

APPENDIX I.
DATA AND WHAT TO DO WITH THEM:
A PRIMER

I. Presentation of Data	2
II. Mean and Standard Deviation	2
III. Statistical Analysis	5
A. What statistics can do	5
B. How statistical tests work: sample and population	5
C. Null hypothesis	5
D. Mechanics of using a test	6
E. Choice of an appropriate test	6
IV. Statistical Tests	7
A. Frequency-type data: Chi-square test	7
Goodness-of-fit	7
Contingency table	8
B. Value-type data:	
1. The t -test	9
Unpaired t -test	9
Paired t -test	12
Non-parametric comparisons: Wilcoxon tests	14
2. Analysis of Variance	14
Non-parametric comparisons: Kruskal-Wallis test	17
3. Correlation analysis	17
Non-parametric correlation: Spearman's rho test	19
V. Statistical Tables	
A. Critical values for correlation coefficient (r)	20
B. Chi-square table	21
C. t table	22

I. Presentation of Data

Measured values are often presented in a table with the **mean** (average) value for each group of measurements. A measure of the variation around this mean value is the **standard deviation** (more on that soon).

A more complete way to present data is in a **histogram** -- or bar graph -- of the number of observed data points in certain categories. The example here presents the number of people in certain age classes in a Biology 2 Lab section. A histogram of your data such as this will also be a help in deciding which statistical tests to use later on.

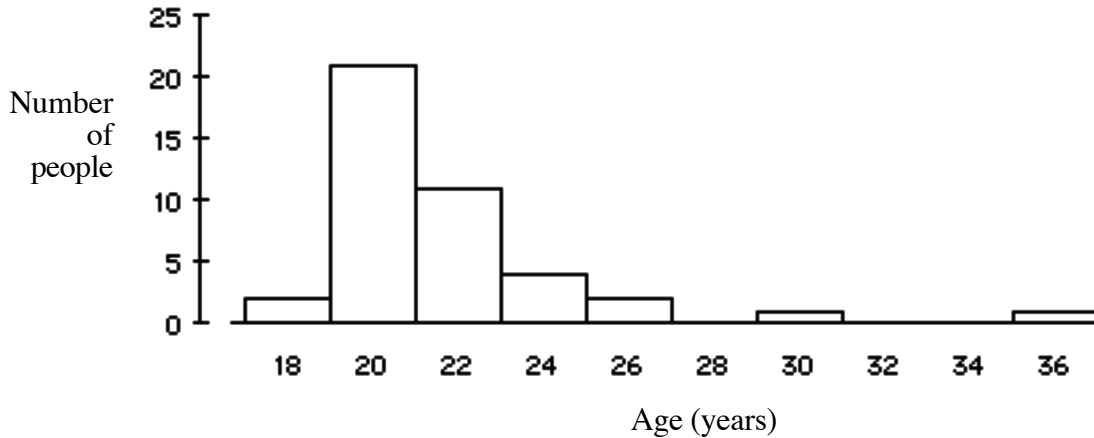


Figure 1. Number of people versus age in years in a sample Biology 2 laboratory.

The numbers along the bottom of a histogram indicate the midpoint of each category, in this case 17 to 18.99 years, 18 to 20.99 years and so on (Figure 1). The height of each bar indicates the number of data points (in this case people) in each category.

What makes a good histogram? A rule of thumb for 10 to 30 data points (such as the data you will collect) is that any histogram has to have at least three bars and that at least half the bars should have more than one data point.

In presenting data be aware of the **sources of error** and the accuracy of your measurements. Do not report units smaller than the accuracy of your measurements. This is especially important when using computers that will compute values for the mean and standard deviation to many decimal places.

II. Mean and Standard Deviation

It is sometimes very useful to condense a set of data to a numerical summary that describes its general pattern and trends. In a size or frequency distribution it is useful to have one number that characterizes the population. So, we often use the average or **mean** value of the population to describe it. The mean is calculated by:

$$\text{Mean value} = \bar{x} = \frac{\text{sum of value of all measurements}}{\text{total number of all measurements}} = \frac{\sum x_i}{n}$$

People are always talking about the mean or average value: batting averages, grade point average, average price of gasoline, but have you ever stopped to wonder why we think so often of average values? Consider a histogram of the height of people in a Bio 2 class (Figure 2).

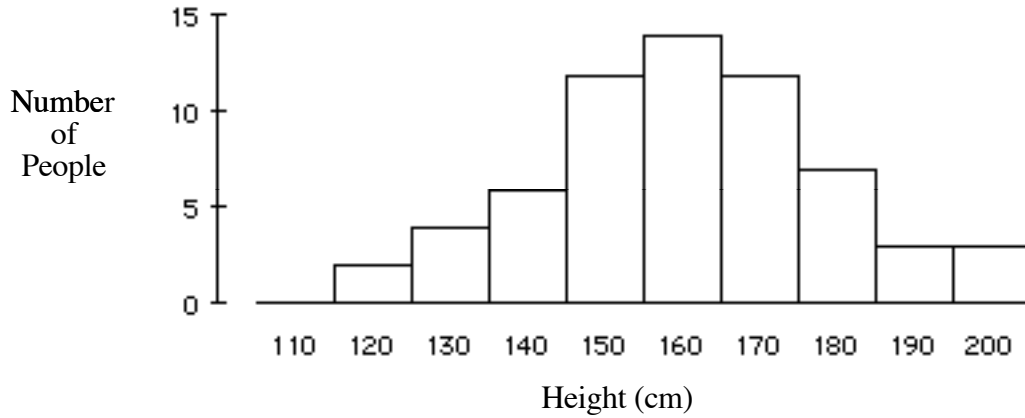
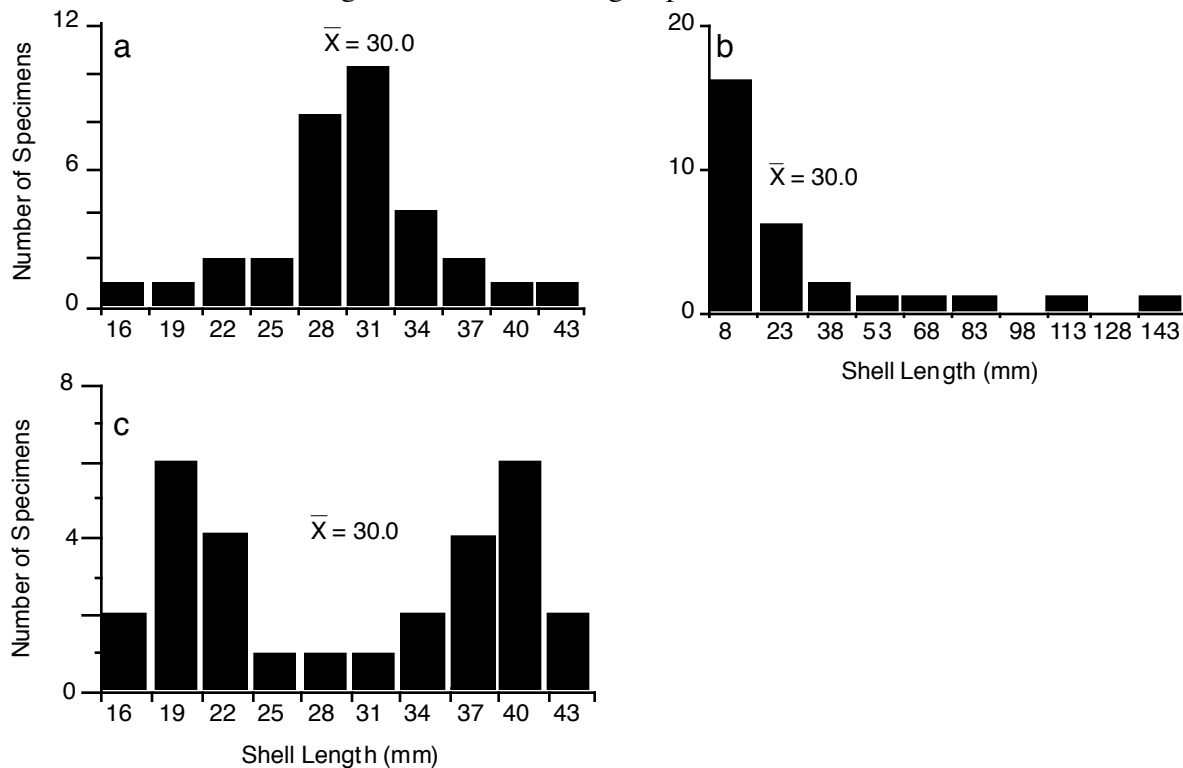


