

Introduction to Statistics Notes (Source: <http://www.highlandsr.spps.org/>)

Why Statistics? Biology uses mathematics as a tool to examine the natural world.

Many phenomena in the natural world can be measured or counted. Indeed, science is often best at explaining things that can be measured or counted.

The results of many investigations in biology are in the form of numbers. These numbers can often be better understood using mathematics.

Example: In an investigation of the heights of the blades of grass in a field, you measure blades of grass with a ruler.

The result is a **set of values**.

These values must be handled appropriately to give us useful information.

The values have **units** (e.g. the height of a particular blade of grass is 25 mm).

The values have **precision**. (E.g. measuring with a ruler might only be precise to ± 1 mm, so the height of a particular blade might be 25 ± 1 mm. In this case, it could have been 24 mm, or 26 mm, or any value in between these, but not more or less than these).

We also express the precision of our values in the number of decimal places that we choose. For example, stating that a blade of grass is 25 mm long is not the same as stating that it is 25.0 mm long. In the first case, it is implicit that the blade could be between 24.5 and 25.5 mm long. In the second case, the blade could be between 24.95 and 25.05 mm long.

In collecting data, we should:

- include the units
- have an appropriate number of decimal places
- have a consistent number of decimal places
- indicate the precision.

When we process data, we should also consider the number of significant figures that is appropriate for our data. Most often, we use three significant figures.

Biologists need statistics

In many investigations of living things, very many numbers are gathered, or could be gathered.

There might be too many numbers for us to easily make sense of the data. We could say that we have a lot of data, but that we do not yet have meaningful information.

Statistics is a branch of mathematics that helps us to handle the large amounts of data that we often obtain in investigations. It helps us to obtain useful information from it and to draw conclusions.

Example. You are comparing the heights of the blades of grass in two fields. One of these fields has a high concentration of potassium ions in the soil and the other has a low concentration of potassium ions in the soil. The hypothesis is that the grass is taller in the field with the high soil potassium level than it is in the field with the low soil potassium level.

We would use statistics in planning the investigation, and in helping us to decide whether the results support the hypothesis, or not.

Sampling

Sometimes, it is possible to measure all of the things that are being considered; for example, the heights of all the oak trees in a very small forest.

In statistics, all the values that could be considered are known as the **population**. (Not to be confused with the ecological use of this term). So the heights of all the oak trees in the forest would be the population.

In most investigations, it is not possible, not practical, or not advisable to measure all the values in a population. In these situations, we measure just some of the values. We call such a group of values a **sample**.

We hope that the values in the sample are **representative** of the population (that is, that they give an accurate picture of the true population).

Example. It would not be practical to measure the heights of all the blades of grass in our two fields (the populations of the two fields).

It would take an enormous amount of time:

- we may not have this much time
- our time could be better spent on other tasks.
- the heights of the blades might actually change during the time

We might also damage each blade as we measure it:

- this act of measuring changes the values that we are examining
- we may be causing environmental damage in a fragile ecosystem

In such a situation, it is better to take a sample of measurements from each of the two fields.

Even in laboratory experiments, we deliberately limit sample sizes. For example, if you are examining the effect of light intensity on the rate of photosynthesis in young bean plants, you could potentially test millions of different bean plants under a range of light intensities, and then repeat the experiment thousands of times. However, in practice you might choose to examine just 50 plants under each different light intensity, and to repeat the experiment three times.

Populations and samples show variation

In a population, we usually find that not all the values are identical. Instead, there are differences between the values even inside a population. We call this **variation**.

The data we obtain from a study has **variability**.

Example. The heights of each of the blades of grass in the two fields differ between each other, even inside the same field.

We could measure the height of one blade of grass from each of the two fields (a very small sample) and find that the blade of grass from the field with high potassium is longer than the blade of grass from the field with low potassium. However, we would still be unsure of whether the difference between the heights of these two blades is due to the field it came from, or was just a difference that occurs anyway within each field.

