Spreadsheets: introduction to working with statistics

IT services
UNIVERSITY OF OXFORD

How to Use this Guide

This handbook accompanies the taught sessions for the course. Each section contains a brief overview of a topic for your reference and then one or more activities.

Your teacher will direct you to the location of files that are needed for any exercises. If you have any problems with the text or the activities, please ask the teacher or one of the demonstrators for help.

Following attendance on the course you may attend follow-up sessions at ITLP called Computer8, where you can get support when using the techniques in your own work

Software Used

Excel 2010

Files Used

Revision Information

Version	Date	Author	Changes made
1.0	15/6/2014	S Albury	Initial version

Copyright

Acknowledgements

Contents

1 Introduction

Welcome to our course on working with statistics in Excel

This booklet accompanies the course delivered by Oxford University's IT Learning Programme. Although the exercises are clearly explained so that you can work through them independently, you will find that it will help if you also attend the taught session where you can get advice from the teacher, demonstrator(s) and even each other!

If at any time you are not clear about any aspect of the course, please make sure you ask your teacher or demonstrator for some help. If you are away from the class, you can get help by email from your teacher or from help@it.ox.ac.uk.

1.1. What You Should Already Know

No previous knowledge of statistics is expected. We will assume that you have some knowledge of *Excel* which may be gained via other courses and that you are familiar with entering text and simple editing, rearranging and formatting - copy and paste, printing and previewing and managing files.

We will also assume that you are familiar with opening files from particular folders and saving them, perhaps with a different name, back to the same or a different folder.

The computer network in our teaching rooms may differ slightly from that which you are used to in your College or Department; if you are confused by the differences please ask for help from the teacher or demonstrator(s).

1.2. What Will You Learn?

This course will help you learn to understand and use statistics when working with data.

In this session we will cover the following topics:

- Using basic averages
- Looking for trends in data
- Looking at variability in data
- Frequency and ranking of items
- Exploring relationships in data

A list of all the statistical functions in Excel can be found at

http://office.microsoft.com/en-gb/excel-help/statistical-functions-HP005203066.aspx

This course looks at a number of the most widely used ones.

These notes provide examples in Excel 2010 using *Windows*. Having worked through these notes, you should also be able to adapt to other spreadsheets including on an Apple Mac or specialist statistics software, since the principles hold true regardless.

Getting to grips with statistics is extremely useful but give yourself plenty of time for practice as applying the techniques correctly can take practice.

1.3. Using Office 2010

If you have previously used another version of *Office*, you may find *Office 2010* looks rather unfamiliar. “Office 2010: What’s New” is a self-study guide covering the Ribbon, Quick Access Toolbar and so on. This can be downloaded from the ITLP Portfolio at http://portfolio.it.ox.ac.uk.

For anyone who prefers not to use the mouse to control software, or who finds a keyboard method more convenient, it is possible to control *Office 2010* applications without using a mouse. Pressing ALT once displays a white box with a letter or character next to each visible item on the Ribbon and title bar (shown in Figure 1).

Figure 1 Keystrokes to Control Ribbon Tabs and Title Bar (Press ALT to Show These)

After you type one of the letters/characters shown, the relevant Ribbon tab or detail appears, with further letters/characters for operating the buttons and controls.

The elements of a dialog can be controlled, as usual with *Windows* applications, by using TAB to navigate between items or typing the underlined character shown beside an item.

1.4. What is the purpose of using statistical analysis?

Statistics are a useful way to understand what is going on in a set of data. For example if you want to know the average income from a particular activity or examine changes in the number of people attending college functions in different terms. Being able to work with and discuss statistics can help to ensure that you have all the necessary information to make or influence decisions. Statistics however are open to interpretation, and learning how to make good use of them is just as important as knowing the techniques.

1.5. Other Tools

There are many other tools that you could use for working with statistics. For example a popular tool on the Mac is *Numbers*. Although these tools will differ in the way that the features are used, the principles described in this session are equally applicable.

There are also a number of web based tools that you might investigate. For example, Google docs has a powerful spreadsheet and includes a lot of statistical analysis features.

2 The purpose of statistics

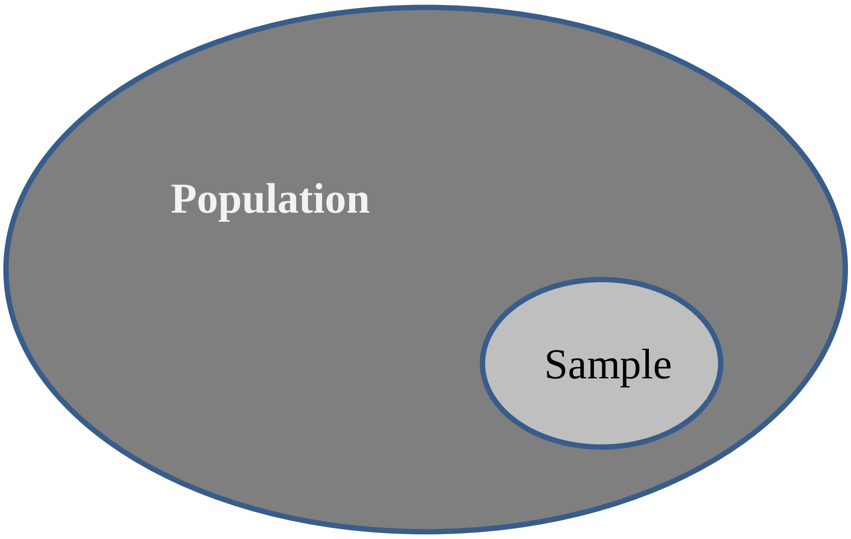
This course is intended to help you develop awareness and understanding of using basic statistics in your work. However, creating statistical information without having a clear motivation and purpose is a waste of effort.

In this part of the course we look at what the role is of statistics and explore what type of questions statistical analysis can help to answer.

2.1. What are 'statistics'

A statistic is a measure of a sample from a population. It is calculated from a sample as access to the whole population is not possible for a number of reasons. If the whole population has been measured it should be called a parameter (ie a known fact about the population). For most purposes (outside of formal research) the word statistic is used for both population and sample measures.

The difference between a sample and a population can be thought of like this:



This diagram also indicates an important point about statistics, they are based on probability. If we took a different sample we might get a different result. This is often overlooked when statistics are reported and can lead to wrong conclusions being drawn.

For example if a survey is carried out of students about the quality of the food in hall that sample may lead to an incorrect conclusion being drawn if most of those students rate the food as 'good'. It might be that many other students don't eat in hall because they think the food is 'poor'. A lot of work in statistics involves making sure a representative sample is obtained and also working out the probability of the sample result reflecting the population. This is often talked about as a 'margin of error'.

2.2. Collecting data for statistical analysis

As statistics are based on probability it is important to understand how surveys and other data collection techniques can be used. This course doesn't cover survey design or data collection more generally but there are some key points that should be kept in mind.

Sampling

When collecting a sample of data it is always a good idea to try and ensure it is representative of the population. To do this decisions about who that population is and how it can be described need to be made.

