Statistics Primer

Contents

Stats in 30 minutes

This primer will briefly answer the following questions. You should read it in advance, bring questions about it to class, and consult it as you complete the statistical analysis worksheet.

A. What is a population, what is a sample, and why do we need statistics?
B. What is a *statistical* hypothesis?
C. What does a statistical test actually do?
D. How is a statistical test carried out?
E. Do these tests assume anything about my data?
F. What are the possible outcomes of the test?
G. How are statistical results reported?
H. Why does P give the probability of making an error, and why do we set $\alpha = 0.05$?

A. What is a population, what is a sample, and why do we need statistics?

In research, we want to say something definitive about a ***population***, but we have only a ***sample*** of the population. Statistics provide a quantitative way to express confidence about a conclusion when we have only limited information from a sample.

As a first step, we try to make sure by our sampling method that our sample is ***representative*** of the population as a whole. But what is "the population," and what kind of sample is representative of it? The answers depend on the question. As an example, we might ask "do biology majors score higher than chemistry majors on their first organic chemistry exam?" The "population" could include any and all biology or chemistry majors that are currently taking, have taken, or could take organic chemistry. We could restrict the population of interest in various ways, for example, to students in the USA, or students at CofC, or students in their sophomore year, or students taking the course this year. If we could measure every one of the students in that population, we could say definitively (without statistics) whether there is an average difference in scores between students in the two majors. It is more realistic, however, to measure representative samples. Then we can make a statement about the ***probability*** that the populations of the two types of majors differ based on (1) the sample means and (2) an estimate of confidence that our sample means represent the population means.

To choose a representative sample, we first try to avoid the potential for ***bias***. In most cases, sampling **at random** from *a population* helps to provide an unbiased sample *of that*

population. Students sampled at random from CofC, for example, would be an unbiased sample of students at CofC but could be a biased sample of students from the USA (if, for example, one of our programs were stronger than the other in an atypical way). Second, we try to avoid the potential for statistical ***noise*** (or "sampling error") by choosing a sample **as large as practical**, to reduce the possibility of getting a set of values in our sample that is atypical of the population. Avoiding ***bias*** and ***noise*** are two of the major issues in designing an experiment or survey.

Given a representative (unbiased and large) sample, we can then use ***inferential statistics*** to draw conclusions about the defined population. In other words, we can *generalize* the results of analyzing a sample to a larger population that the sample appropriately represents.

B. What is a *statistical* hypothesis?

Imagine a bar graph of average organic chemistry scores for biology and chemistry majors. The two bars will differ in height—it is extremely unlikely that the averages will be *exactly* the same. But is this difference in means between *samples* large enough to conclude that the *populations* really differ? To answer this question, we use sample data to determine how likely it is that the difference in means between samples could have been due to *chance* rather than to a *real difference between populations*.

For any statistical test we define two alternative *statistical* hypotheses:

- ***the null hypothesis* (H_0)**: the result expected if there were no relationship between variables
- ***the alternative hypothesis* (H_a)**: the result expected if there were a predicted relationship between variables (either a difference between groups or a correlation between variables)

Why do we bother setting up such formal alternatives? The answer has to do with how science works, by a process called *falsification*: we assume by default that there is **no relationship** (the null hypothesis) unless we have strong enough evidence to reject the null. This process reflects the *conservative* nature of science—we do not accept a new, alternative idea unless the evidence is highly convincing. In fact, a typical criterion for "rejecting the null hypothesis in favor of the alternative"* is that the relationship must be so convincingly strong that it would occur by chance (that is, because of a chance sampling error) no more than 5% of the time.

C. What does a statistical test actually do?

For any statistical test, we start with a simple assumption. We then evaluate whether there is strong enough evidence *to reject* this assumption.

Assumption: **The null hypothesis is true.**

If the null hypothesis **were true**, we would *expect* that a test statistic calculated from the data (a measure of the strength of the relationship between variables) to equal *zero*: a null hypothesis states, for example, that there is no correlation ($r = 0$) or no difference between two group means ($t = 0$). However, because we calculate a test statistic from just a limited (and therefore imperfect) sample of a population, it would not be surprising to get *small* differences from zero (positive or negative) in our test statistic ***just by chance*****,** even if the null hypothesis were true. The question is, how large can the test statistic get before we begin to suspect that it is *not* due to chance (and that the null hypothesis, in fact, is probably *not* true?)