Figure 2. Number of people versus height in a sample Biology 2 class.

For some purposes the average value of 157 cm is a pretty good descriptor of this population; it gives you a reasonable idea of what to expect if you wanted to estimate the height of any person picked at random from the class. Many histograms of biological measurements show a distribution similar to this: highest near the middle and falling off at both ends more or less evenly. These distributions in most cases approximate a special frequency distribution called the **normal distribution** or Gaussian distribution (more information may be found in any statistics textbook).

There are many groups of numbers for which the mean is not a good descriptor: for which the average value is a poor guess of the most common values in the group. Consider the following measurements for shell length of three different groups of clams:



All have the same mean value of 30 mm., but only for the one sample with a normal distribution (where the data are peaked in the middle and falling off at both ends more or less evenly) is the mean a very good descriptor of the sample. For the other distributions the mean does not provide a very useful characterization of the population or give you much sense of what an "average" individual from that population would be like. In one group (bottom), the average is in fact just about the worst guess you could make.

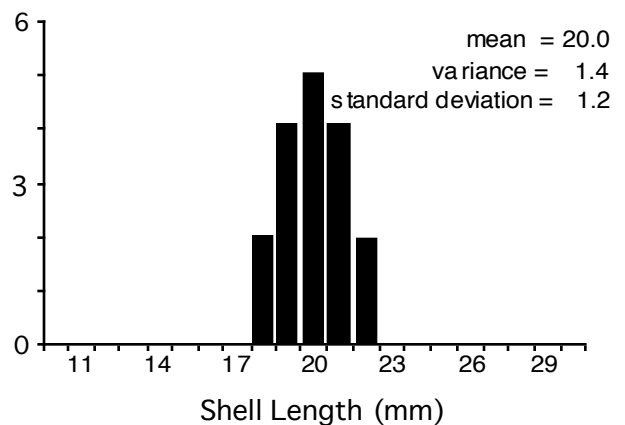
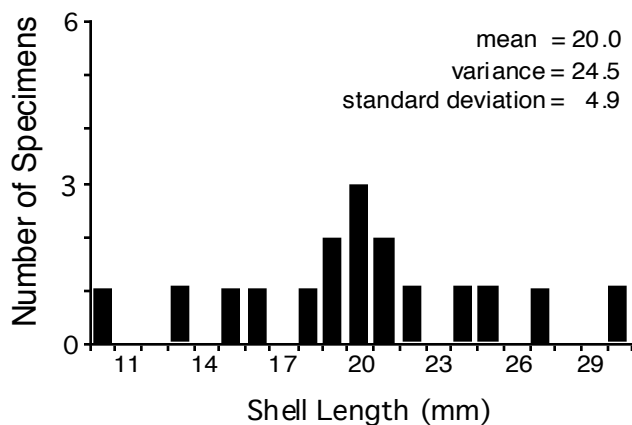
There are rigorous tests that may be applied to data to determine if they may be considered a normal distribution. For our purposes, we will consider a population "normal" if most values fall near the average value and decrease in frequency as you move away from the average on either side, i.e., we will consider a population to be normally distributed if it looks anything like a normal distribution. Believe it or not, this is quite sufficient in most cases. We will consider data to be not from a normal distribution if histograms clearly look "non-normal" (e.g., the distribution is very asymmetric or has more than one pronounced peak).

A second number that we use along with mean to describe a population is the **variance**. It is a numerical way to describe how the population is dispersed on either side of the mean. The variance is calculated by:

$$\text{Variance} = s^2 = \frac{\text{the sum of all (the deviations from the mean)}^2}{\text{total number of measurements}}$$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

If you notice the units you will see that variance is expressed as "units²", which is somewhat difficult to relate to the mean where the units are not squared. For that reason we often take the positive square root of the variance (a quantity that is known as the **Standard Deviation**) to provide us with a measure of dispersion in units that make sense within the context of the particular problem. Below are two graphs that illustrate populations with the same mean, but which vary in variance and standard deviation. In which group is the mean a better predictor of the shell length of a random individual from the population?



Whenever you present a mean or average value for data you should also give the standard deviation. This is important in estimating just how likely it is that any value is significantly different from that mean. Sixty-eight percent of all values in a normal distribution will lie between + and - one standard deviation from the mean, 95% between + and - 1.96 standard deviations, and 99% between + and - 2.58 standard deviations from the mean. In the next section we consider how to compare sets of data to see if they differ significantly.

III. Statistical Analysis

You have done an experiment or made some observations. You have presented the data in the form of histograms or graphs. In many cases you want to compare two sets of data such as experimental and control data. What do you do next? The following is intended to help you analyze those numbers you have collected and determine whether there are significant differences between groups. This is a minimal treatment; for a complete discussion we strongly recommend a course in statistics.