We could measure the heights of 500 blades of grass in each of the two fields (a larger sample). We would obtain a set of 500 values for each field. It is difficult for the human mind to obtain useful information from such a large amount of unprocessed data. However, by using statistics, we can describe the values in various ways that make the information more meaningful for us. Most often, we process the data to estimate an average and to describe the variation in some way.

The mean - This is a measure of average. For most studies, this is the most important item of processed data. We estimate the mean as follows:

$$\text{Mean} = \frac{\text{Sum of values}}{\text{Number of items}}$$

Example. We have measured the heights of 500 blades of grass from each field.

We find that the mean height of the grass in the sample from the field with high potassium is 56.2 mm and the mean height of the grass in the sample from the field with low potassium is 48.5 mm.

The mean height is greater for grass in the sample from the field with high potassium than it is in the sample from the field with low potassium.

This is much clearer than looking at two lists of 500 values.

However, it is still unclear whether the difference between the means of our samples really represents a difference between the two populations.

Measuring variation

In many studies, the variation within the population is itself a very interesting phenomenon and it would be useful to be able to describe it.

We also often need to describe the variation within a population, to help us decide whether a difference between sample means truly represents a difference between population means.

Example. Maybe the heights of the blades of grass are not really different between the two fields. Perhaps it just happens that we measured blades that were longer in the one field than the other. If there are very large differences between all the blades of grass inside each field, this might well be the case.

To help us decide this, we need to describe the variation between the blades of grass in each field.

The range- This is a simple measure of variation.

The range is the difference between the largest and the smallest values.

$$\text{Range} = \text{Largest value} - \text{smallest value}$$

Knowing the range is very useful for some purposes. It gives a simple measure of spread and an idea of the extremes that can exist. However, there may be just a few extreme values (so-called **outliers**) which are very different from all the other values. To more fully describe variation, other measures are needed.

The standard deviation - The standard deviation is a more complete measure of variation. It considers every value in the set.

The standard deviation of a sample is called s and the standard deviation of the population is called σ .

The standard deviation is a number which expresses the difference from the mean of every value in the set.

We can calculate the standard deviation by using a formula, a calculator, or a spreadsheet programme.

A large value for standard deviation indicates that there is a large spread of values. Many values are far from the mean.

A small value for standard deviation indicates that there is a small spread of values. Most values are close to the mean.

Sample size

In designing an investigation, it is important to decide the size of the sample to take and how the sample is to be selected.

In deciding sample size, there is a trade off.

- The larger the sample, the more likely it is to be representative of the population
- The smaller the sample, the more quickly, cheaply and conveniently it can be done, and the less environmental disturbance is caused in the process.

By and large:

- The larger the variation between individual values, the larger the sample that is needed.
- The smaller the difference in the means between different conditions, the larger the sample that is needed.

Sample selection

There is a great risk that if we choose the individuals to measure, we might not be choosing truly representative values.

There are several problems with purposely choosing the individuals to measure:

- Knowing the hypothesis, we might (consciously or unconsciously) choose only individuals to measure that fit the hypothesis.
- We might (consciously or unconsciously) overcompensate in an effort to be fair and pick individuals that do not.
- Outliers might also attract too much attention and we might either over or under represent them.
- We might want to avoid going to inconvenient locations, which should be included.

Choosing evenly spaced samples could also give us an unrepresentative sample, if there is a regular pattern variation in the underlying population.

To overcome these problems, we take **random samples**. In a random sample, every individual has an equal chance of being selected.

There are several techniques for creating random samples, including using random number tables.

Many statistical tests assume that the data has been gathered randomly.