Imagine repeating the same data collection 1000 times, *still assuming the null hypothesis is true*. Each time, you take a new sample from your population and calculate a new test statistic. If you compiled all your test statistics, they would form a normal distribution centered at zero (see right). Most values should be close to zero, and fewer would be extreme (large or small). That is, there is a high probability of getting a small value and a low probability of getting an extreme value, *just by chance*. [Q: Statisticians can generate this probability distribution just by knowing the sample size. *How would you expect the width of the curve to change depending on the size of the samples used to calculate the test statistic?*]

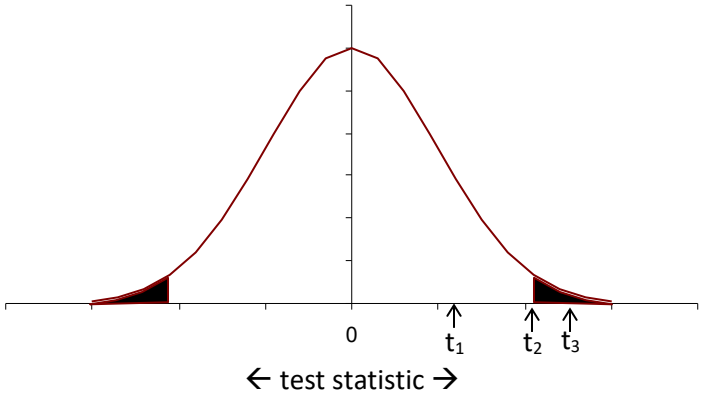


Fig. 1. Distribution of a test statistic given the assumption that the null hypothesis is true. Note that the probability specified by α is distributed equally between the two tails (for a two-tailed test)

The problem is, in research you often have the result of only <u>one</u> such experiment. So, what is the probability of getting the test statistic that *you* got (remember, assuming the null hypothesis is true) *just by chance*? That probability (called the **P-value**) can be found by seeing where your test statistic falls on this distribution. In Fig. 1, test statistic $\mathbf{t_2}$ falls a certain distance from 0, associated with a certain probability (**P)** of getting a value that extreme *just by chance* even if the null hypothesis were true. The value $\mathbf{t_1}$ is closer to 0, so has a higher probability, while $\mathbf{t_3}$ is further from 0, so has a lower probability of occurring by chance *assuming the null hypothesis were true*. Small test statistics have high probabilities of occurring by chance (and large P-values, see Section D), and large test statistics have low probabilities of occurring by chance (and low P-values), again *assuming that the null hypothesis is true*.

Because test statistics fall along a continuum, we need some way to say when our value is so extreme—that is, when it has such a small probability of having occurred by chance—that we question our assumption that the null hypothesis is true. That criterion is based on **α (alpha)**, a threshold probability value that we choose in advance. We use that threshold to judge when we have enough evidence to reject our assumption that the null hypothesis is true. When $\mathbf{P < \alpha}$, we conclude that the probability of our <u>large</u> test statistic occurring by chance is too <u>small</u> to stick with the null hypothesis, and instead we **reject the null hypothesis in favor of the alternative**. Conversely, when $\mathbf{P > \alpha}$, we **fail to reject the null hypothesis.**

By convention in biology, α is usually set at 0.05 (see section I). That is, we decide to reject the null hypothesis only when we expect a test statistic *as large as ours* no more than 5% of the time by chance. In Fig. 1, the shaded areas under the curve together account for 5% of the distribution of test statistics expected by chance (each tail has 2.5% of the probability). The

critical value is the test statistic associated with the probability $\alpha = 0.05$. In this example, the critical value for our test is at $\mathbf{t_2}$, so anything equal to or larger than $\mathbf{t_2}$ (or equal to or smaller than $-\mathbf{t_2}$) provides enough evidence to reject our initial assumption that the null hypothesis is true.