For example if a group of invoices is to be sampled for errors then what type of invoices should be included? What is meant in this case by 'error', does it apply to all invoices or only ones over a certain value or within a certain time period?

There are many questions that are often overlooked due to pressure of being asked for specific information often on a short timescale but time spent asking detailed questions will be repaid later.

Another problem that arises with sampling is bias. This relates to the problem of the sample not being randomly drawn from a population. This can happen for many reasons but convenience is one of them. For example, a sample of library users' opinions about the service might fail to pick up issues if only people actually using libraries is included. The course 'concepts of statistics' includes more help on sampling and understanding how and in what forms statistics can be biased is a useful skill when working with them.

Timeliness

Statistical analysis can be affected by the time period of data collection. An example is student course feedback. If sent out too early responses will be partial and not based on full information about the course, even if the sampling is correct. However, if sent out after the course it can be hard to get responses and the results may not reflect the real opinion of the group.

Hypothesis testing and Type 1 & II Errors

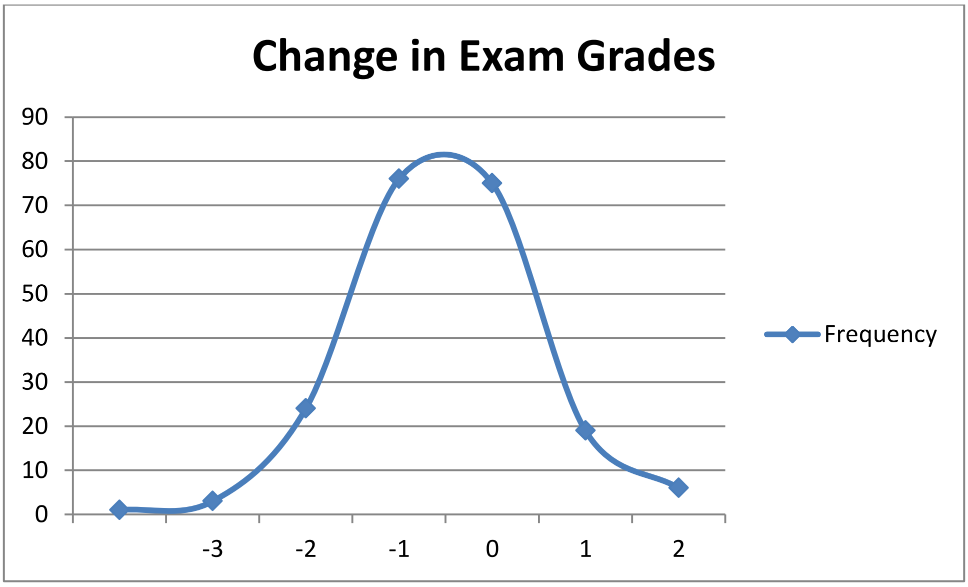
This refers back to the fact that statistics are based on probability and this means we can get a false positive result or a false negative. When carrying out some statistical tests we use a hypothesis, a statement making a claim that we then test to see if the evidence supports it. If we get a false positive and think something is true when it isn't this is a Type I error. If we find from our sample that our hypothesis is false when actually it is true this is a Type II error.

A special hypothesis is also used in statistical analysis, the null hypothesis, often called H0. This is a hypothesis which we use to assume our claim is false and is the basis of many tests. This means that where we have sufficient evidence to suggest the null hypothesis is false (meaning our own alternative hypothesis 'H1' is accepted) we can say that we have evidence to support our claim. This does not mean the claim is true, it means that based on the sample it has a high chance of being a non-random result. This shows that far from statistics ‘proving’ things they are really about interpreting the results to see what is the most likely case to be true. However, if you have made a type I or II error you are unlikely to know about it unless more information becomes available to show you that the original sample was not a good model of the population!

Distributions and Predictions

If the weather forecasts predicts a 90% chance of rain tomorrow with a confidence level of 95% and it doesn’t rain was the prediction wrong? It feels as though the prediction was wrong, but more likely it means that the 10% chance of rain was also predicted with a confidence level of 95% and the weather fell into that part of the distribution. It seems that the weather forecast was wrong but it shows that in a lot of cases people turn a statistical probability (90%) into a physical certainty (“they said it would rain today but it didn’t, they were wrong again”).

A data set being analysed will have values in a range. This distribution will take different forms but commonly creates what is often called a bell curve. When discussing measures of location such as averages it is the position along the range of values that we are looking at. The mean, as an average, always has 50% of the data above and below it. The position from the centre is often reported in terms of standard deviations which shows how data is dispersed around the mean value. A steep bell curve indicates data is closely gathered and a flatter curve suggests a bigger spread of values.



Example of a bell curve showing a normal distribution, created in Excel.

Sometimes a distribution is skewed with more values in one of the ‘tails’ – in this case the median is often a preferred measure of the average as it accounts for these extreme values better. For example if looking at salaries, a well paid employee can make the mean salary look higher than what almost everybody is actually paid. The median helps take account of this as discussed in the next section.

2.3. Types of data

The data used in statistical analysis can come in a variety of forms. There is no universally agreed definition for some of the terms which means that it can be a bit confusing. However the following description should allow you to understand and use the correct terms.

2.3.1. Variable

Simply put a variable is something that varies! It is used as a name for a group of things that are being observed. For example we could have a variable 'Fee type' in a spreadsheet about student enrolments, or one called 'Student Price' on a spreadsheet for managing dining hall meal sales in a college. All of the data that you analyse will be part of a set and will have a variable name. In addition to this it will have the some of the following features .

2.3.2. Discrete and Continuous data

A variable can be discrete, meaning that it can be counted and if the value is zero it means there are none of the item. Continuous variables exist on a number line where any value is possible based on the accuracy of measurement. An example of discrete data would be the number of respondents to a survey who rated something as 'Good'. An example of continuous data would be 'time taken to complete task 1'.

2.3.3. Quantitative and Qualitative data

The term 'quantitative' or 'cardinal' is applied to interval and ratio data as it measures amounts such as 'sales value' or 'books borrowed'. Qualitative or 'categorical' is applied to data where the group's categories are the most important thing and no value is implied, such as 'fee status' or 'satisfaction rating'. Note that some qualitative values are often treated as if they were quantitative but care needs to be taken to ensure the results are reasonable. Nominal and ordinal data is qualitative. (not to be confused with 'qualitative research' though this error is often made in market research).

