

Workbook 1: Introduction to Quercus

Any characteristic that *varies* between elements is a variable. Variables can be Qualitative or Quantitative. **Qualitative**: A variable whose values *cannot be measured* but can be assigned to **categories**. Sometimes referred to as *categorical variables*. Qualitative variables can be split into **2 groups**: Nominal (categories have no *implicit order*, not *hierarchical*) or Ordinal (*do have* an implicit order e.g. *absent, rare, common, dominant*).

Quantitative: If the variable can be *measured* or the amount of it *quantified* then it is said to be a **quantitative** variable. Quantitative variables can also be divided into 2 groups: Discrete (where observations have to be *whole numbers*) or Continuous (Even the *smallest* quantities are not separate — have to *round up or down*).

Workbook 2a: Minitab Part 1

Windows in Minitab: *Session, Data, History, Information*. Column names: no more than **8 characters** long, do not begin/end with a **space**; do not use the name of the **file** you are using, do not use **quotes** (‘) or **octothorpes** (#). To move to the **last cell** in the worksheet, you press *Ctrl+End*.

Workbook 2b: Minitab Part 2

When **retrieving** worksheets, use retrieve ‘c:\files\steroid.mtw’ (Must use **single** quotes). To **import** data from a text file, use the command read ‘O:\TLTP\Quercus\Mfiles>List.txt’ into c1-c6. To have **missing** values in a worksheet, enter “*”. DIR lists the files in any specified *directory*. To create a **table** of e.g. means, use table ‘depth’ ‘nutrient’; means ‘wheat’. (The **semicolon** indicates a *subcommand* is to follow).

Workbook 3: Graphical Presentation

Section 1: An experiment to compare the *effects* of a steroid treatment with a **placebo** on the recovery of white blood cells after *chemotherapy* is discussed. The **expected** outcome is considered. The data in *STEROID.MTW* is loaded and the **variability** of the data is examined.

The first thing to do with **data** is to look at it. In the steroid experiment, it is claimed that steroids would *enhance blood cell recovery* against treatment with a placebo (a pharmacologically inactive substance which is administered as a **drug** in the course of drug trials). Since all the patients are *different*, their responses are likely to be different.

The **white** blood cell counts (10^6ml) for both the steroid treatment and the placebo groups are in columns labelled 'WSteroid' and 'WPlacebo'. It is understandable that variations in the data cannot be detected just by **looking** at the data. In order to gauge the extent of variability within groups, we create *graphical displays* of the data.

Section 2: Graphs and plots are **visual** summaries of the distribution of observations in a *sample*. The type of plot used to present a data set depends on the **type** of variable being measured. Quantitative data can be presented as a *stem and leaf plot*, a *histogram*, a *box plot* or a *dot plot*. Stem and leaf plots are taken as an example and their construction and analysis are discussed in detail using part of the white blood cell count data.

Displaying Quantitative Data

The **variability** in data sets make it hard to *see* differences. Displaying the data as a graph allows us to see **how** the data is distributed. One way to display small sets of quantitative data is a *stem and leaf* plot. The 1st digit of each observation is listed **vertically** in descending order to the *left* of a vertical line, so forming the stem. The remaining digits of each observation are listed (in order) to the **right** of the appropriate stem, creating the *leaves*.

0		5	9
1		1	2 7 9
2		0	0 1
3		3	4 6 6
4		0	1 6
5		2	

Before we can create a stem and leaf plot, we need to *organise the observations* (the observed values recorded for the variable in the sample) in a **frequency distribution**. (Qualitative: the set of values of the variable with their associated frequencies. Quantitative: the frequency distribution is the **set of class intervals** together with the associated frequencies of observations in *each* interval).

To do this, the observations are ordered from **least** to **greatest** into groups (intervals). N.B. the numbers in the data column give a *visual impression* of how the data is distributed.

Creating a Stem and Leaf plot

(1) In the *frequency distribution*, the integer part of each observation is erased and listed in a column on the **left**. These numbers form the “step”. (2) The numbers after the decimal places (tenths) are then arranged from *least to greatest* to form the “leaves”. Note: the effect is similar to that of a histogram lying on its side. In **MINITAB**, Graph > Character Graphs > Stem + Leaf.

In the dialog box, the *increment* is the distance between the **smallest** possible number on one line and the smallest possible number on the **next** line. Minitab output: N = number of observations; the *left* column is a **cumulative count** of the number of observations in each class on either side of the median. Median class count in **brackets**. 2nd column is the *stem*; remaining columns are *leaves*.

Analysing: (1) Look for **gaps**/subgroups. More than one *empty class* may occur between extreme values and the main body of the data, or they may **split** a data group into 2+ subgroups, perhaps indicating a pattern *not previously known*. (2) Look for **outliers**, observations outside the normal range. It is a matter of judgement where an observation *becomes* an outlier. In some cases, outliers may be **atypical** and should be disregarded, but in some cases it may be outliers that are the most instructive cases e.g. the one patient who does **not** die of the disease.

Look at the **shape** of the distribution — is it symmetrical or *skewed* to one side. Skewed: similar to symmetric data except one side is pulled **outward** so that one “tail” is longer than the other. This means that the median is *not the central class*. Also, locate the median and consider the range. A stem and leaf plot is quick & easy for small data sets. The display shows how the data is *distributed* among the classes, the *shape* of the distribution and where the *median* value lies.

Other plots include the **histogram** where the area of each bar is equal to the *proportion* of the total number of observations that fall into that interval; the **box plot** (the bar inside the box indicates the value of the **median**; the sides of the box indicates the **inter quartile range**; the ends of the lines indicate the **range**; outliers are shown by **asterisks**) and the dot plot: each observation in the sample is marked by a *dot* to form a histogram with narrower class intervals.

Section 3: How to write a **report**: there are four parts: *Introduction/Aims; Method; Results; Discussion/Conclusion*.

The **introduction** should give *background information* on the experiment and state the aims. Method: outlines the experimental **method**. Results: contains the results and the concise *description of your data analysis*. Discussion & Conclusion: After describing your results, you need to discuss what they **mean** and say what **inferences** you have drawn from them.

Comparing samples this way does not *prove or disprove* any hypothesis you may have about the effects of the treatment. To determine if differences between samples are significant, you would have to apply certain **statistical tests** as described later.

Customising Histograms

- (1) Get *dialog* box.
- (2) Select *options*.
- (3) *Type* of histogram: 6 types to choose; *type of intervals*: **midpoint/cut point**; define *intervals*: **automatic/no. of intervals/cut point positions**.
- (4) (Optional) Select number of *intervals* in the option box.
- (5) *Annotate* using the Annotation features.
- (6) Create title *etc.*, using (5)'s features.

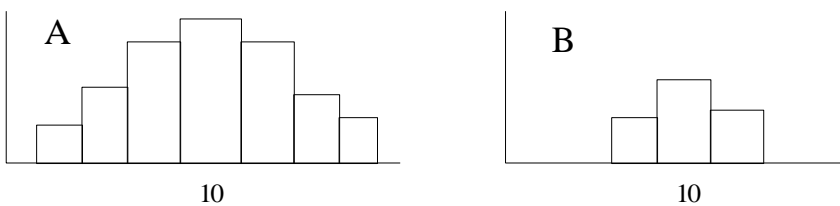
Multiple plots: Graph > Histogram. *Select C1 then C2*. Click on the Frame option, select Multiple graphs; select the **overlay** graphs option; 2 superimposed histograms will appear.

Workbook 4: Measuring Location

Section 1: **Comparing** data sets. To compare data sets, we need *single number summaries* (statistics). The 1st step in this process is to classify the shape of the distribution, skewed or symmetric, and to identify any **outliers**. The shape and presence of outliers may have a significant effect on the value of any statistics we calculate. The interpretation of visual summaries is **subjective**, and can be imprecise. We need more *objective* ways to compare data.

Looking at Distributions

Suppose A and B represent distribution of *response times*. Both have the same location (av. response time) but A's response times are more variable.



C and D have different *locations* but the same *variability*.



First, we need to **describe** objectively. Do this by measuring certain characteristics of the distribution. When describing the distribution of a data set, the 1st 2 characteristics you should mention are Location (where the centre of the data is) and Dispersion (The spread of the data).

Describing Distributions

In order to quantify *differences* between data sets, we need to have **single summary statistics** of location and variability. The simplest summary statistics are the **mode** (or modal class) and the **range**, which can be read directly from most data plots. **Mode/modal class:** most frequently occurring value. In a *dot plot*, it is the highest column. In a *stem and leaf plot* or a histogram, it is the class interval with the **largest** number of observations (frequency). Its value depends on the class interval chosen and so it is *not* unique.

Range = Maximum - Minimum. It is very *sensitive* to outliers. Summary statistics: the *peak* and *range* are useful when initially describing distributions, but are not the **only** and not even the **best** stats. As you'll see later, the most appropriate measures depend on the shape of the distribution. Shapes: the basic shapes are (a) *Symmetrical*, where both tails are of **equal** length; (b) *Skewed*, in the direction of the **longest** tail. Data with a peak occurring to the right is skewed to the **left** because of the long left tail.

Bimodal distributions: these are *distributions* (symmetrical or skewed) with 2 peaks. No single measure of location is appropriate here, and the **best** strategy is to report the 2 peaks. Example: Heights of a *random sample* of men and women.

Section 2: Mean and median are the 2 *most used* measures of location, but represent the data **differently**. The sample mean is derived arithmetically from the values of all the observations and so is dependent on these values. The median is **independent** of the values of the observations. It is dependent only on the *number of observations*.

If **mode** is not a suitable summary statistic, what is? You could take the value of the observation in the *centre* of the distribution, the median, or calculate the “*average value*”, the sample mean. **Median** = middle number when arranged in order of magnitude. **Sample mean:**

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^{i=n} x_i}{n}.$$

To read *text* data, use e.g. MTB> read 'A:\Shapes1.txt' C1. Describe 'react' generates *summary statistics*.

The mean is the **arithmetic average**. Its position within the distribution varies depending on the *values of the observations*. The **median** is the typical value in the sense that its position is always in the middle of the distribution. 50% of the data is above and below the median *always*. If the data is symmetric, the mean and median are **similar**. If the data is skewed, the mean is *different* to the median (e.g. if skewed towards lower values, the mean is **smaller**).

How does the shape affect location?

If the data is **symmetrical** and **unimodal**, (1 peak) you will find all three measures take approximately the *same* value. The mean however uses all of the data and so is the most useful. If the data is **skewed**, the mean can be far removed from the centre of the distribution, therefore the *median* is more appropriate. What other features of the data should you consider before choosing a measure of location?

Outliers. Check their effect before *selecting a measurement* of location. If an outlier increases, the mean increases, but the **median** remains the same. One way to avoid this is to use the *trimmed* mean, but consider its implications. Choosing your location statistic: examine **graphically** before calculating any statistics, taking into account the possible effects of skew and outliers. The mode should only be used for *preliminary descriptions* of data. Q: Is the difference between means of **2 different samples** large compared to the *variation within treatments*?

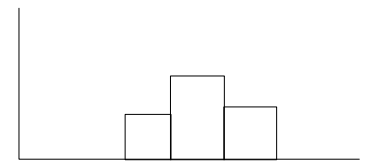
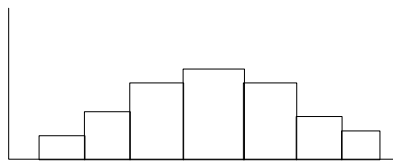
Workbook 5: Measuring Variability

Section 1: A sample consists of a number of **measurements** of a variable. The degree to which these measurements vary *within the sample* (the variability of the sample) is one of its most important characteristics. A comparison of the variability of 2 samples can help us to explore the **degree** to which they differ.

Why do we need to measure variation?

If everything was 'average' e.g. same height, weight etc., then we wouldn't need statistics. In reality, we have to deal with variability rather than **sameness**. Populations are usually too large to deal with, so we take a representative *sample*, and measure the variable(s) you are interested in e.g. height, and calculate the mean, median or modal value.

Knowing the **average** may not be enough. We need a measurement of *variability* as well. For example, a 'medium' sized coat implies a range of values either side of the average. Q:



8

What does an average waiting *time of 8 minutes* mean? Without knowing the **variability**, it is difficult to decide whether to wait or not.

Why is variation important?

Consider **histograms**. There are 2 sources of variation to consider. Variation *within* each data set (sample); differences *between the two samples* (if there is more than one sample).

- (1) **Variation within samples.** Although 2 procedures may be identical, we expect variation in each *sample/procedure*. Good procedure minimises variation, but it will not eliminate it. Consequently, statistical techniques are required to *analyse experiments* in the presence of variability.
- (2) **Differences between samples** (if 2 samples are taken). The differences between samples is important only if it is **large** compared to variation within each sample of data. Variation within samples affects the way we *interpret differences* between samples e.g. the Dentadol example, where a new treatment is compared to Dentadol. The **overlap** between the two distributions is shown. Although the new formula works 6 minutes faster on average, almost a *third* of the subjects had response times in the Dentadol range. The mean response for another formula B was also 9 minutes (as A), but the distribution was **much** less variable. The small overlap (approx. $\frac{1}{6}$) suggests that of the 2 new preparations, more patients showed a *shorter response time* for formula B than for formula A.

Measuring Variability

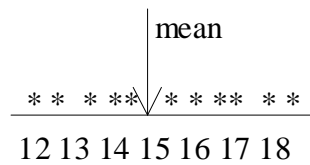
In workbook 4, variability was measured as the **range** of the data. This statistic tells us the difference between the *greatest and the least* values. Because we use measurements of location to indicate variation between data sets, we need to know how the data varies about the chosen **average**.

Section 2: How to measure *dispersion* of data about the mean. If the distribution of the data is **symmetric**, variability can be measured as the s.d. of the data set. This is a measurement of the *mean deviation of the observations* from the mean.

In the **telephone** queue, we need a statistic measuring dispersion of the data and which when added/subtracted to the average, indicates min/max waiting time. How you calculate this dispersion statistic depends on how you calculated the *average*.

Measuring dispersion about the mean

First consider **how** you calculate the mean. (1) Add the values of all the observations (Σx); (2) **Divide** by the total number of observations. As long as there are no outliers and the data is not skewed, then because all the data is used, we can *regard the mean* as being a “middle” value for the data set. As you might expect when measuring dispersion about the mean, we take into account the **difference** between the value of each observation and the mean, and try to arrive at a “*typical*” value.



The difference **between** the value of any observation and the mean is the horizontal difference between the star and the arrow. To measure dispersion about the mean, we start by calculating the *differences* between the value of each observation and the mean, $x_i - \bar{x}$. But, the sum of these deviations approximate to **zero**. Therefore we cannot use $\Sigma(x_i - \bar{x})$ to calculate the dispersion statistic. A way around this is to square each value and use $\Sigma(x_i - \bar{x})^2$. The average value of the expression $(x - \bar{x})^2$ equals the sum of the *squared deviations* divided by $(n-1)$. This statistic is called the **variance** (s^2), where $s^2 = \frac{\Sigma(x - \bar{x})^2}{n-1}$.

Why use (n-1)? When calculating the **average** deviation from the mean, we use $(n-1)$ not n . This is because we are trying to *estimate* the variance within the total population by calculating the variation of a small sample. It has been found that using n in the equation **underestimates** the population variance. Therefore, we subtract 1 from n as a correction factor to *compensate*.

So is the variance the statistic to measure the *dispersion* of data about the mean? NO! To get back to “sensible” units, take the **square root** of the variance. This is called the **Standard Deviation** (s); Where $s = \sqrt{\frac{\Sigma(\bar{x} - x)^2}{n-1}}$. This is the outline of a *continuous data distribution*, the **Normal** or **Gaussian** distribution. 95% of the area beneath the curve (95% of observations) lie within 2 s.d.’s of the mean, $\bar{x} \pm 1.96s$. 99% of the area lies between $\bar{x} \pm 2.58s$.

Section 3: If the distribution of the data is **symmetric**, variability can be measured as the s.d. of the data set. However, this measurement is easily *distorted* by the presence of outliers. If there are outliers or the data is skewed, the **Semi Interquartile Range** is used. This indicates the *mean distance of the quartiles from the median*.

Suppose **9 out of 10 calls** wait between 1-6 mins but a 10th call waits 18 mins. Because \bar{x} and s.d. take all the data into account, the outlier has a *large effect* on these stats. In the data set, we have the statistic shown in the table. The outlier has a **disproportionate** effect on the mean and s.d., but the outlier has hardly affected the *median* value. Why is this? It is because the only value used in calculating the median is that of the **middle** observation when data is in rank order.