D. How is a statistical test carried out?

Scientists have access to a dizzying array of statistical tests. Fortunately, many simple analyses can be performed with knowledge of just three tests: the **correlation analysis**, the **t-test**, and the **chi-square analysis**. Which test to use depends on whether the variables are continuous or categorical. **See Table 1.** to determine when to use each of these tests.

Regardless of which test is used, the procedure is similar:
(1) calculate a **test statistic**,
(2) compare the test statistic to the threshold **critical value** (see Critical Value Tables),
(3) determine a ***P*-value** by comparing the test statistic to other values in the table, and
(4) reach a conclusion to reject the null hypothesis only if ***P*** is less than **alpha**.

Here are the details:

- What is a **test statistic**? A single value computed from your data. For the three tests you will use in this class, the test statistics are *r* (for correlation analysis), *t* (for a t-test), and χ^2 (for a "chi-squared" test). Excel will calculate them or help you to calculate them.
- What is a **critical value**? A threshold value that can be looked up in a table. If your test statistic exceeds the critical value, then the data from which you computed the test statistic are highly different from what you would expect if the null hypothesis were true, so any relationship you found between variables is unlikely to be due to chance. Critical values are tabulated based on the type of test, the **degrees of freedom**, and **alpha**.
- What are the **degrees of freedom**? A number based on the sample size of the data (**see Table 2.** for how to calculate *df* for each test). Because larger samples give greater power to detect a relationship between variables, sample size affects the critical value.
- What is **alpha**? A probability value chosen *before* the data are analyzed. Because the choice of alpha determines the critical value, it also acts as a threshold for decisions about the null hypothesis: when the probability of getting your results by chance is *less than* alpha (that is, your test statistic is *greater than* the critical value), the null hypothesis is rejected in favor of the alternative. Alpha is typically set at 0.05 (see section H).
- What is the ***P*-value**? A probability value calculated from your data. ***P*** is the probability that the relationship between variables measured *from your sample* is due to chance rather than to an actual relationship *in the population*. It is also, therefore, the probability that you are making an error by rejecting your null hypothesis (see Section H).

E. Do these tests assume anything about my data?

Yes, but many tests work even with small violations of these assumptions, so we will not worry here about testing them. The kinds of statistical tests you will use make just a few basic assumptions that are worth knowing about:

- Data points are assumed to be independent of one another. For example, when measuring scores on an organic chemistry exam, we assume that each student's score is independent of the scores of other students (not always the case!).

- Any statistical test fits the data to some kind of model, which is an ideal representation of how the data are patterned. Real data points always deviate from the ideal model. The size of those deviations (known as residuals) are assumed to have a normal (bell-shaped) distribution, with many more measurements close to the average and progressively fewer toward the tails. This kind of distribution will be true for many types of data.
- The **t-test** assumes that measurements for the two groups you are comparing have equal standard deviations. If the standard deviations you calculated are not terribly different, you probably meet this assumption. If they are terribly different, a version of the t-test is available that can account for this difference in standard deviations.

F. What are the possible outcomes of the test?

- Conventionally there are two possible outcomes: (a) *failure to reject* the null hypothesis, or (b) *rejection* of the null hypothesis *in favor of* the alternative hypothesis. It is incorrect to state that the test leads acceptance of the null hypothesis or proof of either hypothesis.
- Rejection of the null hypothesis does not necessarily imply a mechanism for the relationship. The effect could be due to some other mechanism you didn't propose.
- Rejection of the null hypothesis—a statistical outcome—does not necessarily mean that the effect has great biological significance. As a biologist, it is still necessary to consider the magnitude of an effect when judging its biological importance.

G. How are statistical results reported?

To report the outcome of a statistical test, one states a conclusion along with the test statistic, degrees of freedom, and P-value (the latter three often in parentheses). For example:

> "There was no significant difference between the means of the two groups ($t = 0.45$, $df = 134$, $P > 0.05$)".

