Foundation Module Draft

Objectives

1. Understand some of the core terminology of statistics
2. Understand the relationship between populations and samples in education research
3. Understand the basic operation of SPSS
4. Understand descriptive statistics and how to generate them using SPSS
5. Know how to graphically display data using SPSS
6. Know how to transform and compute variables using SPSS
7. Understand the basics of the normal distribution, probability and statistical inference
8. Know how to compare group means

You can jump to specific pages using the contents list below. If you are new to this module start at the **Overview (Page 1.1)** and work through section by section using the 'Next' and 'Previous' buttons at the top and bottom of each page. Do the **Exercise** and the **Quiz** to gain a firm understanding.

Contents

1.1 Overview

For Those of You New to Statistics...

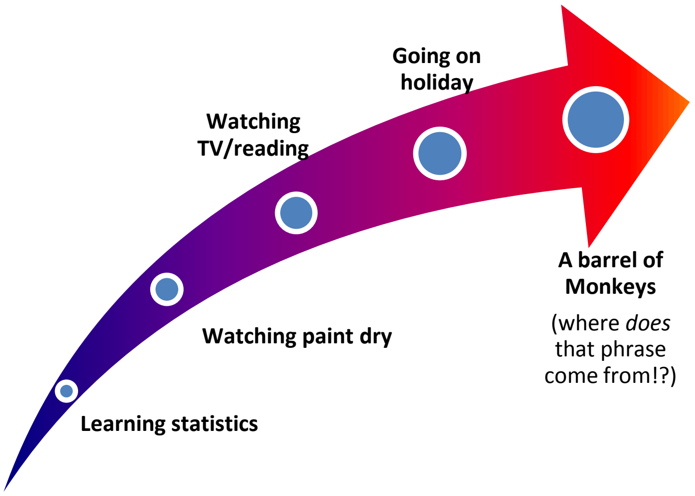
If you're new to this site you will probably fall in to one of two categories:

Category A: You are fairly enthusiastic about learning quantitative research skills and are eager to get stuck in to some data. You may already have some experience with statistics and may even be able to skip this module altogether and get started on one of the regression modules. Good for you!

Category B: You are experiencing a powerful and compelling dread. You have never studied statistics and you did not get involved with social research to get bogged down in all of these meaningless numbers and formulae. Despite this, you feel you need to understand at least some statistics in order to become a more competent researcher.

If you're in category B we're not going to lie to you. Learning to understand and use statistics might not be much fun. Here is the 'Fun Scale' (**Figure 1.1.1**) which was absolutely *not* based on real world research by a team of highly skilled and experienced academics. Please note the position of 'Learning statistics':

Figure 1.1.1: The Fun Scale



As you can see, poor old statistics does not fare well... It can be challenging and boring when you first encounter it. But we're not trying to scare you off – far from it! We're trying to put some fight in to you. Statistics can become a very useful tool in your studies and career and we are living proof that, with the application of a bit of hard work, you can learn statistical skills even if you are not particularly maths orientated. The truth is that, once over the first few hurdles, most people find that stats becomes much easier and, dare we say it... quite enjoyable! We hope that the category B's among you will stick with us.

Why Study Statistics?

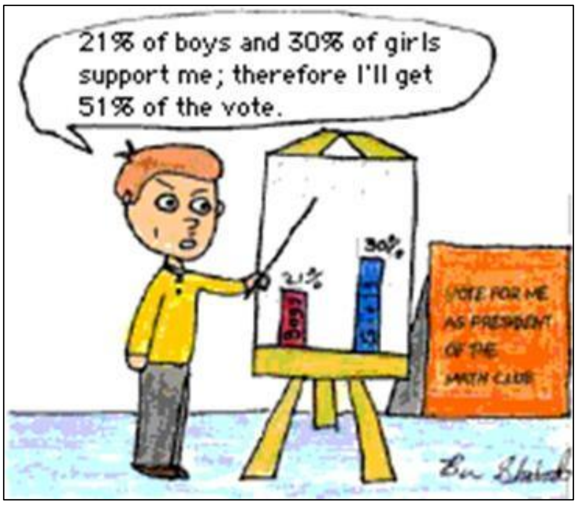
We live in a world in which statistics are everywhere. The media frequently report statistics to support their points of view about everything from current affairs to romantic relationships.

Political parties use statistics to support their agendas and try to convince you to vote for them. Marketing uses statistics to target consumers and make us buy things that we probably don't need (will that sandwich toaster *really* improve my life?). The danger is that we look at these numbers and mistake them for irrefutable evidence of what we are being told. In fact the way statistics are collected and reported is crucial and the figures can be manipulated to fit the argument. You may have heard this famous quote before:

> *"There are three kinds of lies: lies, damned lies, and statistics."*
>
> Mark Twain.

We need to understand statistics in order to be able to put them in their place. Statistics can be naturally confusing and we need to make sure that we don't fall for cheap tricks like the one illustrated below (**Figure 1.1.2**). Of course not all statistics are misleading. Statistics can provide a powerful way of illustrating a point or describing a situation. We only point out these examples to illustrate one key thing: statistics are as open to interpretation as words.

Figure 1.1.2: The importance of mistrusting stats


21% of boys and 30% of girls support me; therefore I'll get 51% of the vote.
21%
Boys
30%
VOTE FOR ME AS PRESIDENT OF THE MATH CLUB

Understanding the prevalence of statistics in our society is important but you are probably here because you are studying research methods for education (or perhaps the social sciences more broadly). Picking up the basics will really help you to comprehend the academic books and papers that you will read as part of your work. Crucially, it will allow you to approach such literature critically and to appreciate the strengths and weaknesses of a particular methodology or the validity of a conclusion.

Perhaps more exciting is that getting your head around statistics will unlock a vast tool box of new techniques which you can apply to your own research. Most research questions are not best served by a purely qualitative approach, especially if you wish to generalize a theory to a large group of individuals. Even if you don't ever perform any purely quantitative research you can use statistical techniques to compliment a variety of research methods. It is about selecting the correct methods for your question and that requires a broad range of skills.

The advantages of using statistics:

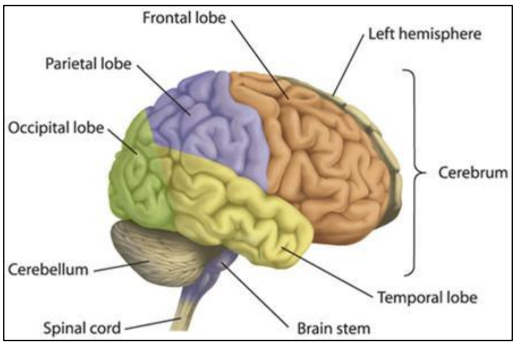
- Stats allow you to summarize information about large groups of people and to search for trends and patterns within and between these groups
- Stats allow you to generalize your findings to a population if you have looked at an adequate sample of that population
- Stats allow you to create predictive models of complex situations which involve a lot of information and multiple variables

What is SPSS?

We won't get in to the fine details yet but basically SPSS (sometimes called IBM SPSS or PASW) is computer software designed specifically for the purpose of data management and quantitative analysis in the social sciences. It is popular at universities all over the world and, though not perfect, it is a wonderful tool for education research. As we shall see it allows you to perform a dazzling array of analyses on data sets of any size and it does most of the heavy lifting for you... You won't need to perform mind-numbing calculations or commit terrifyingly complex formulae to memory.

Sounds great doesn't it!? It is great. *BUT* you have to know what you're doing to use it well. It has an unsettling tendency to spew tables of statistics and strange looking graphs at you and you need to learn to identify what is important and what is not so important. You also need to know how to manipulate the data itself and ask SPSS the right questions. In other words the most important component of the data analysis is *ALWAYS* you!

Figure 1.1.3: Engage Brain Before Touching Keyboard!



Statistical techniques and SPSS are tools for analysing good data based on good research. Even if you're an expert statistician who performs a flawless analysis on a dataset your findings will be pointless if the dataset itself is not good and your research questions have not been carefully thought out. Don't get lost in the methods. Remember you are a researcher first and foremost!

Now that we've set the scene and tried our best to convince you that learning statistics is a worthwhile endeavour let's get started by looking at some of the basic principles that you will need. We are not aiming to provide a full and thorough introduction to statistics (there are plenty of materials available for this, just check out our **Resources** page) but we do hope to provide you with a basic foundation.

1.2 Population and Sampling

The Research Population

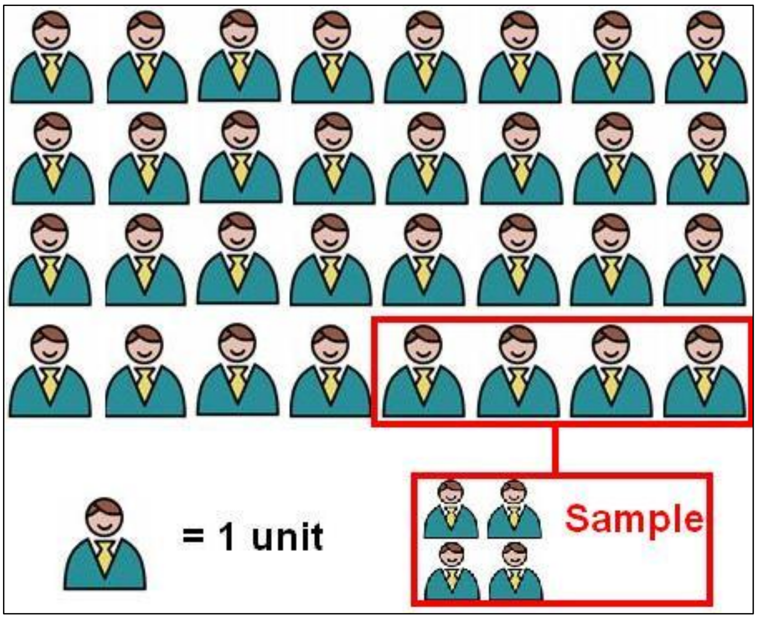
The word ‘population’ is in everyday use and we usually use it to refer to a large group of people. For example, the population of a country or city is usually thousands or millions of individuals. In social research the term can have a slightly different meaning. A population refers to *any* group that we wish to generalize our research findings to. Individual cases from a population are known as units. For example, we may want to generalize to the whole population of 11-12 year old students in the UK in order to research a particular policy aimed at this group. In this case our population is *all* British 11-12 year olds, with every child being a single unit. Alternatively we may be performing a piece of research where the sole objective is to improve the behaviour of a certain year group (say year 7) in a specific school (Nawty Hill School). In this case our population is year 7 of Nawty Hill School. Our research is usually intended to say something about the particular population we’re looking at.

In both of the above examples the populations were made up of individual people as units but this is not necessarily always the case – it depends on how we frame our research question. It may be that we want to compare behaviour at all secondary schools in the South of England (the infamous Nawty Hill Secondary does not come out of this analysis looking good!). In this case each individual *school* is a unit with every school in the South of England making up the population.

In an ideal world we would be able to gather data about every unit in our population but this is usually impractical because of issues of costs in terms of money, time and resources. Returning to an earlier example, what if we wanted to gather achievement data about every 11-12 year old student in the UK? Unless you have a truly enormous budget (sadly unlikely in these credit-crunched times) and plenty of research assistants you will not be able to interview or get a questionnaire back from *all* of these students.

However it is not necessary to gather data on every member of a population in order to say something meaningful, you could draw a sample from the population. **Figure 1.2.1** shows the relationship between samples and populations. A group of individuals (units) is selected from the entire population to represent them. If the sample is drawn well (more on this later) then it should accurately reflect the characteristics of the entire population... it is certainly more efficient and cost effective than contacting everyone!

Figure 1.2.1: Sampling a population



Note that, despite what this image may suggest, most populations are not consisted entirely of featureless male office workers.

Selecting a suitable sample is more problematic than it sounds:

- What if you only picked students who were taking part in an after school club?
- What if you only picked students from schools in the local area so that they are easy for you to travel to?
- What if you only picked the students who actively volunteered to take part in the research?

Would the data you gained from these groups be a fair representation of the population as a whole? We'll discuss this further in the next section. Below is a summary of what we have covered so far:

Population

The population is all units with a particular characteristic. It is the group we wish to generalise our findings to.

Populations are generally too large to assess all members so we select a sample from the population.

If we wish to generalise it is important that the sample is representative of the population. The method used for drawing the sample is key to this.

Selecting a sample

Selecting a representative sample for your research is essential for using statistics and drawing valid conclusions. We are usually carrying out quantitative research because we want to get an overall picture of a population rather than a detailed and contextualized exploration of each individual unit (qualitative approaches are usually better if this is our aim). However, a population is a collection of unique units and therefore collecting a sample is fraught with risk – what if we accidently sample only a small subgroup that has differing characteristics to the rest of the population?

For example, imagine we were trying to explore reading skill development in 6 year olds. We have personal connections with two schools so we decide to sample them. One is based in the centre of a bustling metropolis and the other is based on a small island which has no electricity and a large population of goats. Both of these samples are six year old students but they are likely to differ! This is an extreme example but we do have to be careful with such 'convenience' sampling as it can lead to systematic errors in how we represent our target population.

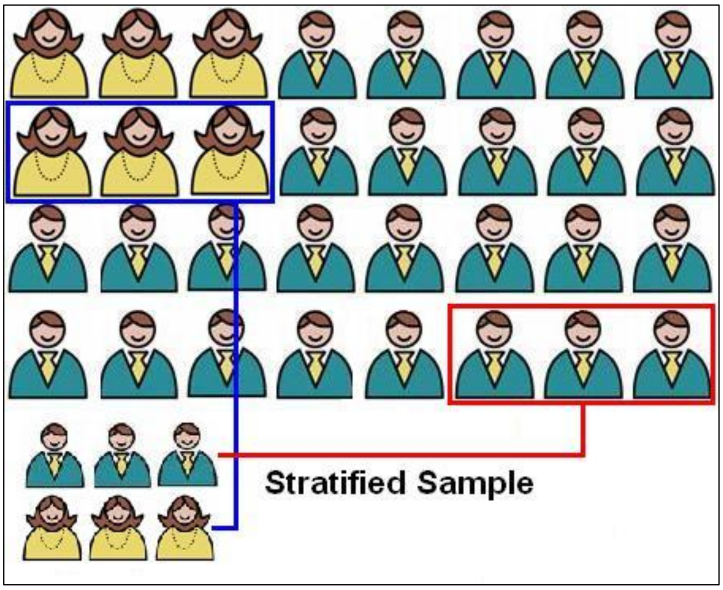
The best way to generate a sample that is representative of the population as a whole is to do it randomly. This 'probability sampling' removes bias from the sampling process because every unit in the population has an equal chance of being selected for the sample. Assuming you collect data about enough participants you are likely to create a sample that represents all subgroups within your population.

For example, returning to Nawty Hill Secondary School, it is unlikely that all of the 2000 students who attend regularly misbehave. A small proportion of the students (let's say 5%) are actually little angels and never cause any trouble for the poor harassed teachers. If we were sampling the school and only chose one student at random there would be a 1 in 20 change of picking out one of these well-behaved students. This means that if we only took a sample of only 10 there would be a chance we wouldn't get one well-behaved student at all! If we picked 100 students randomly we would be likely to get five well-behaved students and this would be a balanced picture of the population as a whole. It is important to realize that drawing samples that are large enough to have a good chance of representing the population is crucial. We'll talk about sample size and probabilities a lot on this website so it is worth thinking about!

There are also more sophisticated types of sampling:

Stratified sampling: This can come in handy if you want to ensure your sample is representative of particular sub-groups in the population or if you are looking to analyse differences between subgroups. In stratified sampling the researcher identifies the subgroups that they are interested in (called strata) and then randomly samples units from within each strata. The number of cases selected from each strata may be in proportion to the size of the strata in the population or it may be larger depending on the purpose of the research. This can be very useful if you want to examine subgroups of units which are not well represented in the overall population. For example, 5% of students in England identify themselves in 'Black' ethnic groups, but a random sample, unless it is very large, may well not include 5% of Black students. A stratified sample might be drawn to ensure that 5% of the sample are from Black ethnic groups. Alternatively a 'boosted' sample might target some groups to ensure enough individuals are selected to form a good basis of comparison. **Figure 1.2.2** illustrates a stratified sampling strategy including a boosted sample for females who are under-represented relative to males in the population (this is not uncommon when looking at course enrolment for degrees in science, technology, engineering and mathematics, for example).

Figure 1.2.2: A stratified sample



Note that, despite what this image may suggest, most females are not featureless and do not have beards.

For example, if you wanted to compare the well-behaved and poorly behaved students at Nawty Hill School and you randomly selected 100 students from the whole population you might get less than 5 well behaved students, but if you stratify by behaviour and select within strata you can guarantee that you will get 5 well behaved students. Indeed you could over select from within the poorly behaved stratum to select a sample of 25 well-behaved and 75 poorly-behaved students so you had large enough samples to make reliable comparisons. It is important though that sampling within the subgroups should still be random where at all possible.

Cluster sampling: When the population is very large (e.g. a whole country) it is sometimes viable to divide it into smaller groups called clusters. First, several of these clusters are randomly selected for analysis. After this individual units from within each selected cluster are randomly selected to make up the sample. For example if we wanted to sample all students in the UK it might be worth first dividing the population into geographic clusters (e.g. South-east, North-west). We would then randomly decide which of these regions we would draw our sample from and this would give us smaller groups to work with (much more practical). For cluster sampling to be viable there should be minimal differences between clusters - any substantial differences must be *within* them.

We've discussed sampling in some depth now. In summary:

Sampling

Probability sampling: There is an element of randomness in how the sample was selected from the population. Can be quite sophisticated (e.g. stratified sampling, cluster sampling).

Non-probability sampling: Convenience sampling (those readily available), or selecting volunteers. Greater risk of a biased or unrepresentative sample.

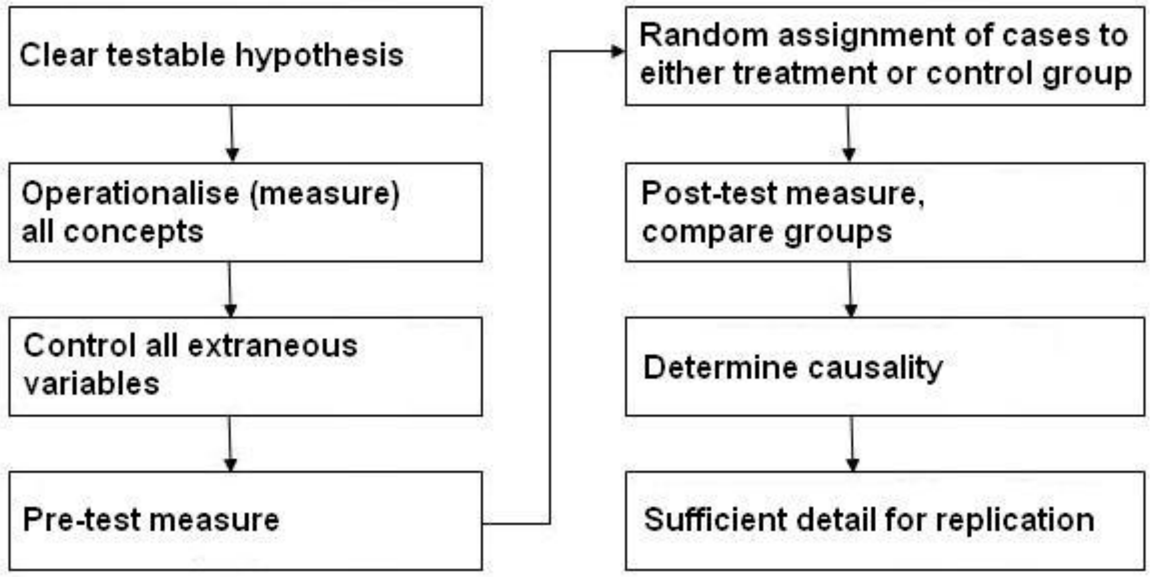
1.3 Quantitative research

Types of research design

This website does not aim to provide an in depth discussion about research methods as there are comprehensive alternative sources available if you want to learn more about this (check out our **Resources** page, particularly **Cohen, Manion & Morrison, 2007; 6th Edition; chapters 6-13**). However, it is worth discussing a few basics. In general there are two main types of quantitative research design.

Experimental designs: Experimental designs are highly regarded in many disciplines and are related to experiments in the natural sciences (you know the type - where you nearly lose your eyebrows due to some confusion about whether to add the green chemical or the blue one). The emphasis is on scientific control, making sure that all the variables are held constant with the exception of the ones you are altering (independent variable) and the ones you are measuring as outcomes (dependent variable). **Figure 1.3.1** illustrates the type of process you may take:

Figure 1.3.1: The process of experimental research


Clear testable hypothesis
Operationalise (measure) all concepts
Control all extraneous variables
Pre-test measure
Random assignment of cases to either treatment or control group
Post-test measure, compare groups
Determine causality
Sufficient detail for replication

A **quasi-experiment** is one where truly random assignment of cases to intervention or to control groups is not possible. For example, if you wanted to examine the impact of being a smoker on performance in a Physical Education exam you could not *randomly* assign individuals into ‘smoking’ and ‘non-smoking’ groups – that would not be ethnical (or possible!). However you could recruit individuals who are already smokers to your experimental group. You could control for factors like age, SEC, gender, marital status (anything you think might be important to your outcome) by matching your ‘smoking’ participants with similar ‘non-smoking’ participants. This way you compare two groups that were matched on key variables but differed with regard to your independent variable – whether or not they smoke. This is imperfect as there may be other factors (confounding variables) that differ between the groups but it does allow you to use a form of experimental design in a natural context. This type of approach is more common in the social sciences where ethical and practical concerns make random allocation of individuals problematic.

Non-experimental designs: These designs gather substantial amounts of data in naturally occurring circumstances, without any experimental manipulation taking place. At one level the research can be purely descriptive (e.g. what is the relationship between ethnicity and student attainment?). However with careful selection and collection of data and appropriate analytic methods, such designs allow the use of **statistical control** to go beyond a purely descriptive approach (e.g. can the relationship between ethnicity and attainment be explained by differences in socio-economic disadvantage?). By looking at relationships between the different variables it can be possible for the researcher to draw strong conclusions that generalize to the wider population, although conclusions about causal relationships will be more speculative than for experimental designs.

For example, secondary schools differ in the ability of their students on intake at age 11 and this impacts very strongly on the pupils attainment in national exams at age 16. As a result ‘raw’ differences in exam results at age 16 may say little about the effectiveness of the teaching in a given school. You can’t directly compare grammar schools to secondary modern schools because they accept students from very different baseline levels of academic ability. However if you control for pupils’ attainment at intake at age 11 you can get a better measure of the school’s effect on the *progress* of pupils. You can also use this type of statistical control on other variables that you feel are important such as socio-economic class (SEC), ethnicity, gender, time spent on homework, attitude to school, etc. All of this can be done without the need for any experimental manipulation. This type of approach and the statistical techniques that underlie it are the focus of this website.

Quantitative/Qualitative methods or Quantitative/Qualitative data?

In some ways we don’t really like to use the term ‘quantitative methods’ as it somehow suggests that they are totally divorced from ‘qualitative methods’. It is important to avoid confusing methods with data. As **Figure 1.3.2** suggests, it is more accurate to use the terms ‘quantitative’ and ‘qualitative’ to describe data rather than methods, since any method can generate both quantitative and qualitative data.

Figure 1.3.2: Research methods using different types of data

Quantitative data	Method	Qualitative data
Highly structured questions	Interviews	Loose script or guide
Closed questions	Questionnaire	Open-ended questions
Detailed coding schemes	Observation	Participant observation
Content analysis	Documents	Impressions & inferences
Standarised test score	Assessment	Formative judgement

You may be conducting face-to-face interviews with young people in their own homes (as is the case in the dataset we are going to use throughout these modules) but choose a highly structured format using closed questions to generate quantitative data because you are striving for comparable data across a very large sample (15,000 students as we shall see later!). Alternatively you may be interested in a deep contextualized account from half a dozen key individuals, in which case quantitative data would be unlikely to provide the necessary depth and context. Selecting the data needed to answer your research questions is the important thing, not selecting any specific method.

