Vocational Aqualabs

TRAINING

# Vocational Aqualabs – Vocational Generic Skills For Researchers

## Experimental Design
## Unit 2a

Trevor Telfer/James Bron
Institute of Aquaculture
University of Stirling

# Statistical analysis and its preliminary considerations

Though this course is about Experimental Design it is important that some understanding of the basic essentials of statistical techniques is also given. Why? Well, an understanding of statistics allows one to critically assess the work of others according to a universally approved framework. Statistical techniques provide criteria for detecting differences or relationships between samples or variables and allow the presentation of experimental or observational results to others. No doubt many of you will have seen a number of quotes surrounding statistics. Some of the popular ones are:

"*An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts - for support rather than for illumination*." Andrew Lang.

"*There are three kinds of lies: Lies, Damn Lies, and Statistics*." A saying attributed to Mark Twain.

It is vital though that we are able to analyse our data properly and be able to comment on our hypotheses. It is unfortunate, however, that the use of statistical tests are either abused or totally neglected. The two quotes below highlight this concern:

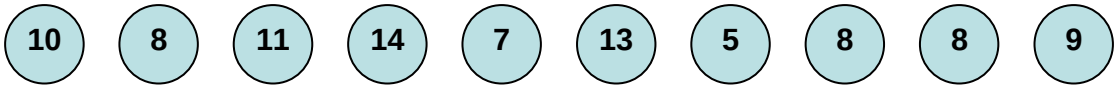
"*…epidemic of flawed statistical analysis plaguing today's research literature. From medical "breakthroughs" that prove to be mirages to grand conclusions drawn from tiny samples, the journals are awash with unreliable findings based on faulty statistics*." Matthews, New Scientist 2385, March 2003

"*A recent survey of 141 papers published in the journal "Infection and Immunity" showed ~50% to contain errors in statistical analysis or reporting of results*". Olsen (2003).

Having designed and run your experiment properly, you will, most likely, now have a lot of data to analyse. To begin with, you may wish to provide some summary stats as a way of easing into your statistical analysis. Your summary stats may include recording the number of observations you have made at each time point, working out averages, ranges and standard deviations among others. Let's briefly look at how to calculate, the mean, median, mode (and its frequency in your data), the range and then calculate the standard deviation and the standard error.

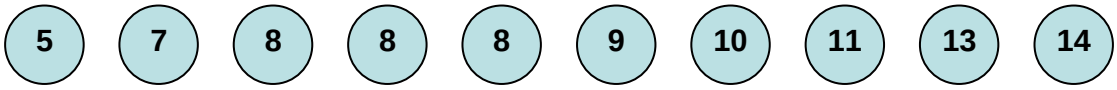
## Mean / median / mode / range

To calculate the **mean**, here are 10 numbers selected at random:



The mean is the sum of these 10 numbers divided by 10 i.e. 93 / 10 = **9.3**

To calculate the **median**, place all the numbers in order of size, like this:




Now find the middle number or middle numbers in your series. In the example above there are two middle numbers "8" and "9", so add these together and divide by 2 = **8.5**

To calculate the **mode**, look for the number that occurs the most in the series. In the example above, it is "8" and its **frequency** (i.e. the number of times it appears), is "3" (i.e. 3 times).

The **range** of your data, is the upper and lower limits of your observations, so from the number series above, your range is 5-14.

## Standard error, standard deviations and variance

### Standard deviations

Standard deviations are used to explain how widely spread the values in your collected data are, normally in relation to the mean of your data. To give you a very quick example, let us consider "how hairy are university lectures?" Here are two lecturers.