H. Why does *P* give the probability of making an error, and why do we set $\alpha = 0.05$?

We have established that the ***P*-value** is the probability of getting a test statistic as extreme as ours by chance if the null hypothesis were true. If ***P*** is small enough (less than alpha), we decide to reject the null hypothesis in favor of the alternative. But of course there is still some probability (given by ***P***) that the null hypothesis *is* true and we just happened to get one of those extreme sampling errors. The P-value is therefore a statement of *confidence* about a decision to reject the null. For example, if the test concludes that $P < 0.02$, we have strong confidence that *less than* 2% of the time, with a test statistic as large as the one we calculated, we will be making an error by rejecting the null hypothesis. This type of error—rejecting the null hypothesis when it is in fact true, known as a "false positive"—is called a **Type-1 error**. Alpha is therefore the upper limit on the type-1 error we are willing to tolerate when performing the test.

A second type of error, called a **Type-2 error** (or a "false negative"), involves failing to reject the null hypothesis when it is in fact false (in the whole population). In science, we generally guard against Type-1 errors more than against Type-2 errors. The reason is that science is conservative—it does not accept new ideas (alternative hypotheses) until there is strong evidence. Setting α higher (e.g., 0.1) would lead to rejecting the null hypothesis more often, but with a higher number of false positives. Setting α lower (e.g. 0.01) would reduce the false positives, but could make it unreasonably hard to reject the null hypothesis, and create more false negatives. The value $\alpha = 0.05$ is a compromise. In some fields, however, a lower value for α is chosen because there are especially high costs of a Type-1 error. For example, a

pharmaceutical company might want especially strong evidence that a new drug works significantly better than the current drug before it invests millions of dollars in R&D. It would set a low $\alpha = 0.01$ in order to conservatively guard against concluding there is a real difference between drugs in case there isn't.

Note: information in section I is optional for this course, but will be interesting and useful for those who want to build their understanding of statistical testing.

.

TABLE 1. Which statistical test should I use?

If you are testing the relationship between…	use this test…	to answer this question…	involving these statistical hypotheses…	to reach this kind of conclusion…
2 continuous variables	Correlation analysis	Is there a statistical tendency for high measures of one variable to be associated with high (or low) measures of another variable?	H_o : there is no association between variables H_a: there is an association (positive or negative) between variables	If the association is stronger than is likely by chance, the variables are said to be significantly positively (or negatively) correlated.
1 categorical variable & 1 continuous variable	t-test	Is there statistical evidence that the mean of one group is significantly different from the mean of a second group?	H_o : there is no difference in the average between groups H_a: there is a difference (positive or negative) in the average between groups	If the difference between means (relative to the standard error) is more extreme than expected by chance, then the difference is said to be statistically significant.
2 categorical variables	Chi-square test	Is there a statistical tendency to belong to a particular category in one variable if a subject belongs to a particular category in the other variable?	H_o : there is no association between two categorical variables H_a: there is an association (positive or negative) between the two categorical variables	If the association is stronger than is likely by chance, the variables are said to be significantly associated with one another.

TABLE 2. Test statistic, calculation of sample size and degrees of freedom for different tests

Statistical test	Test statistic	Sample size	Degrees of freedom
Correlation Analysis	r	N = number of subjects with paired measurements of the two variables	N–2
t-test	t	N = total number of measurements taken in both groups **unless it is a paired t-test in which case it is simply the number of samples in one group**.	N–2
Chi-square test	χ^2	N = total number of subjects measured C_1 & C_2 = number of categories in variables 1 & 2	C_1-1 x C_2-1

Using the t-Test to Test Differences between Means

One very common use of statistics is to figure out whether two things are different. For example, you may wish to know whether one brand of fertilizer causes corn to grow any better than another brand of fertilizer, or whether male frogs weigh more than female frogs, or whether a toxic compound causes mice to grow more slowly (relative to a control). This boils down to a test of differences between two means.

More formally, we wish to test the null hypothesis:

$$H_0 : \mu_1 = \mu_2$$

μ (the Greek letter *mu*) is an abbreviation for the population mean. This refers to the average you would theoretically get if you sampled an infinite population.

The subscripts 1 and 2 refer to population 1 and population 2. In most cases, one of the populations will be the treatment and the other will be the control, but this will not always be the case. For example, 1 could refer to males and 2 to females.

Also "population" as used here is a statistical term, not a biological one. "Population" could mean the set of all test tubes or Petri plates, or it could mean a real population such as all male frogs.