2.3.4. Nominal and Ordinal data

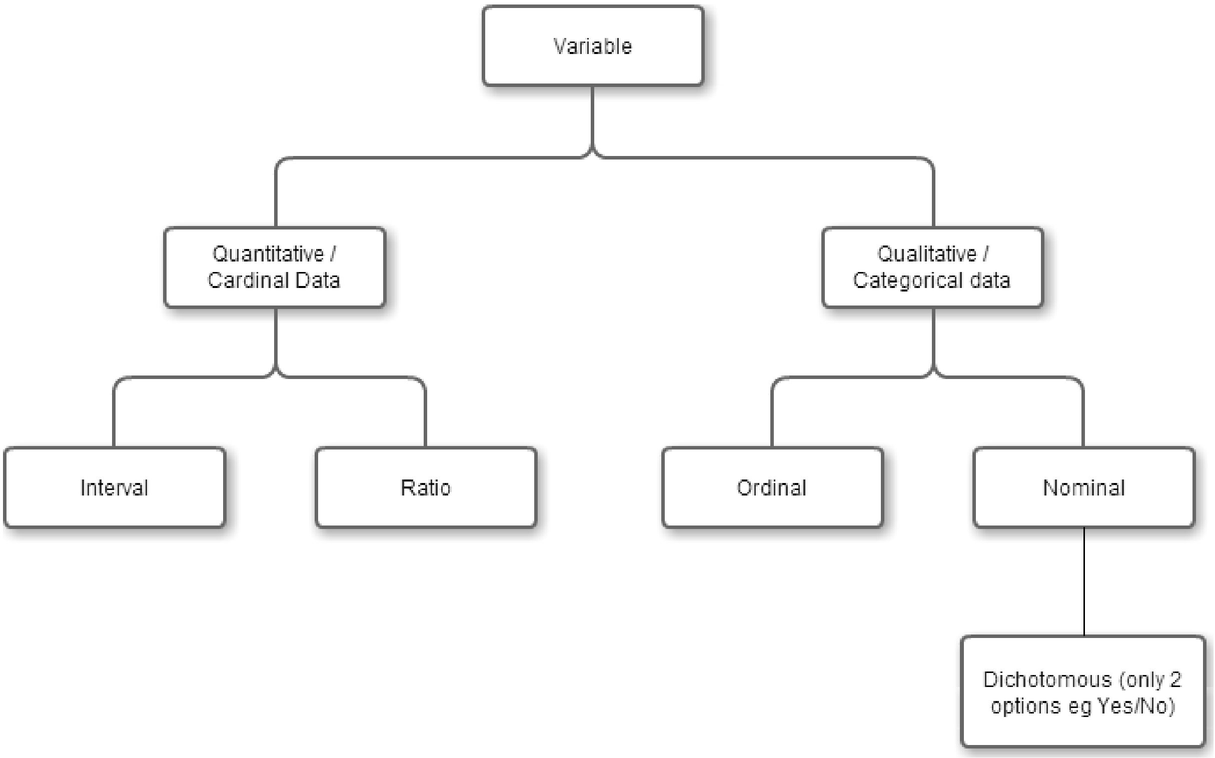
Data that can be grouped such as 'gender', but has no intrinsic order, is nominal. This means we cannot for example talk about the mean to discuss an 'average' gender in a group. However you can use a different measure of the 'average' called the mode, which shows the most frequently occurring category. Ordinal data is nominal data that has an implied order. Degree classification is an example because a doctoral degree is ranked higher than a bachelor's degree. Ordinal and nominal data is very common in survey results and is sometimes treated as if it were quantitative.

2.3.5. Interval and ratio data.

Interval data is a very specialist form of continuous data but is not very common. It means that there is an ordered scale which is constant and has a value but the scale is not an absolute measure as zero does not mean there is none of the item. For example, temperature in degrees Celsius is interval data. If the temperature is zero degrees that does not mean there is no heat. The interval between 20 and 25 degrees is the same as between 40 and 45 degrees but that does not mean that 40 degrees is twice as hot as 20 degrees despite this being common usage. Much more common is ratio data. This is data that can be represented on a continuum and can be treated as an absolute value. An invoice amount of zero for example means no money is owed, and £40 is twice as much money as £20. The key difference between interval and ratio data is the fact that the continuum is both infinite and that zero represents a lack of the item. This is why degrees Kelvin is a temperature scale which is ratio rather than interval, as zero degrees would mean an absence of heat and 40 degrees Kelvin represents twice as much heat as 20 degrees Kelvin.

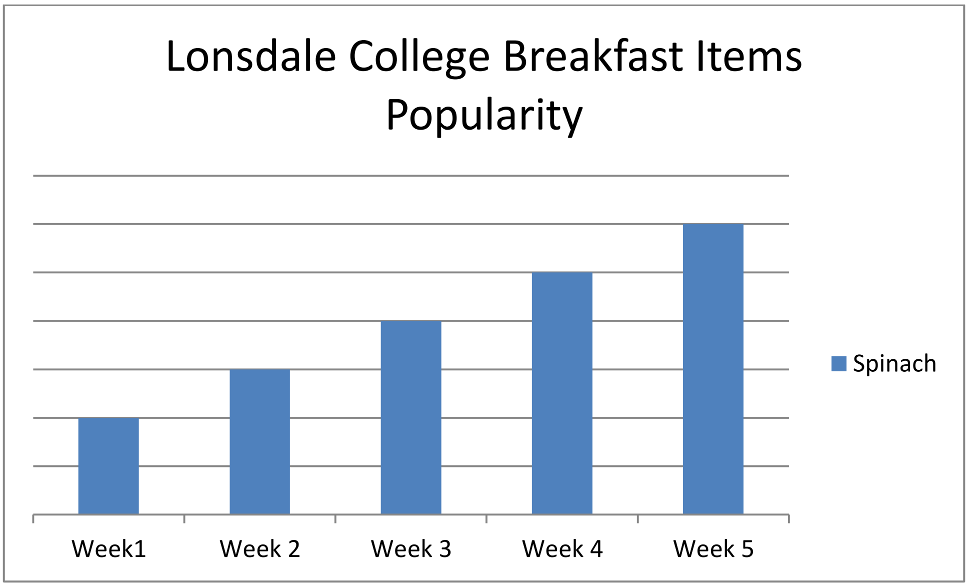
2.3.6. A Basic Overview of Data Types

The following diagram shows a basic overview of how data types are classified. Confusingly it's not always clear what the data type is. For example, survey questions will often have a 5 point scale, from 'very satisfied' to 'very dissatisfied'. Technically this is ordinal data. It is not possible to say that one score is a known amount more than another only that it is more. However these Likert Scales are often treated as values from 1 to 5 and are treated as continuous ratio data. Care should be taken to ensure that results are only treated as indicative, but the practice is so widespread it doesn't matter too much that some statisticians don't accept it.


Variable
Quantitative / Cardinal Data
Qualitative / Categorical data
Interval
Ratio
Ordinal
Nominal
Dichotomous (only 2 options eg Yes/No)

2.4. Uses for statistics

Very often it is helpful to know whether an initiative has had an impact or to report on the effects of a change in business processes. A common use for statistics is to help in making a business case where providing evidence for a requested change in services can enhance its chances of being accepted. Another area where statistics can be useful is spotting patterns and trends and looking for relationships that are not obvious. It is also important however to spot where statistical interpretation has strayed into being misleading such as the use of carefully crafted scales as below, which accurately shows a 'trebling' of demand for 'spinach at breakfast' in the refectory of 'Lonsdale College'



It can be seen that demand has indeed 'trebled' but from 2 people wanting it to 6 people, out of 350 people eating breakfast regularly (i.e. less than 2%). By careful use of scales it is possible to both mislead, and be misled by, statistical reports. Obviously nobody at Oxford would ever choose to use underhand techniques to increase their chances of persuading others with numbers ☺

2.5. Exercises

The exercises for this course are available from the IT portfolio at http://portfolio.it.ox.ac.uk

3 Averages and the 'shape' of data

In this part of the course we will look at the ways we can find a mid-point in data. What we think of as an 'average' can also be compared to a centre of gravity, the point that keeps the data in symmetry with half of it being below and half above. This is sometimes called a measure of central tendency.

Choosing an average can be very useful in getting a proper idea of where the central point is and Excel allows us to identify different types of average without having to understand or use the formulas for calculating them.

The 'shape' of data is also important because it gives us information on the spread and variability of data. For example two sets of data may share the same mean (what most people think of as 'the average') but the spread of the data might be very different. This may or may not be important information and this highlights another important aspect of statistical analysis, what it is being used for.

3.1. Types of average

3.1.1. The mean

This is the most common average calculated and this is reflected in the Excel function name **AVERAGE()**. It is a simple concept but can lead to wrong assumptions being made.

The mean is simply the addition of all the elements in a set and dividing by the number of elements as below:

119
156
131
118
118
112
127
158
184
=(D30+D31+D32+D33+D34+D35+D36+D37+D38)/9

You can use the SUM() function to save adding individual functions either with the : for a range – so SUM(D30:D38) would do the adding up or if the cells aren't next to each other you can use the ',' which allows you to specify individual cells SUM(D30,D33,D35). Note the use of brackets to ensure the addition happens before the division.

It is easier however to simply use the **AVERAGE()** function as this gives the arithmetic mean which represents the central point of a data set and looks like this:

119
156
131
118
118
112
127
158
184
=AVERAGE(D30:D38)

Which calculates as

135.89

The arithmetic mean is based purely on the value of the sum of the components. An alternative mean can also be calculated in Excel, the geometric mean. This is the product of all the items and then the root of the number of items. For example given the following set: 3,3,4,5,5,6 then multiply all the items which is 5400. As there are 6 elements in the set the geometric mean needs the 6th root

$\sqrt[6]{5400}$ which is 4.18. The arithmetic mean is easier to calculate as it just means adding up the numbers to 26 then dividing by 6 which is 4.33. However the geometric mean offers a number of advantages and for data sets where there are outliers (extreme values) it takes better account of them.

The geometric mean function in Excel is GEOMEAN(). Consider using it when working with data that is based on percentages as it provides a better average percentage value than the arithmetic mean. This is also true for comparing different types of item. This is because it equalises the impact of any single data point. For example if working out an ‘average’ salary in a research team then we may have numbers similar to (in thousands of £):

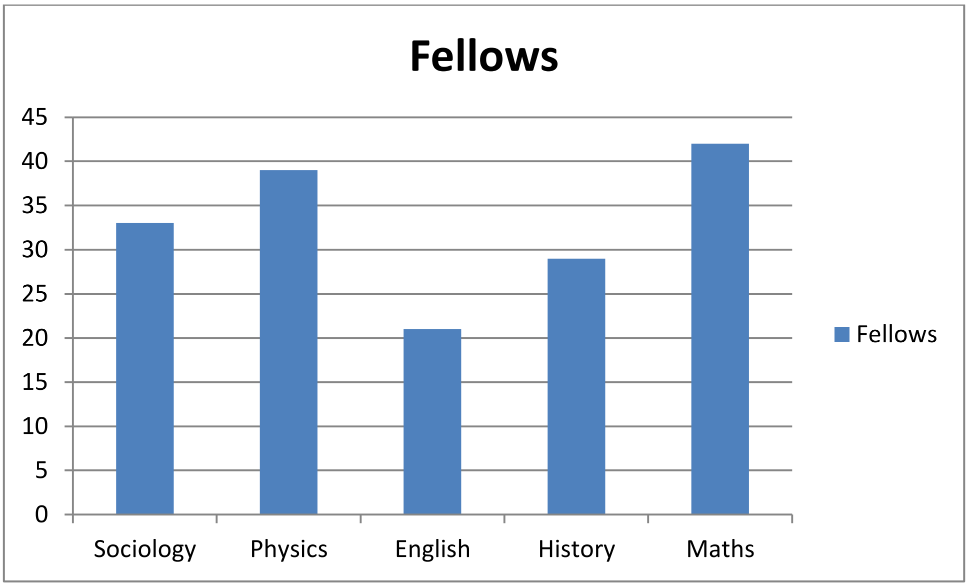
20,20,22,27,27,29,31,31,32,72

The arithmetic mean would be £31,100 but this means the large majority of the team are paid ‘below average’. This is because this is a positively skewed distribution and is distorted by one high salary. The geometric mean however is £28,938. This is a better reflection of what the ‘average’ employee on the project is paid.

3.1.2. The Mode

The mode is simply the most frequently occurring item in a set. It is the only average that should be used for nominal data and is very useful to identify the popularity of a response to a survey question. The mode represents the value that is most likely to be sampled (if you picked one item from your data the modal value is the most likely to be selected) which indicates that any sample where more than two items or present may have more than 1 mode. The Excel functions **`MODE.SG()`** and **`MODE.MULT()`** are the two functions you can use. If there is no mode because all the items have different values this suggests a very unstable data set and might lead to questions about why that is the case and whether it is a problem or not.

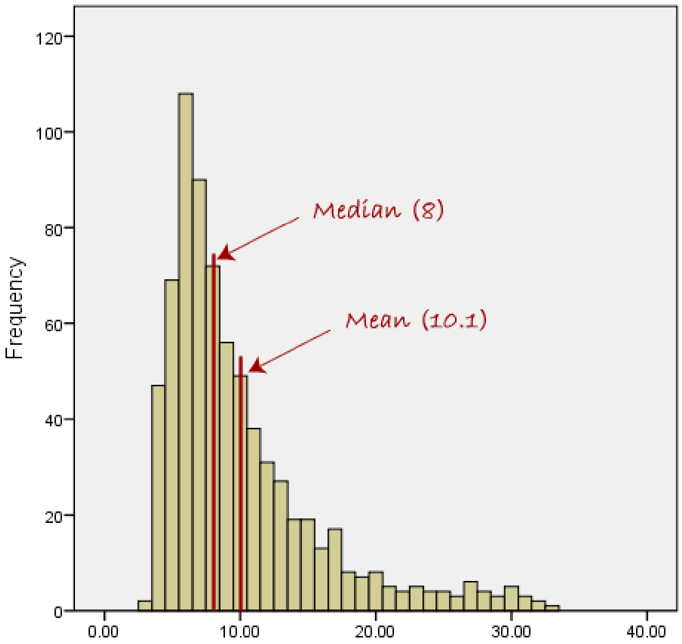
If you plot the mode it is always the highest point in a frequency distribution (because it is the maximum point of density of data and therefore the most likely to be sampled). Looking at the background of Lonsdale College's fellows the modal value is 41 meaning the most common type of fellow is a mathematician.



3.1.3. The Median

The median is the value that is in the centre of an ordered set of data. In the series 2,3,3,5,6,17,19 the median is 5 as it means there are three values above and three below. If there is an even number of data points then the median is simply the mean of the two central ones. For example the series 4,4,5,7,11,12 has a median of (5+7)/2. As a measure of the 'centre' the median is based on an ordered list and the number of items, rather than the value of the items (either their sum or their product as for the mean). The median is a better measure of the average when the data are skewed. If the data are positively skewed (there is a long right tail) then the median will be lower than the mean. If there is a long left tail then the median will be higher. In both cases however the skewed nature of the distribution indicates that there are a large number of outlying values. The median, is much less affected by outliers than the arithmetic mean and therefore provides a better 'average' value.

A skewed distribution is one that looks like this:



The median can be used with quantitative data and is also often used with categorical ordinal data. The use of the median with categorical data is not accepted by all statisticians but the practice of calculating averages of survey results is widespread and the median is often a better tool than the mean in these situations, as it smooths the extreme responses to survey questions.

3.2. Measures of Position and Dispersion

As can be seen from the screenshots data has a 'shape' when graphed. We can discuss the data in terms of where an item is positioned (which part of the distribution it is in) and also in terms of how spread out the data is. When discussing about position in a data set it is usual to talk in percentile terms. A percentile P_i has a greater value than $_i$th percent of values. The median represents the value in the centre of the distribution and so is always the 50th percentile. A value in the 99th percentile would be larger than 99% of all the items. Often these percentiles are grouped into quartiles. A quartile represents a quarter of all the data; so te 1st quartile is the same as the 25th percentile, the 2nd quartile is the 50th percentile and the 3rd quartile is the 75th percentile. Note how the 2nd quartile includes everything above the first and below the 3rd.

Dispersion is related to how peaked or flat the data is; this is usually discussed in terms of variance and standard deviations. The standard deviation is the square root of the variance, and represents the amount by which any value deviates from the mean. To find the variance in a sample we look at the difference of each data point from the mean and then square that number (this removes any negative values). We sum these and divide by the number of data items (which gives us the mean of all the individual variances). If we take the square root of this number we have the standard deviation. About 95% of observations are within two standard deviations of the mean.

Excel has functions for calculating the variance and standard deviation of a data set. It is worth noting that the standard deviation can be calculated for the sample or for the whole population. These are slightly different but a bigger effect is seen with small samples. If you don't have access to the population data you should try to ensure a sample size of at least 20. There are more formal methods for calculating a sample size with the greatest probability of reflecting the population but as a rule of thumb a larger sample is better than a smaller one.

To calculate the variance in a sample use the **`VAR()`**function with the data set. For the standard deviation there are a number of options but mainly you should use either the **`STDEV.S()`** function for a sample or the **`STDEV.P()`** function for a population.

3.3. Exercises

The exercises for this course are available from the IT portfolio at http://portfolio.it.ox.ac.uk

4 Graphical Representation & Trends in Data

4.1. What is a trend?

A trend is where you are looking at either two items of ratio data or a combination of ratio and categorical data. It indicates a movement in a particular direction over time. Most rends can be spotted using a graphical representation and 'eyeballing' it. This part of the course looks at selecting a graph type and making a judgement as well as spotting where trends have been exaggerated or hidden because of the chart design.

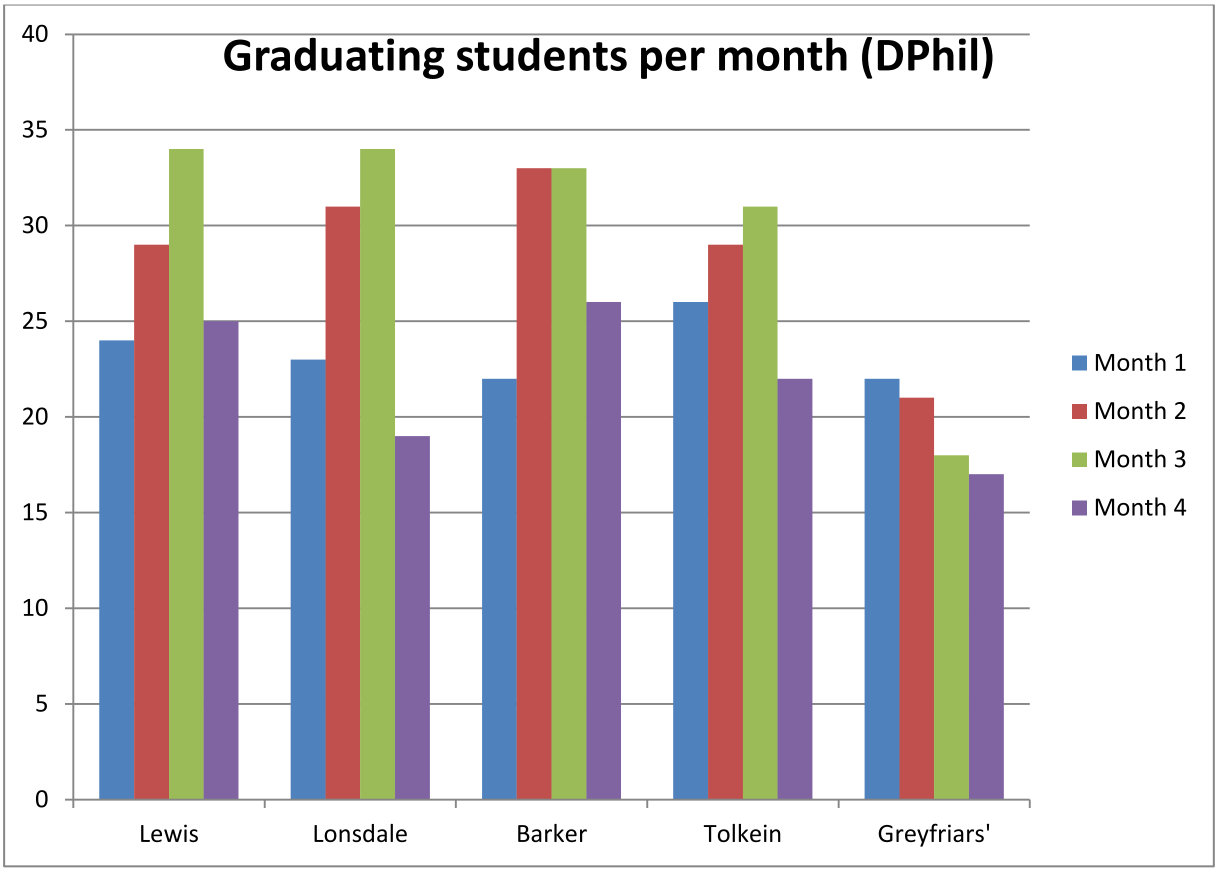
4.2. Types of chart and eyeballing trends

A first step is to work out what is the key message you want to explain using the trend. They can be both positive and negative but the message needn't always just reflect what the trend is saying about the situation now. It can also be used to discuss how things might develop in the future or explain a specific reason why the trend has developed the way it has. Also a trend can be changed and this is often the point of revealing them, to plan on action that will change the direction or steepness of a trend line.