Error bars

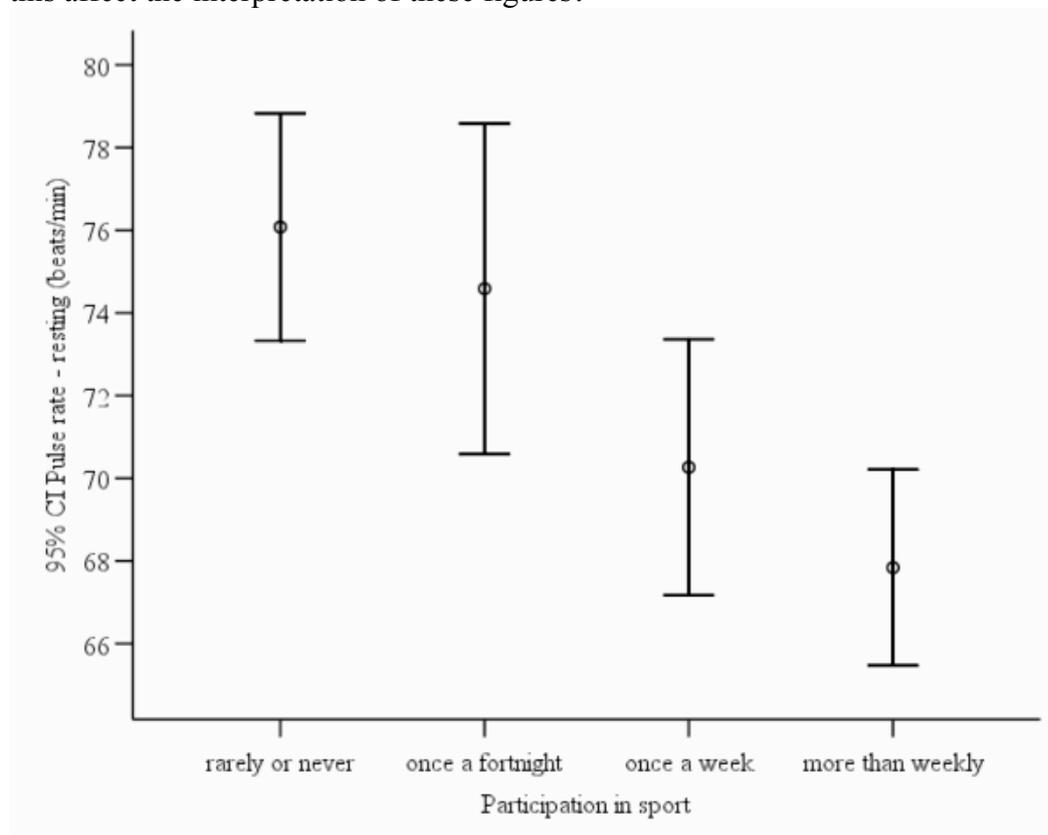
In many charts and graphs, we show the mean values of our samples. Such a chart or graph can clearly show the differences between our conditions and trends may become apparent.

It is useful to also be able to show a measure of the variation inside each of these samples. We do this by adding error bars to the chart or graph.

An error bar is a line that extends above and below a bar in a chart, or a data point in a graph. It could represent the range for that sample, or the standard deviation.

The length of the line represents the size of the range, or the size of the standard deviation. It extends an equal distance above and below the value of the mean. It can be stated that *error bars are a graphical representation of the variability of data.*

Look at the example below. What additional information do the error bars give? How does this affect the interpretation of these figures?



The Normal Distribution

Very often in biology, the variation that are found in samples in biology follow a so-called normal distribution (sometimes also called a bell-curve).