The most widespread method to test for differences in means is the Student's *t*-test. The test statistic for the *t*-test is, not surprisingly, *t*. You get *t* from the following equation:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{s_1^2 / N_1 + s_2^2 / N_2}}$$

where

$\bar{x}_1$ is the mean value for population 1

$\bar{x}_2$ is the mean value for population 2

s^2_1 is the variance for population 1

s^2_2 is the variance for population 2

N_1 is the number of observations for population 1

N_2 is the number of observations for population 2

If the means are similar, then *t* will be close to zero, but if the means are different, then *t* will be very large. To determine whether the differences are significant, you need to calculate the degrees of freedom (*d.f.*):

$$d.f. = N_1 + N_2 - 2.$$

Locate the critical value of *t* in a *t*-table, using the $p = 0.05$ (recall that this is the conventional number) for the column and the degrees of freedom you just calculated as the row. If the value of *t* you calculated is greater than this critical value, then you can reject the null hypothesis of equal means.

The Student's *t*-test as described here is only one kind of *t*-test. Thus you need to be careful that you use the equation here only for testing differences between means.

EXAMPLE

It has been speculated that plants increase the amount of defense compounds in response to herbivory. To test this, an ecologist excluded caterpillars from five randomly selected milkweed plants, and allowed caterpillars to feed on five others. After one day, the caterpillars were removed and alkaloids (a kind of defense compound) were extracted from each plant and measured. The null hypothesis is

H_0: $\mu_c = \mu_{nc}$

where μ is the mean alkaloid content, and c and nc refer to caterpillars and no caterpillars, respectively. More informally (but perfectly acceptable):

H_0: Milkweed plants that have been grazed by caterpillars have the same alkaloid content as ungrazed milkweed plants.

The alternative hypothesis is:

H_A: $\mu_c \neq \mu_{nc}$

or H_A: Milkweed plants that have been grazed by caterpillars do not have the same alkaloid content as ungrazed milkweed plants.

The data obtained from the grazed plants are, in micrograms of alkaloid per gram of leaf tissue:

3.4, 4.5, 3.6, 3.7, 3.9

The mean and standard deviation for this treatment are 3.82 and 0.421, respectively.

The data obtained from the ungrazed plants are:

2.8, 3.4, 2.9, 3.1, 2.8

The mean and standard deviation for this treatment are 3.00 and 0.255, respectively.

To calculate *t*:

$$t = \frac{|\bar{x}_c - \bar{x}_{nc}|}{\sqrt{s_c^2 / N_c + s_{nc}^2 / N_{nc}}} = \frac{|3.82 - 3.00|}{\sqrt{0.421^2 / 5 + 0.255^2 / 5}} = \frac{0.82}{\sqrt{0.0354 + 0.0130}} = 3.726$$

To calculate degrees of freedom: $d.f. = N_c + N_{nc} - 2 = 5 + 5 - 2 = 8$

The critical *t* value (t_{crit}) for 8 degrees of freedom and p=0.05 is 2.306. Since our calculated *t* (3.726) is greater than 2.306, we can conclude:

> "We reject H_0 at p=0.05" or "We reject the null hypothesis that grazed and ungrazed milkweed plants have the same alkaloid level".

In other words, there is less than a 5% chance (or 1 in 20 chance) that we would have found a *t* value as large as 3.726 just due to chance. To compute an exact p-value you must use software like the R programming language.

Testing for Associations between Categorical Variables

Frequently we do not have fine grained information on our independent and dependent variables and instead we have qualitative categorical groupings of variables such as high and low resources or fast and slow speed. If we wish to test for an association between samples from a "high" group to also be "fast" then we can use a Chi-square (χ^2) test.

We arrange our two groups into what is called a ***contingency table.*** Here I've just come up with some pretend counts of samples that were grouped into both groups simultaneously.

	Fast	Slow
High	20	10
Low	8	20

In this pretend example there were 20 samples that were both in the "high" group and that were classified as "fast". The table at least suggests that there may be association between the two groups because most samples in the high group are fast and most in the low group are slow.

We can use a $\chi 2$ test to formally test this hypothesis.

> H_0 = no association between the two sets of groups
>
> H_1 = there is an association between the sets of groups.

Test statistic:

$$X^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

c = number of columns

r = number of rows

$O_{i,j}$ = observed count in the cell of the ith row and the jth column

$E_{i,j}$= expected count in the cell of the ith row and the jth column

Degrees of freedom for the test are

d.f. = (r – 1) * (c – 1)

The expected counts are generated by assuming that counts are arranged randomly in the table while using the observed constrains on the total number of samples that were found in each row and column.

Here is an example in which we have recorded the swim speed of fish (fast or slow) that were provided a diet either High or Low in calories. We can use a chi-square test to test if there is an association between diet and swim speed. Specifically, in this context the null hypothesis is:

H_0: there is no association between diet and swim speed

H_A: there is an association between diet and swim speed

Here is our results including the row and column totals for the ith row (Ri) and the jth column (Cj)

	Fast	Slow	Ri
High	20	10	30
Low	8	20	28
Cj	28	30	58

Under the assumption of no association between the rows and columns we can generate the expected values Ei,j using the product of the row and column totals divided by the grand total of the table

$$E_{i,j} = \frac{\sum_{j=1}^{c} O_{i,j} * \sum_{i=1}^{c} O_{i,j}}{\sum_i \sum_j O_{i,j}} = \frac{R_i * C_j}{T}$$

R_i = total of row i

C_j = total of row j

T = grand total of all observed values

For our example the table of expected value is:

	Fast	Slow
High	14.48	15.52
Low	13.51	14.48

Once the expected values are calculated the $\chi 2$ value can be computed by comparing to the observed values.

For our example the table of chi-squared values is

	Fast	Slow
High	2.10	1.96
Low	2.25	2.10

The sum of all of those value is the chi-squared value, $\chi 2 = 8.41$, the degrees of freedom are 1 (= (# rows – 1) * (# columns - 1)).

For a 2x2 table the critical value is 3.84 and because our $\chi 2$ is so much larger than this we can conclude that statistically we can reject the null hypothesis. Biologically our conclusion is that there is an association between diet and swim speed. Specifically, our example result shows that high calories result in faster swim speeds.

Important Note

Above I provided many different tables to show how the calculation is broken down into steps. However, for the presentation of a chi-square test only the contingency table is needed. Here is how these results would be reported in the Results section of a scientific paper, poster, or talk:

Table 1. The swim speeds (fast or slow) of fish provided a diet that was either high or low in calories ($\chi 2 = 8.41$, $d.f = 1$, $p < 0.05$).

	Fast	Slow
High	20	10
Low	8	20

If you wish to compute the p-value directly rather than comparing your $\chi 2$ value to a critical value then you will need to use statistical software to carry out the calculation like the R programming language.

resiTesting for Correlations between Continuous Variables

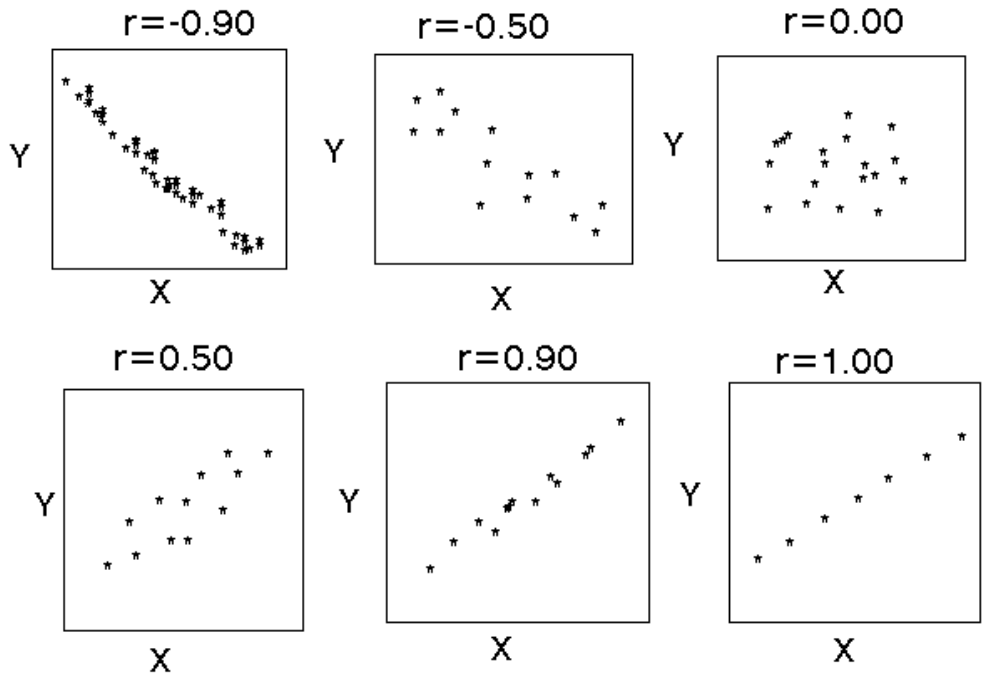
Correlation tests relationships between two variables. You need to identify an independent variable (*x* variable) that is often considered the causative agent, and a dependent variable (*y* variable) that is thought to be affected by *x*. (Do note, however, that correlation does not prove causation. This will be discussed shortly).

For example, *x* might equal the number of leaves on a dandelion plant produced in the fall, and *y* might equal the number of flower heads produced in the spring. We might expect more flowers to be produced if the plant had more leaves during the past growing season - this might be because plants with many leaves would have more carbohydrates stored up for reproduction.

Our null hypothesis is that the number of leaves is unrelated to the number of flowerheads. In other words,

H_0: *x* is uncorrelated with *y*
H_1: *x* is correlated with *y*

If a graph is made, the x variable is plotted on the x (horizontal) axis and the y variable is plotted on the y (vertical) axis. A simple equation (given later) results in a value known as *r* or the correlation coefficient. *r* takes values from -1 to +1.



Note that values close to -1 and +1 indicate very strong relationships, those close to zero indicate very weak relationships. Negative *r*'s indicate negative relationships; positive *r*'s indicate

positive relationships. A value of zero indicates no relationship between the variables. Another way of restating our null hypothesis is that the correlation coefficient will equal zero, and our alternative hypothesis is that the coefficient will differ significantly from zero.

The following is the formula for *r*:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right)\left(\sum (y_i - \bar{y})^2\right)}}$$

$\bar{x}$ (pronounced "x-bar") is the mean (average) of all the *x*'s, and

$\bar{y}$ ("y-bar") is the mean of all the *y*'s. Therefore:

$\bar{x} = \Sigma\ x_i\ /\ N$ and $\bar{y} = \Sigma\ y_i\ /\ N$

Where *N* is the total sample size (e.g. the number of dandelion plants).

One important thing to remember about correlation is that it does not necessarily imply direct causation. For example, suppose that there was an extremely significant positive correlation between the number of dandelion leaves and flower heads. This could be because plants that had many leaves were able to produce enough carbohydrates to produce many flowers. However, it is also possible that some plants were in very fertile areas, and had enough nutrients to produce many leaves as well as many flowers. That is, the number of leaves did not cause any particular number of flowers, but both leaves and flowers were caused by a third factor. As usual in science, results must be interpreted with caution.

As long as some variation exists in *x* and in *y*, there will be a correlation between the two variables. The correlation may be positive, negative, or close to zero. It is extremely unlikely that a calculated value of *r* will equal exactly zero. If it does, then (most likely) the calculations were done incorrectly. Correlation is *not* the same thing as association, a topic that is dealt with elsewhere.

The value r^2 (or R^2) is special: it is known as *the coefficient of determination*. It can vary between zero and one, and can be interpreted as "the proportion of variation in the dependent variable that can be explained by the independent variable". Thus, if you study the relationship between body mass of ladybugs and the number of eggs laid, an r^2 of 0.72 means that 72% of the variation in the number of eggs laid can be explained by body mass. Conversely, 28% of the variation is "unexplained", and one would have to look for other factors such as genetics, habitat, diet, or even chance.

The p-value for the correlation analysis can be determined either by comparing the computed *r* value to a critical *r* value table or by computing the p-value directly in R.