A. What Statistics Can Do

The purpose of an experiment is not limited to investigating the results of the experiment. Scientists use experiments to make estimates about the world outside the laboratory. Thus they are not observing the entire population of organisms or objects with which they are concerned; they are observing only a small **sample** of that total population. Statistics allow us to estimate the probability that our results may correctly reflect aspects of the total population. Statistics also allow us to avoid repeating an experiment unnecessarily when the results have shown an established trend and to avoid very large experiments with large numbers of repeated trials. Thus statistics can be used to reduce the work required to substantiate a point and to allow a scientist to concentrate his or her energies where they will be most useful.

B. How Statistical Tests Work: Sample and Population

Let us say you have a bag with red and white jellybeans. You grab a handful and get 7 red and 3 white jellybeans. Does that show that the bag contains 70% red jellybeans? We call the handful a **sample**. The sum of all the red and all the white jellybeans in the bag is called the **total population**. Obviously you would not be confident about the percentage of red in the total population of jellybeans from just one sample. Suppose you sample repeatedly and get 65%, 72%, and 80%. Would you now feel more confident that the bag contains about 70% red jellybeans? You are now given another bag of jellybeans and grab a handful: 6 red and 4 white. You are asked if this bag contains the same percentage of red jellybeans as the first bag. This is basically the same question asked in most scientific experiments: Given a number of samples from two or more groups, are the total populations different or the same?

C. Null Hypothesis

Statistical tests cannot "prove" that a statement is true. In the example above we could never know the true ratio of red to white jellybeans in the bags until we actually counted all the beans in the two bags. However, we can set up a hypothesis such as "the bags have the same percentage of red and white jellybeans." Implicit in this is an alternative hypothesis, **the null hypothesis**: There is no difference in the percentage of red and white jelly beans. A null hypothesis is the one that makes the *fewest* claims about a situation. If your hypothesis is that people who take vitamin C get over their colds faster than those who do not, the null hypothesis is that there is no difference in the duration of colds for people who take vitamin C and people who do not.

A statistical test works by asking the question, "If the null hypothesis were true, how probable would it be for us to obtain the data at hand?" If this probability is very low we can *reject the null hypothesis with a certain degree of confidence, and conclude that there really was a difference between the populations* (the experimental treatment really did have some effect). A probability value of more than 0.05 ($P > 0.05$) is taken to indicate that the observations or experimental results show no evidence of a difference. If the probability is less than 0.05 ($P < 0.05$) we conclude that a significant difference between the data populations exists. But we must be cautious because $P < 0.05$ means that we could get such a result 5 out of 100 times even if the null hypothesis were correct. A probability of less than 0.01 ($P < 0.01$) is usually considered significant for scientific

work. (A scientist using this level of confidence would misinterpret the results of less than 1 out of every 100 experiments.)

D. Mechanics of Using a Test

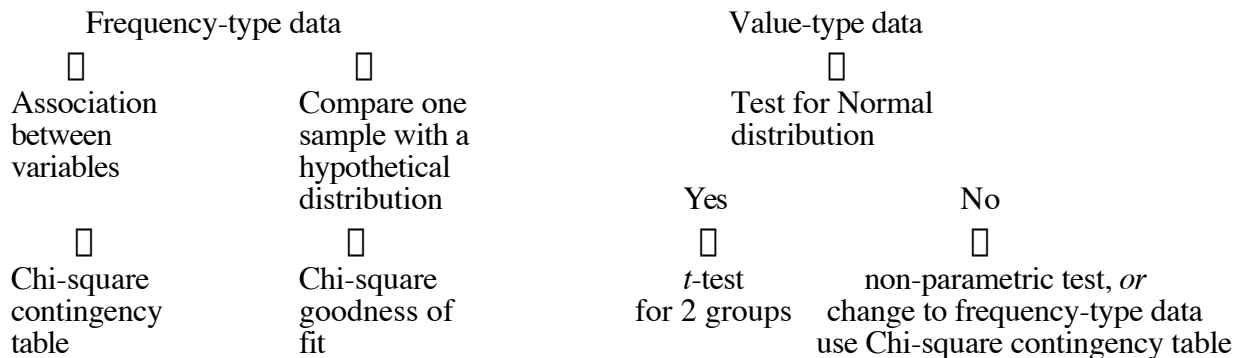
After selection of a suitable test (see below) you can compute a number (the **test statistic**) from your data using the appropriate formula for that test. At the end of this Appendix you will find tables of **critical values** of some test statistics. These are listed by **degrees of freedom** (derived from the experimental design) and probability that the null hypothesis is correct. You compare your computed value of the test statistic (for a given number of degrees of freedom) with values given in the tables. If your value *equals or exceeds* the value in the table you *may reject the null hypothesis* at the probability level indicated. If your value is less than the value given in the table you must accept the null hypothesis at the probability level indicated. This method will be demonstrated in the examples that follow each of the tests in this laboratory.

E. Choice of an Appropriate Statistical Test

The choice of an appropriate test depends on the hypothesis to be tested and on the type of data gathered. All of the tests treated in this section are for differences between groups of data. In the next section we present a test for correlations within one group of data.

There are basically two sorts of data: **Frequency-type data** and **value-type data**. Frequency-type data result from counting the number of times an event occurs: the number of times more than 10% of the birds on a field sprayed with pesticide died, the number of women taller than 180 cm, the number of people who are both left-handed and female. Value-type data result from measurement of some value for each case: The length of time (in hours) that birds live after feeding on a field sprayed with pesticide, the height (in cm) of women. Any time you are reading a meter, or measuring with an instrument, you are collecting value-type data. Note that value-type data can be converted into frequency-type data by dividing the measurements into groups and checking the number of values obtained in each group (for instance women taller than 180 cm) but frequency-type data cannot always be converted into value-type data (there are no degrees of deadness!). The following is a flow chart to help you choose among the statistical tests discussed in this Appendix. Details of each test are given in the next section.

Some Tests for Differences between Groups of Data for Bio 2



IV. Statistical Tests

A. Frequency-type Data: the χ^2 (Chi-square) Test

The test is based upon comparing observed and expected frequencies within a number of discrete categories. The data here consist of the number of times different cases occur.

FORMULA:

$$\chi^2 = \sum \frac{(\text{Obs.} - \text{Exp.})^2}{\text{Exp.}}$$

χ^2 = Chi square (test statistic)
 Obs. = observed values
 Exp. = calculated expected values

The Σ sign means the sum of all the following values.

NULL HYPOTHESIS: There is no difference between the observed and expected frequencies of data. If the value of χ^2 is equal to or greater than the critical value given for your degree of freedom (see below) you may reject the null hypothesis and conclude that there is a significant difference in the distribution of observations among categories between the two populations.

USES: χ^2 tests may be used for actual numbers of observations; they may not be used for percent or average values. No more than 20% of the expected values should be less than 5. No expected value should be less than 1 (see below).

There are two main uses of χ^2 : **goodness-of-fit** and **contingency table**. An example of each is given below.

