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 17. The results of a normality test can be helpful, but not always as helpful as you’d hope. See “Testing for normality” on page 19.

Which post test?

If you are comparing three or more groups, you may pick a post test to compare pairs of group means. It is not appropriate to repeatedly use a t test to compare various pairs of columns (see “Beware of multiple comparisons” on page 15). InStat offers these choices of post test.

- No post test.
- Bonferroni. Compare selected pairs of columns.

- Bonferroni. Compare all pairs of columns.
- Tukey. Compare all pairs of columns.
- Student-Newman-Keuls. Compare all pairs of columns.
- Dunnett. Compare all vs. control.
- Test for linear trend between column mean and column number.

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

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

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

Most often, you will want to compare all pairs of columns. InStat offers you three choices. The only advantage of the **Bonferroni** method is that it is easy to understand. Its disadvantage is that it is too conservative, leading to P values that are too high and confidence intervals that are too wide. This is a minor concern when you compare only a few columns, but is a major problem when you have many columns. Don't use the Bonferroni test with more than five groups.

Choosing between the **Tukey** and **Newman-Keuls** test is not straightforward, and there appears to be no real consensus among statisticians. The two methods are related, and the rationale for the differences is subtle. The methods are identical when comparing the largest group mean with the smallest. For other comparisons, the Newman-Keuls test yields lower P values. The problem is that it is difficult to articulate exactly what null hypotheses the P values test. For that reason, and because the Newman-Keuls test does not generate confidence intervals, we suggest selecting Tukey's test.

Summary of tests to compare three or more columns

Based on your answers...		...InStat chooses a test
Not matched	Gaussian distribution	Ordinary one-way ANOVA
Matched	Gaussian distribution	Repeated measures one-way ANOVA
Not matched	Not Gaussian	Kruskal-Wallis test
Matched	Not Gaussian	Friedman test

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

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 19. 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 19). 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 67.

Questions about experimental design

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 10.

Do you really want to compare means?

One-way ANOVA compares the means of three or more groups. It is possible to have a tiny P value – clear evidence that the population means are different – even if the distributions overlap considerably. In some situations – for example, assessing the usefulness of a diagnostic test – you may be more interested in the overlap of the distributions than in differences between means.

Is there only one factor?

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

Some experiments involve more than one factor. For example, you might compare three different drugs in men and women. There are two factors in that experiment: drug treatment and gender. These data need to be analyzed by two-way ANOVA, also called two factor ANOVA. 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 68.

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 12.

R² value

This is the fraction of the overall variance (of all the data, pooling all the groups) attributable to the difference between 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 R² 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 67). Some programs (and books) don't bother reporting this value. Others refer to it as η^2 (eta squared) rather than R². 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 19.
- 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 "r ² " on page 92.
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.

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 10.

Is the random variability distributed according to a Gaussian distribution?

Repeated measures ANOVA assumes that each measurement is the sum of an overall mean, a treatment effect (the 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 64.

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 effect 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 68, and "Other post tests" on page 69.

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):

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 10.

Are the data unpaired?

If the data are paired or matched, then you should consider choosing the Friedman test instead. If the pairing is effective in controlling for experimental variability, the Friedman test will be more powerful than the Kruskal-Wallis test.

Are the data sampled from 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 19.

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:

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.

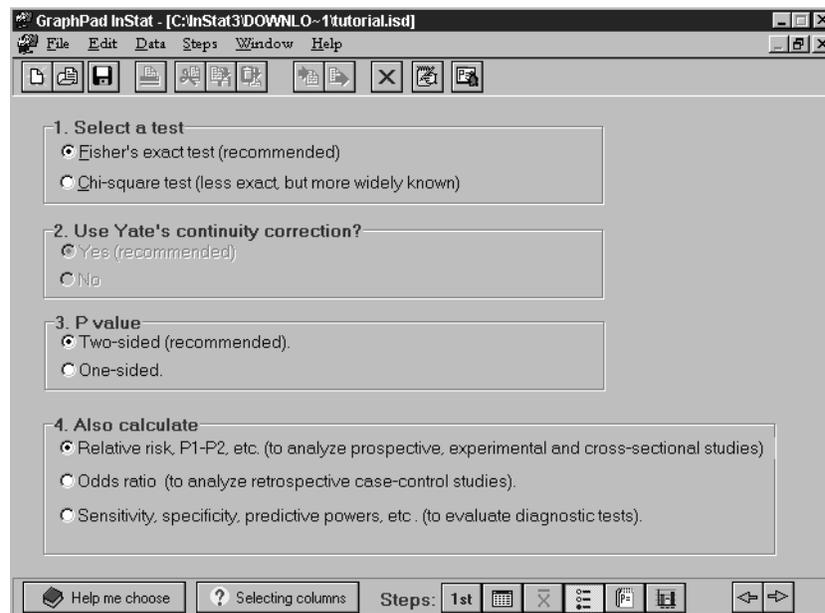
Here is an example contingency table. Subjects with HIV infection were divided into two groups and given placebo or AZT. The result was recorded as disease progression or no progression (from New Eng. J. Med. 329:297-303, 1993).

	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.

Analysis choices for contingency tables

Tables with two rows and two columns



InStat offers two methods for calculating a P value from tables with two rows and two columns: Fisher's exact test and the chi-square test. We recommend always picking Fisher's test, as it calculates a P value that is

exactly correct. The only advantage of the chi-square test is that it is easier to calculate by hand, and so is better known. We don't recommend it.

If you choose a chi-square test, also choose whether to apply Yates' continuity correction. This correction is designed to make the approximate results from a chi-square test more accurate with small samples.

Statisticians disagree about whether to use it. If you always select Fisher's exact test (recommended), Yates' correction is of no concern.

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

Choose a two-sided P value, unless you have a good reason to pick a one-sided P value. (With contingency tables, InStat refers to "two-sided" P values rather than "two-tail P value" -- the distinction is subtle and not worth worrying about.) See "One- vs. two-tail P values" on page 13.

In addition to calculating a P value, InStat can summarize your data and compute a confidence interval. There are many ways to summarize the results of a contingency table. Your choice depends on your experimental design.

Choice	Type of experiment	How to arrange data
Relative risk, P1-P2, etc.	Prospective and experimental studies	The top row is for exposure to risk factor or treatment; the bottom row is for controls. The left column tabulates the number of individuals with disease; the right column is for those without the disease.
Odds ratio	Case-control retrospective studies	The left column is for cases; the right column is for controls. The top row tabulates the number of individuals exposed to the risk factor; the bottom row is for those not exposed.
Sensitivity, specificity, etc.	Determining the accuracy of a diagnostic test	The left column is for people who do have the condition being tested for, and the right column is for people who don't have that condition. Use an established test (or the test of time) to make this decision. Use the top row to tabulate the number of individuals with a positive test result and the bottom row to tabulate the number of individuals with a negative test result.

Contingency tables with more than two rows or columns

If your table has more than two rows or two columns, skip over the choose test step (which will be unavailable). InStat always calculates the chi-

square test. Although statisticians have developed tests analogous to Fisher's exact test for larger tables, InStat doesn't offer them. Yates' continuity correction is never used with larger tables.

If your table has two columns and three or more rows (or two rows and three or more columns), InStat will also perform the chi-square test for trend. This calculation tests whether there is a linear trend between row (column) number and the fraction of subjects in the left column (top row). This test only makes sense when the rows (columns) are arranged in a natural order (i.e. age, dose, time) and are equally spaced.

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:

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 (which InStat does not do, but GraphPad StatMate does). 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 83), 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.

Linear regression and correlation

Introduction to linear regression and correlation

Introduction to correlation

Correlation is used when you have measured two variables in each subject, and wish to quantify how consistently the two variables vary together. When the two variables vary together, statisticians say that there is a lot of covariation or correlation. The direction and magnitude of correlation is quantified by the correlation coefficient, r .

InStat calculates the correlation coefficient, r , and its 95% confidence interval. It also calculates a P value that answers this question: If the two variables really aren't correlated at all in the overall population, what is the chance that you would obtain a correlation coefficient as far from zero as observed in your experiment from randomly selected subjects?

Introduction to linear regression

Linear regression is used to analyze the relationship between two variables, which we will label X and Y. For each subject (or experimental unit), you know both X and Y and you want to find the best straight line through the data. In some situations, the slope and/or intercept have a scientific meaning. In other cases, you use linear regression to create a standard curve to find new values of X from Y, or Y from X.

InStat determines the best-fit linear regression line, including 95% confidence interval bands. You may force the line through a particular point (usually the origin), perform the runs test, and interpolate unknown values from a standard curve determined by linear regression. InStat also creates a notebook-quality graph of your data with the best-fit line. You can not customize this graph.

InStat cannot perform nonlinear or polynomial regression, but GraphPad Prism can (see page 120).

How does linear regression work?

Linear regression finds the line that best predicts Y from X. It does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line.

Why minimize the square of the distances? If the scatter of points around the line is Gaussian, it is more likely to have two points somewhat close to the line (say 5 units each) than to have one very close (1 unit) and one further (9 units). The total distance is 10 in each of those situations. The sum of the squares of the distances is 50 in the first situation and 81 in the second. A strategy that minimized the total distance would have no

preference between the two situations. A strategy that minimizes the sum of squares of the distances prefers the first situation, which is more likely to be correct.

Note that linear regression does not *test* whether your data are linear (except for the runs test). It assumes that your data are linear, and finds the slope and intercept that make a straight line come as close as possible to your data.

Entering data for correlation and linear regression

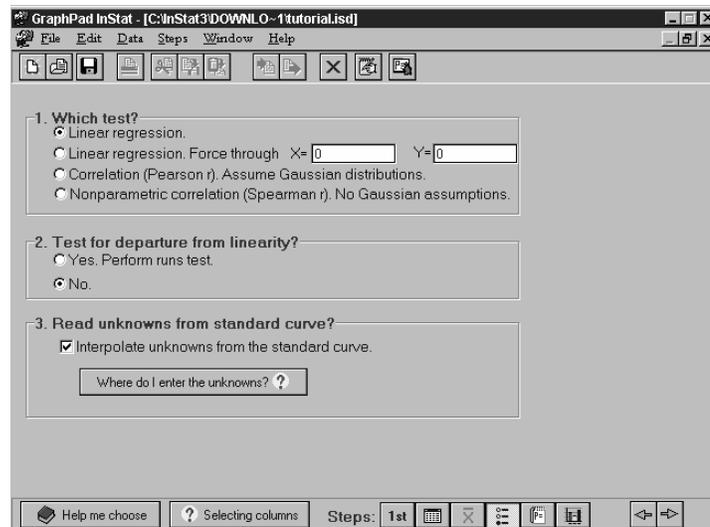
Enter X values into the first column (labeled X), and Y values into the second column (Y1). Only use the remaining columns if you have replicate Y values for each value of X (i.e. triplicate measurements of the same variable).

If you want to interpolate values from a standard curve, enter the unknowns directly below the standard curve. For the unknowns, enter X or Y (but not both). See "Reading unknowns from standard curves" on page 95.

If you want to look at the relationship of more than two variables (for example, if you want to look at how blood pressure is affected by both age and weight), format the data table for multiple regression (see "Introduction to multiple regression and correlation" on page 97) rather than linear regression.

Note that X and Y are asymmetrical for linear regression. If you switch X and Y you'll get a different best-fit regression line (but the same correlation coefficient).

Choosing linear regression or correlation



Regression or correlation?

Linear regression and correlation are related, but different, tests. Linear regression finds the line that best predicts Y from X. Correlation quantifies how well X and Y vary together. When choosing, consider these points:

- If you control X (i.e., time, dose, concentration), don't select correlation. Select linear regression.
- Only choose linear regression if you can clearly define which variable is X and which is Y. Linear regression finds the best line that predicts Y from X by minimizing the sum of the square of the vertical distances of the points from the regression line. The X and Y variables are not symmetrical in the regression calculations (they are symmetrical in the correlation calculations).
- In rare cases, it might make sense to perform both regression and correlation. InStat can only perform one at a time, but you can go back and change the analysis choices.

Pearson or Spearman correlation?

If you choose correlation, choose between standard (Pearson) correlation and nonparametric (Spearman) correlation. Pearson correlation calculations are based on the assumption that both X and Y values are sampled from populations that follow a Gaussian distribution, at least approximately. With large samples, this assumption is not too important. If you don't wish to make the Gaussian assumption, select nonparametric (Spearman) correlation instead. Spearman correlation is based on ranking the two variables, and so makes no assumption about the distribution of the values.

When to force a regression line through the origin?

If you choose regression, you may force the line to go through a particular point such as the origin. In this case, InStat will determine only the best-fit slope, as the intercept will be fixed. Use this option when scientific theory tells you that the line must go through a particular point (usually the origin, $X=0$, $Y=0$) and you only want to know the slope. This situation arises rarely.

Use common sense when making your decision. For example, consider a protein assay. You measure optical density (Y) for several known concentrations of protein in order to create a standard curve. You then want to interpolate unknown protein concentrations from that standard curve. When performing the assay, you adjusted the spectrophotometer so that it reads zero with zero protein. Therefore you might be tempted to force the regression line through the origin. But you don't particularly care where the line is in the vicinity of the origin. You really care only that the line fits the standards very well near the unknowns. You will probably get a better fit by not constraining the line.

Most often, you will let InStat find the best-fit line without any constraints.

Test departure from linearity with runs test?

Linear regression is based on the assumption that the relationship between X and Y is linear. If you select this option, InStat can test that assumption with the runs test.

A run is a series of consecutive points that are either all above or all below the regression line. If the points are randomly distributed above and below the regression line, InStat knows how many runs to expect. If there are fewer runs, it suggests that the data follow a curve rather than a line.

Interpolate unknowns from a standard curve?

InStat can interpolate unknown values from the standard curve created by linear regression. See “Reading unknowns from standard curves” on page 95.

Results of correlation

Checklist. Is correlation the right analysis for these data?

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

Are the subjects independent?

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

Are X and Y measured independently?

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

Were X values measured (not controlled)?

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

Is the covariation linear?

The correlation would not be meaningful if Y increases as X increases up to a point, and then Y decreases as X increases further.

Are X and Y distributed according to Gaussian distributions?

To accept the P value from standard (Pearson) correlation, the X and Y values must each be sampled from populations that follow Gaussian

distributions. Spearman nonparametric correlation does not make this assumption.

How to think about results of linear correlation

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

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

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

Correlation results line by line

Correlation coefficient

The correlation coefficient, r , ranges from -1 to 1. The nonparametric Spearman correlation coefficient is abbreviated r_s but is interpreted the same way.

Value of r or r_s	Interpretation
Zero	The two variables do not vary together at all.
Positive fraction	The two variables tend to increase or decrease together.
Negative fraction	One variable increases as the other decreases.
1.0	Perfect correlation.
-1.0	Perfect negative or inverse correlation.

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

- The X variable helps determine the value of the Y variable.
- The Y variable helps determine the value of the X variable.

- Another variable influences both X and Y.
- X and Y don't really correlate at all, and you just happened to observe such a strong correlation by chance. The P value determines how often this could occur.

r^2

Perhaps the best way to interpret the value of r is to square it to calculate r^2 . Statisticians call the quantity the *coefficient of determination*, but scientists call it *r squared*. It has a value that ranges from zero to one, and is the fraction of the variance in the two variables that is shared. For example, if $r^2=0.59$, then 59% of the variance in X can be explained by (or goes along with) variation in Y. Likewise, 59% of the variance in Y can be explained by (or goes along with) variation in X. More simply, 59% of the variance is shared between X and Y.

Only calculate r^2 from the Pearson correlation coefficient, not from the nonparametric Spearman correlation coefficient.

P value

The P value answers this question: If the two variables really aren't correlated at all in the overall population, what is the chance that you would obtain a correlation coefficient as far from zero as observed in your experiment from randomly selected subjects?

Results of linear regression

Checklist. Is linear regression the right analysis for these data?

To check that linear regression is an appropriate analysis for these data, ask yourself these questions. InStat cannot help answer them.

Can the relationship between X and Y be graphed as a straight line?

In many experiments, the relationship between X and Y is curved, making linear regression inappropriate. Either transform the data, or use a program (such as GraphPad Prism) that can perform nonlinear curve fitting.

Is the scatter of data around the line Gaussian (at least approximately)?

Linear regression assumes that the scatter is Gaussian.

Is the variability the same everywhere?

Linear regression assumes that scatter of points around the best-fit line has the same standard deviation all along the curve. The assumption is violated if the points with higher (or lower) X values also tend to be further from the best-fit line. The assumption that the standard deviation is the same everywhere is termed *homoscedasticity*.

Do you know the X values precisely?

The linear regression model assumes that X values are exactly correct, and that experimental error or biological variability only affects the Y values. This is rarely the case, but it is sufficient to assume that any imprecision in measuring X is very small compared to the variability in Y.

Are the data points independent?

Whether one point is above or below the line is a matter of chance, and does not influence whether another point is above or below the line. See "The need for independent samples" on page 10.

How to think about the results of linear regression

Your approach to linear regression will depend on your goals.

If your goal is to analyze a standard curve, you won't be very interested in most of the results. Just make sure that r^2 is high and that the line goes near the points. Then go straight to the standard curve results. See "Reading unknowns from standard curves" on page 95.

In other cases, you will be most interested in the best-fit values for slope and intercept. Also look at the 95% confidence interval for these values. You can be 95% certain that these ranges include the true best-fit values. If the intervals are too wide, repeat the experiment collecting more data points.

Don't forget to look at a graph of the data by clicking the Graph button (the sixth step button at the bottom of the InStat window). InStat shows you the best-fit line, and an error envelope. You can be 95% sure that the true best-fit line (if you had an infinite amount of data) will lie somewhere within the envelope.

Linear regression results line by line

Slope and intercept

InStat displays the values of the slope and Y-intercept with standard errors and 95% confidence intervals. If the assumptions of linear regression are true, then you can be 95% certain that confidence interval contains the true population values of the slope and intercept.

Goodness of fit

InStat assesses goodness-of-fit by reporting $s_{y,x}$ and r^2 .

The value r^2 is a fraction between 0.0 and 1.0, and has no units. When r^2 equals 0.0, there is no linear relationship between X and Y. In this case, the best-fit line is a horizontal line going through the mean of all Y values, and knowing X does not help you predict Y. When $r^2=1.0$, all points lie exactly

on a straight line with no scatter. If you know X, you can predict Y exactly. With most data, r^2 is between 0.0 and 1.0.

You can think of r^2 as the fraction of the total variance of Y that is “explained” by the linear regression model. More simply, the variation of points around the regression line equals $1.0 - r^2$ of the total variation in Y.

The value $s_{y,x}$ is the standard deviation of the vertical distances of the points from the line. Since the distances of the points from the line are termed residuals, $s_{y,x}$ is the standard deviation of the residuals. Its value is expressed in the same units as Y. You’ll only be interested in its value if you plan to perform more advanced calculations.

Is the slope significantly different than zero?

InStat tests whether the slope differs significantly from zero (horizontal). The null hypothesis is that there is no linear relationship between X and Y overall, so the true best-fit line is horizontal. The P value answers this question: If the null hypothesis is true, what is the probability that randomly selected points would result in a regression line as far from horizontal (or further) as you observed? The P value is calculated from an F test, and InStat reports the value of F and its degrees of freedom.

Residuals and the runs test

The runs test determines whether your data differ significantly from a straight line.

A run is a series of consecutive points that are either all above or all below the regression line. In other words, a run is a series of consecutive points whose residuals are either all positive or all negative.

If the data follow a curve rather than a line, the points will tend to be clustered together above or below the line. There will be too few runs. The P value answers this question: If the data points are randomly scattered around a straight line, what is the chance of finding as few (or fewer) runs as you observed. If there are fewer runs than expected, the P value will be low, suggesting that your data follow a curve rather than a straight line.

Standard Curve

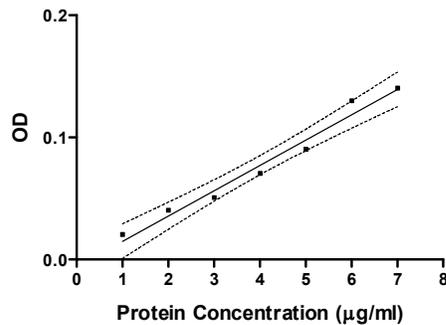
To read unknown values from a standard curve, you must enter unpaired X or Y values below the X and Y values for the standard curve and check the option to interpolate unknowns from a standard curve. See "Reading unknowns from standard curves" on page 95.

InStat's graph of linear regression results

InStat graphs the best-fit line along with the data. It also plots the 95% confidence interval as dotted lines. Assuming that all the assumptions of linear regression are true, you can be 95% sure that the true best-fit line (if you had an infinite amount of data) lies within those confidence limits.

To include uncertainty in both the slope and the intercept, the confidence limits are curved. This does not mean that they include the possibility of a nonlinear relationship between X and Y. Instead, the curved confidence limits demarcate the area that can contain the best-fit straight regression line.

With InStat, you cannot customize this graph in any way. GraphPad Prism (see page 120) can do the same analyses, but lets you customize the graph for publications or presentations. Here is a Prism graph showing a linear regression line with 95% confidence intervals (similar to the graph made by InStat).



Reading unknowns from standard curves

What is a standard curve?

An *assay* is an experimental procedure used to determine the concentration of a substance. The measurement (which might be optical density, radioactivity, luminescence, or something else) varies with the concentration of the substance you are measuring. A *standard curve* is a graph of assay measurement (Y) as a function of known concentrations of the substance (X) used to calibrate the assay. Standard curves can be linear or curved.

Once you have created a standard curve using known concentrations of the substance, you can use it to determine the concentration of unknowns. Perform the same assay with the unknown sample. Then read across the graph from the spot on the Y-axis that corresponds to the assay measurement of the unknown until you intersect the standard curve. Read down the graph until you intersect the X-axis. The concentration of substance in the unknown sample is the value on the X-axis.

Entering standard curve data into InStat

InStat can interpolate unknown values from linear standard curves. Simply enter the unknowns directly below the standard curve. Enter either X or Y, but not both. Most often you will enter Y. This example has four standards, and four unknowns entered underneath.

Row	X	Y
1	1	3.2
2	2	7.1
3	3	12.5
4	4	16.3
5		4.2
6		5.6
7		9.4
8		10.2

On the step where you choose the test, check the option box for standard curve calculations. InStat will include the standard curve results on the result page.

InStat will flag (with an asterisk) any unknowns that are outside of the range of the standard curve. While you may accept the results from unknowns that are just barely out of range, be cautious about results from unknowns far from the standard curve.

InStat does not do any kind of curve fitting. If your standard curve is not linear, you have two choices:

- Transform the data to create a linear standard curve. If you transform Y values of your standards, also be sure to transform the Y values of the unknowns. If you transform the X values of your standards, InStat will report the unknowns in that transformed scale and you'll need to do a reverse transform.
- Use a program, such as GraphPad Prism, that can fit nonlinear curves and read unknown values off that curve.

Multiple Regression and correlation

Introduction to multiple regression and correlation

Uses of multiple regression

In laboratory experiments, you can generally control all the variables. You change one variable, measure another, and then analyze the data with one of the standard statistical tests. But in some kind of experiments, and many observational studies, you need to analyze the interaction of several variables. Multiple regression is one way to do this.

Multiple regression fits an equation that predicts one variable (the dependent variable, Y) from two or more independent (X) variables. For example, you might use multiple regression to predict blood pressure from age, weight and gender.

In some situations, your goal may really be to examine several variables at once. With the blood pressure example, your goal may be to find out which variable has the largest influence on blood pressure: age, weight or gender. Or your goal may be to find an equation that best predicts blood pressure from those three variables.

In other situations, you really only care about one of the independent variables, but your analysis needs to adjust for differences in other variables. For this example, you might ask: Does blood pressure vary with age, after correcting for differences in weight and differences between the sexes? Or you might ask: Does blood pressure differ between men and women, after correcting for differences in age and weight?

InStat, like other multiple regression programs, presents you with many results and it is easy to be overwhelmed. Your approach to the results will depend, in part, on what question you are trying to answer. Before looking at the results, try to clearly articulate your questions.

Multiple regression is more complicated than the other statistical tests offered by InStat, so the results can be confusing and misleading to someone who has never used multiple regression before. Before analyzing your data with multiple regression, find an experienced consultant or consult one of these books:

- SA Glantz and BK Slinker, *Primer of Applied Regression and Analysis of Variance*, McGraw-Hill, 1990.
- LD Fisher and G vanBelle, *Biostatistics. A Methodology for the Health Sciences*, Wiley, 1993.

The multiple regression model and its assumptions

Multiple regression fits your data to this equation, where each X_i represents a different X variable.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \dots + \text{random scatter}$$

If there is only a single X variable, then the equation is $Y = \beta_0 + \beta_1 X_1$, and the “multiple regression” analysis is the same as simple linear regression (β_0 is the Y intercept; β_1 is the slope).

For the example, the equation would be

Blood pressure = $\beta_0 + \beta_1 \text{age} + \beta_2 \text{weight} + \beta_3 \text{gender} + \text{random scatter}$

Gender is coded as 0=male and 1=female. This is called a *dummy variable*.

InStat finds the values of β_0 , β_1 , etc. that make the equation generate a curve that comes as close as possible to your data. More precisely, InStat finds the values of those coefficients that minimize the sum of the square of the differences between the Y values in your data and the Y values predicted by the equation.

The model is very simple, and it is surprising that it turns out to be so useful. For the blood pressure example, the model assumes:

- On average, blood pressure increases (or decreases) a certain amount (the best-fit value of β_1) for every year of age. This amount is the same for men and women of all ages and all weights.
- On average, blood pressure increases (or decreases) a certain amount per pound (the best-fit value of β_2). This amount is the same for men and women of all ages and all weights.
- On average, blood pressure differs by a certain amount between men and women (the best-fit value of β_3). This amount is the same for people of all ages and weights.

The mathematical terms are that the model is *linear* and allows for *no interaction*. *Linear* means that holding other variables constant, the graph of blood pressure vs. age (or vs. weight) is a straight line. *No interaction* means that the slope of the blood pressure vs. age line is the same for all weights and for men and women.

You can sometimes work around the linearity assumption by transforming one or more of the X variables. You could transform weight to square root of weight, for example.

You can sometimes work around the assumption of no interaction by creating a new column by multiplying two variables together (in this example create a new variable defined as weight times age). Including this column as an additional variable in the multiple regression model partially takes into account interactions. Consult a statistician or an advanced statistics book before trying this.

Additionally, the multiple regression procedure makes assumptions about the random scatter. It assumes that the scatter is Gaussian, and that the standard deviation of the scatter is the same for all values of X and Y. Furthermore, the model assumes that the scatter for each subject should be random, and should not be influenced by the deviation of other subjects. See “The need for independent samples” on page 10.

There is an additional complexity with this example, in that the variables are intertwined — weight tends to go up with age, and men tend to be heavier than women. See "Is multicollinearity a problem?" on page 105.

Entering data for multiple regression and correlation

Enter each subject (or experimental unit) into a row, with each variable in a separate column. You don't have to decide (yet) which column contains the dependent (Y) variable, although it is customary to place it in the first column.

Each variable (column) can be:

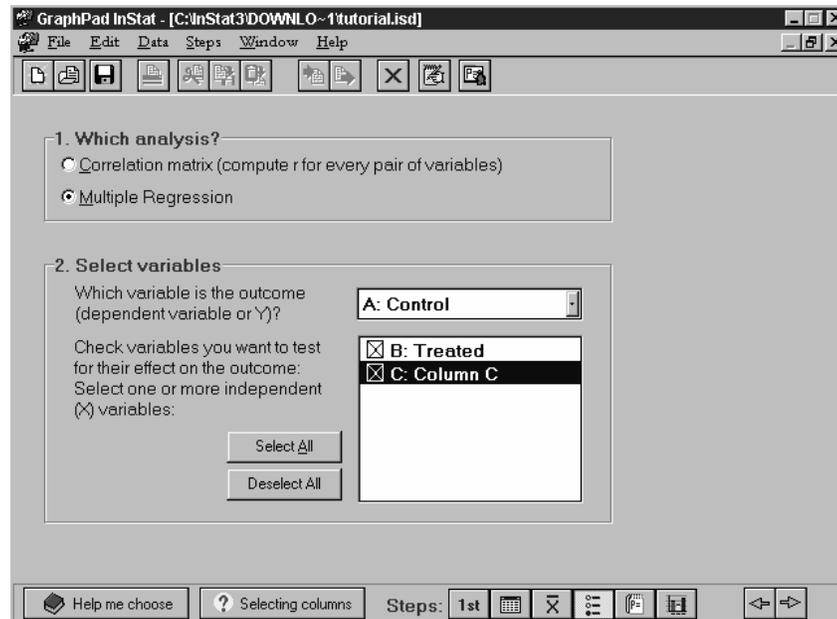
- A measured variable (blood pressure).
- A transformation of a measured variable (i.e., age squared or logarithm of serum LH levels). Since the multiple regression only fits data to a linear equation, you may get better results in some cases by transforming the variables first.
- A discrete variable which has only two possible values. For example a column could code gender; enter 0 for male and 1 for female.

You must enter more rows of data (more subjects) than independent variables. For the results of multiple regression to be useful, you'll need many more subjects than variables. One rule of thumb is that you should have at least 5-10 times more subjects than variables.

Get statistical help (or consult advanced books) before you do any of the following:

- Enter a discrete variable with more than two possible values, for example prior treatment with one of three drugs or residence in one of four states. Don't enter the code into a single variable. Instead, you have to create several dummy variables (several columns) to encode all the possibilities. This is more complicated than it sounds, and you'll need expert guidance to do it properly and to make sense of the results.
- Enter a variable into one column and a function of that variable (perhaps the variable squared) in another column.
- Enter the product of two variables into a column by itself to account for interaction.

Analysis choices for multiple regression and correlation



Decide first whether you want to compute a correlation matrix or perform multiple regression.

Multiple correlation finds the correlation coefficient (r) for every pair of variables. Your only choice is to select the variables (columns) you want to include. The other columns are ignored. InStat computes the correlation coefficient for each pair of columns independently, and shows the results as a correlation matrix. InStat does not compute partial correlation coefficients.

Multiple regression finds the linear equation that best predicts the value of one of the variables (the dependent variable) from the others. To use multiple regression, therefore, you have to designate one of the columns as the dependent (Y) variable and choose which of the remaining columns contain independent (X) variables you want to include in the equation. InStat ignores the rest of the columns. Some programs can decide which X variables to include in the regression model. They do this by performing step-wise multiple regression, using one of several methods. InStat does not perform any kind of stepwise regression.

Note that the Y variable should **not** be a discrete (binary) variable, for example a variable that equals 0 for failure and 1 for success. If you want to find an equation that predicts a binary variable, then you need to use multiple *logistic* regression. InStat does not do this.

Interpreting a correlation matrix

InStat reports the correlation coefficient (r) for each pair of variables (columns). Each r is calculated based on the values in those two columns,

without regard to the other columns. The value of r can range from -1 (a perfect negative correlation) to $+1$ (perfect positive correlation). InStat does not calculate partial correlation coefficients.

If data are missing, those rows are excluded from all calculations. For example if the value for row 5 is missing for column 3, then all values in row 5 are ignored when calculating all the correlation coefficients.

The results of multiple regression

Checklist. Is multiple regression the right analysis?

To check that multiple regression is an appropriate analysis for these data, ask yourself these questions.

Is the relationship between each X variable and Y linear?

In many experiments, the relationship between X and Y is nonlinear, making multiple regression inappropriate. In some cases you may be able to transform one or more X variables to create a linear relationship. You may also be able to restrict your data to a limited range of X variables, where the relationship is close to linear. Some programs (but none currently available from GraphPad Software) can perform nonlinear regression with multiple independent variables.

Is the scatter of data around the prediction of the model Gaussian (at least approximately)?

Multiple regression assumes that the distribution of values from the prediction of the model is random and Gaussian.

Is the variability the same everywhere?

Multiple regression assumes that scatter of data from the predictions of the model has the same standard deviation for all values of the independent variables. The assumption is violated if the scatter goes up (or down) as one of the X variables gets larger. The assumption that the standard deviation is the same everywhere is termed *homoscedasticity*.

Do you know the X values precisely?

The linear regression model assumes that all the X values are exactly correct, and that experimental error or biological variability only affects the Y values. This is rarely the case, but it is sufficient to assume that any imprecision in measuring X is very small compared to the variability in Y.

Are the data points independent?

Whether one value is higher or lower than the regression model predicts should be random. See “The need for independent samples” on page 10.

General approach

Multiple regression is far more complicated than the other analyses offered by InStat, and the information in the manual and help screens may not be sufficient to analyze your data completely and correctly. Consider getting expert guidance.

The results of multiple regression help you answer these questions, each discussed below. Depending on your scientific goals, you may be very interested in some of these questions and less interested in others.

- What is the best fit?
- How good is the fit?
- Which X variable(s) make a significant contribution?
- Is multicollinearity a problem?
- Would a simpler model fit as well?

What is the best fit?

Multiple regression fits an equation to data, so InStat reports the values of the parameters in the equation that make it fit the data best. Each best-fit parameter has the units of the Y variable divided by the units of its X variable.

Here again is the multiple regression model for the blood pressure example.

Blood pressure = $\beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{weight} + \beta_3 \cdot \text{gender} + \text{random scatter}$

Assuming that blood pressure is measured in torr (same as mm Hg) and age is measured in years, the variable β_1 will have units of torr/year. It is the amount by which blood pressure increases, on average, for every year increase in age, after correcting for differences in gender and weight. If weight is measured in kg, then β_2 has units of torr/kg. It is the average amount by which blood pressure increases for every kg increase in weight, adjusting for differences in age and gender. Gender is a dummy variable with no units, coded so that males are zero and females are one. Therefore, the variable β_3 has units of torr. It is the average difference in blood pressure between men and women, after taking into account differences in age and weight.

The variable β_0 , which InStat calls "constant", is needed to complete the equation. It is expressed in units of the Y variable. It is the value of Y when all X variables equal zero. This will rarely make any biological sense, since many X variables are never anywhere near zero. In the blood pressure example, you could think of it as the average blood pressure of men (since gender is coded as zero) with age=0 and weight=0!

The only way you could really know the best-fit values of the parameters in the model would be to collect an infinite amount of data. Since you can't do this, the best-fit values reported by InStat are influenced, in part, by random variability in picking subjects. InStat reports this uncertainty as a

95% confidence interval for each parameter. These take into account the number of subjects in your study, as well as the scatter of your data from the predictions of the model. If the assumptions of the analysis are true, you can be 95% sure that the interval contains the true best-fit value of the variable.

InStat also presents the standard error of each parameter in the model. These are hard to interpret, but are used to compute the 95% confidence intervals for each coefficient. InStat shows them so that its results can be compared to those of other programs.

How good is the fit?

InStat quantifies goodness-of-fit in several ways.

Term	Explanation
R ²	The fraction of all variance in Y that is explained by the multiple regression model. If R ² equals 1.0, then each Y value is predicted perfectly by the model, with no random variability. If R ² equals 0.0, then the regression model does a terrible job of predicting Y values – you’ll get equally accurate predictions by simply predicting that each Y value equals the mean of the Y values you measured. With real data, of course, you won’t see those extreme R ² values, but instead will see R ² values between 0.0 and 1.0.
P value	The P value answers this question: If you collected random data, what is the chance that you’d happen to obtain an R ² value as large, or larger, than you obtained in this experiment. More simply, the P value tests whether the regression model predicts Y in a statistically significant manner – whether the predictions of the model are any better than chance alone.
Sum-of-squares and SD of residuals	Multiple regression finds values for coefficients in the model that minimize the sum-of-squares of the differences between the predicted Y values and the actual Y values. InStat reports the sum-of-squares along with the SD of the residuals (square root of SS divided by N-V, where N is number of subjects, and V is the number of independent variables). These values are used if you perform advanced calculations.

Adjusted R ²	Even if the data are all random, you expect R ² to get larger as you add more variables to the equation. Just by chance the model will predict the data better if it has more components. The adjusted R ² value corrects for this, by correcting for the number of X variables in the model. If you collect random data, you'd expect the adjusted R ² value to be zero on average. If you collected many sets of random data, the adjusted R ² value will be negative half the time, and positive half the time. If the adjusted R ² were really the square of anything, then it would always be positive. But the adjusted R ² is not the square of anything – it is just R ² minus a correction. The adjusted R ² is mostly useful for comparing the fits of models with different numbers of independent variables. You can't compare R ² , because you expect R ² to be smaller in the fit with more variables just by chance.
Multiple R	Multiple R is the square root of R ² . It is not particularly useful, but other programs report it so InStat does too. You can interpret it much like you interpret a correlation coefficient.
F	This F ratio is used to compute the P value. InStat includes it for completeness.

Which variable(s) make a significant contribution?

If the overall P value is high, you can conclude that the multiple regression model does not explain your data. In this case, there is not much point in looking at the results for individual variables. If the overall P value is low, you probably will next want to find out which variables in the model are useful and which are extraneous.

For each independent variable in the model, InStat reports a P value that answers this question: After accounting for all the other independent variables, does adding this variable to the model significantly improve the ability of the model to account for the data? If the P value is small, the variable contributes in a statistically significant manner. If the P value is large, then the contribution of the variable is no greater than you'd expect to see by chance alone. InStat uses the standard threshold (alpha) value of 0.05. If a P value is less than 0.05, then InStat reports that the variable made a statistically significant contribution to the fit. If a P value is greater than 0.05, InStat concludes that the influence of that variable is not statistically significant. This threshold (0.05) is arbitrary but conventional.

A common use of multiple regression is to determine the influence of one independent variable after correcting for others. For example, suppose that you want to compare blood pressure between men and women after correcting for age and weight. In this case, you'll interpret the P value for the main X variable (gender) somewhat differently than the P value for the other X variables (age and weight). What you really care about is the P value for the main variable (gender). If it is low, conclude gender affects blood pressure, after correcting for differences in age and weight. The P value for the other X variables (age and weight) are less interesting. A low P value tells you that there is a linear relationship between that variable and

the outcome, which justifies your decision to include it in the multiple regression model.

For each variable, InStat also reports a t ratio, an intermediate result that won't be of interest to most InStat users. It equals the absolute value of the coefficient divided by its standard error. The P value is defined more precisely in terms of t. If the true best-fit value of this coefficient (given an infinite amount of data) were really zero, what is the chance that analysis of randomly selected data (of the same sample size you used) would lead to a value of t as far from zero (or further) as you obtained here?

Is multicollinearity a problem?

The term *multicollinearity* is as hard to understand as it is to say. But it is important to understand, as multicollinearity can interfere with proper interpretation of multiple regression results. To understand multicollinearity, first consider an absurd example. Imagine that you are running multiple regression to predict blood pressure from age and weight. Now imagine that you've entered weight-in-pounds and weight-in-kilograms as two separate X variables. The two X variables measure exactly the same thing – the only difference is that the two variables have different units. The P value for the overall fit is likely to be low, telling you that blood pressure is linearly related to age and weight. Then you'd look at the individual P values. The P value for weight-in-pounds would be very high – after including the other variables in the equation, this one adds no new information. Since the equation has already taken into account the effect of weight-in-kilograms on blood pressure, adding the variable weight-in-pounds to the equation adds nothing. But the P value for weight-in-kilograms would also be high for the same reason. After you include weight-in-pounds to the model, the goodness-of-fit is not improved by including the variable weight-in-kilograms. When you see these results, you might mistakenly conclude that weight does not influence blood pressure at all since both weight variables have very high P values. The problem is that the P values only assess the incremental effect of each variable. In this example, neither variable has any incremental effect on the model. The two variables are collinear.

That example is a bit absurd, since the two variables are identical except for units. The blood pressure example -- model blood pressure as a function of age, weight and gender – is more typical. It is hard to separate the effects of age and weight, if the older subjects tend to weigh more than the younger subjects. It is hard to separate the effects of weight and gender if the men weigh more than the women. Since the X variables are intertwined, multicollinearity will make it difficult to interpret the multiple regression results.

Multicollinearity is an intrinsic problem of multiple regression, and it can frustrate your ability to make sense of the data. All InStat can do is warn you about the problem. It does this by asking how well each independent (X) variable can be predicted from the other X variables (ignoring the Y variable). There are three ways to express the result.

Term	Explanation
R ² with other X variables.	The fraction of all variance in one X variable that can be predicted from the other X variables.
Variance Inflation Factor (VIF).	If the X variables contain no redundant information, you expect VIF to equal one. If the X variables are collinear (contain redundant information), then VIF will be greater than one. Multicollinearity increases the width of the confidence interval (which is proportional to the square root of variance) by a factor equal to the square root of VIF. If a variable has a VIF of 9, the confidence interval of that coefficient is three times wider than it would be were it not for multicollinearity.
Tolerance	The fraction of the total variance in one X variable that is <u>not</u> predicted by the other X variables.

The three terms measure exactly the same thing – the degree of multicollinearity. InStat reports both R² and VIF, so you can use the value you are more familiar with. For each X variable, the corresponding VIF is computed from R² by this formula: $VIF=1/(1-R^2)$. InStat does not report tolerance, but you can easily calculate it yourself for each variable as $1.0 - R^2$.

If R² and VIF are high for some X variables, then multicollinearity is a problem in your data. How high is high? Any threshold is somewhat arbitrary, but here is one rule of thumb. If any of the R² values are greater than 0.75 (so VIF is greater than 4.0), suspect that multicollinearity might be a problem. If any of the R² values are greater than 0.90 (so VIF is greater than 10) then conclude that multicollinearity is a serious problem.

Don't confuse these individual R² values for each X variable with the overall R². The individual R² values quantify how well each X variable can be predicted from the other X variables. The overall R² quantifies goodness-of-fit of the entire multiple regression model. Generally you want the overall R² value to be high (good fit) while all the individual R² values to be low (little multicollinearity).

If multicollinearity is a problem, the results of multiple regression are unlikely to be helpful. In some cases, removing one or more variables from the model will reduce multicollinearity to an acceptable level. In other cases, you may be able to reduce multicollinearity by collecting data over a wider range of experimental conditions. This is a difficult problem, and you will need to seek statistical guidance elsewhere.

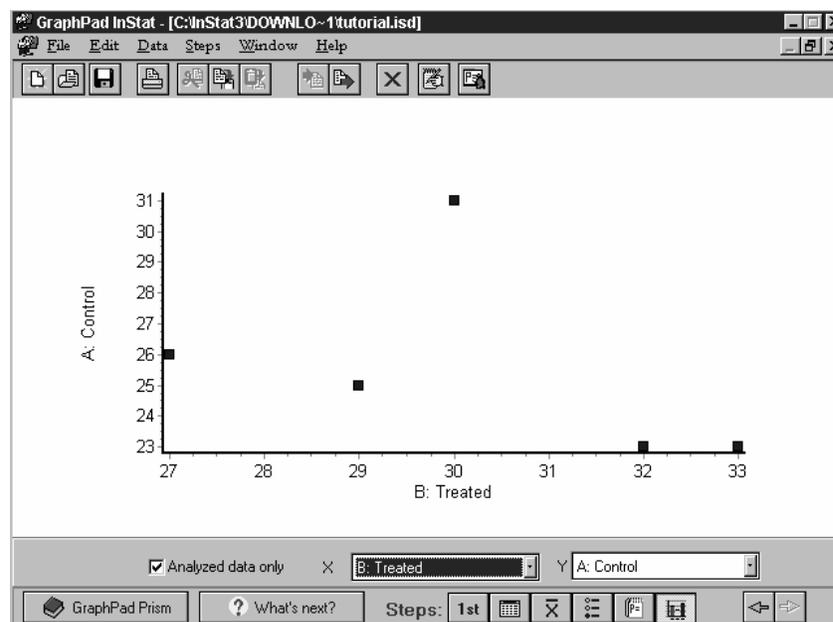
Would a simpler model work as well?

More advanced multiple regression programs can perform variable selection procedures to determine which of the X variables should be kept in the model and which should be omitted. This is trickier than it sounds, and different programs do the job differently, and can wind up with different results. You need to use a great deal of care and experience to use variable selection procedures appropriately, and you may wish to consult with a statistician.

InStat does not do any automatic variable selection, but can help you do one form of variable selection, *backward elimination*, manually. Follow these steps:

1. Perform multiple regression with all potential X variables.
2. Look at the individual P values in the section labeled “Which variable(s) make a significant contribution”. If all of the P values are below a threshold you set in advance (usually 0.05), then you are done. Keep all the X variables in the model.
3. If one or more X variables have a P value higher than the threshold, remove the one with the highest P value (it will also have the lowest t ratio). To do so, go back to the Select Test step and uncheck that variable. Then go back to the results to see the new fit without that variable.
4. Go back to step 2. Keep removing variables until all the P values are less than 0.05.

Graphing multiple regression results



InStat does not graph the results of multiple regression. To graph the best-fit results of a model with two X variables would require a three dimensional graph. The results of models with more than two variables cannot readily be graphed.

InStat does let you look at the relationship between any two variables in the model.

If you check the box "analyzed data only", InStat graphs only the data included in the analysis. This means that it graphs only variables included in the model, and only rows of data with no missing values (if any values are missing, all the values on that row are omitted). InStat can plot any X variable vs. the dependent variable (Y). Or InStat can graph any X variable vs. the residuals. Residuals are defined to be the distance between the independent (Y) value of each row and the Y value predicted by the model. If you've picked an appropriate model, the residuals ought to be random — you should observe no relationship between any of the independent (X) variables and the size or sign of the residual.

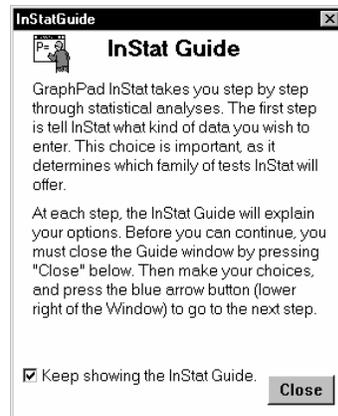
If you uncheck the box "analyzed data only", InStat graphs any column of data vs. any other column. This can be a useful way to examine relationships in your data before selecting variables for multiple regression.

Using InStat

Online Help

The InStat Guide

The InStat Guide window helps you learn InStat. It appears when you first run InStat, and comes back every time you move from step to step until you uncheck the option box "Keep showing the InStat Guide". Show the Guide again by dropping the Help menu and choosing InStat Guide.



Using the help system

The entire contents of this manual are available in the online help system. InStat uses the standard Windows and Mac help engine, so the commands should be familiar to you. Note particularly the button at the right of Help's tool bar labeled like this: >> Click that button to go to the next help screen. Click it repeatedly to step through every InStat help screen.