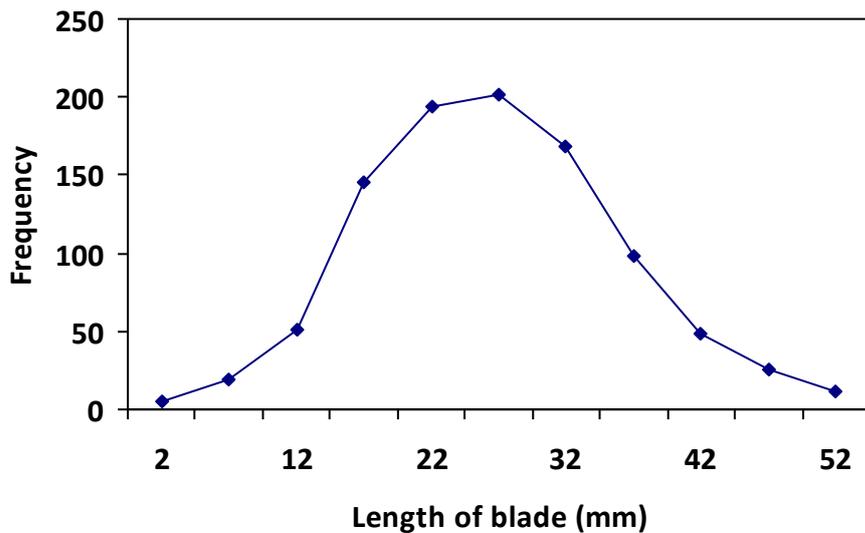
Put simply, most of the values are quite close to the mean, rather fewer are somewhat greater or somewhat less than the mean, and just a few are much greater or much less than the mean.

Example. We have measured the lengths of 969 blades of grass in a field with a medium level of soil phosphorus.

We have then grouped the data into a table, which shows how many blades of grass had a particular length. We call this kind of table a frequency distribution table.

Length of blade of grass (mm)	Number of blades of grass
0 – 4	5
5 – 9	19
10 – 14	51
15 – 19	145
20 – 24	194
25 – 29	202
30 – 34	169
35 – 39	98
40 – 44	49
45 – 49	26
50 – 54	11

We can then plot this result in a diagram. We call this kind of diagram a frequency distribution diagram.



Note that the shape of the curve resembles a bell (hence a bell-shaped curve).

This distribution approximates to a normal distribution.

The variation we find in the world around us very often approximates to a normal distribution.

Some special characteristics of the normal distribution

There are a number of statements that we can make about a true normal distribution. These statements are correct for all data that follows a true normal distribution.

- The graph is always bell-shaped and symmetrical.
- The mean value is the middle value (the median), with as many values above it as there are below it.
- The mean value is also the most frequent value (the mode).
- About 68 % of the values lie within ± 1 standard deviation of the mean.
- About 95 % of the values lie within ± 2 standard deviations of the mean

Provided that the data is normally distributed, this is correct regardless of how large or small the standard deviation is.

A frequency distribution diagram for normally distributed data with a large standard deviation will appear rather flattened.

A frequency distribution diagram for normally distributed data with a small standard deviation will appear rather pointed.

Nevertheless, we can draw the same conclusions.

Many of the tests used in statistics assume that the data follows a normal distribution.

Using the standard deviation in biology

There are several motives for calculating the standard deviation when investigating living things.

- The value provides a description of the variation which considers every data item. For many phenomena, this variation is of interest.
- Large differences between in the sizes of the standard deviations between samples that are being compared can indicate that control variables are not constant. This can be an indication that there is a problem with the validity of the investigation.
- The standard deviation can be used as a support in hypothesis testing.

Statistical significance

In statistics, the word **significance** has a special and very exact meaning, which is different from its meaning in everyday English.

In biology, we often want to compare two or more sets of conditions and to determine if there is a true difference between these conditions in the phenomenon that we are examining.

Example. If we are comparing the heights of blades of grass in a field with low potassium and in a field with high potassium, we will want to determine if there truly is a difference in the mean heights of the blades of grass in these fields.

By a true difference, we are interested in whether there is a difference in the population. That is, if we consider the mean of every single blade of grass in the one field and the mean of every single blade of grass in the other field, are these means the same or different?

If there is a difference between these populations of values, we can say that the difference is significant. If there is no difference between these populations of values, we say that the difference is not significant.