**Dr Smith** **Dr Brown**

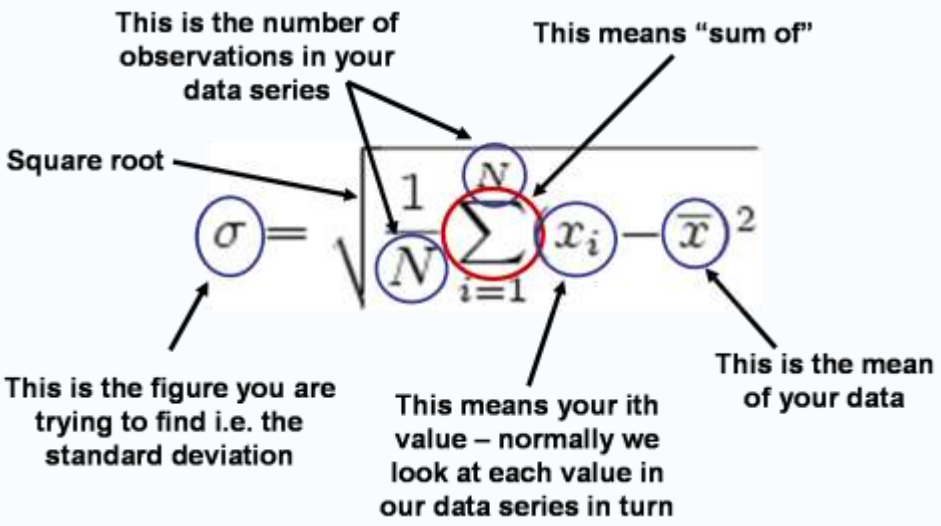
Dr Smith only gets a score of "4" on a scale of 1-10. However, Dr Brown gets a score of "8. If we consider the **average** of their "hairiness" scores (4 + 8 / 2), we get the answer "6", or we can say that their **mean** score is "6". The standard deviation of these scores is "2" (2 from 6 in either direction is

"4" and "8") and so we would write this like this: 6 ± 2 (mean ± standard deviation). In its abbreviated form we can write the mean ± standard deviation like this χ ± s.dev

We can calculate the standard deviation using the following formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Where…….




Okay let's use this equation and calculate the standard deviation but first we need another example. Let us consider the amount of money spent by each lecturer on fish last week.

**£5** **£6** **£8** **£9**

First let us calculate the mean: 5 + 6 + 8 + 9 = 28 / 4 = £7

Each lecturer spent on average £7 on fish last week.

If we put the values into the equation, then:

$\overline{x}$ = 7

N = 4 (4 observations)

$X_i$ = each of the observations in turn

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

$$\sigma = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(x_i - \overline{x})^2}$$

$$\sigma = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(x_i - 7)^2}$$

$$\sigma = \sqrt{\frac{1}{4}\left[(x_1 - 7)^2 + (x_2 - 7)^2 + (x_3 - 7)^2 + (x_4 - 7)^2\right]}$$

$$\sigma = \sqrt{\frac{1}{4}\left[(5 - 7)^2 + (6 - 7)^2 + (8 - 7)^2 + (9 - 7)^2\right]}$$

$$\sigma = \sqrt{\frac{1}{4}\left((-2)^2 + (-1)^2 + 1^2 + 2^2\right)}$$

$$\sigma = \sqrt{\frac{1}{4}(4 + 1 + 1 + 4)}$$

$$\sigma = \sqrt{\frac{10}{4}}$$

$$\sigma = \sqrt{\frac{5}{2}}$$

$$\sigma = 1.5811$$

So from the worked sample above, we can see that the mean ± standard deviation is 7 ± 1.5811.

This of course can be much more easily calculated on a scientific calculator or within a spreadsheet.

**Standard error**

In the "Standard deviation" section above, we looked at the amount of money spent by four lecturers on fish each week. We calculated that on average they spend £7 per week ± £1.58 (mean ± standard deviation). However, if we were to add the amount spent by another four lecturers, then we would certainly get a different mean and a standard deviation. Let's quickly investigate:


If the mean ± standard deviation are re-calculated, you should get a figure of 7.375 ± 2.326.

What the standard deviation is doing is giving us an estimate of the population mean (i.e. all the data that we could ever get on lecturers and their average weekly spending on fish) but what we really want to know is "how good is our estimate of the mean?" and to do this we have to calculate the standard error. If we only use the standard deviation, then we are only looking at the amount of variation around a particular estimate of the mean. If we continue to ask batches of four lecturers about how much they spend on fish each week and calculate their mean spending, we would get a slightly different mean each time. If we then plotted these averages as a frequency histogram, then we would get a normal distribution. To find or calculate the standard error we need to find the "mean of the means" and then calculate a standard deviation of it.

This can be done very simply by using the following formula:

$$\frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation and *n* is the number of observations in the data set. Above we calculated that our new mean and standard deviation was 7.375 ± 2.326 based on 8 observations. If we put these figures into our formula we get:

$$\frac{2.326}{\sqrt{8}} \quad \text{or} \quad \frac{2.326}{2.828} = 0.822$$

**Variance**

In statistics, we use the variance to describe the amount of dispersion that there is in our data and how it is spread about the expected mean value. A more understandable measure of this is the square root of the variance or the standard deviation which gives us an idea of the possible deviations from the mean. There are two ways that we can calculate the variance ($S^2$).

***Method 1:***

First we can find out how much each of our observations differs from the mean of all our observations (i.e. calculate the **deviation** (***d***) of each variable). Then, if we square each of these deviations to get $d^2$ and then add them all up to get the sum of the squares or $Ed^2$ and then divide this number by the number of observations we made minus 1 (i.e. divide by ***n-1***) then we get the variance (which is also the square of the standard deviation).

Let's apply this to the data we already have on the amount of money each lecturer spends on fish each week:

There are eight lecturers (i.e. 8 observations)

The average amount they spend is: 5 + 6 + 8 + 9 + 4 + 7 + 9 + 11 = 59 / 8 = 7.375

Now calculate the deviation (d) for each observation minus the mean:

| | | |
|---|---|---|
| **Lecturer 1** | = 5 – 7.375 | = -2.375 |
| **Lecturer 2** | = 6 – 7.375 | = -1.375 |
| **Lecturer 3** | = 8 – 7.375 | = +0.625 |

| | | |
|---|---|---|
| **Lecturer 4** | = 9 – 7.375 | = +1.625 |
| **Lecturer 5** | = 4 – 7.375 | = -3.375 |
| **Lecturer 6** | = 7 – 7.375 | = -0.375 |
| **Lecturer 7** | = 9 – 7.375 | = +1.625 |
| **Lecturer 8** | = 11 – 7.375 | = +3.625 |

Now square each deviation (***d***):

| | |
|---|---|
| = -2.375 × -2.375 | = 5.641 |
| = -1.375 × -1.375 | = 1.891 |
| = +0.625 × +0.625 | = 0.391 |
| = +1.625 × +1.625 | = 2.641 |
| = -3.375 × -3.375 | = 11.391 |
| = -0.375 × -0.375 | = 0.141 |
| = +1.625 × +1.625 | = 2.641 |
| = +3.625 × +3.625 | = 13.141 |

Now sum (***E***) all these $d^2$ values:

$Ed^2$ = 5.641 + 1.891 + 0.391 + 2.641 + 11.391 + 0.141 + 2.641 + 13.141 = 37.878

Now divide by the number of observation minus 1 (***n – 1***):

Variance ($S^2$) = 37.878 / (8-1) = 37.878 / 7 = **5.411**

**$S^2$ = 5.411**

***Method 2:***

We can use a couple of formulae to help us work this out quickly:

$$\sum d^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$\text{sample variance } (S^2) = \frac{\sum d^2}{n-1}$$

Where: ***Ex*** = the sum of all our observations

**n** = the number of our observations

Let's use our data again:

Our lecturers gave us 8 observations: 5, 6, 8, 9, 4, 7, 9, 11

The sum of these (***Ex***) = 59

$\boldsymbol{Ex^2} = 5^2 + 6^2 + 8^2 + 9^2 + 4^2 + 7^2 + 9^2 + 11^2 = 25 + 36 + 64 + 81 + 16 + 49 + 81 + 121 = 473$

If we put these into formula 1:

$$\sum d^2 = 473 - \frac{(59)^2}{8}$$

$$\sum d^2 = 473 - \frac{3481}{8}$$

$$\sum d^2 = 473 - 435.125$$

$$\sum d^2 = 37.875$$

Now put this value into formula 2:

$$\text{Sample variance } (S^2) = \frac{37.875}{(8-1)}$$

$$\text{Sample variance } (S^2) = 5.411$$

## The distribution of data

Now that you have used a variety of basic descriptive stats to describe your data you may also wish to describe how the variation is distributed in statistical terms. For example, if you were measure the height of all the humans in the world and then plot the number of people you found of a certain height into each category as a histogram, then you would get height of each person then it would look like this in Figure 1:

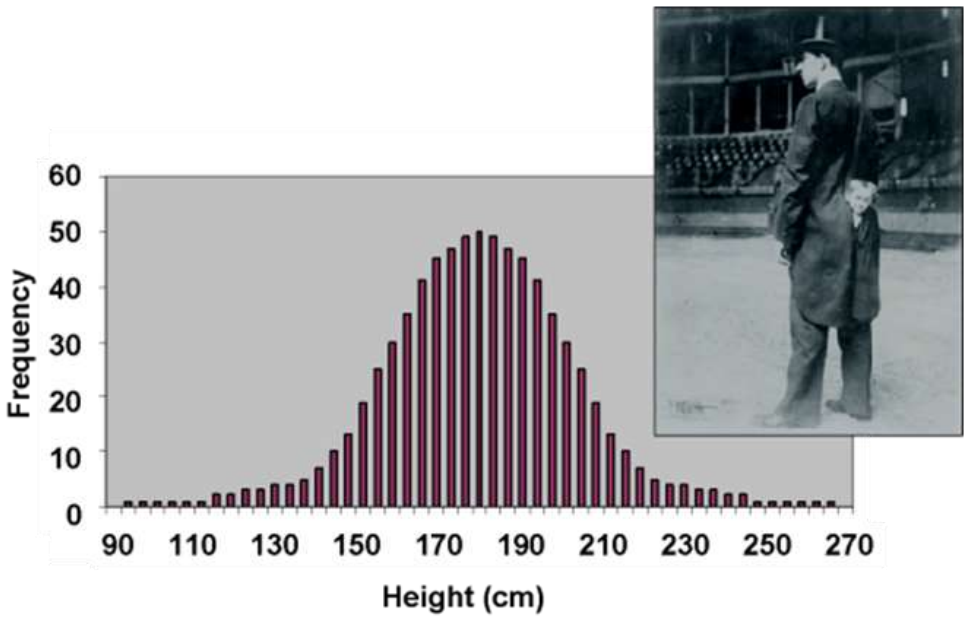



***Figure 1: A normal distribution of human height showing a typical "bell-shaped" curve.***

The inset shows individuals that might be found on the outer limits of the distribution. One is the Welsh giant George Auger (255 cm tall), with Tom Sordie (72.5 cm tall) poking his head out through the tails of Auger's coat (c. 1916). Image taken from www.missioncreep.com/mundie/gallery/gallery13.htm. You can see from Figure 1 above, that you get a symmetrical typically "bell-shaped" curve. In this sort of distribution, most of the people that you measured would be close to mean height of say 180 cm (if we are to believe that this could be the average height of humans) with progressively fewer people as you moved away from the mean. At the very end of the distributions you would have only very, very short people or giants at the other end of the spectrum. Many of the statistical tests that we will describe in this unit are for data that are normally distributed, however, we will also look at some tests for not normally distributed data.

Let us briefly think about other types of distribution. If we were to take the average height of adults belonging to the Masai people in Africa and the average height of adults inhabiting the Arctic region (e.g. the Inuit people), then we might get a **bimodal distribution** – two peaks with a dip between them (Figure 29). The reason we get a bimodal distribution is because Masais are known to be a very tall race of people while the Inuit are known to be quite short in stature.

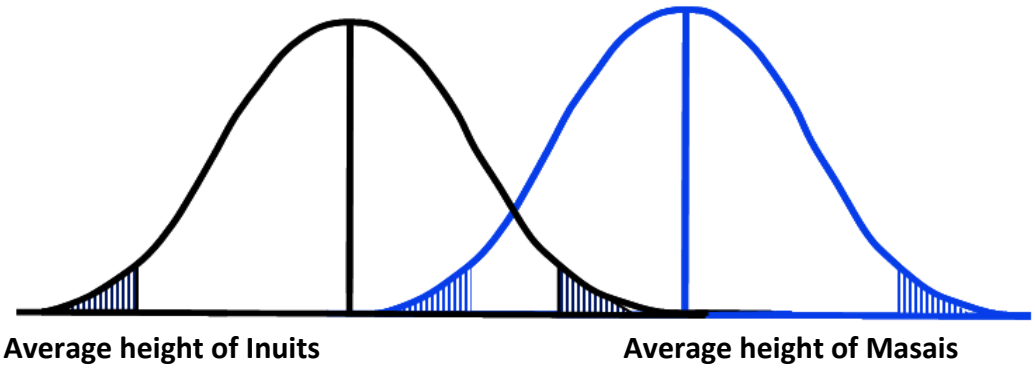



***Figure 2: A bimodal distribution – the average height of Inuit people versus the average height of Masais. Source of the Inuit woman from www.findthevision.com/archives/Eskimo%20woman%20fishing.jpg, Masai warriors from http://daryllang.com/africa/1007.html.***

Ideally, before we start to use one statistical test or another we should check the distribution of our data. As previous, if our data is normally distributed and it is plotted as a frequency histogram, then we should get a nice symmetrical "bell-shaped" curve. Our mean is the mid point of this symmetrical data and the standard deviation tell us tightly clustered our data is around the mean.

If we look at Figure 3 below, we can see that within one standard deviation of the mean (i.e. the red area) this accounts for about 68% of the observations or the measurements on height that we made. Two standard deviations from the mean (i.e. the green and the red area) accounts for 95% of the observations, while three standard deviations (red, green and blue areas) account for about 99.7% of the variation.

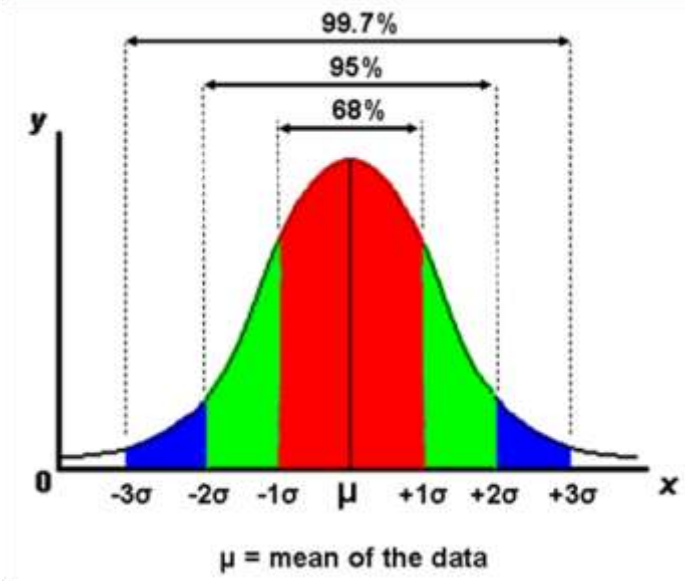



***Figure 3: A normal distribution and the standard deviations around the mean. Image adapted from www.robertniles.com/stats/stdev.shtml.***

**Skewness**

The use of certain statistical tests that depend on normally distributed data (i.e. the so-called parametric tests) may lead to the wrong conclusions being drawn if the sample distribution is skewed (Figure 4). Let us consider another example. If you set out to find the average height of people living in a city, and to do this you need to measure the height of 1000 people. If you did your study correctly, you may get a normal distribution with most people around, say, 170cm but your data set will also include a small number of very short people and a few very tall people (see Figure 4a). You may, however, get a skewed distribution if your sample was biased towards measuring children. When you plot out the data, you may find that you have a large number of "people" in your smaller height classes causing your data to skew to the left (see Figure 4b).

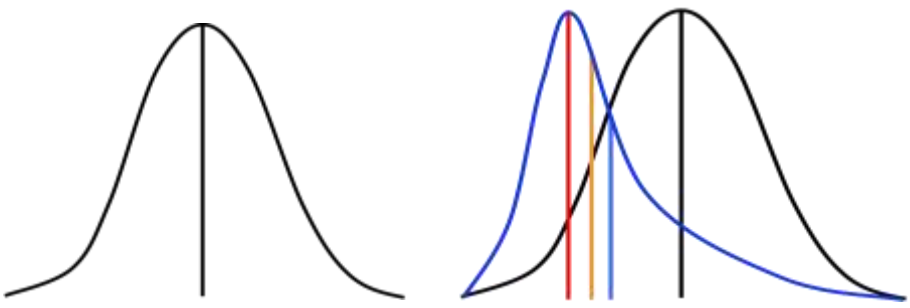



***Figure 4: A comparison between normally distributed data (black line) and data that is not normally distributed but skewed to the left (blue line).***

We will show you a little later on how to use calculated values of skew to see if your data is normally distributed or not.

**Testing for normality**

In most cases or through experience you will be able to look at your data or plot it and ascertain whether it is normally distributed. What should you do, however, if it is not clear whether your data is not normally distributed? Well fortunately, there is a statistical test you can do to test for normality.

Let's perform a step by step basic statistical summary of your data in Microsoft Office Excel:

1. Open up the Excel programme on your computer

2. Now type in the following data into the first column. Imagine that this data represents the first ten marks that a student "Mr X" received on an MSc course in fish husbandry (Figure 5). His marks were: 56, 73, 78, 67, 91, 71, 74, 69, 68, 77.

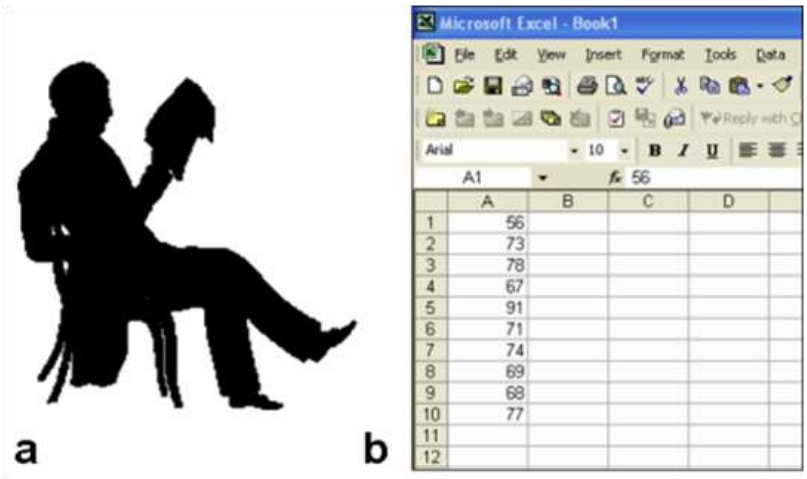



***Figure 5: a. "Mr X" studying his course materials – his image has been blacked out to protect his identity. (Source: www.stopshrinks.org/images/maninchairreadingbook.gif). b. Mr X's data entered into the first column in an Excel spreadsheet.***

3. Now run some basic descriptive statistics on this data.

Go to the **Tools** menu → **Data Analysis**

If the **Data Analysis** option is not there then:

Go to **Tools** menu → **Add-Ins** → select the **Analysis ToolPax** box → click **OK (**Fig. 6a)

If you have already have this or have loaded it up then:

Go to **Tools** menu → **Data Analysis** → **Descriptive Statistics** → click **OK**

Now click on "**Summary Statistics**" (arrowed) and under the "**Input Range**" put in where the data is in your table i.e. A1:A10 (i.e. column A cell 1 to column A cell 10) (arrowed, Fig. 6b). Then click "**OK**".

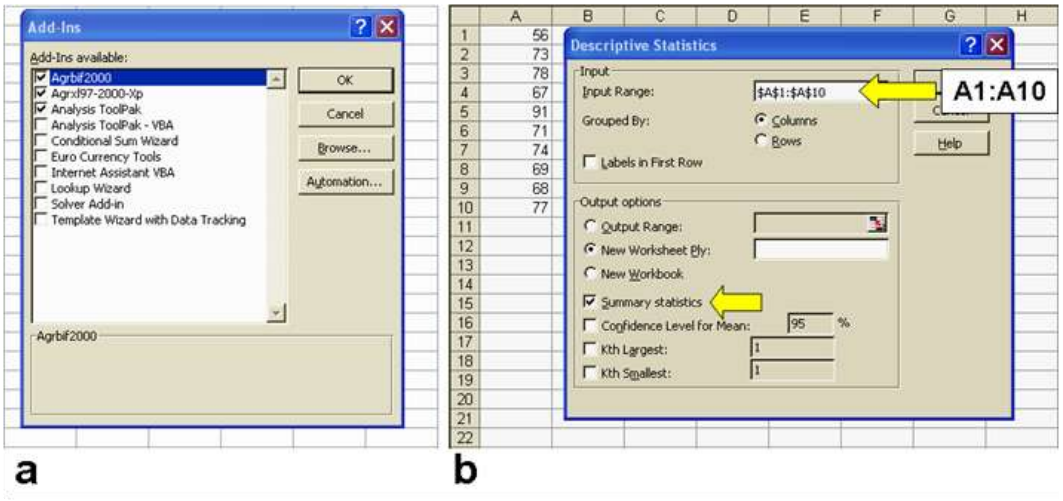



***Figure 6: a. The Add-Ins window in Excel. b. Entering the details into the Descriptive Statistics window in Excel.***

4. You should get the following (Figure 34) – the results of the analysis may appear on another "**Sheet**" so remember to have a look at all sheets. Your data will still be present but on the first sheet.

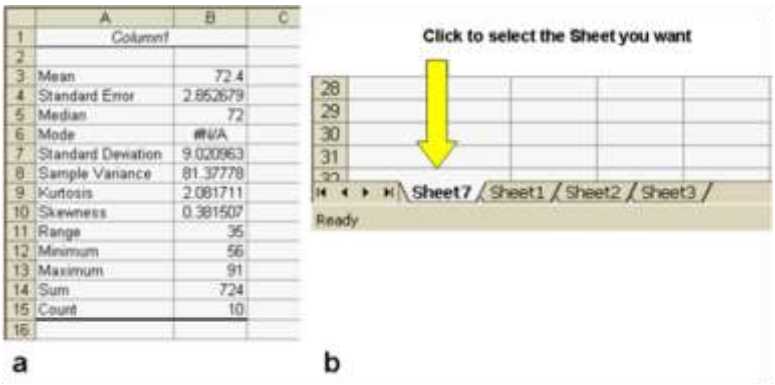



***Figure 7: a. The summary produced from the descriptive statistics option in Excel. b. Selecting the relative Sheet in Excel.***

From the descriptive statistics, you can see that we have the mean (72.4), standard error (2.852679), median (72), standard deviation (9.020963), sample variance (81.37778), the range of the data given as a maximum (91) and minimum (56) and the distance between them (range = 35), and the sum of all the marks (724).

5. What about normality? We can also use the "Skew" value given in the summary of the Descriptive Statistics to help look at normality by explaining the degree of asymmetry of a distribution around the mean. If you remember "Skew" means that the distribution is uneven and may be slanting to one side. A normal distribution normal has a skew value of about zero but a value close to zero be an acceptable value for data that is normally distributed. This value changes the more the data is skewed to one side or another. If the value of skew, however, is more than two standard deviations to the side of the mean in either direction then this is a good indication that the data is skewed and not normally distributed. Fortunately, we can approximate this using a formula provided by Tabachnick & Fidell (1996):

$$\sqrt{\frac{6}{N}}$$

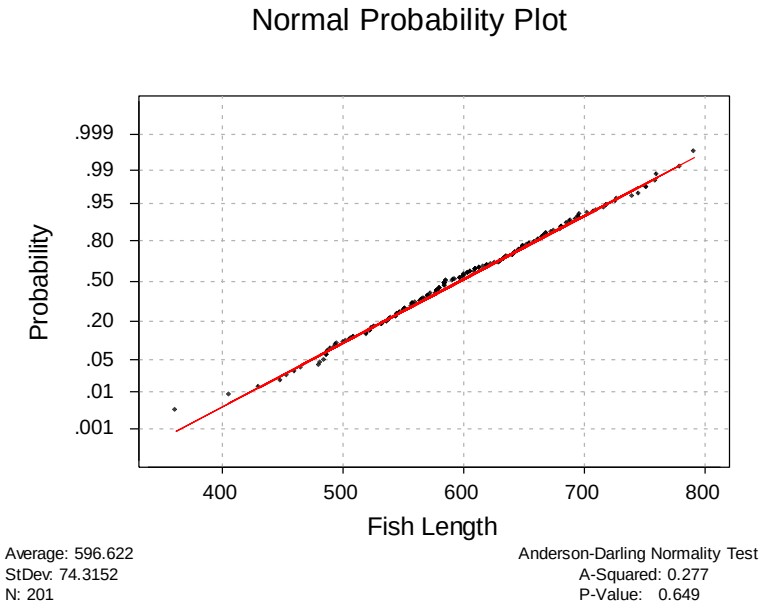
If we look back at the value of skew we calculated for Mr X's course assignment marks in the summary of the Descriptive Statistics, we calculated a skew value of 0.381507 based on the 10 marks he has so far. Therefore, if we put this into the equation, we get:

$$\sqrt{\frac{6}{10}} = \sqrt{0.6} = 0.7745966$$

As two times the standard deviation of the skewness is 1.5491932 and the value of the skew we calculated in our summary statistics was 0.381507 which is less than the two standard deviations we have calculated. So we can say that our data is not skewed and normally distributed.

It is important to note that the sign of the skew value indicates which way the data is skewed. If the sign is "+" (positive) it indicates that the data is skewed to the left, while a "-" (negative) indicates it is skewed to the right.