There are numerous chart types in Excel and more information can be fdound in the course Spreadsheets: Organising and Displaying data. Excel allows you to plot a trend line on a graph. However care should be taken if using this feature as there are different types of trend that can be monitored.

4.2.1. Bar Charts

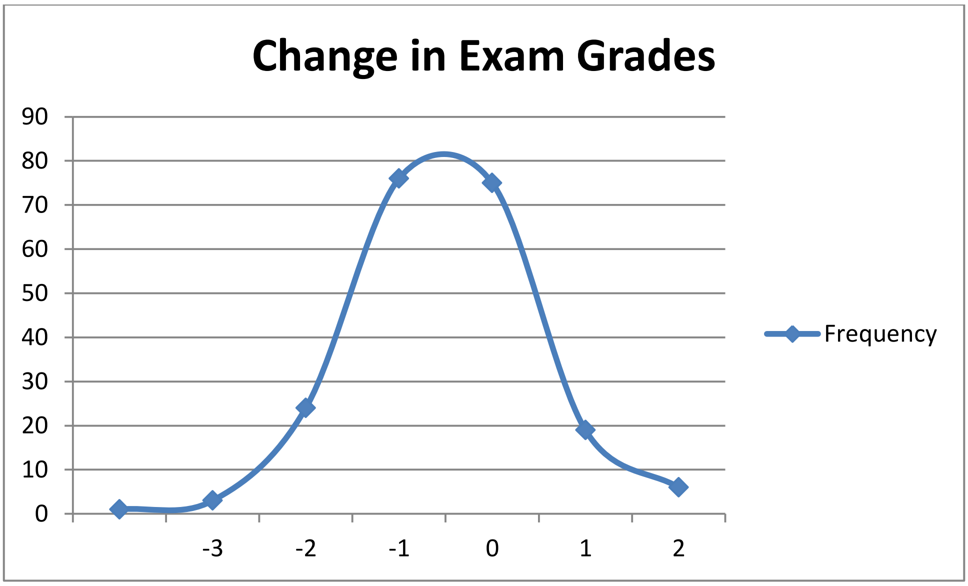
Bar graphs use columns to display discrete data. Categorical data doesn't have an intrinsic order and can be rearranged. You can also group items into categories and show a grouped bar chart. This shows how dependent and independent variables interact. The independent variable is the cause or input into a test and the dependent variable is the output – it depends on the independent variable.

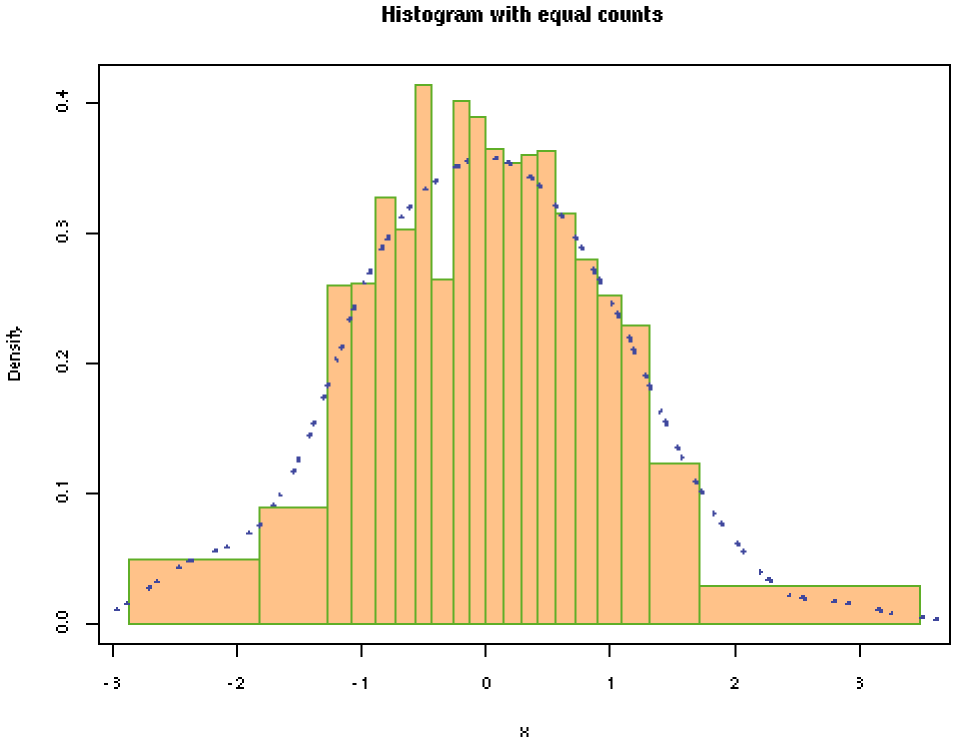


This chart shows grouped categorical variables on the ‘x’ axis and a continuous variable on the ‘y’ axis. It can be seen that for each college, except Greyfriars there is an upward trend until month 4. This might raise questions about why that is, or lead to an examination of what is different in that month. For Greyfriars it could mean more investigation into why a declining number of graduations is taking place. There could be very good explanations for these points, but looking for trends is an important part step of understanding what the data is saying.

4.2.2. Histogram

A histogram is used to collect continuous variables into categories and plot them as a frequency distribution. A histogram can provide a quick view of whether a data set is normally distributed. Unfortunately although you can create histograms in Excel if you plot them it shows as a normal bar chart. As a histogram is of continuous data there should be no gaps between bars and they also have an order. When plotting a histogram in Excel therefore you should create the graph as a scatter plot with smoothed lines. This is how the chart shown earlier was created



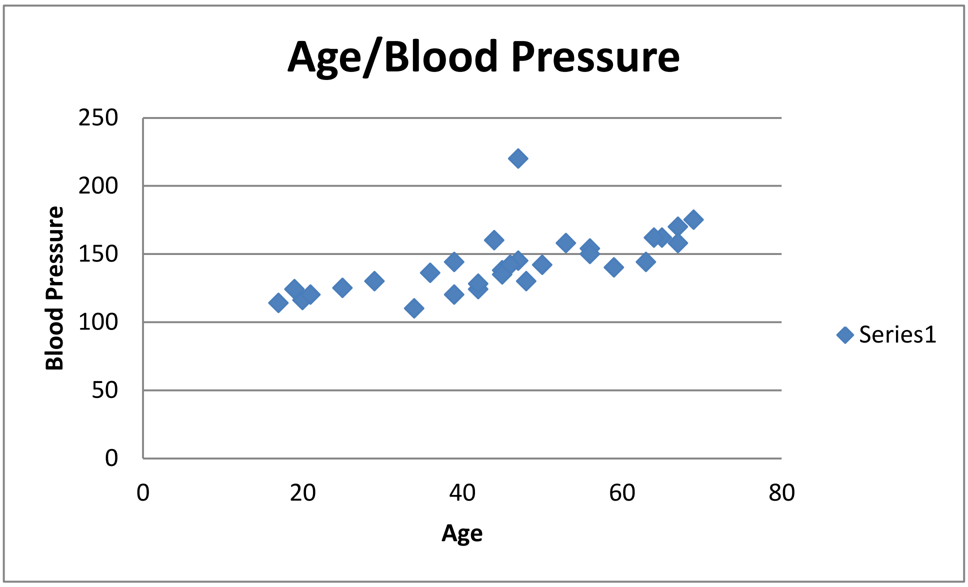


This image shows a histogram plotted using statistics software. It shows how a histogram represents a distribution and also that the width of the bars is not necessarily equal as the range covered by one 'bin' can vary. Excel does not mimic this behaviour precisely but does have functions in the data analysis add-on to create frequency tables, the charting is the main issue.

4.2.3. Scatter Plots

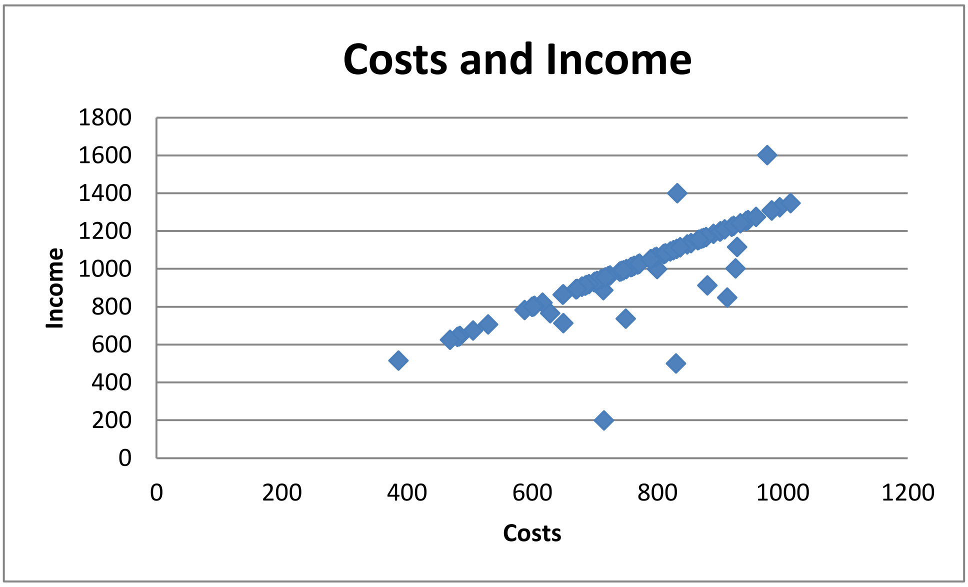
A scatter diagram can help show a trend in the relationship between two continuous variables. It helps quickly identify clusters of data and outliers but also a direction of relationship, either positive or negative. It is also possible with a scatter plot to explore two other important relationships, regression and correlation. Regression is where is where we want to see the effect on one variable (the dependent variable) when we change another one (the independent variable). A correlation is one continuous variable changes in line with another one. The first rule of correlation club is that correlation is not causation. Things can appear to change in line with each other but this does not show that a change in one causes the change in the other.

For example hot weather can be shown to cause two phenomena, an increase in crime and an increase in ice cream sales. They are closely correlated. It would not be sensible however to suggest an increase in ice cream sales causes an increase in crime!



In the above chart it is possible to see that blood pressure appears to be higher as people get older but that there are also some outliers with 1 person aged about 50 having the highest blood pressure of all.

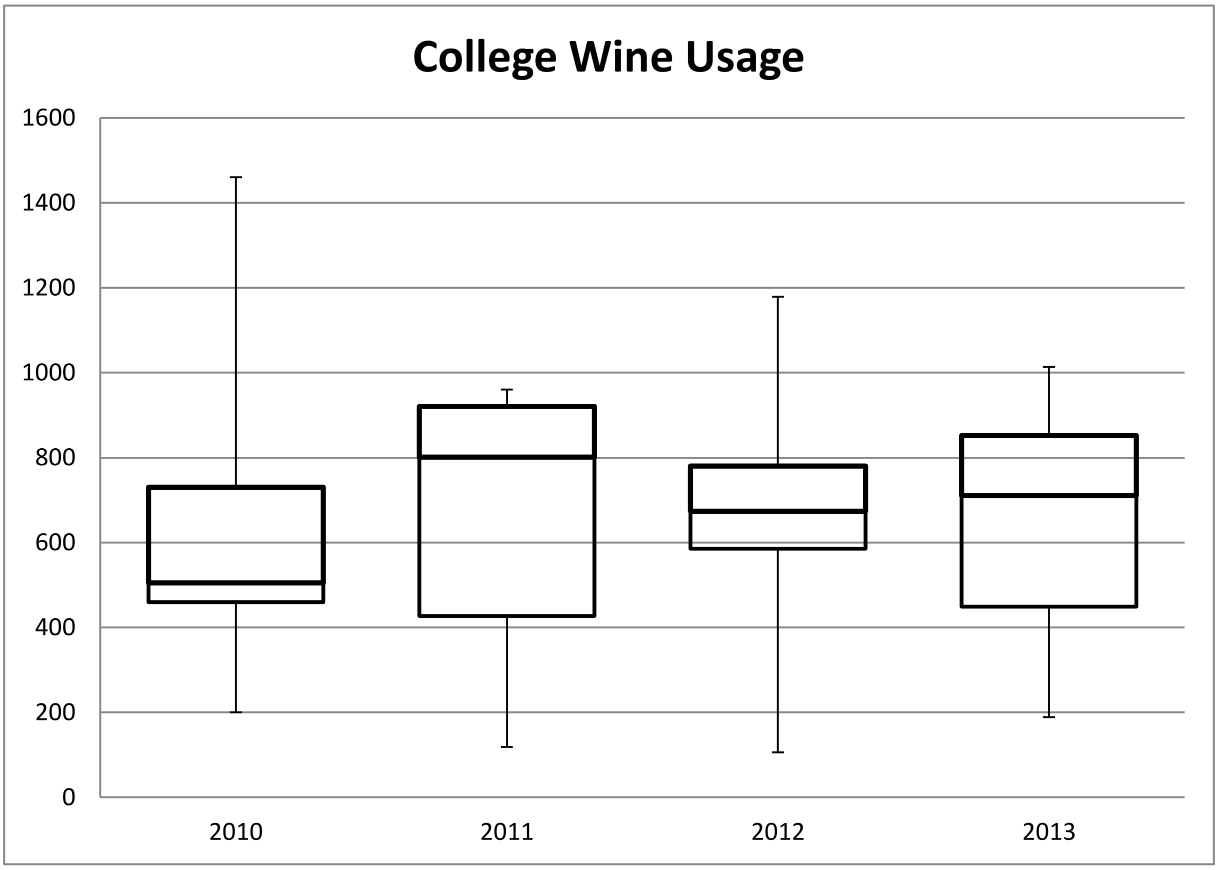
This scatter chart shows a stronger correlation



This chart suggests that, excluding outliers, higher costs are associated with higher income and that they move in tandem so as costs increase income also moves up. What we cannot say about this chart though is that they are connected and an increase in income is dependent on costs going up (e.g. we need to invest more to increase income). There is a correlation and we can calculate a value for that called Pearson's correlation coefficient. A value of 1 means things move perfectly together positively (an increase in 1 leads to an equivalent increase in the other) and a value of -1 means they diverge perfectly (a movement of 1 leads to a decrease of 1 in the other). You can do this using the **`CORREL()`** function and in this case it is 0.814 which is a high level of correlation.