Operational measures

The hallmark of quantitative research is measurement - we want to measure our key concepts and express them in numerical form. Some data we gather as researchers in education are directly observable (biological characteristics, the number of students in a class etc.), but most concepts are unobservable or 'latent variables'. For any internal mental state (anxiety, motivation, satisfaction) or inferred characteristic (e.g. educational achievement, socio-economic class, school ethos, effective teachers etc) we have to operationalise the concept, which means we need to create observable measures of the latent construct. Hence the use of attitude scales, checklists, personality inventories, standardised tests and examination results and so on. Establishing the reliability and validity of your measures is central but beyond the scope of this module. We refer you to *Muijs (2004)* for a simple introduction and any general methods text (e.g. *Cohen et. al., 2007, Newby, 2009*) for further detail.

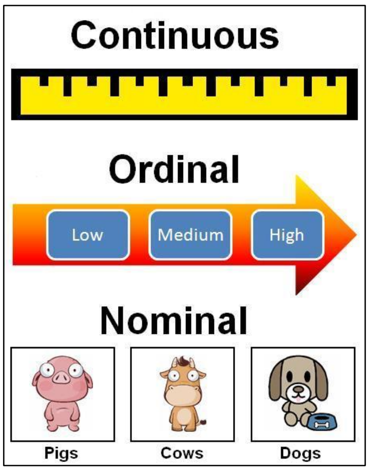
Variables and values

The construct we have collected data on is usually called the variable (e.g. gender, IQ score). Particular numbers called values are assigned to describe each variable. For example, for the variable of IQ score the values may range from 60-140. For the variable gender the values may be 0 to represent 'boy' and 1 to represent 'girl', essentially assigning a numeric value for each category. Don't worry, you'll get used to this language as we go through the module!

Levels of measurement

As we have said, the hallmark of quantitative research is measurement, but not every measurement is equally precise: saying someone is 'tall' is not the same as saying someone is 2.0 metres. **Figure 1.3.3** shows us that quantitative data can come in three main forms: continuous, ordinal and nominal.

Figure 1.3.3: Levels of quantitative measurement



Apologies for the slightly childish cartoon animals, we just liked them! Particularly the pig - he looks rather alarmed! Perhaps somebody is trying to make him learn something horrible... like regression analysis.

Nominal data is of a categorical form with cases being sorted into discrete groups. These groups are also mutually exclusive; each case has to be placed in one group only. Though numbers are attached to these categories for analysis the numbers themselves are just labels - they simply represent the name of the category. Ethnicity is a good example of a nominal variable. We may use numbers to identify different ethnic groups (e.g. 0= White British, 1= mixed heritage, 2=Indian, 3=Pakistani etc) but the numbers just represent or stand for group membership, '3' does not mean Pakistani students are three times more of ethnicity than White British students!

Ordinal data is also of a categorical form in which cases are sorted into discrete groups. However, unlike nominal data, these categories can be placed into a meaningful order. Social economic class is a good example of this. Different social economic groups are ranked based on how relatively affluent they are but we do not have a precise measure of how different each category is from one another. Though we can say people from the 'higher managerial' group are better off than those from the 'routine occupations' group we do not have a measure of the size of this gap. The differences between each category may vary.

Continuous data (scale) is of a form where there is a wide range of possible values which can produce a relatively precise measure. All the points on the scale should be separated by the same value so we can ascertain exactly how different two cases are from one another. Height is a good example of this. Somebody who is 190cm tall is 10cm taller than somebody who is 180cm tall. It is the exact same difference as between someone who is 145cm tall and someone who is 155cm tall. This may sound obvious (actually that part is obvious!) but although collecting data which is continuous is desirable surprisingly few variables are quantified in such a powerful manner! Test score is a good example of a scale variable in education.

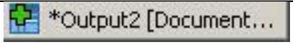
All of these levels of data can be quantified and used in statistical analysis but must usually be treated slightly differently. It is important to learn what these terms mean now so that they do not return to trip you up later! *Field (2009), pages 7-10* discusses the types of data further (see the **Resources** page).

1.4 SPSS: An introduction

This section will provide a brief orientation of the SPSS software. Don’t worry; we’re not going to try to replicate the user manual - just run you through the basics of what the different windows and options are. Note that you may find that the version of SPSS you are using differs slightly from the one we use here (we are using Version 17). However the basic principles should be the same though things may look a little different. The best way to learn how to use software is to play with it. SPSS may be less fun to play with than a games console but it is more useful! Probably...

If you would like a more in depth introduction to the program we refer you to *Chapter 3 of Field (2009)* or the *Economic and Social Data Service (ESDS) guide to SPSS*. Both of these are referenced in our **Resources**. SPSS also has a *Help* function which allows you to search for key terms. It can be a little frustrating and confusing at times but it is still a useful resource and worth a try if you get stuck. Okay let’s show you around... Why not open up the LSYPE 15,000 dataset and join us on our voyage of discovery? We have a few examples that you can work through with us.

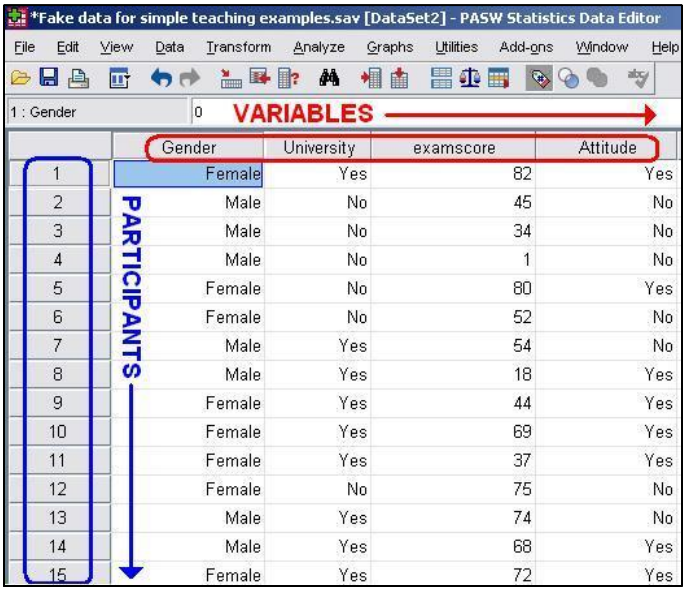
There are two main types of window, which you will usually find open together on your computer’s task bar (along the bottom of the screen). These windows are the *Data Editor* and the *Output Viewer*.

Data Editor	Output Viewer
	

The Data Editor

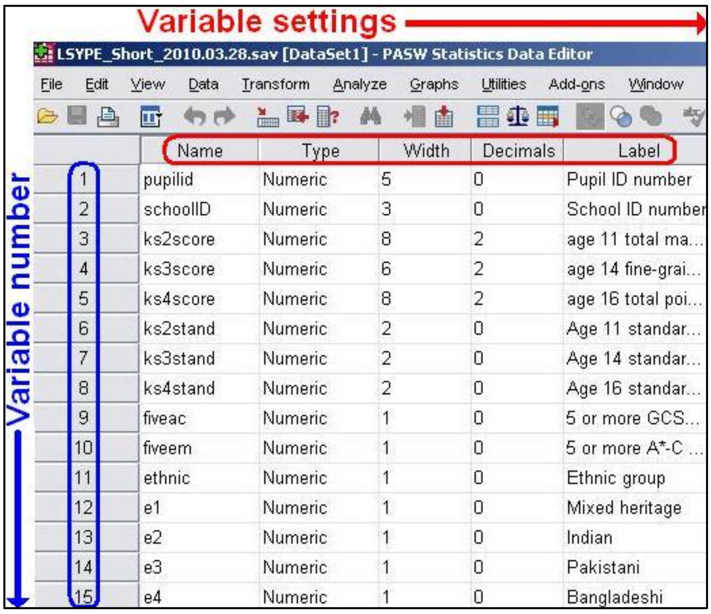
The data editor is a spreadsheet in which you enter your data and your variables. It is split in to two windows: *Data View* and *Variable View.* You can swap between them using the tabs in the bottom left of the Data Editor.

> ***Data View****:* Each row represents one case (unit) in your sample (this is usually one participant but it could be one school or any other single case). Each column represents a separate variable. Each case’s value on each variable is entered in the corresponding cell. So it is just like any 2 x 2 (row by column) spreadsheet!


*Fake data for simple teaching examples.sav [DataSet2] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : Gender
0
VARIABLES
PARTICIPANTS

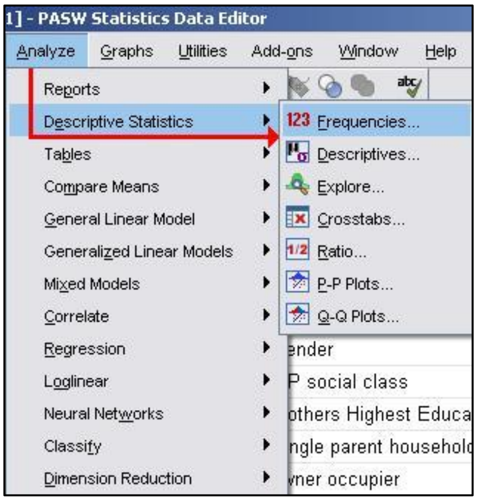
	Gender	University	examscore	Attitude
1	Female	Yes	82	Yes
2	Male	No	45	No
3	Male	No	34	No
4	Male	No	1	No
5	Female	No	80	Yes
6	Female	No	52	No
7	Male	Yes	54	No
8	Male	Yes	18	Yes
9	Female	Yes	44	Yes
10	Female	Yes	69	Yes
11	Female	Yes	37	Yes
12	Female	No	75	No
13	Male	Yes	74	No
14	Male	Yes	68	Yes
15	Female	Yes	72	Yes

***Variable View**:* This view allows you to alter the settings of your variables with each row representing one variable. Across the columns are different settings which you can alter for each variable by going to the corresponding cell. These settings are characteristics of the variable. You can add labels, alter the definition of the level of measurement of the data (e.g. nominal, ordinal, scale) and assign numeric values to nominal or ordinal responses (more on this in **Page 1.7**).


Variable settings
LSYPE_Short_2010.03.28.sav [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window
Variable number

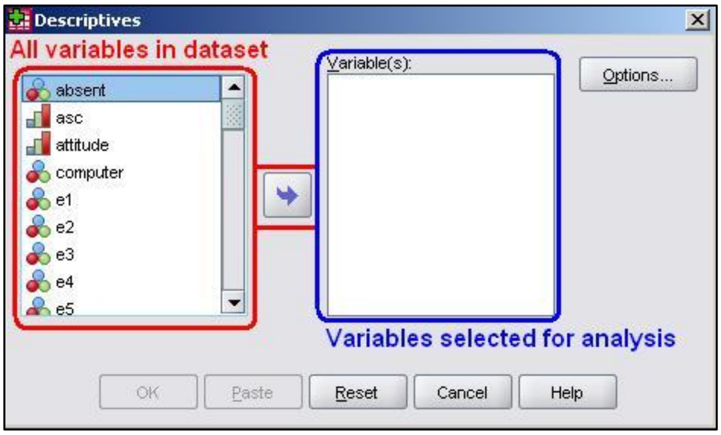
	Name	Type	Width	Decimals	Label
1	pupilid	Numeric	5	0	Pupil ID number
2	schoolID	Numeric	3	0	School ID number
3	ks2score	Numeric	8	2	age 11 total ma...
4	ks3score	Numeric	6	2	age 14 fine-grai...
5	ks4score	Numeric	8	2	age 16 total poi...
6	ks2stand	Numeric	2	0	Age 11 standar...
7	ks3stand	Numeric	2	0	Age 14 standar...
8	ks4stand	Numeric	2	0	Age 16 standar...
9	fiveac	Numeric	1	0	5 or more GCS...
10	fiveem	Numeric	1	0	5 or more A*-C ...
11	ethnic	Numeric	1	0	Ethnic group
12	e1	Numeric	1	0	Mixed heritage
13	e2	Numeric	1	0	Indian
14	e3	Numeric	1	0	Pakistani
15	e4	Numeric	1	0	Bangladeshi

The lists of options at the top of the screen provide menus for managing, manipulating, graphing, and analysing your data. The two most frequently used are probably *Graphs* and *Analyze*. They open up cascading menus like the one below:



Analyze is the key for performing regression analyses as well as for gaining descriptive statistics, tabulating data, and exploring associations between variables. The *Graphs* menu allows you to draw the various plots, graphs and charts necessary to explore and visualize your data*.*

When you are performing analyses or producing other types of output on SPSS you will often open a pop-up menu to allow you to specify the details. We will explore the available options when we come to discuss individual tasks but it is worth noting a few general features.

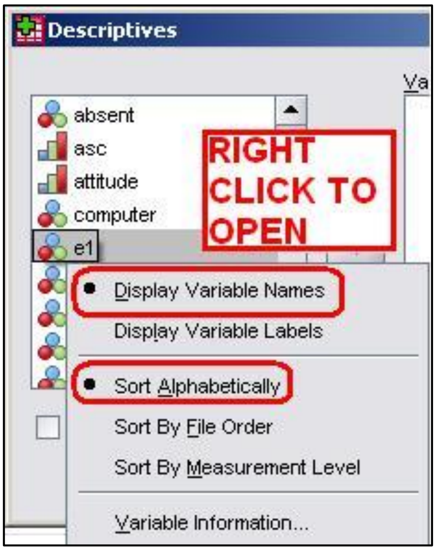


On the left of the pop-up window you will see a list of all the variables in your dataset. You will usually be required to move the variables you are interested in across to the relevant empty box or boxes on the right. You can either drag and drop the variable or highlight it and then move it across with the arrow:

You will become very familiar with these arrows and the menu windows in general the more you use SPSS!

On the far right there are usually buttons which allow you to open further submenus and tinker with the settings for your analysis (e.g. *Options*, as above). The buttons at the bottom of the window perform more general functions such as accessing the *Help* menu, starting again or correcting mistakes (*Reset* or *Cancel*) or, most importantly, running the analysis (*OK*). Of course this description is rather general but it does give you a rough indication of what you will encounter.

It is useful to note that you can alter the order that your list of available variables appear in along with whether you see just the variable names or the full labels by right clicking within the window and selecting from the list of options that appears (see below). This is a useful way of finding and keeping track of your variables! We recommend choosing 'Display Variable Names' and 'Sort Alphabetically' as these options make it easier to see and find your variables.



RIGHT CLICK TO OPEN

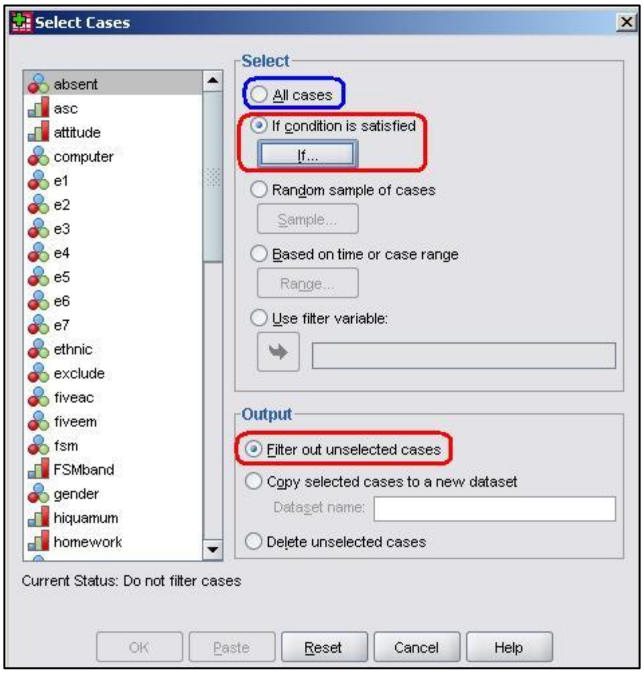
On the main screen there are also a number of buttons (icons) which you can click on to help you keep track of things. For example, you can use the pair of snazzy *Find* binoculars to search through your data for particular values (you can do this with a focus on individual variables by clicking on the desired column). The *label* button allows you to switch between viewing the numerical values for each variable category and the text label that the value represents (for ordinal and nominal variables). You can also use the *select cases* button if you want to examine only specific units within your sample.

Find	**Label**	**Select cases**

That last one can be important so let's take a closer look...

Selecting Cases

Clicking on the *Select Cases* button (or accessing it through the menus, **Data > Select Cases**) opens up the following menu:



Select
○ All cases
● If condition is satisfied
If...
○ Random sample of cases
Sample...
○ Based on time or case range
Range...
○ Use filter variable:

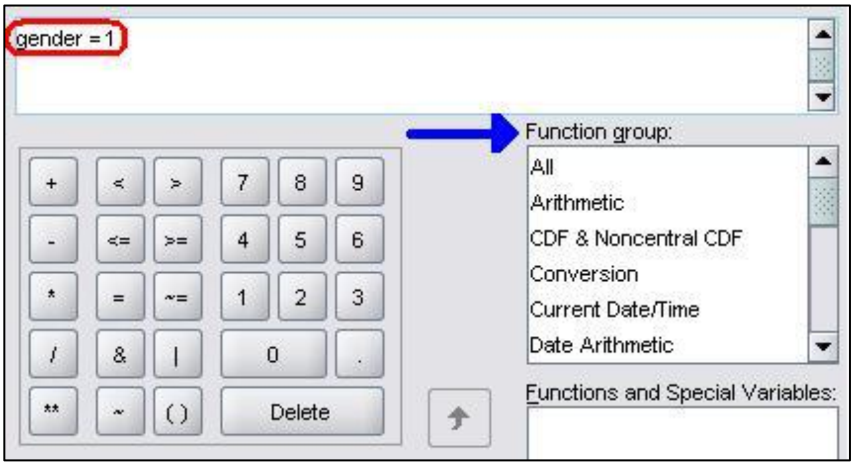
Output
● Filter out unselected cases
○ Copy selected cases to a new dataset
Dataset name:
○ Delete unselected cases

Current Status: Do not filter cases

This menu allows you to select specific cases for you to run analysis on. Most of the time you will simply have the *All cases* option selected as you will want to examine the data from all of your participants. However, on occasion it may be that you only want to look at a certain sub-sample of your data (all of the girls for example), in this case the *If* option will come into play (more on this soon!). In addition you can select a sub-sample completely at random (*Random sample of cases*) or select groups based on the order in which they are arranged in the data set (*Based on time or case range*). These last two options are rarely used but they are worth knowing about.

It is also important to note that you have a number of options regarding how to deal with your selection of cases (your sub-sample). The *Output* options allow you to choose what happens to the cases that you select. The default option '*Filter out unselected cases'* is best – all this does is temporary exclude unselected cases, placing a line through them in the data editor. They are not deleted - you can reintroduce them again through the select cases menu at any time. '*Copy selected cases to a new dataset'* can be useful if you will be working with a specific selection in detail and want to store them as a separate dataset. Your selected cases will be copied over to a new data editor window which you can save separately. Finally the option to '*Delete unselected cases'* is rather risky – it permanently removes all cases you did not select from the dataset. It could be useful if you have a huge number of cases that needs trimming down to a manageable quantity but exercise caution and have backup files of the original unaltered dataset. If, like us, you tend to make mistakes and/or change your mind frequently then we recommend you avoid using this option all together!

The most commonly used selection option is the *If* menu so let's take a closer look at it. To be honest the *If* menu (shown in part below) terrifies us! This is mainly because of the scientific calculator keypad and the vast array of arithmetic functions that are available on the right. The range of options available is truly mind-blowing! We will not even attempt to explain these options to you as most of them rarely come into use. However we have highlighted our example in the image.



gender =1

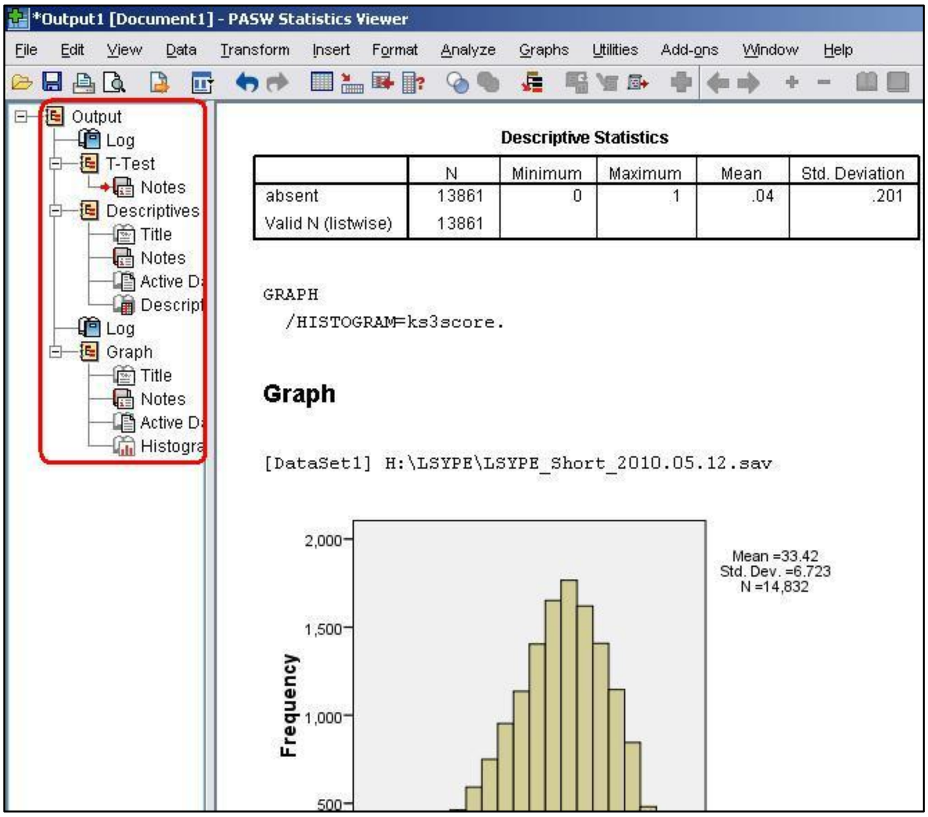
Most uses of the *If* menu really will be this simple. Girls are coded as '1' in the LSYPE dataset. If we wish to select only girls for our analyses we need to tell SPSS to select a case only if the gender variable has a value of '1'. So in order to select only girls we simply put 'gender =1' in the main input box and click *Continue* to return to the main *Select Cases* menu. This is a simple example but the principles are simple. We only briefly describe these functions here but you can calculate almost any 'if' situation using this menu. It is worth exploring the possibilities yourself to see how the '*If*' menu can best serve you! This calculator like setup will also appear in the *Compute* option which we discuss later (**Page 1.7)**, so we are coming back to it if you are confused.

Once back at the main *Select Cases* menu simply click **OK** to confirm your settings and SPSS will do the rest. Remember to change it back when you are ready to look at the whole sample again!

The Output Viewer

The output viewer is where all of the statistics, charts, tables and graphs that you request will pop into existence. It is a scary place to the uninitiated... Screen spanning 'pivot tables' which are full of numbers rounded to three decimal places. Densely packed scatterplots which appear to convey nothing but chaos. Sentences that are written in a bizarre computer language that appear to make absolutely no sense whatsoever (For example: 'DESCRIPTIVES VARIABLES = absent STATISTICS = MEAN STDDEV MIN MAX'... yes SPSS, whatever you say – actually we come to learn about this so-called 'Syntax on **Page 1.7**, so hold on to your hats).

Trust us when we say that those who withstand the initial barrage of confusion will grow to appreciate the output viewer... it brings forth the detailed results of your analysis which greatly informs your research! The trick is learning to filter out the information that is not important. With regard to regression analysis (and a few other things!) this website will help you to do this. Below is an example of what the output viewer looks like.


*Output1 [Document1] - PASW Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Output
Log
T-Test
Notes
Descriptives
Title
Notes
Active D
Descript
Log
Graph
Title
Notes
Active D
Histogra

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
absent	13861	0	1	.04	.201
Valid N (listwise)	13861				

```
GRAPH
  /HISTOGRAM=ks3score.
```

Graph

```
[DataSet1] H:\LSYPE\LSYPE_Short_2010.05.12.sav
```

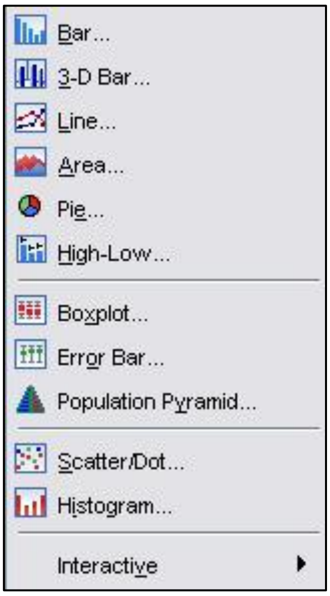
Tables and graphs are displayed under their headings in the larger portion of the screen on the right. On the left (highlighted) is an output tree which allows you to jump quickly to different parts of your analysis and to close or delete certain elements to make the output easier to read. SPSS also records a log in the output viewer after each action to remind you of the analyses you have performed and any changes you make to the dataset.

One very useful feature of the output is how easy it is to manipulate and export to a word processor. If you double-click on a table or graph an editor window opens which gives you access to a range of options, from altering key elements of the output to making aesthetic changes. These edited graphs/tables can easily be copied and pasted into other programs. There is nothing better at grabbing your reader's attention than presenting your findings in a well-designed graph! We will show you how to perform a few useful tricks with these editors on **Page 1.5** and in **Extensions C and E** but, as always, the best way to learn how to use the editor is simply to experiment!

Let us now move on to talk a little bit more making graphs.

1.5 Graphing data

Being able to present your data graphically is very important. SPSS allows you to create and edit a range of different charts and graphs in order to get an understanding of your data and the relationships between variables. Though we can't run through all of the different options it is worth showing you how to access some of the basics. The image below shows the options that can be accessed. To access this menu click on **Graphs > Legacy Dialogs >**:

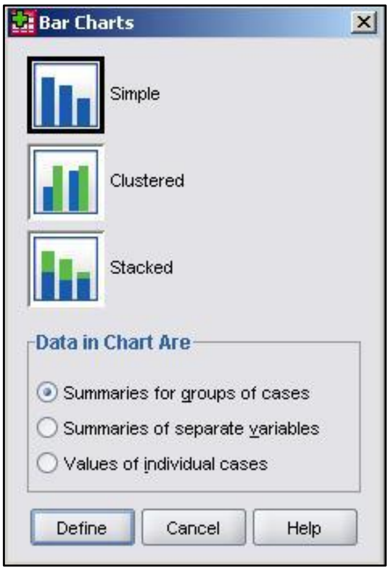


You will probably recognize some of these types of graph. Many of them are in everyday use and appear on everything from national news stories through to cereal boxes. We thought it would be fun (in a loose sense of the word) to take you through some of the LSYPE 15,000 variables to demonstrate a few of them.

<u>Bar charts</u>

Bar charts will probably be familiar to you – a series of bars of differing heights which allow you to visually compare specific categories. A nominal or ordinal variable is placed on the horizontal x-axis such that each bar represents one category of that variable. The height of each bar is usually dictated by the number of cases in that category but it can be dictated by many different things such as the percentage of cases in the category or the average (mean) score that the category has on a second variable (which goes on the horizontal y-axis).

Let's say that we want to find out how the participants in our sample are distributed across ethnic groups - we can use bar charts to visualize the percentage of students in each category of ethnicity. Take the following route through SPSS: **Graphs > Legacy Dialogs > Bar**. A pop-up menu will ask you which type of bar chart you would like to create:

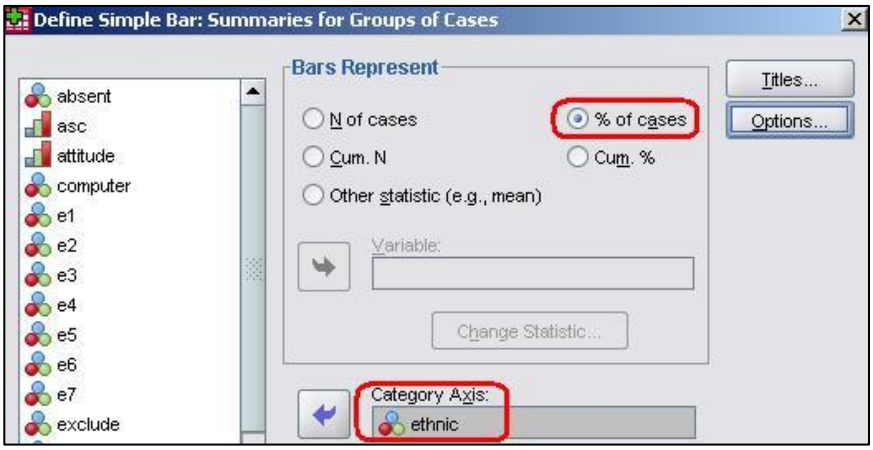

Bar Charts

Data in Chart Are

● Summaries for groups of cases
○ Summaries of separate variables
○ Values of individual cases

In this case we want the *simple* version as we only want to examine one variable. The clustered and stacked options are very useful if you want to compare bars for two variables, so they are definitely worth experimenting with.

We could also alter the '*Data in Chart Are*' options using this pop-up window. In this case the default setting is correct because we wish to compare ethnic groups and each category is a group of individual cases. There may be times when we wish to compare individuals rather than groups or even summaries of different variables (for example, comparing the mean of age 11 exam scores to the mean of age 14 scores) so it is worth keeping these options in mind. SPSS is a flexible tool. When you're happy, click **Define** to open the new window:


Define Simple Bar: Summaries for Groups of Cases

Bars Represent

○ N of cases
● % of cases
○ Cum. N
○ Cum. %
○ Other statistic (e.g., mean)

Variable:

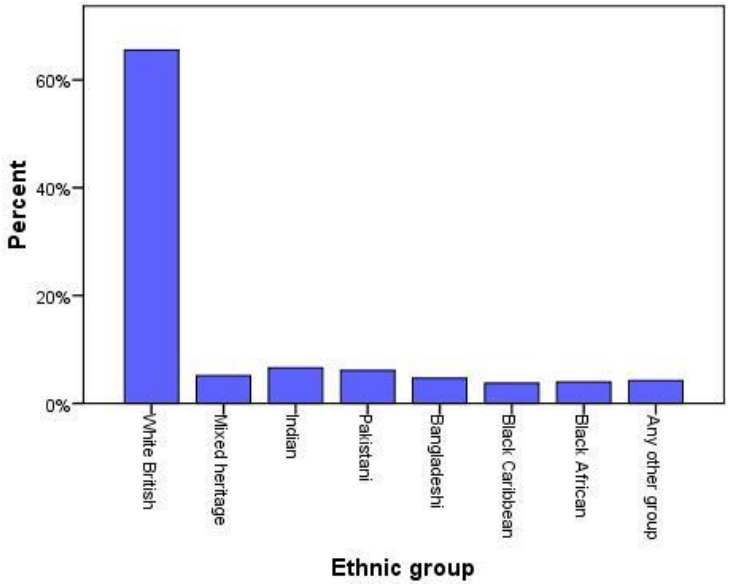
Category Axis: ethnic

The '*Bars represent*' section allows you to select whether you want each bar to signify the total number (N) of cases in the category or the percentage of cases. You can also look at how cases accumulate across the categories (*Cum. N* and *Cum. %*) or compare your categories across another statistic (their mean score on another variable, for example). In this instance we wish to look at the percentage of cases so click on the relevant option (highlighted in red).

The next thing we need to do is tell SPSS which variable we want to take as our categories. The list on the left contains all of the variables in our dataset. The one labelled *ethnic* is the one we're after and we need to move it into the box marked '*Category axis*'.

When you are happy with the settings click **OK** to generate your bar graph:

Figure 1.5.1: Breakdown of students by ethnic group



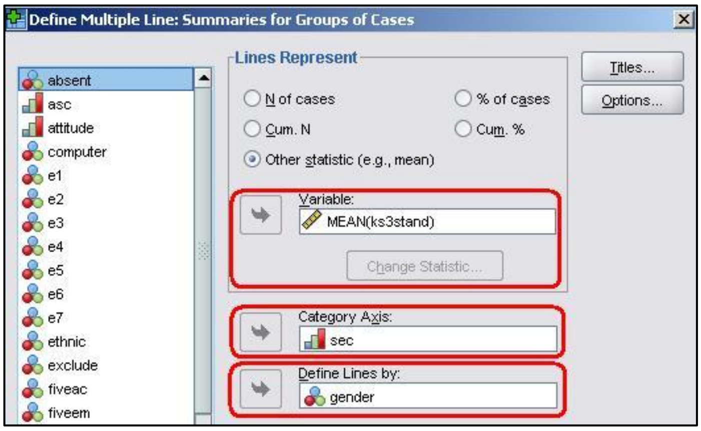
As you can see all categories were represented but the most frequent category was clearly White British, accounting for more than 60% of the total sample. Note how our chart looks somewhat different to the one in your output. We're not cheating... we simply unleashed our artistic side using the **chart editor**. We discuss the chart editor and how to alter the presentation of your graphs and charts in **Extension C**. It is a very useful tool for improving the presentation of your work and sometimes for clarifying your analysis by making certain effects easier to see.

Line charts

The line chart is useful for exploring how different groups fluctuate across the range of scores (or categories) of a given variable within your dataset. It is hard to explain in words (which are why graphs are so useful!) so let's launch straight in to an example. Let's look at socio-economic status (*sec*) but this time compare the different groups on their achievement in exams taken at age 14 (*ks3stand*). We also want to see if males and females are different in this regard.

This time take the route **Graphs > Legacy Dialogs > Line**. You will be presented with a similar pop-up menu to before. We will choose to have *Multiple* lines this time:

As before we want to select *'summaries for groups of cases'.* Click **Define** when you are happy with the setup to open the next option menu. This time we are doing something slightly different as we want to represent three variables in our chart.



Lines Represent

○ N of cases ○ % of cases

○ Cum. N ○ Cum. %

● Other statistic (e.g., mean)

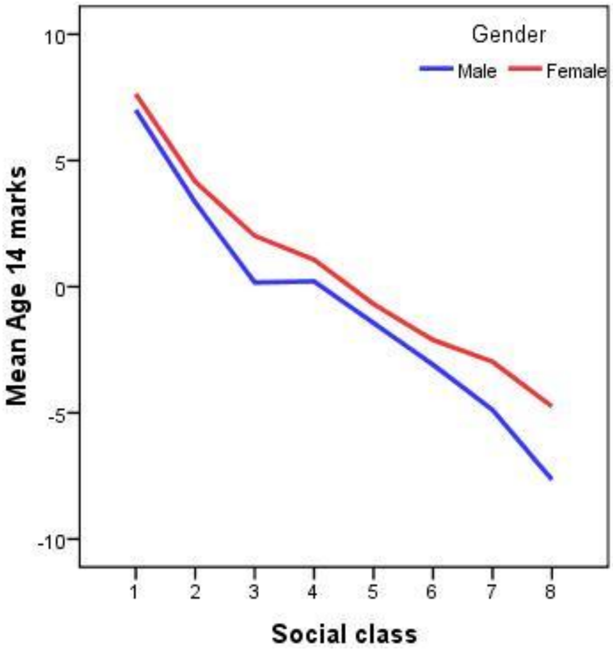
Variable: MEAN(ks3stand)

Category Axis: sec

Define Lines by: gender

You will notice that the '*Lines Represent*' section provides identical options to those that were offered for bar graphs. Once again this section basically dictates what the vertical (y-axis) will represent. For this example we want it to represent the average exam score at age 14 for each group so select '*other statistic*' and move the variable *ks3stand* from the list on the left into the box marked *Variable*. You can select a variety of summary statistics instead of the mean using the '*Change Statistic*' button located below the variable box but more often than not you will want to use the default option of the mean (if you are uncomfortable with the concept of the mean do not worry – we discuss it in more detail on **page 1.8**). The variable *sec* goes in the box marked '*Category Axis*'. This time we are going to break the output down further by creating separate lines for males and females – simply move the variable *gender* into the '*Define Lines by*' box. Click **OK** to conjure your line graph into existence, as if you were a statistics obsessed wizard.

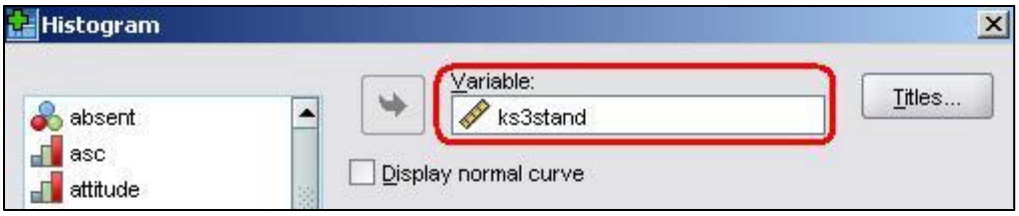
Figure 1.5.2: Line chart of age 16 exam score by gender and maternal education



The line chart shows how average scores at age 14 for both males and females are associated with SEC (the category number decreases as the background becomes less affluent). Students from more affluent backgrounds tend to perform better in their age 14 exams. There is also a gender difference, with females getting better exam scores than males in all categories of SEC. What a useful graph!

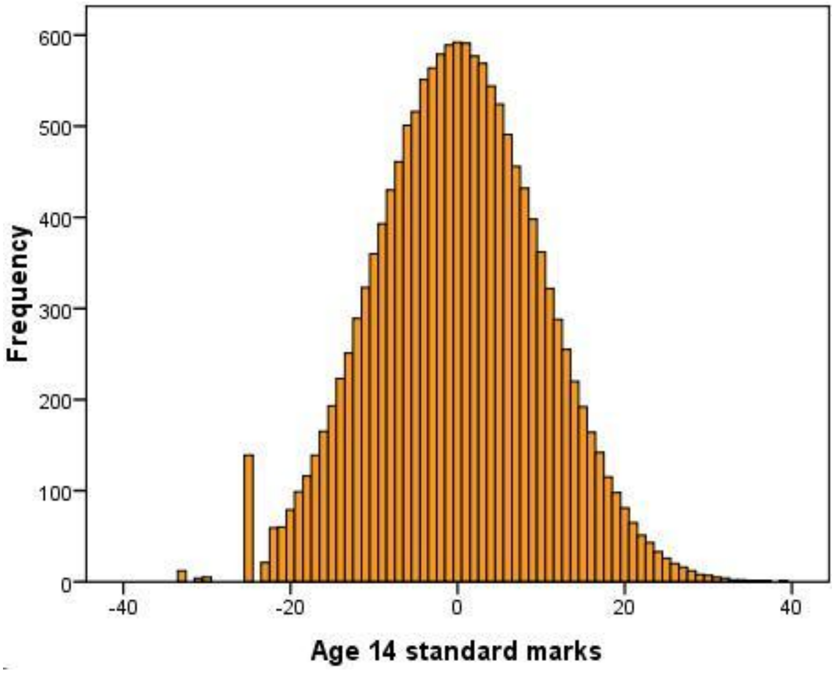
Histograms

Histograms are a specific type of bar chart but they are used for several purposes in regression analysis (which we will come to in due course) and so are worth considering separately. The histogram creates a frequency distribution of the data for a given variable so you can look at the pattern of scores along the scale. Histograms are only appropriate when your variable is continuous as the process breaks the scale into intervals and counts how many cases fall into each interval to create a bar chart. Let's show you by creating a histogram for the age 14 exam scores. Taking the route **Graphs > Legacy Dialogs > Histograms** will open the following menu:



We are only interested in graphing one variable, *ks3stand*, so simply move this into the variable box. There are options to 'panel' your graphs but these are usually only useful if you are trying to directly compare two frequency distributions. The '*Display normal curve*' tick box option is very useful if you are using your graph to check whether or not your variable is normally distributed. We will come to this later (**Page 1.8**). Click **OK** to produce the histogram:

Figure 1.5.3: Histogram of Age 14 Exam scores



The frequency distribution seems to create a bell shaped curve with the majority of scores falling at and around ‘0’ (which is the average score, the mean). There are relatively few scores at the extremes of the scale (-40 and 40).

We will stop there. We could go through each of the graphs but it would probably become tedious for you as the process is always similar! We have encouraged you to use the **Legacy Dialogs** option and haven’t really spoken about is the **Chart Builder**. This is because the legacy options are generally more straight forward for the beginner. That said the chart builder is more free form, allowing you to produce charts in a more creative manner, and for this reason you may want to experiment with it. We will now turn our attention on to another way of displaying your data: by using tables.

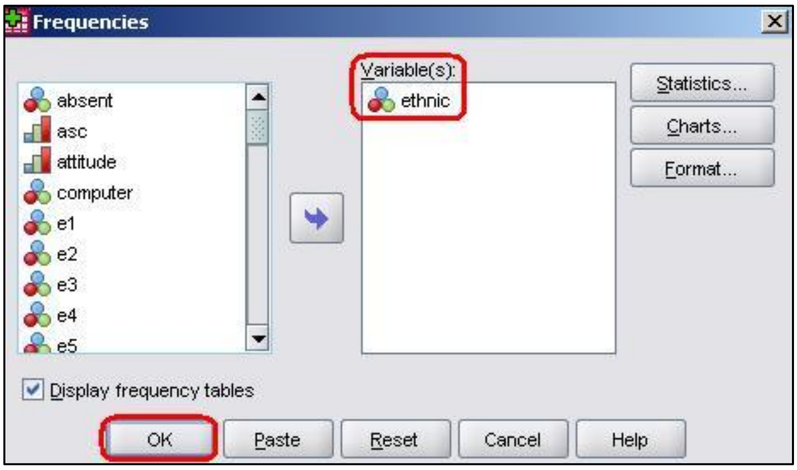
1.6 SPSS: Tabulating data

Graphing is a great way of visualizing your data but sometimes it lacks the precision which you get with exact figures. Tables are a good way of presenting precise values in an accessible and clear manner and we run through the process for creating them on this page. Why not follow us through on the LSYPE 15000 dataset?

Frequency Table

The frequency table basically shows you how many cases are in each category or at each possible value of a given variable. In other words it presents the distribution of your sample among the categories of a variable (e.g. how many participants are male compared to female or how many individuals from the sample fall into each socio-economic class category). It can only usually be used when data is ordinal or nominal – there are usually too many possible values for continuous data which results in frequency tables that stretch out over the horizon!

Let us look at the frequency table for the ethnicity variable (*ethnic*). It will be good to see how the table related to the bar chart we created on the previous page. Take the following route through SPSS **Analyse > Descriptive Statistics > Frequencies** to access the following menu:

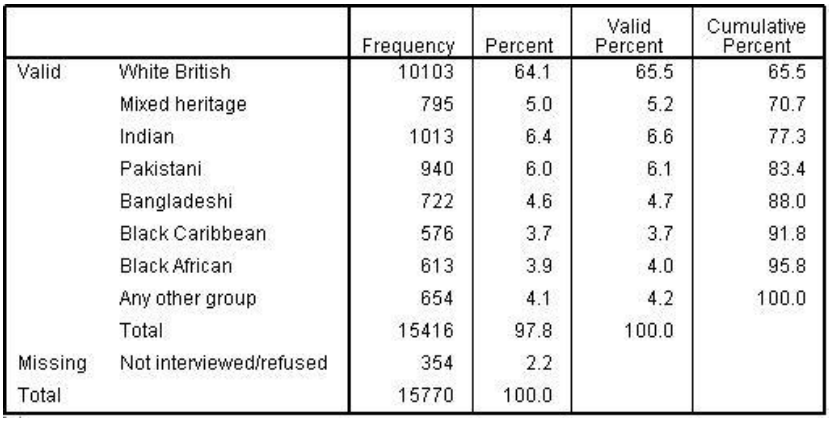


Variable(s): ethnic

☑ Display frequency tables

This is nice and simple as we will not be requesting any additional statistics or charts (you will use these options, the buttons on the right hand side of the menu box, when we come to tackle regression). Just move *ethnic* over from the list on the left into the box labelled *Variable(s)* and click **OK**.

Figure 1.6.1: Frequencies for ethnic groups



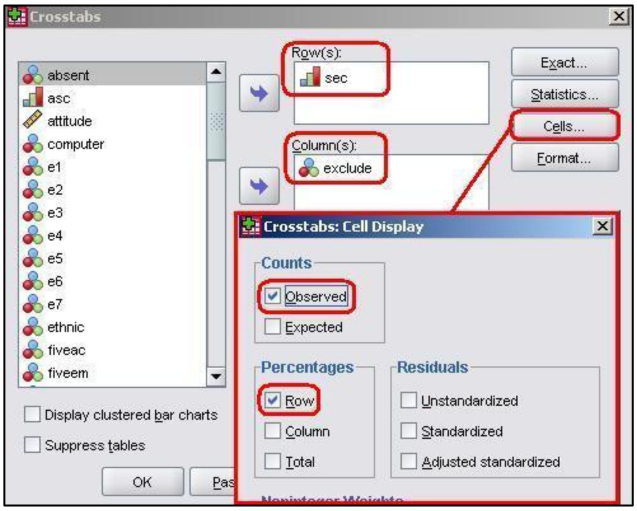
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White British	10103	64.1	65.5	65.5
	Mixed heritage	795	5.0	5.2	70.7
	Indian	1013	6.4	6.6	77.3
	Pakistani	940	6.0	6.1	83.4
	Bangladeshi	722	4.6	4.7	88.0
	Black Caribbean	576	3.7	3.7	91.8
	Black African	613	3.9	4.0	95.8
	Any other group	654	4.1	4.2	100.0
	Total	15416	97.8	100.0	
Missing	Not interviewed/refused	354	2.2		
Total		15770	100.0		

Our table shows us both the count and percentage of individual students in each ethnic group. '*Valid Percent*' is the same as '*Percent*' but excluding cases where the relevant data was missing (see our missing data **Extension B** for more on the mysteries of missing data). 'Cumulative Percent' is occasionally useful with ordinal variables as it adds each category individually from the first category to provide a rising total. Overall, it is important to understand how your data is distributed.

<u>Crosstabulation</u>

Crosstabs are a good way of looking at the association between variables and we will talk about them again in detail in the **Simple Linear Regression Module** (**Page 2.2**). They allow you to put two nominal or ordinal variables in a table together, one with categories represented by rows and the other with categories represented by columns. Each cell of the table therefore represents how many cases have the relevant combination of categories within the sample across these variables.

Let us have a look at an example! We will see how socio-economic class (*sec*) relates to whether or not a student has been excluded in the last 12 months (*exclude*). The basic table can be created using **Analyse > Descriptive Statistics > Crosstabs**. The pop-up menu below will appear.



Row(s): sec

Column(s): exclude

□ Display clustered bar charts

□ Suppress tables

Crosstabs: Cell Display

Counts

☑ Observed

□ Expected

Percentages

☑ Row

□ Column

□ Total

Residuals

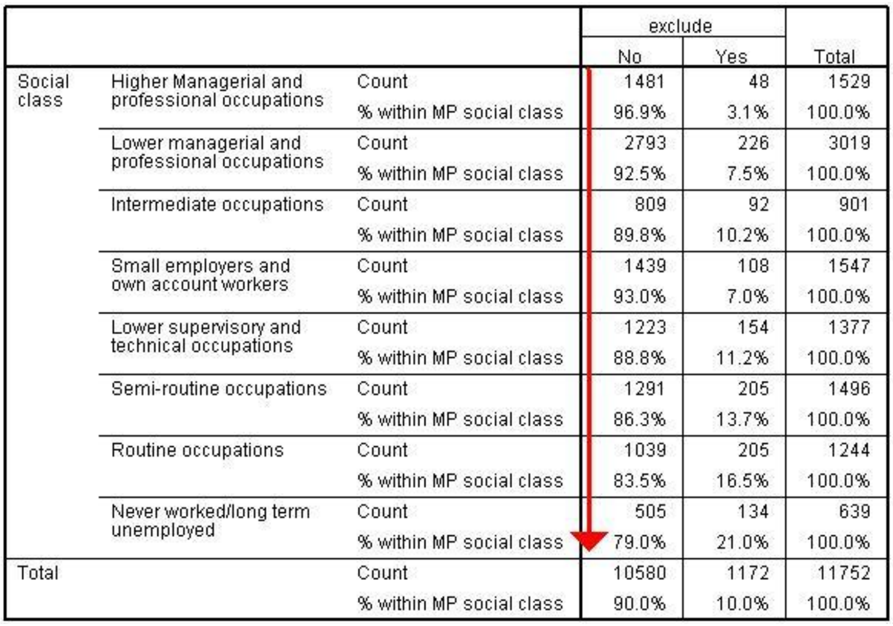
□ Unstandardized

□ Standardized

□ Adjusted standardized

As you can see we need to add two variables, one which will constitute the rows and the other the columns. Put *sec* in the box marked *Row(s)* and *exclude* in the box marked *columns*. Before continuing it is also worth accessing the *Cells* menu by clicking on the button on the right hand side. This menu allows you to include additional information within each cell of the crosstab. 'Observed' is the only default option and we will keep that – it basically tells us how many participants have the combination of scores represented by that cell. It is useful to add percentages to the cells so that you can see how the distribution of participants across categories in one variable may differ across the categories of the other. This will become clearer when we run through the example. Check *Row* in the *Percentages* section as shown above to add the percentages of students who have and have not been excluded to each category of maternal education. Click **OK** to create the table.

Figure 1.6.2: Crosstabulation of SEC and exclusion within last year



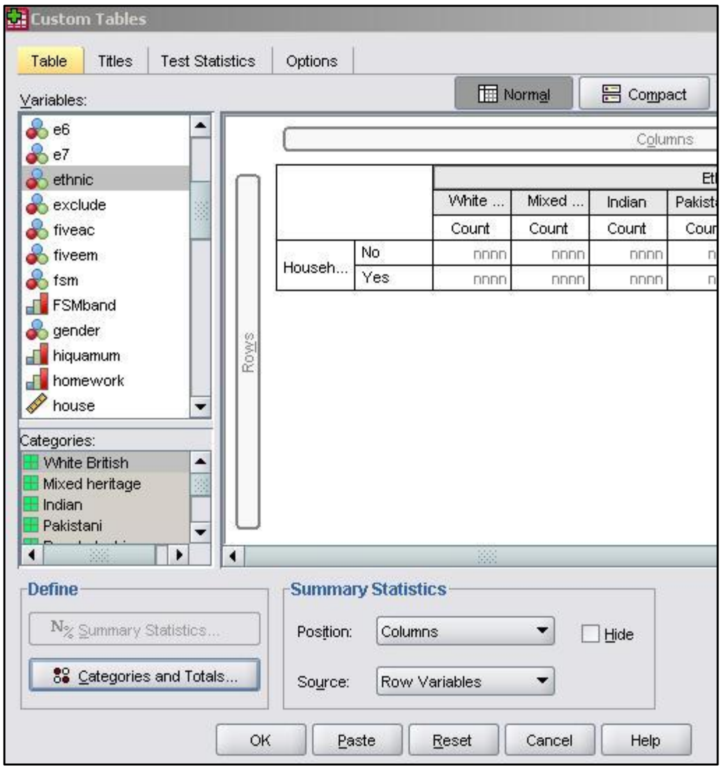
			exclude		
			No	Yes	Total
Social class	Higher Managerial and professional occupations	Count	1481	48	1529
		% within MP social class	96.9%	3.1%	100.0%
	Lower managerial and professional occupations	Count	2793	226	3019
		% within MP social class	92.5%	7.5%	100.0%
	Intermediate occupations	Count	809	92	901
		% within MP social class	89.8%	10.2%	100.0%
	Small employers and own account workers	Count	1439	108	1547
		% within MP social class	93.0%	7.0%	100.0%
	Lower supervisory and technical occupations	Count	1223	154	1377
		% within MP social class	88.8%	11.2%	100.0%
	Semi-routine occupations	Count	1291	205	1496
		% within MP social class	86.3%	13.7%	100.0%
	Routine occupations	Count	1039	205	1244
		% within MP social class	83.5%	16.5%	100.0%
	Never worked/long term unemployed	Count	505	134	639
		% within MP social class	79.0%	21.0%	100.0%
Total		Count	10580	1172	11752
		% within MP social class	90.0%	10.0%	100.0%

As you can see the 11,752 valid cases (those without any missing data) are distributed across the 16 cells in the middle of the table. By looking at the '*% within MP social class*' part of the row we can see that the less affluent the background of the family the more likely the student is to have been excluded: 21.0% of students from 'Never worked / long term unemployed backgrounds' have been excluded compared to 3.1% of students 'Higher managerial and professional backgrounds'. We will talk about associations like this more on **Page 2.3** of the **Simple Linear Regression Module** but this demonstrates how useful crosstabs can be.

Creating Custom Tables

SPSS allows you to create virtually any table using the '*Custom Tables*' menu. It is beyond the scope of the website to show you how to use this feature but we do recommend you play with it as it allows you to explore your data in creative ways and to present this exploration in an organized manner.

To get to the custom table menu go **Analyse > Tables > Custom Tables**. The custom table menu looks like this:



		Et			
		White ...	Mixed ...	Indian	Pakist
		Count	Count	Count	Cour
Househ...	No	nnnn	nnnn	nnnn	n
	Yes	nnnn	nnnn	nnnn	n

Position: Columns

Source: Row Variables

□ Hide

It is worth persevering with if there are specific tables you would like to create. Custom tables and graphs have a lot of potential!

1.7 Creating and manipulating variables

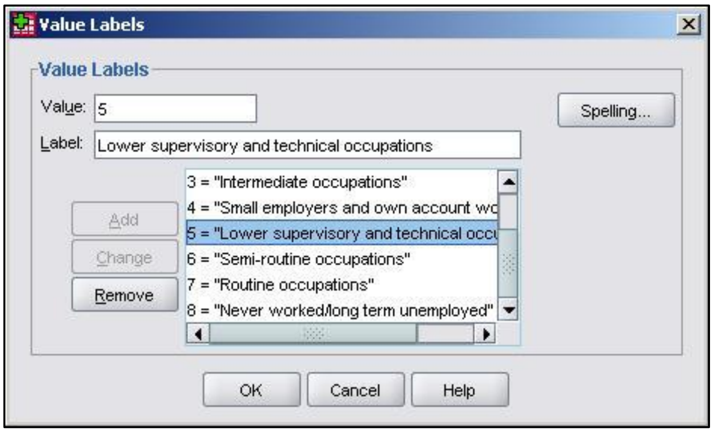
It is important that you know how to add and edit variables into your dataset. This page will talk you through the basics of altering your variables, computing new ones, transforming existing ones and will introduce you to syntax: a computer language that can make the whole process much quicker. If you would prefer a more detailed introduction you can look at the *Economic and Social Data Service SPSS Guide, Chapter 5* (see **Resources**).

Altering Variable Properties

We briefly introduced the *Variable View* on **Page 1.4** but we need to take a closer look. Correctly setting up your variables is the key to performing good analysis – your house falls down if you do not put it on a good foundation!

Each variable in your dataset is entered on a row in the *Variable View* and each column represents a certain setting or property that you can adjust for each variable in the corresponding cell. There are 10 settings:

> ***Name:*** This is the name which SPSS identifies the variable by. It needs to be short and can't contain any spaces or special characters. This inevitably results in variable names that make no sense to anyone but the researcher!
>
> ***Type:*** This is almost always set to numeric. You can specify that the data is entered as words (string) or in dates if you have a specific purpose in mind... but we have never used anything but numeric! Remember that even categorical variables are coded numerically.
>
> ***Width:*** Another option we don't really use. This allows you to restrict the number of digits that can be typed into a cell for that variable (e.g. you may only want values with two significant figures – a range of -99 to 99).
>
> ***Decimals:*** Similar to *Width*, this allows you to reduce the number of decimal places that are displayed. This can make certain variables easier to interpret. Nobody likes values like 0.8359415247... 0.84 is much easier on the eye and in most cases just as meaningful.
>
> ***Label:*** This is just a typed description of the variable, but it is actually very important! The *Name* section is very restrictive but here you can give a detailed and accurate sentence about your variable. It is very easy to forget what exactly a variable represents or how it was calculated and in such situations good labelling is crucial!
>
> ***Values:*** This is another important one as it allows you to code your ordinal and nominal variables numerically. For example you will need to assign numeric values for gender (0 = boys, 1 = girls) and ethnicity (0 = White British, 1 = Mixed Heritage, 2 = Indian, etc.) so that you can analyse them statistically. Clicking on the cell for the relevant variable will summon a pop-up menu like the one shown below.



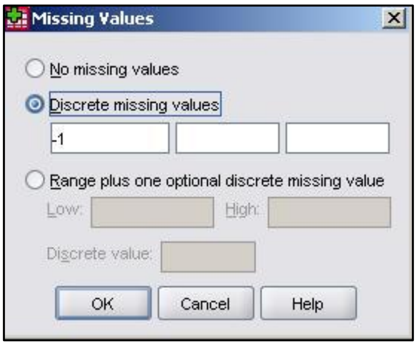
Value: 5

Label: Lower supervisory and technical occupations

3 = "Intermediate occupations"
4 = "Small employers and own account wo
5 = "Lower supervisory and technical occu
6 = "Semi-routine occupations"
7 = "Routine occupations"
8 = "Never worked/long term unemployed"

This menu allows you to assign a value to each category (level) of your variable. Simply type the value and label you want in the relevant boxes at the top of the menu and then click *Add* to place them in the main window. You can also *Change* or *Remove* the value labels you have already placed in the box. When you are satisfied with the list of value labels you have created click **OK** to finalize them. You can edit this at any time.

Missing: This setting can also be very important as it allows you to tell SPSS how to identify cases where a value is missing. This might sound silly at first – surely SPSS can assign a value as missing when a value is well... not there? Actually there are lots of different types of missing value to consider and sometimes you will want to include missing cases within your analysis (**Extension B** talks about missing data in more detail). Clicking on the cell for the relevant variable will summon the pop-up menu shown below.



○ No missing values

● Discrete missing values

-1

○ Range plus one optional discrete missing value

Low: High:

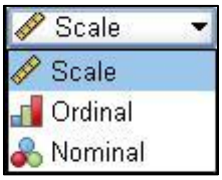
Discrete value:

You can type in up to three individual values (or a range of values) which you wish to be coded as missing and treated as such during analysis. By allowing for multiple missing values you can make distinctions between types of missing data (e.g. N/A, Do not know, left blank) which can be useful. You can give these values labels in the normal way using the *Values* setting.

Columns: This option simply dictates how wide the column for each variable is in the *Data View*. It makes no difference to the actual analysis it just gives you the option of hiding or emphasising certain variables which might be useful when you are looking at your data. We rarely use this!

Align: This is another aesthetic option which we don't usually alter. It allows you to align values to the left, right or centre of their cell.

Measure: This is where you define what type of data the variable is represented by. We discuss different types of data in detail on **Page 1.3** if you want more detail. Simply select the data type from the drop down menu in each cell (see below).



Getting the type of data right is quite important as it can influence your output in a number of ways and prevent you from performing important analyses.

This was a rather quick tour of the variable view but hopefully you know how to enter your variables and adjust or edit their properties. As we said, it is crucial that time is taken to get this right – you are essentially setting the structure of your dataset and therefore all subsequent analyses! Now you know how to alter the properties of existing variables we can move on to show you how to compute new ones.

Transforming Variables

Sometimes you may need to calculate a new variable based on raw data from other variables or you may need to transform data from an existing variable into a more meaningful form. Examples of this include:

- *Creating a general variable based on several related variables or items*:
  For example, say we were looking at our LSYPE data and are interested in whether the parent and the student BOTH aspired to continue in full time education after the age of 16 (e.g. they wanted to go to college or university). These are two different variables but we could combine them. You would simply compute a new variable that adds all the values of the other two together for each participant.
- *Collapsing the categories of a nominal or ordinal variable:*
  There are occasions when you will want to reduce the number of categories in an ordinal or nominal variable by combining ('collapsing') them. This may be because you want to perform a certain type of analysis.
- *Creating 'dummy' variables for regression* (**Module 3**, **Pages 3.4 and 3.6**): We'll show you how to do this later so don't worry about this now! However, note that dummy variables are often a key part of regression so learning how to set them up is very important.
- *Standardizing a measure* (**Extension A**)*:*
  Again, this is not something to worry about yet... but it is an important issue that will require familiarity with the recoding process.

- *Refining a variable:*
  It may be that you want to make smaller changes to a variable to make it easier to analyse or interpret. For example, you may want to round values to one decimal place (**Extension A**) or apply a transformation which turns a raw exam score into a percentage.

We'll show you the procedure for these first two examples using the LSYPE dataset, why not follow us through using LSYPE 15,000?

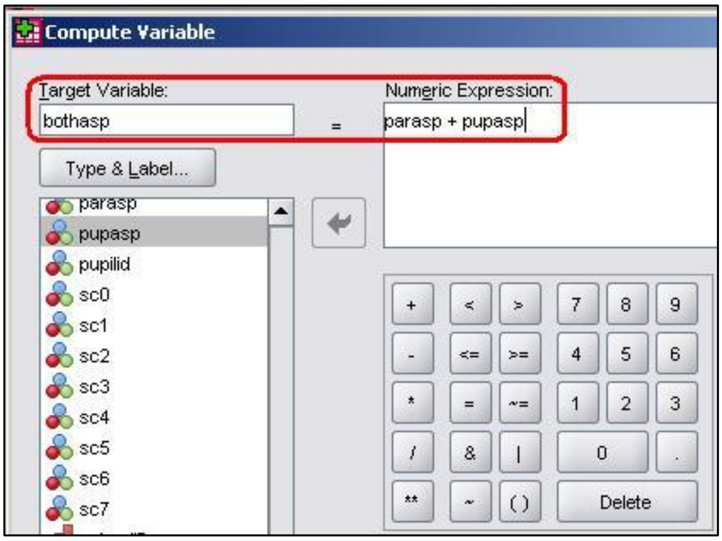
Computing variables

We use the *Compute* function to create totally new variables. For this example let's create a new variable which combines the two existing questions in the LSYPE dataset:

1) Whether or not the parent wants their child to go to full-time education after the age of 16 (the variable named *parasp* in SPSS, 0 = no; 1= yes).

2) Whether or not the student themselves want to go into full-time education post-16 (*pupasp;* 0 = no, 1= yes).

The new variable will provide us with a notion of the general educational aspirations of *both* the parents and the student themselves. We will therefore give it the shortened name in SPSS of '*bothasp*'. Let's create this new variable using the menus: **Transform > Compute**. The menu below will appear, featuring the calculator like buttons we saw when we were using the *If* menu (**Page 1.4**).

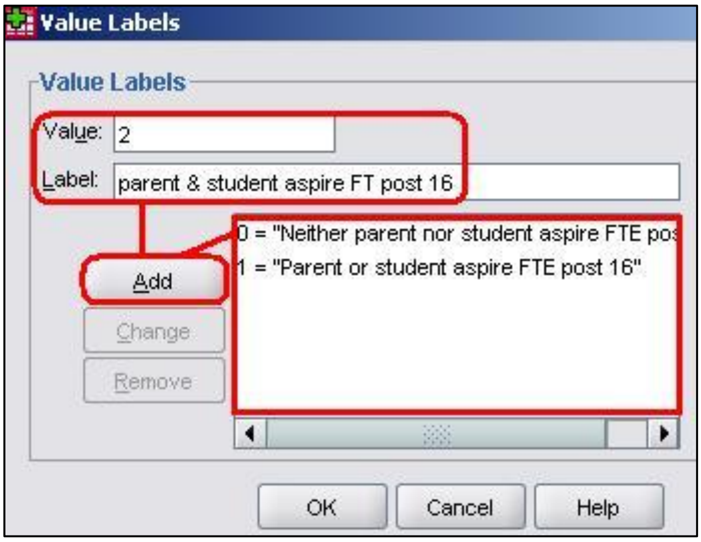


Compute Variable

Target Variable: bothasp = Numeric Expression: parasp + pupasp

The box marked *Target Variable* is for the name of the variable you wish to create so in this case we type '*bothasp*' here. We now need to tell SPSS how to calculate the new variable in the *Numeric Expression* box, using the list of variables on the left and the keypad on the bottom right. Move *parasp* from the list on the left into the *Numeric Expression* box using the arrow button, input a '+' sign using the keypad, and then add *pupasp*. Click **OK** to create your new variable...

If you switch to the *Variable View* on the main screen you will see that *bothasp* has appeared at the bottom*.* Before you begin to use it as part of your analysis remember that you will need to define its properties. It is a nominal variable not a scale variable (which is what SPSS sets as the default) and you will need to give it a label. You will also need to define *Missing* values of -1 and -2 and define the *Values* as shown:



Value: 2

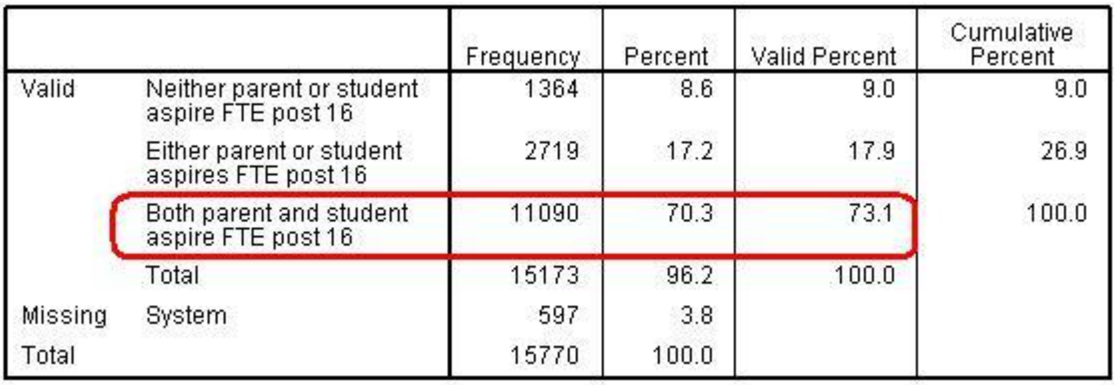
Label: parent & student aspire FT post 16

0 = "Neither parent nor student aspire FTE pos

1 = "Parent or student aspire FTE post 16"

It is worth checking that the new variable has been created correctly. To do this we can run a frequency table of our new variable (*bothasp)* and compare it to a crosstabulation of the two original variables (*parasp* and *pupasp).* See **Page 1.6** if you can't remember how to do this. **Figure 1.7.1** shows the frequency table for the *bothasp* variable. As you can see there were 11090 cases where both the pupil and the parent had aspirations for full-time education after age 16.

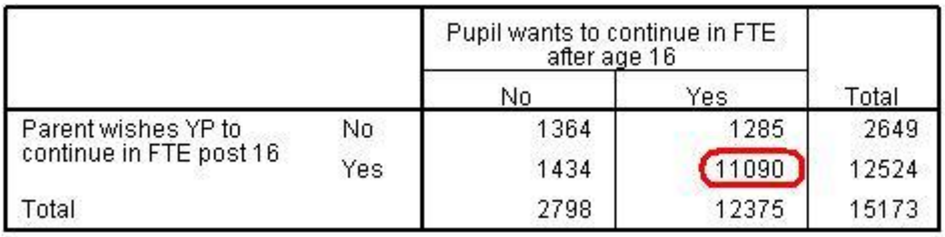
Figure 1.7.1: Frequency table for single variable Full-Time Education Aspiration



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Neither parent or student aspire FTE post 16	1364	8.6	9.0	9.0
	Either parent or student aspires FTE post 16	2719	17.2	17.9	26.9
	Both parent and student aspire FTE post 16	11090	70.3	73.1	100.0
	Total	15173	96.2	100.0	
Missing	System	597	3.8		
Total		15770	100.0		

Figure 1.7.2 show a crosstabulation of the original aspiration variables. If you look at the cell where the response to both variables was 'yes' you will see the value of 11090, which is the same value as saw when looking at the frequency of responses for the *bothasp* variable. It seems the process of computing our new variable has been successful... yay!

Figure 1.7.2: Crosstabulation for both Full-Time Education Aspiration variables

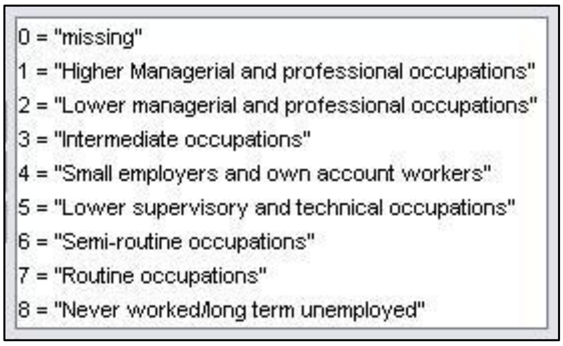


		Pupil wants to continue in FTE after age 16		Total
		No	Yes	
Parent wishes YP to continue in FTE post 16	No	1364	1285	2649
	Yes	1434	11090	12524
Total		2798	12375	15173

Once you have set up your new variable and are happy with it you can use it in your analysis!

Recoding variables

We use the *recode into same variable* or *recode into different variable* options when we want to alter an existing variable. Let's look at the example of the SEC variable. There are 8 categories for this variable, and a ninth category for missing data so the values range between 0 and 9. You can check this in the *Values* section of the variable view:



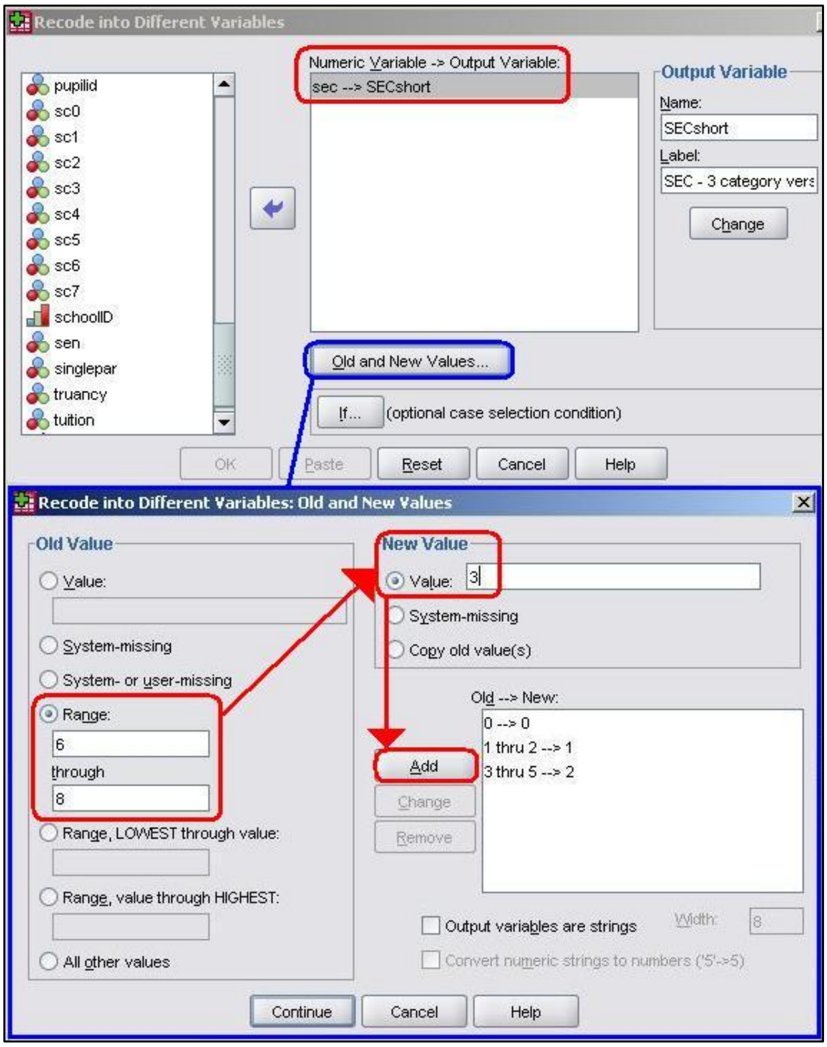
0 = "missing"
1 = "Higher Managerial and professional occupations"
2 = "Lower managerial and professional occupations"
3 = "Intermediate occupations"
4 = "Small employers and own account workers"
5 = "Lower supervisory and technical occupations"
6 = "Semi-routine occupations"
7 = "Routine occupations"
8 = "Never worked/long term unemployed"

SEC is a very important variable in the social sciences and in many circumstances this fairly fine-grained variable with 9 categories is appropriate. However sometimes large numbers of categories can overcomplicate analysis to the point where potentially important findings can be obscured. A reasonable solution is often to combine or 'collapse' categories. SEC is often collapsed to a three class version, which combines higher and lower managerial and professional (categories 1 and 2), intermediate, small employers and lower supervisory (categories 3 to 5) and semi-routine, routine and unemployed groups (categories 6 to 8). These three new categories are called (1) Managerial and professional, (2) Intermediate and (3) Routine, Semi-routine or Unemployed.

Let's do this transformation using SPSS! We want to create an adapted 3 category version of the original *SEC* variable rather than overwriting the original so we will recode into different variables: **Transform > Recode into Different Variables**. You will be presented with the pop-up menu shown below, so move the *SEC* variable into the box marked *Numeric Variable -> Output Variable*. You then need to name (and *Label,* as you would in the *Variable View*) the *Output Variable*, which we have named

SECshort (given we are essentially shortening the original SEC variable). Click the *Change* button to make it appear in the *Numeric Variable -> Output Variable* box.

We now need to tell SPSS how we want the variable transformed and to do this we click on the button marked *Old and New Values* to open up (yet another!) pop-up menu. This one requires you to recode the old values into new ones. Moving left to right you enter the old value(s) you want to change and the new value you want to represent them (as shown). We are using the *Range* option because we are collapsing multiple values so that they are represented by one value (e.g. values 1 and 2 become 1, values 3, 4 and 5 become 2, etc.) You need to click on the *Add* button after each change of value to move it into the *Old -> New* window in the bottom right.


Numeric Variable -> Output Variable:
sec --> SECshort

Output Variable
Name: SECshort
Label: SEC - 3 category vers

Old Value
● Range:
6
through
8

New Value
● Value: 3

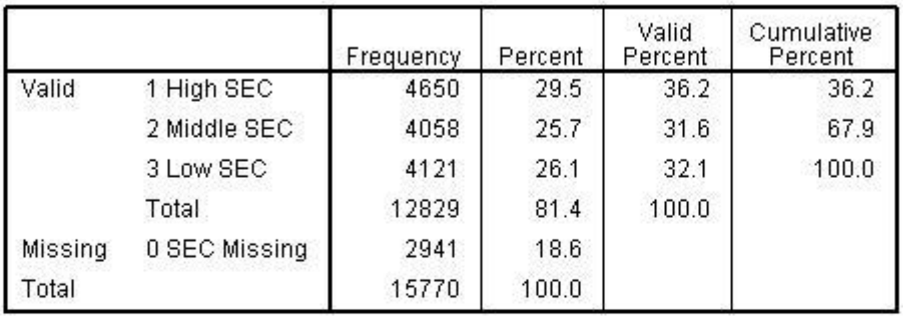
Old --> New:
0 --> 0
1 thru 2 --> 1
3 thru 5 --> 2

Recode into Different Variables
pupilid
sc0
sc1
sc2
sc3
sc4
sc5
sc6
sc7
schoolID
sen
singlepar
truancy
tuition
Change
Old and New Values...
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Recode into Different Variables: Old and New Values
○ Value:
○ System-missing
○ System- or user-missing
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values
○ System-missing
○ Copy old value(s)
Add
Change
Remove
□ Output variables are strings
Width:
8
□ Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help

Click *Continue* to shut the *Old and New Values* window and then **OK** on the main recode window to create your new variable... as before, remember to check that the properties are correct and to create *value labels* in the *Variable View*. As we will see, this new *SECshort* variable will become useful when we turn out attention to multiple regression analysis (**Module 2, Page 2.12**).

Let's generate a frequency table of our new variable to check that it looks okay (See **Page 1.6** if you need to refresh your memory about this). **Figure 1.7.3** shows that our new variable contains 3 levels as we would expect and a good spread of cases across each category. If you would like to know more about the *Office of National Statistics* SEC coding system see our **Resources** page.

Figure 1.7.3: Frequency table for 3 category SEC



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 High SEC	4650	29.5	36.2	36.2
	2 Middle SEC	4058	25.7	31.6	67.9
	3 Low SEC	4121	26.1	32.1	100.0
	Total	12829	81.4	100.0	
Missing	0 SEC Missing	2941	18.6		
Total		15770	100.0		

We have whizzed through the process of computing and recoding variables. We wanted to give you a basic grounding as it will come in handy later but realize we have only scratched the surface. As we said, if you want to know more about these processes we recommend you use some of the materials we list on our **Resources Page**, particularly the *Economic and Social Data Service SPSS Guide*.

Let us turn our attention to another pillar of SPSS: feared by some, cherished by others, it is time to meet Syntax!

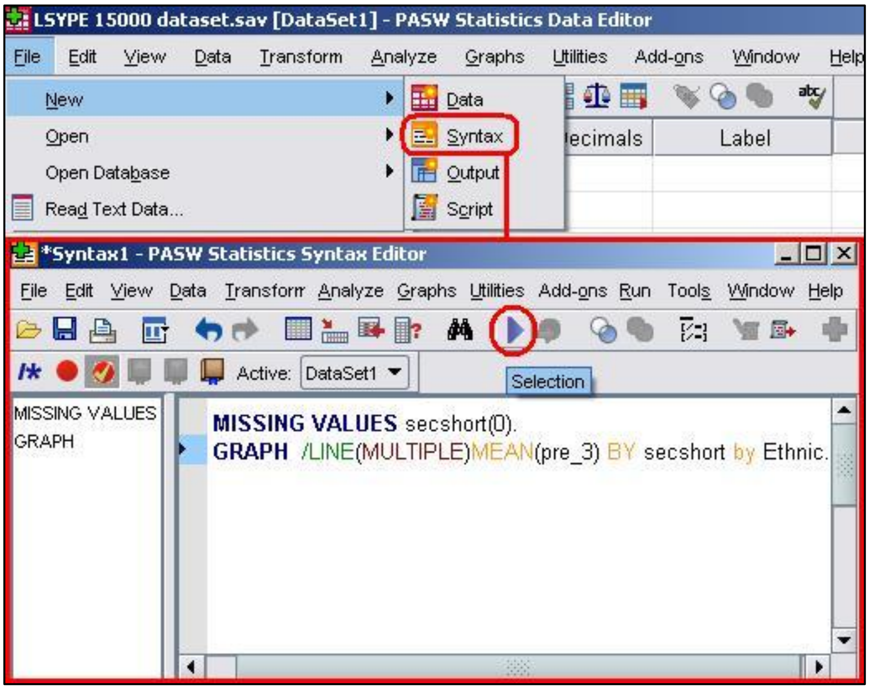
What is Syntax?

Syntax, in the context of SPSS, is basically computer language. Luckily it is quite similar to English and so is relatively easy to learn – the main difference is the use of grammar and punctuation! Basically it is a series of commands which tell SPSS what to do. Usually you enter these commands through the menus. We have already seen that this can take a while! If you know the commands and how to input them correctly then syntax can be very efficient, allowing you to repeat analyses with minor changes very quickly.

Syntax is entered and operated through the *Syntax Editor* which is a third type of SPSS window.

Syntax Editor

Syntax files can be saved and opened in the exact same way as any other file. If you want to open a new syntax window simply go **File > New > Syntax**. The image below shows you this along with an example of a Syntax window in operation.


LSYPE 15000 dataset.sav [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
New
Open
Open Database
Read Text Data...
Data
Syntax
Output
Script
ecimals
Label
*Syntax1 - PASW Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet1
Selection
MISSING VALUES
GRAPH

```
MISSING VALUES secshort(0).
GRAPH /LINE(MULTIPLE)MEAN(pre_3) BY secshort by Ethnic.
```

Syntax is ‘run’, as you would run computer code. To do this you highlight the syntax you would like to use by clicking and dragging your mouse over it in the syntax window and then clicking on the highlighted ‘*Run*’ arrow. Whatever you have requested in your syntax, be it the creation of a new variable or a statistical analysis of existing variables – will then appear in your *Data Editor* and *Output* windows.

Throughout the website we have provided SPSS Syntax files and we have occasionally provided little boxes of syntax like this one:

Syntax Alert!!!

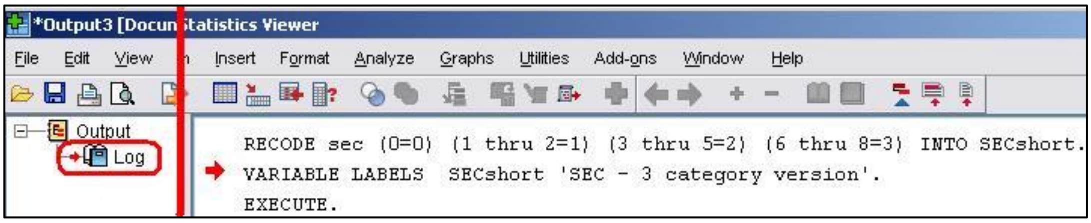
RECODE sec (0=0) (1 thru 2=1) (3 thru 5=2) (6 thru 8=3) INTO SECshort.

VARIABLE LABELS SECshort 'SEC - 3 category version'.

EXECUTE.

These boxes contain the syntax that you will need to paste into the *Syntax Editor* in order to run the related process. It may appear as though we are giving you some sort of shortcut. In a way this is true – once you have the correct syntax it is much quicker to perform processes and analyses in SPSS by using it rather than by navigating the menus. However there are other benefits too as it allows you to view more concisely the exact process that you have requested that SPSS perform.

An easy way to get hold of syntax is to copy it from the *Output Window*. Whenever you perform an action on SPSS it is interpreted as syntax and saved to the output window. There is an example below – the syntax taken from the process of recoding the SEC variable (also shown in the above syntax alert box):


*Output3 [Docun Statistics Viewer
File Edit View Insert Format Analyze Graphs Utilities Add-ons Window Help
Output
Log

```
RECODE sec (0=0) (1 thru 2=1) (3 thru 5=2) (6 thru 8=3) INTO SECshort.
VARIABLE LABELS  SECshort 'SEC - 3 category version'.
EXECUTE.
```

If you want to run the syntax again simply copy and paste it into the *Syntax Editor*. If you look at the commands you can see where you could make quick and easy edits to alter the process: VARIABLE LABELS is where the name and label are defined for example. If you wanted ‘1 thru 3’ rather than ‘1 thru 2’ to be coded as 1 you could change this easily. You may not know the precise commands for the processes but you don’t need to – run the process using the menus and examine the text to see where changes can be made. With time and perseverance you will learn these commands yourself.

Attempting to teach you how to write syntax would probably be a fruitless exercise. There are hundreds of commands and our goal is to introduce you to the concept of syntax rather than throw a reference book at you. If you want such a reference book, a recommendation can be found over in our **Resources**: try *Economic and Social Data Service SPSS Guide (Chapter 4)*. We just want you to be aware of syntax – how to operate it and how to get hold of it from your output. You do not need to worry about it but learning it in tandem with learning SPSS will really help your understanding so don’t ignore it! Let us now turn our attention to a crucial pillar in the... erm... mansion of statistics: the normal distribution.

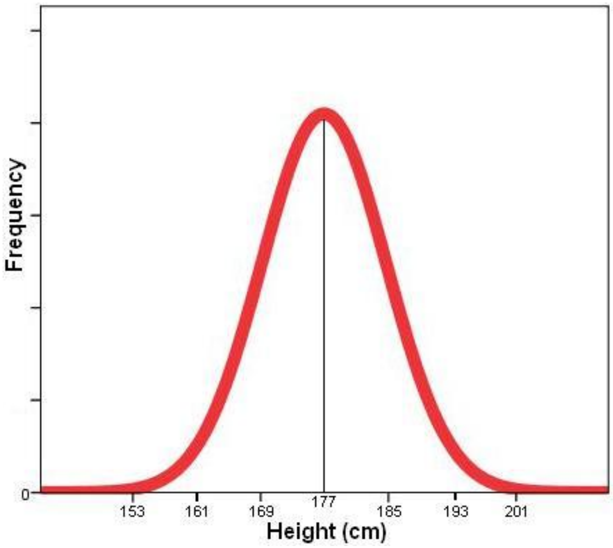
1.8 The Normal Distribution

We have run through the basics of sampling and how to set up and explore your data in SPSS. We will now discuss something called the normal distribution which, if you haven't encountered before, is one of the central pillars of statistical analysis. We can only really scratch the surface here so if you want more than a basic introduction or reminder we recommend you check out our **Resources,** particularly *Field (2009), Chapters 1 & 2* or *Connolly (2007) Chapter 5*.

The Normal Distribution

The normal distribution is essentially a frequency distribution curve which is often formed naturally by scale variables. Height is a good example of a normally distributed variable. The average height of an adult male in the UK is about 1.77 meters. Most men are not this exact height! There are a range of heights but most men are within a certain proximity to this average. There are some very short people and some very tall people but both of these are in the minority at the edges of the range of values. If you were to plot a histogram (see **Page 1.5**) you would get a 'bell shaped' curve, with most heights clustered around the average and fewer and fewer cases occurring as you move away either side of the average value. This is the normal distribution and **Figure 1.8.1** shows us this curve for our height example.

Figure 1.8.1: Example of a normal distribution 'bell' curve



Assuming that they are scale and they are measured in a way that allows there to be a full range of values (there are no ceiling or floor effects), a great many variables are naturally distributed in this way. Sometimes ordinal variables can also be normally distributed but only if there are enough categories. The normal distribution has some very useful properties which allow us to make predictions about populations based on samples. We will discuss these properties on this page but first we need to think about ways in which we can describe data using statistical summaries.

Mean and Standard Deviation

It is important that you are comfortable with summarizing your variables statistically. If we want a broad overview of a variable we need to know two things about it:

1) The ‘average’ value – this is basically the typical or most likely value. Averages are sometimes known as measures of *central tendency*.

2) How spread out are the values are. Basically this is the range of values, how far values tend to spread around the average or central point.

Measures of central tendency

The ‘mean’ is the most common measure of central tendency. It is the sum of all cases divided by the number of cases (see formula). You can only really use the Mean for continuous variables though in some cases it is appropriate for ordinal variables. You cannot use the mean for nominal variables such as gender and ethnicity because the numbers assigned to each category are simply codes – they do not have any inherent meaning.

Mean: = ——————

Note: N is the total number of cases, x1 is the first case, x2 the second, etc. all the way up to the final case (or nth case), xn.

It is also worth mentioning the ‘median’, which is the middle category of the distribution of a variable. For example, if we have 100 students and we ranked them in order of their age, then the median would be the age of the middle ranked student (position 50, or the 50^{th} percentile). The median is helpful where there are many extreme cases (outliers). For example, you may often here earnings described in relation to the national median. The median is preferred here because the mean can be distorted by a small number of very high earners. Again the median is only really useful for continuous variables.

Measures of the spread of values

One measure of spread is the range (the difference between the highest and lowest observation). This has its uses but it may be strongly affected by a small number of extreme values (outliers). The inter-quartile range is more robust, and is usually employed in association with the median. This is the range between the 25th and the 75th percentile - the range containing the middle 50% of observations.

Perhaps more important for our purposes is the standard deviation, which essentially tells us how widely our values are spread around from the mean. The formula for the standard deviation looks like this (apologies if formulae make you sad/confused/angry):

Standard Deviation: s =

Note: means 'sum of'.

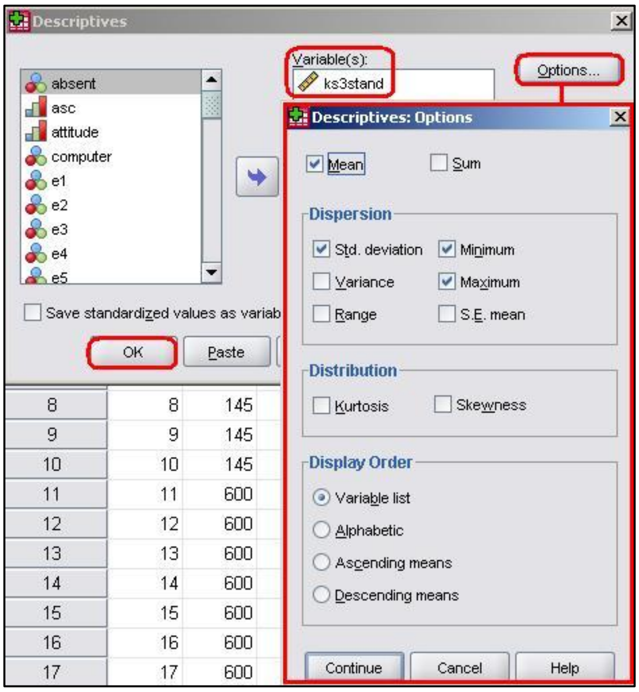
This looks more horrible than it is! Essentially all we're doing is calculating the gap between the mean and the actual observed value for each case and then summarising across cases to get an average. To do this we subtract the mean from each observed value, square it (to remove any negative signs) and add all of these values together to get a total sum of squares. We then divide this by the number of cases -1 (the '-1' is for a somewhat confusing mathematical reason you don't have to worry about yet) to get the average. This measure is often called the *variance,* a term you will come across frequently. Finally we take the square root of the whole thing to correct for the fact that we squared all the values earlier.

Okay, this may be slightly complex procedurally but the output is just the average (standard) gap (deviation) between the mean and the observed values across the whole sample. Understanding the basis of the standard deviation will help you out later.

Getting Descriptives using SPSS

Let's show you how to get these summary statistics from SPSS using an example from the LSYPE dataset (LSYPE 15,000). Let's have a closer look at the standardized age 14 exam score variable (*ks3stand*).

To access the descriptive menu take the following path: **Analyse > Descriptive Statistics > Descriptives**.



Variable(s): ks3stand

□ Save standardized values as variab

☑ Mean □ Sum

Dispersion: ☑ Std. deviation ☑ Minimum □ Variance ☑ Maximum □ Range □ S.E. mean

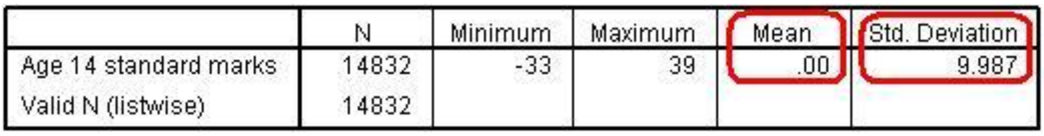
Distribution: □ Kurtosis □ Skewness

Display Order: ● Variable list ○ Alphabetic ○ Ascending means ○ Descending means

8	8	145
9	9	145
10	10	145
11	11	600
12	12	600
13	13	600
14	14	600
15	15	600
16	16	600
17	17	600

Move *ks3stand* from the list of variables on the left into the *Variables* box. We only need the default statistics but if you look in the *Options* submenu (click the button the right) you will see that there are a number of statistics available. Simply click **OK** to produce the relevant statistics (**Figure 1.8.2**).

Figure 1.8.2: Descriptive statistics for age 14 standard marks



	N	Minimum	Maximum	Mean	Std. Deviation
Age 14 standard marks	14832	-33	39	.00	9.987
Valid N (listwise)	14832				

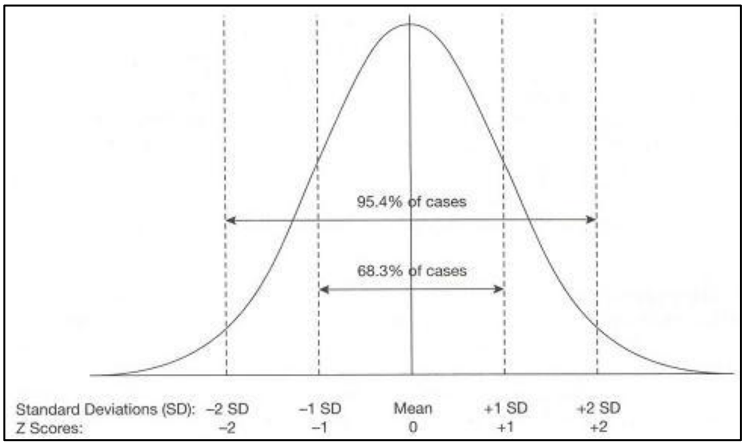
Figure 1.8.2 shows that age 14 marks range between -33 and 39 and the mean score is 0. This is because the score has been standardized – transformed in such a way that the mean score is zero and the value for each case represents how far above or below average that individual is (see **Extension A** for more about the process of standardizing variables). The standard deviation is 9.987 which means that the majority of individuals differ from the mean score by no more than plus or minus 10 points. The interpretation of standard deviation will become more apparent when we discuss the properties of the normal distribution.

Properties of the Normal Distribution

If data is normally distributed, the mean is the most commonly occurring value. The standard deviation indicates the extent to which observations cluster around the mean. Because the normally distributed data takes a particular type of pattern, the relationship between standard deviation and the proportion of participants with a given value for the variable can be calculated.

Because of the consistent properties of the normal distribution we know that two-thirds of observations will fall in the range from one standard deviation below the mean to one standard deviation above the mean. For example, for age 14 score (mean=0, SD=10), two-thirds of students will score between -10 and 10. This is very useful as it allows you to calculate the probability that a specific value could occur by chance (more on this on **Page 1.9**). **Figure 1.8.3** shows how a normal distribution can be divided up.

Figure 1.8.3: Proportion of cases by standard deviation for normally distributed data



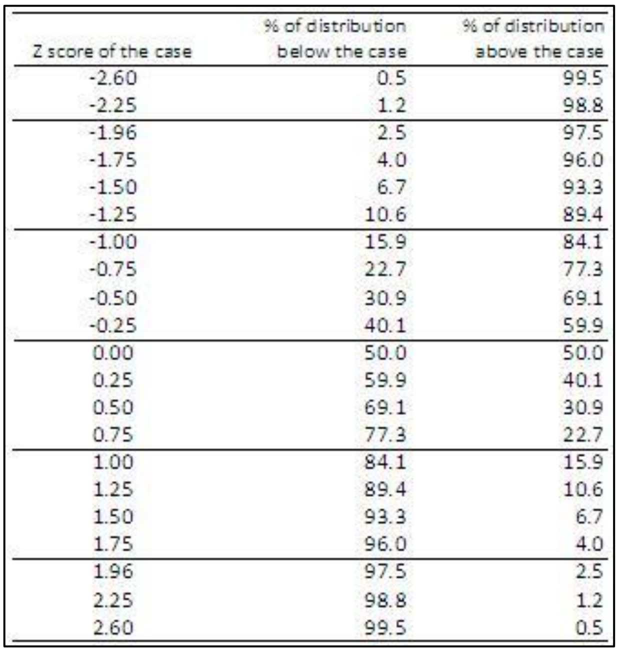
These known parameters allow us to perform a number of calculations:

- We can convert our values to a standard form where the mean=0 and the SD=1 (We do this by subtracting each value from the mean and dividing by the SD).

- Each standardized value can be assigned a *Z score*, which is a direct measure of the number of standard deviations a value is from the mean.

- The Z score gives you an idea where a case sits in a distribution whatever the metric (be it age, marks on a maths test or scores on an attitude scale). **Figure 1.8.4** is a table of these z-scores and the proportions of the population that they represent.

Figure 1.8.4: Table of Z scores



Z score of the case	% of distribution below the case	% of distribution above the case
-2.60	0.5	99.5
-2.25	1.2	98.8
-1.96	2.5	97.5
-1.75	4.0	96.0
-1.50	6.7	93.3
-1.25	10.6	89.4
-1.00	15.9	84.1
-0.75	22.7	77.3
-0.50	30.9	69.1
-0.25	40.1	59.9
0.00	50.0	50.0
0.25	59.9	40.1
0.50	69.1	30.9
0.75	77.3	22.7
1.00	84.1	15.9
1.25	89.4	10.6
1.50	93.3	6.7
1.75	96.0	4.0
1.96	97.5	2.5
2.25	98.8	1.2
2.60	99.5	0.5

For example, an individual who scores 1.0 SD below the mean will be in the lower 15.9% of scores in the sample. Someone who scores 2.6 SD above the mean will have one of the top 0.5% of scores in the sample.

Now that we have seen what the normal distribution is and how it can be related to key descriptive statistics from our data let us move on to discuss how we can use this information to make inferences or predictions about the population using the data from a sample.

1.9 Probability and Inferential statistics

We discussed populations and sampling on **Page 1.2**. As researchers we are often trying to create a model of the world around us from the data we collect and generalize this to our population of interest, making statements that we can be confident extend beyond the confines of our sample. The properties of the normal distribution allow us to cautiously make such inferences in order to test our hypotheses and calculate how confident we can be about our results. *Field (2009), Chapters 1 & 2* and *Connolly (2007) Chapter 5* from our **Resources** page might help you with this topic if our introduction is too brief.

Hypothesis Testing and Making Inferences

Inferential statistics are used to make generalisations about the characteristics of your sample, or associations between variables in your sample, to the characteristics/associations in the wider population. Such inferences require you to have a suitably large and representative sample. They also require you to make certain assumptions about your data, many of which can be directly tested.

Usually when you are conducting research you wish to test a hunch or a hypothesis that you have about a population. There are several steps for testing your hypothesis:

Steps for hypothesis testing

1. Decide on your hypothesis and then derive the null hypothesis

2. Consider the measurement level of the variables you are analysing, and select an appropriate statistical test

3. Select your confidence level

4. Conduct the test, derive and evaluate the *p-value*

Let's start by talking about hypotheses. You probably noticed that there are two types of hypothesis mentioned in these steps; your initial hypothesis (often called the *alternate hypothesis)* and something called the *null hypothesis*. In order to explain these, let us take an example of a specific research question:

Do girls have higher educational achievement than boys at age 14?

Fortunately a measure of educational achievement at age 14 is available through national tests in English, mathematics and science which can be used to create a continuous (scale) outcome variable. We can use a particular statistical test called an *independent t-test* (see **Page 1.10**) to compare the mean test score for boys with the mean test score for girls. But what do we expect to discover from this?

- **Alternate hypothesis**: There is a relationship between gender and age 14 test score.

- **Null hypothesis**: There is no relationship between gender and age 14 test score. This is the default assumption (even if you do not think it is true!).

In essence the process of hypothesis testing works like the UK legal system; you assume t*hat the effect or relationship* you are looking for does not exist, unless you can find sufficient evidence that it does… **Innocent until proven guilty!** We test to see if there is a difference between the mean scores for boys and girls in our sample and whether it is sufficiently large to be true of the population (remembering to take into account our sample size).

Imagine we find a difference in the age 14 test scores of boys and girls in our sample such that boys have, on average, lower scores than girls. This could be a fair representation of the wider population **or** it could be due to chance factors like sampling variation. There is a chance, however small, that we inadvertently selected only the boys with low attainment so our sample does not represent the whole population fairly. The independent t-test, like many statistical analyses, lets us compute a test of statistical significance to find out how likely it is that any difference in scores resulted just from sampling variation. To understand this properly you will need to be introduced to the p-value...

Statistical Significance - What is a P-value?

A p-value is a probability. It is usually expressed as a proportion which can also be easily interpreted as a percentage:

P = 0.50 represents a 50% probability or a half chance.

P = 0.10 represents a 10% probability or a one in ten chance.

P = 0.05 represents a 5% probability or a one in twenty chance.

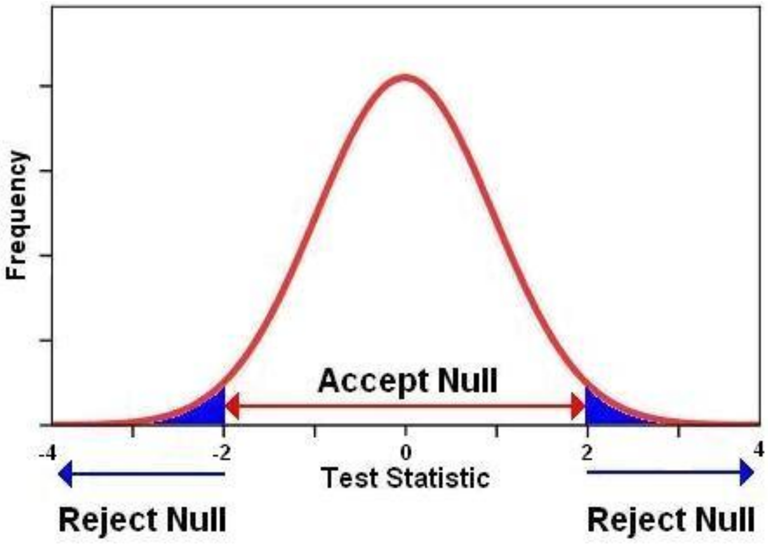
P = 0.01 represents a 1% probability or a one in a hundred chance.

P-values become important when we are looking to ascertain how confident we can be in accepting or rejecting our hypotheses. Because we only have data from a sample of individual cases and not the entire population we can never be absolutely (100%) sure that the alternative hypothesis is true. However, by using the properties of the normal distribution we can compute the probability that the result we observed in our sample could have occurred by chance. To clarify, we can calculate the probability that the effect or relationship we observe in our sample (e.g. the difference between boys and girls mean age 14 test score) could have occurred through sampling variation and in fact does not exist in the population as a whole. The strength of the effect (the size of the difference between the mean scores for boys and girls), the amount of variation in scores (indicated by the standard deviation) and the sample size are all important in making the decision (we will discuss this in detail when we report completing independent t-tests on **Page 1.10**).

Conventionally, where *there is less than a 5% probability that the results from our sample are due to chance* the outcome is considered statistically significant. Another way of saying this is that we are 95% confident there is a ‘real’ difference in our population. This is our *confidence level*. You are therefore looking for a p-value that is less than .05, commonly written as $p < .05$. Results significant at the 1% level ($p < .01$), or even the 0.1% level (*p<.001),* are often called "highly" significant and if you want to be more sure of your conclusions you can set your confidence level at these lower values. It is important to remember these are somewhat arbitrary conventions - the most appropriate confidence level will depend on the context of your study (see more on this below).

The way that the p-value is calculated varies subtlety between different statistical tests, which each generate a *test statistic* (called, for example, t, F or X^2 depending on the particular test). This test statistic is derived from your data and compared against a known distribution (commonly a normal distribution) to see how likely it is to have arisen by chance. If the probability of attaining the value of the test statistic by chance is less than 5% (p<.05) we typically conclude that the result is statistically significant. **Figure 1.9.1** shows the normal distribution and the blue 'tails' represent the standardized values (where the mean is 0 and the SD is 1) which allow you to reject the null hypothesis. Compare this to **Figure 1.8.3** and you can see that obtaining a value of less than -2 or more than 2 has a probability of occurring by chance of less than 5%. If we attain such a value we can say that our result is unlikely to have occurred by chance – it is statistically significant.

Figure 1.9.1: Choosing when to Accept and When to Reject the Null Hypothesis



In other words, if the probability of the result occurring by chance is p<.05 we can conclude that there is sufficient evidence to reject the null hypothesis at the .05 level. There is only a 5% or 1 in 20 likelihood of a difference of this size arising in our sample by chance, so is likely to reflect a 'real' difference in the population. Note that either way we can never be **absolutely certain**, these are probabilities. There is always a possibility we will make one of two types of error:

Type of Error

Type I error: When we conclude that there is a relationship or effect, but in fact there is not ***(false positive).***

Type II error: when we conclude there is no relationship or effect, when in fact there is ***(false negative).***

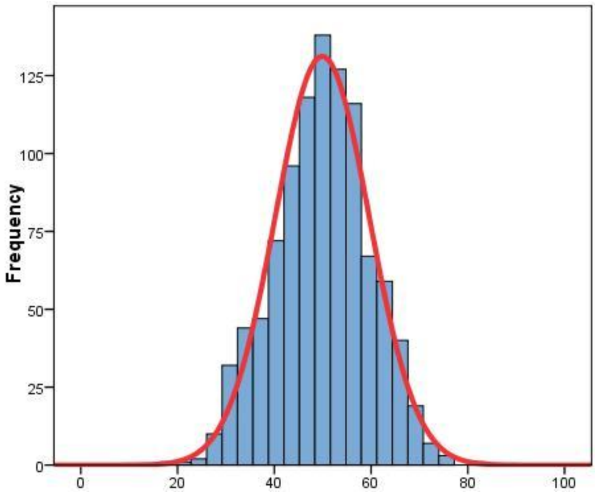
The balance of the consequences of these different types or error determines the level of confidence you might want to accept. For example if you are testing the efficacy of a new and very expensive drug (or one with lots of unwanted side effects) you might want to be very confident that it worked before you made it widely available, you might select a very stringent confidence level (e.g. p<.001) to minimize the risk of a false positive (type I) error. On the other hand if you are piloting a new approach to teaching statistics to students you might be happy with a lower confidence level (say p<.05) to determine whether it is worth investigating the approach further.

Before leaving p-values we should note that the p-value tells us nothing about the size of the effect. In large samples even very small differences may be statistically significant (bigger sample sizes increase the statistical power of the test). See **Page 1.10** for a discussion of effect size. Also, remember that *statistical* significance is not the same as *practical* importance - you need to interpret your findings and ground them in the context of your field.

Standard error and confidence intervals

A core issue in generalising from our sample to the wider population is establishing how well our sample data fits to the population from which it came. If we took lots of random samples from our population, each of the same number of cases, and calculated the mean score for each sample, then the sample means themselves would vary slightly just by chance. Suppose we take 10 random samples, each composed of 10 students, from the Year 11 group in a large secondary school and calculate the mean exam score for each sample. It is probable that the sample means will vary slightly just by chance (*sampling variation).* While some sample means might be exactly at the population mean, it is probable that most will be either somewhat higher or somewhat lower than the population mean. So these 10 sample means **would themselves have a distribution with a mean and a standard deviation** (we call this the *sampling distribution*). If lots of samples are drawn and the mean score calculated for each, the distribution of the means could be plotted as a histogram (like in **Figure 1.9.2**).

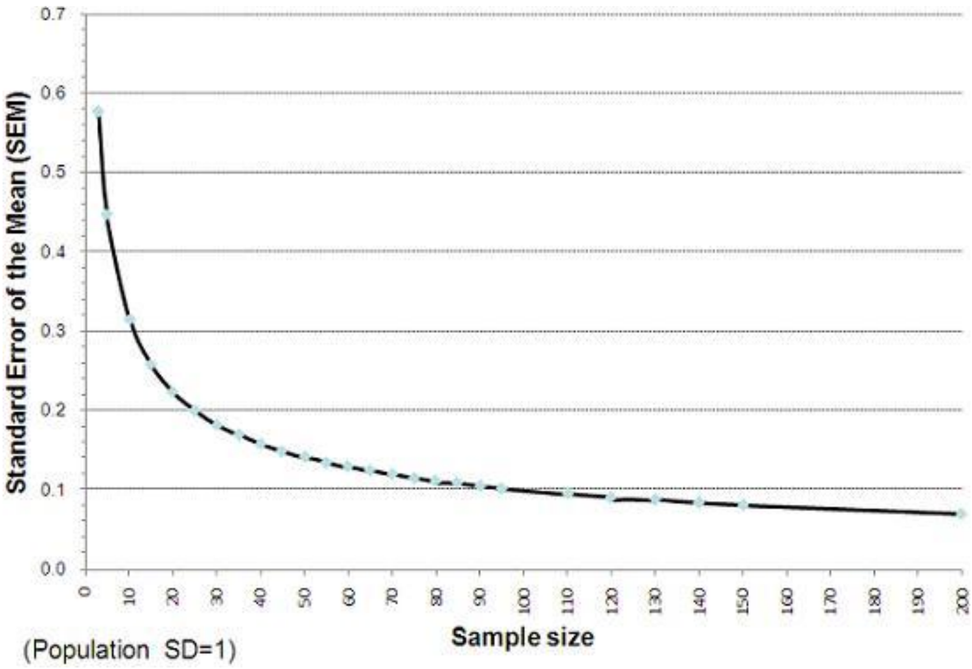
Figure 1.9.2: Histogram of mean scores from a large number of samples



The standard deviation of the distribution of the sample means is called the *standard error (SE)*. The SE is extremely important in determining how confident we can be about the accuracy of the sample mean as a representation of the population mean. Suppose **Figure 1.9.2** was the result of drawing many random samples, each composed of 10 cases, from a population where the mean score was 50. The standard deviation of the distribution of the sample means (the *standard error*) is approximately 10 score points. We can use the properties of the normal distribution to calculate the range above or below the population mean within which we would expect any given sample mean to lie (given our sample size). Two-thirds (68%) of the sample means would lie between +/- 1 SE of the population mean and 95% of samples means would lie within +/- 2 SE of the population mean. For the example in **Figure 1.9.2** we can say that 68% of the means (from random samples of 10 cases) would lie between 40 and 60, and 95% of the means (from random samples of 10 cases) would lie between 30 and 70. These 'confidence intervals' are very useful and will crop up frequently (in fact we say more about their uses below).

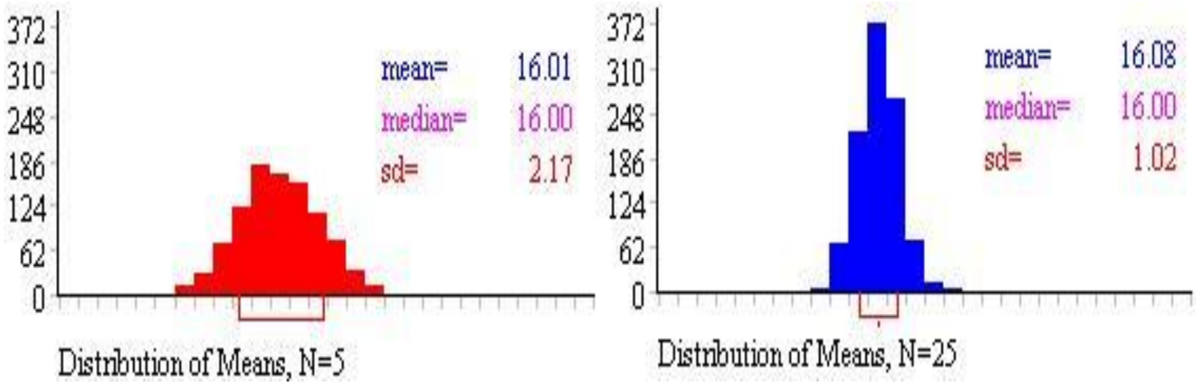
Crucially the SE will vary depending on the size of the samples. With larger samples we are more likely to get sample mean scores that cluster closely around the population mean, with smaller samples there is likely to be much more variability in the sample means. Thus the greater the number of cases in the samples the smaller the SE. **Figure 1.9.3** shows the relationship between sample size and the SE.

Figure 1.9.3: Relationship between standard error of the mean (SEM) and sample size



We understand that taking the 'mean of the means' and all that this entails may be a fairly complicated idea so we thought you might like to use this online toy which allows you to model sampling distributions. We call it a toy to make it more enticing but it is probably less fun than a Transformer or Buckaroo. We thank *David Lane* and his wonderful *Onlinestatsbook* (see **Resources**) for the creation of this helpful little application! To demonstrate how the sample size influences the SE of the sampling distribution, look at the difference between the histograms of the sample means when you draw 1000 samples each composed of 5 cases compared to when you draw 1000 samples each composed of 25 cases. We have adapted the output from the Onlinestatbook application in **Figure 1.9.4** to demonstrate this... but don't take our word for it, experiment for yourself!

Figure 1.9.4: Influence of sample size on SE



***Note**: the value labelled SD in the figure is actually the SE, because it is the SD of the distribution of means from several samples.*

Note that the means of the two sampling distributions are very similar. With a sufficient number of samples the mean of the sampling distribution will be centred at the same value as the population mean. However look at the SE. You can see how the SE shrinks when the larger sample size is used. In the first case, when each sample is composed of just 5 cases (N=5) the SE is 2.17. For the second case (where each sample has 25 observations, N=25) the SE is much smaller (1.02). This means the range of values within which 95% of sample means will fall is much more tightly clustered around the population mean.

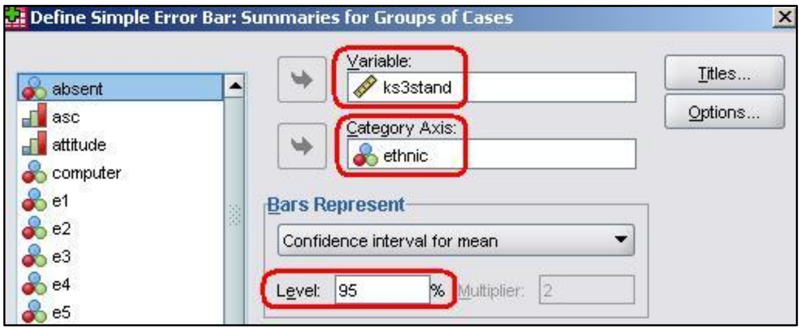
In practice of course we usually only have one sample rather than several: we do not typically have the resources to collect hundreds of separate samples. However we can estimate the SE of the mean quite well from knowledge of the SD and size of our sample according to the simple formula:

$$se = \frac{SD}{\sqrt{n}}$$

It turns out that as samples get large (usually defined as 30 cases or more) the sampling distribution has a normal distrubution which can be estimated quite well from the above formula. You will notice this chimes with **Figure 1.9.3**. The reduction in the SE as we increase our sample size up to 30 cases is substantial while the incremental reduction in the SE by increasing our sample sizes beyond this is much smaller. This is why you will often see advice in statistical text books that a minimum sample size of 30 is advisable in many research contexts.

Practical uses of confidence intervals

Let's take a practical look at confidence intervals. An error bar plot can be drawn to help you visualize confidence intervals. Let's use the LSYPE dataset (LSYPE 15,000) to compare the mean standardized test score at age 14 for different ethnic groups but take into account the margin of error (error bar) for each group mean. Use **Graphs > Legacy Dialogs > Error Bar** and select the default *Simple* and *Summaries for groups of cases option* to open the following menu:



Define Simple Error Bar: Summaries for Groups of Cases

Variable: ks3stand

Category Axis: ethnic

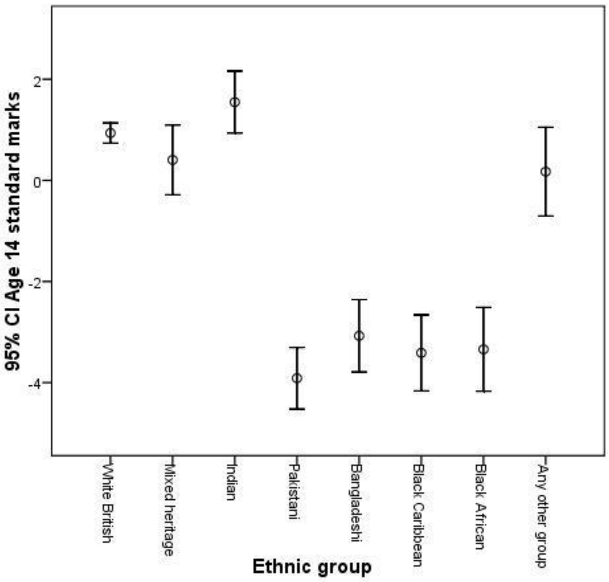
Bars Represent: Confidence interval for mean

Level: 95 %

Multiplier: 2

We will move age 14 test score (*ks3score*) into the *Variable* box and *ethnic* onto the *Category Axis*. Note the section titled *Bars Represent* which allows you to define the confidence interval – the default of 95% is the most commonly used, so we'll stick to that, but it is useful to know it can be altered to match the context. Click **OK** when you are happy with the settings and **Figure 1.9.5** should appear.

Figure 1.9.5: Mean age 14 score by ethnicity with 95% Confidence intervals



The circle in the middle of each line represents the mean score for that ethnic group. The extension of the line represents the range in which we are 95% confident that the 'true' mean lies for the group (+/- 2 SE). Note how the confidence interval for White British students is comparatively narrower than the intervals for the other ethnic groups. This is because the sample size for this group is much larger than for the other groups (see **Figure 1.5.1, Page 1.5**). Everything else being equal the larger the sample the more likely it is to represent the population and the more precisely we can estimate the 'true' population mean.

We can see from these error bars that, even though there are differences in the mean scores of the Pakistani, Bangledeshi, Black African and Black Caribbean groups, their confidence intervals overlap, meaning that there is insufficient evidence to suggest that the true population means for these groups differ significantly. However, such overlap in confidence intervals does not occur when we compare, for example, the White British students with these ethnic groups. The White British students score more highly at age 14 on average and the confidence intervals do not overlap. Overall the error bar plot suggests that on average White British, Mixed Heritage and Indian groups achieve a significantly higher age 14 test score than the Pakistani, Bangladeshi, Black African and Black Caribbean groups.

We have shown you some of the basics of probability and started to consider how to analyse differences between group means. Let’s now expand on this and show you some of the different methods of comparing means using SPSS.

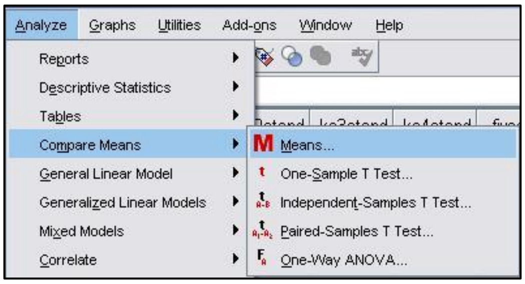
1.10: Comparing Means

On **Page 1.9** we discussed the use of an independent t-test to test the hypothesis that there is a difference between boys and girls age 14 test scores. This page teaches you about the t-test along with other ways of comparing the mean scores of groups to ascertain if there are statistically significant differences. The statistical tests work on the principle that if the two samples are drawn from the same population they will have fairly similar but not identical means, since there will be random variation between samples selected from the population (see **Page 1.9** about the standard error). However if the differences between the means are large enough in relation to the sample size we can conclude that the groups are drawn from populations with different means (e.g. boys and girls).