In addition to skewness, there are a number of tests which test normality of the data. One of the most commonly used is the Anderson-Darling test, which can be accessed from statistical packages such as Minitab. This will give an output such as below:



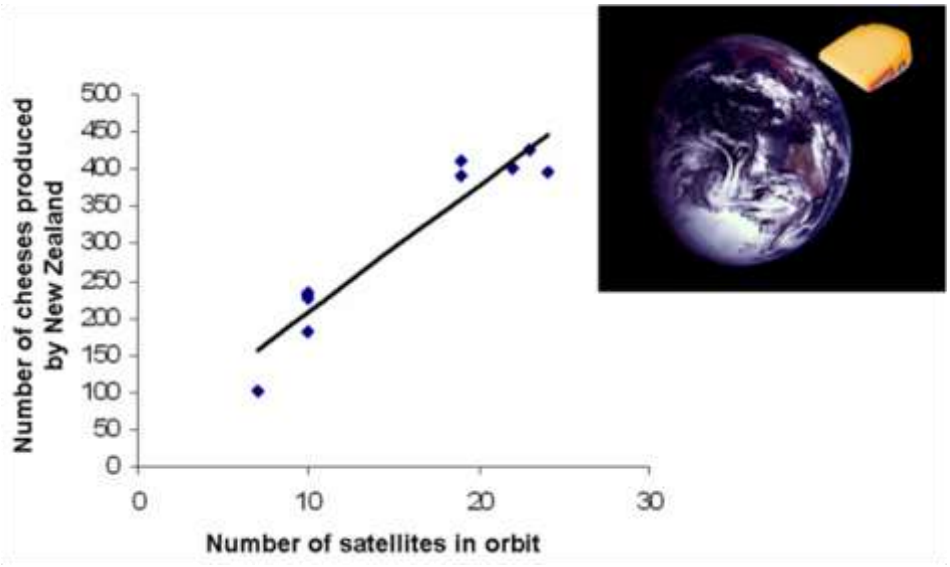

If the probability value for the test is GREATER THAN 0.05 then your sample is NOT significantly different from a normal distribution i.e. it IS normally distributed. If your data are normally distributed then they are suitable for parametric (normal) statistical analysis.

## Statistical Traps

Statistical analyses must be used sensibly and have clear causality for example:

- "A statistical test is not an infallible guide and can never *prove* a particular hypothesis. There is always a possibility that an alternative hypothesis exists" Elliot (1993)
- Causal associations may be two way – small fish may be more susceptible to high infection levels **OR** high infection levels may give rise to smaller fish.
- A correlation between two variables **does not** mean that they are associated.

A correlation between two variables does not mean that they are associated. The graph below shows the correlation between the number of satellites orbiting the Earth in the period 1994-2003 (figures from www.ceip.org) and the number of cheeses produced in New Zealand over the same period (figures taken from www.thecheeseweb.com).




A correlation between two variables does not mean that they are associated. The graph above shows the correlation between the number of satellites orbiting the Earth in the period 1994-2003 (figures from www.ceip.org) and the number of cheeses produced in New Zealand over the same period (figures taken from www.thecheeseweb.com).

## Guides to identifying causality

It is important that for a statistical analysis to be relevant there must be causality, in other words one measured parameter is directly controlled or associated to the other. There are several considerations which can illustrate or determine causality, these are (after Hill, 1965):

1) **Strength of Association**
2) **Consistency** with other knowledge
3) **Specificity** e.g. only one recognised cause
4) **Temporality** – Exposure to risk before effect
5) **Biologic gradient** – e.g. Dose response
6) **Biologic plausibility** – Scientific theory
7) **Biologic coherence** – No conflict with knowledge
8) **Experimental evidence** – Test cause and effect
9) **Analogy** – Similarity to known association

## Probability vs. Prediction

Statistics allows us to evaluate the probability of a given event - it is **impossible** to predict the outcome of a given event **for certain** using statistics. It is only possible to predict the probable outcome. For the biological sciences a probability of **95%** is by convention accepted as **significant** i.e. $P<0.05$. This is because biological and ecological systems are rarely static and consistently prone to externalities. Therefore trying to achieve a "perfect" 100 % probability is virtually impossible. By accepting $p<0.05$ investigators are acknowledging that their statistical proof of a tested hypothesis will be erroneous one time out of every twenty.

The probability taken to be significant in a given study should **always** be stated.

## Problems Faced in Aquaculture Data Analysis

- Much of the data encountered in aquaculture is non-normal such that classical tests such as t-tests, ANOVAs and correlation coefficients are not justified on the raw data as collected or measured.
- Often, one is interested in comparing multiple (>2) groups e.g. three morphotypes / species, treatments and controls etc. Such multiple comparisons cannot be properly be carried out using multiple paired comparisons e.g. t-tests.

Vocational Aqualabs

TRAINING

# Many thanks for your attention today

## Experimental Design
## Unit 2a

For further information please contact:

Trevor Telfer/James Bron
Institute of Aquaculture
University of Stirling

Education and Culture DG
Lifelong Learning Programme

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained herein.