4.2.4. Box and Whisker Charts

A boxplot or box-and-whisker chart is way of looking at a range of descriptive statistics in one graphical representation. Unfortunately Excel has no inbuilt way to show these types of charts but it is possible to edit a stacked bar chart to achieve the same effect. The benefit of a boxplot is that it allows you to see changes in a range of data points. For example if you have exam grades for a number of years a boxplot provides a quick view of how for each year the maximum and minimum, the quartiles and the median have changed. In many statistical analysis tools the upper and lower ranges can be set to exclude true outliers but this cannot be done in Excel.



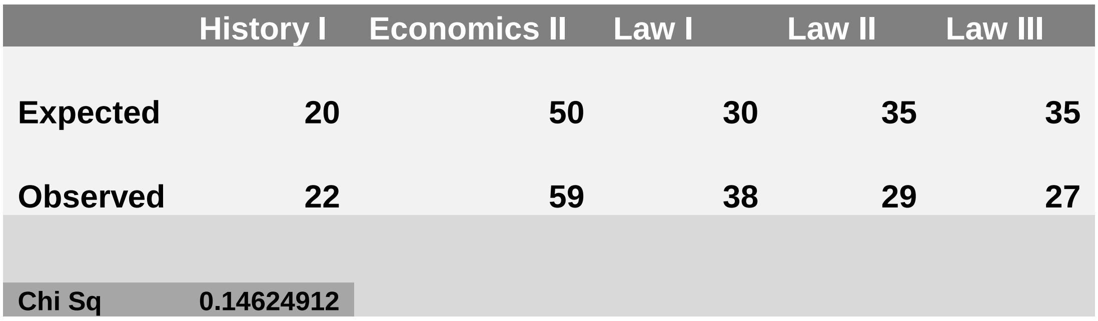
5 Categorical Variables and the Chi Squared Test

The Chi squared test (pronounced Kai squared) is a way to compare categorical variables. It provides a way to see the 'goodness of fit' between categories and should only be used where you have at least 5 items in each group. The test tells us whether the relationship between an observed set of results compares to a standard set (are these results consistent with what we might expect). It really tells you whether the distribution of your data is consistent with expected results and can act as an indicator of where problems or outperformance might be present.

We also make a number of basic assumptions about the test. One of the most important is the level of statistical significance, which can feel a little backwards. If a base level of 5% is accepted this means that there is a 5% chance that results are not what you would expect. A chi squared score of less than 0.05 (which is 5%) indicates a value outside the expected range and this then needs further investigation.

Like all statistical results however this should be viewed as indicative. If a value of 0.049 is scored or 0.055 is this a problem? Results of such tests are useful information for decision making but each situation needs to be seen in context. Any value can be set for the acceptable level but 5% is common.

For example a faculty may look at results for a set of collections and know the average results for these based on 30 years of data:



	History I	Economics II	Law I	Law II	Law III
Expected	20	50	30	35	35
Observed	22	59	38	29	27
Chi Sq	0.14624912				

The value for Chi squared is 0.14 and because this is higher than 0.05 we do not reject the null hypothesis that the results are in line with what we would expect.

6 Exercises

The exercises for the statistics course are available from the IT portfolio at http://portfolio.it.ox.ac.uk

7 What Next?

Courses offering training in statistics and related topics are described below. In all cases, please refer to the IT Learning Programme web page (via www.it.ox.ac.uk/courses/) for further details.

7.1. Further Courses in spreadsheets and statistics

Now that you have some basic *statistics* skills you may want to develop them further or develop more advanced spreadsheet skills. Other spreadsheets courses include:

Official Title and Code	Topics
Spreadsheets: essential techniques for working with data	Understanding formulas and functions Creating basic formulas Copying data and using ranges Sorting and Filtering data Introduction to Charts Worksheet formatting
Spreadsheets: Techniques for managing and checking data	Organising and cleaning up data Testing and documenting spreadsheets Auditing a spreadsheet Making changes to complex workbooks Designing a spreadsheet solution
Spreadsheets: Creating professional data views	Creating formulae Named cells, ranges and constants Functions and how to use them; logic and lookups Formatting spreadsheet data Viewing and printing large workbooks Working in 3D Charts: creating and controlling
Spreadsheets: An introduction to working with statistics	Using basic averages Looking for trends in data Looking at variability in data

	Frequency and ranking of items Exploring relationships in data
Spreadsheets: Organising and displaying data	Working with ranges and tables of data Filtering your data Subtotal and group your data Charting non-adjacent worksheet areas and charting filtered data Working with mixed chart types Formatting charts Charting on a secondary axis Adding data series Creating histograms and sparklines
Spreadsheets: Advanced data analysis and modelling	
Spreadsheets Summarising Data using pivot tables	Creating and formatting pivot tables Expanding and collapsing pivot table data Filter sort and group pivot table data Use report filters with pivot table data Using Claculated fields in pivot tables Creating pivot charts Using Slicers to analyse pivot table data

The IT Learning Programme offers other courses you may find helpful and the courses catalogue can help identify them – http://courses.it.ox.ac.uk

7.2. Computer8

We encourage everyone to work at their own pace. This may mean that you don't manage to finish all of the exercises for this session. If this is the case, and you would like to complete the exercises while someone is on hand to help you, come along to one of the Computer8 sessions that run during term time. More details are available from www.it.ox.ac.uk/courses/

7.3. IT Services Help Centre

The Help Centre is a good place to get advice about any aspect of using computer software or hardware. You can contact the Help Centre on (2)73200 or by email using help@it.ox.ac.uk

7.4. Downloadable Course Materials and More – the ITLP Portfolio

These course materials are available through the ITLP Portfolio, at http://portfolio.it.ox.ac.uk .

Each course pack includes the course handbook in pdf form and a zip folder of the exercise files that you need to complete the exercises. Archive versions of the course book may also be useful if you use an earlier version of the software.

The ITLP Portfolio helps you find articles, videos, resources and weblinks for further IT study. For some resources, you will be asked for your Oxford (SSO) username and password.

8 References

Curwin J, Slater R (2007) Quantitative Methods for Business Decisions

http://onlinestatbook.com/2/introduction/levels_of_measurement.html

(accessed 25 November 2014)

http://www.usablestats.com/index.php

(accessed 25 November 2014)

http://www.excel-easy.com/functions/statistical-functions.html

(Accessed 25 November 2014)