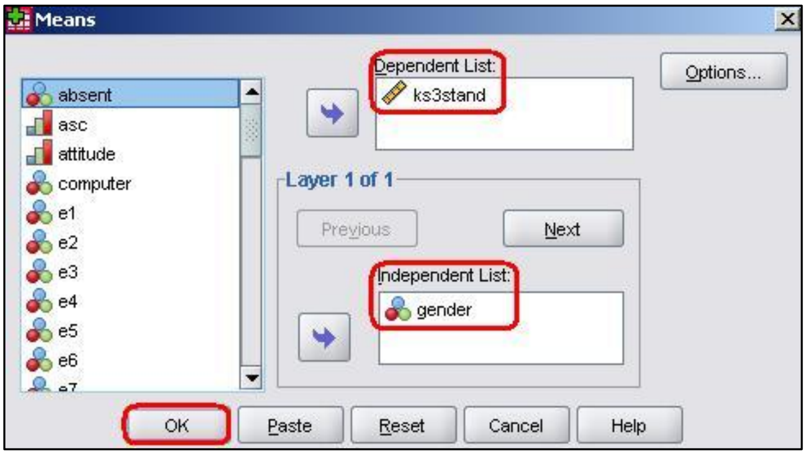
Field (2009), Chapter 9 (see **Resources**) covers the comparison of means in some detail should you wish to learn about the topic in more depth. Let's start by showing you a simple mean comparison and how to do it on SPSS.

Simple Means Comparisons

The first thing to do is just look at the mean score on the test variable for the two groups you are interested in. Let's see how girls and boys differ with regard to their age 14 test score (*ks3stand*). You can follow us through using the LSYPE 15,000 dataset: **Analyze > Compare Means > Means**.



This will access a pop-up window which allows you to define your variables. Age 14 standardized exam score (*ks3stand*) goes in the *Dependent List* box because this is the variable we will be comparing our categories on. *Gender* goes in the *Independent List* because it contains the categories we wish to compare.

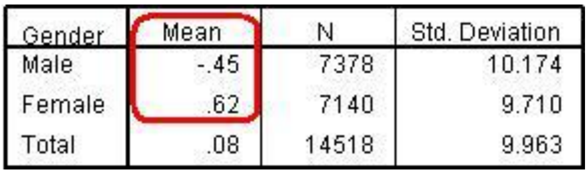

Dependent List: ks3stand

Layer 1 of 1

Independent List: gender

You can access the *Options* sub-menu to select a number of different statistics to add to your output. These are useful options and worth exploring but for now we only need the basic statistics so click **OK** to run the analysis.

Figure 1.10.1: Basic Mean Comparison Report Output



Gender	Mean	N	Std. Deviation
Male	-.45	7378	10.174
Female	.62	7140	9.710
Total	.08	14518	9.963

The **Case Processing Summary** just tells you the number of participants used for the analysis and those who were excluded (missing) so we haven't shown it here. **Figure 1.10.1** is the **Report** and shows us separate mean scores on age 14 exams for boys and girls. We can see that the female mean is .62, which seems a lot higher than the male mean of -.45. When males and females are not treated separately the mean score for the students included in this analysis is .08. We can also see the number of students of each gender and the standard deviation for each gender in this table. Note that the SD for boys is slightly higher than it is for girls, demonstrating that the boy's scores were more variable.

Though there seems to be a clear difference in the means we need to check that this difference is statistically significant by providing evidence that it is unlikely to be a result of sampling variation. This is where T-tests come in.

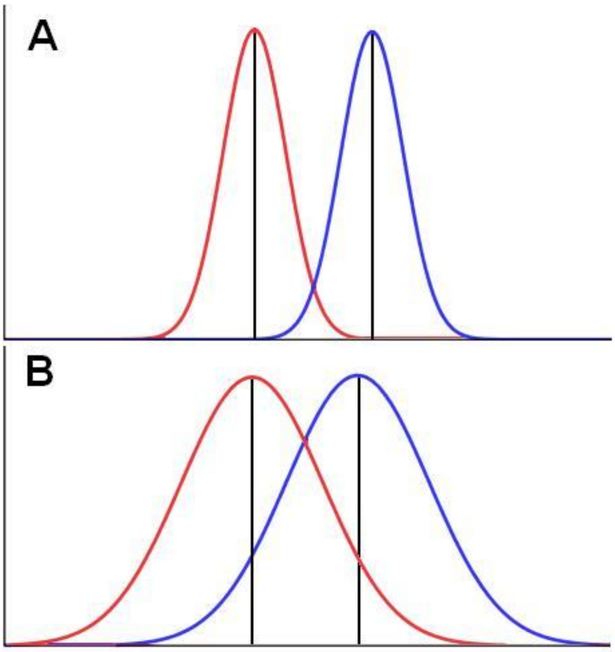
T-tests

We will not get into the formula – we try to minimize our involvement with such things! Besides, there are plenty of sources available which explain the mechanics of T-tests far better than we could (don't believe us? Check out our **Resources**, *Field, 2010; pages 334-341* in particular). However it is important to understand the basic principles underlying the t-test so you can perform one correctly and interpret the output accurately.

T-tests allow you to test the statistical significance (calculate the p-value) of the difference between two means on a scale variable. Statistical tests work on the principle that if two samples are from the same population they will have fairly similar means, but not identical since there will be random variation inherent in the sampling process. Basically we are asking if the two means are far enough apart from one another that we can be confident that they were drawn from separate populations.

It is slightly more complicated than simply looking at the difference between means because we also need to consider the *variance* (in the form of standard deviation) within our groups along with the *sample size* of the two groups. **Figure 1.10.2** should help you to visualize the importance of both mean and standard deviation. Imagine that boys and girls are each given their own frequency distribution of age 14 scores. The red frequency distributions represent boys and the blue ones girls, with age 14 exam score running along the horizontal x-axis. Two possible cases are outlined in the figure that illustrates the role of standard deviation. In **Cases A and B** the difference in male and females mean scores are the same but in **Case B** the standard deviations for the groups are much higher:

Figure 1.10.2: The role of standard deviation when comparing means



- The means in **Case A** are likely to differ significantly because there is little overlap in the distributions, the difference in means is therefore large relative to the variance (standard deviation). In **Case A** the difference in age 14 exam scores between boys and girls is likely to be statistically significant.

- The means in **Case B**, while roughly the same as **Case A**, may not be significantly different because the difference between means is small relative to the considerable overlap in the distributions. In **Case B** the difference in age 14 exam scores between boys and girls is unlikely to be statistically significant, they are more likely to exist simply through chance factors during sampling (e.g. a disproportionate number of less able boys were selected).

Statistical significance is ascertained by returning to the properties of the normal distribution. As you can see in **Figure 1.10.2 Case A**, the mean boys score appears to be somewhere beyond two standard deviations from the mean girls score. This is outside of the 95% confidence interval and therefore unlikely to have come from the same population (p < .05). We have shown this visually but the T-test crunches the numbers to calculate it precisely.

T-tests are a powerful tool but they do require you to be using something called *parametric* data. To be defined as parametric, your data needs to meet certain assumptions and if these are violated your conclusions can become wildly inaccurate. These assumptions are:

Parametric assumptions
1) Data are normally distributed in the population

2) Data are measured at least at interval (continuous) level

3) Variance in the groups to be compared are roughly equal (there is ‘homogeneity of variance’)

4) Scores are independent (the behaviour of one participant does not influence the behaviour of another)

To complicate matters there are also three forms of t-test, each designed to deal with a specific type of research question:

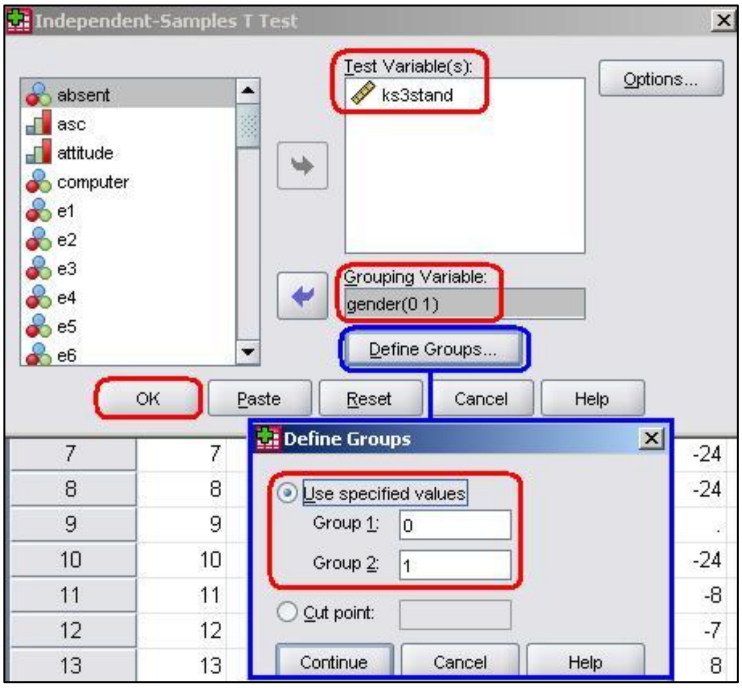
> ***One sample t-test***: to compare one sample to a known population mean (e.g. an IQ test with an established mean score of 100)
>
> ***Independent samples t-test***: to compare two separate (independent) groups (e.g. males vs. females)
>
> ***Paired samples t-test***: when the same cases are assessed on two different occasions (e.g. a group of infants’ reading test scores are compared before and after a specially designed classroom activity)

Luckily the basic principles of all three tests are very similar, with only the methods tweaked to suit each type of research question. All three types are easily performed using SPSS but the most common is probably the independent samples T-test.

Example

Let’s run one using our example research question from the LSYPE 15,000 dataset: *Do girls do better in exams at age 14 than boys?*

Go **Analyze > Compare Means > Independent Samples T Test** to access the following menu:



Test Variable(s): ks3stand

Grouping Variable: gender(0 1)

Define Groups...

● Use specified values

Group 1: 0

Group 2: 1

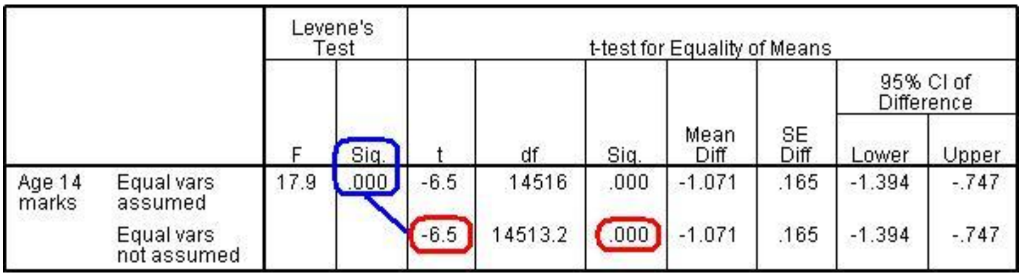
○ Cut point:

The variable we wish to compare boys and girls on is age 14 exam score (*ks3stand*) and needs to be placed in the *Test Variable(s)* window (notice how they keep changing the name of this window for different tests? Though there are good reasons for this it can get disorientating!). Our *Grouping Variable* is *gender*. Before you can proceed to run the test you will need to click on the button marked *Define Groups* to tell SPSS which categories within the variable you need to compare. This seems silly because we only have two categories (boys and girls) but there are times when you may want to compare two specific categories from a variable which has more than two. Also, SPSS is occasionally quite silly.

Simply enter the numeric codes '0' (for boys) and '1' (for girls) into the *Group 1* and *Group 2* fields, clicking *Continue* when you are satisfied. Note SPSS does allow you to set a *Cut point* which means you can divide up scale data into two categories if you wanted to. Once all the variables are defined click **OK** to run the analysis.

The first table contains the **Group Statistics**, which basically gives us the same information we saw when we ran a simple means comparison. It is the second (unwieldy long) table that we are interested in here, the **Independent Samples Test** (**Figure 1.10.3**):

Figure 1.10.3: Independent samples T-test comparing age 14 exam score across gender



		Levene's Test		t-test for Equality of Means						
									95% CI of Difference	
		F	Sig.	t	df	Sig.	Mean Diff	SE Diff	Lower	Upper
Age 14 marks	Equal vars assumed	17.9	.000	-6.5	14516	.000	-1.071	.165	-1.394	-.747
	Equal vars not assumed			-6.5	14513.2	.000	-1.071	.165	-1.394	-.747

Note: *We have cut some of the terms down slightly to fit the table on our page so it may look slightly different to your version.*

Let us work through this table and interpret it. The first matter to address is rather confusing but it is important. *Levene's Test* tells us whether or not we are safe in our assumption that our two groups have equal variances (if you recall, this tackles point 3 of our parametric assumptions). If the test is not statistically significant then we can assume there are equal variances and use a normal T-test. However if Levene's test is statistically significant (as is the case here) then we need to use a corrected version of the T-test... Luckily SPSS does this for you! All you need to do is use the *Equal variances not assumed* row of the table. Had Levene's test been non-significant we would use the top row, *Equal variances assumed*.

Now we know which row to examine we need only move along to the column marked '*Sig*' to ascertain whether the differences between the boys and girls is statistically significant. We can see from the table that it is highly significant – the p-value is .000, so small it is less than 3 decimal places (p < .001)! The actual T-statistic is included which is important to report when you write up your results, though it does not need to be interpreted (it is used to calculate the p-value).The table also tells us the difference between the means (-1.071, meaning boys '0' score less than girls '1') and provides us with a confidence interval for this figure.

Effect Size

We are now confident that the difference we observed between the age 14 exam scores of males and females reflects a genuine difference between the subpopulations. However what the p-value does not tell us is the how big this difference is. Given our large sample size we could observe a very small difference in means but find that it is statistically significant. P-values are about being confident in your findings but we also need a gauge of how strong the difference is. This gauge is called the effect size.

Effect size is an umbrella term for a number of statistical techniques for measuring the magnitude of an effect or the strength of a relationship. In a few cases it can be straightforward. If the dependent variable is a natural or well understood metric (e.g. GCSE grades, IQ score points) we can tell just by looking at the means – if males are scoring an average of 50% on an exam and females 55% than the effect is 5 percentage points. However, in most cases we wish to standardize our dependent variable to get a universally understood value for effect size. For T-tests this standardized effect size comes in the form of a statistic called *Cohen's d*.

SPSS does not calculate Cohen's d for you but luckily it is easy to do manually. Cohen's d is an expression of the size of any difference between groups in a standardised form and is achieved by dividing this difference by the standard deviation (SD):

Cohen's d = (Mean group A – Mean group B) / pooled SD

Pooled SD = (SD group A x n group A + SD group B x n group B) / N

The output statistic is a value between 0 and 1. Effect size is powerful because it can compare across many different outcomes and different studies (whatever the measure is we can calculate an effect size by dividing the difference between means by the standard deviation). Values of Cohen's d can be interpreted as follows:

Figure 1.10.4: Interpreting Effect size (Cohen's d)

Cohen's d	Description
0.0 – 0.2	"Weak"
0.2 – 0.5	"Moderate"
0.5 - 0.8	"Strong"
0.8+	"Very strong"

Alternatively the Cohen's d value can be viewed as equivalent to a Z score. You can then use the normal distribution (**Page 1.8, Figure 1.8.4**) to indicate what percentage of one group score below the average for the other group. For example, if we found that group B had a higher mean than group A with an effect size of 0.50, this would correspond to 69.1% of group A having a score *below* the group B mean (but remember that 50% of group B members do as well!).

Example

Let's calculate the effect size for the difference in age 14 scores between males and females in the LSYPE dataset. **Figure 1.10.3** tells us that the difference between the means scores for girls and boys is 1.07. We also know the standard deviations and sample sizes (*n*) for each group from **Figure 1.10.1**. All we need to do is plug these values into our formula:

Pooled SD = *(SD group A x n group A + SD group B x n group B) / N*
= (10.174 x 7378 + 9.710 x 7140) / 14518 = 9.946
Cohen's d = (Mean group A – Mean group B) / pooled SD

= 1.07/ 9.946 = **.108**

According to **Figure 1.10.4** the value of .108 actually corresponds to a weak effect. Even though we have observed a gender difference that is highly statistically significant it is not hugely powerful. We can be confident that there is a gender difference but the difference is relatively small.

Overall the results of the T-test could be written up like this:

> Male and female students differed significantly in their mean standardized age 14 exam score (t= -6.5, df =1453, p<.001). The male mean (mean = -.45, SD=10.2) was 1.07 standard points lower than for females (mean= .62, SD=9.7), indicating an effect size (Cohen's d) of 0.11.

Let’s now move on to look at how to handle means comparisons when there are multiple categories.

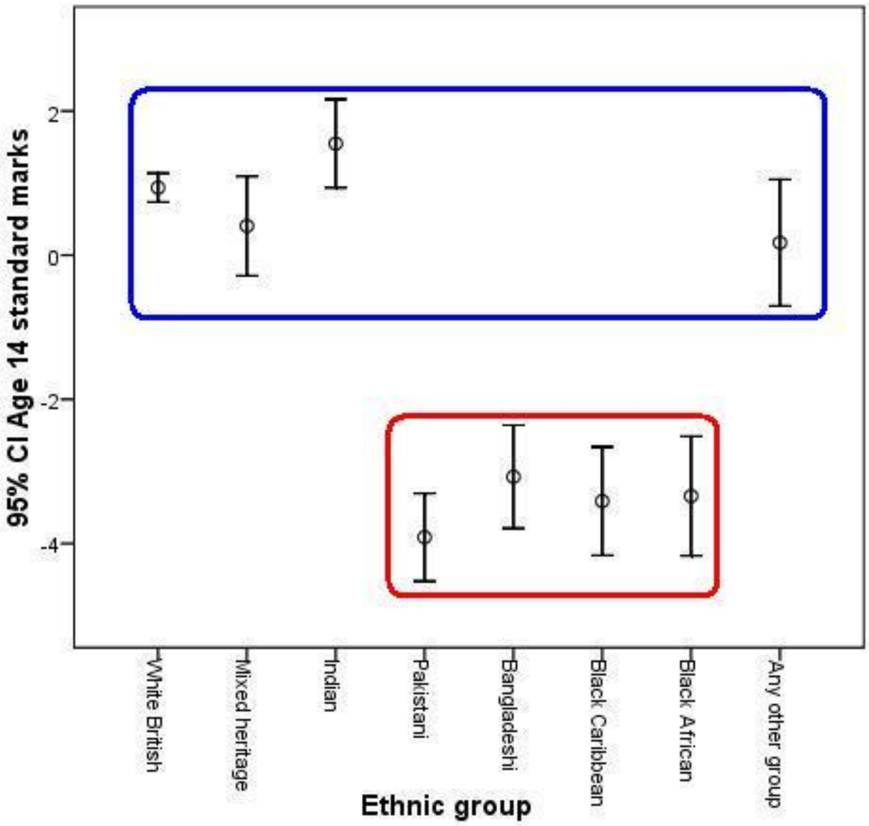
One-Way ANOVA

ANOVA stands for ‘Analysis of Variance’. The one-way ANOVA is very similar to the t-test but it allows you to compare means between more than two groups, providing an overall test of whether there is any significant variation in scores between these groups (producing something called an F test statistic). Basically it tests the null hypothesis that all of the group means are the same. Again, we want to only introduce the important concepts and practicalities in this module so we do not provide an explanation of how the ANOVA works (Check out our **Resources**, *Field (2009), Chapter 10* if you are of a curious mind!). However we will say that it comes from a very similar family of analytical methods as regression and so your understanding of ANOVA may stealthily grow as you carry on further down the regression rabbit hole.

We use T-tests to compare two group means but if we are interested in comparing the scores of multiple groups, we need to use a one-way ANOVA. When a variable, like gender, has two categories (male and female), there is only one comparison. However, if you have an independent variable with five categories (e.g. social science, science, art, humanities, other) then 10 comparisons, one for each pair of variables, are needed. When the overall one-way ANOVA result is significant, that does not necessarily mean that all of these comparisons (known as pair-wise comparisons) are significant. Thus, we need to find out whether all ten comparisons are significant, or just some of them. You can make such comparisons between the pairs of categories using ‘Post-hoc’ tests. These are a bit like individual T-tests which back up and elaborate upon the overall ANOVA result.

Figure 1.10.5 is an adaptation of **Figure 1.9.5** which illustrates the need for an ANOVA (called an ‘omnibus’ test because it provides an overall picture) to be backed up with post-hoc tests. The error bars show that there clearly is an overall effect of gender on age 14 exam scores: some ethnic groups clearly outperform others (e.g. the comparison between Indian and Pakistani students). However it is not the case that every single pair-wise comparison is statistically significant. For example, the Bangladeshi and Black Caribbean students do not appear to score much differently. We have highlighted two sets of category on the error bars below which appear to demonstrate significant differences between some pair-wise comparisons (between categories in the blue and red sets, for example White British and Pakistani) but not others (within each set, for example White British and Indian).

Figure 1.10.5: Mean age 14 score by ethnicity with 95% CI Error Bars and illustration of statistically significant comparisons



Remember that when making a large number of pair-wise comparisons some are likely to be significant by chance (at the 5% level we would find 1 in 20 comparisons statistically significant just by chance).

There are 18 different forms of post hoc tests (which is rather intimidating!). Your choice of post-hoc test depends on whether the group sample sizes and variances are equal and the practical significance of the results (See *Field, 2009, p372-374* in our **Resources** for a full discussion). The following are the ones which are most frequently used:

- **Bonferroni** and **Tukey** are conservative tests in that they are unlikely to falsely give a significant result (type I error) but may miss a genuine difference (type II error). See **Page 1.9** for more on these error types.

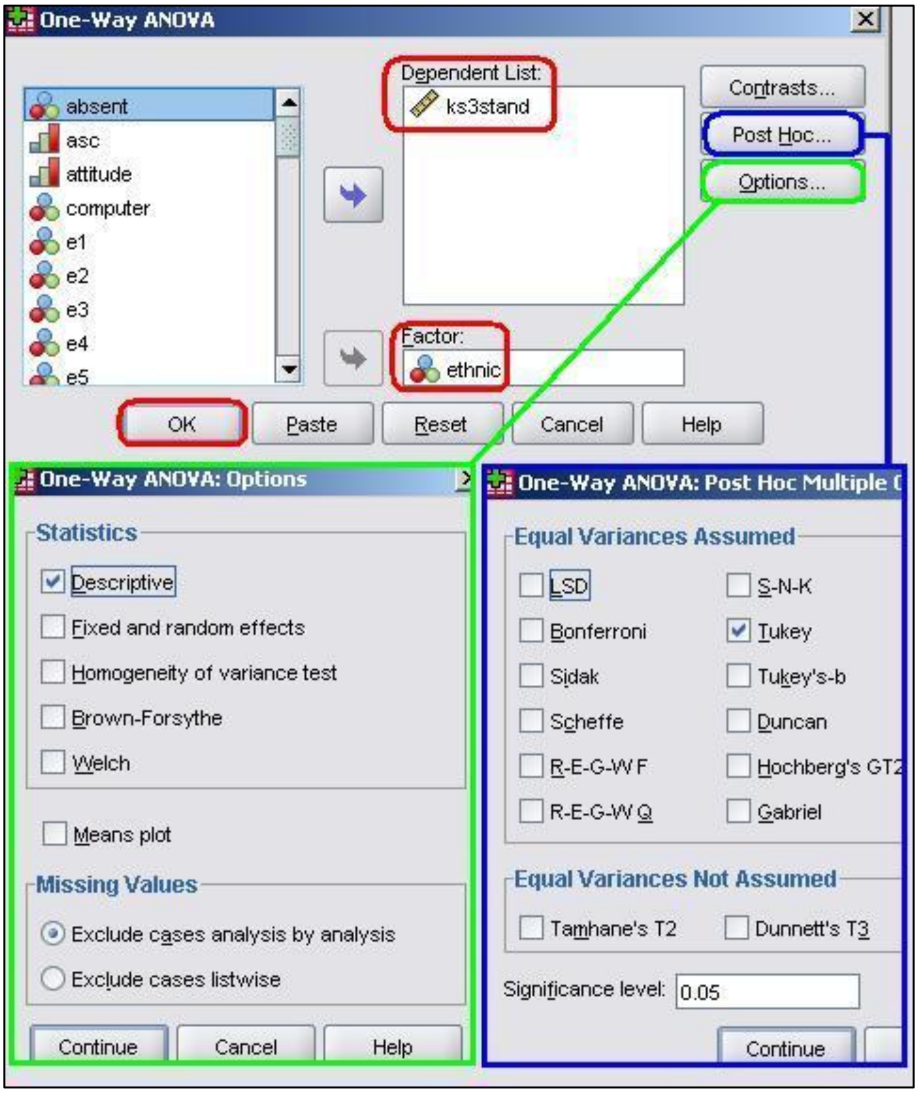
- **LSD and SNK** are more liberal tests, which means that they may give a false positive result (type I error) but are unlikely to miss a genuine difference (type II error).