EXAMPLE 1: Goodness-of-Fit - χ^2 Test

You watch birds arriving at a feeder and count 56 sparrows, 71 finches and 83 chickadees. Is there a significant difference in the number of individuals of these three species visiting the feeder?
Null hypothesis: There is no significant difference between the number of individuals of each species visiting the feeder.

Degrees of freedom (d.f.) = (number of categories) - 1. d.f. = 3 - 1 = 2

Formula:

$$\chi^2 = \sum \frac{(\text{Obs.} - \text{Exp.})^2}{\text{Exp.}}$$

Since the total number of individuals that visited the feeder was 210, the expected values for the null hypothesis are 70, 70, and 70 for each species.

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(71-70)^2}{70} + \frac{(83-70)^2}{70} = 2.80 + 0.01 + 2.41 = 5.22$$

Critical value is 5.991 for d.f. = 2 at P = 0.05. Thus you cannot reject the null hypothesis and must conclude that there was no statistically significant difference in the numbers of birds you saw visiting the feeder.

The goodness-of-fit χ^2 test is difficult to do on a computer but easy to do by hand as above.

EXAMPLE II: Contingency Table - χ^2 Test

One of the more common uses of the χ^2 test is in a **contingency table** in which we test for an **association** between two sets of categories. For example, suppose we wanted to test whether women were more apt to be left-handed than men. We tested 30 people and found the following data:

10 left-handed women	7 right-handed women
4 left-handed men	9 right-handed men

Null hypothesis: There is, in fact, no significant relationship between the categories and our observed distribution of data was due to chance alone.

We can calculate how many people would be expected in each category if there were no association between sex and left-handedness. To do this in a contingency table, first add up data in each row and column and get the totals for each row and column (called marginal totals) as in the example below. The computer will do this for you but you must understand what it is doing.

OBSERVED VALUES:	Left	Right	Row Total
Women	10	7	17
Men	4	9	13
Column Total	14	16	30 = Grand Total

We see that 17/30 of the total subjects are women. Under the null hypothesis, i.e., no association between sex and handedness, we would expect that 17/30 of all 14 left-handed people in the study would be women. Thus, $17/30 \times 14 = 7.9$ for the expected value in the upper left hand box. In general the expected value in a contingency table for the hypothesis of no association between variables is:

$$\frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}} = \text{Expected Value}$$

Repeating the calculation for each box of the contingency table we get:

EXPECTED VALUES:	Left	Right	Row Total
Women	7.9	9.1	17
Men	6.1	6.9	13
Column Total	14	16	30 = Grand Total

Degrees of freedom: The number of degrees of freedom in a contingency table is given by (number of rows - 1) x (number of columns - 1).

Formula:

$$\chi^2 = \sum \frac{(\text{Obs.} - \text{Exp.})^2}{\text{Exp.}}$$

Once you have calculated χ^2 and the degrees of freedom, you can compare your χ^2 value with those in the attached table (p. I-22). Under 1 degree of freedom and $P = 0.05$ the table gives 3.841. This means that if we did an experiment 100 times we would get a χ^2 value of 3.841 or greater an average of 5 times due to chance alone. The value under $P = 0.01$ shows that if we did an experiment with 1 degree of freedom 100 times we would be likely to get a χ^2 value of 6.635 only once due to chance alone. In our example the value of χ^2 we obtained was smaller than 3.841 which indicates that our observed distribution of data would occur more often than 5 times out of each 100 experiments. We therefore cannot reject the null hypothesis that there is no association between handedness and sex. We could decide to repeat the experiment interviewing more people, but without other evidence the best course would be to try a different type of experiment. If our value of χ^2 had fallen between 3.841 and 6.635 we would conclude that there was evidence to support an association between handedness and sex, but we would be hesitant to base our reputation or plan a large scale research project without further work because, as discussed above, this result could happen by chance alone 5 times out of 100. If the χ^2 value were greater than 6.635 we would confidently reject the null hypothesis that there was no association between handedness and sex, knowing that result could have occurred by chance only once in 100 experiments.

B. Value-type Data: Normally Distributed

The t -Test

The t -test is used to test a hypothesis about the mean. One type of t -test, the unpaired t -test, is used to determine whether two samples (e.g., control and experimental data groups from an experiment or observations) are likely to have been drawn from the same population (no significant difference between the means of two samples) or from different populations (a significant difference between the means of two samples). A related type of t -test, the paired t -test, is a special case in which we can make use of the fact that data in the two groups we wish to compare are paired in some way.

Paired or Unpaired? Deciding which test to use

If you have unequal numbers of data points in your two samples OR if the data points in the two samples come from different individuals, use an unpaired t -test. If the two samples consist of pairs of data, one from each individual tested in each sample, use a paired t -test.

Unpaired t -Test

The unpaired t -test asks the following question: "Is the difference between the means of two groups significantly different from zero?" It tests the null hypothesis that the difference between the means is zero by constructing a confidence interval around the difference between the means of the two groups. If we wish to examine the hypothesis at the level of $P < 0.05$, we need to construct a 95% confidence interval around the mean difference. Once we have constructed this interval, it means that we are 95% certain that the actual mean difference for the population lies within this interval. If zero lies within this interval, then we must accept the null hypothesis. If zero does not lie within the 95% confidence interval around the mean difference, then we can reject the null hypothesis and conclude that there is a significant difference between the two samples at the level of $P < 0.05$.

Since the *t*-test relies on mean values, the samples must be from a population of data that are normally distributed in order for the mean value to be useful.

There are three assumptions:

1. sample 1: $x_1, x_2, x_3, \dots, x_n$ is a random sample from a normally-distributed population.
2. sample 2: $y_1, y_2, y_3 \dots y_m$ is a random sample from a normally-distributed population.
3. the variances (see Appendix I-4) of the two parent populations are equal (this can be tested but, for now, we will make this assumption).

FORMULAS:

lower boundary of confidence interval

$$= (\text{mean difference of two samples}) - (t\text{-value for d.f.})(\text{standard error of mean difference})$$

$$= (\bar{x} - \bar{y}) - (t) \sqrt{s_p^2 \left(\frac{1}{n} + \frac{1}{m} \right)}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

- \bar{x} = the mean for the first sample
- \bar{y} = the mean for the second sample
- n = number of data points in the first sample
- m = number of data points in the second sample
- s_x = standard deviation of first sample
- s_y = standard deviation of second sample

Degrees of Freedom: d.f. = $n + m - 2$

upper boundary of confidence interval

$$= (\text{mean difference of two samples}) + (t\text{-value for d.f.})(\text{standard error of mean difference})$$

To perform the test by hand:

1. Check the histograms you have made for the two samples to be tested. If they do not meet the criteria for normal distribution outlined previously, divide your data into classes and use a contingency table Chi-square test.
2. If the samples appear to be normal, calculate the degrees of freedom and use the table at the end of this Appendix entitled "Student's *t*-distribution" to find the appropriate *t*-value for your computation. To construct a 95% confidence interval ($P < 0.05$ if the null hypothesis is rejected) use the *t*-values in the column labeled "0.05." To construct a 99% confidence interval ($P < 0.01$ if the null hypothesis is rejected), use the *t*-values in the column labeled "0.01." Compute the upper and lower boundaries of the confidence interval as shown above.
3. Your confidence interval goes from the lower boundary to the upper boundary. Does zero lie within this range? If so, you cannot reject the null hypothesis at the P-level you have selected. If not, you can reject the null hypothesis and conclude that there is a significant difference between the two means at the P-level you have selected.