Importing and exporting data

Importing data tables from other programs

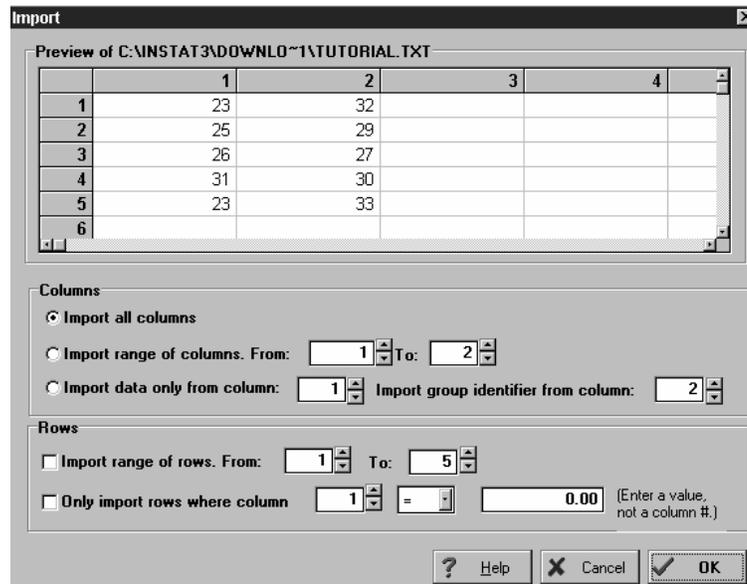
If you've already entered your data into another program, there is no need to retype. You may import the data into InStat via a text file, or copy and paste the values using the clipboard.

InStat imports text files with adjacent values separated by commas, spaces or tabs. Some programs refer to these files as ASCII files rather than text files. To save a text file from Excel (versions 4 or later) use the File Save As command and set the file type to Text or CSV (one uses tabs, the other commas to separate columns). With other programs, you'll need to find the

appropriate command to save a text file. If a file is not a text file, changing the extension to .TXT won't help.

To import data from text (ASCII) files:

- Go to the data table and position the insertion point. The cell that contains the insertion point will become the upper left corner of the imported data.
- Choose Import from the File menu.
- Choose a file.
- Choose import options.



If you have trouble importing data, inspect the file using the Windows Notepad to make sure it contains only numbers clearly arranged into a table. Also note that it is not possible to import data into a 2x2 contingency table.

Importing indexed data

Some statistics programs save data in an indexed format (sometimes called a stacked format). Each row is for a case, and each column is for a variable. Groups are not defined (as in InStat) by different columns, but rather by a grouping variable.

InStat can import indexed data. On the import dialog, specify one column that contains all the data and another column that contains the group identifier. The group identifiers must be integers (not text), but do not have to start at 1 and do not have to be sequential.

For example, in this sample indexed data file, you may want to import only the data in column 2 and use the values in column 3 to define the two groups.

Row #	Col. 1	Col. 2	Col. 3
1	12	123	5
2	14	142	9
3	13	152	5
4	12	116	9
5	11	125	9
6	15	134	5

In the Import dialog, specify that you want to import data only from column 2 and that column 3 contains the group identifier. InStat will automatically rearrange the data, so they look this like:

Row #	Group A	Group B
1	123	142
2	152	116
3	134	125

Filtering data

You don't have to import all rows of data from a file. InStat provides two ways to import only a range of data. You can specify a range of rows to import (i.e. import rows 1-21). Or you can filter data by applying criteria. For example, only import rows where column 3 equals 2, or where column 5 is greater than 100. InStat filters data by comparing the values in one column with a value you enter. It cannot compare values in two columns. For example, it is not possible to import rows where the data in column 3 is larger than the value in column 5.

Exporting data

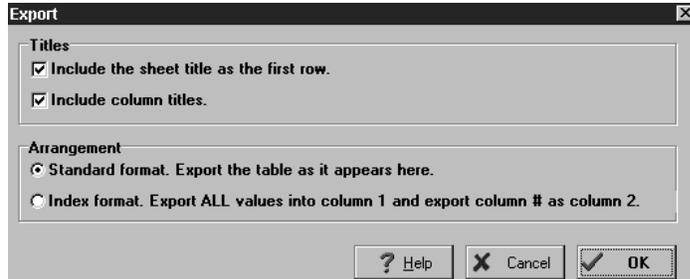
Transfer data from InStat to other programs either by exporting the data to disk or copying to the clipboard. Other programs cannot read InStat data (ISD) files.

InStat exports data formatted as plain ASCII text with adjacent values separated by commas or tabs. These files have the extensions *.CSV or *.TXT.

To export data:

- Choose Export from the File menu.

- Choose the disk and directory, if necessary. Enter a file name. Press OK.
- Specify whether the exported file should contain column titles or the overall data set (sheet) title (entered on top of the data table).
- Choose whether you want to export in standard format (so the exported table is arranged the same as the data table) or indexed format (for importing into other statistics programs).



Here is an example of exported indexed data. The data look like this in InStat:

Row #	Group A	Group B	Group C
1	126	123	124
2	142	142	165
3	135	152	174

Here is what the file looks like when exported in index format:

Row #	Col. 1	Col. 2
1	126	1
2	142	1
3	135	1
4	123	2
5	142	2
6	152	2
7	124	3
8	165	3
9	174	3

Working with the data table

Editing values

To move the insertion point, point to a cell with the mouse and click, or press an arrow key on the keyboard. Tab moves to the right; shift-Tab

moves to the left. Press the Enter (Return) key to move down to the next row.

When you move the insertion point to a cell in the data table, you also select the number (if any) in that cell. To overwrite that number, just start typing. To edit that number, click once to go to the cell and then click again to place the insertion point inside the cell. Then you can edit individual digits within the number.

The InStat data table has 1000 rows and 26 columns.

Number format

Initially, InStat automatically chooses the number of decimal points to display in each column. To change the number of decimal points displayed, select the column or columns you wish to change. Then pull down the Data menu and choose Number Format and complete the dialog. It is not possible to change the numerical format of selected cells. InStat displays all data in each column with the same number of decimal places.

Altering the numerical format does **not** change the way InStat stores numbers, so will not affect the results of any analyses. Altering the numerical format **does** affect the way that InStat copies numbers to the clipboard. When you copy to the clipboard, InStat copies exactly what you see.

Missing values and excluded data

If a value is missing, simply leave its spot on the data table blank. InStat handles missing values appropriately. If you pick multiple regression, InStat will ignore an entire row of values if one or more is missing. InStat does not allow missing values with a paired t test, repeated measures ANOVA, or the analogous nonparametric tests.

If a value is too high or too low to be believable, you can exclude it. Excluded values are shown in blue italics on the data table, but are not included in analyses and are not shown on graphs. From the point of view of analyses and graphs, it is just as if you had deleted the value. But the number remains on the data table to document its value.

To exclude data:

- Select the cell or cells you wish to exclude.
- Pull down the Data menu and choose exclude. The excluded values appear in blue Italics.
- Repeat the process to include the value again.

Tip: If you want to run some analyses both with and without the excluded values, duplicate the window (Window Duplicate command). Then exclude values from one of the copies.

Row and column titles

Enter column titles on the data table right below the column identifiers (A, B, C...).

InStat labels each row with the row number, but you can create different row labels. When you enter paired or matched data, this lets you identify individual subjects.

To add your own row label:

1. Click to the left of the row number, in the small area labeled "...".
2. Enter or edit the row label. Initially the insertion point appears after the row number. Type additional characters if you want to label the row with both row number and label. Press backspace to delete the row number.
3. After entering or editing one row number, press the down arrow key or Enter to move down to the row label for the next row.

Using the clipboard

InStat uses the clipboard in a standard way to copy data from one location and to paste it somewhere else. Before copying, you must select a region on the data table.

To Select	Mouse	Keyboard
A range of data.	Point to one corner of the block. Hold down the left mouse button and drag to the opposite corner.	Move to one corner of the block. Hold down the Shift key and move to the opposite corner (using arrow keys).
One or more columns.	Click on one of the column headers ("A", "B", etc.). Drag over the desired range of columns.	Hold Ctrl, and press the spacebar (Windows only).
One or more rows.	Click on one of the row headers ("1", "2", etc.). Drag over the desired range of rows.	Hold Shift, and press the spacebar (Windows only).
All data on the table.	Click on the rectangle to the left of column A and above row 1.	Ctrl-A (Windows only)

Cut or copy the selection, then paste, using the toolbar buttons, commands on the Edit menu, commands on the shortcut menu (click the right mouse button) or keyboard shortcuts (using Windows hold Ctrl and using Mac hold Command, and then press X for cut, C for copy, V for paste).

Note: InStat copies exactly what you see. Changing the number (decimal) format will alter what is copied to the clipboard.

When you paste data, InStat maintains the arrangement of rows and columns. You can also transpose rows and columns by selecting Transpose

Paste from the Edit menu. InStat will paste what was the first row into the first column, what was the second row into the second column and so on.

Deleting data

Pressing the DEL key is not the same as selecting Delete from the Edit menu.

After selecting a range of data, press the DEL key to delete the selected range. InStat does not place deleted data on the clipboard and does not move other numbers on the table to fill the gaps.

Select Delete Cells from the Edit menu to delete a block of data completely, moving other data on the table to fill the gap. If you have selected one or more entire rows or columns, InStat will delete them. Remaining numbers move up or to the left to fill the gap. If you have selected a range of values, InStat presents three choices: Delete entire rows, delete entire columns, or delete just the selected range (moving other numbers up to fill the gap).

To delete an entire data table, pull down the Edit menu and choose Clear All.