Example

We realize we have sprinted through this explanation so let's run an example one-way ANOVA using a research question from the LSYPE 15,000 dataset.

How do White British students do in exams at age 14 compared to other ethnic groups?

As before, we access the One-Way ANOVA using the compare means menu: **Analyze > Compare Means > One-Way ANOVA**. The pop-up menu below will appear. Once again we put the variable we are comparing the categories on, age 14 exam scores (*ks3stand*), in the field marked *Dependent List*. Our independent variable, ethnicity (*ethnic*) goes in the field marked *Factor*.



Dependent List: ks3stand

Factor: ethnic

☑ Descriptive

☑ Tukey

● Exclude cases analysis by analysis

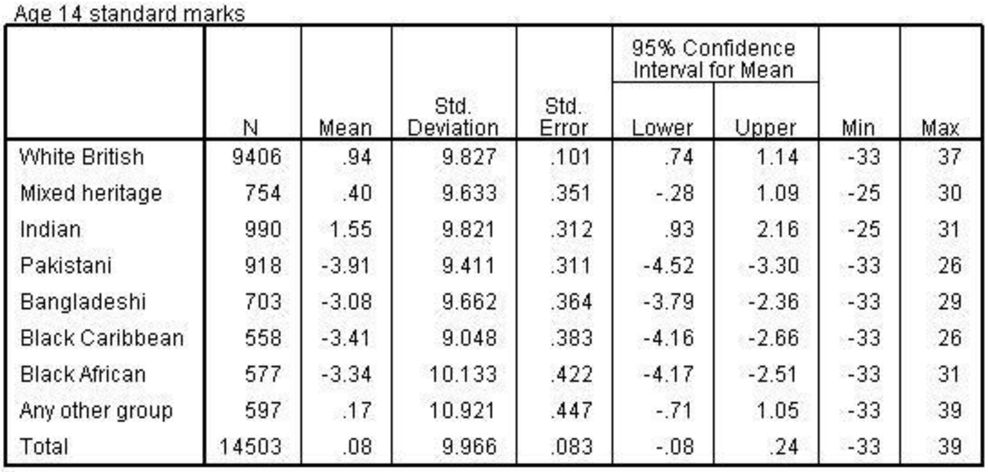
Significance level: 0.05

Before we continue we need to request that SPSS performs post-hoc analysis for us. Click on the button marked **Post Hoc** to open the relevant submenu. There is a mind boggling array of tests listed here and if you intend to perform ANOVAs in your own research we recommend you find out more about them through our **Resources**. For our purposes though, which is to perform a simple run through of the one-way ANOVA, let's just choose the **Tukey** *test.* Click *Continue* to shut this menu.

It is also worth checking the **Options** submenu. There are a number of extra statistics that you can request here, most are related to checking the parametric assumptions of your ANOVA. For now we will request only the basic **Descriptive** statistics to compliment our analysis. Click *Continue* to shut this menu and then, when you are happy with the settings click **OK** to run the analysis.

Figure 1.10.6 shows the first table of output, the **Descriptives**. This is a useful initial guide as it shows us the mean scores for each ethnic group.

Figure 1.10.6: Descriptives – Mean Age 14 Exam score by Ethnicity

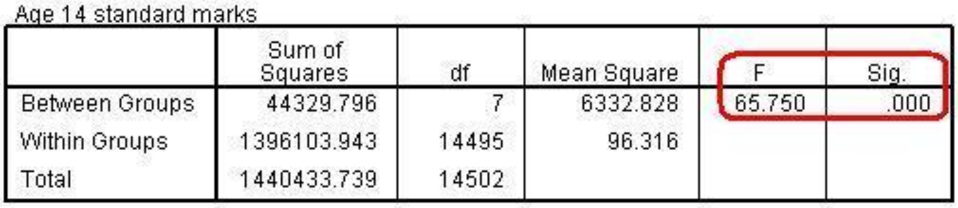


Age 14 standard marks

					95% Confidence Interval for Mean			
	N	Mean	Std. Deviation	Std. Error	Lower	Upper	Min	Max
White British	9406	.94	9.827	.101	.74	1.14	-33	37
Mixed heritage	754	.40	9.633	.351	-.28	1.09	-25	30
Indian	990	1.55	9.821	.312	.93	2.16	-25	31
Pakistani	918	-3.91	9.411	.311	-4.52	-3.30	-33	26
Bangladeshi	703	-3.08	9.662	.364	-3.79	-2.36	-33	29
Black Caribbean	558	-3.41	9.048	.383	-4.16	-2.66	-33	26
Black African	577	-3.34	10.133	.422	-4.17	-2.51	-33	31
Any other group	597	.17	10.921	.447	-.71	1.05	-33	39
Total	14503	.08	9.966	.083	-.08	.24	-33	39

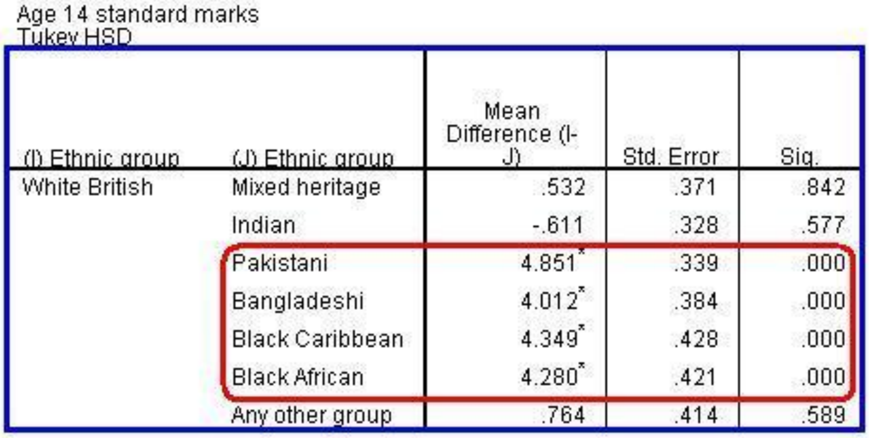
Figure 1.10.7 shows the **ANOVA** output along with a truncated version of the massive table marked **Multiple Comparisons**. We have included only the comparisons between White British students and the other groups but you will notice that the table you have is much bigger, providing pair-wise comparisons between all of the ethnic groups. The final two columns of the **ANOVA** table tell us that there are statistically significant differences between the age 14 scores of at least some of the different ethnic groups (F = 65.75, p < .001). This means we can reject the null hypothesis that all the means are the same.

Figure 1.10.7: ANOVA Output – Age 14 Exam score by Ethnicity



Age 14 standard marks

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	44329.796	7	6332.828	65.750	.000
Within Groups	1396103.943	14495	96.316		
Total	1440433.739	14502			



Age 14 standard marks
Tukey HSD

(I) Ethnic group	(J) Ethnic group	Mean Difference (I-J)	Std. Error	Sig.
White British	Mixed heritage	.532	.371	.842
	Indian	-.611	.328	.577
	Pakistani	4.851*	.339	.000
	Bangladeshi	4.012*	.384	.000
	Black Caribbean	4.349*	.428	.000
	Black African	4.280*	.421	.000
	Any other group	.764	.414	.589

The highlighted section of the **Multiple Comparisons** table shows the results of the post-hoc Tukey tests for the pair-wise comparisons between the White British students and the other ethnic groups. Looking down the column on the far right we can see that there are statistically significant differences with four of the seven groups. There are no significant differences between the White British students and the Mixed Heritage, Indian or 'Other' categories. However White British students score significantly higher than Pakistani, Bangladeshi, Black Caribbean and Black African groups (note that the stats only show that there is a difference – we had to check the means in the **Descriptives** table to ascertain which direction the difference was in).

We would report the ANOVA results as follows:

> There was a significant overall difference in mean standardized age 14 exam scores between the different ethnic groups F(7, 14495) = 65.75, p<.001. Pair-wise comparisons using Tukey post-hoc tests revealed multiple statistically significant comparisons. Students from White British (Mean = .94) backgrounds scored higher than those from Pakistani (Mean =-3.91), Bangladeshi (Mean = -3.08), Black Caribbean (Mean = -3.41) and Black African (Mean = -3.34) backgrounds.

Factorial ANOVA

Before wrapping this module up it is worth mentioning the Factorial ANOVA. The one-way ANOVA can be extended to simultaneously look at the influence on the outcome measure of multiple independent variables (e.g., gender, hours spent doing homework and A-level subjects). This is important because it lets you estimate both the *unique influence* of each variable and whether there are any *interactions* between the independent variables.

We are not going to cover this here because Factorial ANOVA is typically only used in experimental designs. We can do the same analyses using **regression**, which is arguably a more flexible and adaptable tool (though both ANOVA and regression are based on the same underlying procedures). Regression analysis is the primary focus of this website! Still, Factorial ANOVA is very useful and we suggest that you learn about it by using one of our recommended **Resources**, we suggest *Field (2009), chapter 12* (we do love Field!).

Conclusion

That's it for the foundation module. Please remember that this module has given a relatively superficial coverage of some of the important topics... it is not intended to fully prepare you for regression analysis or to give you a full grounding in basic statistics. There are some excellent preparatory texts out there are we recommend you see our **Resources** section for further guidance. If you are new to research with quantitative data you will need to read more widely and to practice simple data manipulation and exploratory data analysis with your data. We hope that you now have the confidence to start getting stuck into the world of regression...

Now it is time to take our **Quiz** and perhaps work through the **Exercise** to consolidate your understanding before starting on the next module. Go on... Have a go!

Foundation Module Exercise

Welcome to the first exercise! The following five questions can be worked through using the LSYPE 15,000 dataset. We recommend that you answer them in full sentences with supporting tables or graphs where appropriate – this will help when you come to report your own research. There is a link to the answers at the bottom of the page.

Note: The variable names as they appear in the SPSS dataset are listed in brackets.

Question 1

What percentage of students in the LSYPE dataset come from a household which has a home computer (*computer*)?

Use frequencies.

Question 2

Let's say you are interested in the relationship between achievement in exams at age 16 and computer ownership. Create a graph which compares those who own a computer to those who do not (*computer*) with regard to their average age 16 exam score (*ks4score*).

Use a bar chart.

Question 3

Is the difference between the average age 16 exam scores (*ks4score*) for those who do and do not own a computer (*computer*) statistically significant?

Use a T-test.

Question 4

Let's look at the relationship between social-economic class (*secshort*) and achievement in exams at age 16. Is there a difference between the three SEC groups (high, medium and low SEC) with regard to their average achievement in age 16 exams (*ks4score*)? If so which groups differ significantly?

Perform a oneway ANOVA with Scheffe post-hoc tests.

Question 5

Create an error bar graph which illustrates the difference between SEC groups (*secshort*) with regard to their average achievement in age 16 exams (*ks4score*).

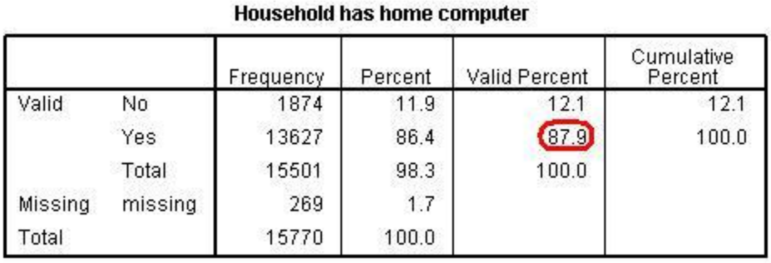
Use an error bar chart.

Answers

Question 1

What percentage of students in the LSYPE dataset come from a household which owns a computer (*computer*)?

By using **Analyze > Descriptive Statistics > Frequencies** the following table can be produced:



Household has home computer

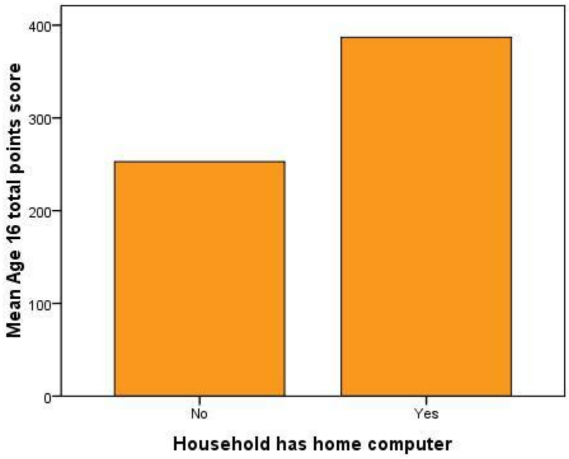
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	1874	11.9	12.1	12.1
	Yes	13627	86.4	87.9	100.0
	Total	15501	98.3	100.0	
Missing	missing	269	1.7		
Total		15770	100.0		

It shows that 87.9% of students who answered this question (the *valid cases*) come from a household which owns a computer.

Question 2

Let’s say you are interested in the relationship between achievement in exams at age 16 and computer ownership. Create a graph which compares those who own a computer to those who do not (*computer*) with regard to their average age 16 exam score (*ks4score*).

A bar chart can be produced by using **Graphs > Legacy Dialogs > Bar**. You need to select *Other Statistic (Mean)* for ‘bars represent’ and choose age 16 exam score.

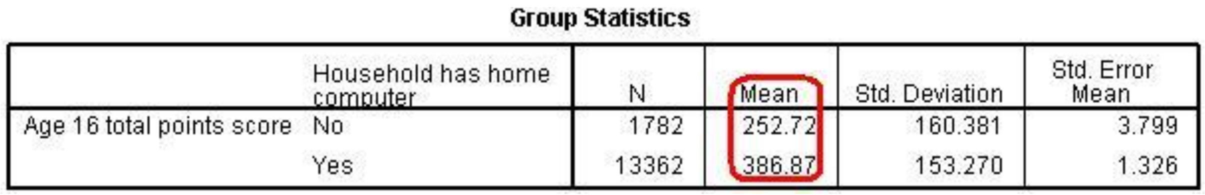


It seems that those from families who do own a computer have a higher mean score in age 16 examinations.

Question 3

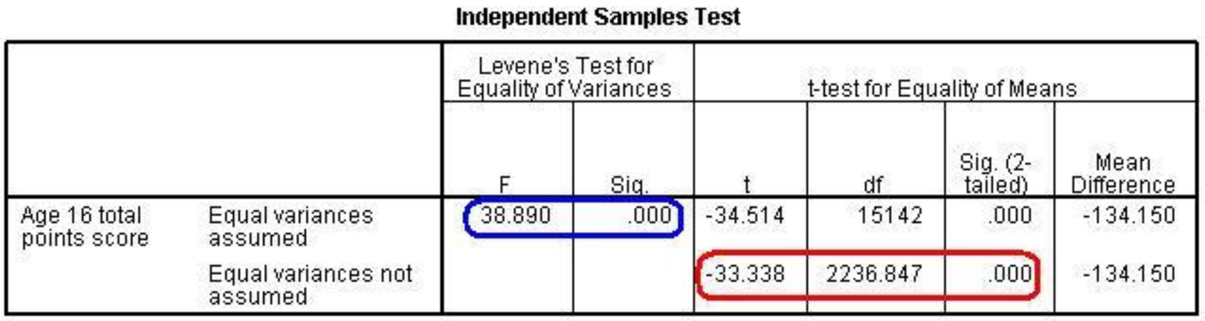
Is the difference between the average age 16 exam scores (*ks4score*) for those who do and do not own a computer (*computer*) statistically significant?

The t-test can be performed using **Analyze > Compare Means > Independent Samples T-test**. *ks4score* is the test variable and *computer* is the grouping variable. You should get the following output:



Group Statistics

	Household has home computer	N	Mean	Std. Deviation	Std. Error Mean
Age 16 total points score	No	1782	252.72	160.381	3.799
	Yes	13362	386.87	153.270	1.326



Independent Samples Test

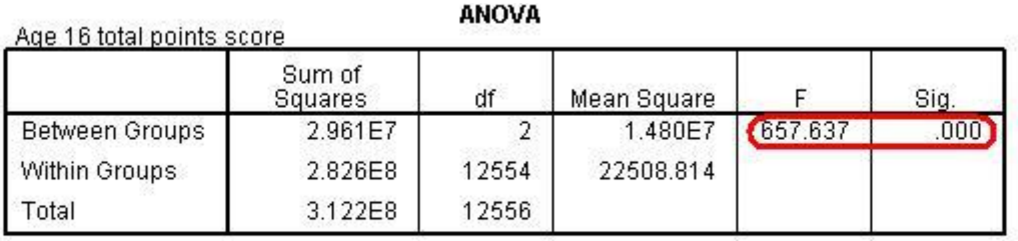
		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
Age 16 total points score	Equal variances assumed	38.890	.000	-34.514	15142	.000	-134.150
	Equal variances not assumed			-33.338	2236.847	.000	-134.150

The first table displays the descriptive statistics which tells us the mean Age 16 exam score and standard deviation for each group. There appears to be a substantial difference between the groups. The second table shows the T-test itself. Note that Levene's test is statistically significant which means we should not assume equal variances in the two groups and should use the adjusted figures in the second row (highlighted in red). The T-test shows that there is indeed a statistically significant difference between the mean age 16 exam scores of those from families with a computer compared to those from families without one (*t* = 33.3, df = 2237, *p* < *.0005*).

Question 4

Let’s look at the relationship between social-economic class (*secshort*) and achievement in exams at age 16. Is there a difference between the three SEC groups (high, medium and low SEC) with regard to their average achievement in age 16 exams (*ks4score*)? If so which groups differ?

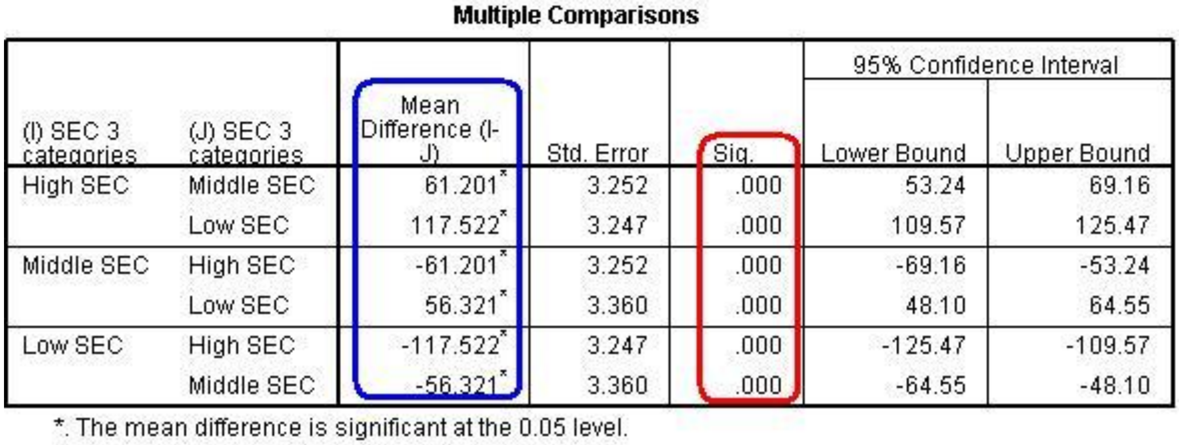
A one way ANOVA can be performed using **Analyze > Compare Means > One-Way ANOVA**. We use *ks4score* as the dependent variable and *secshort* as the factor. From the ‘Post-Hoc’ submenu you should select ‘Scheffe’ in order to perform the relevant pair wise post-hoc comparisons between SEC groups. You should generate the following output:



ANOVA

Age 16 total points score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.961E7	2	1.480E7	657.637	.000
Within Groups	2.826E8	12554	22508.814		
Total	3.122E8	12556			



Multiple Comparisons

(I) SEC 3 categories	(J) SEC 3 categories	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
High SEC	Middle SEC	61.201*	3.252	.000	53.24	69.16
	Low SEC	117.522*	3.247	.000	109.57	125.47
Middle SEC	High SEC	-61.201*	3.252	.000	-69.16	-53.24
	Low SEC	56.321*	3.360	.000	48.10	64.55
Low SEC	High SEC	-117.522*	3.247	.000	-125.47	-109.57
	Middle SEC	-56.321*	3.360	.000	-64.55	-48.10

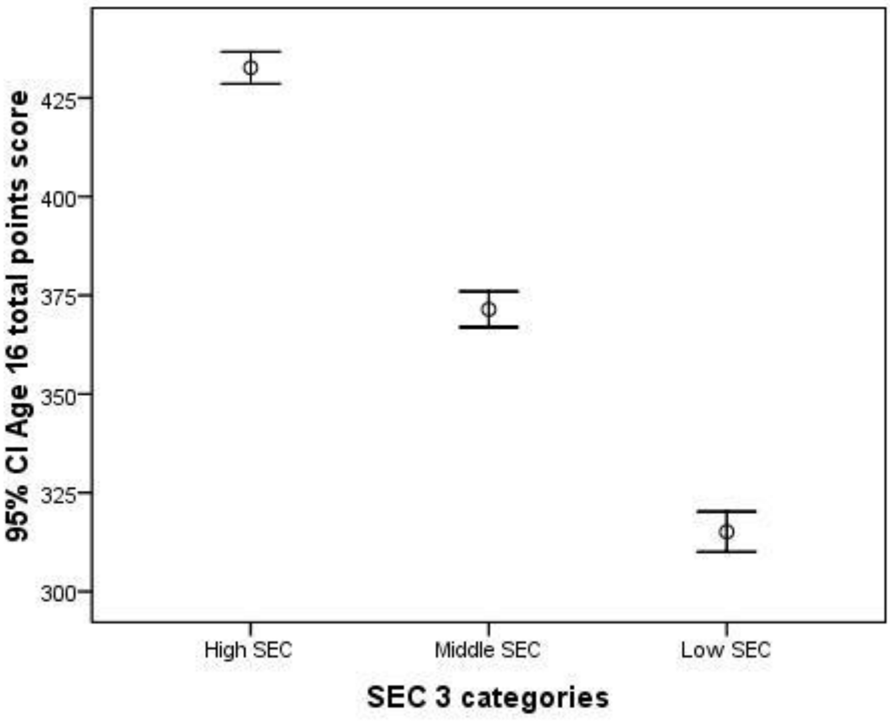
*. The mean difference is significant at the 0.05 level.

The first thing to notice is that, according to the omnibus F-test, there is a statistically significant difference between the groups *overall*, *F* = 657.6, df = 2, 12554, < .0005. We need to look at the post-hoc analysis to explore where these differences actually are. It appears that all three SEC groups are different from one another! The mean difference column shows us the ‘High SEC’ group scores an average of 61 more points than the ‘Middle SEC’ group and 117.5 more than the ‘Low SEC’ group. The ‘Middle SEC’ group scores 56 more points on average than the ‘Low SEC’ group. As shown in the column headed ‘Sig.’ all of these differences are highly statistically significant.

Question 5

Create an error bar graph which illustrates the difference between SEC groups (*secshort*) with regard to their average achievement in age 16 exams (*ks4score*).

An error bar chart can be produced by using **Graphs > Legacy Dialogs > Error Bar**. You should be able to produce a chart which looks like this:



From this chart you can see that there are clear differences between the mean age 16 exam scores for each group (the circle in the centre of each error bar), with the 'High SEC' group outperforming the 'Middle SEC' group, who in turn outperform the 'Low SEC' group. The error bars themselves encompass the range of scores within which we are 95% sure that the true mean in the population lies. The fact that the error bars do not overlap implies that the differences between groups are statistically significant (something we actually know to be true based on question 4).