	\bar{x}	s	median
no outlier	3.6	1.4	3.4
with outlier	4.9	4.5	3.6

So how do we measure *dispersion* about the median? Note that the median divides the data so that 50% of the observations lie **below** the median, and 50% **above**. The 1st quartile divides the data in the ratio 25:75; the 3rd quartile divides in the ratio 75:25. 50% of the data lies between the **1st and 3rd** quartiles. The difference between the first and third quartiles is called the *interquartile range*. The mean difference between the quartiles and the median is called the semi interquartile range and is the statistic used to measure *dispersion* about the mean.

To calculate the **SIR**, (1) *Order the data* according to rank; (2) Identify the *Quartiles* Q1 and Q3; (4) Calculate $SIR = \frac{1}{2}(Q3-Q1)$.

Choosing which measure of dispersion to use

- **Interquartile range** — indicates the spread of the *middle 50%* of the data and the SIR expresses this as the average distance of the quartiles from the median. Does not utilise as much of the available data as the...
- **Standard Deviation**, which measures the average spread of **ALL** the observations about the mean, but like the mean, it is sensitive to *outliers* and is unsuitable for *skewed* data or data with *outliers*.

If **S1 and S2** are two standard deviations, then their *ratio* is $S1/S2$. It is usual to put the **smaller** standard deviation as the *denominator*. Interpretation: One is $S1/S2$ times more variable than the other.

Workbook 6: Sampling Theory

Section 1: Random Sampling. In order to test the **effect** of treatment, we usually test the response of a sample group against that of a control group. The result of this comparison however is only *relevant* if we can be sure that the individuals selected for the control group are typical of the population. How we select samples can affect their reliability. The importance of random sampling is investigated.

Test a **sample** against a control. When testing for the effects of a treatment, we have to assume that there will be a level of response due to **chance**. The effects of the treatment need to be distinguished from this background response. In order to measure the background or chance level of response, a control group which receives no treatment or a placebo is used.

Before you select subjects for your experiments, think... a comparison of the reaction of a sample group against that of a control group is only **meaningful** if the control group is typical of the population. (1) What steps can you take to ensure that your control group is *typical* (i.e. avoid bias); (2) How can you measure the **reliability** of your control sample?

How should you select a sample from a population?

Populations are usually very large, which of course is why we have to take samples. For the purpose of this demonstration, we have created a population of 30 individuals from which we select 6. Results after *selection*: **Population** mean = 1.68; s.d. = 1.19; **Control** mean = 1.57; s.d. = 1.72.

Avoiding Bias: although you used a rule to try to select an *unbiased* sample, you may have had difficulty in choosing a control group which accurately represented the population. This was because you were unaware of an **underlying** pattern in the response of the subjects. If your rule coincided with this *response gradient*, you may have inadvertently biased your sample e.g. choosing the first 6 subjects means the mean is biased too **high**; and choosing the last 6 subjects means the mean is biased too **low**.

How do we avoid bias when selecting a control sample?

When there is no *known patterns* of variation within a population, the best way to avoid inadvertently selecting a biased sample is to select the sample units at **random**. (random sampling).

Section 2: The **sampling distribution**. Suppose we take a number of random samples from a population. These samples will all have *different means and s.d.'s*. This is called sampling variability. The distribution of sample means is called the *sampling distribution*. If we know the mean and s.d. of the sampling distribution, we can predict how reliable a sample is likely to be i.e. how **close** its mean will be to the sample mean.

How do you select a random sample?

One way: label each unit in the population with a **number**, and place these in a box. Pick out as many as required. Often computer *random number generators* are used to select labels. But if randomly selected, we still need some way of determining how representative a random sample is of the population. To do this, we need to estimate how much the **random** samples are likely to vary between *themselves*.

Variation between samples. If samples are selected at random, it is unlikely than any 2 samples will be the same. If we take a *number* of samples, how much are they likely to vary between themselves? Each time a sample is **selected**, different individuals are used — different sample statistic results. This illustrates *sampling variability* as a consequence of random sampling.

The Distribution of Sample Statistics

Previously, we discussed the idea of a **sample** as a number of observations. You can now see that if we take a number of samples from a population, and calculate their means, we would have a *sample of sample means*. We can investigate the distribution of the means just as we would a series of observations. To examine the distribution of the sample means, type the means data into a worksheet and calculate the **mean** of means and the s.d. of the sample *means*.

Comparing Population and Sample Statistics

Create an **annotated** histogram (Select Annotate in the Hist. dialog box; Select Text; in the text box select the co-ordinates of the point you want the text to appear in, and of course type in the text itself. Draw lines also from the Annotate menu — specify the co-ordinates e.g. (15 6 15 0)). Include the location of the **population mean** (μ), the mean of means, and the range of values one s.d. on either side of this mean.

Think: how close is the mean of the *sample means* compared to μ ? How wide is $\bar{x} \pm s$? Suppose you take another sample. Can you see now how an **appreciation** of sampling variability will help you determine how representative this sample will be of the population?

Section 2: Simulate taking **large** numbers of samples from a population using Minitab. (Population is normal, of *known mean and s.d.*) The effect of sample size on sampling distribution is explored. The **Standard Error** is calculated and this is shown to be a good *approximation* for the s.d. of the sampling distribution.

We can use Minitab to simulate the effect of taking a *large number* of random samples from a population of known mean and s.d. We can then calculate the mean and s.d. of these samples and so explore the **sampling distribution**.

Generating Random Samples using Minitab

The **RANDOM** command generates a random sample of any size you choose from a “population” whose mean and s.d. you specify. Generated data appears in the worksheet. Here, there will be *10 columns, each 500 rows long*. Each row of random numbers represents a “sample” of 10 observations. *CALC > RANDOM DATA > NORMAL*. Specify the number of samples (500 rows), **population** mean ($\mu = 100$); population standard deviation ($\delta = 15$). The **INFO** box shows that the data has been created.

Create **means** for samples. *CALC > ROW STATISTICS*. Indicate you wish to calculate means. Set input variables to C1-C10 and store the results in C11. Calculate means for smaller samples: use only **columns** C1-C5 (In C12).

Explore and Compare Sample Distributions

Calculate *mean and s.d.*; create dot plots of C11 and C12: the dot plots are roughly normal in shape. Note: if you take a **large** number of samples (e.g. 500), the mean of the sample means approximates to the mean of the population. Also, the variability in the sampling distribution as measured by the s.d. *decreases* as sample size increases.

At the start of this workbook we asked: **how** can we know if the control sample *represents* the population? We have shown that (i) we can avoid **deliberate** or **accidental** bias by using *random* samples; (ii) if we know how the sample means vary about the population mean, (i.e. what is the *s.d.* of the sampling distribution) then we can predict **how** close the mean of a randomly selected sample is likely to be to the *population* mean. (iii) If we increase the size of the sample, then we decrease the **variability** of the sampling distribution. This implies that the larger your *sample* size, the closer the sample mean is likely to be to the **population** mean.

But — how do we **measure** sampling variability? Do we have to take a large number of samples and calculate the s.d. of the sampling distribution every time we do an experiment? *Of course not!* We use the **standard error**, $SE = \frac{\sigma}{\sqrt{n}}$. The SE measures the *precision* of the sample mean (i.e. it measures how well \bar{x} approximates μ). As n (the sample size) increases, the SE decreases, implying that **larger** samples are more precise than *smaller* samples.

If you take samples of **size** n from a normal distribution, then the mean of the sample also follows a *normal distribution*, with the same mean but with a smaller dispersion i.e. Standard Error. Now investigate the effect of taking a **large** number of random samples from a skew population of known mean and s.d. We can then calculate the *mean* of each of these samples and so explore the **Sampling Distribution** of the sample mean.

Exponential distribution: used as a model for *lengths* of time, *survival* from an intervention. Use Minitab to generate random samples from an exponential distribution. *CALC > RANDOM DATA > EXPONENTIAL*. **Specify** 500 rows, columns 1-10, $\mu=10$. Note: each row in the worksheet contains 10 numbers taken at random from an exponential population whose mean is 10 and *whose s.d. is consequently* 3.16 i.e. $\sqrt{10}$. There are 500 rows, therefore you have 500 samples, **each** of 10 observations.

Calculating Means for Samples

CALC > ROW STATISTICS. Check the mean box. **Input** = C1-C10; **Output** C11. Means for smaller samples: Input C1-C5; Output to C12. Calculate **mean, s.d.**, create dot plots for C11 and C12. The shape of these plots are *skewed*, but not as skewed as the original exponential distribution.

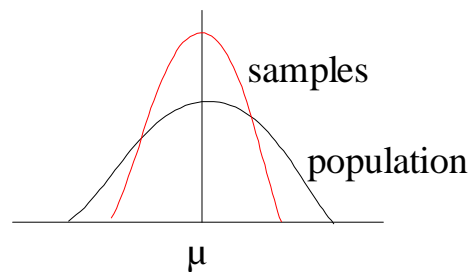
Note: with **larger** samples, the means of the sample means are close to the mean of the population. Note also how the *variability* in the sampling distribution as measured by the SE, given by the formula $\frac{\delta}{\sqrt{n}}$, decreases as the sample size increases.

Summary: when taking **small random** samples from a skew population, the distribution of the sample mean is still skew, although *not as skew* as the original distribution. Repeat the experiment using **larger** samples (20-30). In this exercise, it has been demonstrated that when sampling at *random* from a skew population, the distribution of the sample mean is (i) Normal, provided the sample size is sufficiently **large** ($n > 30$ is usually sufficient); (ii) SE is $\frac{\delta}{\sqrt{n}}$. This decreases as the sample size *increases*, indicating that the mean is more *precise* in large samples.

Workbook 7: Confidence Intervals

Section 1: **Why** are CI's necessary? We need some *indication* of how representative a sample is of the population it is drawn from. A CI is a range of values about the mean of a sample, based on our estimation of the SE of the sampling distribution. A percentage value is *attached to the interval*, which indicates the proportion of intervals which will include the **population** mean.

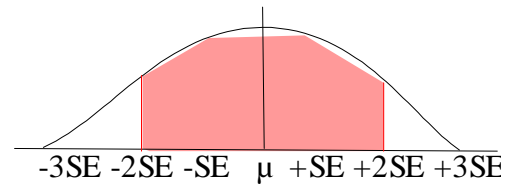
How can 1 sample tell us about the **sampling distribution** of a *statistic* in a population? Suppose we have a dot plot of the “*Reaction Time*” for a population. Suppose we select a number of samples and calculate the **mean** of each of these samples. If the number of samples were infinitely large, then the outline of the sample distribution would look like as shown in red. Note: both distributions are normal; population mean and the mean of the sampling distribution are *equal*. The relationship between the s.d. of the population (δ) and the s.d. of the sampling distribution ($\delta_{\bar{x}}$) [SE] is given by $\delta_{\bar{x}} = \frac{\delta}{\sqrt{n}}$.



What if we *don't know* the value of δ ? Then we use the best **estimate** that we have for the population mean i.e. the sample *standard deviation* s . In **practise**, we usually estimate the s.d. of the sampling distribution (the SE) as $SE = S_{\bar{x}} = \frac{s}{\sqrt{n}}$. Therefore from just 1 sample, we can estimate the s.d. of a sampling distribution.

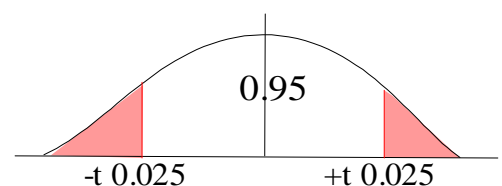
Consider having to select 10 subjects as a control. They're given CureAll, with *mean reaction time* 30 minutes. The results were: mean response time: 31.78 minutes, s.d. 4.11. To determine whether a sample is typical of the population, first calculate the SE. A difference between the sample and population mean may be due to **sampling variability**. If this is the case, then the mean of the sampling distribution will be the same as the mean of the population.

Consider the *control mean* to be a single random case drawn from an infinite population of sample means, whose mean is equal to μ . We **expect** that 95% of all possible control sample means will lie within 1.96SE of μ . We use \bar{x} to estimate the value of μ . We **interpret** this to mean that it is expected that for 95% of samples, the interval $\bar{x} \pm 1.96SE$ will *contain the population mean* (μ). But, because when we calculate the SE, we use the estimate $\hat{\delta}$, we must therefore **introduce** a correction factor, $t_{\alpha/2}$. **Therefore** the 95% CI is calculated as $\bar{x} \pm t_{\alpha/2}SE$.



Note: t is a theoretical continuous distribution based on the *differences* between the sample and population means divided by $\hat{\delta}/\sqrt{n}$. The **shape** of this distribution depends upon the value of n . It is *wider than the normal curve* at low values of n , but becomes indistinguishable from it when $n = 30$ or more.

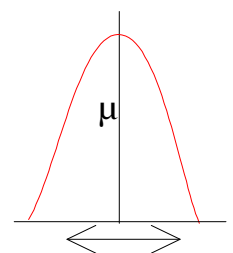
Note: $\alpha/2$, the “critical value”. Alpha (α) is the area *under the curve* outside the CI. For e.g. a 95% interval, $\alpha = 5\%$. 95% of the area **beneath** the curve lies in the unshaded area. The **remaining** 5% is divided between the 2 ‘tails’ of the distribution. Each tail *contains* $\alpha/2$ (0.025) of the area *under the curve*.



95% Confidence Intervals

CI’s indicate the **precision** with which the mean of a sample estimates the mean of the population. The *interval is in the range* of values about the sample mean which we are 95% confident will contain the population mean. The **narrower** the interval, the more precise the estimate. For the experiment mentioned, the corrected *confidence interval* (t interval) for the control sample is calculated as $\bar{x} \pm t_{0.025}SE$; $31.78 \pm 2.26 \times 1.3$; 31.78 ± 2.94 .

There’s a 95% chance that the interval 28.84 to 34.72 contains the mean of the population from which the control sample was taken. Since this interval *includes* the expected population mean, 30, it is probable that the sample is **typical** of the population. Suppose we simulate taking 100 samples. The arrows indicate the 95% CI for each sample as shown. *Approximately* 95% of the intervals will include μ . Suppose we increase the sample size to 30 — the larger n is, the narrower the interval in which we are 95% **confident** μ will lie. The width of the CI depends on the SE of the sample. The *narrower the CI* is, the *greater* its precision. (Accuracy = Bias + Precision).



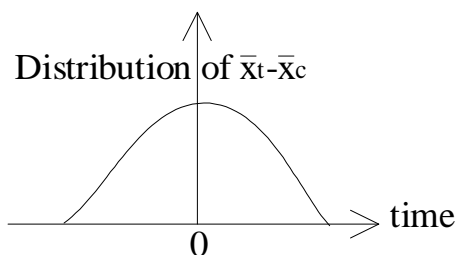
Section 2: **Calculating** CI’s. Use the t -interval command to calculate the CI (95%, 99% or 99.9%) for a column of data. To calculate the CI for the difference in the means of 2 samples, either use the two sample command or subtract one column from the other and use the t -interval command on the difference data.

Calculating the CI for a **single** sample: The 95% CI is *calculated* as $\bar{x} \pm t_{0.025} \times s/\sqrt{n}$, $df = n-1$. (*Degrees of freedom*: the exact shape of the t-distribution depends on a quantity known as “degrees of freedom”, df . Here, the **value** of df is one less *than the size of the sample* i.e. $df = n-1$. This is the number of observations which are **free** to vary as a result of chance alone).

Using **Minitab**, STAT > BASIC STATISTICS > 1 SAMPLE T. In the dialog box, select the variable, set CI = 95.0 (or any other value); and click OK. *The output appears in the session window*: 95.0 PERCENT CI (28.84, 34.72). You can repeat for CI's of 99% and 99.9%. Q: Is it **expected** that 95% of all intervals will contain μ ? A: True. Q: Does μ have a 95% chance of being in the interval? A: False (We have not calculated $P(\mu \text{ occurring within CI for any 1 sample})$).

Comparing Two Samples

In the **experiment**, another 10 subjects were given a new drug; their response time was: mean = 25.53; $s = 5.6$. To *compare the control* and the *treatment* samples, subtract the mean of one from that of the other. If the treatment had no effect, then we would **expect** $\mu_c - \mu_t = 0$. Due to **sampling variability** however, it is unlikely that $\bar{x}_c - \bar{x}_t = 0$. Suppose that you were to repeat the experiment a **large** number of times, each time calculating the difference between the *control* mean and the *treatment* mean.

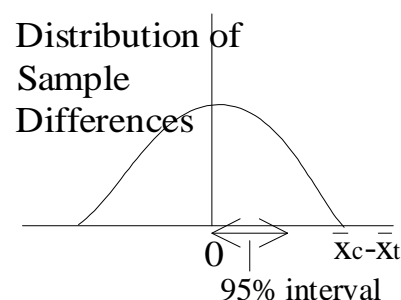


If the set of **differences** between the sample means is normally distributed, you can *calculate the SE* and hence the CI of the distribution: $CI = \bar{x}_t - \bar{x}_c \pm t_{\alpha/2} \times SE_c$. 2 sample t-test: the SE of the difference between the 2 means is calculated as $SE_c = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. We expect that for **95%** of samples, the population mean will lie *within this CI*. If zero lies within the CI, it is probable that the reaction time of the new drug is the **same** as that of the old one.

In Minitab, STATS > BASIC STATISTICS > 2-SAMPLE-T. In the *dialog box*, select the 2 variables, set the CI and click OK. **Output**: 95% C.I. for μ Control - μ Test: (1.6, 10.9). T-test μ Control = μ Test (vs not=): T = 2.85 P = 0.012 DF = 16. The first line is read as “95% CI for μ_1 minus μ_2 (**where** μ_1 and μ_2 are the means of the *control* and *treatment* populations as estimated by x_c and x_t) [CI is calculated as $CI = (\bar{x}_c - \bar{x}_t) \pm t_{0.025}SE$).

Interpreting the Confidence Interval

The **difference** between the *mean response times* to the standard (control) preparation and the new preparation is estimated to be 6.25. If the difference was **close** to 0, this would suggest that both preparations had *similar response* times. As the interval is wholly above 0, then it appears that the new treatment response time is **shorter** than that for the standard treatment. The *width of the CI* indicates the precision of the estimate.



Section 3: Before performing any **analysis**, consider any assumptions you may be making about the data. Consider also how *sensitive* the statistics you may be calculating are to these assumptions. When **calculating** CI's, we assume that the data is normally distributed, but where data sets are *large*, we can relax this assumption.

Experiments: the first thing you should do with any data set is to *explore it graphically* i.e. create a **histogram, dot plot**, etc., and consider the distribution of the data. Why is it important to determine the *shape* of the distribution before calculating the CI — to determine if the data is normally distributed so you can **calculate** SE or CI's. If the data is not normally distributed, if the sample size is *relatively large* (> 30), the normal assumption is not crucial and can be relaxed.

Example 1: pH in Bull Meat

The pH in **home produced meat** has a mean of 5.5 with a normal range of 5.3-5.7. 63 samples of **imported** bull meat were tested. Is there any evidence to suggest that the pH of imported meat is *significantly different* from the pH of home produced meat? **NH:** The mean pH of the sample of imported beef *is not significantly* different from the mean pH of home produced meat. **AH:** The mean pH of the sample of imported beef *is* significantly different from the mean pH of *home produced meat*.

Which **test** shall we use to calculate t , the *test statistic*? A: a 1 sample t-test because in this case, you have 1 sample to compare against a **standard result**, the mean pH of home produced meat. N.B. When doing the test, remember to *specify* the AH. Assumptions about the data: assume the data is normally distributed, include a histogram or other graph to prove this.

In this experiment, $t = -1.07$, $P = 0.29$. Conclusion: **because** $P > 0.05$, we accept the NH that there is no evidence that the mean pH of the sample of imported beef is different from that of *home produced meat*.

Example 2: The Effect of Steroids on Blood Cell Recovery

The effect of **steroid** (+ chemotherapy) treatment on *blood cell recovery* was compared against the effect of a placebo (+ chemotherapy). It was hoped that steroid treatment would enhance the number of blood cells. NH: The mean number of red blood cells is not significantly greater in patients treated with steroids than in patients treated with a *placebo*. AH: The mean number of red blood cells is **not** significantly higher in patients treated with steroids than in patients treated with a placebo. N.B. a test for a *greater than / less than* effect is called a **one-tailed** test.

Which test to use? A 2-sample t-test. In this case you wish to **compare** the means of 2 independent samples. N.B. Remember to *specify the AH*. We assume the data is normally distributed. For the experiment, $T = 6.96$, $P = 0.0000$ (This seems to imply that there is no chance of observing this event if NH is true, i.e. it's impossible. In fact, Minitab only **reports** to 4 d.p., so it means $p < 0.0001$ (write this in *reports*). Conclusion: $P < 0.005$ so **accept** AH.

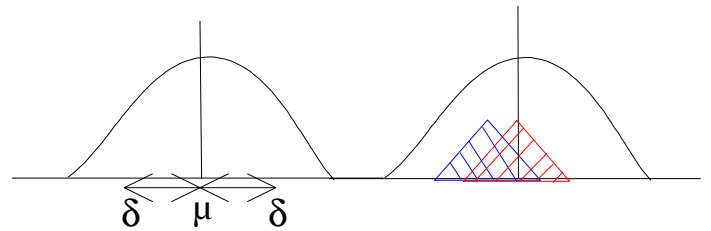
Example 3: Effect of a Drug on the Blood Clotting Time

A drug which was **believed** to hasten blood clotting time was tested by comparing the drug group ($n=64$) with a placebo group ($n=30$). **Method used:** Pooled t-test. When the **variances** of the two samples are approximately *equal*, we can use the pooled method of calculating the SE of the differences between them. The advantage of using this method is that it is statistically more **powerful**. To do a pooled test, in the 2-sample t-test dialog box, check the *Assume equal variances* box.

- Notes:** (i) To **calculate** the pooled SE, first calculate the SUMS OF SQUARES for each data set, $SS_1 = \sum(x - \bar{x}_1)^2$ and $SS_2 = \sum(x - \bar{x}_2)^2$. **Pooled** variance = $s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$. Pooled SE = $s^2 \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$.
- (ii) The *power of a t-test* is a measure of its ability to detect a difference between the means of 2 samples (More in **workbook 8**).

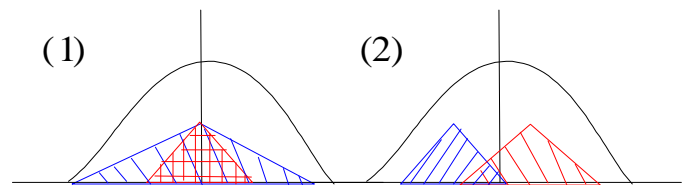
Patients were treated with a drug. After 10 weeks, a random sample were monitored for *side effects*. One measurement: ceratine, average 1.20g/24h. Suppose the curve shown below represents the distribution of the **ceratine** levels. The **Red** and **Blue** dots on the 2nd graph indicate the **sampling distributions** of 2 separate *estimates* of the population mean μ .

Red is **unbiased** (doesn't consistently over/under estimate the population parameter). N.B. most *common sample stats* e.g. mean, variance, proportion are unbiased. However, the sample s.d. is *slightly* biased as its mean is the same as μ . **Blue** is **biased** (the tendency to over/under estimate the population parameter) as its mean is *not the same* as μ .



Both have the same **precision*** as the dispersions of the *sampling distributions* are the same. The SE's are the same. *: the width of a **CI** depends on the SE of the estimate. This is a measure of the *precision* of the estimate. The **precision** can usually be increased by increasing the **sample size**.

In (1), the **Red** and **Blue** are *both unbiased* as their means are the same as μ , but the **blue** is less *precise* as its dispersion is **greater** than the **red** distribution. In (2), the **blue** is more **precise** but is more **biased**. **Red** is less **biased** but is less **precise**. Overall, the **red** distribution is *slightly more accurate*.



Accuracy = Bias + Precision. Accuracy is often *confused* with precision, but is more **general** as it takes bias into account as well. Precision is *solely concerned* with variation.

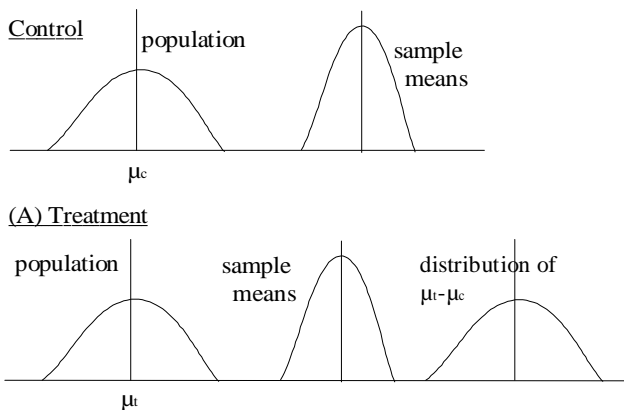
Workbook 8: Significance Testing

Section 1: Null and Alternate Hypothesis. The NH states that the means of 2 samples (or of a *sample* and a *standard*) will be equal. The **t-test** summarises the difference in the means and s.d.'s of two samples and allows us to *calculate the probability* that the NH is true. If this p-value is **less than 5%** (the critical value), we *reject* the NH and *accept* the AH.

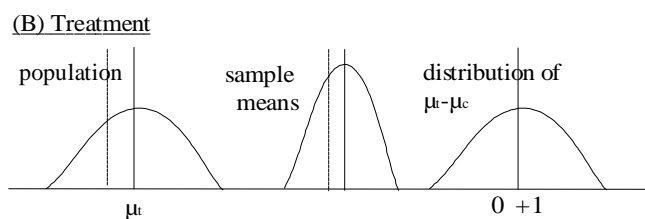
Is **one set of data** different from another e.g. does one treatment withstand microbes *more* than another? Are any differences in data obtained **significant**? Think — there could be 2 conflicting hypothesis about the relative effects of 2 preparations: one person could think there will be *no change*; another could think the new treatment will be *more effective*. An experiment could be done to see which **hypothesis** is true (A **hypothesis** is a *logical proposition* to be tested. A hypothesis shown to be true becomes a **theory**).

We can *generalise* about the hypotheses experiments are **designed** to test. The NH represents the experimenter's belief that there is *no significant difference* between the means of the samples, $\mu_t = \mu_c$. (No significant difference between the *treatment* and *control* means). Before the experiment is done, the values for the **control** and **treatment** means are not known, so we use the μ notation to *indicate* these values.

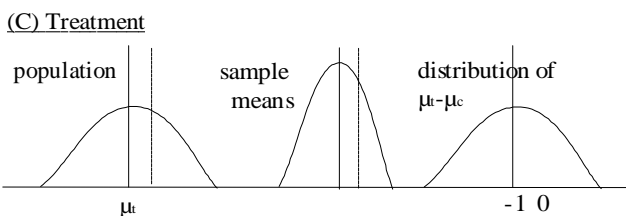
Example. In an experiment, NH = "when a **new** preparation is compared with the old preparation, the mean diameter of the *zone of resistance* will not be significantly different". AH is the hypothesis the experimenter will **accept** if the evidence causes the *NH to be rejected*. For this experiment, AH is $\mu_t > \mu_c$. We are only **interested** in the treatment if $\mu_t > \mu_c$.



If sample and treatment distributions are the same, $\mu_t - \mu_c = 0$ so we accept the Null Hypothesis.



When $\mu_t > \mu_c$, mean of $\mu_t - \mu_c > 0$, so reject NH.



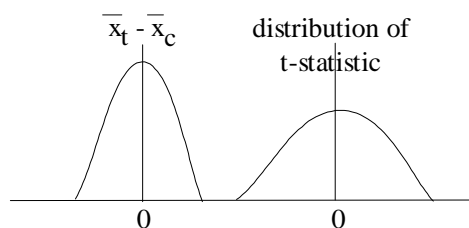
If $\mu_t < \mu_c$, the difference in the sample means will be less than zero and again we reject the Null Hypothesis.

Comparing the Means of 2 Samples — The t-test

To **compare** the means of 2 samples, we need to *calculate a test statistic*. The 2 sample t-test will generate a **suitable** statistic, calculated as $t = \frac{\bar{x}_t - \bar{x}_c}{s \cdot \sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$, with $df = n_t + n_c - 2$.

To do a **2 sample t-test**, STAT > BASIC STATISTICS > 2 SAMPLE T TEST. Select the data and the *appropriate* AH. Note: if variances are similar, use “Assume equal variances”.

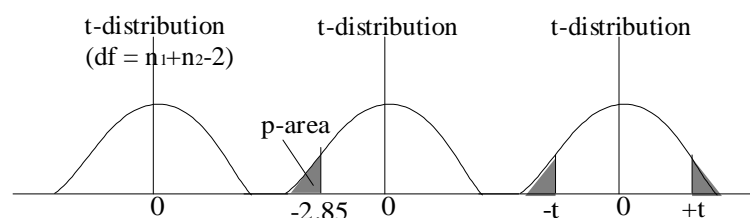
What can the t-statistic tell us?



Consider the **distribution** of $\bar{x}_t - \bar{x}_c$. We assume that it will be *normal* and that if the NH is true, then the mean of the sample means will be zero — but we *don't know SE*. However, we do know exactly what the distribution of t is when $df = n_t - n_c - 2$ and when the NH is true: $t = \frac{\bar{x}_t - \bar{x}_c}{s \cdot \sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$. Using these

probabilities (known as *p-values*), evidence against the null hypothesis can be assessed.

The Significance Level



When the **NH** is true, the mean value of t is zero. The t -statistic calculated from the 2-sample t -test was $t = -2.85$. The p -value is the area of the *shaded portion* of the tail. The p -value is the probability of obtaining a more extreme value of t , assuming the NH is

true. If the NH is true, then t -statistics at the **extremes** of either tail are unlikely to be observed i.e. small p -values. If the p -value for t is small, then an *unlikely event* has been observed; this suggests that the NH is not true. If the p -value is **large**, then the observed test statistic is not unlikely, and there is *little evidence* against the NH.

The 5% Significance Level

Small p -values provide evidence against the NH. “Small” is set at an *arbitrary level* known as the Significance Level (N.B. it corresponds to $P(\text{Rejecting NH when true})$. Ideally this should be as **small** as possible and 5% is the accepted level). Based on the p -value of the t -statistic we have calculated, we should *accept the* NH at the 5% level. The p -value of 0.99 or 99% indicates that there was only a **1%** chance that the new preparation was more effective than the old, therefore we **accept** the NH.

Section 2: There are *5 stages in performing a significance test*: (1) **State** the NH or AH; (2) Select the **appropriate test** with due consideration for the assumptions made about the data; (3) **Calculate** the test statistic; (4) **State** the p -value; (5) **Conclude** whether or not to *accept* the NH.

To **summarise**, when testing a *single mean* against a standard result, use a 1-sample t -test (assuming that the data is normally distributed). When testing **equality** of two independent samples, (see next page for definition of **independent**) use a 2-sample t -test. Check variances for equality, but if unsure use the *unpooled method*.

Independence: e.g. if 2 drugs are to be compared by testing one *on one set of subjects*, and the other on a *different* set, then the samples are **independent**. If, however, the same set of subjects try both drugs, then the samples are **not** independent and you should use a paired t-test. (Paired t-test: for each subject, *subtract the response to treatment A from the response to treatment B*. If the NH is true and the population is **normally** distributed, then we *expect* μ to be zero. Use a 1-sample t-test to compare the mean of the differences between the two treatments against $\mu = 0$).

Section 3: With any **significance** test, there is the possibility of *rejecting the NH when it is true*. This is controlled by setting the significance level to a **small** level (5%). If you try to make the significance level smaller, then you *increase the risk* of accepting the NH when it is false, so the significance level should not be made **arbitrarily** small. The power of a statistical test is defined as the *probability* of rejecting the NH if the NH is false.

Errors: What's the probability of getting it wrong?

You may have **noticed** that we stress that p-values provide *evidence* for accepting or rejecting the NH, not proving or disproving. There is always the possibility that we will either **accept** the NH when it is false, or *reject it when it is true*. Think like this: in law, a defendant is **innocent** until proven guilty. At the start of the trial, there are 4 *outcomes* as shown in the table. Suppose the defendant is innocent. If the jury decides “**guilty**” then this is a mistake. It is also a mistake if the jury decides “*innocent*” when guilty. In an experiment, $\mu_1 = \mu_2$ is analogous to the presumption of innocence in a trial. We want to avoid mistakenly accepting it when it is **false** or rejecting it when it is **true**.

verdict	didn't do it	did do it
not guilty	✓	✗
guilty	✗	✓

What is the probability of rejecting the NH when the NH is true?

Even if the NH is true, it *may be possible to get* t-values so extreme that we might mistakenly reject the NH. How **probable** is this event? We can use Minitab to investigate the probability of making this error. In a *new worksheet*, choose CALC > RANDOM DATA > T DISTRIBUTION. In the dialog box, specify the generation of **1000** randomly chosen t-values, with 10 degrees of freedom, stored in C1. Create a *histogram* of these simulated t-values in C1.

Think! Each value in C1 is a *simulated t-statistic* summarising differences in the means and standard deviations between 2 samples. Look at the histogram — the mean value of t is 0. This indicates that the NH is true; the **differences** between the populations is zero. Because of the variability of random samples, *some t-statistics* will have P-values < 0.05; NH will be rejected.

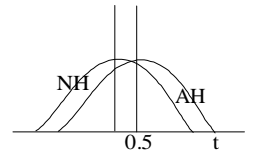
Calculating P-values

If you **calculate** the P-values associated with each t statistic in C1, you can count how many have values *less than* 0.05 and thus are likely to be rejected. For each randomly generated value of t, we can **calculate** P(NH is true). For example, when $df = 10$, P(getting a t-value of 1.82) is approximately 5%. This *implies a 5% probability* that the t-value belongs to a t distribution whose mean is zero.

CALC > PROBABILITY DISTRIBUTIONS > T DISTRIBUTIONS. In the dialog box, check the **cumulative probability option**, 10 df, input C1, output C2. N.B. what Minitab calculates is not P, but (1-P), so you need to know *how many values are greater than* 0.95. The easiest way to do this is to create a histogram of the data in C2. N.B. in the histogram dialog box, press **options**, choose the percent graph type, and set interval size to 0.1 (see help). Look at the histogram (P values are on the x-axis). How many **observations** lie within the interval 0.95 - 1.05? Calculate this as a percentage — 5%.

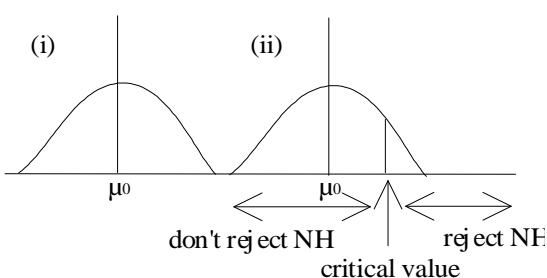
Classifying Errors

Rejecting the NH when it is true is known as a TYPE 1 error; it's the **worst** error — finding an *innocent person guilty*. Accepting the NH when false is known as a TYPE 2 error, and is a safety decision when **not enough** data is known. What is the probability of accepting the NH when false? We first need a *set of statistics* from the AH distribution i.e. when the mean of the distribution is not zero. Then calculate P(getting these t-values if the NH is true). Only t-values which have less than 5% chance of coming from the NH distribution will cause the NH to be **rejected**. To do this, *generate 1000 random t-values*; add 0.5 to all the values e.g. LET C4 = C3 + 0.5. Make a **histogram** of these t-statistics — note that the mean of the distribution is around 0.5. Then use the *probability distribution* command as before to calculate the P-values for C4, putting the result in C5. Look at **C5's** histogram. How many **observations** lie in the interval 0.95-1.05? What percentage of the t-statistics will be rejected — about 10%. Although the AH is true, only 10% of the t statistics indicate this. In 90% of cases, the t-test will be unable to detect a small difference.

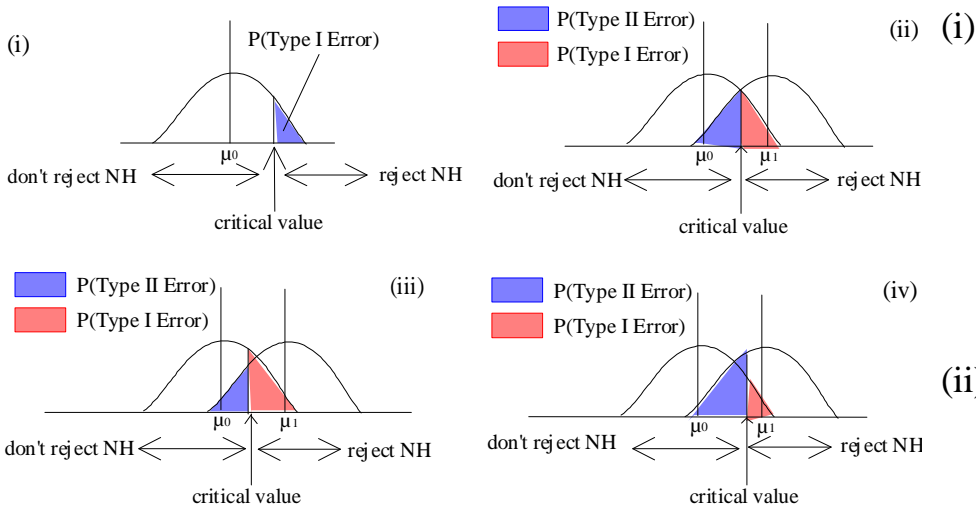


Power

The **Power** of a statistical test is defined as the probability of *rejecting the NH when it is false*. We can increase the **power** of a test by making the significance level smaller, thus reducing the risk of making a Type I error (But **increasing** the probability of making a Type II error), so the significance level cannot be *arbitrarily small*: we set it to 5%.



When **testing** the NH: $\mu = \mu_0$ against the AH: $\mu > \mu_0$, diagram (i) represents the *distribution of the test statistic* assuming the NH is true. (ii) In the **critical value** approach to significance testing, the decision is reached by *comparing* the calculated value of the test statistic to the critical value.



(i) A **decision error** is made if the NH is rejected when it is true. This is known as a Type I Error. The sig. level is the prob. of a Type I Error. Usually it is set at 5%.

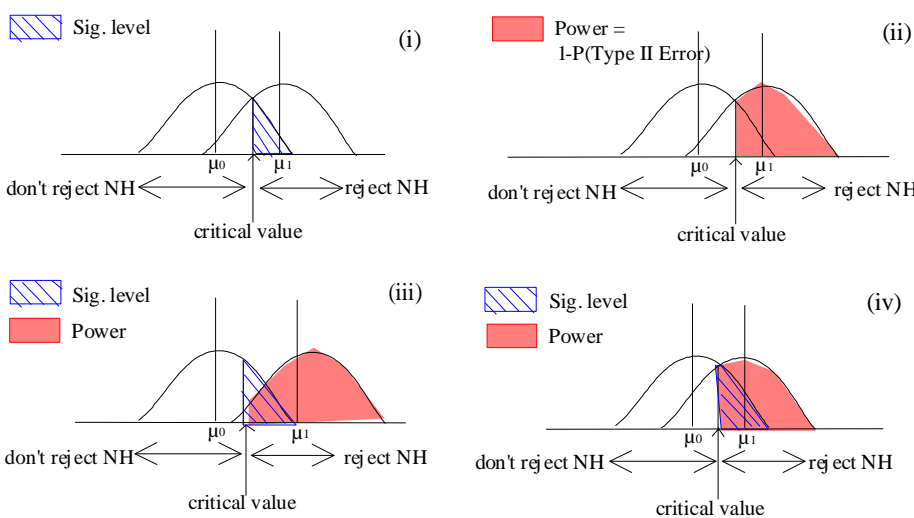
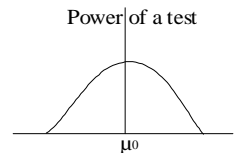
(ii) If the **true** value of the mean is actually μ , then the *test statistic* will have the 2nd distribution. There

is then the possibility of a **Type II Error** — not *rejecting* the NH when it is false.

- (iii) If the *sig. level* if increased, then $P(\text{Type I Error})$ also increases, and in this one tailed test, the critical value moves to the **left**. However, $P(\text{Type II Error})$ *decreases*.
- (iv) If the **significance level** is reduced, then $P(\text{Type I Error})$ also reduces, and in this one-tailed test, the critical value moves to the right. However, $P(\text{Type II Error})$ **increases**.

The **significance level** is the probability of *rejecting the NH* when it is true, By reducing the sig. level, the prob. of a Type II Error **increases**. Or, if the prob. of rejecting the NH when true is set too small, the consequence is that the prob. of not rejecting the NH when false is too high. Making **Type I errors** less likely means that **Type II Errors** become *more* likely, and vice-versa.

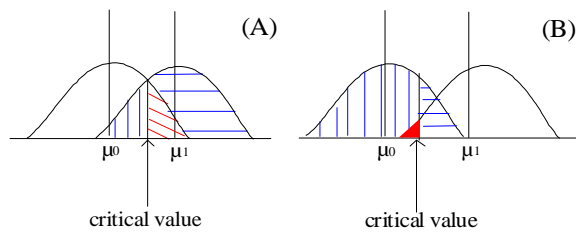
When **testing** the NH, $\text{NH}: \mu = \mu_0$ against the AH, $\text{AH}: \mu > \mu_0$, the diagram represents the *distribution* of the test statistic assuming the NH is true.



In (i), we show that if the *true value* of the mean is μ_1 , the test statistic will have the **2nd** distribution. In (ii), we show that the power of a test is the *probability* of rejecting the Null Hypothesis when it is false. In (iii) and (iv), we show that as μ_1 changes, so too does the *Power*, but the **significance level** remains the same throughout.

- The **further away** μ_1 is from μ_0 , the **greater** the power of the test, and the smaller the probability of *making a Type II Error*.
- The closer μ_1 is to μ_0 , the greater the **probability** of making a Type II Error, and the *smaller the Power*.

- For a **fixed** value of μ_1 the Power of a test can be increased by *increasing* the sample size.



Test A: **NH:** $\mu = \mu_0$; **AH:** $\mu > \mu_0$. μ_1 is a *possible value* of μ if the NH is not true. Q: Which area is P(reject NH when true). A: **Red strips** (this is P(Type I Error), *usually set at 5%*). Q: Which area is P(not rejecting NH when false). A: **Blue vertical strips** (this is also P(Type II Error). It can be *quite large*, and in

general, if the significance level is decreased, then the probability of not rejecting the NH when it is not true is **increased**).

Test B: **NH:** $\mu = \mu_0$; **AH:** $\mu < \mu_0$. μ_1 is a possible value of μ if *NH is not true*. Q: Which is the area representing the **significance** level? A: **Red** area, the P(Type I Error), rejecting NH when true. It is always calculated *assuming that the NH is true*. Q: Which area represents the **power** of the test? A: **Blue vertical strips** — this area represents the probability of *rejecting the NH when false*, and so is the power of the test. It is always calculated assuming that the *NH is not true*.

Workbook 9: Regression

Summary: **Exploring** relationships using scatter plots; Understand *elements* of the linear model. Use Minitab to calculate coefficients of a linear regression model and the confidence interval for *predicted* values.

Relationships between dependent variables

Independent variable plotted on x-axis (manipulated by *experimenter*); dependent variable plotted on y-axis (measured *variable*). Relationships. **Linear** suggests a straight line e.g. $y = a + bx$; *Exponential* suggests an *exponential* curve e.g. $y = e^x$. Draw a best-fit line. Note: the fit is defined as the **square** of the vertical distances between the points and the line (these lines are called *residuals*). The best fit minimises this quantity.

Straight lines: $y = a + bx$, where $a = \mathbf{y\text{-intercept}}$, and $b = \mathbf{gradient}$, $\Delta y / \Delta x$. An equation is a *model of the data* in that it describes important features of a relationship: shape (linear), the degree to which changes in one are related to changes in the other (gradient); position (intercept).

The Statistical Model of the Data

$$Y = a + bx + \epsilon.$$

Analysis: $a + bx$ is the *systematic part* (prediction), while ϵ is the *random part* (errors not included in the model). x is the *predictor/independent* variable; Y is the dependent variable/response; a is the *intercept* (predicted value when $x=0$); b is the gradient, while ϵ is the random error, the residuals. Generally, it's Response Data = Model + Residuals.

Estimating the least squares model

The **estimate** of b is denoted as $\hat{b} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$. The estimate of a is $\hat{a} = \bar{y} - \hat{b}\bar{x}$. The estimate of e is denoted by $e = y - \hat{y}$. e is known as the **residual**; and in its formula, y is the obtained y value at any value of x , and \hat{y} is the fitted y at any value of x . Therefore, the **estimated** model is denoted as $Y = \hat{a} + \hat{b}x + e$.

Minitab: plot scatter plots; estimate a , $b = y/x$. To calculate the model, use the regress *command* (STAT > REGRESSION > REGRESSION, selection, OK). **Regression Output**:

```
MTB> regress 'absrbnc' 1 'protein'           (1 indicates that we have 1 predictor variable)

Predictor   Coef      Stdev      t-ratio      P
constant    0.001579   0.008811    0.18         0.864
protein     0.98421   0.006979   14.10        0.000

S = 0.01521          R-sq = 97.1%      R-sq(adj) = 96.6%
```

The regression equation is the *predicted model*. \hat{a} and \hat{b} are the coefficients of the equation (Coef). Remember these are estimates and so will have **standard deviations**. The t-tests test the hypothesis that each coefficient = 0 and p is the **two sided** p-value.

How good a model of the data is the regression line?

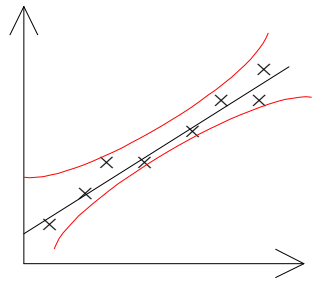
Variation is measured as the sums of squares (SS). SS explained by the fitted model = **Total SS - Residual SS** ($Y = \hat{a} + \hat{b}x$). Explanation: **Total SS** is the sum of squares of all response values $\sum(y - \bar{y})^2$. Total variance = $\frac{1}{n-1} \sum(y - \bar{y})^2$. The Residual SS is the sums of squares = e^2 . Residual variance = $s^2 = \frac{1}{n-2} \sum e^2$. Division by $(n-2)$ to get an unbiased estimate — 2 parameters are estimated, a and b .

The **percentage** of the total variance of Y explained by the *fitted model* is called “R-squared”. It’s usually taken as an indication of **how well** the model fits the data. Back to the data: S is the *standard deviation of the points* about the line; **R-sq** is “R-squared”, the percentage of the total variation in Y explained by the model. The *analysis of variance* table details how the variance is divided between the regression model and the residuals (see later).

The *regression line* has 2 main uses: (1) **Summarising**, $Y = a+bX$. The most important element is b . (2) **Prediction**. Given X , what is the *predicted value of Y*? Also back prediction: Given Y , what X value(s) produced it? **Regression**: A regression equation is a model of the relationship *between two variables*. Such models are often used to predict what the value of the dependent variable will be given any value of the **independent** variable (predictor). It should be remembered that the *coefficient values* of the model are estimates, and therefore for any predicted y value, the **CI** should be included. Note that beyond the range of the data, CI’s become very large, so *extrapolation leads to unreliable predictions*.

Extrapolation: Can you extend the line?

Consider a value x_3 outside the range (x_1, \dots, x_2) . It is tempting to extend the line, but is the *model* still valid beyond the predictor range? We need to consider the Confidence Intervals. Calculating confidence intervals: the curved red lines indicate the CI for fitted values of Y i.e. the range where 95% of estimates of Y will fall for each value of X. To calculate the 95% CI, choose **Regression**, specify the variables, select Options, a box marked Prediction Intervals for New Observations — click on the **dependent** variable for the box. The subsequent output lists the CI for fitted values for each value of the predictor variable.



Fit	Stdev.fit	95% C.I.	95% P.I.
0.00158	0.00881	(-2E-02, 0.02314)	(-4E-02, 0.04460)
0.02618	0.00750	(0.00782, 0.04455)	(-2E-02, 0.06770)
0.05079	0.00565	(0.03510, 0.06648)	(0.01039, 0.09199)

Fitted values of the dependent variable.

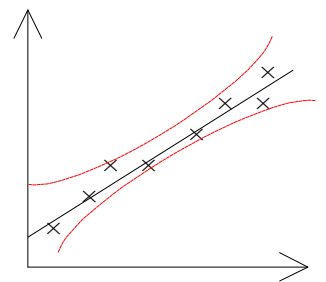
Standard deviation of the column on the left.

The **range** of 95% CI for each fitted value.

95% Prediction intervals for a *single predicted observation* are listed. The PI for a single value is much wider **than for CI** for the fitted mean value on the line.

Plotting the Confidence Intervals

We could enter the **95% CI's as data**; redraw the scatter plot, plotting regression lines and CI's using the *Lines Subcommand*. To save time, use macros (to invoke type % followed by the name). To use, type % O:\TLTP\QUERCUS3\MFILES\PLOTTC1 'absorbnc' 'protein'. The new scatter plot looks as shown. Note how the lines indicating the CI *curve outwards at the ends of the regression line*. What does this **imply**? The CI's are wider at the extremes of the regression line. The narrower the CI, the more precise the estimate of y. Predictions based on extrapolations of the regression line will therefore be much *less precise* than those based in the middle of the x-values.



Workbook 10: Correlation (Exploring Relationships, Part 2)

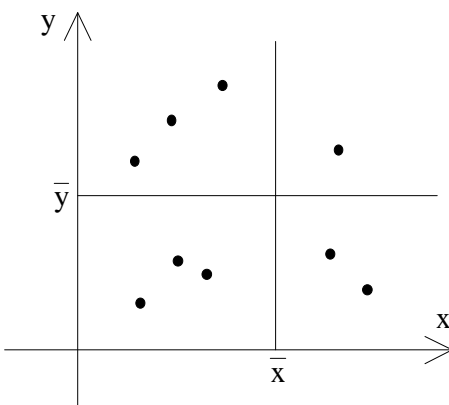
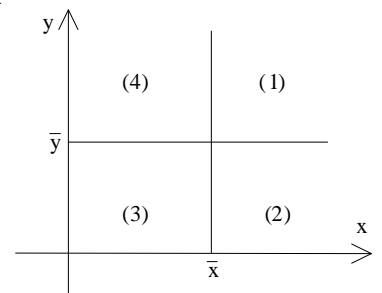
Correlation is a *technical* term that quantifies the tendency of one quantity to **vary** with another. A *low correlation coefficient* i.e. approx. zero, would suggest no relationship between variables, but a non zero value would only indicate that a relationship was possible. If the correlation coefficient is **approx.** -1 or 1, then a statistical association is implied. Correlation stats do not infer a *casual link*. Moreover, correlation measures linear associations specifically. This means a straight line relationship only. We can investigate relationships using **scatter** plots. If they are linear, we are *justified in calculating a correlation coefficient*. We use STAT > BASIC STATISTICS > CORRELATION.

Interpretation

The **correlation coefficient**, r , quantifies the *tendency of the y variable* to change with increasing values of x . (1) The value of r lies between -1 and 1 . (2) If $r = 0$, then there is no correlation. When $r = 1$, points lie on a st. line **with +ve gradient**. When $r = -1$, we have points on a st. line **with -ve gradient**. If $r = 0$, then there is *no linear relationship* — but there may be a curved relationship. Always plot before calculation.

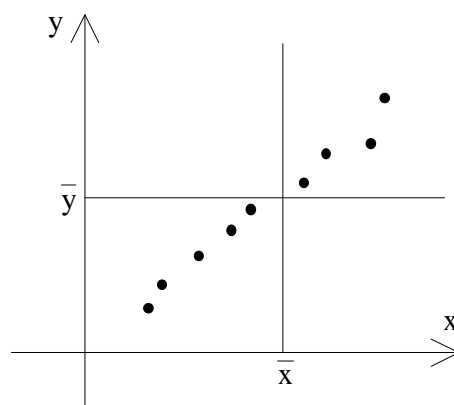
$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$
. **Top line:** Covariance of x and y , with $(n-1)$ df. This measures how x and y vary together. **Bottom left:** Variance of x , with $(n-1)$ df. Measures variability of x . **Bottom right:** similarly for y .

Each point in a scatter plot has an **attribute that depends** on its position relative to the means of x and y . This attribute can be used in *determining linear associations*. On the scatter plot graph, we draw $x = \bar{x}$ and $y = \bar{y}$. For any given point, the sign of $(x - \bar{x})(y - \bar{y})$ is related to which quadrant the point is plotted in. In (1), it's **+ve** as $y > \bar{y}$ and $x > \bar{x}$. In (2), it's **-ve** as $y < \bar{y}$ and $x > \bar{x}$. In (3), it's **+ve** as $y < \bar{y}$ and $x < \bar{x}$. In (4), it's **-ve** as $y > \bar{y}$ and $x < \bar{x}$. Note: the **sign** of an individual point depends on the positions of all other points as well as its own.



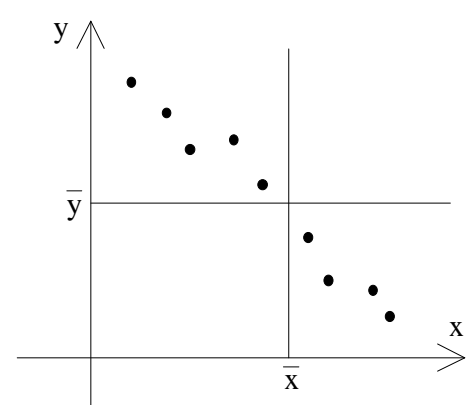
No linear association.

$\sum(x - \bar{x})(y - \bar{y})$ is approx. 0, so r is approx. 0.



+ve linear assoc.

$\sum(x - \bar{x})(y - \bar{y})$ +ve; so r +ve.



-ve linear association.

$\sum(x - \bar{x})(y - \bar{y})$ -ve; so r -ve.

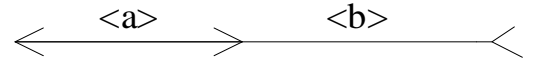
Section 2: How **significant** is r ? We determine this *using the t-test*. Samples are drawn from a population with correlation coefficient p . How confident can we be that the sample correlation coefficient, r , **accurately describes** p .

If there's *no linear association* in the population from which the sample is drawn from, then $p = 0$. If r is **significantly** different from zero, that would indicate a *genuine linear association* in the population.

Hypothesis Testing

Test the **NH**, $H_0: \rho = 0$. The *test statistic* is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, with $df = n-2$. Example: when t is calculated, it is **found that** 4.77 is within the top 2.5%. Based on this, your decision should be “do not reject H_0 at the 5% level” (**This is incorrect:** since t is in the upper 2.5% critical region, there’s *evidence that the correlation is significant*) or “reject H_0 at the 5% sig. level” (**Correct:** there is evidence to suggest a **significant** linear association between the variables).

Example: The Muller-Lyer experiment. Both line segments are the same length, but one appears *longer*. An experimenter wishes to test a hypotheses that there is no relationship between a’s length and your ability to **estimate** it. Measurements made: (*independent*) **length** of fixed arrow; (*dependent*) **length** of adjustable arrow, **time** to complete test.



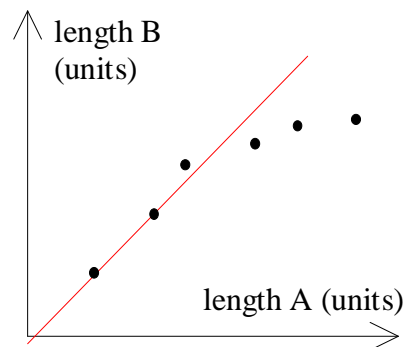
Absolute Difference ($abs(\Delta)$). Another way to look at data is to plot *absolute difference* (‘fixed’ - ‘adjustable’) *against fixed*. When calculating correlation coefficients for the 5 columns of data, ‘print m1’ will cause a **matrix** of correlation coefficients to be printed (The matrix lists corr. coeffs. *for all possible combos*. Note: the matrix is **symmetrical** along the diagonal — each variable has a perfect +ve correlation with itself).

Section 3: Are the correlation coefficients significant?

A correlation may be statistically **significant**, but there may be other factors to be taken into consideration before any *conclusions can be drawn*. The experiment chosen was badly designed. At the end of the discussion section, suggest how the experiment could be changed and improved. (Experiment: to test the **NH** that the Muller-Lyer illusion is unaffected by the length of the **left-hand** segment).

Guidelines for using Correlation Coefficients

(1) **Corr. coeffs.** describe how closely (x, y) data points cluster around a *straight* line. If, however, the data points lie on a curve, or if there are any outliers, you will find that the correlation coefficient gives a **poor** description of the relationship between the variables. For example, did the graph have a *slight curve as shown*? The r value would have been significant, but the important point is that the illusion get stronger as segment A gets longer. **Correlation** implies the effect is constant.



(2) When using correlation to make influences about *relationships between data* i.e. testing H_0 , your conclusions will only be justified if each pair (x,y) have been randomly chosen from a joint population **distribution** of (X,Y) . This implies that each pair (x,y) must be *independent of every other pair*.

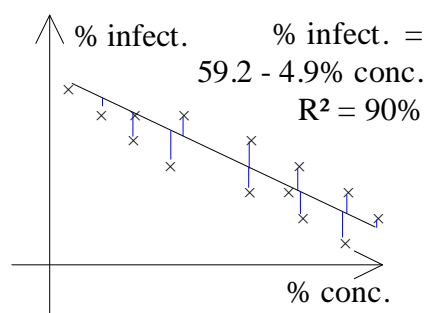
Use correlation **only** when conditions of linearity, random sampling and normality (this can be relaxed when *n is large*) have been met. Interpret correlation with care. Remember, “significant” is only used in a **statistical** sense, and correlation does not imply **causation**.

Workbook 11: Residuals and Transformations

Section 1: Even when the R^2 value is high, a **linear regression** may not always yield the most appropriate model of the relationship between 2 data sets. It is essential to consider the *residuals*. If these are not normally distributed, then the model may not be valid. There are a number of ways in which plots of the residuals and/or the standardised residuals can be used as regression diagnostics.

Analysing the relationship between 2 variables

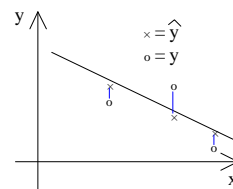
(a) *Plot the data*; (b) calculate a *linear regression model*; (c) consider the *Confidence Intervals*; (d) consider the *residuals*. The general form of the fitted model is as follows: $\hat{y} = \hat{a} + \hat{b}x + e$ ($e = \text{residual} = y - \hat{y}$). Were we to **consider** only the fitted line, we might conclude that % infection decreases with *increased concentration* of the fungicide. The residuals (blue vertical lines) are approximately the same size for % conc. between 1% and 7%. **But for > 7%**, you may not observe a *decrease* in the % infection with P. Infestants on the leaves.



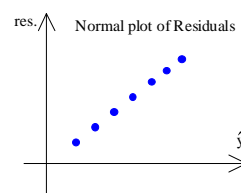
Using residuals to test the validity of a linear model

A regression line is only a *model of the relationship* between 2 variables. As we saw in the previous example, an examination of the residuals can lead us to question the **validity** of such a model. We'll now look at how to use the residual information to test the validity of linear models.

(1) Plot the **standard residuals**. The model is valid if the plot of the standardised residuals looks like a *random sample from a standard normal distribution*. (To calculate this, (a) select the **regression** command, (b) select the data **column** for regression; (c) check the **Standard Resids.** in the dialog box. The residuals will be stored in a column where the regression equation is being calculated. Note: The characteristics of a *standardised normal distribution* are: mean = 0, s.d. = 1, 95% of all observations lie in the range -2 to 2.



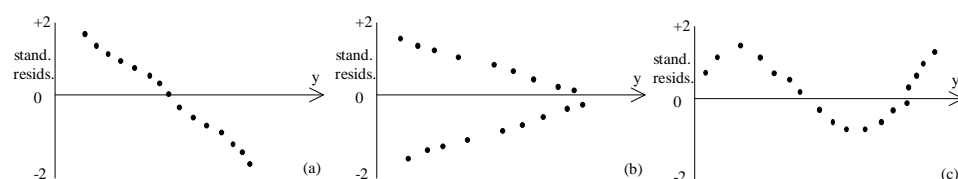
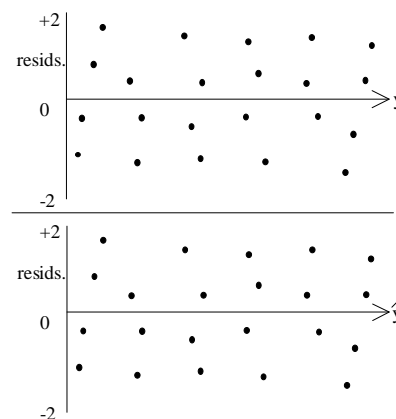
(2) Alternatively, Create a **normal plot** of the data (There is a *macro* to do this: select the Normal Plot command from the **Graph** menu; make a normal plot of the standardised residuals). If the residuals are normally distributed, all the points will *lie in a straight line*. Any non-linearity suggests that the model and any predictions made from it may **not** be valid.



If either the standard residuals or the normal plot of the residuals do not indicate a normal distribution, then the linear model is *not* valid.

Using residuals as regression diagnostics

Residual plots can be used to help us **diagnose** problems with linear regression models. There are two kinds of residual plot you can plot the standard residuals against: either the *observed values* (y) or the *fitted values* (\hat{y}). If the linear model is all right, then these plots should show a *random scatter* with about 95% of the residuals in the range -2 to 2. Any non-random pattern in the distribution of the residuals indicates that the model is **not** valid.



In (a), the graph suggests that the *slope of your regression line is too shallow* — need to increase \hat{b} . In (b), the graph suggests that the magnitude of the residuals **decreases** as the y -values **increase**. This suggests that the model is not a good fit for the data at low values of y . In (c), the graph shows a *definite pattern*: a “wave” shape, and therefore a straight line model is **not** valid. Other examples: a “U” shaped graph: this suggests the data is curved and should **not be modelled** by a straight line; a graph like (b) but reflected horizontally: the increase in the *scatter of the residuals* indicates that the model is not such a good fit to the data at higher values of the response variable. And a graph with an **outlier** will have a major effect on the fit of the model.

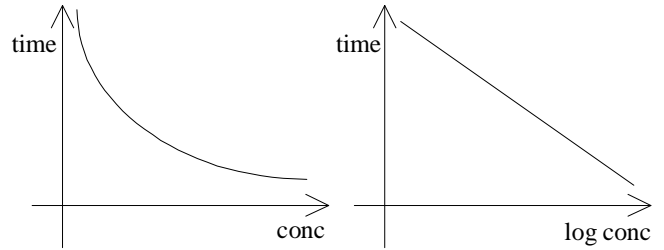
If the residual plots indicate a problem due to (i) *outliers*; go back to the experimental data and check for any recording or **transcription** errors. If you find that the experiment was not correctly conducted at this point, then it is permissible to *omit* the observation. Otherwise you **must** include it in your analysis. (ii) *Due to curves*: Even if there is a curved relationship between the response and the predictor variables, then you may still be able to use a linear model if you can find an appropriate **transformation** for the predictor variable.

Conclusions: (1) After calculating a regression model, use the *standard residuals* to check that the model is valid. (2) **Residual plots** are used to identify problems which can be sometimes be corrected. (3) Always check *outliers and influential points*, and don't forget CI's.

Section 2: Where there is a **non-linear transformation** between 2 variables, we can sometimes transform the x-variable so that a *linear model* can be used. There are a number of functions which can be applied to the x-variable to “*unbend*” non-linear relationships. Which one depends on where the **gradient** of the curve is steepest, and on how *curved* the relationship is.

Non-linear Relationships

Consider the **left** hand graph. Clearly a relationship exists — but it is **not** linear. We could replace “conc” with a related value, but which had a *linear relationship with time* — then we could calculate a linear model. Trying $\log_{10}\text{conc}$, we see that there is a *reasonable linear relationship* between these variables ($R = 99.3$).



Transforming the Predictor Variable

To **unbend** a relationship, you could try using: logs/antilog, raise x to the power n ; or multiplying by some other quantity. This process is known as transformation. When the first part of the curve is **steeper** than the 2nd part, we use a transformation that will cause the x -values to *clump*, such as \sqrt{x} , $\log x$, $-1/x$, $-2/x$, etc. (increasing **strength** as we go along). When the 2nd part of the curve is *steeper than the first part*, we use a transformation which will cause the higher x -values to ‘spread’, such as x^2 , x^3 , antilog x , e^x , etc. (**increasing** strength as we go along).

Choosing a Transformation

Select a **transformation** according to which part of the curve has the steepest gradient, and the strength required to “unbend” the curve. Using Minitab to transform data: CALC > MATHEMATICAL EXPRESSIONS. For example, to find the square root of “time”, type the expression “sqrt(‘time’)” into the Expression box. Note: where there is one bend to the curve, simple transformations as described may be used. But more **complex** curves require more complex transformations (beyond the scope of this course).

Section 3: Modelling the relationship between 2 variables *has 3 states*: (1) **select** a model, which may require transformation; (2) **test** the validity of the model using the regression diagnostic techniques; (3) **interpret** the model. In the example in this section, you must be able to *draw conclusions* about the effects of inter-specific competition.

Two questions: (1) How does **inter-specific competition** affect the dry-weight of each species; (2) Which species is the **better** competitor?

Presenting Regression Analysis

When preparing *reports*, the figures showing the data should be able to stand alone as a complete summary of analysis. If you draw the model onto the scatter plot, clearly label the points as “*actual data*” and the line as “*fitted*”. The title should contain (1) a **number** e.g. figure 1(a); (2) a **brief description** e.g. Model of relationship between...; (3) list of **abbreviations** e.g. bmix (B.Sterillis in mixture); (4) the **equation** of the model and the **R^2 value**.

Residual plots are essential as regression diagnostics to help you establish the **validity** of your model — or to find a better one. If the residual plots indicate that there is a problem with the model, you should specify this in the *results section*. Otherwise, it is not usually necessary to **detail** your examination of residuals, or to present the plots in the final report.

Workbook 12: Analysis of Variance (ANOVA)

Section 1: ANOVA is a technique used to determine *whether the difference* between the effects of three or more treatments are **statistically significant**. It works on the principle of dividing the total variation in some data according to its sources. In a one-way ANOVA, there is variability due to treatment and residual variability (due to random variation between subjects). We use an **F-test** to determine if treatment variability is large compared to *residual variability*.

Suppose we wanted to investigate the effects of **2 different levels** of food additive on the live weights (lwg) of turkeys. We want to compare the treatments with each other and to a control (3 samples). Our resources allow us to use a *total of 15 turkeys*. Allocate the treatments at random to the turkeys and assume that all other factors are constant. Even before measuring the lwgs of the turkeys, we can see that we are going to have to deal with **2 sources of variability** — variability within treatments and variability between treatments.

Variability within each treatment: all of the turkeys are kept under the same conditions, so we attribute this to the natural *variability* among turkeys. Variability between treatments: random allocation of treatments mean that we consider this to be due to *food additive levels*. The difference in nutrients can only be said to be significant if the variability between treatments is **large** relative to the variability within treatments.

Compare the **means** of the treatments. Previously we compared the means of 2 samples using a t-test, $t = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$ i.e. the *ratio of the differences* in the means to the SE. Note: to calculate the **pooled SE**, first calculate the SS for each data set, $SS_n = \sum(x - \bar{x}_n)^2$. The pooled variance is $s^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$. The *pooled SE* is $s^2 \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}$. In this case however, we have 3 samples to compare. We could set up **3 pairwise tests** where $H_01: \mu_1 = \mu_2$; $H_02: \mu_1 = \mu_3$; $H_03: \mu_2 = \mu_3$, but...

- The probability of getting **at least 1 wrong** answer, and thereby interpreting all the results incorrectly, *increases* as the number of **samples** increases;
- Also the number of pairwise tests needed *increases* e.g. if you have 10 samples, you will have to do **45 tests** (The formula for working out the number of combinations is $\frac{n!}{(n-c)!(c)!}$, where n is the *total number of samples* and c is the number *to be drawn* from the total. In this case (drawing 2 from 10), we have $10! / (10-2)!2! = 45$ pairs.

Therefore, to compare the means of *many samples*, we will use ANOVA. Before discussing this, we need **notation** to use. The observed x is the live weight of the turkey. Let's give each observation a code to identify which turkey and which treatment it was from.

Each observation (x) will have a *subscript*, i : the treatment code, where i has a value between 1 and t . t : The number of **treatments**, here $t = 3$. Any observation from e.g. treatment 2, *could* be referred to as x_2 . To identify a **single** observation from a treatment, add a 2nd subscript, j . j : the *observation code*, with value between 1 and m . m : the total number of turkeys in each treatment. In this **case**, $m = 5$. The observation on the 5th turkey could be *referred* to as $x_{2,5}$.

A model for turkey live weight x_{ij}

Turkey lwg has 3 **components**: μ , the mean of the population; t_i , the mean deviation in lwg *due to the effects of treatment*; r_{ij} , the **random** variation among individuals. Therefore $x_{ij} = \mu + t_i + r_{ij}$. If the NH ($\mu_1 = \mu_2 = \mu_3$) is true, then *treatment effect* (t_i) is zero and therefore all the observed **variation** in the experiment is due to random variation among individuals (r_{ij}). It follows that in order to **test** the NH, we need to consider how much of the observed variation in an experiment *comes from simple random variability*, and how much is due to real **differences** between the treatments.

Notes: t_i : t is the amount of *weight that a turkey* may gain or lose due to treatment i , where it is any number from 1 to 3. r_{ij} : there's going to be a certain amount of '**random**' weight gain/loss observed in each turkey. Even if all the other factors are held constant: light, water, space, etc., this variability *will* be observed. It may be due to the *generic make-up* of each turkey which influences size, etc., or to behavioural factors such as competitiveness.

Measuring and Comparing Sources of Variation

Total Variation = Treatment Variation + Random Variation (Residuals). The idea that we can partition observed variation according to sources lies at the heart of the ANOVA technique. We know how to measure variation and so we can compare the relative sizes of variability from each source. To measure variation, use the same Sums of Squares (SS) method described in workbook 5. Random Variation = Total Variation - Treatment Variation.

The ANOVA table — and how to read it

It is customary to lay out the ANOVA table in a **standard** way, which records the amount of variability due to *different sources* and shows how the test statistic is calculated.

Source of Variation	Df	SS	MS	VR
Between Treatments	$t-1$	SS_{between}	MS_{between}	F
Residuals	$N-t$	SS_{residual}	MS_{residual}	
	$N-1$	SS_{total}		

Explanation: $(t-1)$ = df *within the treatment group*; $(N-t)$ = Residual df; $(N-1)$ = Total df [$(N-1) = (t-1) + (N-t)$]. SS_{between} = Sum of *squared variations* between treatments; SS_{residual} = Sum of squared deviations *within treatment* groups; SS_{total} = total Sum of Squares. MS_{between} = Mean **Square** for the treatments. To *calculate Mean Square*, divide the appropriate SS by df. MS_{residual} = MS for the residuals. F: To compare the *sizes of the mean squares*, divide MS_{between} by MS_{residual} . The ratio is called the *F-statistic*.

The F-statistic **is the test** of the NH. If the value of F in the table exceeds the critical F value ($F_{0.05}$), then *reject the NH* that the sample means are equal.

Section 2: The ANOVA technique is only **suitable** if (i) the observations are *independent*; (ii) the data is *normally distributed*; (iii) the **variances** are equal. If these conditions are met, use Minitab to analyse the data. The Minitab output gives the F-stat, P-value and summary stats for the data, **according to treatment**, so you can see how the treatment effects varied from each other.

Example: in a *nutrition experiment*, a total of 29 rats were assigned at random to 4 groups. Each group was fed a different diet for a fixed period of time, and the weights of each individual's liver as a percentage of body weight was **obtained**. To use ANOVA, three assumptions must be satisfied: (1) all the observations are **normally distributed**; (2) the **variances** in each level are equal; (3) all the observations are **independent**.

The 3rd condition depends on *good experiment design*; assume it is correct for the example. We don't need to explicitly measure **normality** or variances to test the first 2 assumptions — it is enough for the moment to plot histograms of the data from the 4 treatments (diets). Is the data **normally** distributed?

If the histogram is not *noticeably skewed*, and there are no outliers, assume approximately normal. For a more **objective** test, use a normal plot (see Workbook 11). The output produces the Anderson-Darling Normality Test Statistic and its P-value. Are the variances roughly equal: when sample sizes are small, it's difficult to make a decision about how much they can differ before we say they aren't **equal**. They can vary by 30% by chance alone. If however one SD is more than *twice the size* of the others, assume unequal variances. F-squared and chi-squared tests will give more **objective** estimates of equality.

Make a **preliminary judgement** while viewing the histograms. Consider the means of the 4 treatments as well. How does the *difference between treatment* means (i.e. variance between diets) compare to variance within diets (i.e. residual variation)? To **produce** an ANOVA table, STAT > ANOVA > ONEWAY (UNSTACKED). Select all columns, click OK. Click on the *session* window to see the results. Output obtained: (i) **Table**; (ii) Table of **descriptive statistics** for treatments; (iii) **Plot** of means and 95% CI's for treatments.

Interpreting the Output

Source	df	SS	MS	F	P
Factor	3	1.1601	0.3867	10.7416	0
Error	25	0.9012	0.036		
Total	28	1.0613			

Factor = treatment; **Error** = Residuals. Divide SS by df to get MS. Divide MS_{factor} by MS_{error} to get the F-statistic. If $P < 0.05$, then F is significant. **Note:** P is not zero as shown — only given to 2 d.p. in Minitab. If the F is *significant*, reject the NH ($\mu_1 = \mu_2 = \mu_3$).

There are 4 farms, **therefore** 4 blocks.

	Vaccine 1	Vaccine 2	Vaccine 3	
Farm 1	91	95	67	Block 1
Farm 2	84	90	56	Block 2
Farm 3	95	106	76	Block 3
Farm 4	99	107	69	Block 4

This design introduces a *new source of variability*: variability between blocks.

The Randomised Block Design Model

Total Variability = Treatment Variability + Variability Between Blocks + Variability Within Blocks.

If we treated the data as if it came from a completely **randomised** design, then the error would be (*Total Variability - Treatment Variability*). But since we can measure the variability between blocks, the error is (*Total Variability - Treatment Variability - Block Variability*) i.e. only the variability **within** blocks contributes to the error term.

ANOVA for Randomised Block Design

Source of Variation	df	SS	MS	F
Block	2	2,328.5	1,164.25	133.91
Treatment	3	385.5	161.86	18.62
Error	6	52.17	8.69	
Total	11	2,866.25		

For the data in the *vaccine example*, the ANOVA is as shown. Note that the F statistic is calculated for Block and Treatment by **dividing** the Error MS. The P-values for the Block And Treatment are both *less than 0.05*. This indicates that there is a significant difference between the vaccines. That the block variation was also significant suggests that we were right to **group** the data according to the farm.

For the Vaccine.MTW file, C1 = response data; C2 = codes for treatment; C3 = codes for farms. Using Minitab: STAT > ANOVA > BALANCED ANOVA. Indicate that the **responses** are in C1, **codes** on C2 and C3. An ANOVA table similar to the one above appears. Note that this doesn't produce *summary statistics* (you can do this using MTB> TABLE C2; SUBC> STATS C1).

Workbook 3: we are told that a **report** should contain: (1) *Introduction*, (2) *Methods e.g completely random or randomised block design*; (3) *Results*; (4) *Discussion*. You should include information about **experimental design** as a subsection in the methods section. In the results section, mention which **kind** of ANOVA you have used.

More About ANOVA

Once you've understood the *basic principles* underlying ANOVA, you can extend this technique to more **complex** experimental designs e.g. we might wish to measure turkey live weight at different stages in their development. This adds another factor to the analysis (time), so we could use a two-way ANOVA (see below) to **analyse** the data. 2 and 3 factor experiments are not *uncommon*, but beyond the scope of this course.

Two Way ANOVA

As well as measuring the effects of 2 **different factors** on a response variable, 2 way ANOVA's can measure the *interaction* between 2 factors. You will find a two-way command in the ANOVA sub menu. You could use this command to reanalyse the vaccine data (provided you check the additive model box in the dialog box), but because you only have one observation for each farm and vaccine, it won't be able to **calculate** the interaction effect between farms and vaccines.

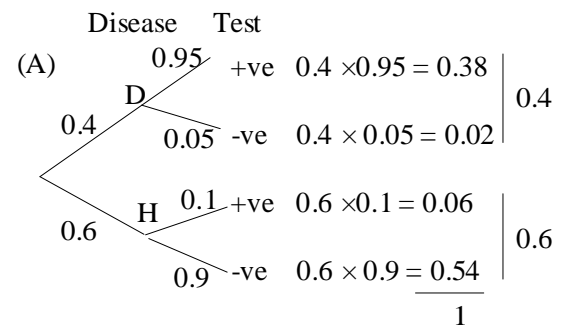
Lecture Notes

Probability

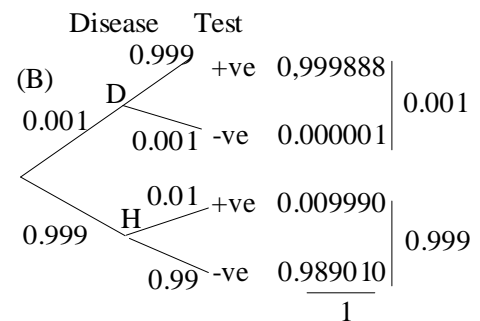
For a **coin**, $P(\text{Heads}) = 0.5$. In an *experiment*, $P(\text{Heads}) = (\text{no. of heads})/(\text{no. of tosses})$. This probability levels out towards 0.5 as more tosses are made. Bayesian probability: measure of **belief**. Tree diagrams can be used to analyse the probability of e.g. getting two consecutive heads. In a *tree diagram*, all possible outcomes are shown, but it can become very complicated when e.g. showing 8 tosses of the coin. For **2 tosses**, the probability distribution for the number of heads is as follows: 2 Heads: Probability = 0.25. 1 Head = 0.5; 0 Heads = 0.25.

In Minitab, Call *constants* μ . "To the power": use two stars, **. It uses **scientific** notation e.g. 1.024E-007.

Consider the following medical situation: **screening** for a disease. In the first test, 95% of tests are positive if the person has the disease; and 90% of tests are negative if the person does not have the disease. In the population, 40% *have the disease*. Using tree diagram (A), we see that the proportion with the disease among those with a +ve test result is $\frac{0.38}{0.38+0.06} = \frac{0.38}{0.44} = 0.86$. The **probability** of a disease given a +ve test result is $\frac{0.38}{0.38+0.06} = 0.86$.

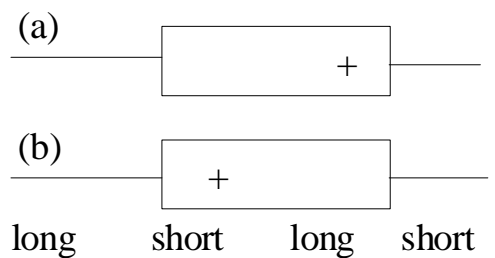


In **situation (B)**, $\frac{1}{1000}$ people have the disease. In the test, 0.999 of people *with the disease* give a +ve result; 0.99 of people without the disease give a -ve result. For those with a +ve result, what is the *probability/proportion* of having the disease? A: $= \frac{0.000999}{0.000999+0.009990} = 0.09090\dots = \text{only } 9\%$.



The above is an **example** of conditional probability: probability of a +ve test **GIVEN** the disease. $P(+/D)$ is not the same as $P(D/+)$ (the probability of having the disease given a +ve test result).

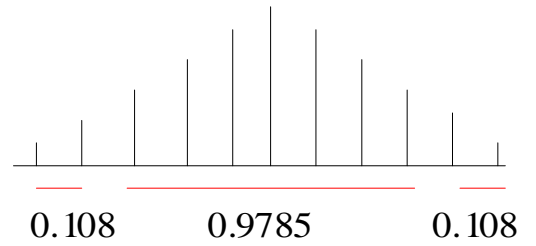
Directory for Data: O:\STATDATA\HAMPTON (or DATA). A *dot plot* is ideal for small samples. A stem and leaf plot is ideal for **30-50** samples; a histogram ideal for large samples. In a boxplot, * = near outlier; O = far outlier. The "hinges" are the quartiles. If the box plot is *symmetric*, the median is in the middle of the box. If **skewed**, the median is at the end with the shorter line, as in (a). If there is no consistency as in (b), take it as **symmetric**.



In the **normal** distribution, the *mean, median and mode* occur in the middle.

An Example

Germination % of seeds is 50%. Therefore, the **probability** of a seed germinating is 0.5. If we plant 10 seeds, how many germinate? Assumptions: (i) *Fixed number of seeds (10)*; (ii) *germinate or not*; (iii) *probability is the same*; (iv) *seeds are independent*. In this example, use the Binomial Distribution. If $p = 0.5$, then with probability 0.9785, you will observe between 2 and 8 seeds germinating. If $p = 0.5$, then with probability 0.216 you will observe 1 or less or 9 or more seeds germinating.



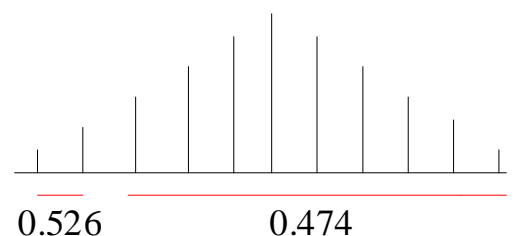
Now consider a more **realistic** situation. Plant 10 seeds, record *how many germinate*. On the basis of the number germinating, are we willing to **decide** that the seeds are representative of a batch where $p = 0.5$? **If** $p = 0.5$, we know that with a probability of 0.9785, we will get between 2 and 8 seeds germinating. If $p = 0.5$, then with probability 0.216 we'll get *between 0 and 1 or between 9 and 10*. If between 2 and 8 seeds germinate, then the assumption $p = 0.5$ is **acceptable**. If 1 or less or 9 or more germinate, then the assumption is not valid.

Hypothesis Test

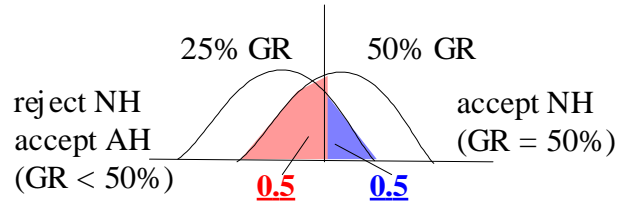
NH: the probability that a seed **germinates** is 0.5 ($p = 0.5$). AH: $p \neq 0.5$. *Perform the experiment*. Get data: n out of m germinate. (A) 4 out of 10 germinate. As 4 is between 2 and 8, then the probability of **getting** the data (4) assuming NH is true is large (0.9785), so we accept the NH. (B) 0 out of 10 germinate. The probability of getting the data assuming that the NH is true is small (0.0216), so we reject the NH and accept the AH. *Convention*: if $p < 0.05$, then reject the NH and accept the AH (at the 5% significance level).

The **significance level** of a test is the probability of rejecting the NH when it is true. It is the probability of *making an error*, and often taken to be 0.05 or 5%. For this test, the values 2,...,8 are collectively called the acceptance region (NH accepted). The **values** 0,1,9,10 are collectively called the rejection (or critical) region — NH is rejected. In a 2 tail test, NH: $p = 0.5$; AH: $p \neq 0.5$. In a 1 tail test, NH: $p = 0.5$; AH: $p < 0.5$.

The **other possible error** is accepting the NH when it is false. If the *germination rate* is as low as 25%, then we want to be able to reject the NH that the germination rate is 50%. Here there is a large probability (0.474) of **accepting** the NH (G.R. = 50%) even if the germination rate is as low as 25%. With $n=20$ and $p=0.5$, $P(\text{between } 0 \text{ and } 6 \text{ germinate})$ is 0.061. With $n=20$ and $p=0.25$, $P(\text{between } 0 \text{ and } 6 \text{ germinate})$ is 0.786. Here there is *still a large probability* (0.214) of accepting the NH (G.R. = 50%) when in **fact** it is 25%. BUT, with $n=38$, there is a small probability (0.38) of getting 0 to 14 if $p=0.5$ AND there is only a **small** probability (0.071) of getting 15 to 38 if $p=0.25$.



0.5 is the *significance level*, the probability of rejecting the NH when the NH is correct. **0.5** (1-Power of the test) is the probability of *accepting the NH when the NH is false*. Tests are often designed to have a significance level of around 0.5; with the **power** of a test around 0.80. (Therefore the probability of making the mistake of accepting the NH when it is false is $1 - 0.80 = 0.20 = 20\%$).



NH: sample is from a *normally distributed population* e.g. where the pulse rate is normally distributed. AH: sample is from a population where the pulse rate is **not** normally distributed. Normality Test: $A^2 = 2.395$; P-value = 0.000. [close to zero = close to being normal]. The probability of *getting this data under the NH is small* (If $p < 0.5$ then reject the NH). If $p > 0.5$, then accept the NH (at the 5% significance level).

2 sample t-test: 2 methods: (1) **Samples** and **Subscripts**; (2) By **columns** e.g. Male PR, Female PR. μ is the population mean. Notes: *Sex: 1 = male; 2 = female*. 1-power = prob. of accepting the NH when it is wrong. Hypothesis: **mean** = 80; want to see if a difference of 1.19 is significant. Power of the test taken to be 80%. (If the mean pulse rate is 80 and we get a sample mean of 78.81, then we *need a sample size* of 1014 to get 80% power).

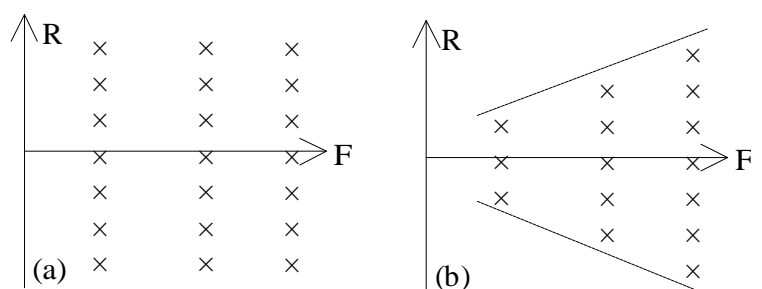
If you have to **estimate** standard deviation, think of a possible *range*, and then $s.d. = \text{range} / 4$ or 5 or 6 (Roughly) e.g. $120-60/4 = 15$. The correlation coefficient is **between -1 and 1**. NH: No linear relationship between 2 variables in the *population*; AH: There is a... Do not just look at the correlation coefficients — use the graphs. A **fitted line plot** plots a “best fit” on the graph.

Examples 1998/1999

(1) 30 **similar mice** were tested for the effect of diet on weight change (3 diets: Control, Junk, Health). Assign *10 at random to each diet*. Analysis of data: use One Way ANOVA. Total $df = 29$, therefore 30 observations. Group $df = 2$, therefore 3 groups. Error MS (Mean Square) [the mean of the 3 diet variances] = 0.789. $F = 41.81 = \text{groupMS} / \text{errorMS} =$ “how *different* the 3 diet means are”.

If there really are **no differences** in mean weight change due to the 3 diets (i.e. NH is true), then F is approximately 1. If there *are* differences in weight change due to the 3 diets, then F is significantly more than 1, $F \gg 1$.

Carry out a **normal plot** of the residuals, Residuals vs. Fits. For a one way ANOVA, fitted value = mean weight change for the animals that *received the same diet*. The residual is the observed **weight** change — fitted. We want graph (a) showing roughly equal spread (equal variability), but a **common problem** is graph (b).



In a *normal plot of the residuals*, if the residuals are normal, then we expect a straight line. If they are not normal, we expect e.g. a **curve**. Don't be misled by *extreme values* i.e. a reasonably straight line in the middle.

Tukey's *Pairwise Comparison Test*: Diet 1 - Diet 2; $10.78 - 12.21 = -1.93$. The 95% CI for the difference *between diet 1 and diet 2* is -2.9159 to -0.9441. 0 is **not** in this interval, so the difference is significantly different from 0; *diet 1 and diet 2 are significantly* different from each other.

Descriptive	Test(s)
Mean	t-test: 1 sample
s.d. / variance	t-test: 2 sample / 2 independent samples
CI	t-test: paired / matched
Pictures / graphs / plots	ANOVA
	Regression

A lot of **tests** assume that data is *normally distributed*. Tip: use the graphs option in the descriptive statistics dialog box.

[Material Gained using Quercus and Minitab](#)

Exercise Book Output

Exercise Book 1

Question 1

Correct. You correctly identified the variable 'reaction to ointment' as an ordinal variable.

Question 2

- Correct. Number of peas in a pod is a discrete variable.
- Correct. nMoles mg^{-1} of ADP in a sample of plant tissue is a continuous variable.
- Correct. The % area of a bacterial plaque in a petri dish is a continuous variable.
- Correct. Units of alcohol per week drunk by students is a discrete variable.

Question 3

Correct. The top-right diagram is the odd one out, because the variable on the horizontal axis is qualitative. The variables on all the other axes are quantitative.

Question 4

- Correct. A nominal variable can have more than two values.
- Correct. Ranked data, or that an order is implied, is a characteristic of ordinal data.
- Correct. Discrete variables may have fractional values.
- Correct. An observation of a continuous variable may be expressed to the nearest whole number.

Question 5

Correct. Graded response to a drug treatment is a qualitative variable, the others are quantitative.

Question 6

- Correct. A data set is a collection of observations of a variable.
- Correct. A variable may have many states or values.

- Correct. Qualitative variables consist of a number of categories which describe the observations.
Correct. Whether a variable is qualitative or quantitative is less important than whether it is appropriate and informative in the context of your study or experiment.

Consider your test results. At the beginning of WorkBook 1, 'Learning about Variables', it was stated that at on the completion of the WorkBook, you should be able to demonstrate how variables are classified. Are you satisfied that you have achieved this objective? If not, we suggest that you (i) revise WorkBook 1, (ii) read the appropriate chapters in the recommended text book, (iii) discuss your difficulties with fellow students and/or your tutor.

Exercise Book 2

Question 1

- Correct. It is the 'History', not the 'Info' window that lists all the commands used in the session.

Question 2

- Correct. The rows are numbered but they are not labelled in the same way the columns are.
Correct. This key combination will make the data window active.
Correct. Each column must have a unique name.
Correct. To move to the end of the worksheet, press Ctrl + End.

Question 3

- Correct. The command `sort c1 c2`, sorts the data in c1 and copies the sorted data to c2.

Question 4

- Correct. The command to exit Minitab is Stop, not Quit.
Correct. To make the session window active, type Ctrl + M, not Ctrl + S.
Correct. The command Help will start the online help system.
Correct. The command is not valid because the column name is in double quotes rather than single quotes.

Question 5

- Correct. To indicate that you wish to use a subcommand, type a semicolon at the end of the command, not a colon.

Question 6

- Correct. The ? in the dialog box for any command will open the help file at the page for that command.
Correct. Typing the command Help 'describe' will not access the help information on the describe command. Putting single quotes around the word describe makes it appear to be the name of a column or constant.
Correct. Describe will be listed under Commands in the help file, by clicking on describe you can access the help information on this command.
Correct. Typing describe and pressing F1 will have no effect.

Question 1

- Correct. The general syntax of a pathname is `drive:\directory\subdirectory\filename`.

Question 2

- Correct. Selecting the menu command "Open Worksheet" and correctly completing the dialog boxes will retrieve a saved worksheet.
Correct. The Retrieve command followed by a pathname in single quotes will retrieve a saved worksheet.

Question 3

Correct. The retrieve command is for saved Minitab worksheets, the other commands copy data from ASCII files into the worksheet.

Question 4

Correct. The Save Worksheet command creates a file with a .mtw extension.
Correct. When a filename is not specified, Minitab uses minitab.mtw as a default filename.
Correct. The “Save Worksheet As” command will save the worksheet in the Minitab format regardless of the file extension used. Only the “Write” or “Export as ASCII” commands will create ASCII files.
Correct. The Write command will save the data as text but not the column names.

Question 5

Correct. The commands to Cut, Copy and Paste are listed in the Edit menu, not the Editor menu.

Question 6

Correct. The Read, Set or Insert commands will read data in the ASCII format into the Minitab data window.
Correct. Text and graphs copied from the session window can be edited in the NoteBook window. Graphics copied from Graph windows however cannot be edited in the NoteBook window.
Correct. Minitab Worksheets exported as ASCII can be edited by a text editor.
Correct. Files created by the Save Worksheet commands cannot be edited by a text editor.

Exercise Book 3

Question 1

Correct. Qualitative data can be presented as barcharts, piecharts, pictograms, etc.

Question 2

Correct. Each observation is represented by 1 leaf, therefore the total number of leaves is equal to the total number of observations.
Correct. The total number of leaves in each group is the frequency of the group.
Correct. The stem is to the left of the dividing line and the leaves to the right.
Correct. The objective is to choose a division between stem and leaf which will show the data piling up in the stems.

Question 3

Correct. The lowest class has an interval between 15 and 19, but there is an observation $X = 14$. This observation is therefore not included in the frequency distribution.

Question 4

Correct. The appropriate division between stem and leaf for this data set is $1 | 100$, where the stem = 1000.

Question 5

Correct. The stem should represent 1000; therefore the appropriate leaf unit for this data set is 100.

Question 6

Correct. Each outlier in the data set is indicated by an asterisk.
Correct. The data is not divided into classes — each dot represents one observation.
Correct. The sides of the box indicate the values of the first and third quartiles, not the range.
Correct. Increasing the size of the intervals widens the range of values included within each interval, and therefore decreases the number of bars in the histogram.

Exercise Book 4

Question 1

Correct. With skew data, use the median as it is not influenced by the extreme values in the tails.

Question 4

Correct. This is the best course of action as both distributions are skew and A has an outlier.

Question 5

Correct. Although the distributions are a little skew, it is not so serious as to warrant the use of the medians in both groups.

Question 6

Correct. Range = Maximum - Minimum and so is bigger in B than in A.

Correct. The range depends on the extreme values and so is influenced by outliers.

Correct. Although most of B is spread over a shorter range than A, the outlier means that the range of B is greater than that of A.

Exercise Book 5

Question 3

Correct. The standard deviation is the 'average' distance of each observation from the mean.

Question 4

Correct. The distributions are skew, so this is the correct response.

Question 5

Correct. The distributions are reasonably symmetric, so use the standard deviation.

Question 6

Correct. The distributions are reasonably symmetric, so use the standard deviation.

Exercise Book 6

Question 1

Correct. There will always be sampling variability whenever samples selected as different samples yield different data observations, and so different summary statistics.

Question 2

Correct. The elimination of bias and the aim of getting a representative sample are strong reasons for using random samples.

Correct. Only when using these statistics to provide population estimates are random samples required.

Correct. The precision of the sample mean increases as the sample size increases.

Correct. This is the Central Limit Theorem which plays such an important role in statistical inference.

Question 3

Correct. The standard error of the sample mean measures variability in the sample mean from one sample to another.

Question 4

Correct. As the distribution is skew, a large sample is required.

Question 5

Correct. This is $9/3$ i.e. σ/\sqrt{n} .

Question 6

Correct. Precision refers to the spread of the sampling distribution; the standard error is a measure of precision. The smaller the standard error, the more precise the estimate.

Exercise Book 7

Question 1

Correct. All you know is that in repeated sampling experiments, 90% of such intervals are expected to contain the population mean.

Question 2

Correct. As the sample size increases, the width decreases.
Correct. The sample mean has nothing to do with the width.
Correct. In larger samples, the t coefficient gets smaller.
Correct. The smaller the standard deviation, the smaller the confidence interval.

Question 3

Correct. Location and Precision are the two important bits of information provided by confidence intervals.

Question 4

Correct. This is a proper interpretation of a confidence interval.

Question 5

Correct. The narrower the interval, the more precise the estimate for a fixed confidence.

Question 6

Correct. One Variance is 4 times the other, so this is the correct choice.

Exercise Book 8

Question 1

Correct. Using a pooled test is correct as the variation was the same in the two groups. Also a two tailed test was required as you were testing for a difference.

Question 2

Correct. The p-value is greater than the significance level, so there is no evidence to support a difference in means.

Question 3

Correct. A Type I Error corresponds to rejecting the null hypothesis when it is true, leading to a small p-value.
Correct. Values of t close to zero are consistent with the null hypothesis. Large positive or negative values of t are consistent with the alternative hypothesis.
Correct. As the p-value is the probability of getting a more extreme t value than that calculated, these two quantities are directly related.

Correct. There is no such thing as a statistical proof. Statistical tests provide evidence against the null hypothesis. In any test, there are two types of errors, either of which may occur, so there is never any statistical proof.

Question 4

Correct. Always quote the p-value, and refer to the direction of the test as specified in the alternative hypothesis.

Question 5

Correct. Take the inverse transformation of the CI for the differences in mean log(MPN) values. Differences in log(means) correspond to ratios of means on the MPN scale.

Question 6

Correct. The alternative which reflects values consistent with a hypothesis is μ not equal to the specified value, as high and low values are associated with inconsistency.

Exercise Book 9

Question 1

Correct. Although there is a great deal of variability, a straight line appears to be a reasonable model.

Question 2

Correct. In this case, the Time is chosen by the experimenter and so it is the independent variable (X). The log Bacterial Count is to be predicted, and so is the response variable (Y).

Question 3

Correct. The intercept is estimated to be 4.60.
Correct. The equation is $\log_{bc} = 4.60 + 0.0207 * \text{days}$.
Correct. 4.60 is the intercept. The slope is 0.0207.
Correct. The regression is the wrong way around as log bacterial Count should be the dependent variable and time the independent one.

Question 4

Correct. This is a relatively wide interval as a single observation is predicted.

Question 5

Correct. As 35 days is well outside the range of observed times, predictions should not really be made as the linear association in this region has not been checked.

Question 6

Correct. $t = 0.0206/0.0035 = 5.83$, and this is significantly different from zero with a p-value less than 0.0001.
Correct. 0.0207 is the estimated change in mean log Bacterial Count per day.
Correct. A confidence interval for the slope is the estimate $\pm t^*(\text{standard error of the estimate})$.
Correct. The standard error for the slope is 0.0035, not 0.1609. The latter is the standard deviation of the log Bacterial Counts about the straight line.

Exercise Book 10

Question 1

Correct. The correlation coefficient must lie between -1 and +1, inclusive.

Question 2

Correct. A correlation of zero means no linear association.

Question 3

Correct. The t statistic is $(r \cdot \sqrt{n-2})/\sqrt{1-r^2}$, where n is the sample size and r is the correlation coefficient.

Question 4

Correct. This is a two sided test and you have reached the correct conclusion.

Question 5

Correct. You need normal distributions and random samples.

Question 6

Correct. The points will lie on a straight line with a negative slope.

Selected Output from Minitab

From Workbook 7...

Workbook 7:

Report on the 2nd experiment.

Effect of Drug on Blood Clotting Time.

A drug which was believed to hasten blood clotting time was tested by comparing a drug group (n=64) with a placebo group (n=30)

Minitab output follows...

```
'O:\TLTP\QUERCUS3\MFILES\WRKBK7.MTW'.
```

```
MTB > TwoSample 95.0 'Drug' 'Placebo';  
SUBC> Alternative 0.
```

Two Sample T-Test and Confidence Interval

Twosample T for Drug vs Placebo

	N	Mean	StDev	SE Mean
Drug	64	6.46	3.01	0.38
Placebo	30	7.62	3.09	0.56

95% C.I. for mu Drug - mu Placebo: (-2.52, 0.20)

T-Test mu Drug = mu Placebo (vs not =): T = -1.71 P = 0.093 DF = 55

Stem-and-Leaf 'Drug' 'Placebo'

Character Stem-and-Leaf Display

Stem-and-leaf of Drug N = 64
Leaf Unit = 0.10

```

 2   1 59
 6   2 5699
14   3 01222557
23   4 003366668
(12) 5 001334667788
29   6 3699
25   7 1234778
18   8 0279
14   9 04569
 9  10 11357
 4  11 4
 3  12
 3  13 5
 2  14 02
```

Stem-and-leaf of Placebo N = 30
Leaf Unit = 0.10

```

 1   1 0
 3   2 26
 4   3 6
 5   4 1
 9   5 3689
13   6 3356
(3)  7 266
14   8 48
12   9 225689
 6  10 0257
 2  11
 2  12 5
 1  13
 1  14 7
```

Describe 'Drug' 'Placebo'

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SEMean
Drug	64	6.458	5.740	6.301	3.011	0.376
Placebo	30	7.615	7.645	7.614	3.086	0.564

Variable	Min	Max	Q1	Q3
Drug	1.570	14.240	4.115	8.585
Placebo	1.010	14.700	5.832	9.863

```
MTB > TwoSample 95.0 'Drug' 'Placebo';
SUBC>    Alternative 0;
SUBC>    Pooled.
```

Two Sample T-Test and Confidence Interval

Twosample T for Drug vs Placebo

	N	Mean	StDev	SE Mean
Drug	64	6.46	3.01	0.38
Placebo	30	7.62	3.09	0.56

95% C.I. for μ Drug - μ Placebo: (-2.49, 0.18)

T-Test μ Drug = μ Placebo (vs not =): T = -1.72 P = 0.088 DF = 92

Both use Pooled StDev = 3.03

From Workbook 12...

THE VACCINE EXPERIMENT

Results and Discussion

I analysed the data using Minitab.
The following is MINITAB Output.

NOVA 'response' = vaccine farm.

Analysis of Variance (Balanced Designs)

Factor	Type	Levels	Values			
vaccine	fixed	3	1	2	3	
farm	fixed	4	1	2	3	4

Analysis of Variance for response

Source	DF	SS	MS	F	P
vaccine	2	2328.50	1164.25	133.91	0.000
farm	3	485.58	161.86	18.62	0.002
Error	6	52.17	8.69		
Total	11	2866.25			

TABULATED STATISTICS

ROWS: vaccine

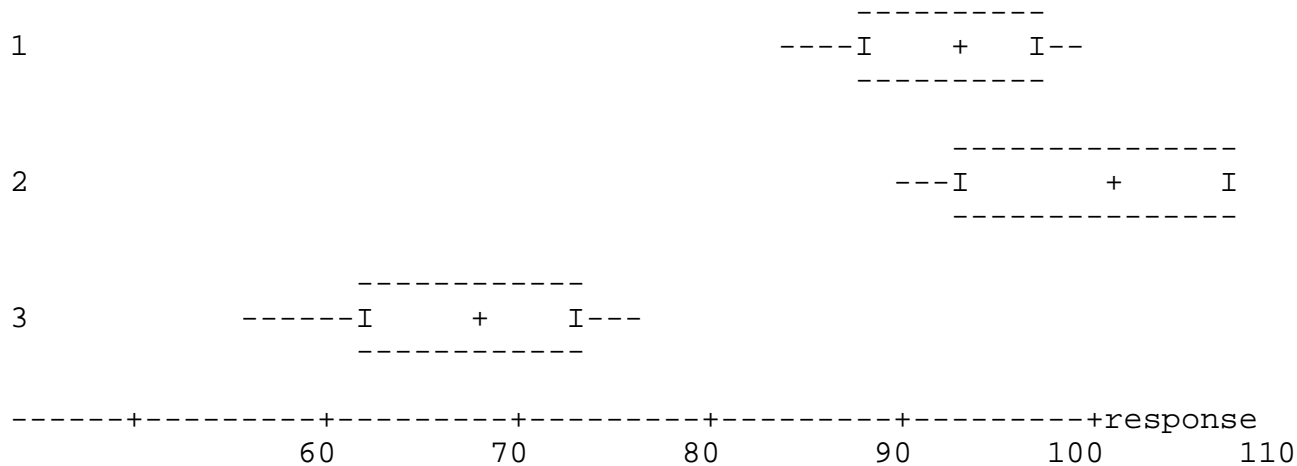
	response N	response MEAN	response STD DEV
1	4	92.25	6.40
2	4	99.50	8.35
3	4	67.00	8.29
ALL	12	86.25	16.14

GRAPHS OF DATA

```
BoxPlot 'response';  
By 'vaccine'.
```

Character Boxplot

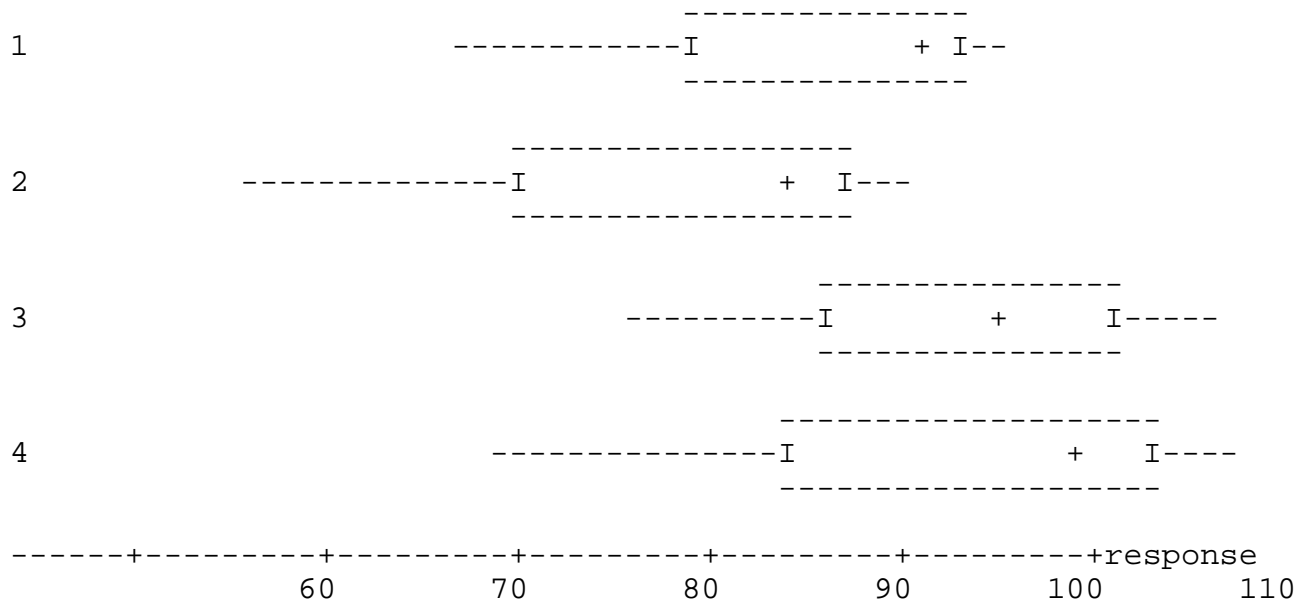
vaccine



```
BoxPlot 'response';  
By 'farm'.
```

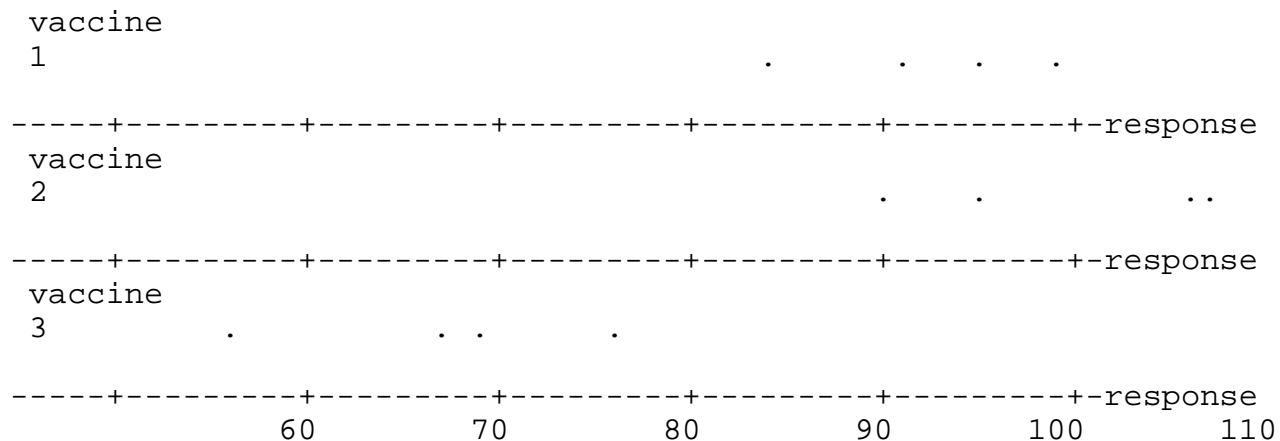
Character Boxplot

farm



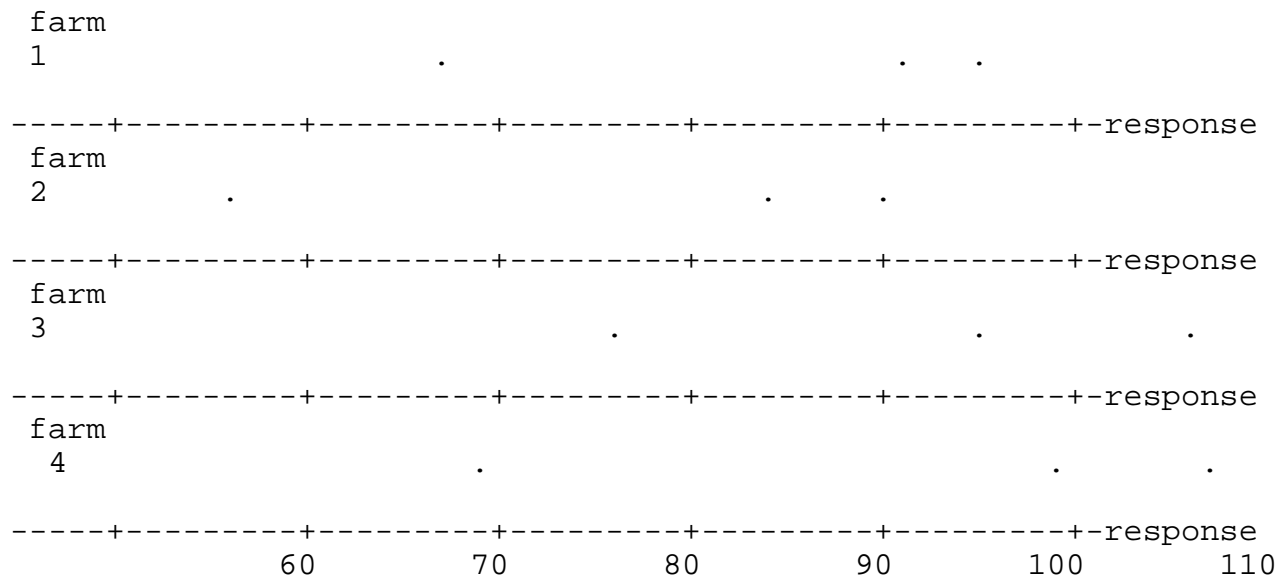
```
DotPlot 'response';
By 'vaccine'.
```

Character Dotplot



```
DotPlot 'response';
By 'farm'.
```

Character Dotplot



From Workbook 8...

Example 1: A Comparison of Bull Meat pH in Imported and Home Produced Beef

09 December 1998

Introduction

Outline Aim of the Experiment. Q: What hypothesis is being tested? A: The mean pH of the sample of imported beef is not significantly different from the mean pH of home produced meat.

Method

Outline experimental method.

Analysis of Data.

Describe method of analysis, mention and justify any assumptions about the data.

1 sample t-test.

Normally distributed .

Results.

Cut and paste graphs.

The t statistic is -1.07 $p = 0.29$.

Conclusions

Because $p > 0.05$, we accept the Null hypothesis that there is no evidence that the mean pH of the sample of imported beef is different from that of home produced beef.

Example 2: The Effect of Steroid Treatment on Red Blood Cell Recovery

Introduction

Outline Aim of the Experiment. Q: What hypothesis is being tested? A: The mean number of red blood cells is not significantly greater in patients treated with steroids than in patients treated with a placebo.

Method

Outline experimental method.

Analysis of Data.

Describe method of analysis, mention and justify any assumptions about the data.

2 sample t-test.

Normally distributed .

Results

The t statistic is 6.96 $p = 0.0000$.

Conclusions

The P-value for the statistic is less than 0.05, therefore we reject the Null hypothesis and accept the Alternate hypothesis that the evidence indicates that steroid treatment enhances red blood cell recovery.

Exam Paper: January 1999

SECTION 1 (Compulsory: Answer ALL Questions)

- (1) The following data are the wool weights (ounces) of lambs born and bred at the UWB farm in 1980: 32 37 40 41 44 45 46 49 50 52 53 63 79
- (a) Test whether the sample comes from a population with a mean of 45 ounces at the 5% significance.
 - (b) What is the 99% confidence interval for the population mean?
 - (b) Is there any evidence that one or more of the observations are outliers?
 - (c) Explain the terms significance level and confidence interval in your own words.
- (2) You are about to undertake the following experiment. You have a large number of plant pots available. You are waiting to compare the height to which grass will grow over a certain time period when planted in one of two fertilisers. One fertiliser is the standard fertiliser currently used which has given a mean and standard deviation of 35mm and 10mm respectively. It is hoped the new fertiliser will give an increase of 5mm in height over the standard treatment.
- (a) State the name of the statistical test you intend to use to analyse the data you will collect.
 - (b) State the assumptions that the test chosen in question (a) above makes and state how you intend to check the assumptions.
 - (c) How many plant pots do you intend to use in your experiment. Justify your choice.

- (3) The file `o:\statdata\data\capefear.mtw` contains data on the experiment described below. The data summarises a piece of research to identify soil characteristics and relationships in the Cape Fear estuary in North Carolina.

There are 3 factors, Type and Location.

Type is a label for 3 areas, 1 - revegetated areas, 2 - short grass area, 3 - tall grass area.

Location is a label for 3 geographic locations, 1 - Oak Island, 2 - Smith Island, 3 - Snows Marsh.

There are 5 chemical measurements:

soil salinity - measured in part per thousand.

pH - acidity as measured in water

Kk - potassium in part per million

Na - sodium in part per million

Zn - zinc in part per million

There is one biological measurement, Biomass, which is the aerial biomass of the new grass in gm^{-2} .

Answer **ONE** of the following questions.

- (a) Is the salinity concentration found in the plots the same or different when compared at the three locations?
- (b) Is it possible to predict the concentration of zinc from that of the pH?

In your answer, include summary statistics from any output you use, give clear statements about the statistical tests you have used. I only expect one analysis to be performed answering the central question, together with any checks of the assumption. If the assumptions do not hold, make sensible recommendations about what you would do. Do not reanalyse the data.

Credit will be given for a clearly written and reasoned conclusion.