Transforming data

To transform data:

1. Select a block of data you wish to transform. Or to transform a single value, place the insertion point in that cell.
2. Pull down the Data menu and choose Transform.
3. Select a transformation from this list:

Function	Comments
Y=Y squared	
Y=Log(Y)	Logarithm base 10
Y=Ln(Y)	Natural logarithm
Y=10^Y	Antilog of log base 10
Y=exp(Y)	e ^Y (antilog of natural log)
Y=1/Y	
Y=Sqrt(Y)	Square root of Y
Y=Logit(Y)	ln[Y/(K-Y)]
Y=sin(Y)	Y is in radians
Y=cos(Y)	Y is in radians
Y=tan(Y)	Y is in radians
Y=arcsin(Y)	Result is in radians
Y=abs(Y)	Absolute value
Y=K*Y	
Y=Y+K	
Y=Y-K	
Y=Y/K	

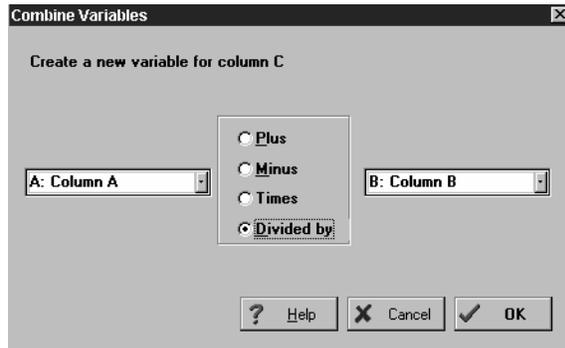
4. Enter K if necessary.

Note: InStat transforms in place, so erases the original data. There is no undo command, so the only way to bring back the original data is to perform the inverse transformation.

Combining variables

Add, subtract, multiply or divide two columns to create a new column:

- Place the cursor in an empty column.
- Pull down the Data menu and choose Combine variables.
- Choose two columns and how you want to combine them.



Arranging data

Details on how to arrange various kinds of data appear elsewhere:

- "Entering t test data into InStat" on page 38.
- "Entering ANOVA data" on page 58.
- "Entering data for correlation and linear regression" on page 88.
- "Entering standard curve data into InStat" on page 95.
- "Creating contingency tables" on page 78.
- "Entering data for multiple regression and correlation" on page 99.

Selecting columns to analyze

With InStat, you select columns to analyze on a dialog. Selecting columns on the spreadsheet – as you would to copy to the clipboard – has no effect on the analyses.

Type of test	How to select columns
Compare means or medians	By default, InStat will analyze all the columns you entered. To analyze a subset of columns, click the "select other columns" button on top of the screen where you choose a test. InStat displays a dialog listing all the columns. Check the ones you wish to analyze.
X and replicate Y values for linear regression and correlation	InStat will analyze all the columns. There is no way to select columns.
Large contingency table	InStat will analyze all the data. There is no way to select columns.
Y and 2 or more X variables for multiple regression	To pick columns, see "Analysis choices for multiple regression and correlation" on page 100.

After you view the results

After you read the results, you may want to do the following:

Print or export the results

Print or export the results (as a text file) using commands on the File menu. Or select a portion of the results, and copy to the Windows clipboard as text.

View a graph

InStat displays a notebook quality graph of your data to help you see differences and spot typographical errors on data entry. You cannot customize the graph to create a publication quality graph. Print the graph or export it as a Windows Metafile (wmf) using commands on the File menu. Or copy to the clipboard, and paste into another program.

Record notes or append the results to the notes window

Click the Notes button, or pull down the Edit menu and choose View Notes, to pop up a notes editor. Use the notes page to record where the raw data are stored, why you excluded values, what you concluded, or why you chose certain tests. InStat saves the notes as part of the ISD file, so you can refer back to them after opening a file.

To append portions of the results to the notes window, first select a portion of the results. Then pull down the Edit menu and select Append to notes. If you don't select a portion of the results first, InStat appends all the results.

To print notes, click the alternate (right) mouse button and choose Print.

Analyze the same data with a different test

Each InStat file contains a single data table and the results of a single statistical test. If you want to analyze data in several ways (say to compare a parametric test with a nonparametric test), you have two choices.

The easiest approach is to simply replace one set of results with another. After completing the first analysis, consider appending the results to the Notes window. Then go back to the Choose test step and make different choices. Then click Results to see the answer. The new results will replace the previous results.

To view both sets of results at once, follow these steps.

1. Enter data and do your first analysis.
2. Pull down the Windows menu and choose Duplicate. You now have two identical open documents in separate windows.
3. Change the analysis choices in one of the windows and go to results.
4. The two analyses now exist in separate windows. Switch between them using the Windows menu. Save and print the two windows separately. Each Window will be saved as a separate file.

Perform the same analysis on new data

To perform the same analyses on a new set of data, go back to the data table and replace the data. Then go straight to results. You don't have to select a test, as InStat will remember your choices. The new results will replace the previous results, which you can no longer view.

To view both sets of results at once, follow these steps.

1. Enter data and do your first analysis.
2. Pull down the Windows menu and choose Duplicate. You now have two identical open documents in separate windows.
3. Change the data in one of the windows and go to results.
4. The two analyses now exist in separate windows. Switch between them using the Windows menu. Save and print the two windows separately.

Start InStat again

Pull down the File menu and choose New to start InStat again. You'll be able to start fresh (erase the current data set) or start a new window while keeping the current one.

Create an analysis template

An InStat file contains not only a data table, but also your analysis choices. This lets InStat recalculate the results when it opens a file. If you perform the same analysis often, create an analysis template. To do so, simply save a

file after deleting the data. The file will contain only analysis choices. To use the template, open the file, enter data and go straight to results. You can skip the Choose Test screen, as InStat reads the choices from the file. The Windows and Macintosh versions of InStat use identical file formats, so you can move files between platforms with no special conversion.

InStat files

Save an InStat file using the File Save command, then open it using File Open. The toolbar has shortcut buttons for both commands.

InStat files store your data table along with analysis choices and notes. InStat files are denoted by the extension .ISD (InStat Data). Note that each file contains only one data table.

If you want to share data with other programs, use the File Import and Export commands. See "Importing and exporting data" on page 109. Other programs will not be able to open InStat ISD files.

GraphPad Software

Technical support

Do you have the current version?

Like all software companies, GraphPad occasionally issues minor updates to Prism. If you are having trouble with InStat, check that you are running the current release.

The full version number is not on the manual cover or the CD label. You have to run the program and find out which version it is. Drop the Help menu (Windows), Apple menu (Mac OS8-9) or Prism menu (Mac OS X) and choose About InStat. Windows versions have two digits after the decimal point (i.e. 3.05). Mac versions have a single digit after the decimal followed by a letter (i.e. 3.0a).

Go to the Support page at www.graphpad.com to find out what version is most current. Download and install the updater if your version is not the most current. Updates (interim versions of GraphPad software containing bug fixes or minor improvements) are free to owners of the corresponding major releases. In contrast, upgrades (a new version with many new features) must be purchased.

Is the answer to your question on www.graphpad.com?

If you need help using InStat and can't find the answers in this manual, please visit our web site at www.graphpad.com. Your solution is very likely in the searchable Quick Answers Database in the Support section.

You can browse the list of most frequently asked questions, browse questions by topic or search for particular words. We update the Quick Answers database almost every week, and the answer to your question is very likely to be there.

If you have questions about data analysis, also browse the library of statistical articles and links on www.graphpad.com

Personal technical support

If you need personal help, contact us via email at support@graphpad.com or use the form on the support page. Be sure to mention the version of InStat you are running and if you are using InStat for Windows or for Mac.

If you really think that your issue is better solved by a phone call, please email your phone number. We give much higher priority to emailed questions, and you may not get a return call the same day. You will get faster personal support by email than by phoning.

While we reserve the right to charge for support in the future, we promise that you'll receive free support for at least one year.

We can't predict how computer hardware and system software will change in the future, so cannot promise that Prism 3, will work well with future versions of Windows or the Mac OS.

Note that your InStat license does not include free statistical consulting. Since the boundary between technical support and statistical consulting is often unclear, we will usually try to answer simple questions about data analysis.

GraphPad Prism

GraphPad Prism (for Windows and Macintosh) combines scientific graphics, curve fitting (nonlinear regression) and basic statistics.

Instant scientific graphs. Click one button to instantly graph your data. Prism even chooses an appropriate type of graph and creates error bars and legends automatically. Easily change symbols and annotate your graph (including Greek, math and international characters). Once you've created several graphs, arrange them using Prism's unique page layout tools. You can even include data and analysis results on the same page.

Instant curve fitting. Nonlinear regression couldn't be simpler. Just select the equation you want from the list (or enter your own equation) and Prism does the rest automatically - fits the curve, displays the results as a table, and draws the curves on the graph. Even better, Prism will automatically fit all related data sets at once. You don't have to repeat commands for each experimental condition. Prism also gives you many advanced fitting options - automatically interpolate unknown values from a standard curve, compare two equations with an F test, and plot residuals. To review the principles of nonlinear regression, go to www.graphpad.com and read the GraphPad Guide to Nonlinear Regression and the companion GraphPad Guide to Analyzing Radioligand Binding Data.

Clear statistical help. Prism performs the same tests as InStat (except for multiple regression), as well as two-way ANOVA and survival analysis. Like InStat, Prism explains the choices and results in plain language.

Intelligent automation. When you fix a data entry mistake, Prism automatically reanalyzes your data and updates your graph. You don't have to do anything. Because Prism links data to results and graphs, you can analyze data from a repeated experiment in a single step. Just plug in the new data and Prism handles all the graphing and analysis steps automatically - without programming or scripting! Every file you save can be a template for a repeated experiment.

Everything is automatically organized. Prism lets you store multiple data tables in one file, linked to analysis results, graphs, and layouts. Even your most complicated projects stay organized and easy to manage. Unlike other scientific graphics programs, Prism stores analysis results with your

data and remembers your analysis choices. When you open a Prism file, you can retrace every step of every analysis.

Try Prism with your own data. The demo version of Prism has some limitations in printing, exporting and saving - but no limitations in graphing or data analysis. Download it from www.graphpad.com.