It is usually not practical to measure every single possible value in a population. Instead, we take a sample from each condition being considered, and then compare the means of these samples.

Usually, we do find a difference in the means between different samples. However, this presents us with a problem of interpretation:

- Do the means of the samples differ because the means of the populations from which they were collected differ? That is, do the differences between the sample means represent a difference in the means of the underlying populations?

Or

- Do the means of the samples differ only because of differences that exist within each field? That is, do the differences between the sample means result only from variations within the populations, with there being no difference in the means of the underlying populations?

We are asking:

- Are the means of the samples significantly different?

Or

- Are the means of the samples not significantly different?

A difference between sample means is significant if it represents a true difference in the means of the underlying populations.

Example. In an investigation of the two fields, we found that the mean height of the blades of grass in the sample from the field with high potassium was 56.2 mm and the mean height of the blades of grass from the field with the low potassium was 48.5 mm.

We have found thus a difference between the means of our two samples. However, we need to determine if this difference is significant. That is, does it indicate a true difference in the means of the populations of the two fields.

Determining statistical significance

There are many techniques for determining statistical significance. This is often called testing for significant difference. Much of the branch of statistics is concerned with significance testing.

Statistical significance and hypothesis testing

In investigations in biology, we are usually testing a hypothesis. Statistical significance is our main tool in deciding whether the data supports the hypothesis.

Example: It is our hypothesis that the grass is taller in the field with the high soil potassium level than it is in the field with the low soil potassium level.

In our investigation, we found that the mean blade height for the sample from the field with the higher level of soil potassium is greater than it is for the sample from the field with the lower level of soil potassium.

We carry out a statistical test to determine whether the difference between these sample means is significant.

If we find that the difference between these means is significant, we can say that the data supports the acceptance of the hypothesis.

If we find that the difference between these means is not significant, we must say that the data does not support the acceptance of the hypothesis.

Using the standard deviation to indicate possible significance

A large difference between the means of samples, and small standard deviations for these samples, indicates that it is likely that the difference between the means is statistically significant.

A small difference between the means of samples, and large standard deviations for these samples, indicates that it is likely that the difference between these means is not statistically significant.

Confidence levels

It is seldom possible to say with absolute certainty that the difference between sample means is significant with complete certainty (100 % confidence).

Instead, we determine if the difference between the sample means is probably significant.

Most often in biology, we decide that we want to be 95 % confident that the difference between the samples is significant.

This means that there is only a 5 % chance that the samples could be as different as they are because of chance, and not because of a real difference between the populations.

We could also say that we are confident that the probability (p) that chance alone produced the difference between our sample means is 5 % ($p = 0.05$).

We determine whether the samples are significantly different at the 95 % confidence level.

The t-test

The t-test determines whether the difference observed between the means of two samples is significantly, at a chosen confidence level.

The test assumes that the data is normally distributed (remember that there are special rules that describe the variation in the data). The sample size must be at least 10.

The test works by considering the following:

- The size of the difference between the means of the samples.
- The number of items in each sample.
- The amount of variation between the individual values inside of each sample (the standard deviation).

Briefly, the test works as follows:

- A value of t is calculated from the data.
- The value of t is found that would be needed to indicate that the observed difference between the means is significance at a chosen confidence level.
- The value of t that was calculated from the data is compared with the value that would be needed to indicate that the observed difference between the means is significant.
- If the calculated value for t is larger than the required value for t , the difference between the means is significant at this confidence level.
- If the calculated value for t is smaller than the required value for t , the difference between the means is not significant at this confidence level.

Using the t-test

Calculating the value for t from the data

This may be done in several ways, such as:

- From a formula
- Using a scientific calculator
- Using a spreadsheet

In an exam, you may be given a value for t that has been calculated in one of these ways.

Finding the value for t that would be needed to indicate significance

This is found from a table.