To perform the test by computer:

1. Check the histograms you have made for the two samples to be tested. If they do not meet the criteria for normal distribution outlined previously, divide your data into classes and use a contingency table Chi-square test, or use a non-parametric test.
2. Follow the instructions for computing the unpaired t -test in Appendix II, "Using JMP". Note that data for the unpaired t -test must be arranged in a single column, with a second column that acts as an indicator column.

EXAMPLE:

Seven plants of wheat grown in pots and given a standard fertilizer treatment produced 8.4, 4.5, 3.8, 6.1, 4.7, 11.2 and 9.6 g dry weight of seed.

Another 8 plants from the same source are grown with new Super-Whammo fertilizer and produced 11.6, 7.5, 10.4, 8.4, 13.0, 7.0, 9.6, 13.2 g of seed.

Is Super-Whammo worth the extra cost?

$$\text{For the first sample} \quad \bar{x} = 6.9, \quad s^2 = 8.1, \quad n = 9$$

$$\text{For the second sample} \quad \bar{y} = 10.1, \quad s^2 = 5.6, \quad n = 8$$

$$\text{Mean difference} \quad \bar{x} - \bar{y} = 3.2$$

$$s_p^2 = \frac{(7-1)8.1 + (8-1)5.6}{7+8-2} = \frac{(6)8.1 + (7)5.6}{13} = \frac{87.8}{13} = 6.75$$

$$\sqrt{s_p^2 \frac{1}{n} + \frac{1}{m}} = \sqrt{6.75 \left(\frac{15}{56} \right)} = 1.34$$

The degrees of freedom = $7 + 8 - 2 = 13$. Consulting the t -table we find that at d.f. = 13 and $P = 0.05$, $t = 2.16$. Therefore,

$$\text{lower boundary of 95\% confidence interval} = 3.2 - (2.16)(1.34) = 0.31$$

$$\text{upper boundary of 95\% confidence interval} = 3.2 + (2.16)(1.34) = 6.09$$

Zero does *not* lie within the range 0.31 to 6.09; we therefore reject the null hypothesis at the level of $P < 0.05$.

At $P = 0.01$, however, $t = 3.01$.

$$\text{lower boundary of 99\% confidence interval} = 3.2 - (3.01)(1.34) = -0.83$$

$$\text{upper boundary of 99\% confidence interval} = 3.2 + (3.01)(1.34) = 7.23$$

Zero *does* lie within the range -0.83 to 7.23; we therefore cannot reject the null hypothesis at the level of $P < 0.01$.

We therefore conclude that Super-Whammo can be considered better, but that we might come to that conclusion 5 times out of 100 tests by chance alone when there is no real difference between the two groups.

Paired *t*-Test

There are always differences among the individuals in a population. For example, although some lizards run faster than others, all are affected the same way by large changes in temperature. Say we race 10 lizards at each of two temperatures. To compare the running speeds at these two temperatures, we can make use of the fact that we have data for each lizard at each temperature by using a paired *t*-test. The paired *t*-test takes the difference of the measurements for each lizard *at the start of the analysis*, with the result that the test "takes into account" the differences in the natural running abilities of the individual lizards in the experiment. This is an especially powerful tool when there are large individual differences in the basal levels of the variable you wish to investigate.

The paired *t*-test relies on the same reasoning as the unpaired *t*-test, except that in this case the confidence interval is constructed around the mean of the paired differences for each individual.

As with the unpaired *t*-test, an assumption must be met before you can apply this test:

The paired differences $(x_1 - y_1), (x_2 - y_2), (x_3 - y_3) \dots (x_n - y_n)$ represent a random sample from a normally-distributed population.

FORMULAS:

lower boundary of confidence interval

= (mean of differences) - (t-value for d.f.)(standard error of mean of differences)

$$= (\bar{x} - \bar{y}) - (t) \frac{s}{\sqrt{n}}$$

\bar{x} = the mean for the first sample

\bar{y} = the mean for the second sample

n = number of pairs of data points

s = standard deviation of mean difference

Degrees of Freedom: d.f. = n - 1

upper boundary of confidence interval

= (mean of differences) + (t-value for d.f.)(standard error of mean of differences)

To perform the test by hand:

1. Create a single column of numbers that are the differences for each pair of data. Check the histogram of these data to ensure that the data are normally distributed. If they do not meet the criteria for normal distribution outlined previously, divide your data into classes and use a contingency table Chi-square test, or use a non-parametric test.
2. If the differences appear to be normally distributed, calculate the degrees of freedom and use the table at the end of this Appendix to find the appropriate *t*-value for your computation. To construct a 95% confidence interval ($P < 0.05$ if the null hypothesis is rejected) use the *t*-values in the column labeled "0.05." To construct a 99% confidence interval ($P < 0.01$ if the null hypothesis is rejected), use the *t*-values in the column labeled "0.01." Compute the upper and lower boundaries of the confidence interval as shown above.

3. Your confidence interval goes from the lower boundary to the upper boundary. Does zero lie within this range? If so, you cannot reject the null hypothesis at the P-level you have selected. If not, you can reject the null hypothesis and conclude that there is a significant difference between the two groups of data at the P-level you have selected.

To perform the test by computer:

1. Check the histograms you have made for the two samples to be tested. If they do not meet the criteria for normal distribution outlined previously, divide your data into classes and use a contingency table Chi-square test.
2. Follow the instructions for computing the paired t -test in Appendix II, "Using JMP." Note that data for a paired t -test must be arranged in two matching columns.

EXAMPLE:

In your days as a youth by the swimming pool, you notice that you can hold your breath under water longer if you hyperventilate first. Does hyperventilation really make a difference? You assemble some friends and ask them to hold their breath with and without prior hyperventilation, timing the length of breath-holding (in seconds) in each case for each person. You collect the following data:

Name	With Prior Hyperventilation	Without Prior Hyperventilation	Difference (x-y)
Buster	60	45	15
Pele	30	32	-2
Sara	50	45	5
Poindexter	75	65	10
Phoebe	45	35	10
Mary	65	60	5
Mean	54.2	47.0	7.2
s	15.94	13.19	5.84
s ²	254.1	174.0	34.1
s.e.m.	6.51	5.39	2.39

The degrees of freedom = $6 - 1 = 5$. Consulting the t -table we find that at d.f. = 5 and $P = 0.05$, $t = 2.57$. Therefore,

lower boundary of 95% confidence interval = $7.2 - (2.57)(2.39) = 1.03$

upper boundary of 95% confidence interval = $7.2 + (2.57)(2.39) = 12.13$

Zero does *not* lie within the range 1.03 to 12.13; we therefore reject the null hypothesis at the level of $P < 0.05$ and conclude that prior hyperventilation does increase breath holding ability.

Now compare the result for the paired t -test you have just performed with the result for an *unpaired* t -test performed on the same data:

$$s_p^2 = \frac{(6-1)254.1 + (6-1)174.0}{6 + 6 - 2} = \frac{(5)254.1 + (5)174.0}{10} = \frac{2140.5}{10} = 214.0$$

$$\sqrt{s_p^2 \left(\frac{1}{n} + \frac{1}{m} \right)} = \sqrt{214.0 \left(\frac{12}{36} \right)} = 8.45$$

The degrees of freedom = $6 + 6 - 2 = 10$. Consulting the t -table we find that at d.f. = 10 and $P = 0.05$, $t = 2.23$. Therefore,

$$\text{lower boundary of 95\% confidence interval} = 7.2 - (2.23)(8.45) = -11.64$$

$$\text{upper boundary of 95\% confidence interval} = 7.2 + (2.23)(8.45) = 26.04$$

Zero *does* lie within the range -11.64 to 26.04; we therefore cannot reject the null hypothesis at the level of $P < 0.05$, and we would have to conclude that prior hyperventilation does not increase breath-holding time.

In this example, there was great variability among the subjects, both in their basal breath-holding ability and in the effect of prior hyperventilation on breath-holding ability. However, in all but one case, prior hyperventilation resulted in an increase in breath-holding duration. The paired t -test, which examines the mean of the individual differences for the two treatments, shows that the effect of prior hyperventilation is statistically significant at $P < 0.05$, but the unpaired t -test, which examines the difference between the means of the two treatments, does not allow us to reject the null hypothesis. Which result should you use? Whenever you have the ability to use a paired t -test, you should do so because, by taking into account individual differences, the paired t -test makes the best use of the data you have collected.

Non-Parametric Comparisons

The t -test assumes that both of the samples being compared are drawn from normally distributed populations (the unpaired t -test), or that the differences between paired measurements are normally distributed. Non-parametric comparisons are statistical tests that do not rely on an assumption of normality.

The **Wilcoxon rank-sum test** compares two unpaired samples, and the **Wilcoxon signed rank test** evaluates the differences between paired samples. Both of these tests rank the data (from lowest value to highest) and are based on the ranks, rather than the actual values observed.

Analysis of Variance

An Analysis Of Variance (**ANOVA**) is similar to a t -test, but it **compares the mean value between more than two groups**. For example, you want to compare mean feeding rate of birds observed under trees, in the open, and in the tree canopy.

ANOVA allows you to compare all three (or more) groups simultaneously. Here's how it works:

Remember that **variance** is a numerical way to describe how the population is dispersed on either side of the mean. The variance is calculated by:

$$\text{Variance} = s^2 = \frac{\text{the sum of all (deviations from the mean)}^2}{\text{total number of measurements}}$$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The standard deviation is the square root of the variance.

ANOVA detects differences between means of groups by comparing the variance **within** groups to the variance **between** groups. You will appreciate that if there are differences between the means, the variance between groups will be much greater than the variance within groups. If there are no significant differences between means, then the variance between groups will be approximately equal to the variance within groups.

ANOVA makes use of a quantity called the "sum of squares" to calculate the variance. The sum of squares is:

$$\sum (x_i - \bar{x})^2$$

You can see that dividing the sum of squares by n-1 will give you the variance. The *total* sum of squares is calculated for the entire data set, ignoring group membership. The total sum of squares can be partitioned into the *treatment* sum of squares (between groups) and the *error* sum of squares (within groups). Divided by the appropriate quantity, each of these measures of variance is called the "mean square".

This simple example will illustrate how effective this approach really is. Consider some measurement made on three groups of specimens:

Group A (feed under trees)	Group B (feed in open)	Group C (feed in canopy)
2	4	6
3	5	7
4	6	8
Group A	Group B	Group C
means: 3	5	7
overall mean:	5	

The sum of squares for each group is:

$$\begin{matrix} (2-3)^2+(3-3)^2+(4-3)^2 & (4-5)^2+(5-5)^2+(6-5)^2 & (6-7)^2+(7-7)^2+(8-7)^2 \\ =2 & =2 & =2 \end{matrix}$$

Adding these together, you get a measure of variability within groups:

$$SS_{\text{within}} = 2+2+2=6$$

The total sum of squares is calculated similarly, from the differences between each observation and the overall mean:

$$SS_{\text{total}} = (2-5)^2 + (3-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + \dots = 30$$

It is clear that the estimate of the variability based on the overall mean is much larger than the within-group variability. The reason of course is that there is a difference between means of the three groups. Since variance within groups is due to factors other than group membership (the effect that you are interested in), this quantity is often referred to as the "residual" or "error".

The ratio of the variance between groups ($SS_{\text{treatment}}/df$) and the variance within groups (SS_{error}/df) is the **F-value**. This statistic is evaluated objectively by comparison to a critical value, just as the t -value is evaluated in a student's t -test. Values of F greater than the critical value at some level of error (again, we usually use $\alpha=0.05$) indicate that the probability of error in rejecting the null hypothesis (there is *no* difference between means) is low.

The ANOVA table that JMP or some other statistical software would generate for the data given above would look like this:

Source	df	Sum of Squares	Mean Square	F-Value	P-Value
Treatment	2	24	12	12	0.0080
Residual	6	6	1		
Total	8	30			

Because the total sum of squares can be partitioned in this way, ANOVA can be used to evaluate the variance due to multiple factors in a more complicated experiment, as well as interaction between factors. For example, suppose that in addition to the effect of group membership, we wished to evaluate the effect of the sex of each specimen on the value of our measurement. We could set up a "two-way" ANOVA to analyze those effects, and the significance of the interaction between sex and group membership (if the effect of group membership depends on the sex of the specimens considered, then there is a significant interaction between the two factors).

ANOVA is more appropriate than t -tests for multiple groups, even though the calculations are more complicated!

Here's why:

When a student's t -test is used to compare the means of two groups of measurements, the t -statistic is computed from the means of both groups, the standard deviations, and the sample sizes. This statistic is evaluated by comparison to a critical value of t from a table of the t -distribution. The critical value is based on a given level of confidence (usually 0.05 for most scientific work) for a given number of degrees of freedom (based on the number of observations used to calculate the means and standard deviations).

The confidence (α) is the probability that you will conclude that there *is* a difference between groups when there really is none ("Type I error"). We would like to minimize this (of course), so we tend to be conservative and set α low. Especially if there is a large amount of variation *within* groups, a smaller probability of error requires a larger difference between means to detect a significant difference between groups. *Each* comparison involves that probability of error.

So if you wish to compare three groups of measurements, you would have a certain probability of error for **each** of the three necessary comparisons:

$$0.05 \times 3 = 0.15 \text{ probability of error overall! Unacceptably high....}$$

ANOVA provides better discrimination and lower probability of error than a *t*-test.

Non-Parametric Comparisons -- Kruskal-Wallis test

Like the *t*-test, ANOVA assumes that the samples being compared are drawn from normally distributed populations. A non-parametric equivalent to ANOVA that does not rely on an assumption of normality is the **Kruskal-Wallis test**. Like the non-parametric Wilcoxon tests that are equivalent to *t*-tests, the Kruskal-Wallis test ranks the data (from lowest value to highest) and comparisons between samples are based on the ranks, rather than the actual values observed.

Correlation Analysis

So far we have compared populations using Chi-squared and analyses comparing means or mean ranks between groups. Often we want to look at the relationship between two variables within a population. Correlation analysis can sometimes be used to estimate the degree to which two variables vary together, i.e., the degree to which one variable increases or decreases as the other increases. Figure 5A is an example of positively correlated variables; Figure 5B is an example of negatively correlated ones. Figure 5C shows a case in which the variables are not correlated. The variables must be numerical value-type data or at least be able to be treated as such, i.e., they cannot be frequency-type data. Thus, we can look for a correlation between people's height and weight, but we cannot use correlation to look for a relationship between weight and gender or left versus right handedness; this would require a different statistical method.

Correlation, like all statistical methods, makes certain assumptions about the data. The statistical analysis may not be valid if the assumptions are not met. Suppose we ask whether height is correlated with weight in humans. We gather a random sample of humans and measure height and weight of each. The first step in data analysis is to make a plot of the data to see if they are consistent with the assumptions of the correlation model. For correlation analysis, the data should have a certain type of distribution called a **bivariate normal distribution**. For our purposes this means that a plot of variable 1 versus variable 2 with a reasonable sample size should be circular or elliptical in overall shape, and there should be more points toward the center of the circle or ellipse than toward the periphery. Figures 5A-C show data that are consistent with the assumptions of correlation analysis, while Figures 5D-F show data that are not consistent with the assumptions of correlation; in the latter cases correlation analysis would not be appropriate.

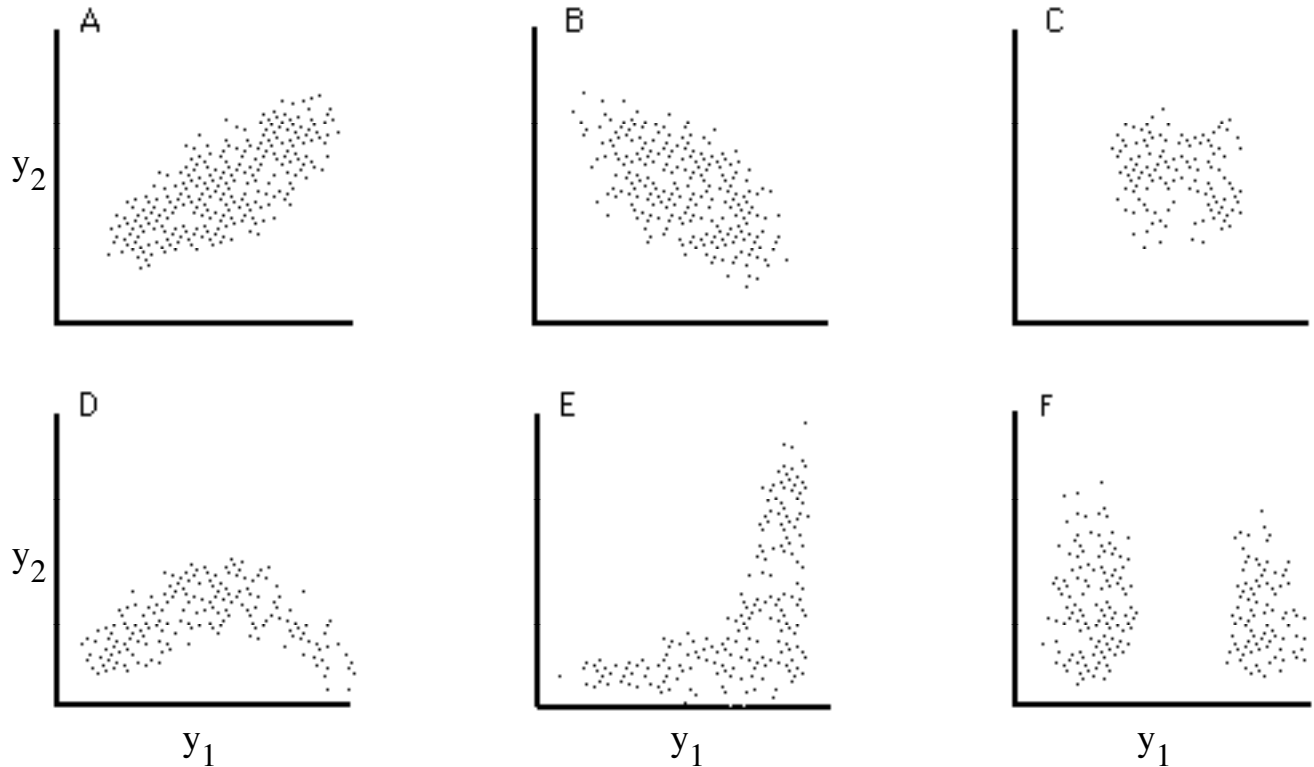


Figure 5. Examples of scatter plots.

If the data are distributed in a way that is consistent with the assumptions of correlation analysis, the correlation coefficient, r , will be meaningful

$$r = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2 y_2^2}}$$

where

r is the correlation coefficient

y_1 and y_2 are the deviations of variable 1 and variable 2 from the mean for each case

r is a measure of how much y_1 and y_2 vary together in the sample. $r=1.0$ means a perfect positive correlation (Figure 6A); $r=-1.0$ means a perfect negative correlation (Figure 6B). An r of 0.7 might look like Figure 5A and an r of -0.7 might look like Figure 5B.

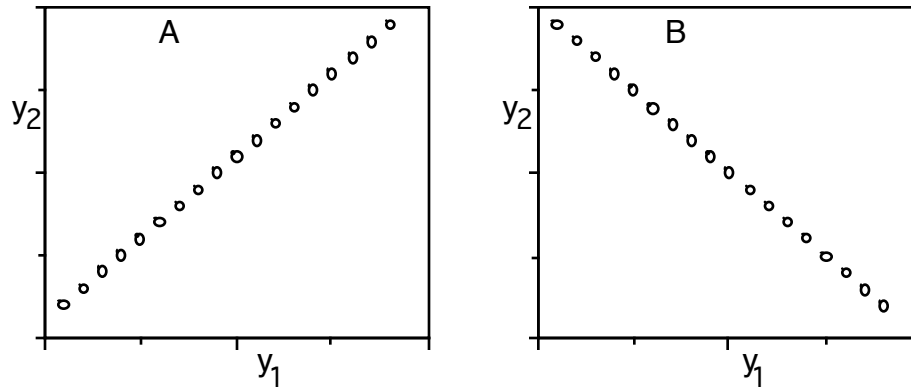


Figure 6. A. Perfect positive correlation; $r = 1.0$. B. Perfect negative correlation, $r = -1.0$.

A correlation coefficient can be calculated from the equation given above. Alternatively, many calculators are programmed to calculate a correlation coefficient. Finally, one can use a statistical package on a Macintosh to perform the calculations. .

To reiterate the discussion on pp. 5-6 of this Appendix, we use statistics to analyze a sample with the hope of being able to say something about the whole population from which the sample was taken. We are ultimately interested in the population, not the sample itself. The first and most important question to ask about a calculated correlation coefficient is "Is the correlation coefficient significantly different from zero?" Put another way this question is "What is the probability of getting a sample similar to the one we are analyzing from a population in which there is no correlation between the variables?" (i.e., if the null hypothesis were true) or "What is the probability that the correlation coefficient calculated from our sample is due to chance alone, and does not reflect a correlation in the population?" For example, a correlation coefficient can be high yet not significant if the sample size is too small. There is a certain probability of randomly choosing three points in a straight line from a totally random distribution. This is why the correlation coefficient must be greater than 0.95 to be significant if the sample size is only four. Conversely, a correlation coefficient can be very low yet still significant (i.e., significantly different from zero) if the sample size is large. In this case we may still be able to say that less than one time in twenty ($P < 0.05$) or less than one time in a hundred ($P < 0.01$) would we get such a correlation coefficient (or one more extreme) if there was no correlation between the variables in the population. The significance of the correlation coefficient says nothing about the strength of the relationship, it just tells us how likely it is to get such a correlation coefficient if there were no correlation between the variables in the population. The probability that a correlation observed in a sample reflects the existence of one in the population is a function of the value of the calculated correlation coefficient and the number of "degrees of freedom" (which equals the sample size minus two). For each correlation coefficient and degrees of freedom there is a probability that the correlation coefficient is significantly different from zero. We can calculate this probability from known distributions or find it in a table that has been calculated from such distributions. Table 1 on the next page is an example of such a table.

Non-parametric correlation -- Spearman's Rho test

If the scatterplot of the data does not show a bivariate normal distribution, then **Spearman's rho test** is a more appropriate analysis than pairwise correlation. Like the other non-parametric statistical tests that we have discussed, Spearman's rho is calculated from the ranks of the data instead of from the values.

Critical Values for Correlation Coefficients

This table furnishes 0.05 and 0.01 critical values for the correlation coefficient r . These values are given for every degree of freedom, ν , between 1 and 30 and for selected values of ν 30 and 1000 ($\nu = n-2$). This table is used to test the null hypothesis that the correlation coefficient of the population from which the sample is taken is zero. To test the significance of a correlation coefficient, the sample size n upon which it is based must be known. Look up $\nu = n-2$ degrees of freedom and consult the column headed " r ". For example, for a sample size of 28, $\nu=26$, the critical values of r are found to be 0.374 at the 5% level and 0.478 at the 1% level. Thus, for an observed correlation coefficient $r=0.29$ in a sample of 28 paired observations, one would have to conclude that the correlation between the variables in question is not significantly different from zero. Negative correlations are considered as positive for the purposes of this test.

ν	P	r	ν	P	r	ν	P	r
1	.05	.997	16	.05	.468	35	.05	.325
	.01	1.000		.01	.590		.01	.418
2	.05	.950	17	.05	.456	40	.05	.304
	.01	.990		.01	.575		.01	.393
3	.05	.878	18	.05	.444	45	.05	.288
	.01	.959		.01	.561		.01	.372
4	.05	.811	19	.05	.433	50	.05	.273
	.01	.917		.01	.549		.01	.354
5	.05	.754	20	.05	.423	60	.05	.250
	.01	.874		.01	.537		.01	.325
6	.05	.707	21	.05	.413	70	.05	.232
	.01	.834		.01	.526		.01	.302
7	.05	.666	22	.05	.404	80	.05	.217
	.01	.798		.01	.515		.01	.283
8	.05	.632	23	.05	.396	90	.05	.205
	.01	.765		.01	.505		.01	.267
9	.05	.602	24	.05	.388	100	.05	.195
	.01	.735		.01	.496		.01	.254
10	.05	.576	25	.05	.381	125	.05	.174
	.01	.708		.01	.487		.01	.228
11	.05	.553	26	.05	.374	150	.05	.159
	.01	.684		.01	.478		.01	.208
12	.05	.532	27	.05	.367	200	.05	.138
	.01	.661		.01	.470		.01	.181
13	.05	.514	28	.05	.361	300	.05	.113
	.01	.641		.01	.463		.01	.148
14	.05	.497	29	.05	.355	500	.05	.088
	.01	.623		.01	.456		.01	.115
15	.05	.482	30	.05	.349	1000	.05	.062
	.01	.606		.01	.449		.01	.081

χ^2

Probability that χ^2 , with N Degrees of Freedom,
Will be Exceeded if Null Hypothesis is True

<u>Degrees of Freedom</u>	<u>Probability</u>	
	<u>0.05</u>	<u>0.01</u>
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
8	15.507	20.090
9	16.919	21.666
10	18.307	23.209
11	19.675	24.725
12	21.026	26.217
13	22.362	27.688
14	23.685	29.141
15	24.996	30.578
16	26.296	32.000
17	27.587	33.409
18	28.869	34.805
19	30.144	36.191
20	31.410	37.566
21	32.670	38.932
22	33.924	40.289
23	35.172	41.638
24	36.415	42.980
25	37.652	44.314
26	38.885	45.642
27	40.113	46.963
28	41.337	48.278
29	42.557	49.588
30	43.773	50.892

Student's t -distribution

Values exceeded in two-tailed test with probability P.

<u>Degrees of Freedom</u>	<u>Probability</u>	
	<u>0.05</u>	<u>0.01</u>
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845
21	2.080	2.831
22	2.074	2.819
23	2.069	2.807
24	2.064	2.797
25	2.060	2.787
26	2.056	2.779
27	2.052	2.771
28	2.048	2.763
29	2.045	2.756
30	2.042	2.750
40	2.021	2.704
60	2.000	2.660
120	1.980	2.617
	1.960	2.576