Intuitive Biostatistics (book)

If you like the style of this manual, you'll probably also like *Intuitive Biostatistics*, by Harvey Motulsky, president of GraphPad Software and author of this manual. Here is the publisher's description:

"*Intuitive Biostatistics* provides a nonmathematical introduction to biostatistics for medical and health sciences students, graduate students in biological sciences, physicians and researchers. Using nontechnical language, this text focuses on explaining the proper scientific interpretation of statistical tests rather than on the mathematical logic of the tests themselves. *Intuitive Biostatistics* covers all the topics typically found in an introductory statistics text, but with the emphasis on confidence intervals rather than P values, making it easier for students to understand both. Additionally, it introduces a broad range of topics left out of most other introductory texts but used frequently in biomedical publications, including survival curves, multiple comparisons, sensitivity and specificity of lab tests, Bayesian thinking, lod scores, and logistic, proportional hazards and nonlinear regression. By emphasizing interpretation rather than calculation, *Intuitive Biostatistics* provides a clear and virtually painless introduction to statistical principles, enabling readers to understand statistical results published in biological and medical journals."

You can see the table of contents and read five complete chapters at www.graphpad.com. You may order the book from GraphPad Software with software purchases only. To order from a bookstore or the publisher (Oxford University Press), cite this number: ISBN 0-19-508607-4. *Intuitive Biostatistics* is also available from the online bookstore www.amazon.com.

Index

—A—

Adjusted R²104
Analysis templates..... 119
ANCOVA, InStat doesn't do 8
ANOVA table67
ANOVA table from repeated
measures ANOVA..... 71
ANOVA, checklist 62
ANOVA, choosing..... 60
ANOVA, entering data for 58
ANOVA, introduction to..... 58
ANOVA, one-way vs. two-way 58
ANOVA, repeated measures,
checklist 69
ANOVA, repeated measures,
results 70, 71
ANOVA, results 64, 66
Appending results to notes 117
Average data, entering into InStat..... 28

—B—

Bartlett's test for equal variances.....67
Book, Intuitive Biostatistics122

—C—

Case-control study..... 78
Central limit theorem..... 17
Chi-square test for trend 86
Chi-square test, checklist81
Chi-square test, results 83
Chi-square vs. Fisher's test 80
Citing InStat ii
Cluster analysis, InStat doesn't do..... 8
Coefficient of determination,
defined 92
Column titles 114
Comparing models in multiple
regression106
Confidence interval of a mean12
Confidence interval, definition of27
Constrained linear regression, when
to choose..... 89
Contingency table analyses,
checklist81
Contingency tables, creating 78, 116
Correlation coefficient, defined91
Correlation matrix vs. multiple

regression..... 100
Correlation matrix, interpreting.....100
Correlation vs. regression.....89
Correlation, checklist.....90
Correlation, introduction.....87
Correlation, results 91
Cross-sectional study..... 78

—D—

Dallal-Wilkinson method20
Data filtering 111
Data mining, problems with..... 16
Data missing113
Data tables, moving the insertion
point113
Data tables, number format113
Data tables, selecting a range114
Data, excluding113
Data, exporting111
Decimal format of data tables113
Delete command.....115
Deleting data.....115
Dunn's post test, following
Friedman test..... 77
Dunn's post test, choosing 61
Dunn's post test, following
Kruskal-Wallis test 74

—E—

Excel 109
Excluding data113
Exporting data 111

—F—

F test to compare variance, from
unpaired t test.....46
Factor analysis, InStat doesn't do 8
Feature list 7
Files, .ISD..... 120
Filtering data..... 111
Fisher's test vs. Chi-square80
Fisher's test, checklist..... 81
Fisher's test, results83
Format of data tables, numerical113
Friedman test, checklist..... 76
Friedman test, how it works 76
Friedman test, posttests 77

Friedman test, results 76

—G—

Gaussian distribution, defined 17
Gaussian distribution, testing for 19
Gaussian distribution, transforming data to create 19
Graph, viewing 117
GraphPad Prism 121
Graphs, creating with GraphPad Prism 121
Guide, InStat 109

—H—

Hypothesis testing, defined 14

—I—

Ignoring data 113
Importing data 109
Importing indexed data 110
Independence, statistical use of the term 10
Indexed data, importing 110
InStat Guide 109
Intuitive Biostatistics, book 122
ISD files 120

—K—

Kolmogorov-Smirnov test 19
Kruskal-Wallis test 72
Kruskal-Wallis test, checklist 72
Kruskal-Wallis test, how it works 74
Kruskal-Wallis test, posttests 74
Kruskal-Wallis test, results 73

—L—

Likelihood ratio, defined 83
Lilliefors method 20
Limits, maximum number of rows and columns 113
Linear regression vs. correlation 89
Linear regression, checklist 93
Linear regression, introduction 87
Linear regression, results 93
Linear regression, runs test 94
Linear regression, theory 87
List of features 7
Logistic regression, InStat doesn't do 8

—M—

Mann-Whitney test, checklist 53

Mann-Whitney test, how it works 53
Mann-Whitney test, results of 53
Mantel-Haneszel test, InStat doesn't do 8
Maximum number of rows and columns 113
Mean data, entering into InStat 28
Median, defined 28
Missing values 113
Multicollinearity 105
Multiple comparison post tests, choosing 60
Multiple comparisons 16
Multiple regression results, comparing models 106
Multiple regression results, multicollinearity 105
Multiple regression results, which variables count? 104
Multiple regression vs. correlation matrix 100
Multiple regression, checklist 101
Multiple regression, entering data into InStat 99
Multiple regression, how to view results 102
Multiple regression, model and assumptions 97
Multiple regression, uses 97
Multiple regression, what is the best-fit? 102

—N—

Negative predictive value, defined 83
Newman-Keuls post test, choosing 61
Newman-Keuls vs. Tukey post test 61
Nonlinear regression with GraphPad Prism 121
Nonlinear regression, InStat doesn't do 8
Nonparametric posttests 74
Nonparametric tests, choosing 17, 31, 40, 60
Normality test 17, 19
Normalizing transformations 19
Notes editor 117
Null hypothesis, defined 13, 15
Number format of data tables 113

—O—

Odds ratio 83
One sample t test vs Wilcoxon test 30
One sample t test, checklist 32
One sample t test, choosing 30
One-sample t test, results of 32, 34
One-tail P value, defined 13

—P—

P value, one- vs. two-tailed defined	13
P values, common misinterpretation	13
P values, general	12
P1-P2, interpreting	82
Paired t test, checklist	47
Paired t test, results	48, 50
Paired t test, results of	47
Paired tests, when to use	40
Pairing, testing if adequate, paired t test	51
Paste, transpose	115
Pasting data	114
Pearson correlation, introduction	87
Pearson vs. Spearman correlation	89
Population, use of the term in statistics	10
Positive predictive value, defined	83
Post test for linear trend, results	68
Post tests following ANOVA, results	69
Post tests, choosing	60
Prism	121
Prospective studies	78
Prospective study	78

—R—

R squared from correlation	92
R squared from linear regression	94
R squared, adjusted, from multiple regression	104
R squared, from ANOVA	66
R squared, from multiple regression	103
Referring to InStat	ii
Regression vs. correlation	89
Relative risk, interpreting	82
Repeated measures ANOVA, checklist	69
Repeated measures ANOVA, results	70, 71
Resampling approach to statistics	11
Results, appending to notes	117
Retrospective case-control study	78
Row titles	114
Run, defined	90
Runs test following linear regression	94

—S—

Sample, use of the term in statistics	10
Sampling from a population	9
SD vs. SEM	28
SD, definition of	27
Selecting data	114
SEM, definition of	27
Sensitivity, defined	83

Significance, defined	15
Spearman vs. Pearson correlation	89
Specificity, defined	83
Spreadsheet programs, creating text files	109
Stacked data, importing	110
Standard curves, defined	95
Standard curves, entering into InStat	95
Standard deviation, definition of	27
Standard error of the mean, definition of	27
Statistical significance, defined	15
Stepwise regression, InStat doesn't do	8
Support, technical	120
Survival analysis, InStat doesn't do	8

—T—

t ratio, from one-sample t test	34
t ratio, from paired t test	51
t ratio, from unpaired t test	45
t test, one sample, checklist	32
t test, one sample, choosing	30
t test, one sample, results of	32, 34
t test, paired, checklist	47
t test, paired, results	48, 50
t test, unpaired, checklist	41
t test, unpaired, results of	43, 45
t test, Welch's for unequal variances	40
t tests, choosing	39
t tests, entering data into InStat	38
Technical support	120
Templates, creating	119
Test for adequate pairing, paired t test	51
Test for linear trend, choosing	61
Test for trend, chi-square	86
Text, Intuitive Biostatistics	122
Titles, row and column	114
Tolerance	106
Transforming data	115
Transforming data to create a Gaussian distribution	19
Transpose paste	115
Tukey post test, choosing	61
Tukey vs. Newman-Keuls post test	61
Tutorial	21
Two-tail P value, defined	13
Two-way ANOVA, InStat doesn't do	8
Type I error, defined	15

—U—

Unpaired t test, checklist	41
Unpaired t test, results of	43, 45

—V—

VIF (Variance Inflation Factor) 106

—W—

Welch's modified t test..... 40
Welch's t test 40
Wilcoxon matched pairs test, how it
works56
Wilcoxon matched pairs test,

results of 56
Wilcoxon signed rank test, checklist... 35
Wilcoxon signed rank test, choosing... 30
Wilcoxon signed rank test, how it
works..... 36
Wilcoxon signed rank test, results of .. 36
Wilcoxon test vs. one sample t test 30

—Y—

Yates' continuity correction, choosing 80