Find the appropriate column

The columns represent different confidence levels.

$p = 0.05$, which is the 95 % confidence level, is the most common choice.

Find the correct row

Each row is represents a so-called degrees of freedom.

$$\text{Degrees of freedom} = (\text{size of sample 1} + \text{size of sample 2}) - 2$$

The value for t that would be needed to indicate significance is in the intersection between this row and this column.

Example of using the t-test

A study has been conducted comparing the level of a particular plant secondary product (a scent) produced in the flowers of a species of plant on two successive years, the first during a summer with little rainfall and the second during a summer with heavy rainfall. The hypothesis is that the level of this scent is different between these two summers.

We determined the level of the scent in each of 14 flowers on each occasion. The mean levels of the scent were 23 mg per gram plant dry weight for year one and 32 mg per gram plant dry weight for year 2.

We want to know if this difference is significant.

The value of t was calculated using a calculator. It was found to be 3.43.

The 95 % confidence level is chosen ($p = 0.05$).

The degrees of freedom were calculated.

$$\begin{aligned}\text{Degrees of freedom} &= (\text{size of sample 1} + \text{size of sample 2}) - 2 \\ &= (14 + 14) - 2 \\ &= 26\end{aligned}$$

Using the table of t values, the value of t is found that corresponds to $p = 0.05$ and 26 degrees of freedom. This is a t value of 2.06.

The calculated value for t is compared with the value from the table.

3.43 is larger than 2.06.

Thus, the difference between the sample means is significant at the 95 % confidence level.

The results thus support the hypothesis that the level of the scent is different between the two successive years.

Task

Use the t-test to test for statistical significance in the exercise.

Correlation

Correlation is a measure of the association between two factors.

If an increase in one factor is associated with an increase in another factor, we say that there is a **positive correlation**.

Example. The number of hours a week that a group of students spent training was compared with the fastest speed that they could run. A trend was found. With increasing time spent training, the faster the student could run.

If an increase in one factor is associated with a decrease in another factor, we say that there is a **negative correlation**.

Example. The number of cigarettes that a group of students smoked each week was compared with the speed that they could run. A trend was found. With increasing number of cigarettes smoked each week, the slower the student could run.

If an increase in a factor is not associated with a consistent change in another factor, we say that there is **no correlation** between them.

Example. The speed at which a group of students could solve mathematical tasks was compared with the speed that they could run. No trend was found. With increasing speed in solving mathematical tasks, no consistent change was found in the speed at which the students could run.

The **strength of the association** between two factors can be measured.

An association in which all the values closely follow the trend is described as being a strong correlation.

An association in which there is much variation, with many values being far from the trend, is described as being a weak correlation.

A value can be given to the strength of the correlation, r .

$r = +1$ a complete positive correlation

$r = 0$ no correlation
 $r = -1$ a complete negative correlation

Correlation and causation

Often, a correlation exists between two factors, because one of the factors is causing a change in the other factor.

For *example*:

- Training develops muscles and improves the cardio vascular and respiratory systems, and thereby causes better athletic performance.
- Smoking damages the cardio-vascular and respiratory systems, and so causes poorer athletic performance.

Many important relationships have been uncovered by studies of the correlation between different factors. This is particularly important in studies involving human health.

Example. Much of the early evidence that smoking may be harmful to human health was uncovered by finding correlations between amount of smoking and the frequency of different diseases.

However, **correlation does not necessarily indicate causation.**

The change in one factor is not necessarily causing the change in the other factor.

It may be a coincidence that the two factors appear associated in a particular situation, or there may be another underlying factor that is causing the change.

Example. An early study found a negative correlation between women taking hormonal replacement therapy incidence of heart disease. This implied that hormonal replacement therapy caused improved heart health. However, it was later observed that the women taking hormonal replacement therapy tended to be from higher socio-economic groups than women who did not. It was this underlying factor which was probably giving the better heart health.

(http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation)