

# Statistics for Health, Life and Social Sciences

Denis Anthony



Download free books at

[bookboon.com](http://bookboon.com)

Denis Anthony

# Statistics for Health, Life and Social Sciences

---

Statistics for Health, Life and Social Sciences  
© 2011 Denis Anthony & Ventus Publishing ApS  
ISBN 978-87-7681-740-4

# Contents

	<b>Preface</b>	<b>6</b>
<b>1</b>	<b>Designing questionnaires and surveys</b>	<b>7</b>
<b>2</b>	<b>Getting started in SPSS – Data entry</b>	<b>14</b>
<b>3</b>	<b>Introducing descriptive statistics</b>	<b>39</b>
<b>4</b>	<b>Graphs</b>	<b>55</b>
<b>5</b>	<b>Manipulating data</b>	<b>78</b>
<b>6</b>	<b>Chi square</b>	<b>101</b>
<b>7</b>	<b>Differences between two groups</b>	<b>115</b>
<b>8</b>	<b>Differences between more than two groups</b>	<b>124</b>
<b>9</b>	<b>Correlation</b>	<b>132</b>
<b>10</b>	<b>Paired tests</b>	<b>135</b>
<b>11</b>	<b>Measures of prediction: sensitivity, specificity and predictive values</b>	<b>149</b>

www.sylvania.com

**We do not reinvent the wheel we reinvent light.**

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM **OSRAM SYLVANIA**



12	<b>Receiver operating characteristic</b>	159
13	<b>Reliability</b>	168
14	<b>Internal reliability</b>	175
15	<b>Factor analysis</b>	183
16	<b>Linear Regression</b>	197
17	<b>Logistic regression</b>	216
18	<b>Cluster analysis</b>	227
19	<b>Introduction to power analysis</b>	238
20	<b>Which statistical tests to use</b>	249
	<b>Endnotes</b>	253
	<b>Answers to exercises</b>	254
	<b>Appendix Datasets used in this text</b>	277

**CHALLENGING PERSPECTIVES**

## Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

**We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.**

To find out more and apply, visit [www.jobs.eads.com](http://www.jobs.eads.com). You can also find out more on our **EADS Careers Facebook page**.

**AIRBUS    ASTRIUM    CASSIDIAN    EUROCOPTER**

**EADS**



# Preface

There are many books concerned with statistical theory. This is not one of them. This is a **practical** book. It is aimed at people who need to understand statistics, but not develop it as a subject. The typical reader might be a postgraduate student in health, life or social science who has no knowledge of statistics, but needs to use quantitative methods in their studies. Students who are engaged in qualitative studies will need to read and understand quantitative studies when they do their literature reviews, this book may be of use to them.

This text is based on lectures given to students of nursing, midwifery, social work, probation, criminology, allied health, podiatry at undergraduate to doctoral level. However the examples employed, all from health, life and social sciences, should be understandable to any reader.

There is virtually no theory in this text, almost no mathematics (a bit of simple arithmetic is about as far as it gets). I do not give every statistical test, nor do I examine every restriction of any test. There are texts, good ones, that do all of these things. For example Field's (2009) text is excellent and gives a fuller treatment, but his book is 820 pages long!

All books contain errors, or explanations that are not clear. If you note any email me on danthony@talktalk.net and I will fix them.

FIELD, A. 2009. *Discovering statistics using SPSS (and sex and drugs and rock 'n' roll)*, London, Sage.

# 1 Designing questionnaires and surveys

## Key points

- A well designed questionnaire is more likely to be answered
- A well designed questionnaire is more likely to give valid data
- Constructing your sample from the population is crucial

## At the end of this unit you should be able to:

- Create a sampling strategy
- Design a questionnaire

## Introduction

In this chapter we will learn how to create a questionnaire. Questionnaire design is well covered in many web sites (see list at the end for a few examples) and any good basic research text (there are many). This chapter presents an example and asks you to work on it to produce a questionnaire.

The main example we will use will be infection control in hospitals. There is considerable evidence that single rooms in hospitals are associated with fewer hospital acquired infections. The reasons are fairly obvious. However what proportion of hospital beds are in single rooms in the UK? It turns out that this was not known, and in 2004 I conducted a study to find out.

Suppose we want to find out how many single rooms there are in UK hospitals. What would be the best way to do this?

## Surveys

One way is to send a survey instrument out to hospitals in the UK requesting this information. But to which hospitals, and to whom in each hospital?

If you look at [www.statpac.com/surveys/sampling.htm](http://www.statpac.com/surveys/sampling.htm) you can see several ways to pick a sample. Let us consider each in turn:-

**Random sampling:** We could put all the names of all hospital trusts in a bag and take out the number required to make our sample. In practice we are more likely to select them with a random number generator, but in essence this is the same thing. Since the trusts are being selected randomly there should be no bias, so (for example) all the sites should not be in England, but a similar proportion in Wales, Scotland and Northern Ireland.

**Systematic sampling:** If we chose every tenth (say) record in our database this would probably not introduce a bias, unless there was some intrinsic pattern in the database, for example if the Welsh trusts were first, eleventh, twenty first etc. we would end up with a sample biased to Welsh sites.

**Stratified sampling:** To avoid bias we could split the population into strata and then sample from these randomly. For example if we wanted small, medium and large hospitals in our sample we could have size of hospital as the three strata and ensure we got required amounts of each.

**Convenience sampling:** We could use hospitals we have access to easily. We could use those in the East Midlands of the UK (for example).

**Snowball sampling:** We might use our convenience sample of the East Midlands and ask respondents to give us other contacts.

In this case, provided we do not introduce bias it is not critical which method we employ. We can rule out two methods. Convenience is not helpful here as we want a national picture of single rooms. Snowball sampling could introduce bias as the contacts known by a trust are not likely to be random, and it is not necessary as there are mailing lists of hospital trusts obtainable. Systematic is probably not going to be problematic as there is no reason to assume every  $n$ th contact will be related in any way. However it is not necessary, so why risk the possible bias? These three sampling methods are useful, but not here. Convenience sampling might be employed by a student on a limited budget, who realises the sample is likely to be biased, and states this as a limitation. Systematic sampling might be employed in a study of out-patient patients, where every tenth or twentieth patient was invited for interview until one got the number required (say thirty). You could sample the first thirty, but this might get you all patients in one morning clinic, which may not be generalisable. So random or stratified sampling are probably the way forward in this case.

While we often for practical reasons have to select a sample from a population, sometimes there is no need. In our case there are about 500 hospital trusts in the UK (figures vary as trusts are not static entities but merge and fragment from time to time) in the database at our disposal. Each ward manager will know how many single rooms are in his or her ward. But if we intended to contact every ward in every hospital then we would be sending out tens of thousands of questionnaires. This would be very expensive. However if we sent the questionnaire to the estates department of each hospital then we could possibly send out a questionnaire to every trust. This is a census study as every subject is included. In that case there may be no need to create a sample, or in other words the population is the sample. This is because a postal survey is cheap, and we could therefore afford to send a questionnaire to every trust. Entry of data for only a few variables would also be relatively fast and therefore cheap.

## Questionnaire design: data variables

To get valid and reliable data we need a well designed research instrument, in this case a questionnaire. We could send out a survey with one question on it, *How many single rooms are in your hospital?* However this is insufficient, we need to know how many beds in total to work out the percentage, so at least one other question is needed, *How many beds are in your hospital?*

However there are several other related issues which we may want to explore. For example single rooms are probably better than other ward designs in terms of infection control, but bays (4-6 beds in a room or area) might be better than Nightingale wards (30 or so beds in one long ward). We might therefore ask about the different types of ward designs.

We are more likely to get responses if the questionnaire is easy to fill in. A tick list is quick and simple. We could have a list of all ward designs and allow the respondent to just tick the ones in use.

But we do not know every type of ward design, so while a tick list is useful we might miss data. We could add an extra category “other” and invite the respondent to fill it in.

We might wonder why single rooms are not used in all hospitals for all patients. There are countries where single rooms are the norm, but not the UK. So finally we are interested in the views of respondents on the utility of single rooms. Is it for example the case that financial constraints limit the use of single rooms, or could it be that single rooms are not seen as optimal. For example are Nightingale wards thought to be superior in terms of observation of patients, or patient communication?

### Layout of the questionnaire

The same items laid out badly will get fewer responses than one laid out well. The design of the questionnaire is important. If it looks professional it is more likely to be responded to.

Sometimes we want precise answers, sometimes we are interested in ranges of values. If we do not care about someone’s exact age, we could give ranges of ages to tick. This may be important if exact data could in principle identify subjects, but data in groups might not. For example if I ask for a person’s age, gender and ethnic group there may only be one person of a given ethnicity, gender and age. This may make the respondent less willing to complete it, and has obvious ethical implications.

Simple things like lining up tick boxes in a question aid the respondent. Consider the layout of following segments of a questionnaire on following pages:-

#### Design 1

First Name ..... Family name ..... Number single rooms .....

Total number of beds in hospital ..... Postcode .....

**Design 2**

Number of single rooms	<input style="width: 80%;" type="text"/>
Total number of beds in hospital	<input style="width: 80%;" type="text"/>
Post Code	<input style="width: 80%;" type="text"/>

**Design 3**

Total number of beds in hospital	<input style="width: 90%;" type="text"/>
----------------------------------	--

% of beds in in hospital that are in single rooms	0-19 %	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
Please tick one (✓)	<input type="checkbox"/>								

Post code	<input style="width: 90%;" type="text"/>
-----------	--

Which one would you prefer to fill in? Would it make any difference it were:-

**Design 4**

Number of beds in hospital	Number of beds in single rooms
<input style="width: 95%;" type="text"/>	<input style="width: 95%;" type="text"/>

<b>Post code</b> (e.g. LE6 0AR)	<input style="width: 95%;" type="text"/>
------------------------------------	--

Please give any view you have on the use of single rooms with respect to privacy, infection control, noise, patient safety or any other issue you feel important
<input style="width: 100%; height: 100%;" type="text"/>

What difference would it make if the questionnaire was printed on high quality (bond) paper or coloured paper, or had an institutional logo/s (such as the university and/or hospital trust logo/s)?

Which of the above four questionnaires would you fill in? How could it be improved upon? Which gives more precise data? How precise do the data need to be? Personally I would use design four as it is neater than designs one and two, and has more precise information than design three, and has a space for qualitative comments.

Below are some activities that you may use in (say) a course. There are no definitive answers so I have not given correct answers in this text. In other chapters I use the term exercises (which do have correct answers).

## Activity 1

You interview some managers in the local trust and also conduct a few focus groups of nurses. You ask them about ward design, and they all tell you single rooms are a bad idea because of safety issues, and Nightingale wards are best for observation, and patients prefer them because they can communicate with other patients and be more involved. You now want to find out what patients think about this. Create a survey questionnaire that asks them about the following issues:-

1. Can nurses observe me?
2. Is privacy a problem?
3. Is lighting adequate?
4. Is the ward too noisy?
5. Is the bed space (bed, curtains etc.) pleasant?

Now think about the following issues:-

1. Would patients from certain minority ethnic groups have different views than (say) the majority white population?
2. Do older patients have different attitudes to younger ones?
3. Are men and women the same in their views?
4. Are patients in different types of ward likely to be the same in terms of attitude to ward layout (think of a very ill elderly cancer patient or a fit young man with a hernia being repaired)?

Create some variables to add to the questionnaire to gather data that could explore these issues.

While the question of ward design was explored with staff in one area, we want to know what patients across the NHS in the UK think. I.e. we are interested in national (UK) not local views. However it is possible there are regional differences. What variable/s would be needed to identify the area where patients come from so we could explore these?

Finally what type of bed space do they have? Are they in a single room, a bay or a Nightingale ward? Create a variable that captures these data.

## Activity 2

How are you going to conduct a survey of the patients using the questionnaire you designed above in activity 1? What sort of survey will it be? Will it be a census survey, or will you need to sample? If you sample how will you ensure your data are generalisable? Remember this is a national survey. If you sample, how big a sample will enable you to be reasonably sure you have valid data? How big a sample can you afford (what resources do you have)? How will you deliver the questionnaire? What ethical procedures will you need to address prior to conducting the survey?

## Activity 3

We want to do a survey of students on an online course. We want to know the following:-

1. Do students value the online course?
2. Which components (discussion board, email or course notes) are most useful?
3. Have the students previous experience of using email or the Web?
4. How confident are the students in terms of their computing skills?

However we also want to explore whether any of the following are relevant in student attitudes:-

- Ethnic group
- Age
- Gender



360°  
thinking.

**Deloitte.**

Discover the truth at [www.deloitte.ca/careers](http://www.deloitte.ca/careers)

© Deloitte & Touche LLP and affiliated entities.

Remember that the questionnaire has to:-

- Give you the data needed
- Be as easy to fill in as possible
- Look clean, neat and professional

## Resources (websites)

### Surveys

StatPac, (2010)

### Questionnaires

A workbook from the University of Edinburgh (Galloway, 1997)

A list of resources on questionnaire design (The University of Auckland)

### General

BUBL is an academic catalogue at [bubl.ac.uk/](http://bubl.ac.uk/) and has a research methods page (BUBL) including resources on surveys and questionnaires.

## References

BUBL. Available: [bubl.ac.uk/LINK/r/researchmethods.htm](http://bubl.ac.uk/LINK/r/researchmethods.htm) [Accessed 11 Nov 2010].

GALLOWAY, A. 1997. *Questionnaire Design & Analysis* [Online]. Available: [www.tardis.ed.ac.uk/~kate/qmcweb/qcont.htm](http://www.tardis.ed.ac.uk/~kate/qmcweb/qcont.htm) [Accessed 11 Nov 2010].

STATPAC. 2010. *Survey & Questionnaire Design* [Online]. Available: [www.statpac.com/surveys/](http://www.statpac.com/surveys/) [Accessed 11 Nov 2010].

THE UNIVERSITY OF AUCKLAND. *Developing questionnaires websites* [Online]. Available: <http://www.fmhs.auckland.ac.nz/soph/centres/hrmas/resources/questionnaire.aspx> [Accessed 2010 11 Nov ].

# 2 Getting started in SPSS – Data entry

## Key points

- SPSS is a statistics package
- You can enter data and load data
- You can perform statistical analysis and create graphs

At the end of this chapter you should be able to:

- Create a datafile
- Enter data
- Edit variables
- Save a datafile
- Conduct very basic descriptive statistics

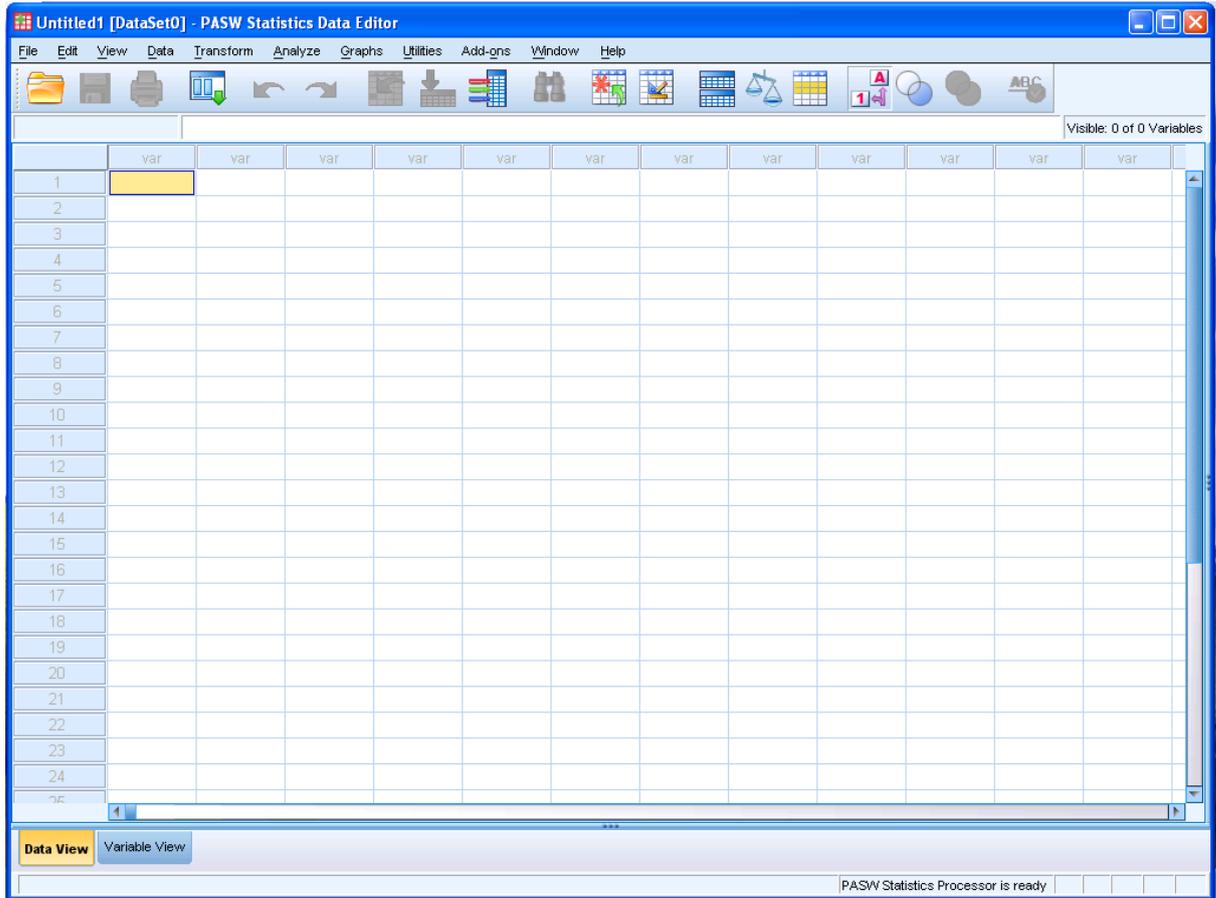
## Starting SPSS

SPSS was renamed PASW Statistics in 2009, but most people continues to call it SPSS. In 2010 it was renamed to SPSS: An IBM Company. In this text all mention to SPSS means either earlier SPSS versions or PASW version 18 or SPSS: An IBM Company. I am using PASW version 18 throughout in this text.

You may find other free books helpful, and to start with Tyrell (2009) is useful.

When you start SPSS (how this is done may vary, for example on Windows may select it from a menu of all programs, but there could be an icon on the desktop) you will be asked what you want to do, the default option is to open a datafile. You can select one of these options (one of them is the tutorial), but if you hit ESC you will get the screen as seen in Figure 1.

Figure 1: Data screen



SIMPLY CLEVER

ŠKODA



We will turn your CV into an opportunity of a lifetime



Do you like cars? Would you like to be a part of a successful brand? We will appreciate and reward both your enthusiasm and talent. Send us your CV. You will be surprised where it can take you.

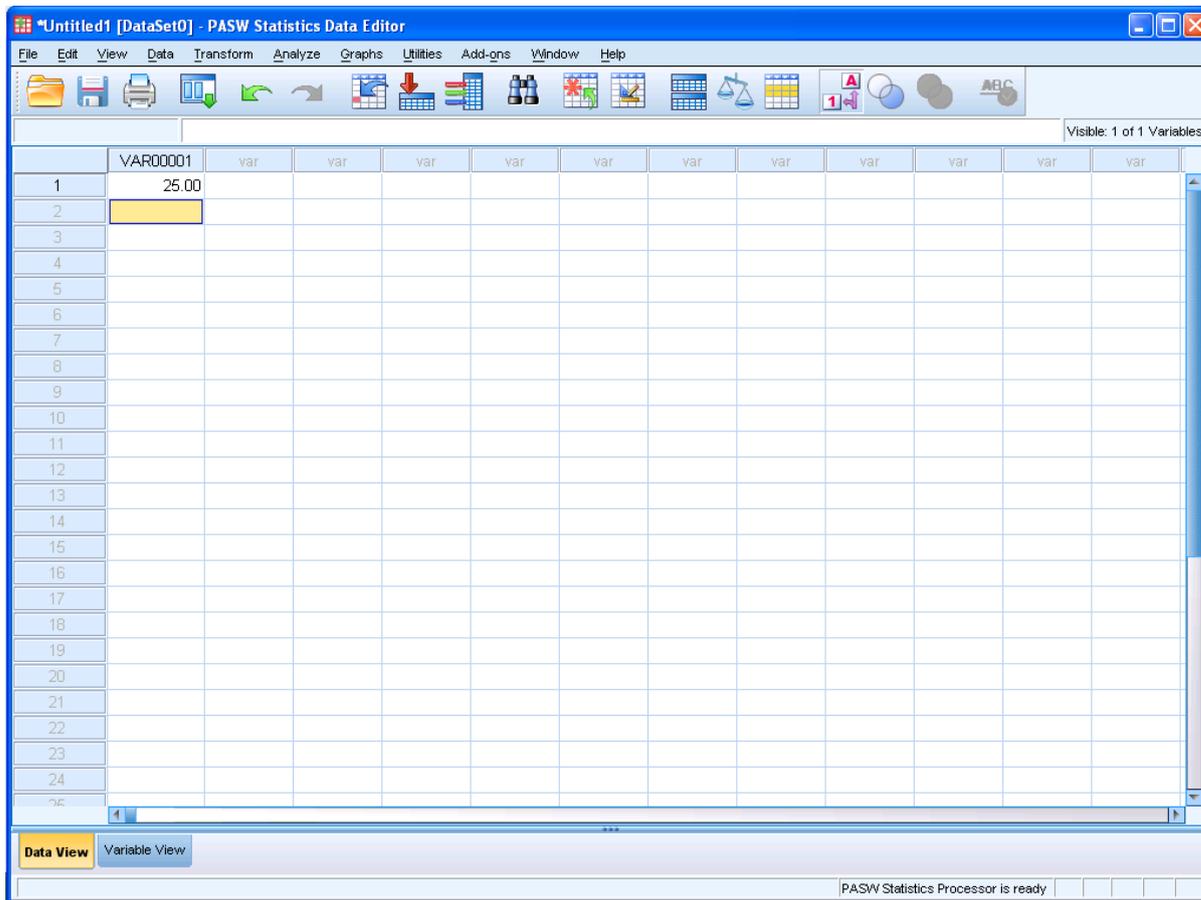
Send us your CV on [www.employerforlife.com](http://www.employerforlife.com)



### Data entry in SPSS

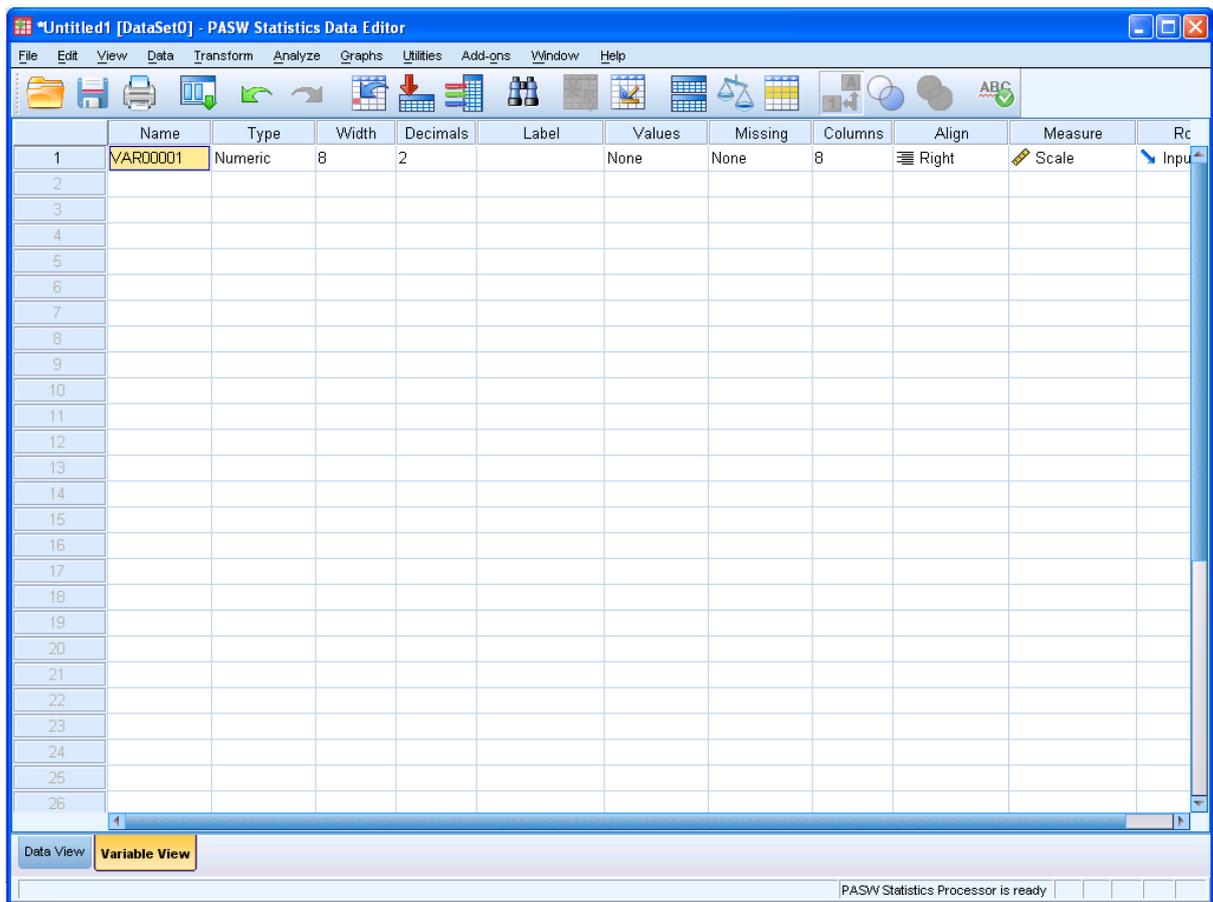
Note there is a data view, where you can enter or edit data, and a variable view here you create new variables or edit existing ones. You can simply enter data (numbers, dates, names etc.) into the data window. If you try (for example) entering an age of a subject, say 25, you will find the data is changed to 25.00.

**Figure 2: Entering data**



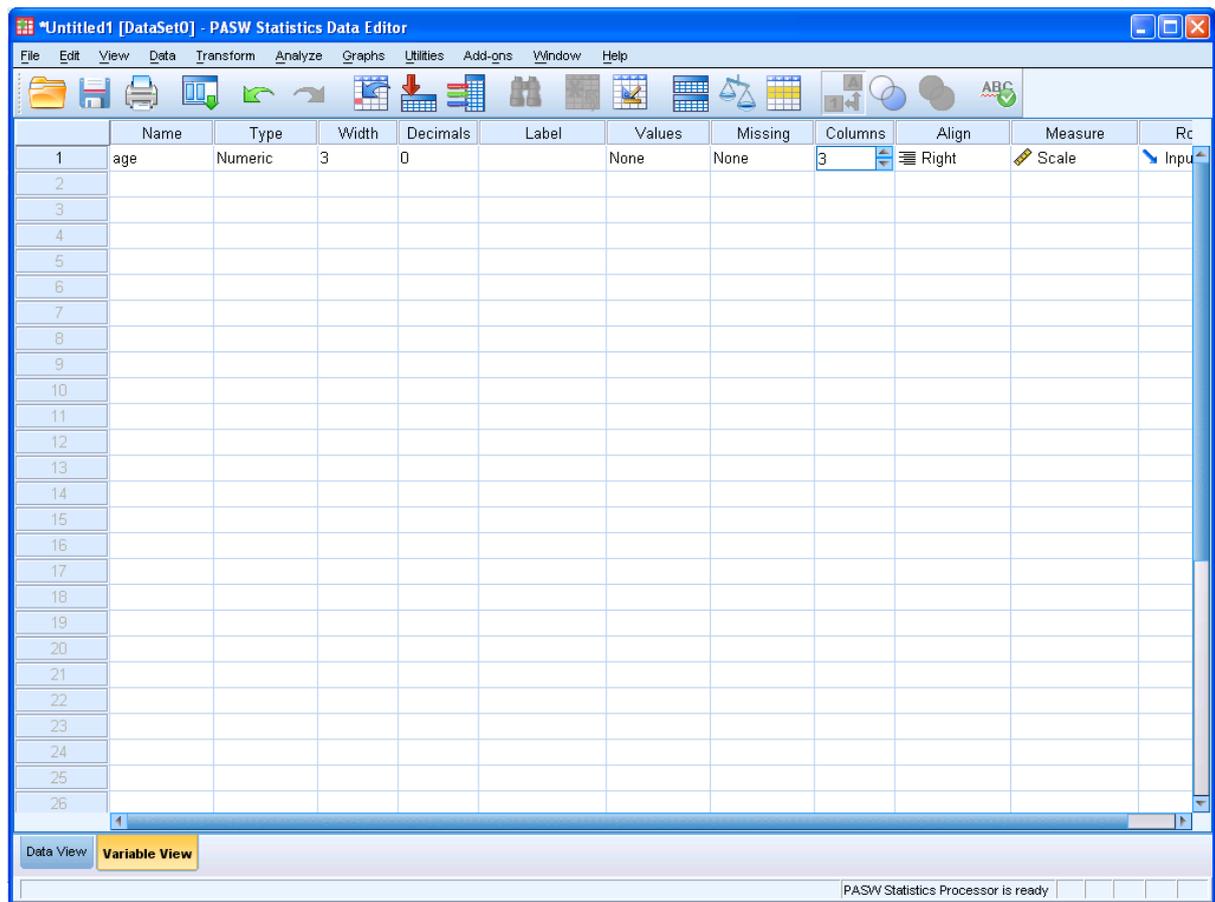
If you switch to variable view by clicking on the tab labelled Variable View, you will see the screen as in Figure 3. Note that the variable name is VAR0001, it is defined as a number and given a default value of length 8, 2 decimal places and column size of 8. Most of this is suboptimal, it would be better for the name to be something like “Age” and since we do not use decimal places for ages, and most people live less than millions of years, a width of 3 (ages of 100 are possible) and no decimal places. The column width is simply how wide the display is on the data tab.

**Figure 3: Variable view**



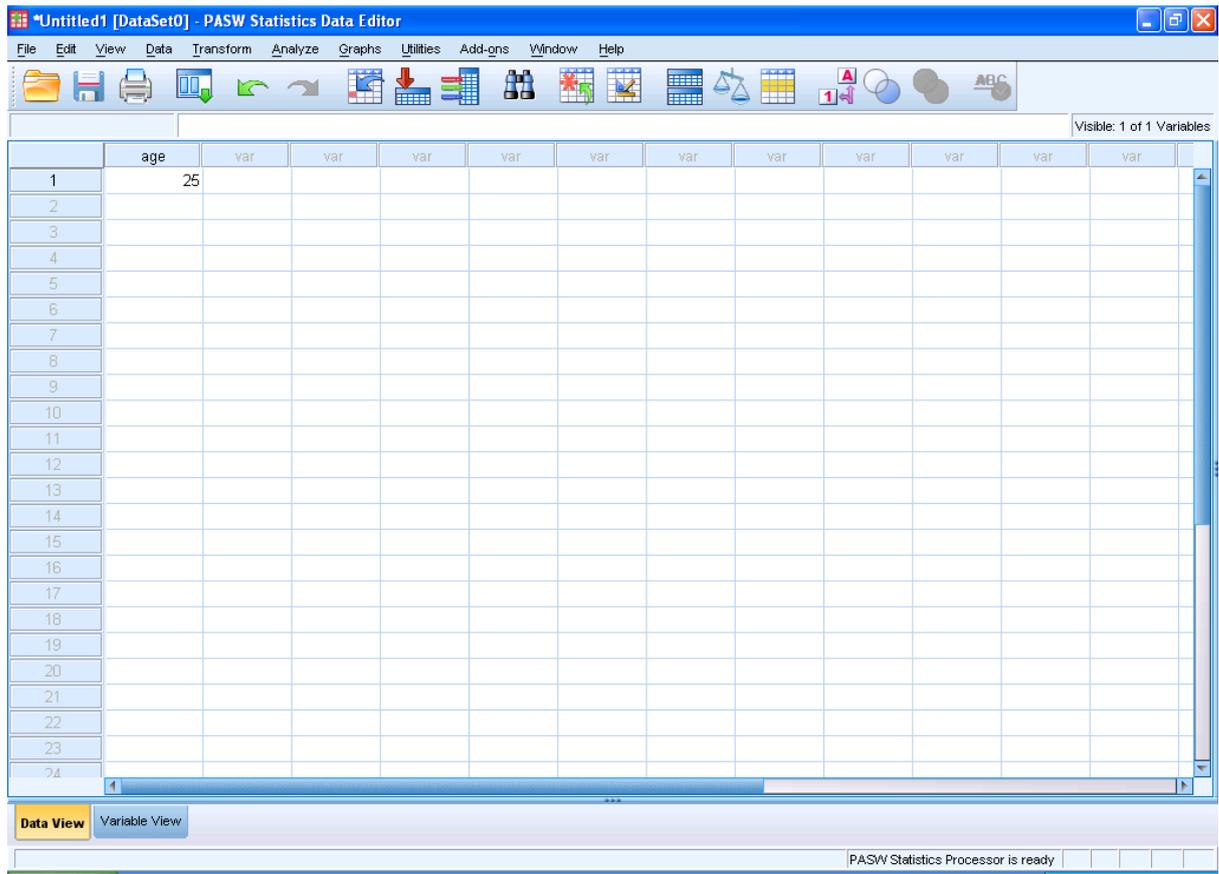
Rather than accept these you can define your new variables in Variable View as seen in Figure 4. Here I have named the variable age, defined it to be a number, of width 3 and no decimal points.

**Figure 4:** Defining a variable



If you click to on Data View the display now seems much more sensible, see Figure 5.

Figure 5: Entering data



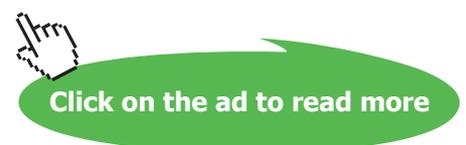
Sweden  
Sverige

Linköping University –  
innovative, highly ranked,  
European

Interested in Computer Science? Kick-start your career  
with an English-taught master's degree.

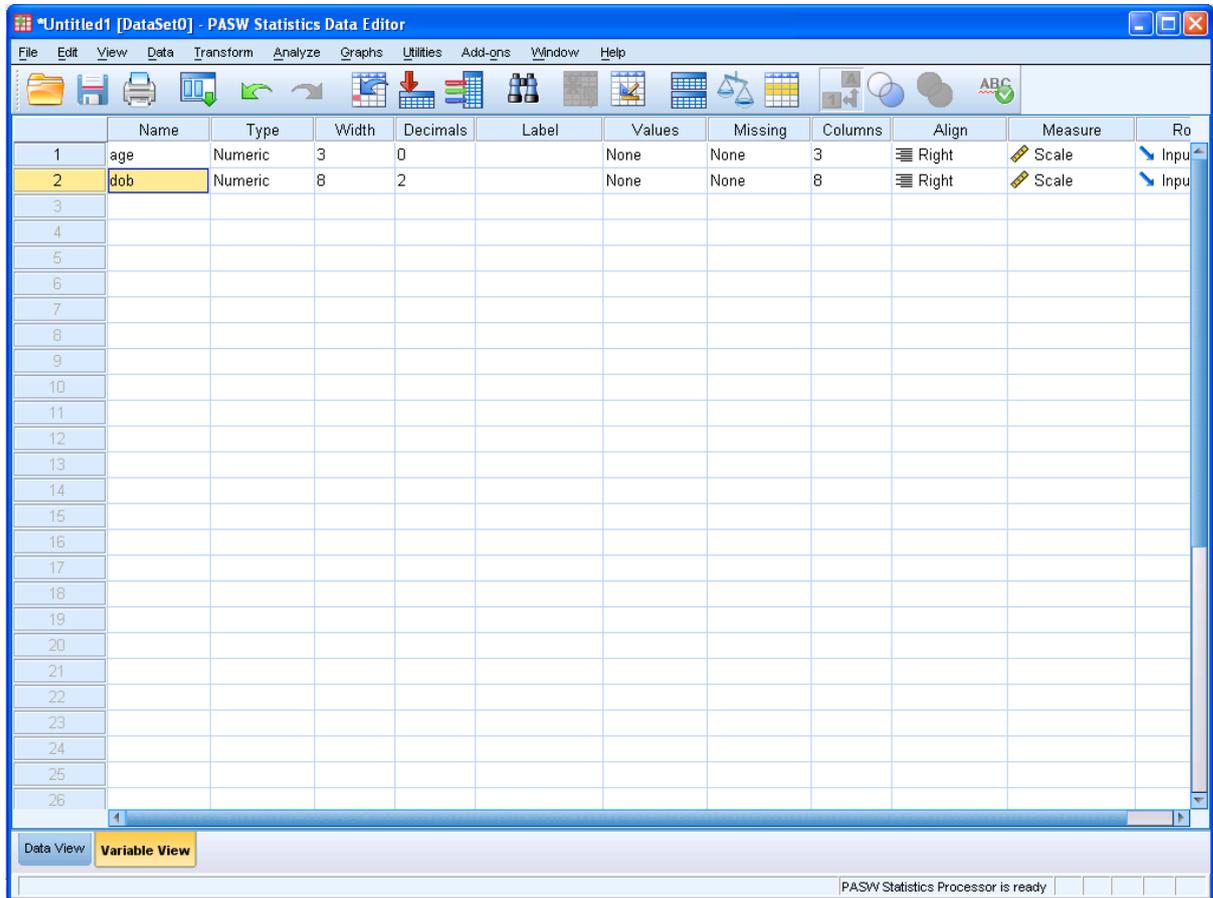
→ Click here!

**li.u** LINKÖPING  
UNIVERSITY

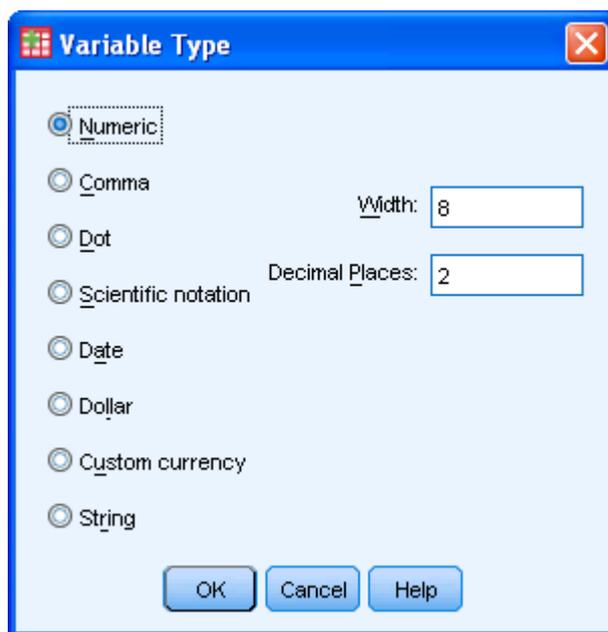


There are many other types of data. If in Variable View you click on the three dots on the right of the Type column (here I have created a new variable called dob), see Figure 6, you will see the dialogue box as in Figure 7.

**Figure 6:** Exploring data type

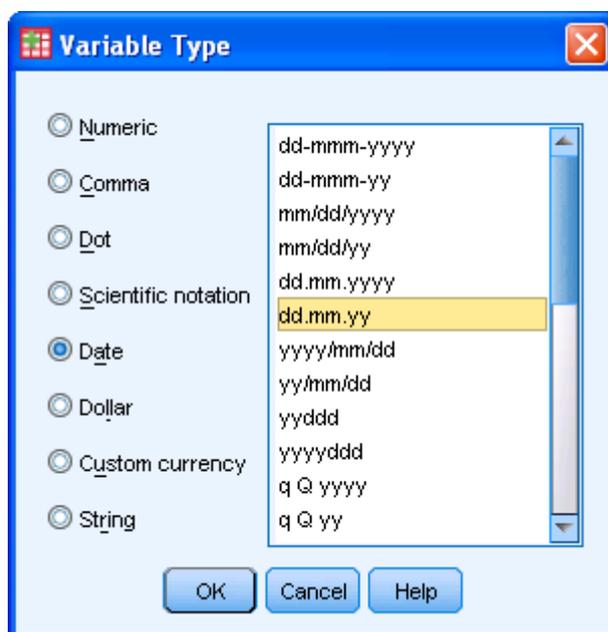


**Figure 7:** Data type dialogue box



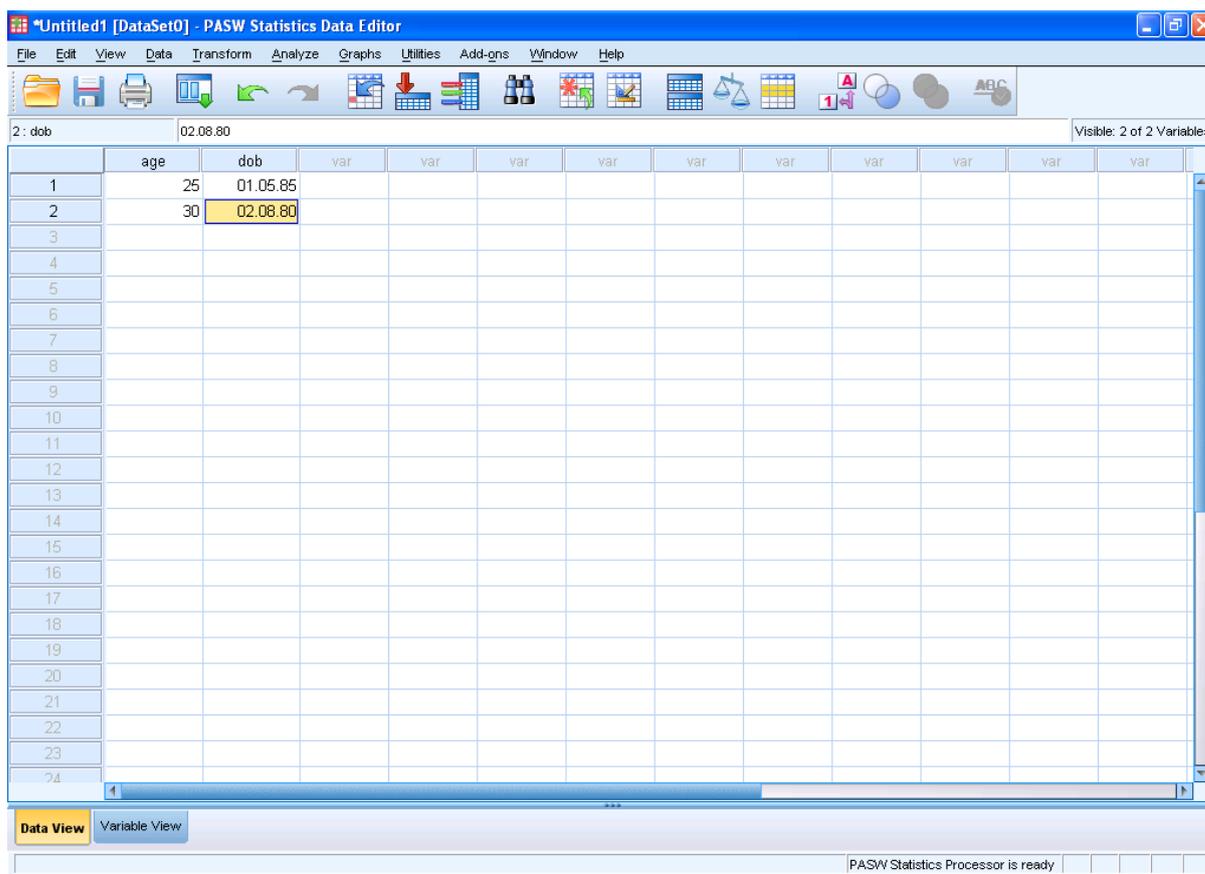
You can change the default (Numeric) to many other types, though in practice I only ever use Numeric, Date and String (alphanumeric, i.e. letters and/or numbers). Below I change variable dob to be a Date, which can be in a variety of formats, see Figure 8. I have chosen to enter dates in UK format as numbers separated by dots.

**Figure 8:** Date format



Then you can enter dates as in Figure 9.

Figure 9: Entering dates



If you change the date format the data display also changes, for example if you select as below (Figure 10) then the data will appear as in Figure 11.

Figure 10: Changing date format

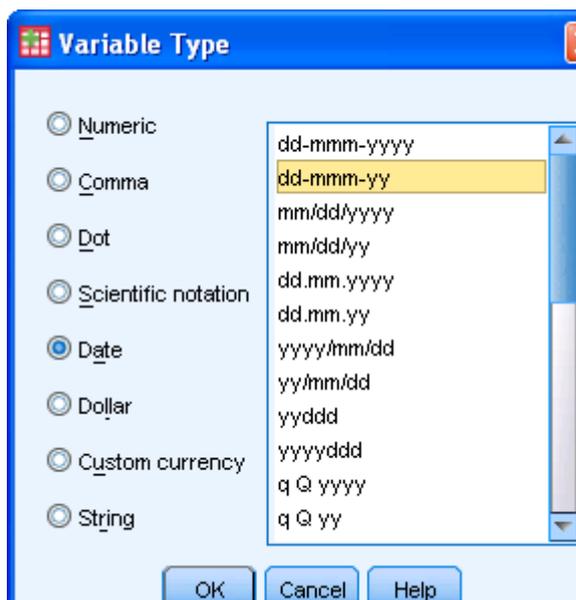
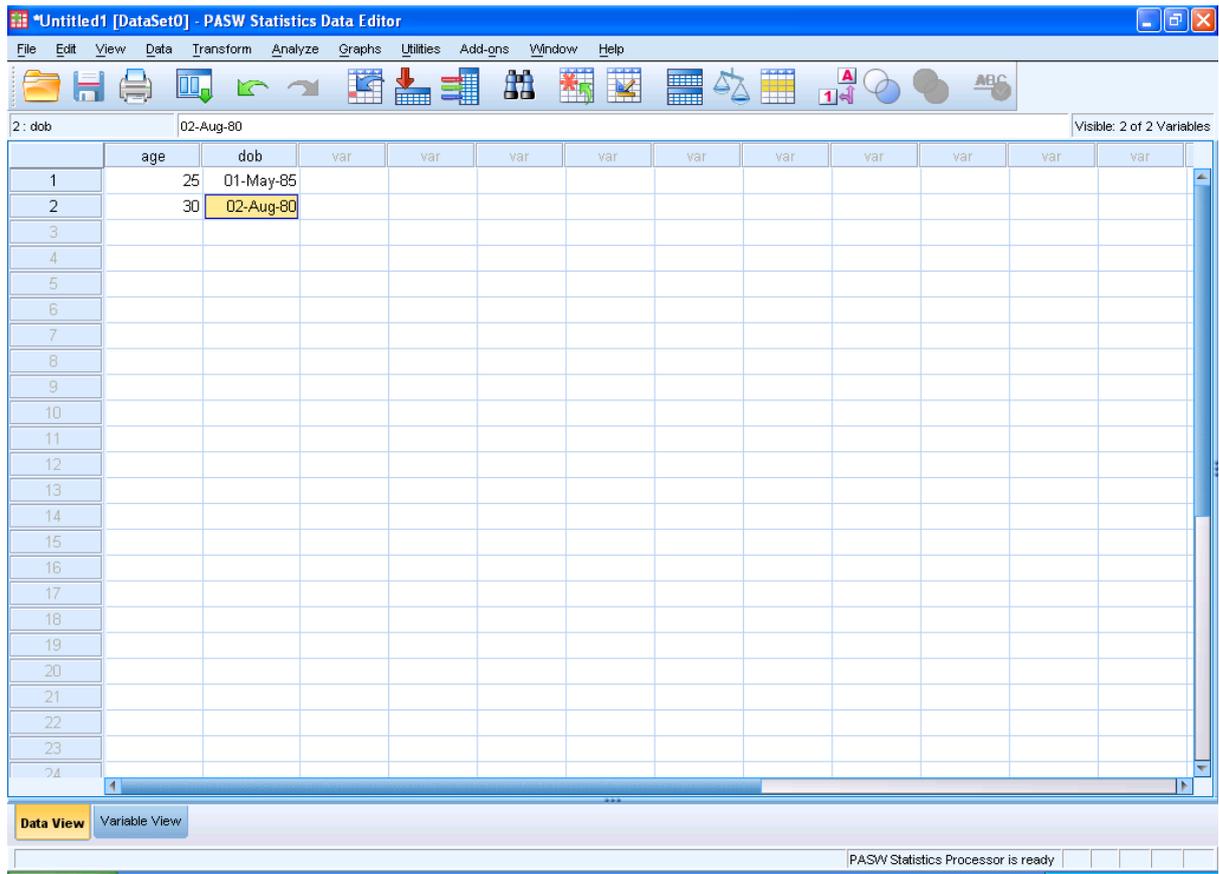


Figure 11: Date in Data View



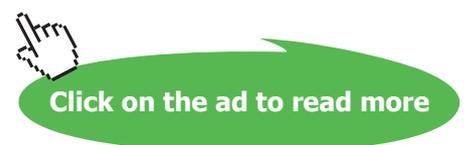
I joined MITAS because I wanted **real responsibility**

The Graduate Programme for Engineers and Geoscientists  
[www.discovermitas.com](http://www.discovermitas.com)

**Month 16**  
 I was a construction supervisor in the North Sea advising and helping foremen solve problems

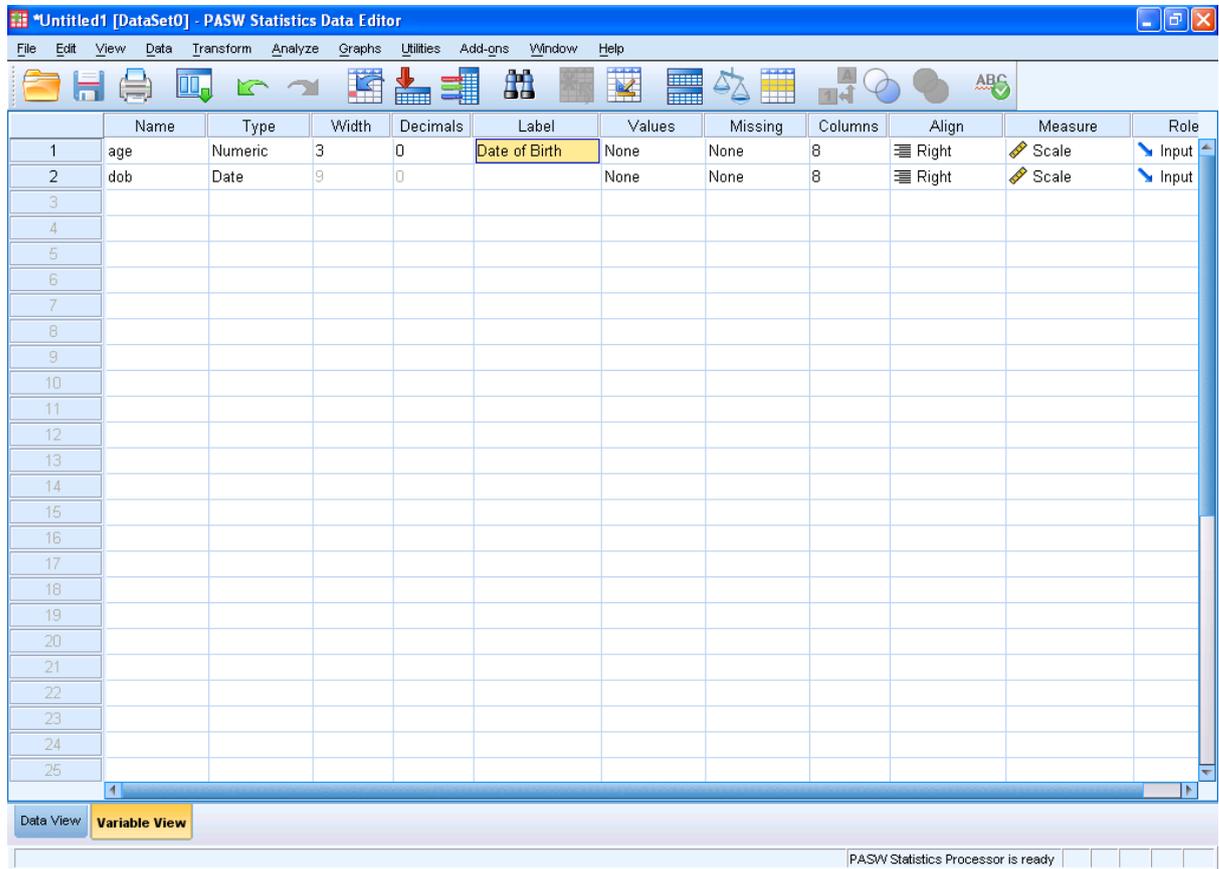
Real work  
 International opportunities  
 Three work placements

**MAERSK**

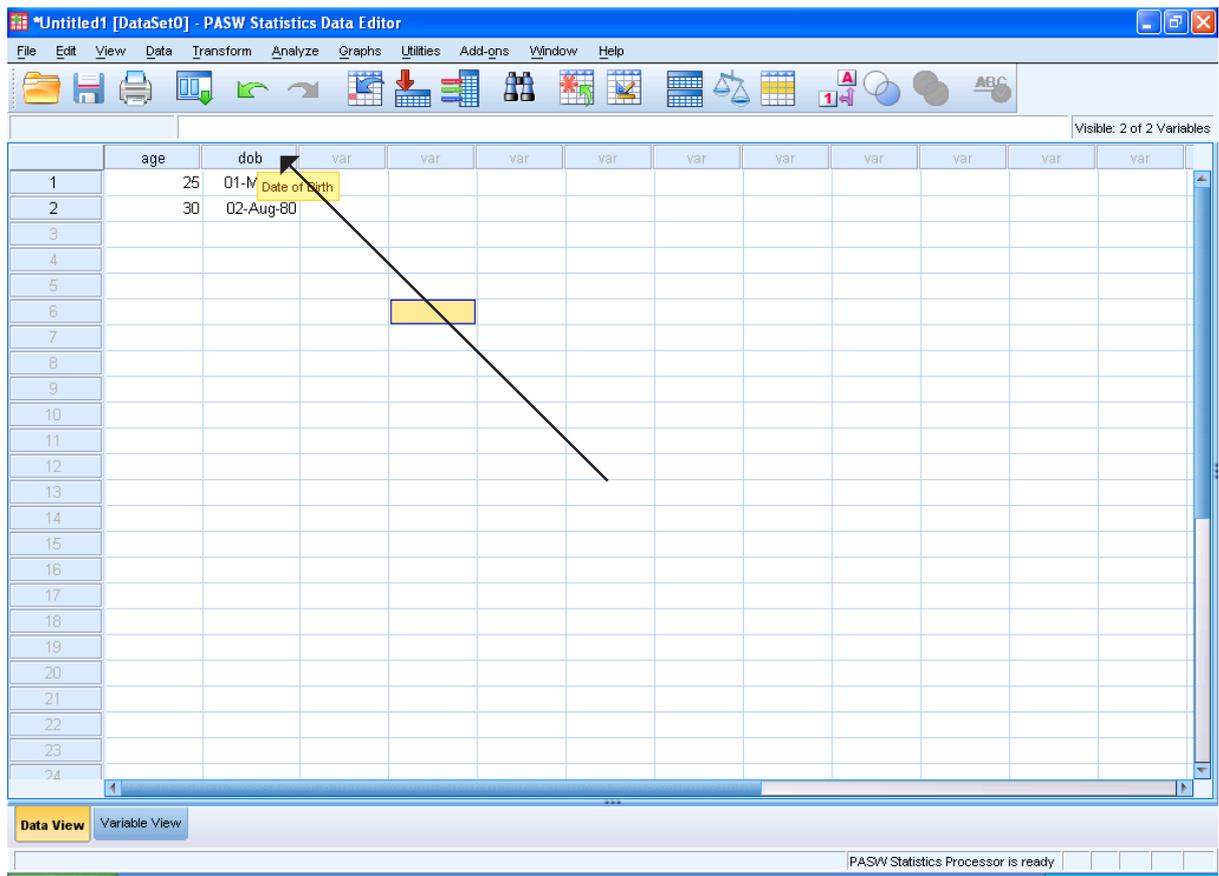


However what is dob? In Figure 12 I give a label to show what this means. This is necessary as you cannot use spaces in variable names, but labels are much less restricted. This label can be seen when you move the mouse over the title in the second column in Data View.

**Figure 12: Giving a label**

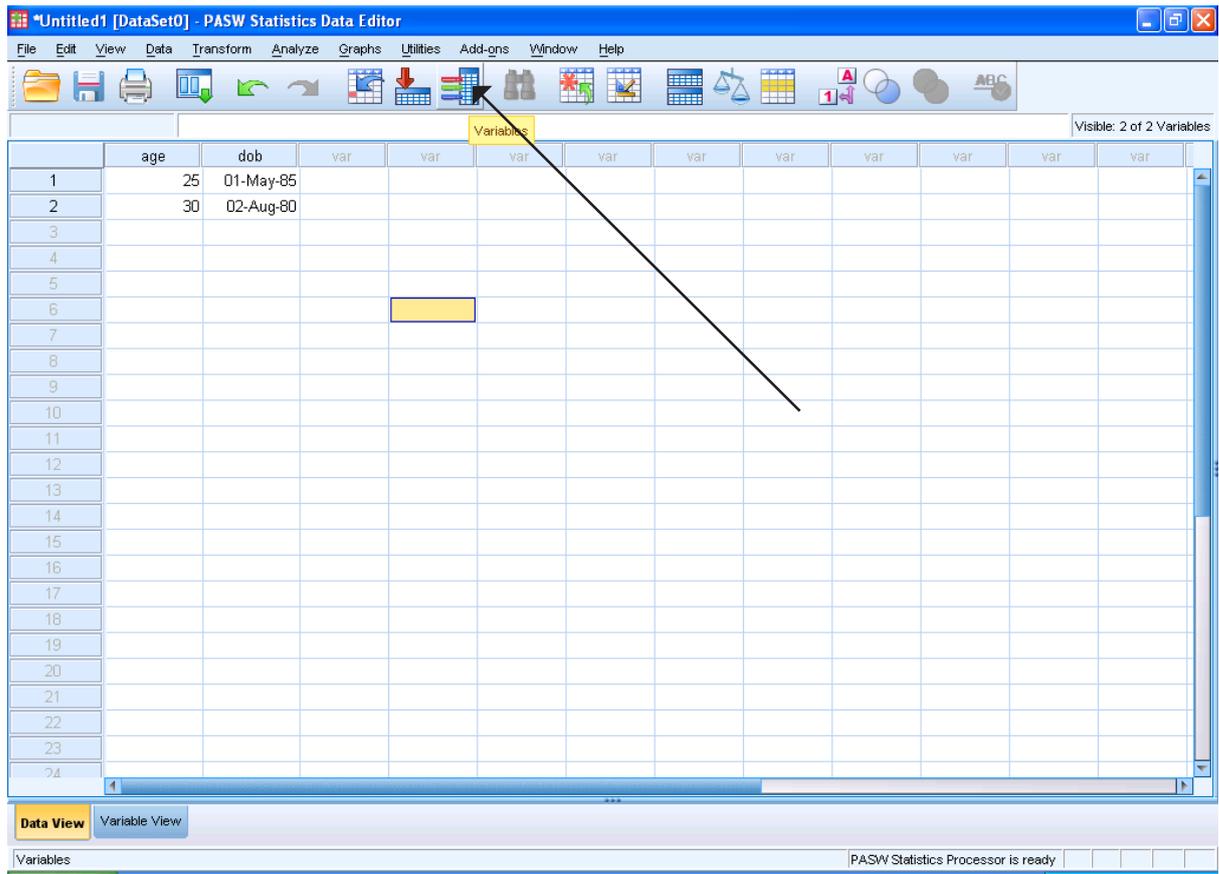


**Figure 13: Seeing label**



You can see all aspects of variables by hitting on the Variables button (shown as a column icon) as being done in Figure 14.

Figure 14: Variable button



“I studied English for 16 years but...  
...I finally learned to speak it in just six lessons”  
Jane, Chinese architect

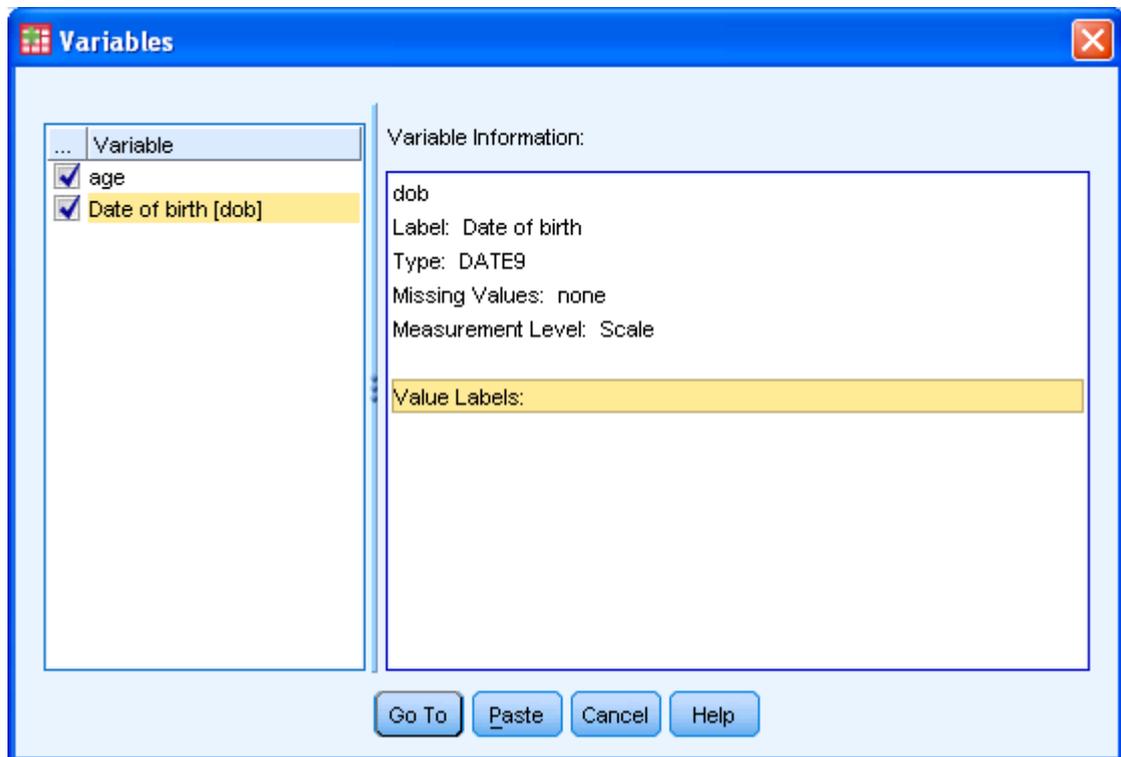
ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



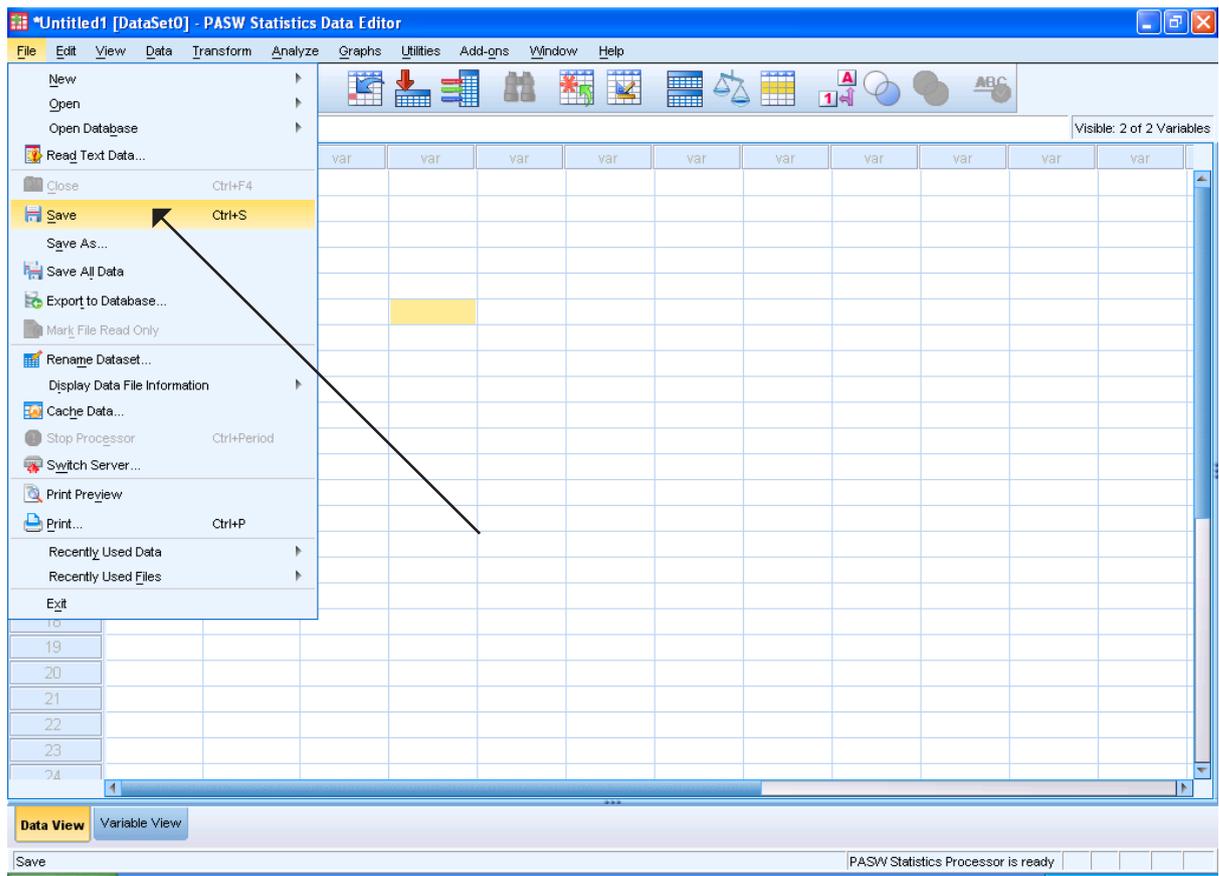
Then you see all the variables I have created so far, see Figure 15. Note the variables with a label is described first by name, and then by its label.

**Figure 15:** Variables



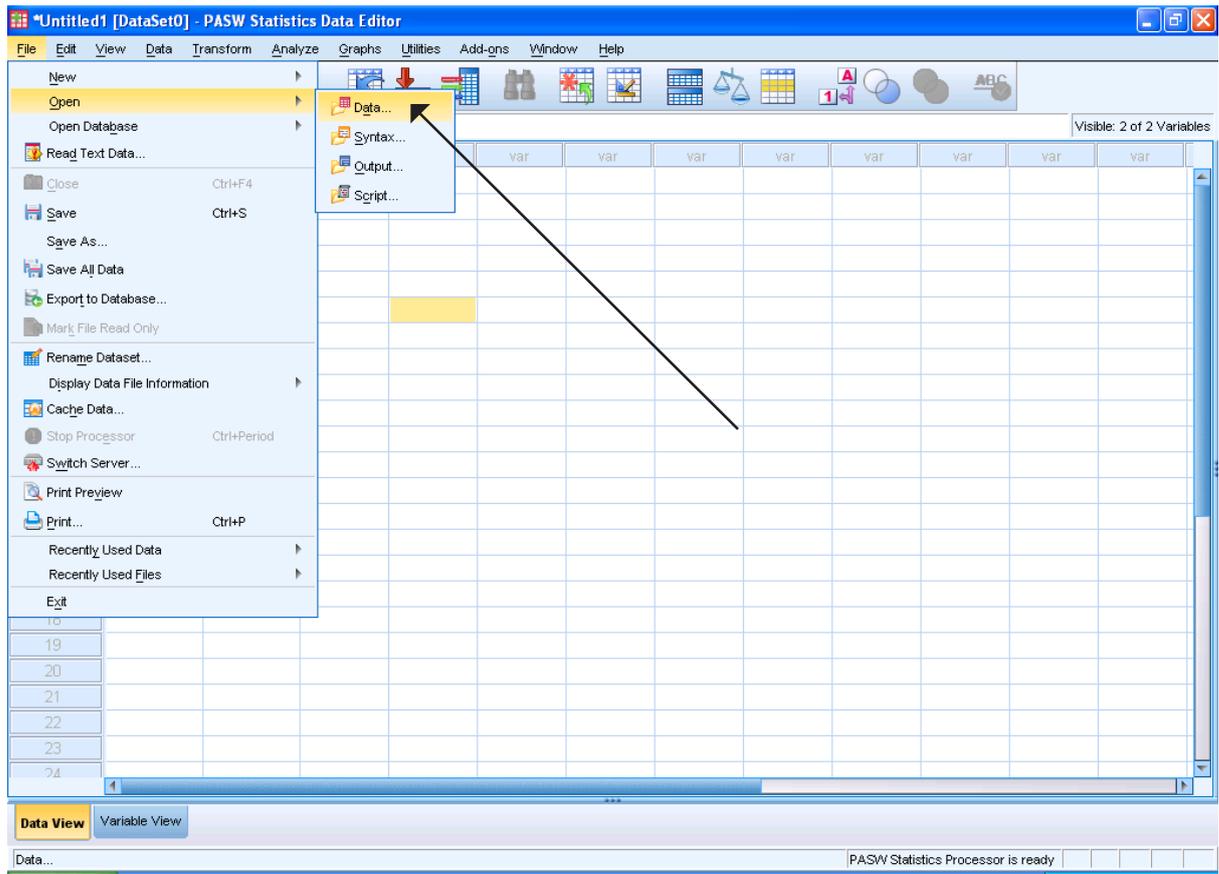
I can save the file at this stage. SPSS allows several types of file to be saved, here I am saving a Data file, see Figure 1 6.

Figure 16: Saving data



I can leave SPSS now, and later pick this data file up, see Figure 17, note there are other types of files you can open, but for the moment we will just look at Data.

Figure 17: Opening a datafile





In the past four years we have drilled

# 89,000 km

That's more than **twice** around the world.

**Who are we?**  
We are the world's largest oilfield services company<sup>1</sup>. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

**Who are we looking for?**  
Every year, we need thousands of graduates to begin dynamic careers in the following domains:

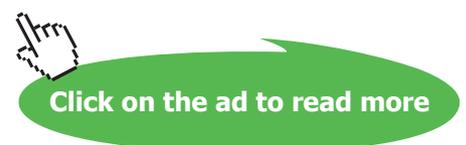
- Engineering, Research and Operations
- Geoscience and Petrotechnical
- Commercial and Business

**What will you be?**

**Schlumberger**

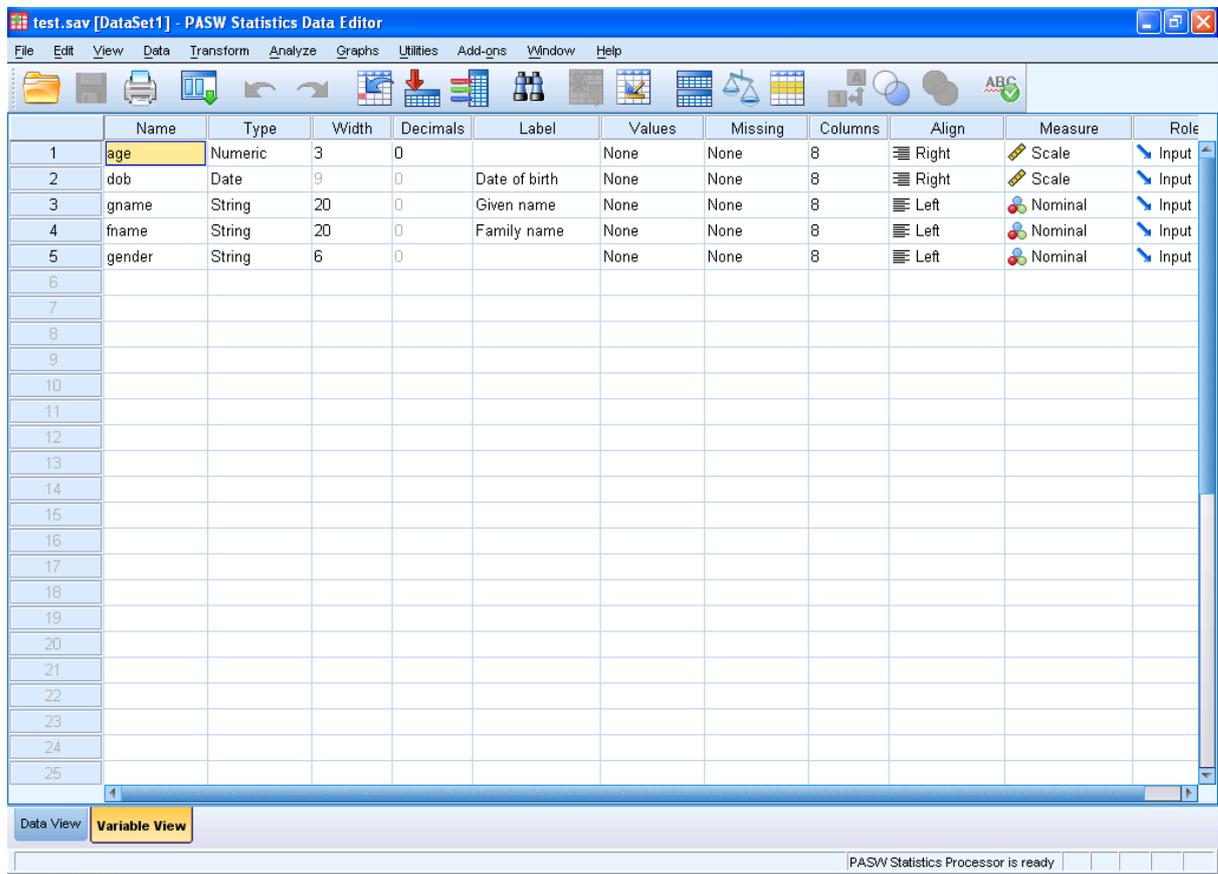
[careers.slb.com](http://careers.slb.com)

<sup>1</sup>Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.



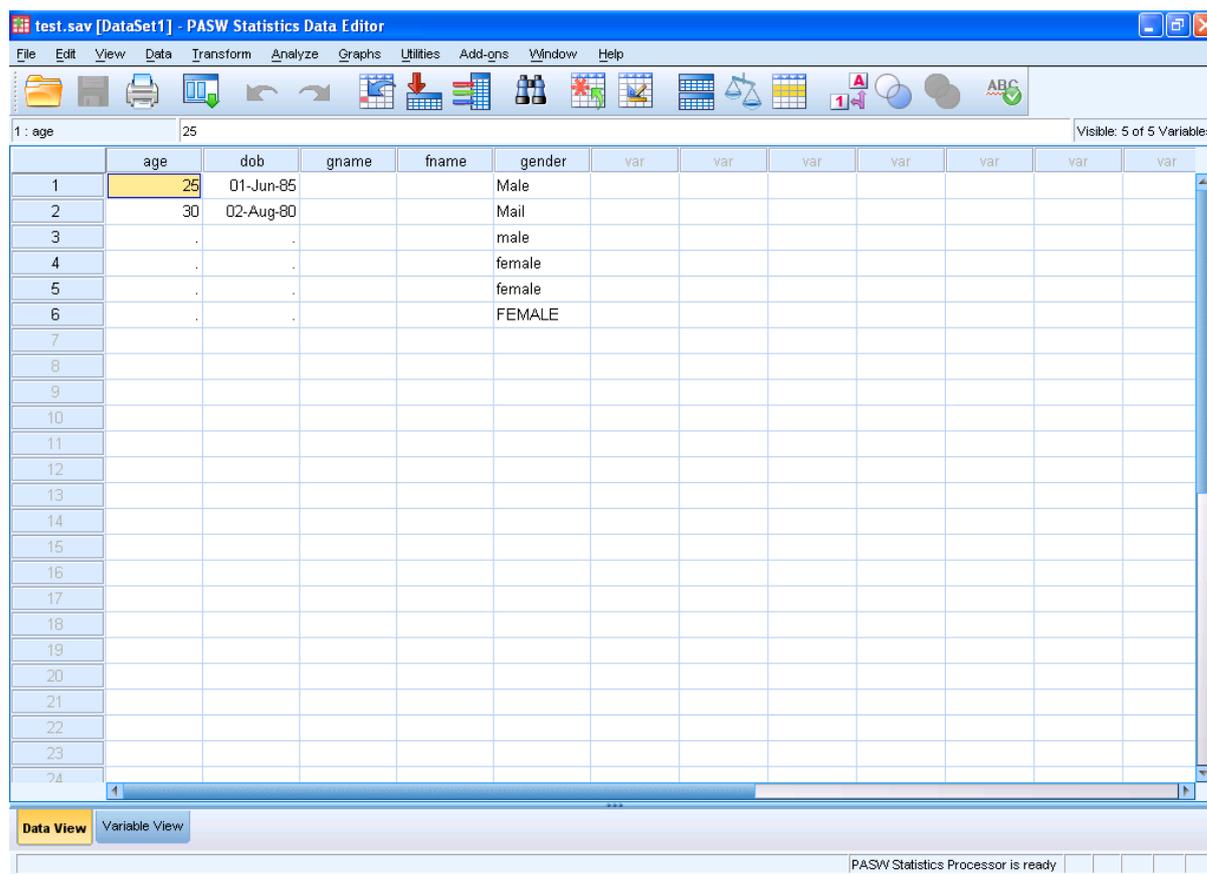
I can now add some new variables, see Figure 18. Here I add two string variables which identify names of subjects (though in many studies you would not want to do this for reasons of confidentiality) and one for gender.

**Figure 18:** Adding more variables



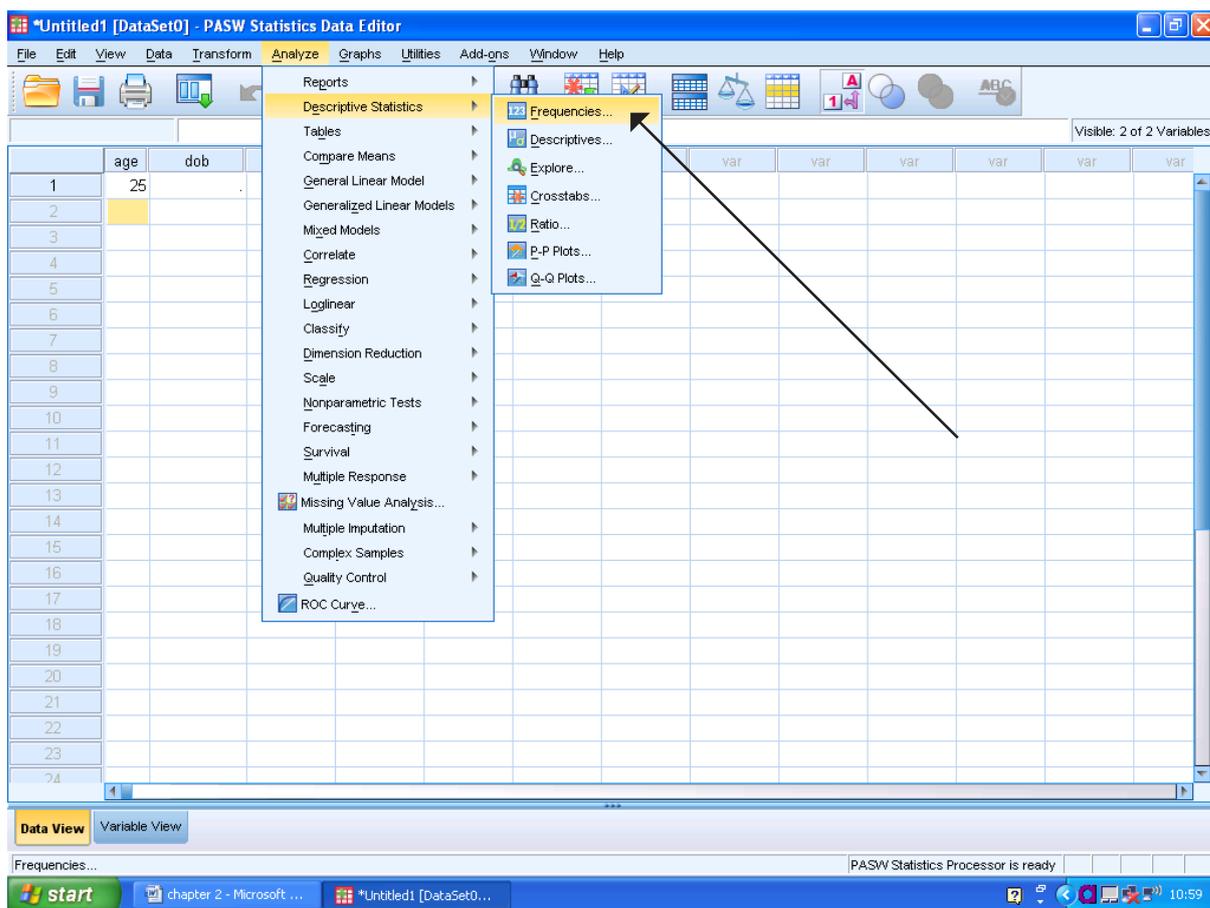
I could have used a string variables for gender, and written down Male and Female for each subject. I have done this below, see Figure 19.

Figure 19: Gender as string variable



I have used a lot of screen dumps to demonstrate what SPSS looks like in practice, but I will start now using the convention of **Pull down** -> **Further Pull down**, to indicate interactive commands in SPSS, where the underlined letter means you can use the Alt key combined with that letter to get that dialogue box (if you prefer keyboard strokes to the mouse). So for example the sequence shown in Figure 20 is **Analyze** -> **Descriptive Statistics** -> **Frequencies**. In this chapter I will continue to show the screen dumps for all interactions but in later chapters I not show the initial menus and pull downs where this is not necessary.

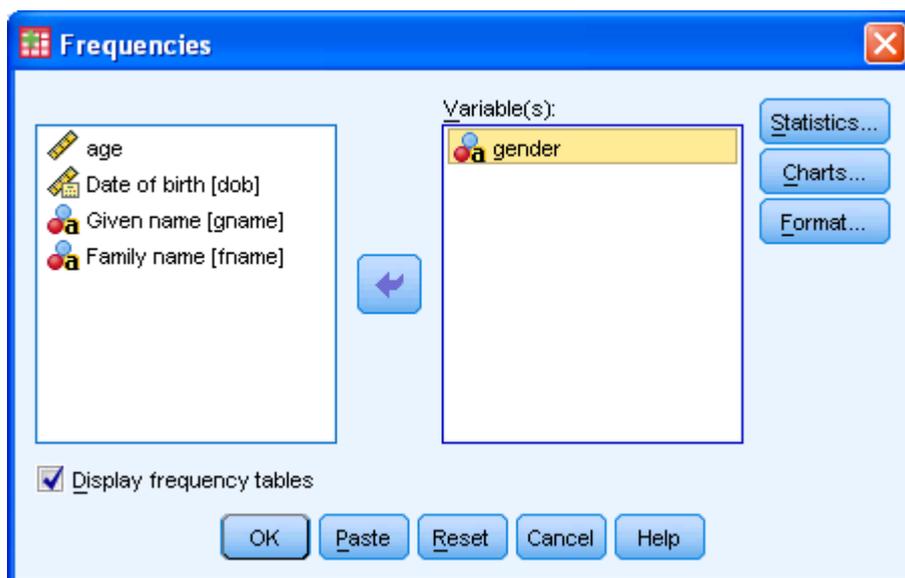
Figure 20: Example of interactive command



I can now use SPSS to count the males and females (not necessary here, but think if you had thousands of subjects!).

You then get a dialogue box as in Figure 21. Here you can enter gender as a variable. You then hit the OK button and get a new Output window. This is separate from the Data window, and all output goes into this. It can be separately saved and later opened if you want to keep the results.

Figure 21: Frequencies dialogue box



The Output window contains the frequency table of gender, and above it is shown how many valid cases there are (i.e. for how many subjects we have information on gender, here all of them). Note there are more than two types of gender since in a string variable capitalised (e.g. Male) is different to lower or upper case, and one (Mail) is misspelt. To avoid these data errors it is better to code nominal data (see next chapter), and best to use numbers as codes. For example Male=0, Female=1. However how would you remember which code relates to which gender? You can use value labels in the Variable tab, see Figure 23 and the dialogue box Figure 24, where I have entered Male as 0 already and about to enter Female as 1, what I need to do next is click Add to add this value, then OK to place both value labels into the variable.

WHILE YOU WERE SLEEPING...

DUKE  
THE FUQUA  
SCHOOL  
OF BUSINESS

www.fuqua.duke.edu/whileyouweresleeping



Figure 22: Output window

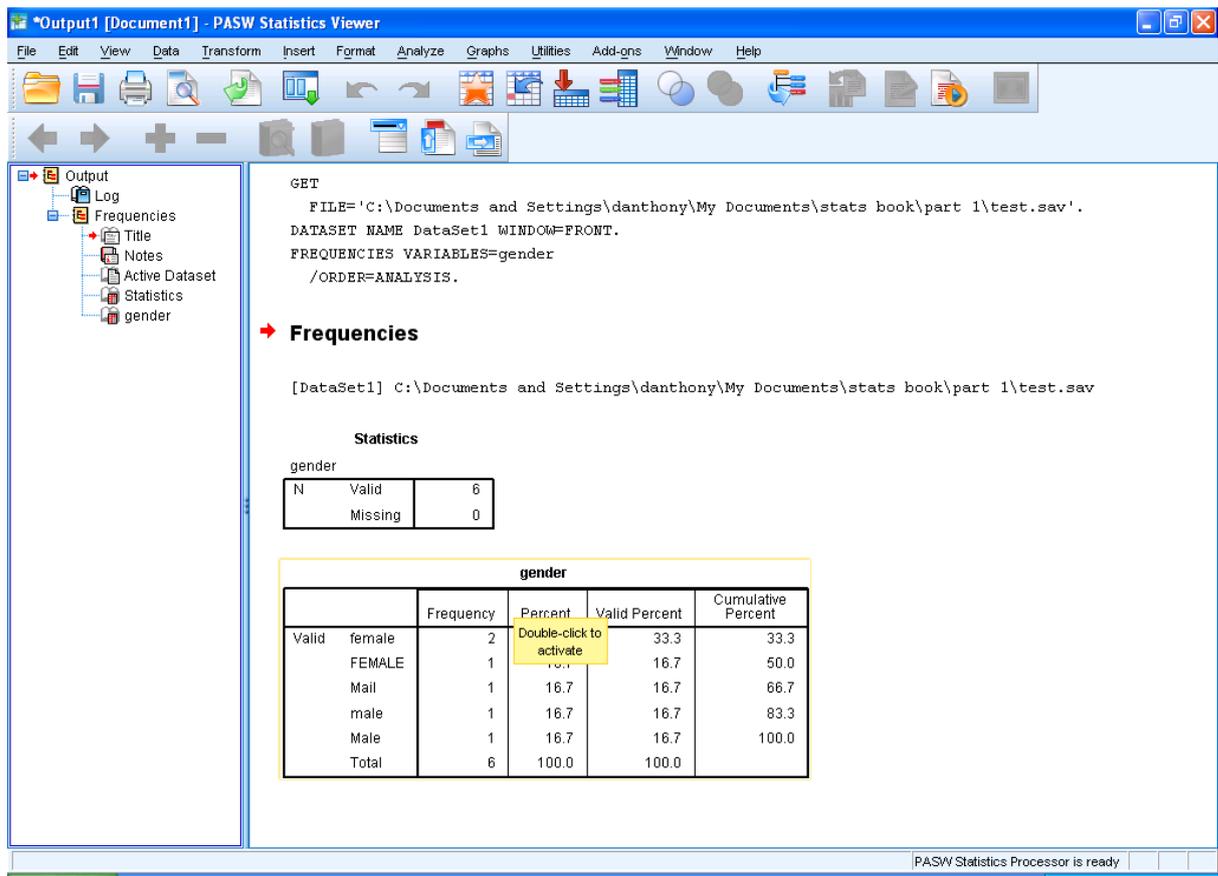


Figure 23: Adding value labels

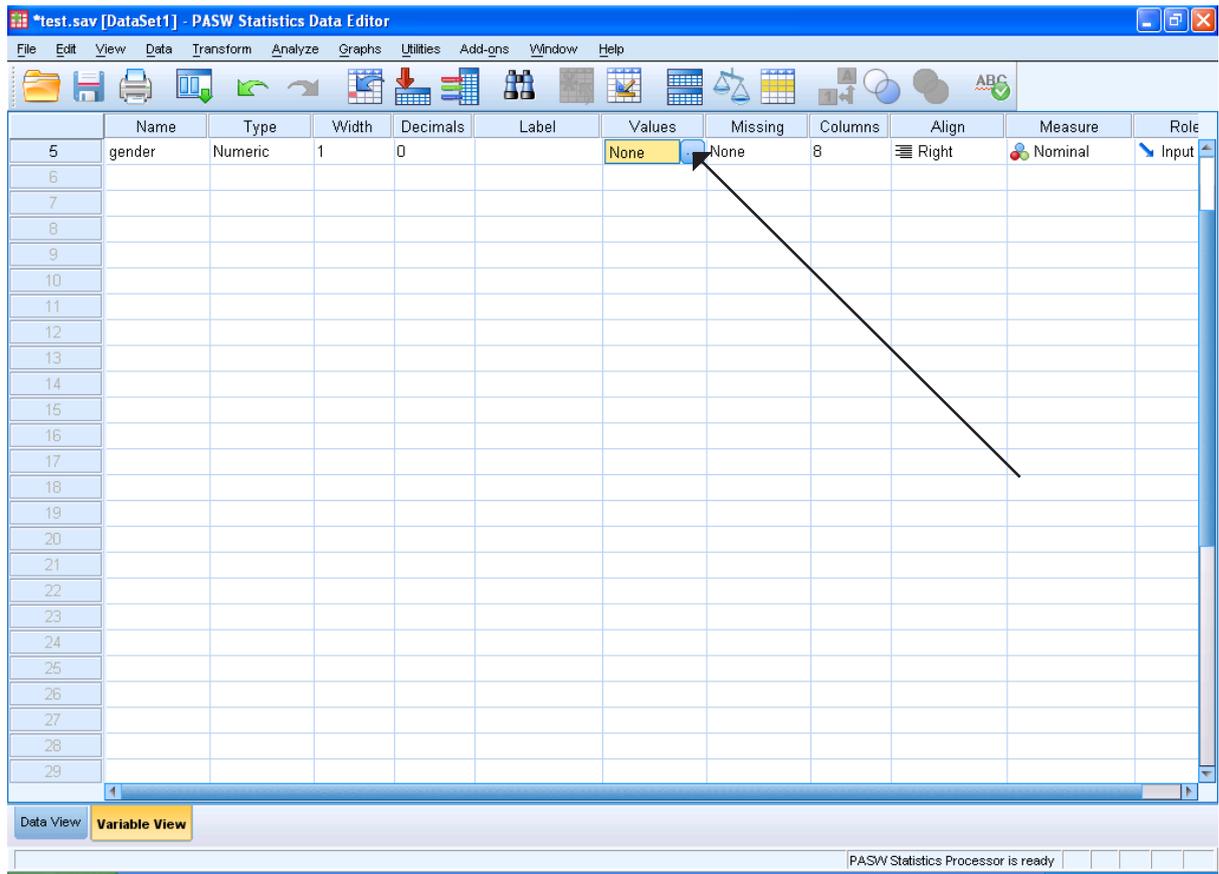
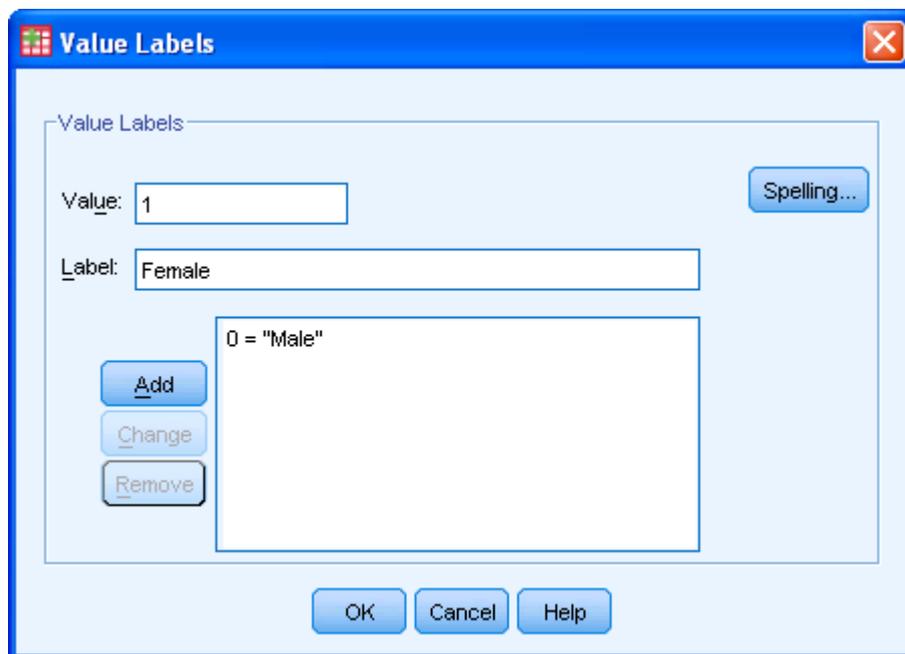
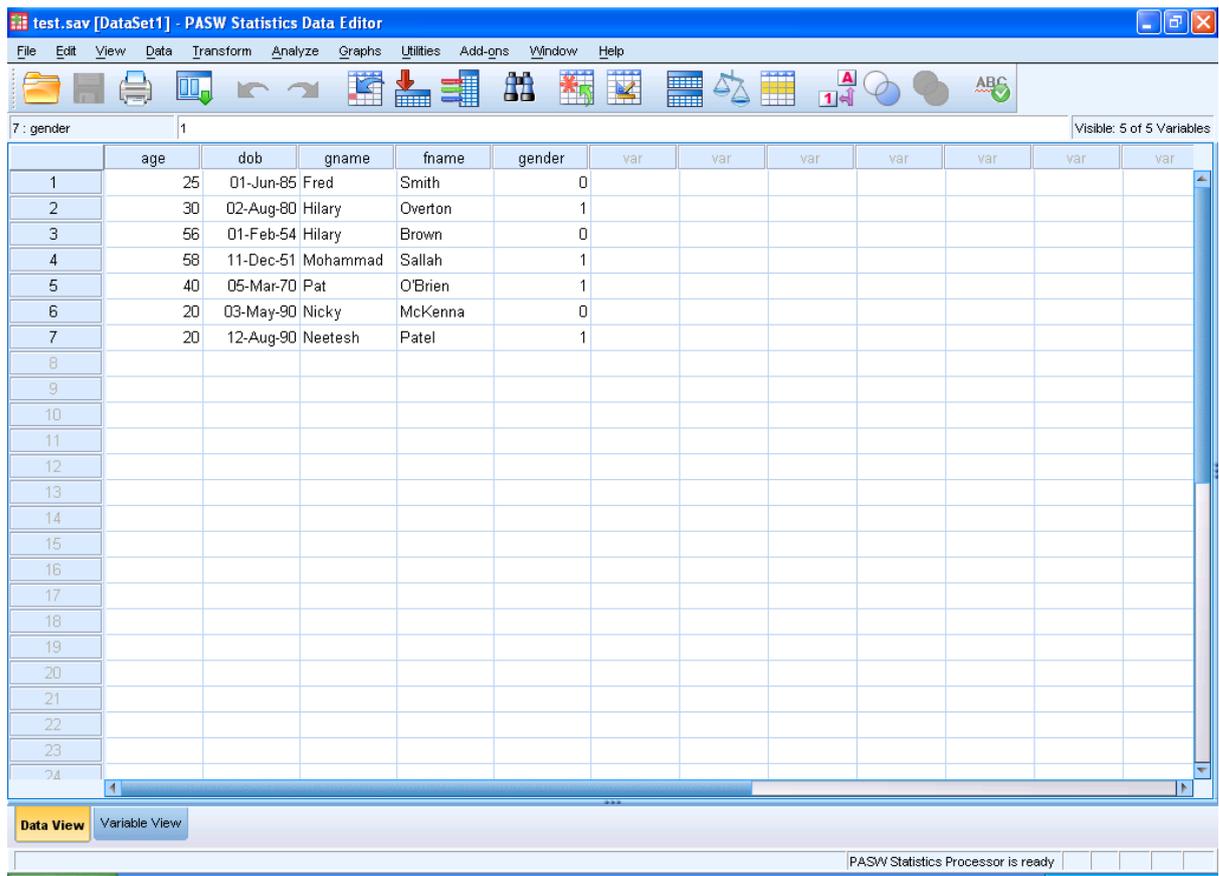


Figure 24: Value labels dialogue box



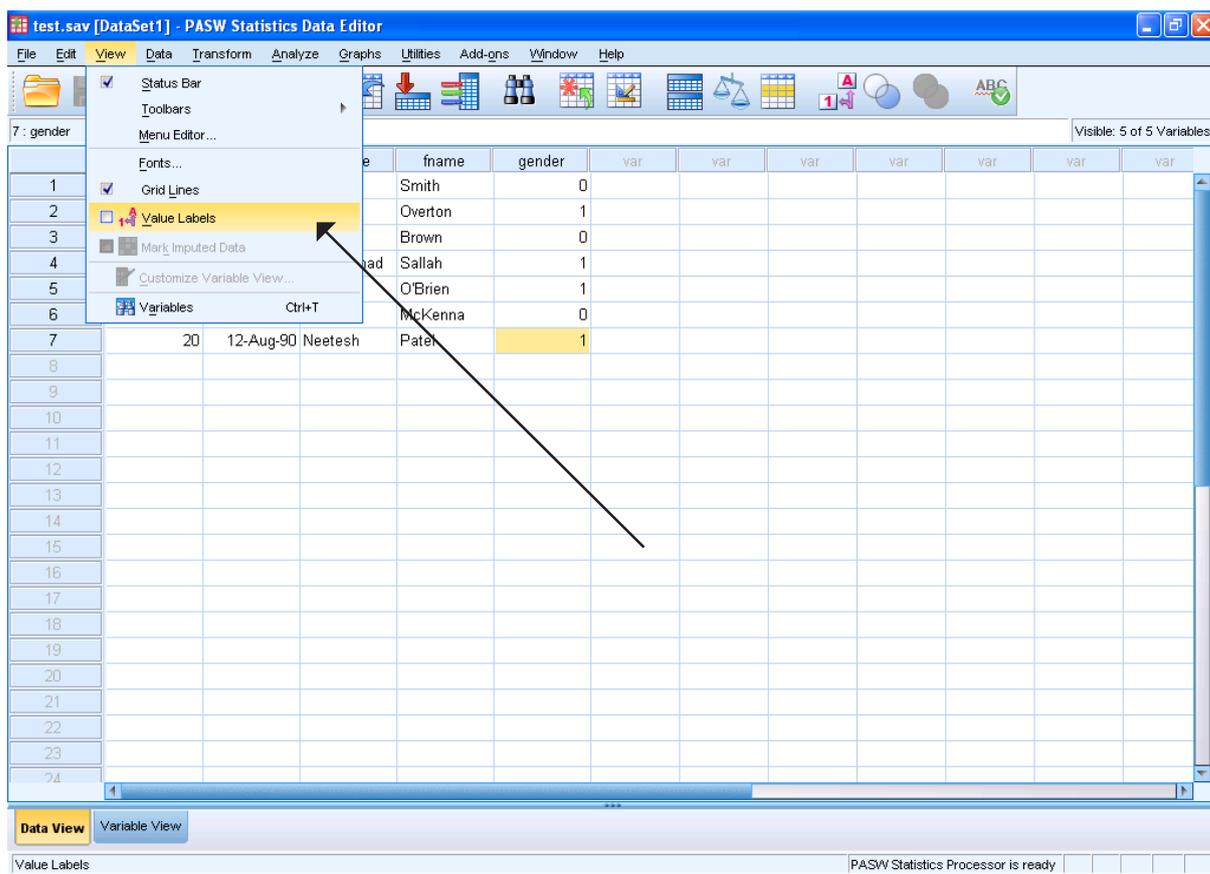
If I add more data I could see in the Data tab the results as in Figure 25. Note you cannot guess some people's gender from their given name!

Figure 25: More data



If you want to see what the labels represent you can select **View->Value Labels** (Figure 26) and then you will see Figure 27.

Figure 26: Viewing labels



Excellent Economics and Business programmes at:



**university of  
groningen**



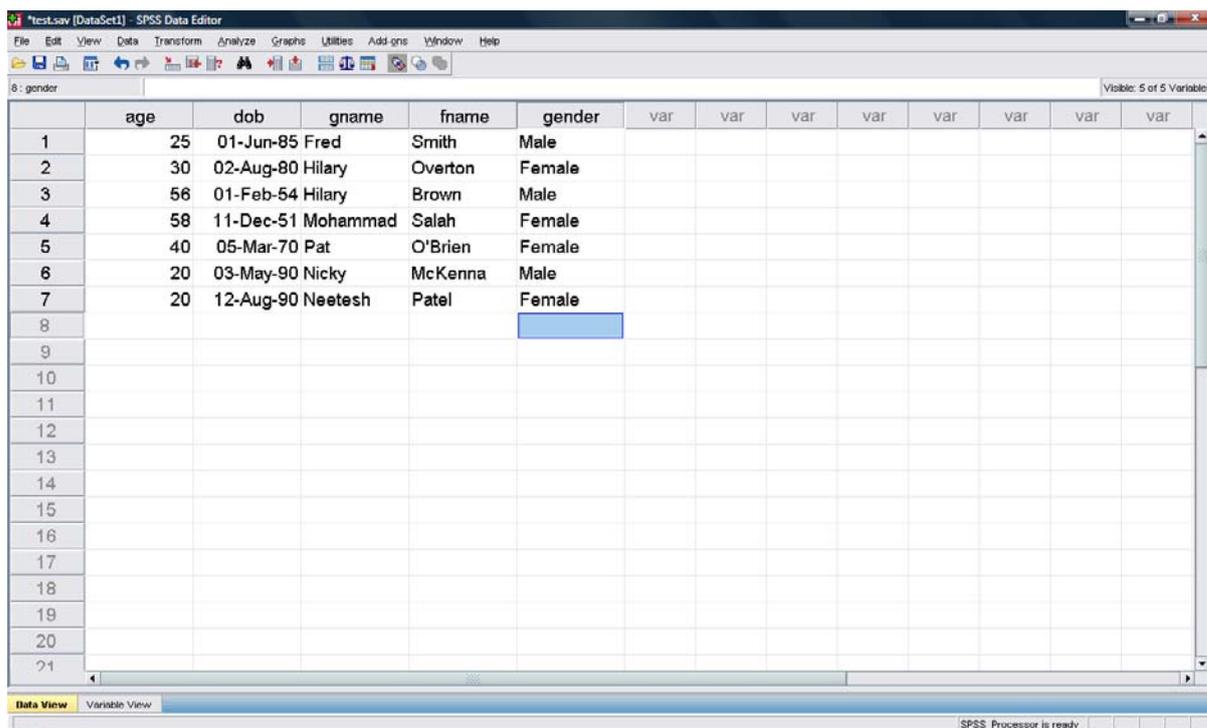
**“The perfect start  
of a successful,  
international career.”**

**CLICK HERE**  
to discover why both socially  
and academically the University  
of Groningen is one of the best  
places for a student to be

[www.rug.nl/feb/education](http://www.rug.nl/feb/education)



**Figure 27:** Showing labels in variable gender



If we redo the frequencies for gender we get Table 1.

**Table 1:** Frequencies for gender with clean data

		gender			Cumulative
		Frequency	Percent	Valid Percent	Percent
Valid	Male	3	42.9	42.9	42.9
	Female	4	57.1	57.1	100.0
Total		7	100.0	100.0	

**In this chapter you have learned:-**

- How to enter data
- How to create variables
- How to change variables to suit the data they contain
- How to code data for greater data integrity
- How to view variables, including labels associated with a variable
- How to view frequencies of data in a variable

**References**

Tyrell, S. (2009). *SPSS: Stats practically short and simple*. Copenhagen, Ventus.

# 3 Introducing descriptive statistics

## Key points

- You need to know the *type* of data you have
- Descriptive statistics and graphs are best suited to certain types of data

## At the end of this chapter you should be able to

- Conduct more complex descriptive statistics
- Explore data
- Generate graphs with the graph builder

In this chapter I am going to consider a dataset I created for a small study of the attitudes of students to online learning. For details see the *Appendix Datasets used in this text*, though I repeat most of the information in this chapter as it is your first dataset. In later chapters I will just tell you to open a dataset, and you will find all details of the variables in the Appendix. All the datafiles are downloadable from the BookBoon website.

I gave out a survey questionnaire (see Figure 28 ) over the next two pages (I gave this out as a single sheet double-sided) and entered the data into SPSS.

This survey was conducted in 2005, when some of my nursing students were not very computer literate, many were mature students who were re-entering the education system after many years. I could have used an online method, however I wanted to know about students who would NOT engage with online learning and using IT, so I needed to use a paper based survey.

**Figure 28:** Questionnai

<b>Name</b> (optional, you do not have to fill this in if you prefer to remain anonymous)					
<b>Age</b> (years)					
<b>Gender</b> (please tick)	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; text-align: center;">Male</td> <td style="width: 50%; text-align: center;">Female</td> </tr> <tr> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </table>	Male	Female	<input type="checkbox"/>	<input type="checkbox"/>
Male	Female				
<input type="checkbox"/>	<input type="checkbox"/>				
<b>Nationality</b> (please state, e.g. British)					

Ethnic origin (as proposed by the Commission for Racial Equality, please tick)

<b>White</b>	British	
	Irish	
	Any other White background,	
<b>Mixed</b>	White and Black Caribbean	
	White and Black African	
	White and Asian	
	Any other Mixed background	
<b>Asian or Asian British</b>	Indian	
	Pakistani	
	Bangladeshi	
	Any other Asian background	
<b>Black or Black British</b>	Caribbean	
	African	
	Any other Black background	
<b>Chinese or other ethnic group</b>	Chinese	
	Any other Chinese background	

<b>Academic and professional qualifications (please state, e.g. GNVQ)</b>	
---	--

I feel confident in using computers (please tick one)

Strongly agree	Agree	Disagree	Strongly disagree

I have used (please tick one box in each line)

Item	Never	A few times ever	Less than weekly	At least once a week	Most days
The Web					
Email					
Word (or other word processor)					
Acrobat					

I have accessed the PRUN 1100 online module (please tick one)

Never	Once	2-5 times	More than 5 times

With regard to my nursing studies I have found the PRUN 1100 online module (please tick one)

Very useful	Useful	Slightly useful	Useless	I did not use it

The parts of the PRUN 1100 online course were (please tick one box for each part)

Online course component	Very useful	Useful	Slightly useful	Useless	I did not use it
Staff information (contact details of staff)					
Module documents (handouts, lecture notes etc.)					
Assessments (example assignment)					
External links (e.g. to resuscitation council)					
Discussion board					
Email					
Tools (e.g. calendar)					
Module guide (handbook and work & safety workbook)					

Please feel free to make any comments below about the PRUN 1100 online course:

Before we start however there are some concepts that you will need, specifically data types, measures of central tendency and variance.

Data types.

- Nominal variables are those whose outcomes are categorical, not meaningfully put into numbers (other than coded). For example gender has two outcomes male and female, but while you can code males as 1 and females as 2, it does not follow that females are twice male. SPSS calls variables of this type **Nominal**.
- Ordinal variables are those that are naturally ordered. For example Likert scales (Likert, 1932) are those that allow a subject to give a response typically something like “strongly agree” through “agree” and “disagree” to “strongly disagree”. Clearly a response of “agree” is between “strongly agree” and “disagree”; i.e. there is an order. Thus if “strongly agree” is coded 4 it is more than agree coded 3 and not merely labelled differently. However the distance between “strongly agree” and “agree” is not obviously the same as between “agree” and “disagree”. SPSS calls variables of this type **Ordinal**
- Interval/ratio variables are where the numbers have even more meaning, for example age where the difference between consecutive numbers is always the same. For example a person aged 20 is not just older than one of 19, they are one year older. Ratio variables have in addition an absolute zero. You cannot have negative height or age, these are ratio data. You can have negative temperature in Celsius, these are interval data but not ratio. SPSS calls variables of either of these types **Scalar**.

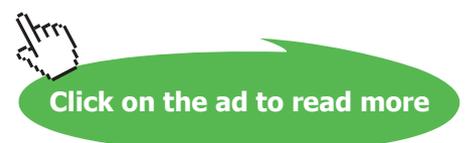
N.B. SPSS allows you to define variables as Scalar, Ordinal or Nominal, however its type checking is not always helpful, and I have sometimes to define an Ordinal variable as Scalar (for example) before SPSS Graph Builder will allow me to create a graph using such ordinal data.

.....Alcatel-Lucent 

[www.alcatel-lucent.com/careers](http://www.alcatel-lucent.com/careers)

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



## Measures of central tendency

These describe the middle of a distribution of data, and the three commonly used measures are mean, median and mode. The mean is what most people think of as the average. It is the sum of all values divided by the number of values. The mode is the most common value and the median is the value in the middle of the distribution. If we had eleven PhD students with the following ages:-

22, 22, 22,, 22, 22, **23**, 23, 23, 23, 30, 43

Then the mean is  $(22+22+22+22+22+23+23+23+23+30+43)/11 = 25$ . The median is 23 (shown in bold) as it is the sixth value of eleven in ascending order. The mode is 22 as there are five students with this age.

Which measure we use depends on the question to be answered. For example if you wanted the “average” wage earned in the UK you might think the mean is a good measure, and this was in 2008 £26,020, and for full time workers £31,323. However this does not imply half the full time workers in the UK get £31K or more. The median (also called 50th percentile) is the figure for the mid-point, and this is £20,801 or £25,123 for full time workers (source <http://news.bbc.co.uk/1/hi/8151355.stm>). The median is the salary below which 50% of people earn, and above which the other 50% earn, it splits the salaries into two equal numbers of people. The reason for this difference between mean and median is that a few very well paid people push the mean value up, but this has little effect on the people in the middle of the distribution. If you want to know the middle earning worker then the median is better than the mean (and this is what is typically used in statistics on salaries). The problem with salaries is that they are skewed, with a lot of poorly paid people and a few very highly paid people.

## Variability

Data may be clustered tightly around the mean value or very spread out. Variance is measured by essentially adding together the differences between the actual salaries and the mean (in fact the squared differences are used, so a value below the mean and one above the mean do not cancel each other out). It is defined the mean of the sum of the squared deviation of each value from the mean.

The variance of the students’ ages above would be (where the symbol  $\Sigma$  means sum of all values and n is the number of subjects)

$$\begin{aligned} & \Sigma(\text{value}-\text{mean})^2/n \\ & ((22-25)^2 + (22-25)^2 + (22-25)^2 + (22-25)^2 + (22-25)^2 + (23-25)^2 + (23-25)^2 + (23-25)^2 + (23-25)^2 + (30-25)^2 + (43-25)^2)/11 \\ & = (5 * 3^2 + 4 * 2^2 + 5^2 + 18^2)/11 \\ & = (45 + 16 + 25 + 324)/11 \\ & = 410/11 \\ & = 37.3 \end{aligned}$$

However this is only correct if our sample is the whole population, if it is a sample drawn from the population (typically is) then we divide not by the number of observations  $n$  but by  $n-1$

Then variance is  $410/10 = 41$

Note here the few values that are very different contribute massively to the variance, the one age of 45 contributes more than the rest of the ages put together. So variance is sensitive to skewed samples.

For an example of how variance might be used, in a very equal society most people would have a salary close to the mean. Variance would be small. In an unequal society with many poor and rich people earning respectively much less or much more than the mean, variance would be high. In practice the standard deviation is often used which is the square root of the variance. As we noted above the variance, and therefore the standard deviation is not a good measure in some distributions (salary is one such case) and a better measure of variability of data is the inter-quartile range. This is the difference between the 25<sup>th</sup> percentile and the 75<sup>th</sup> percentile. In salaries this would mean finding the salary below which 25% of the population earn (25<sup>th</sup> percentile) and that above which only 25% earn (75<sup>th</sup> percentile).

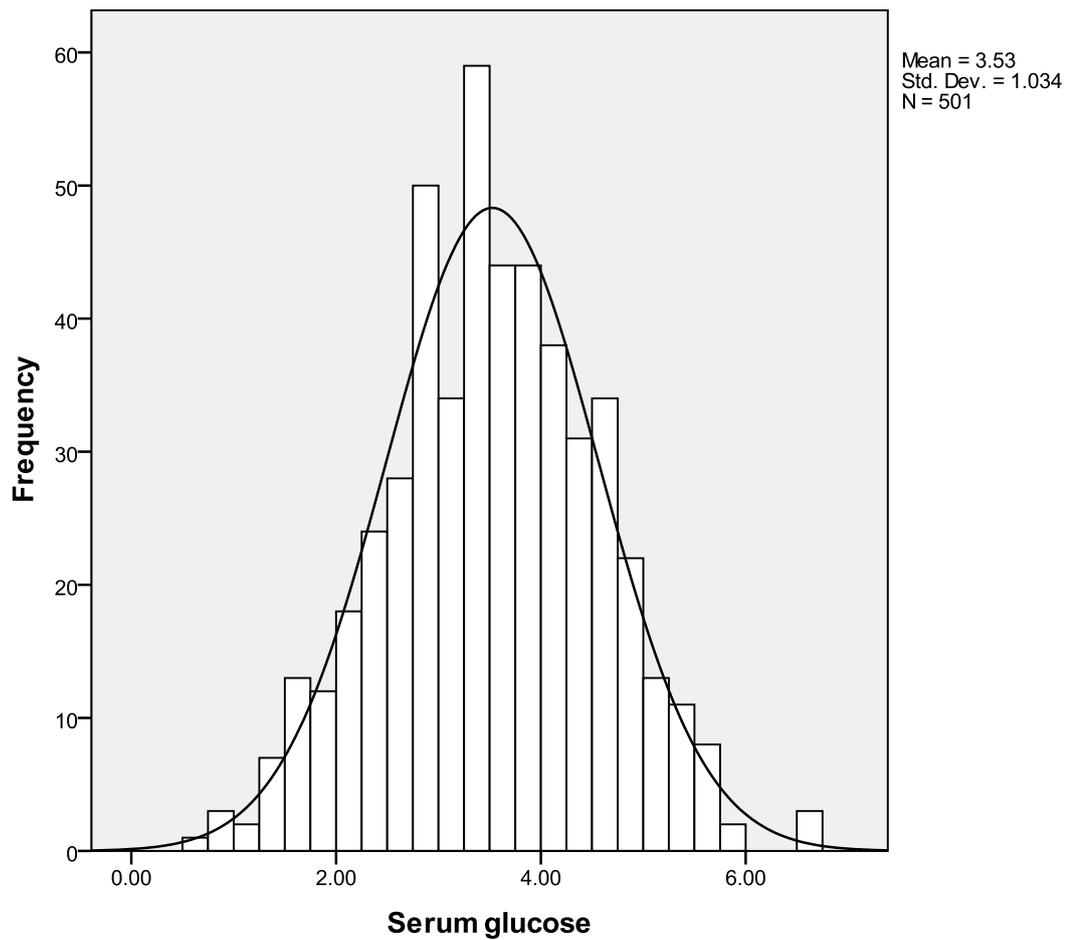
## Normal distributions

A normal distribution is characterised by large numbers of values in the middle near the mean, and decreasing numbers as one gets further from the mean in either direction. This gives a bell shaped curve. In a perfectly normal distribution mean, median and mode are all the same. A roughly (fictional) normal distribution is seen in Figure 29. In a normally distributed sample about two thirds of values lie within one standard deviation, and about 95% within two standard deviations, and almost all within three standard deviations. So here with a mean of about 3.5 and a standard deviation of about 1.0, almost all values lie between 1.5 and 6.5, and 95% between 2.5 and 5.5.

A related concept to standard deviation is standard error. Whereas standard deviation tells us about the range of values, standard error informs us about the range of mean values of different samples drawn from the population. A common error is to confuse the two, and as standard error is always less (usually much less) than standard deviation this can be problematic.

The perfect normal curve has been superimposed on the histogram (I explain histograms later, but they show the frequency of subjects that lie within ranges of values, so here numbers of subjects in ascending ranges of blood glucose), the actual values are not quite in line with this, but are close enough to be seen as normally distributed. The main thing to note is very few people have either very low or high blood sugar (fortunately) and most values bunch in the middle.

**Figure 29:** Normal distribution



### Descriptive statistics for online survey

I entered the data manually typing in the responses from the paper questionnaire (as above). The data are found in datafile *online survey.sav* (all datafiles in SPSS have a suffix “.sav”). If you load this into SPSS you will see something like Figure 30.

Figure 30: Data for online course

	no	course	Age	gender	q1	q2a	q2b	q2c	q2d	q3	q4
1	1	Degree	32	Female	Disagree	At least on...	At least on...	At least on...	A few time...	Once	Useless
2	2	Degree	18	Female	Strongly a...	Most days	Most days	Most days	Most days	More than 5 ti...	Useful
3	3	Degree	32	Female	Agree	Most days	Most days	Most days	Most days	Never	I did not us...
4	4	Degree	33	Female	Strongly a...	Most days	Most days	Most days	Most days	2-5 times	Slightly us...
5	5	Degree	32	Male	Strongly a...	Most days	Most days	Most days	Most days	More than 5 ti...	Useless
6	6	Degree	34	Male	Agree	At least on...	At least on...	At least on...	At least on...	2-5 times	Slightly us...
7	7	Degree	29	Female	Agree	Most days	At least on...	At least on...	Never	Never	I did not us...
8	8	Degree	31	Female	Strongly a...	Most days	Most days	Most days	A few time...	2-5 times	Slightly us...
9	9	Degree	20	Female	Agree	At least on...	Most days	Most days	At least on...	2-5 times	I did not us...
10	10	Degree	29	Male	Disagree	A few time...	Most days	A few time...	Never	Never	Useful
11	11	Degree	24	Female	Strongly a...	Most days	Most days	Most days	Less than ...	2-5 times	Useless
12	12	Degree	.	Female	Agree	Most days	Most days	Most days	Most days	2-5 times	Slightly us...
13	13	Degree	30	Female	Agree	Most days	At least on...	At least on...	At least on...	2-5 times	Useless
14	14	Degree	18	Female	Strongly a...	At least on...	At least on...	Less than ...	A few time...	More than 5 ti...	Useless
15	15	Degree	18	Female	Agree	Most days	At least on...	Most days	At least on...	2-5 times	Useless
16	16	Degree	38	Female	Agree	At least on...	At least on...	Less than ...	Less than ...	More than 5 ti...	Useful
17	17	Degree	18	Female	Agree	Most days	Most days	Most days	A few time...	2-5 times	Slightly us...
18	18	Degree	20	Female	Agree	Most days	Most days	Most days	Never	Never	Useful
19	19	Degree	18	Male	Strongly a...	Most days	Most days	Most days	At least on...	2-5 times	I did not us...
20	20	Degree	19	Female	Strongly di...	Most days	Most days	Most days	Less than ...	2-5 times	Slightly us...
21	21	Degree	19	Female	Agree	Most days	Most days	Most days	A few time...	Once	Useless
22	22	Degree	19	Female	Strongly a...	Most days	Most days	Most days	Less than ...	Once	Useless
23	23	Degree	21	Female	Agree	Most days	Most days	Most days	A few time...	2-5 times	Useless
24	24	Degree	41	Female	Strongly a...	Most days	At least on...	Most days	At least on...	Never	I did not us...

**Maastricht University** *Leading in Learning!*

**Join the best at the Maastricht University School of Business and Economics!**

**Top master's programmes**

- 33<sup>rd</sup> place Financial Times worldwide ranking: MSc International Business
- 1<sup>st</sup> place: MSc International Business
- 1<sup>st</sup> place: MSc Financial Economics
- 2<sup>nd</sup> place: MSc Management of Learning
- 2<sup>nd</sup> place: MSc Economics
- 2<sup>nd</sup> place: MSc Econometrics and Operations Research
- 2<sup>nd</sup> place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

**Maastricht University is the best specialist university in the Netherlands (Elsevier)**

**Visit us and find out why we are the best!**  
**Master's Open Day: 22 February 2014**

[www.mastersopenday.nl](http://www.mastersopenday.nl)

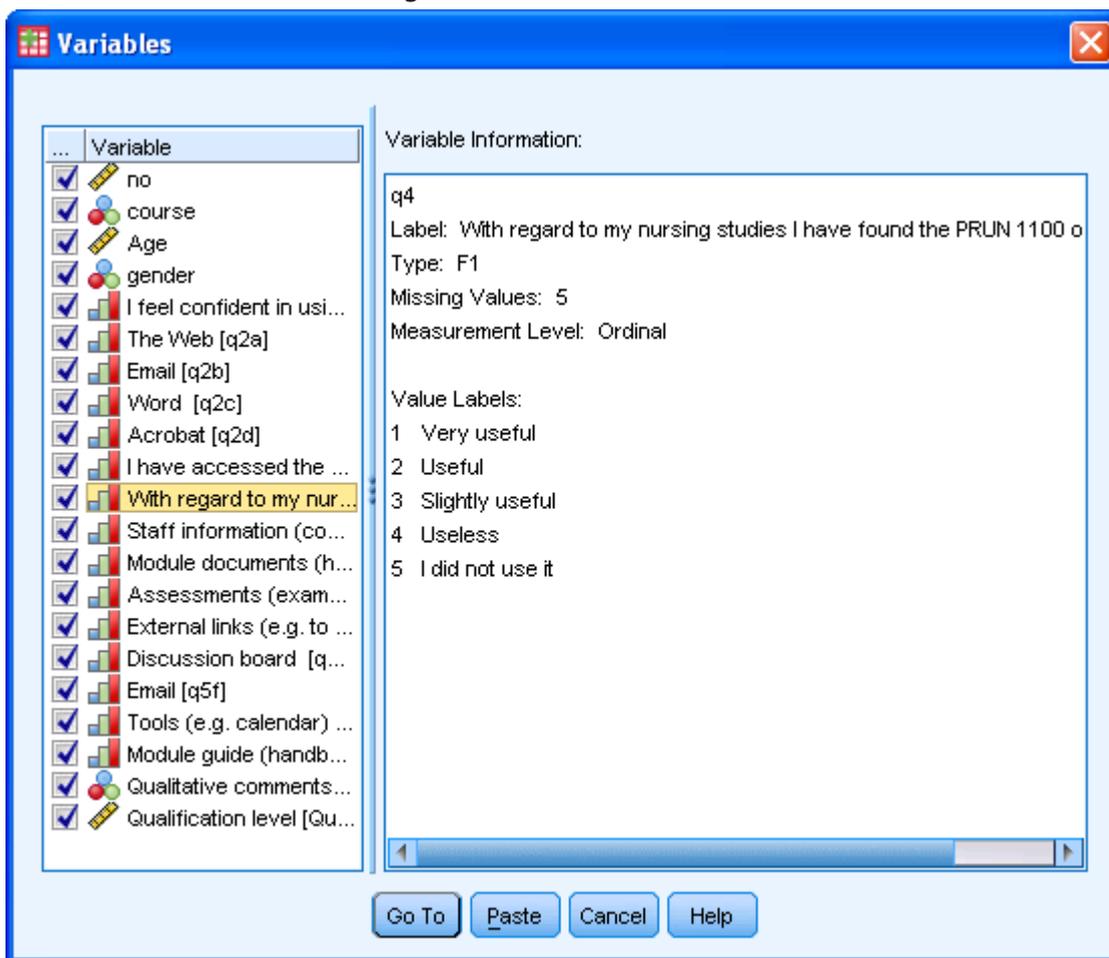


The variables are as in Figure 31, where I show a summary as seen in the Variable View, and then using the Variable popup I show how the labels for one variable are identified, see Figure 32.

Figure 31: Variables

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	no	Numeric	3	0		None	None	8	Right	Scale	Input
2	course	Numeric	1	0		{1, PDip}...	None	8	Right	Nominal	Input
3	Age	Numeric	3	0		None	None	8	Right	Scale	Input
4	gender	Numeric	1	0		{0, Female}...	None	8	Right	Nominal	Input
5	q1	Numeric	1	0	I feel confident i...	{1, Strongly ...	None	8	Right	Ordinal	Input
6	q2a	Numeric	1	0	The Web	{1, Never}...	None	8	Right	Ordinal	Input
7	q2b	Numeric	1	0	Email	{1, Never}...	None	8	Right	Ordinal	Input
8	q2c	Numeric	1	0	Word	{1, Never}...	None	8	Right	Ordinal	Input
9	q2d	Numeric	1	0	Acrobat	{1, Never}...	None	8	Right	Ordinal	Input
10	q3	Numeric	1	0	I have accesse...	{1, Never}...	None	10	Right	Ordinal	Input
11	q4	Numeric	1	0	With regard to ...	{1, Very use... 5	5	8	Right	Ordinal	Input
12	q5a	Numeric	1	0	Staff informatio...	{1, Very use... 5	5	8	Right	Ordinal	Input
13	q5b	Numeric	1	0	Module docum...	{1, Very use... 5	5	8	Right	Ordinal	Input
14	q5c	Numeric	1	0	Assessments (...	{1, Very use... 5	5	8	Right	Ordinal	Input
15	q5d	Numeric	1	0	External links (...	{1, Very use... 5	5	8	Right	Ordinal	Input
16	q5e	Numeric	1	0	Discussion board	{1, Very use... 5	5	8	Right	Ordinal	Input
17	q5f	Numeric	1	0	Email	{1, Very use... 5	5	8	Right	Ordinal	Input
18	q5g	Numeric	1	0	Tools (e.g. cale...	{1, Very use... 5	5	8	Right	Ordinal	Input
19	q5h	Numeric	1	0	Module guide (...	{1, Very use... 5	5	8	Right	Ordinal	Input
20	qual	Numeric	1	0	Qualitative com...	{1, Access ...	None	14	Right	Nominal	Input
21	Qual1	Numeric	1	0	Qualification level	{1, 0 level o...	None	8	Right	Scale	Input
22											
23											
24											
25											

Figure 32: Variables information



### The data

I was interested in how the students rated an online course. Thus I asked them to tell me how useful each component of the course was from very useful to useless. I also collected data on demographics such as gender, age, ethnic group. I obtained 151 replies. But how do I show my results to (say) the teaching team that gives the course? I could let them see all of the 151 questionnaires, but this would not be practical or even very useful. For the tutors would have too much information, and not necessarily be able to draw out the main points.

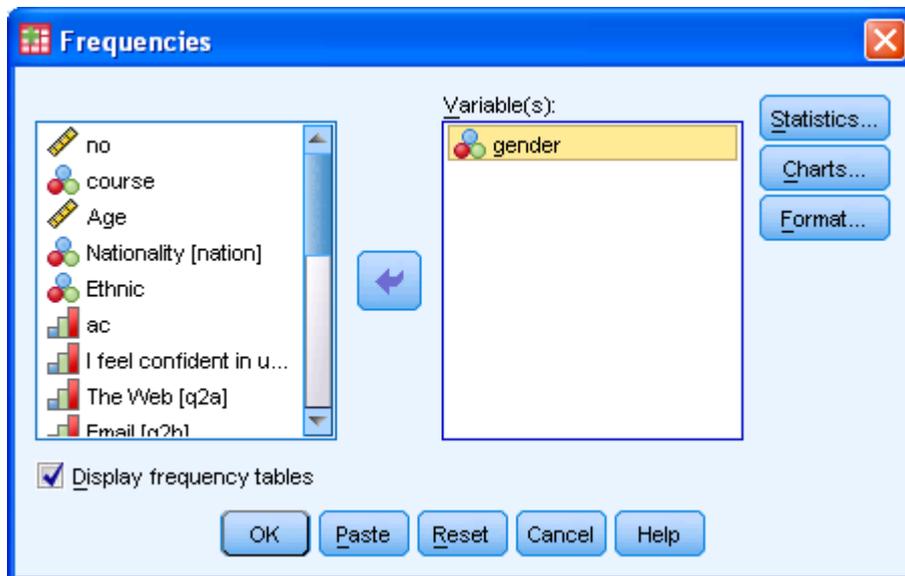
### Descriptive statistics

The role of descriptive statistics is to reduce the information to manageable proportions; it is to summarise data in a meaningful way.

A simple example is a frequency table. A frequency table counts the number of subjects that have a particular attribute. For example gender has two possible outcomes, male and female.

To obtain a frequency table use **Analyze -> Descriptive Statistics -> Frequencies** and then in the dialogue box move across the variable (here gender) see Figure 33, then click **OK** (in many occasions you will need to click on **OK** to get output, I won't necessarily state this).

**Figure 33: Frequencies**



The very simple frequency table is shown in the **Output** window as in Table 2.

**BI NORWEGIAN BUSINESS SCHOOL**

EFMD EQUIS ACCREDITED

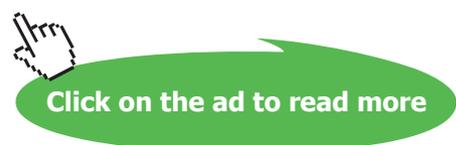
## Empowering People. Improving Business.

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

[www.bi.edu/master](http://www.bi.edu/master)



**Table 2:** Frequencies of variable

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	133	88.1	88.7	88.7
	Male	17	11.3	11.3	100.0
	Total	150	99.3	100.0	
Missing	System	1	.7		
Total		151	100.0		

The first column with a heading is frequency, it shows you how many stated they were female (133) and male (17) making 150 in total. However I had 151 replies. For some reason I did not have a response for gender from one student, probably they had not filled in this part of the questionnaire. This is shown as “missing”. Then there is a column for percent, showing 88.1% are female, and 11.3% male, and 0.7% did not state gender. Of those who did answer, 88.7% were female, and this is shown in “valid percent” i.e. the percentage once missing values are removed. In this case as so few data are missing, there is not much difference between the columns.

Let us look at another variable, that of nationality, see Table 3.

**Table 3:** Frequencies of nationalit

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	British	115	76.2	95.8	95.8
	German	1	.7	.8	96.7
	Nigerian	1	.7	.8	97.5
	Zimbabwean	3	2.0	2.5	100.0
	Total	120	79.5	100.0	
Missing	System	31	20.5		
Total		151	100.0		

There are more possible values for this variable. Note that there are many more missing data. You may want to ponder why many students did not give their nationality. However clearly there is a big difference in the columns percent and valid percent. It is clearly more meaningful to say about 96% of those who stated their nationality were British than to say about 76% of all responses were from those who said they were British. Of course we do not know if the 31 missing values were also largely British, and would need to be careful in interpreting these results.

We could now look at age employing a frequency table, see Table 4.

**Table 4:** Frequencies of age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	17	1	.7	.7	.7
	18	22	14.6	15.8	16.5
	19	26	17.2	18.7	35.3
	20	6	4.0	4.3	39.6
	21	3	2.0	2.2	41.7
	22	6	4.0	4.3	46.0
	23	7	4.6	5.0	51.1
	24	5	3.3	3.6	54.7
	25	2	1.3	1.4	56.1
	26	3	2.0	2.2	58.3
	27	4	2.6	2.9	61.2
	28	2	1.3	1.4	62.6
	29	3	2.0	2.2	64.7
	30	3	2.0	2.2	66.9
	31	3	2.0	2.2	69.1
	32	5	3.3	3.6	72.7
	33	4	2.6	2.9	75.5
	34	6	4.0	4.3	79.9
	35	4	2.6	2.9	82.7
	36	4	2.6	2.9	85.6
	37	4	2.6	2.9	88.5
	38	1	.7	.7	89.2
	39	3	2.0	2.2	91.4
	40	1	.7	.7	92.1
	41	2	1.3	1.4	93.5
	43	2	1.3	1.4	95.0
	44	1	.7	.7	95.7
	45	1	.7	.7	96.4
	46	4	2.6	2.9	99.3
	52	1	.7	.7	100.0
	Total	139	92.1	100.0	
Missing	System	12	7.9		
Total	151	100.0			

This is trickier, as there are many more ages than either gender or nationality. Again some students did not give their age. I will look at the valid percent column. You will note the final column gives a cumulative percent. Thus 0.7% are 17, 15.8% are 18, but adding these gives 16.5% who are 18 or younger. This becomes useful as we can immediately see 66.9% (or over two thirds) are 30 or younger, and 92.1% are no older than 40.

Yet another way to view age is with mean and standard deviation. **Using Analyze -> Descriptives -> Frequencies** and then entering in the variable age gives the output in Figure 33. The mean is the average value, and may differ from the median. Here it does, the mean is in the later twenties. The standard deviation is of limited value in this set of data.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	139	17	52	26.48	8.584
Valid N (listwise)	139				

A more useful estimate of variability for these data is the inter-quartile range, which is where the middle 50% of data lie. You can get this and other values using **Analyze -> Descriptive Statistics -> Explore** and putting *age* into the Dependent List, see Figure 34. You will get output as shown in Table 5. The median is shown as 23. Thus we can see that the typical student is young (median 23) and half of all students are in a 14 year (inter-quartile) range.

## Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

**Get Help Now**



Go to [www.helpmyassignment.co.uk](http://www.helpmyassignment.co.uk) for more info





Figure 34: Analyse using Explore

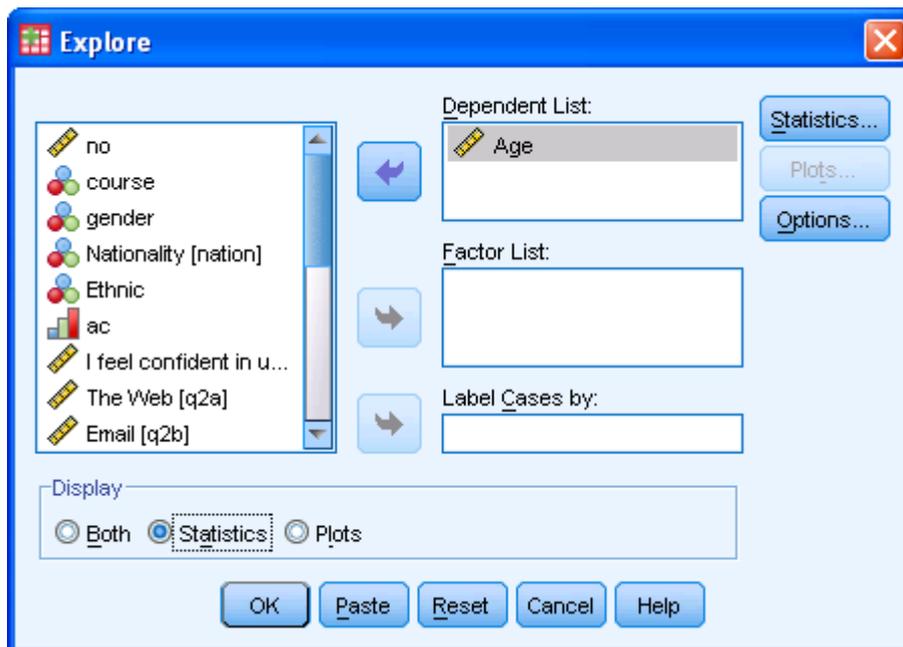


Table 5: Explore data output

		Statistic	Std. Error
Age	Mean	26.48	.728
	95% Confidence Interval		
	for Mean	Lower Bound	25.04
		Upper Bound	27.92
	5% Trimmed Mean	25.85	
	Median	23.00	
	Variance	73.686	
	Std. Deviation	8.584	
	Minimum	17	
	Maximum	52	
	Range	35	
	Interquartile Range	14	
	Skewness	.819	.206
	Kurtosis	-.369	.408

Additional items include 95% confidence interval - where we expect the “real” mean to lie, since this is a mean of a sample of the population, not the population itself and the 5% trimmed mean – the mean once the top and bottom 5% of values have been removed, i.e. removing outliers that can distort the mean. Skewness measures how skewed data are, in a normal distributions this should be zero, and positive skew means there is a larger tail of values on the right than the left of the histogram, and negative is the opposite. Here with a positive skew there is a skew towards older students. Kurtosis measures how peaked the histogram is, with high kurtosis having a high peak and low kurtosis a low one. In a normal distribution kurtosis is three, but it is often reported as excess kurtosis which is kurtosis-3 or zero. Kurtosis is not very important in statistics, but skewness is. To see if skewness is a problem check whether the value for skewness is within the range of minus to plus twice the standard error of skewness. Here it is, as 0.819 is way out of the range +/- 0.512 so the distribution is significantly skewed and can be considered not normally distributed.

## Further Reading

World of statistics contains links to many statistical sites across the world via a clickable map.

Statsoft electronic textbook <http://www.statsoft.com/textbook/stathome.html>. Statsoft create a statistics package we are not using here. However the electronic textbook is not specific to their products and is a handy free resource.

SPSS home page <http://www.spss.com/>

A thesaurus of statistics <http://thesaurus.maths.org/mmkb/view.html>

Any number of books in your library (for example the university library or medical library) on basic statistics and/or SPSS. Choose ones that suit your learning style. I recommend any of three texts I have found useful (Bryman and Cramer, 2005, Marston, 2010, Pallant, 2010) for good introductions and to explore the more detailed aspects Field (2009).

But there are literally dozens of others. Choose one that is recent and try to get those that are looking at later versions as the interface changes between some versions, and there have been other changes at each version (though not very substantial in the things we are considering). The interface for creating graphs changed massively after version 16, but there is a legacy option for those familiar with the older graph interface.

## References

BRYMAN, A. & CRAMER, D. (2005) *Quantitative data analysis with SPSS 12 and 13 : a guide for social scientists*, Hove, Routledge.

FIELD, A. (2009) *Discovering statistics using SPSS (and sex and drugs and rock 'n' roll)*, London, Sage.

LIKERT, R. (1932) A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140, 1-55.

MARSTON, L. (2010) *Introductory statistics for health and nursing using SPSS*, London, Sage.

PALLANT, J. (2007) *Title SPSS survival manual : a step by step guide to data analysis using SPSS for Windows* Maidenhead Open University Press.

# 4 Graphs

## Key points

Graphs are alternatives to descriptive statistics

At the end of this unit you should be able to:

Create barcharts (including with error bars), histograms, boxplots and scatterplots.

## Introduction

I introduced descriptive statistics in chapter three. Here I show you how to use similar information to create graphs.

## Graphics interface for SPSS

Some people like to see data in tables, some prefer graphs. In some cases it is not crucial which you use, so nationality is perfectly well shown either in a table, or (as below) in a bar chart. In other cases graphs are better than tables, so while the frequency table is meaningful for age, maybe a better way to show the data is in a graph. The one that is useful here is not the bar chart. Bar charts are good for nominal data and sometimes ordinal data but not optimal for or interval/ratio data. What is needed here is a histogram or a box plot (also both shown below).

**Brain power**

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

**The Power of Knowledge Engineering**

Plug into The Power of Knowledge Engineering. Visit us at [www.skf.com/knowledge](http://www.skf.com/knowledge)

**SKF**



To create graphs you need to understand the **Chart Builder**

The graphics interface in SPSS underwent a total redesign in version 16 onwards. You set up a graph in Chart Builder using **Graph -> Chart Builder**. You will be prompted to check the data typing of your variables, see Figure 35. This is because the graph builder will only allow certain types of variable to be used in some graphic displays. You may find the variable you want to use is not offered in a list of variables. When this occurs you need to change the type in the **Variable View**. After dealing with this you will see a **Canvas** where the graph is drawn with **Drop Zones** in the x and y axes, where variables are placed from the **Variables List**, there is a **Categories List** (shown under **Variables**, if the given variable has any value labels) to show you what categories are in the selected variable (see Figure 36). You want to choose the **Simple Bar** which is the left top option in the **Galley** and put **Nationality** into the x-axis **drop zone**, see Figure 37. The resultant graph in the **Output** window is in Figure 37.

**Figure 35:** Checking data types in graph builder

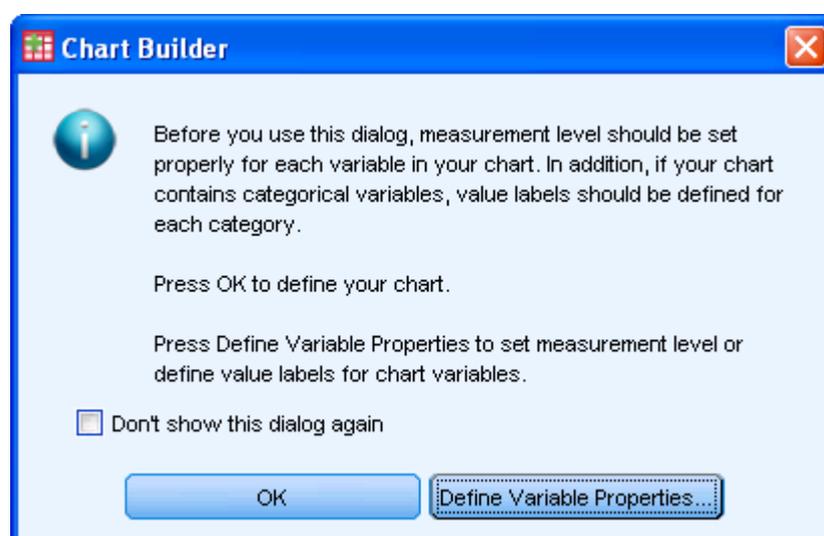
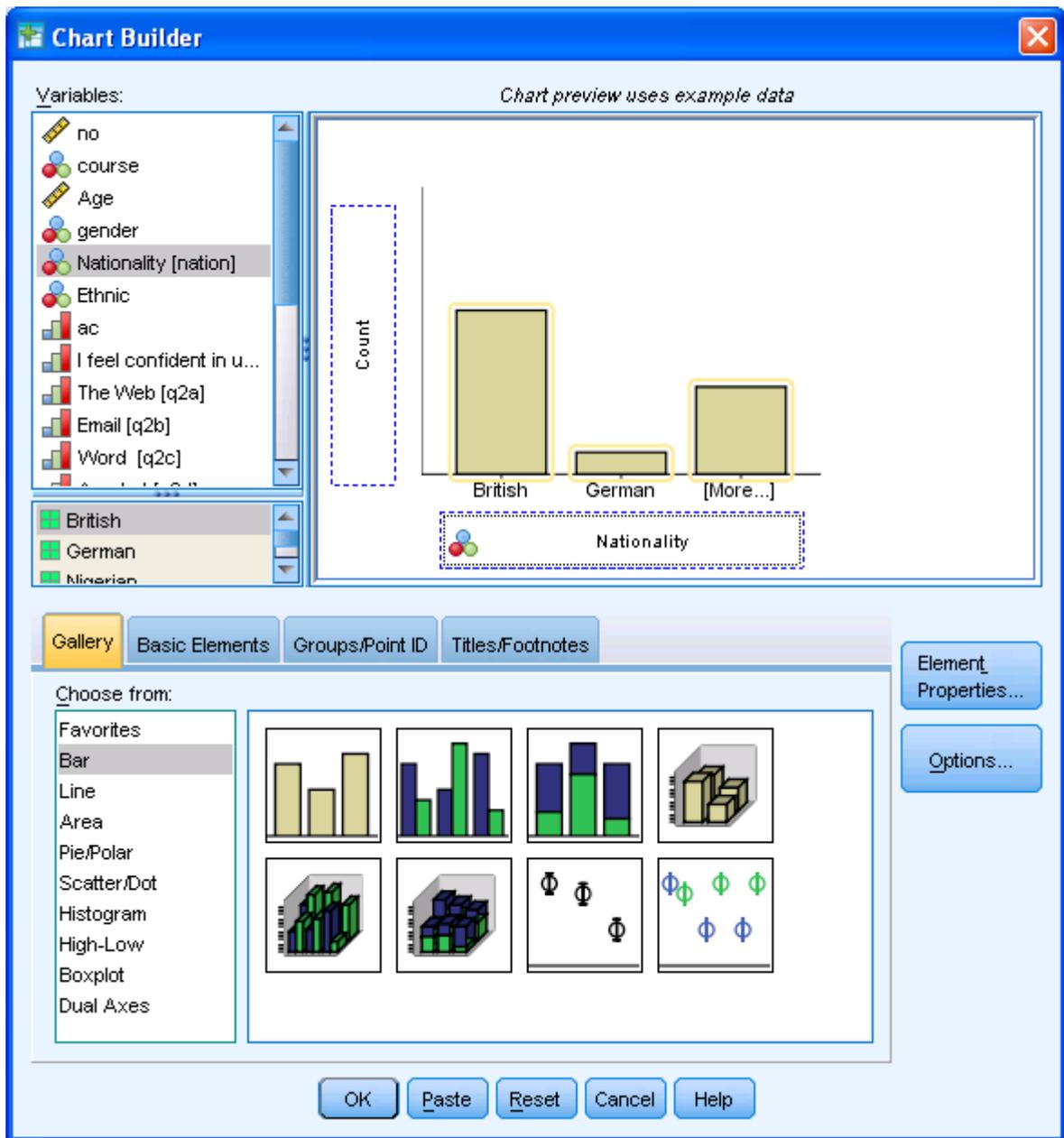
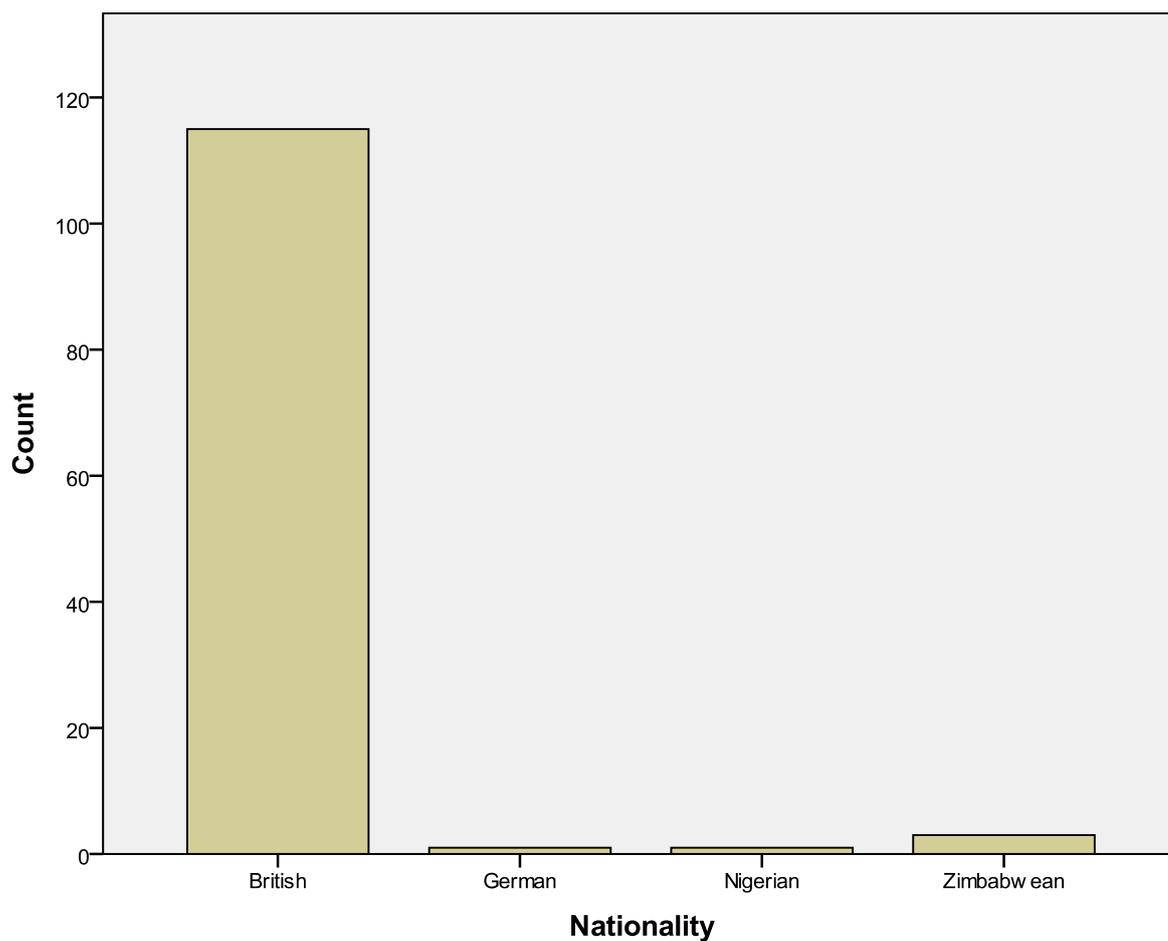


Figure 36: Setting up a bar chart

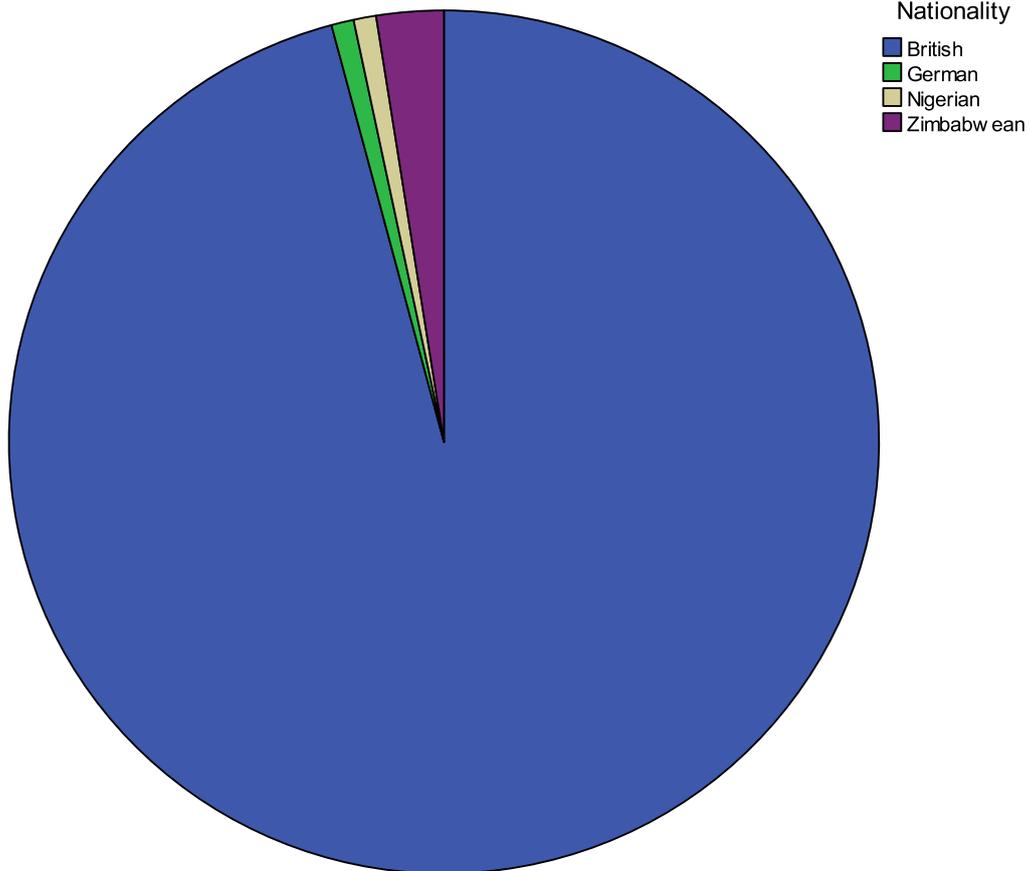


**Figure 37:** Bar chart of nationality



A pie chart can be done in similar way, by dragging a pie chart from the Gallery and again choosing *Nationality* to go into the x-axis see Figure 38.

Figure 38: Pie chart



"I studied English for 16 years but...  
...I finally learned to speak it in just six lessons"  
Jane, Chinese architect

ENGLISH OUT THERE

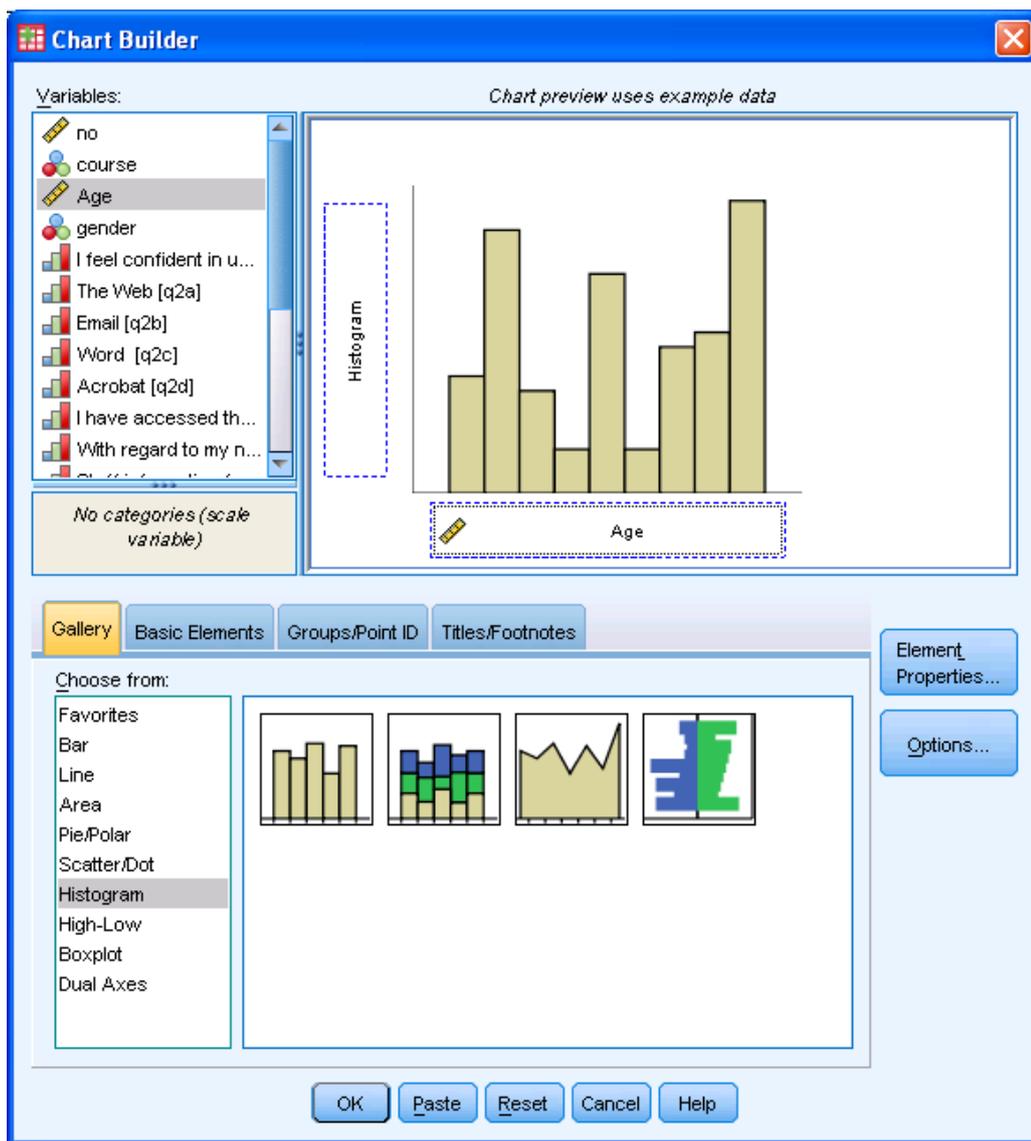
Click to hear me talking before and after my unique course download

The advertisement features a woman, Jane, a Chinese architect, smiling. A green speech bubble contains the text 'ENGLISH OUT THERE'. Below the speech bubble is a call to action: 'Click to hear me talking before and after my unique course download'. The background is a blurred image of a classroom or office setting.



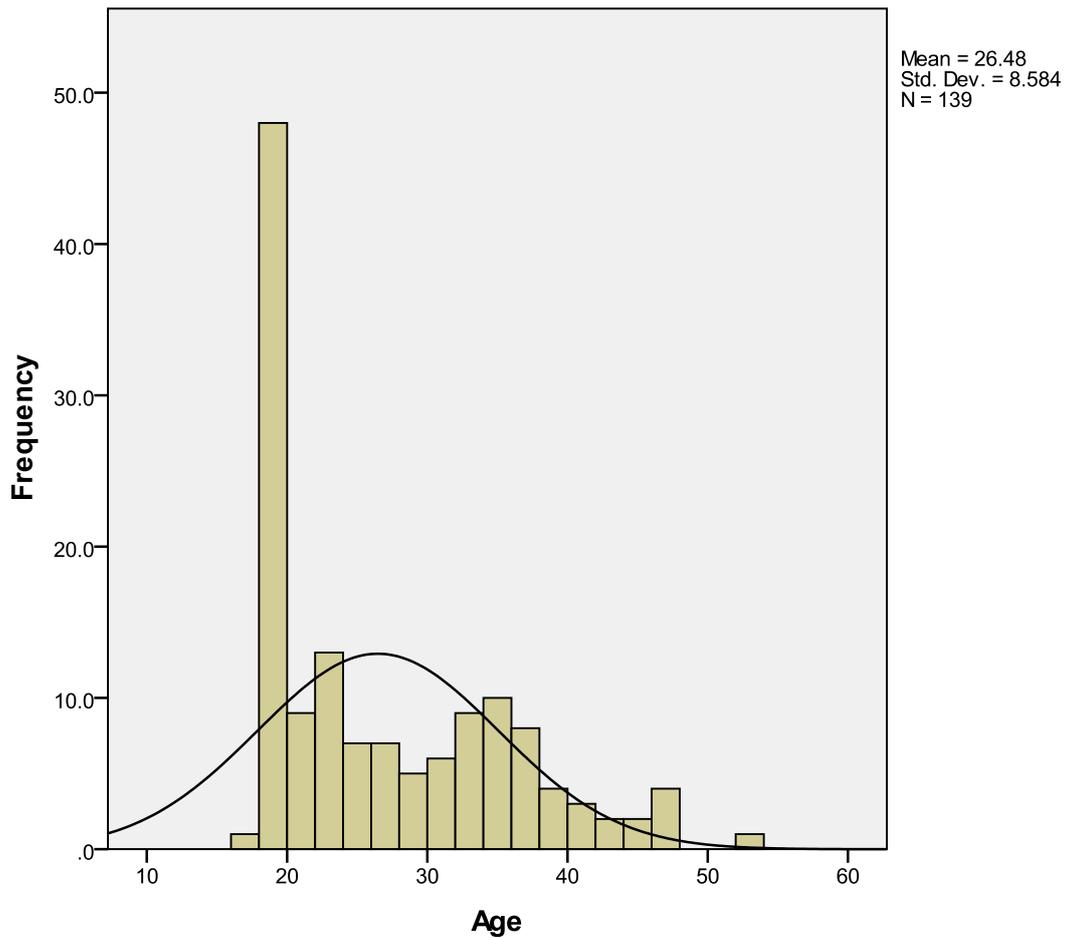
While for nominal and some ordinal data it makes good sense to use bar charts, for interval/ratio data a more specialised graph, the histogram is better, which shows you the number of subjects in each of several increasing ranges. Now to make a histogram of age we choose histogram from the **Gallery** . To create our histogram first choose from the possible (here four) designs, select **Simple Histogram** by clicking on it with the mouse. Then we drag the **Histogram**, using the mouse, from the **Gallery** to the **Canvas**, see Figure 39. Finally drag the variable *age* into the x axis **drop zone**, then the Output window will show the histogram, see Figure 40. Note we are given the mean and standard deviation.

Figure 39: Creating a histogram



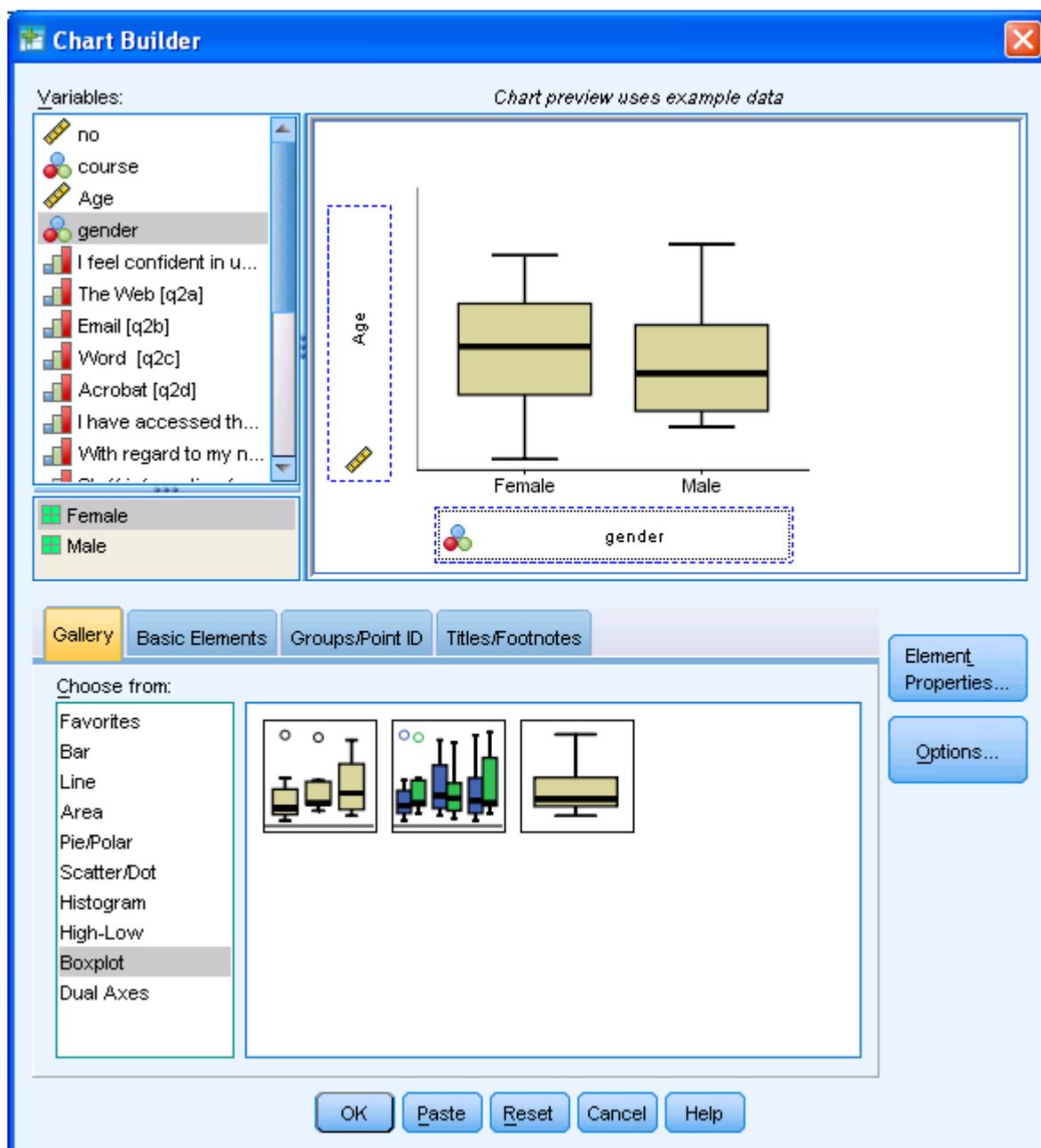
You can see immediately if the data are normally distributed, i.e. if they show a bell shaped curve. Clearly the histogram of age shows it is not normally distributed, note the normal curve I have superimposed over the data (this is added on the **Element Properties**). The distribution looks nothing like it, it is obviously skewed to younger ages. I.e. we can see that there is a big cluster of students just below 20, and a long tail of older students of various ages. This is important as we will see later that some statistical methods assume normal distribution.

**Figure 40:** Output of histogram

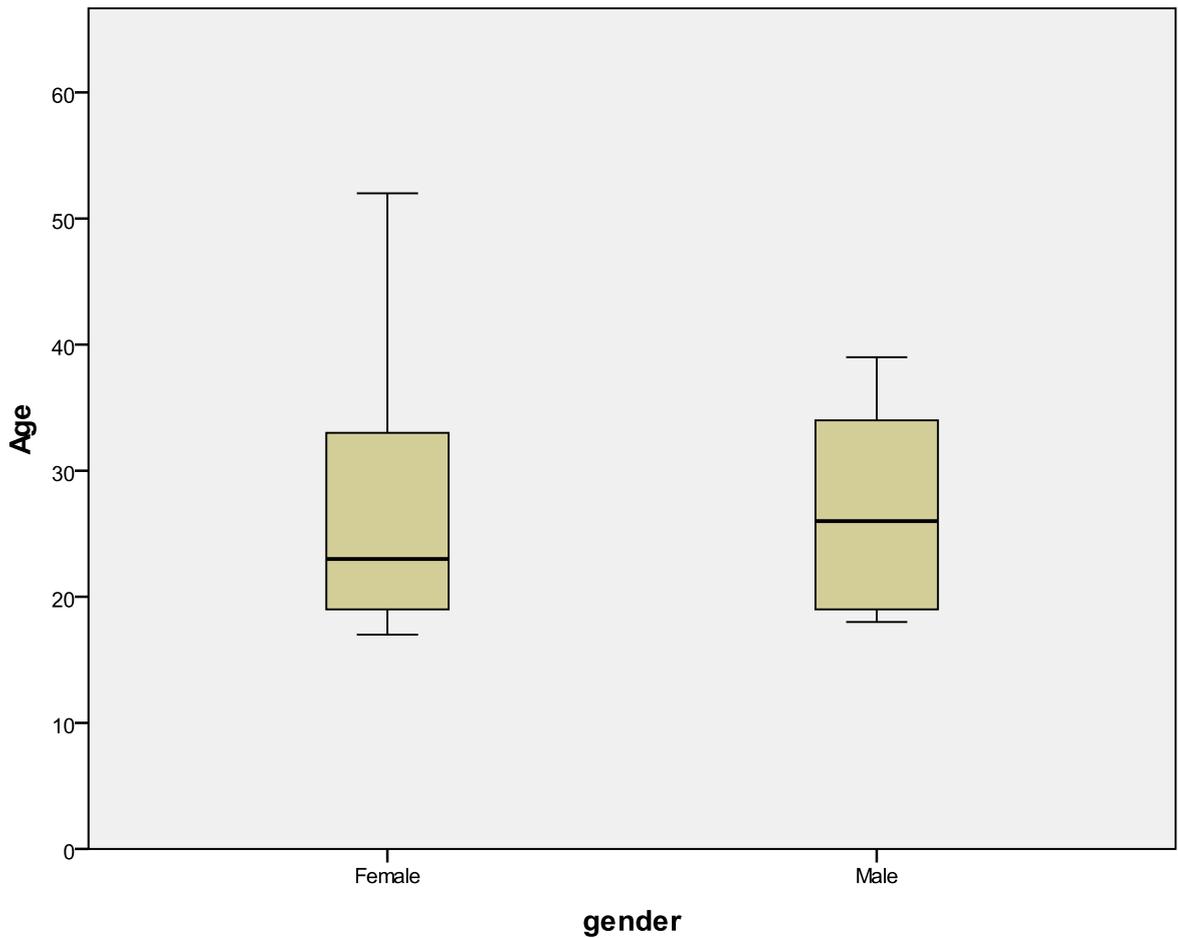


Another way to look at interval/ratio data is with a box plot (see Figure 41). Drag the **Simple Boxplot** (the leftmost) onto the **Canvas**. Then drag variables *age* and *gender* onto the y and x axes respectively, giving the output in the **Output** window. This shows you the range (see the vertical line showing ages from about 18 to a little over 50 for females and up to about 40 for males), and the median value shown as the black line in the box. The box itself is the inter quartile range where half the data lie. So for females 50% of students are between about twenty to early thirties, the median is somewhere in the early twenties.

Figure 41: Creating a box plot



**Figure 42:** Output of boxplot



### Exercise

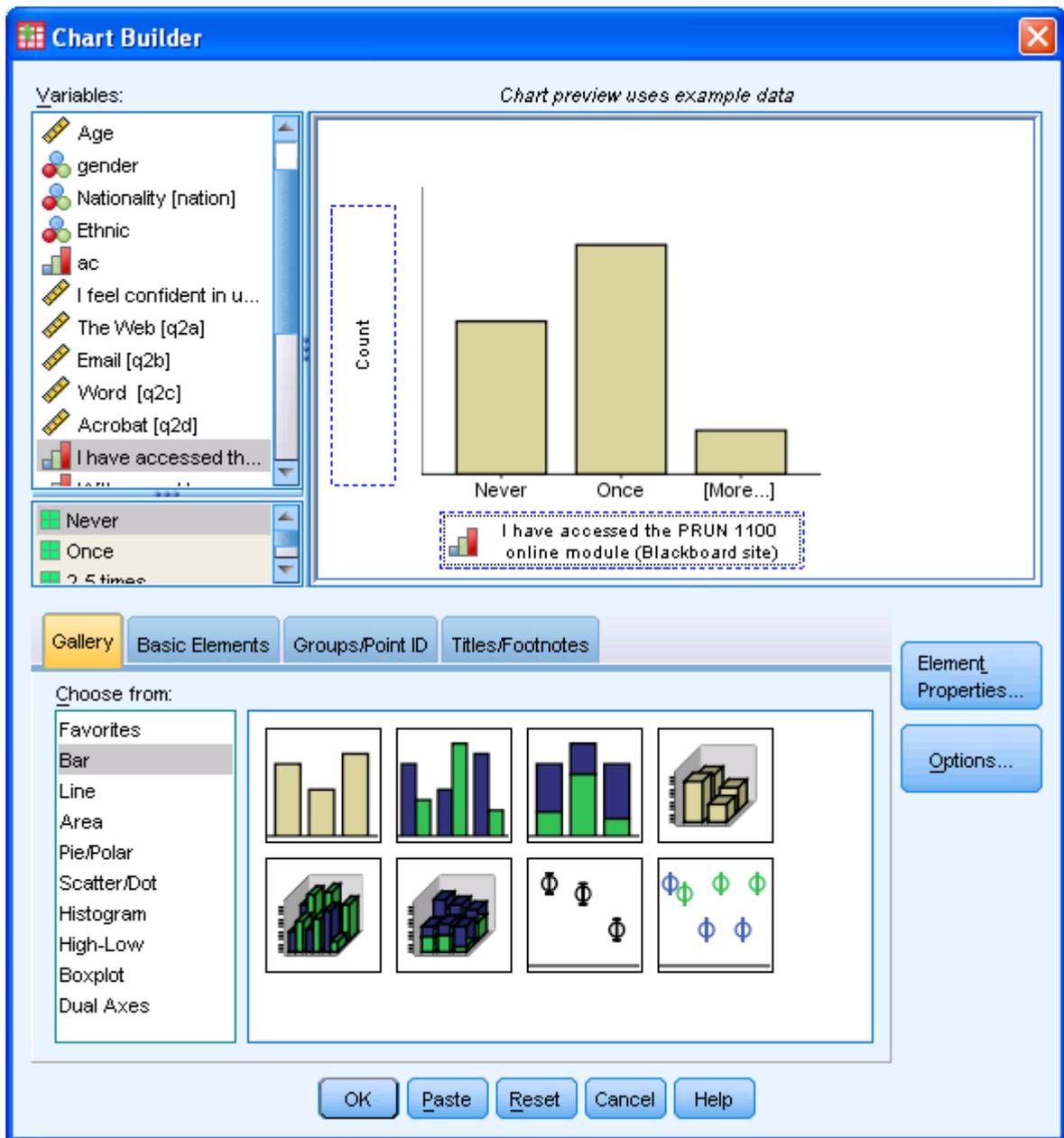
I collected data on ethnic groups (this is different from nationality) e.g. White British, Black or Black British, Asian or Asian British. Separately I obtained data on attitude to the online course ranging from very useful to useless on a four point scale. How would I best show these data to a teaching colleague to show them the general picture of ethnicity in this course, and also describe their attitudes to the relevance of online courses. What type of data is ethnic group, and what type of data is the attitudinal variable?

### Bar charts

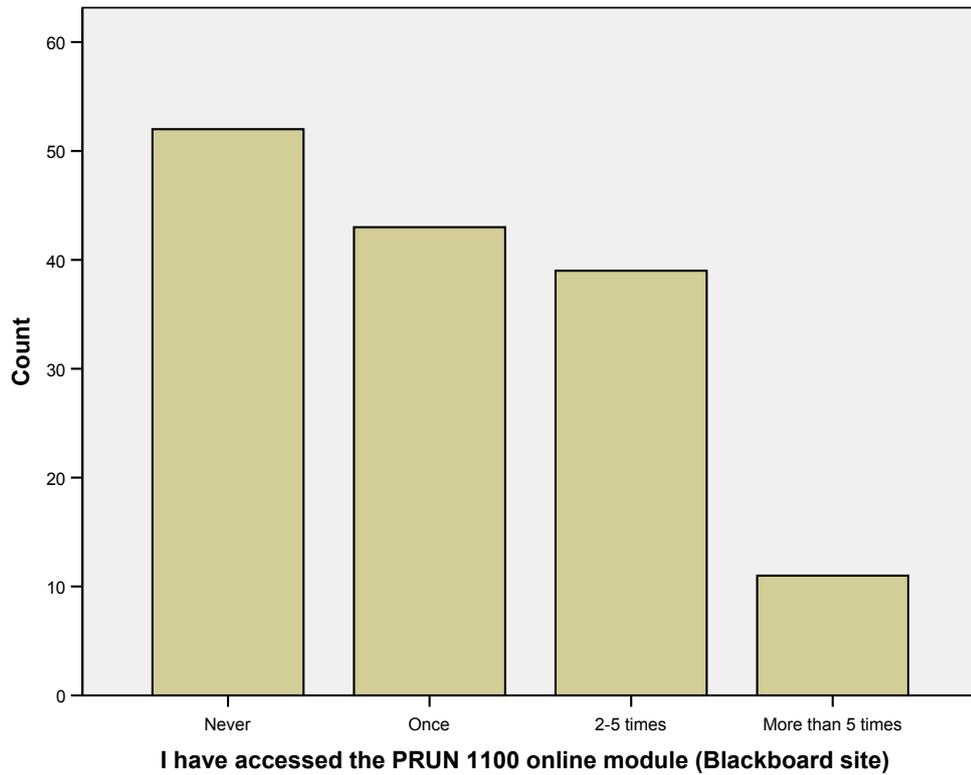
Let us consider the bar chart of accesses to the course, which is a graphical way to present data that otherwise would be a frequency table.

Access the **Chart Builder** and move **Simple Bar** into the **Canvas**, then move  $q^3$  into the horizontal drop zone as in Figure 43, then hit **OK**. This might look like Figure 44.

Figure 43: Creating a simple bar chart



**Figure 44:** Bar chart of access to online course

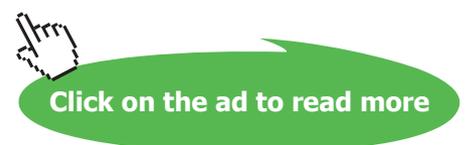


**What do you want to do?**

No matter what you want out of your future career, an employer with a broad range of operations in a load of countries will always be the ticket. Working within the Volvo Group means more than 100,000 friends and colleagues in more than 185 countries all over the world. We offer graduates great career opportunities – check out the Career section at our web site [www.volvogroup.com](http://www.volvogroup.com). We look forward to getting to know you!

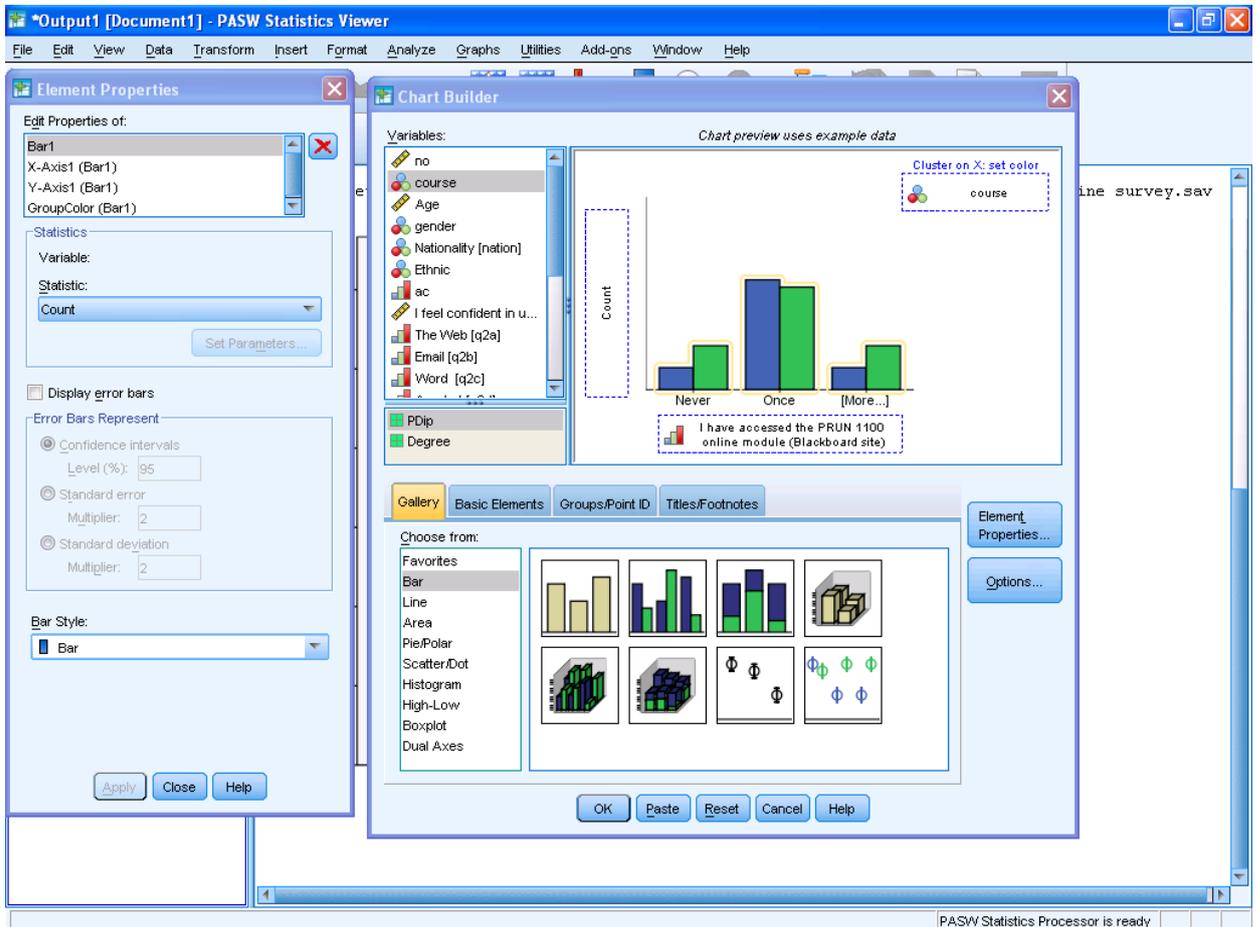
**VOLVO**  
 AB Volvo (publ)  
[www.volvogroup.com](http://www.volvogroup.com)

VOLVO TRUCKS | RENAULT TRUCKS | MACK TRUCKS | VOLVO BUSES | VOLVO CONSTRUCTION EQUIPMENT | VOLVO PENTA | VOLVO AERO | VOLVO IT  
 VOLVO FINANCIAL SERVICES | VOLVO 3P | VOLVO POWERTRAIN | VOLVO PARTS | VOLVO TECHNOLOGY | VOLVO LOGISTICS | BUSINESS AREA ASIA

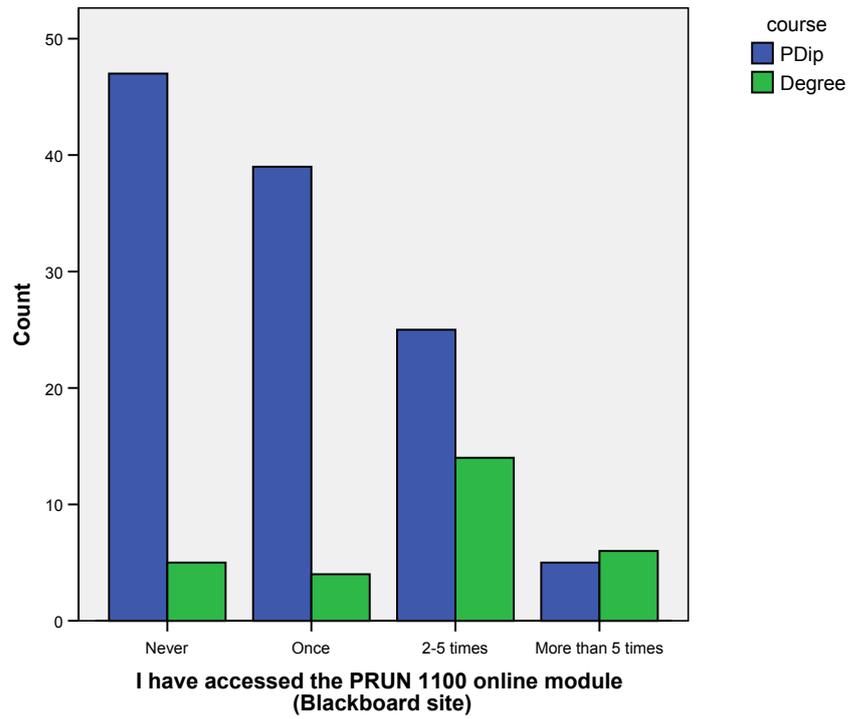


What if we wanted to show graphically data that were cross tabulated into access and (say) type of course. Access the **Chart Builder** and then select **Bar** in the **Gallery** and drag the clustered (second top left) design onto the **Canvas**, as in Figure 45. Then choose the two relevant variables, one to go into the **x-axis** the other into the **Cluster on X set color**. This will give you the output as in Figure 46.

**Figure 45:** Getting clustered bar chart from chart builder

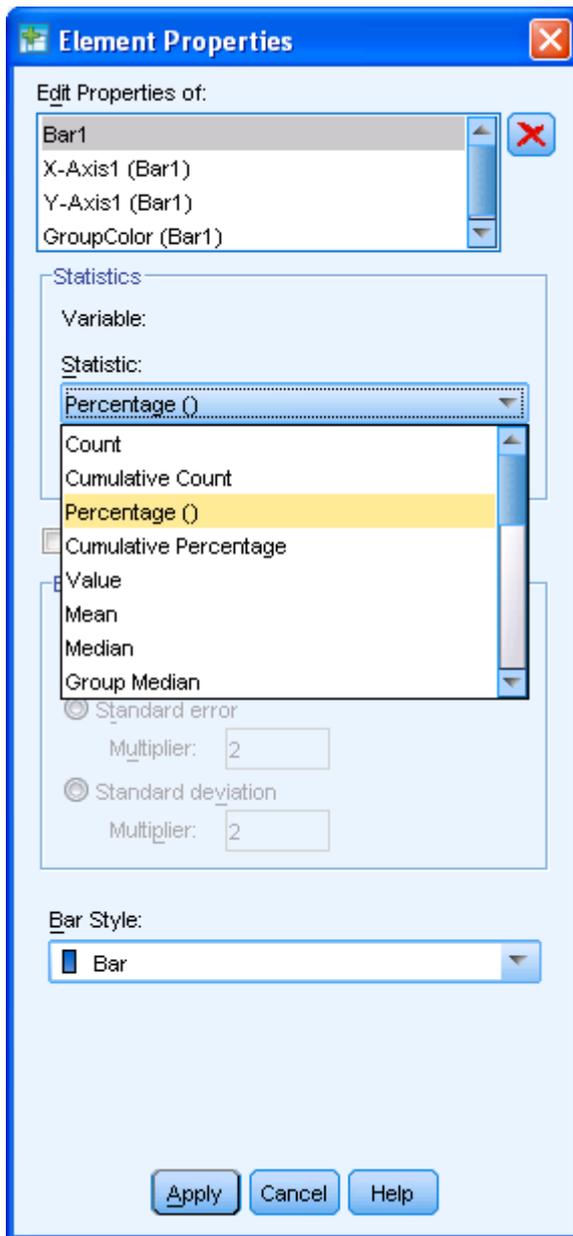


**Figure 46:** Clustered bar chart

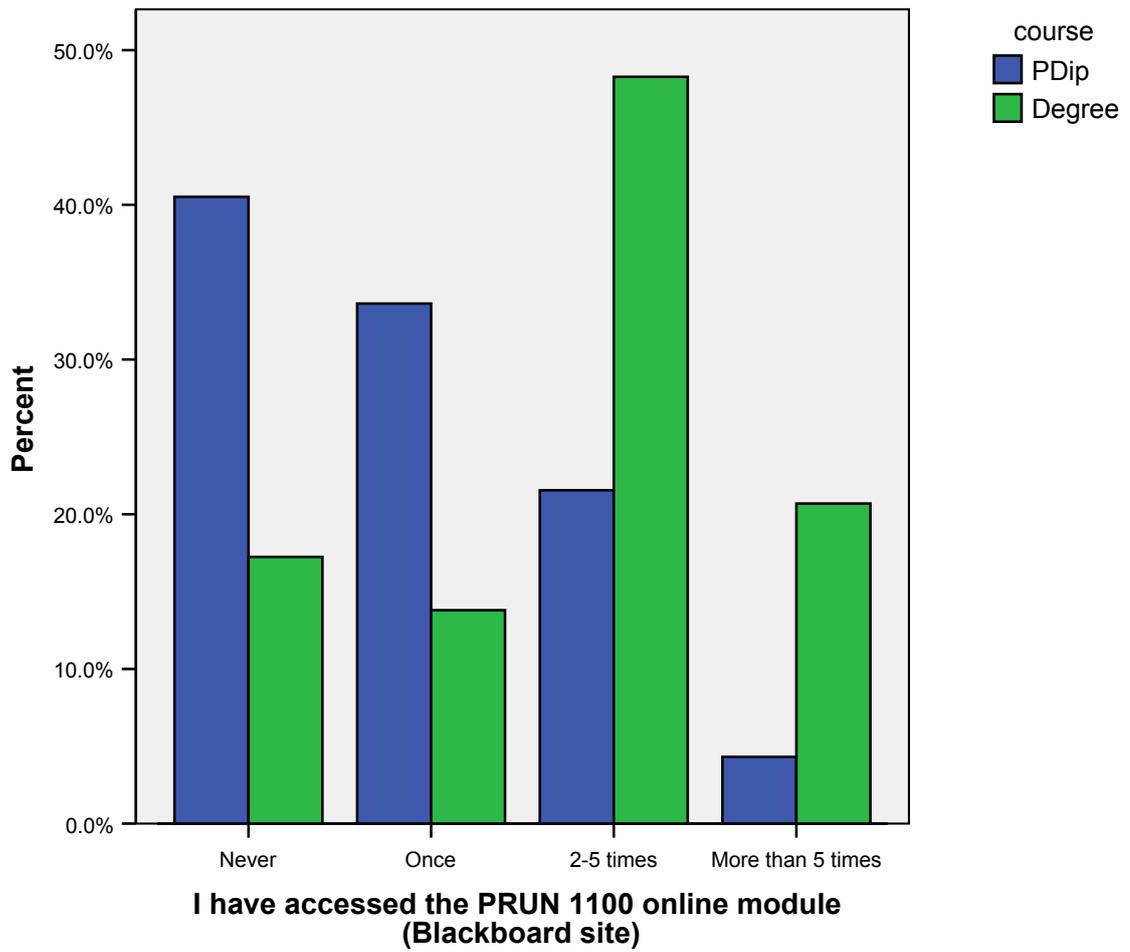


This is an instructive chart, it shows us that degree students are more likely to access the course several times. However this can be made even clearer if we use percentages instead of raw numbers (N of cases). To get percentage select in **Element Properties** the statistic **Percentage** rather than the default (**Count**), see Figure 47. Then we get Figure 48. Note how the difference is much more striking. The difference was a little masked before as the number of degree students were much lower. Using % allows a direct comparison of each bar. This is especially notable for the last bar, labelled *more than five times*.

**Figure 47:** Setting up percentage in bar chart



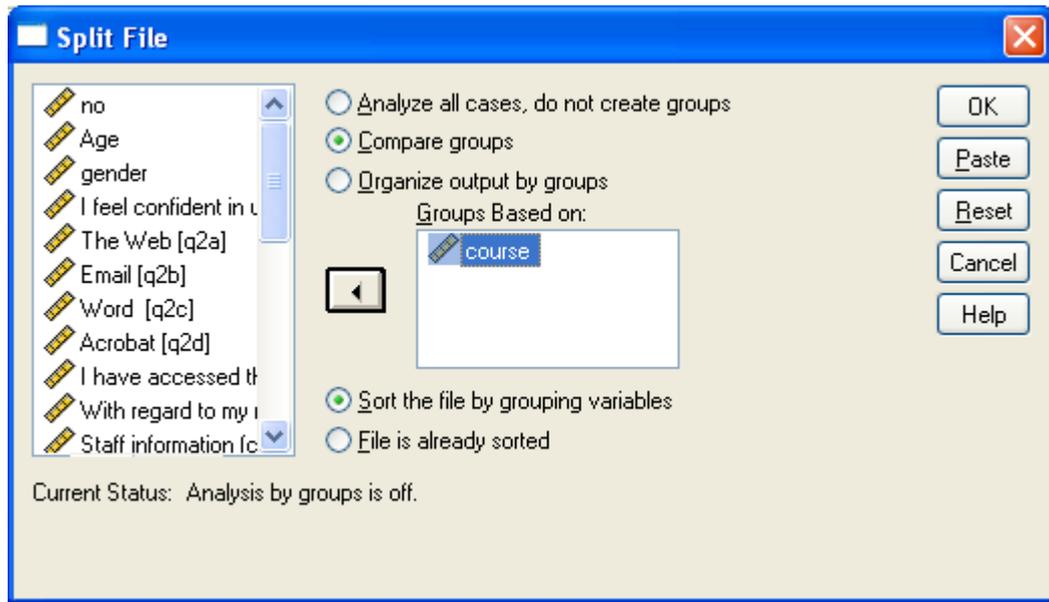
**Figure 48:** Clustered bar chart using %



### Splitting data

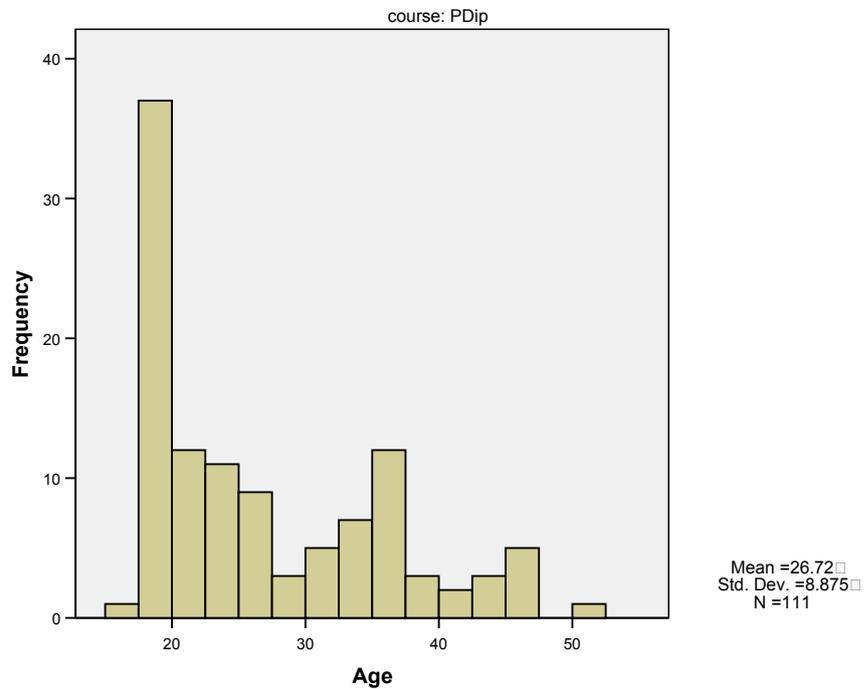
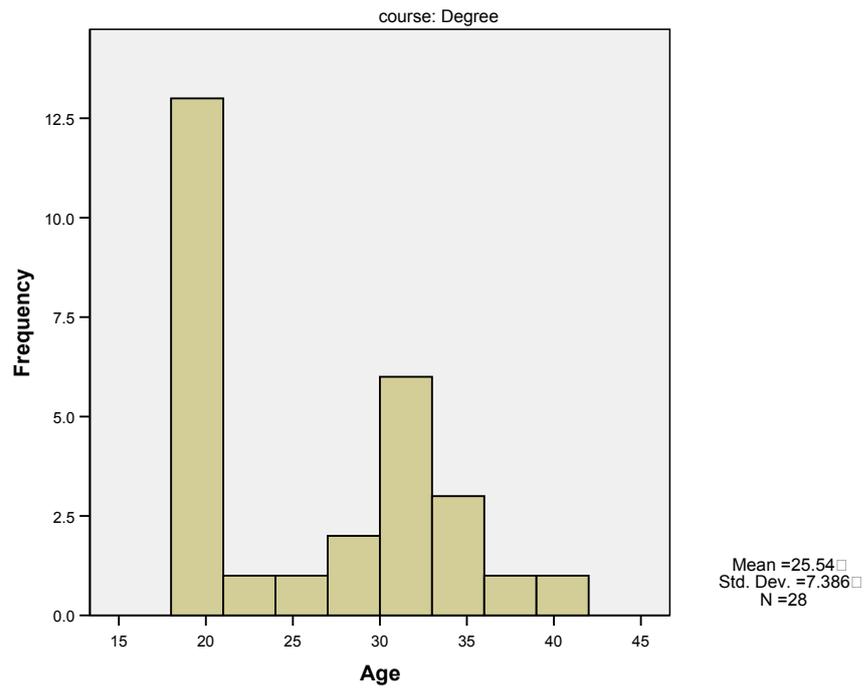
Let us consider now the histogram of age from week 1. What if you wanted a separate histogram for each course. You can do this in many ways, one is to split the data. Use **Data -> Split File**, then follow Figure 49.

Figure 49: Dialogue box for splitting data



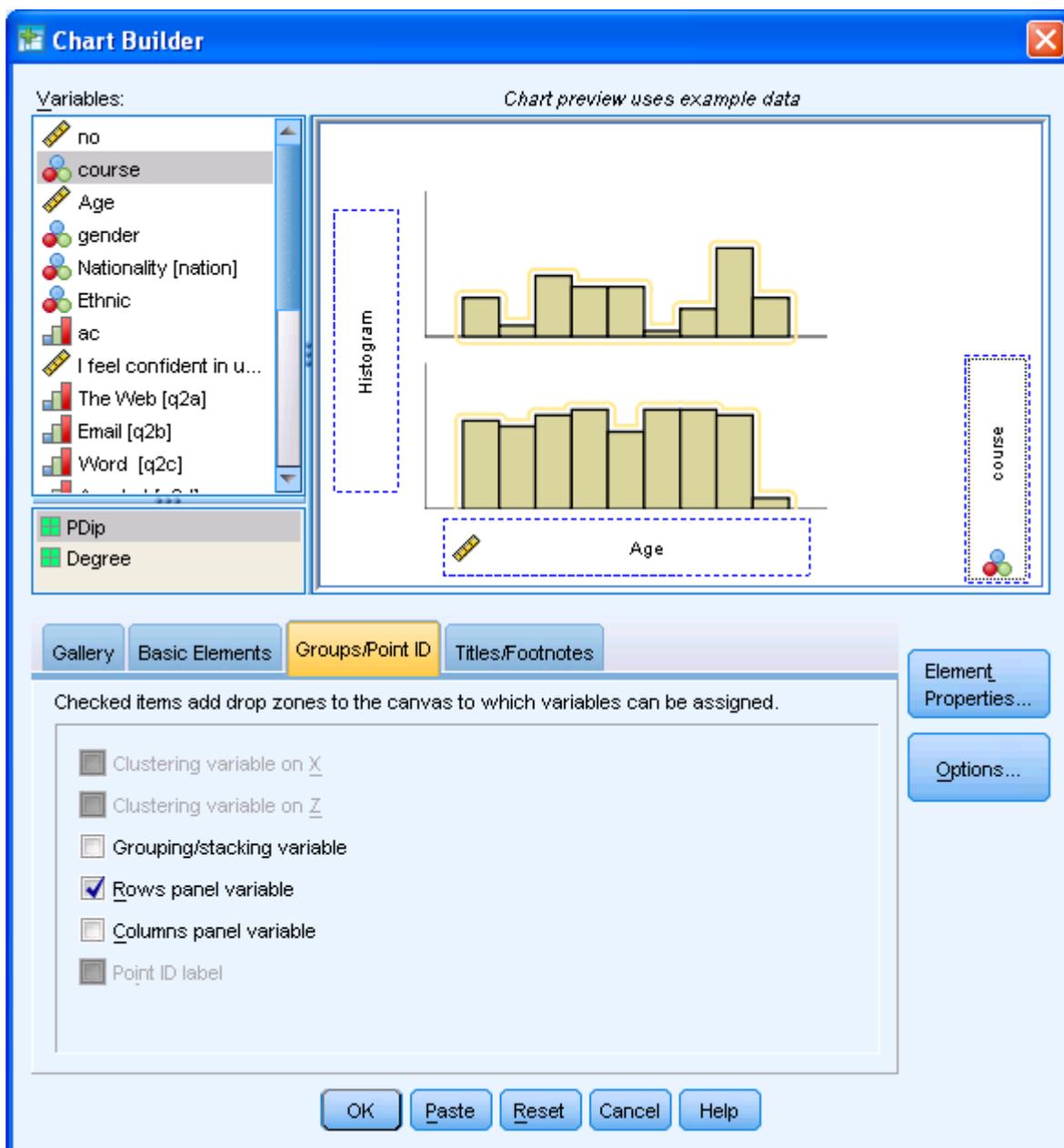
Nothing much happens, though you may note the data gets sorted by *course*. However what happens if you do a histogram of age? You get the histogram of age for each type of course. Indeed any analysis or graph you perform (until you unsplit the data, by ticking *analyze all cases* in Figure 49) is done on each value of variable used to compare groups (here there are only two), see Figure 50.

**Figure 50: Histogram for split data**



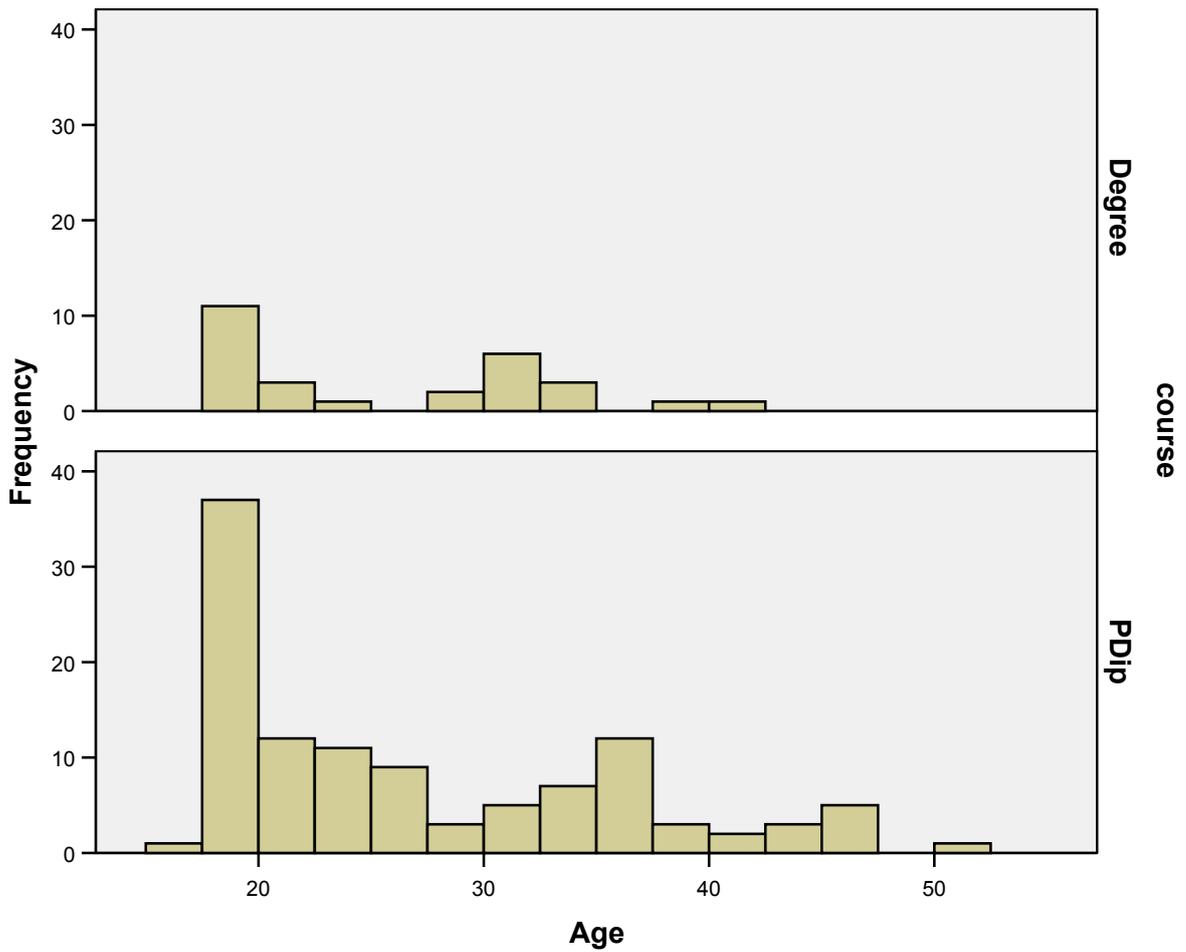
In later versions of SPSS (14.0 onwards) you can create graphs in many cases without needing to use the split data method. See Figure 51 for method in v18 where you click on **Groups/Point ID** and click on **Rows panel variable** (or columns if you prefer).

Figure 51: Using rows with graphs



The output of this is shown in Figure 52.

**Figure 52:** Output of histogram using rows



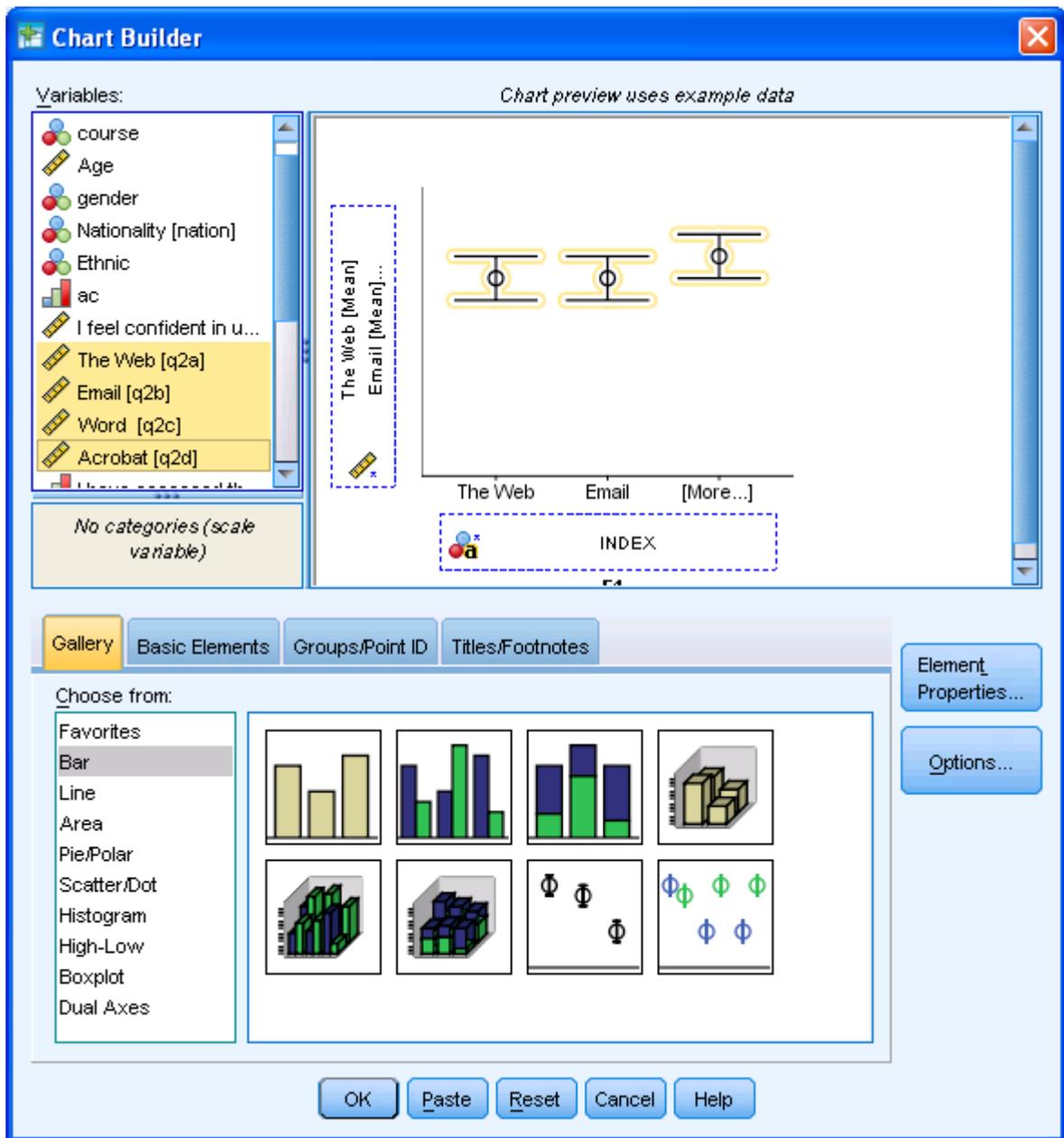
**Exercise**

Split data by gender and do a histogram of age for each gender. Unsplit the data and perform a clustered bar chart using percentages of gender against frequency of accessing the course. Interpret the results of both operations.

**Looking at several variables**

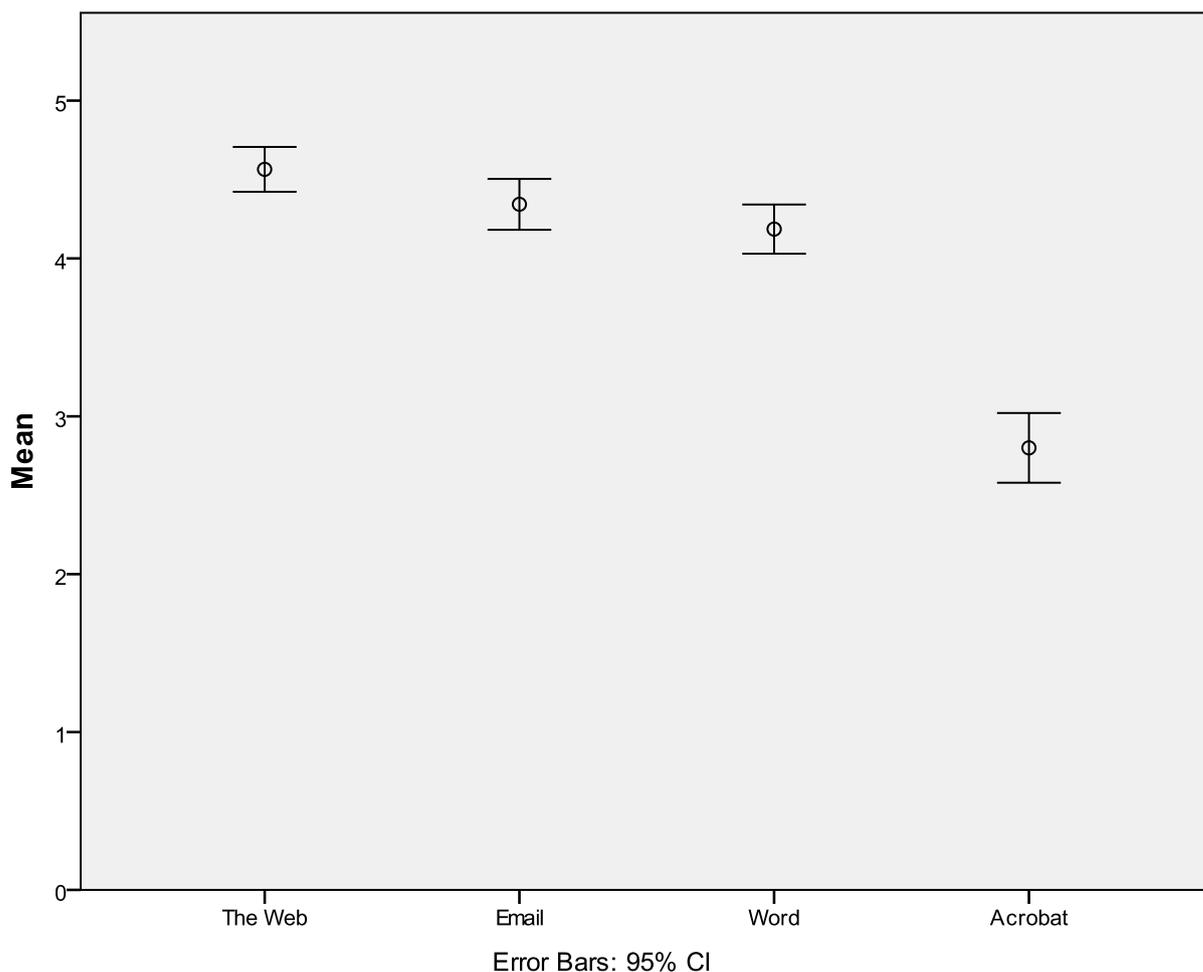
We might like to do a boxplot of several variables at once. This used to be possible in older versions of SPSS, but not now (though there is a legacy option, so actually you can). A similar effect can be done with **Bar** and **Simple Error Bar**, see Figure 53 where I have dragged four variables onto the y-axis.

Figure 53: Dialogue box for boxplots



The output is seen in Figure 54. While this graph does show the general trend, boxplots would be better on data that are not normally distributed, which these are not, so means and standard deviations are less useful.

**Figure 54:** Error bars of several variables

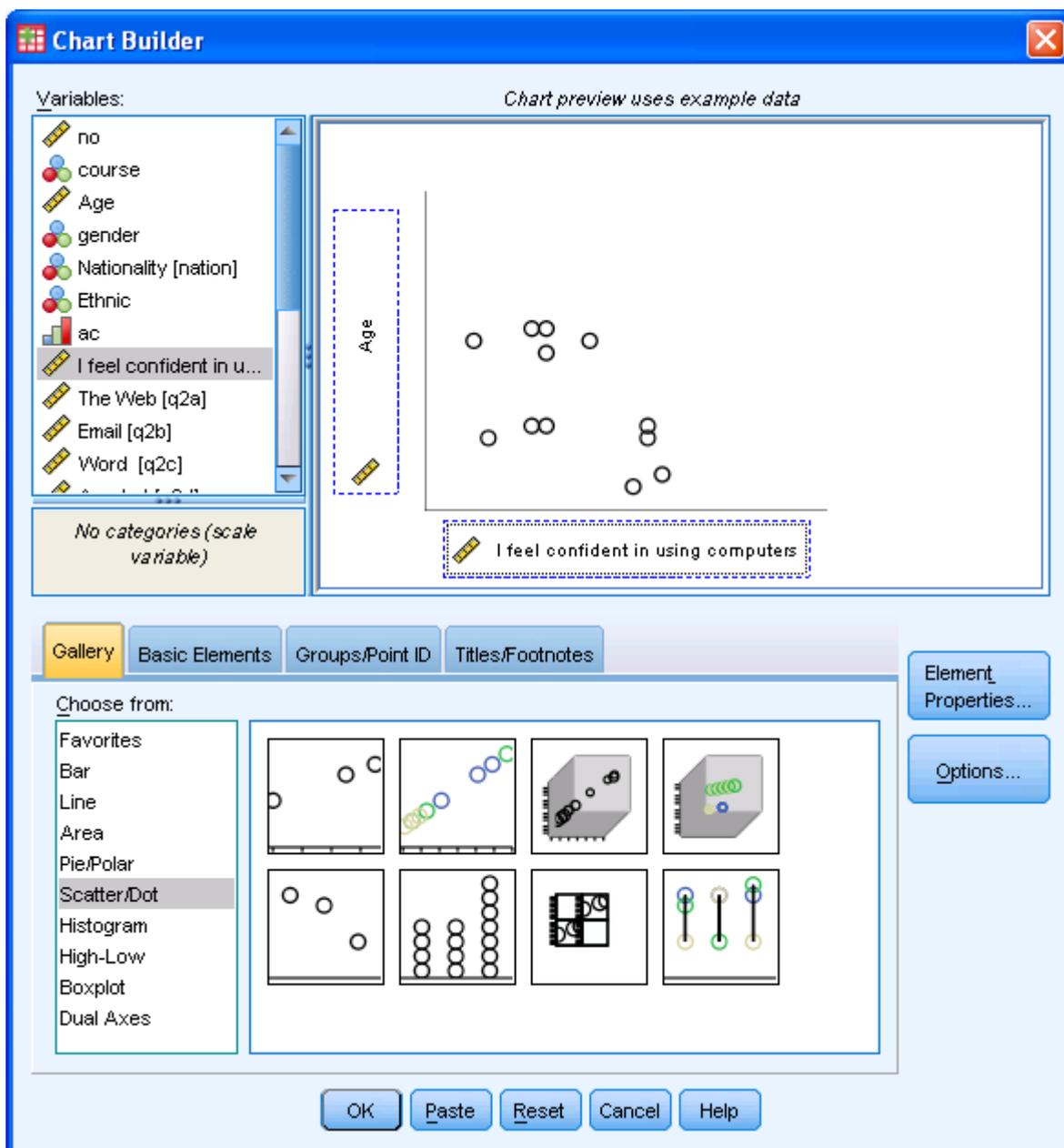


What this shows is that (as 1 = *never* to 5 = *most days* for all these variables) that Acrobat is used less than the other three applications. Whereas boxplots consider ranges and inter quartile ranges, error bars give a mean and (by default) 95% confidence intervals (i.e. where you expect 95% of values to lie). If this was a normally distributed sample it would indicate that the web, email and word are all used with much the same frequency, and Acrobat much less than the others. This is probably correct, but as the data are probably not normally distributed we would need to explore this with different methods, which will be done in later chapters (non-parametric tests).

### Scatterplots

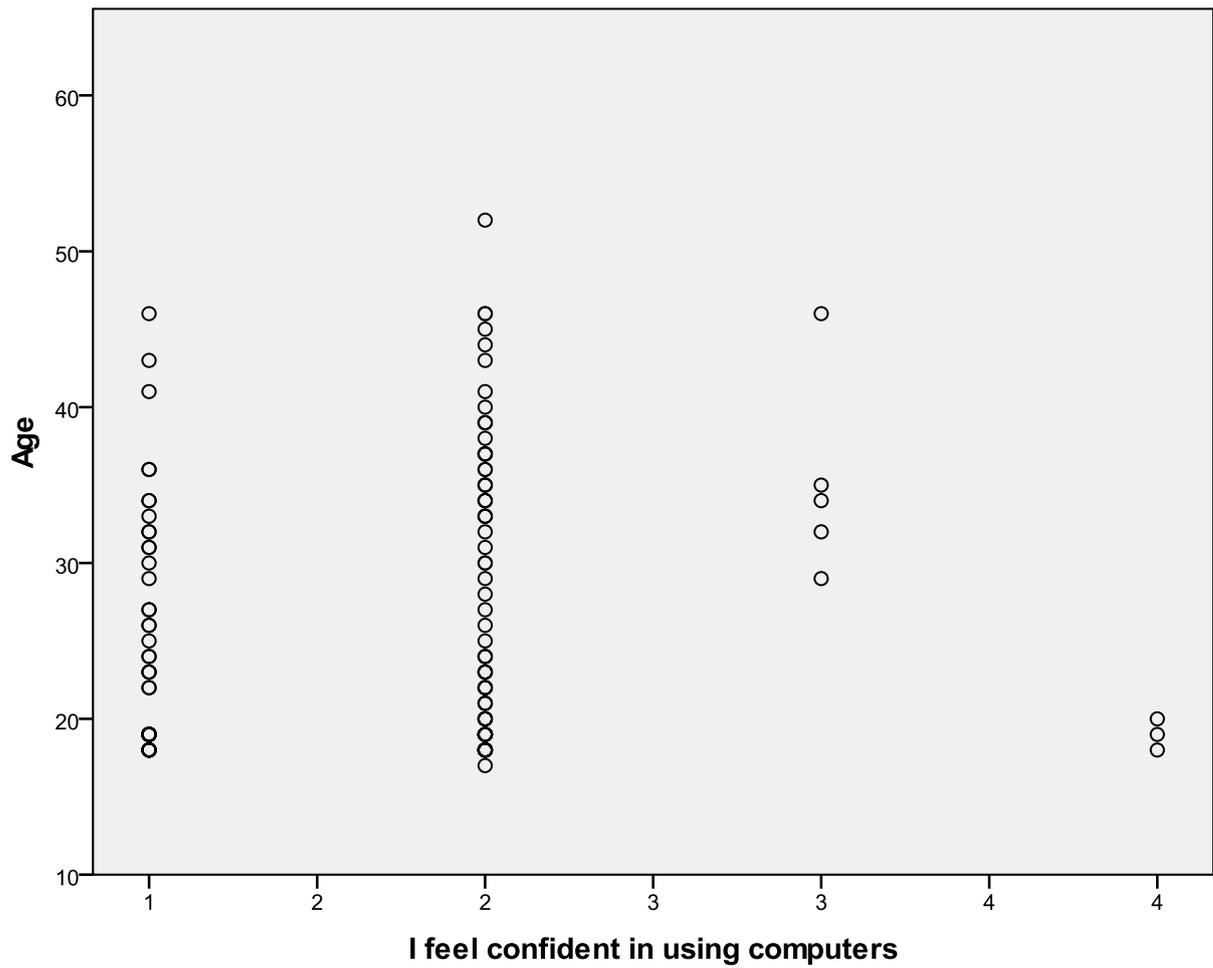
If you have two variables that you think are related, provided neither is nominal a scatterplot is one answer to see if they are. See Figure 55 where I have chosen *Scatter/Dot* from the **Chart Builder** and then selected the two variables) and following this is the graph (Figure 56).

Figure 55: Steps in making a scatterplot



The scatterplot (see Figure 56) appears to show a weak relationship between generally higher computer confidence with lower age of student (as you might expect). But the picture is hardly clear cut. In the lower levels of computer confidence all ages are shown, only in the very highest where only a few (three) are seen is age apparent, as all three are young. But three young people who are highly confident in using computers, against a sea of students of all ages with all levels of confidence a relationship does not make. This needs further exploration, and is so explored in correlation later in the book.

**Figure 56:** Scatterplot



### Conclusion

We have looked at some ways of showing data. In no case are we doing significance testing, this is more to show you how you can get a message across. If you need to test an hypothesis you should use one of the inferential tests covered in other chapters.

# 5 Manipulating data

## Introduction

What I want to look at now are a series of methods that do not really fall into basic or advanced, but could be described as *useful*. Many of the methods relate to ordering material so you can get to grips with it.

This chapter is relatively long but most of this is graphical and the text is not unduly lengthy.

The datafile we are using in this chapter is *assets.sav*, see the *Appendix Datasets used in this text* for details.

## Sorting data

This can be useful to look at raw data and get an intuitive feel of what is going on. Suppose in the asset data you wanted to look at only the more severe cases. You can sort by rank outcome, using **Data -> Sort Cases**. If I accept the default it will give the less severe cases at the top of the file, so I select the descending sort order instead, see Figure 57.

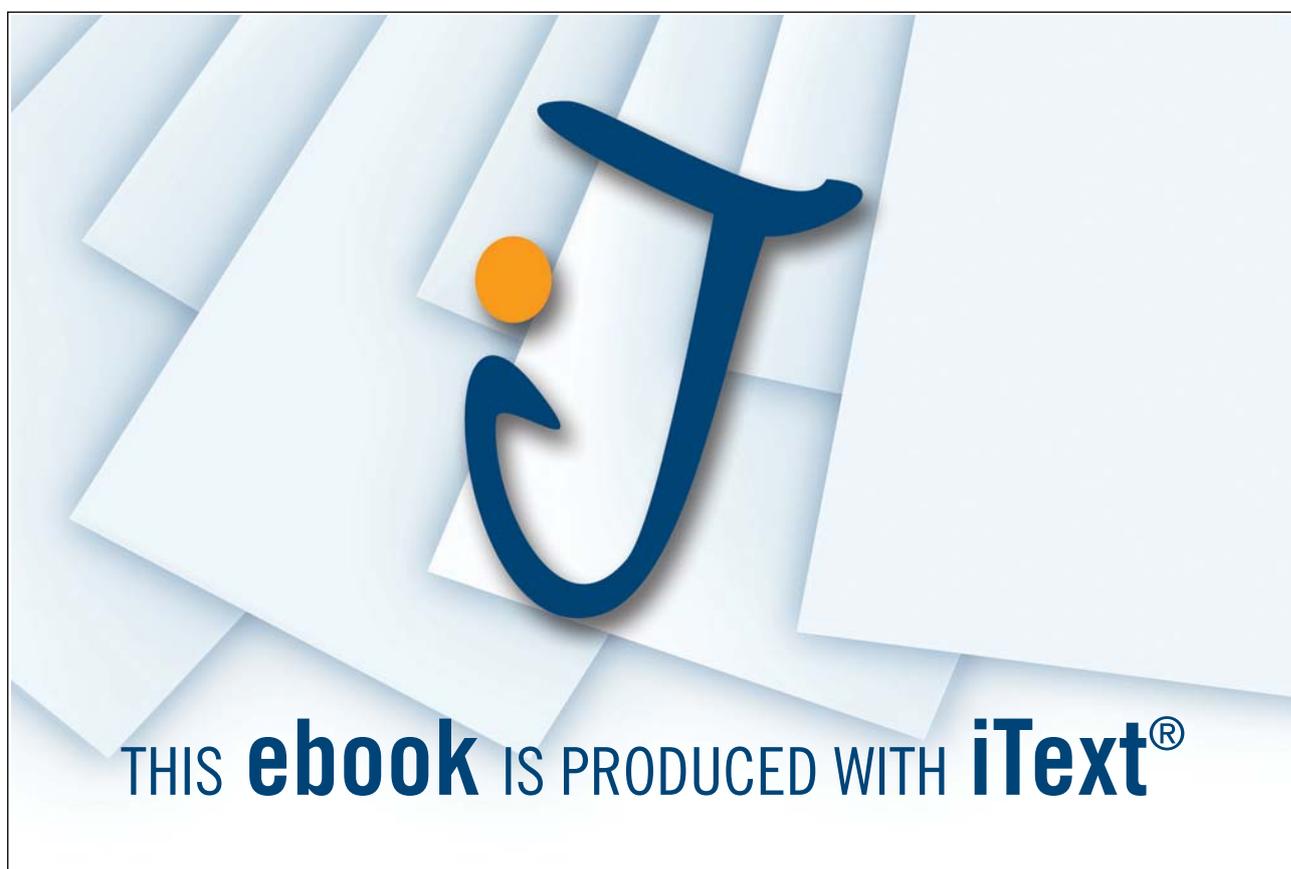


Figure 57: Sort dialogue

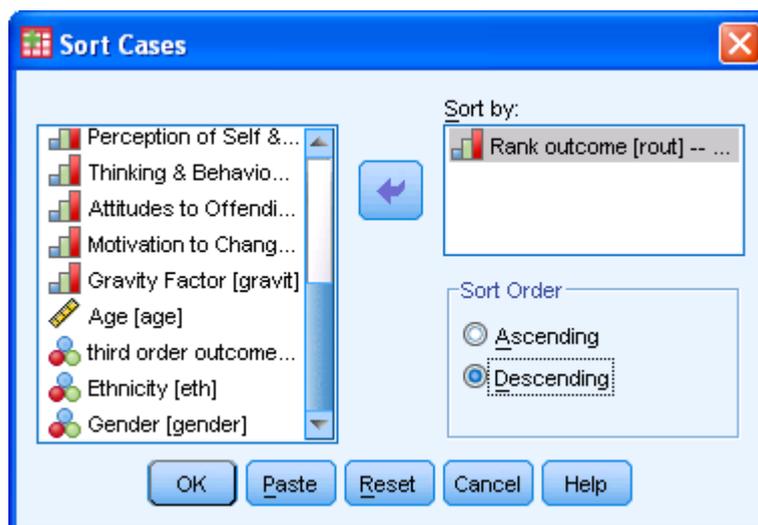


Figure 58: Sorted by rank outcome

	ass8	ass9	ass10	ass11	ass12	gravit	age	outcome	rou	eth	gender
1	2	2	2	2	2	5	19	Custodial s...	3rd order	White	Female
2	0	2	1	2	0	6	19	Custodial s...	3rd order	White	Male
3	2	3	4	2	3	6	19	Custodial s...	3rd order	White	Female
4	1	4	4	3	4	4	19	Custodial s...	3rd order	Mixed	Male
5	0	3	3	2	4	5	19	Custodial s...	3rd order	White	Male
6	0	0	3	2	0	5	19	Custodial s...	3rd order	White	Male
7	0	2	2	3	3	4	19	Custodial s...	3rd order	White	Male
8	4	3	4	4	2	5	19	Custodial s...	3rd order	White	Male
9	2	1	2	3	3	2	19	Custodial s...	3rd order	White	Male
10	0	2	3	3	3	6	19	Custodial s...	3rd order	White	Male
11	1	0	3	0	0	3	19	Custodial s...	3rd order	White	Male
12	2	1	2	0	1	7	19	Custodial s...	3rd order	White	Male
13	1	2	1	3	2	6	19	Custodial s...	3rd order	White	Male
14	0	0	3	3	2	3	19	Custodial s...	3rd order	White	Male
15	0	2	2	2	1	6	18	Custodial s...	3rd order	Black or Bl...	Male
16	0	1	3	3	1	5	18	Custodial s...	3rd order	White	Male
17	0	2	3	2	2	7	18	Custodial s...	3rd order	Black or Bl...	Male
18	3	1	3	3	1	7	18	Custodial s...	3rd order	White	Female
19	2	3	3	3	1	4	18	Custodial s...	3rd order	White	Female
20	2	4	3	3	3	6	18	Custodial s...	3rd order	Black or Bl...	Male
21	0	1	1	1	1	4	18	Custodial s...	3rd order	Mixed	Male
22	3	3	4	4	3	5	18	Custodial s...	3rd order	White	Male
23	3	1	3	3	2	5	18	Custodial s...	3rd order	Mixed	Male
24	2	4	3	4	4	5	18	Custodial s...	3rd order	White	Male

I can see (Figure 58) that the custodial sentencing occurs in many ethnic groups and both genders.

We could sort by more than one variable, for example in Figure 59 I have also sorted on ethnicity, this time ascending, so all the Asians with custodial sentences are at the top. We can see at once there are only five Asians, furthermore two cases have no ethnic group identified, which might prompt a review of the records to see if the data are missing or just not entered, see Figure 60.

Figure 59: Sort on two variables

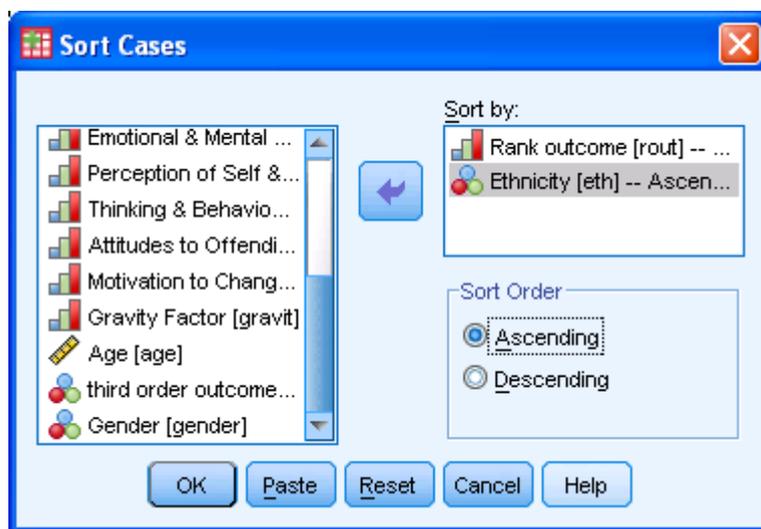


Figure 60: Sorted on custody and ethnic group

www.sylvania.com

We do not reinvent the wheel we reinvent light.

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM OSRAM SYLVANIA



3: eth 1 Visible: 18 of 18 Variables

	ass8	ass9	ass10	ass11	ass12	gravit	age	outcome	rout	eth	gender
1	0	2	2	2	2	.	18	Custodial s...	3rd order	.	.
2	4	4	4	3	4	6	18	Custodial s...	3rd order	.	.
3	0	1	3	2	2	6	17	Custodial s...	3rd order	Asian	Male
4	1	1	2	2	3	6	17	Custodial s...	3rd order	Asian	Male
5	0	1	3	2	3	6	16	Custodial s...	3rd order	Asian	Male
6	0	3	3	4	2	6	14	Custodial s...	3rd order	Asian	Male
7	0	0	3	2	2	6	13	Custodial s...	3rd order	Asian	Male
8	0	2	2	2	1	6	18	Custodial s...	3rd order	Black or Bl...	Male
9	0	2	3	2	2	7	18	Custodial s...	3rd order	Black or Bl...	Male
10	2	4	3	3	3	6	18	Custodial s...	3rd order	Black or Bl...	Male
11	0	3	3	3	4	5	18	Custodial s...	3rd order	Black or Bl...	Male
12	0	2	2	2	0	6	18	Custodial s...	3rd order	Black or Bl...	Male
13	3	0	3	3	2	8	18	Custodial s...	3rd order	Black or Bl...	Male
14	0	0	3	3	2	6	18	Custodial s...	3rd order	Black or Bl...	Male
15	3	3	3	3	4	7	17	Custodial s...	3rd order	Black or Bl...	Male
16	0	2	3	2	2	6	17	Custodial s...	3rd order	Black or Bl...	Male
17	1	2	2	2	2	4	17	Custodial s...	3rd order	Black or Bl...	Male
18	2	2	4	2	1	6	17	Custodial s...	3rd order	Black or Bl...	Male
19	1	1	3	2	1	5	15	Custodial s...	3rd order	Black or Bl...	Male
20	1	4	4	3	4	4	19	Custodial s...	3rd order	Mixed	Male
21	0	1	1	1	1	4	18	Custodial s...	3rd order	Mixed	Male
22	3	1	3	3	2	5	18	Custodial s...	3rd order	Mixed	Male
23	0	2	3	2	0	5	18	Custodial s...	3rd order	Mixed	Female
24	1	3	2	3	1	6	18	Custodial s...	3rd order	Mixed	Male

Data View Variable View PASW Statistics Processor is ready

## Aggregating

Sometime you want to have a more condensed dataset, for example if you wanted the highest gravity case for each ethnic group and gender. Putting gravity into the aggregated variable, note it defaults to function “mean” which is often what you want, but not in this case, use **Data -> Aggregate** and then Figure 61. So we choose “function” and then “maximum” as in Figure 62. Finally the default position is to aggregate variables into the current dataset, which is never what I want to do, so I choose to put the aggregated data in a new dataset, see Figure 63. N.B. you can save it in a new file, but here I have not done this. Note in versions of SPSS before 14 you MUST save to a new file.

Figure 61: Dialogue for aggregating

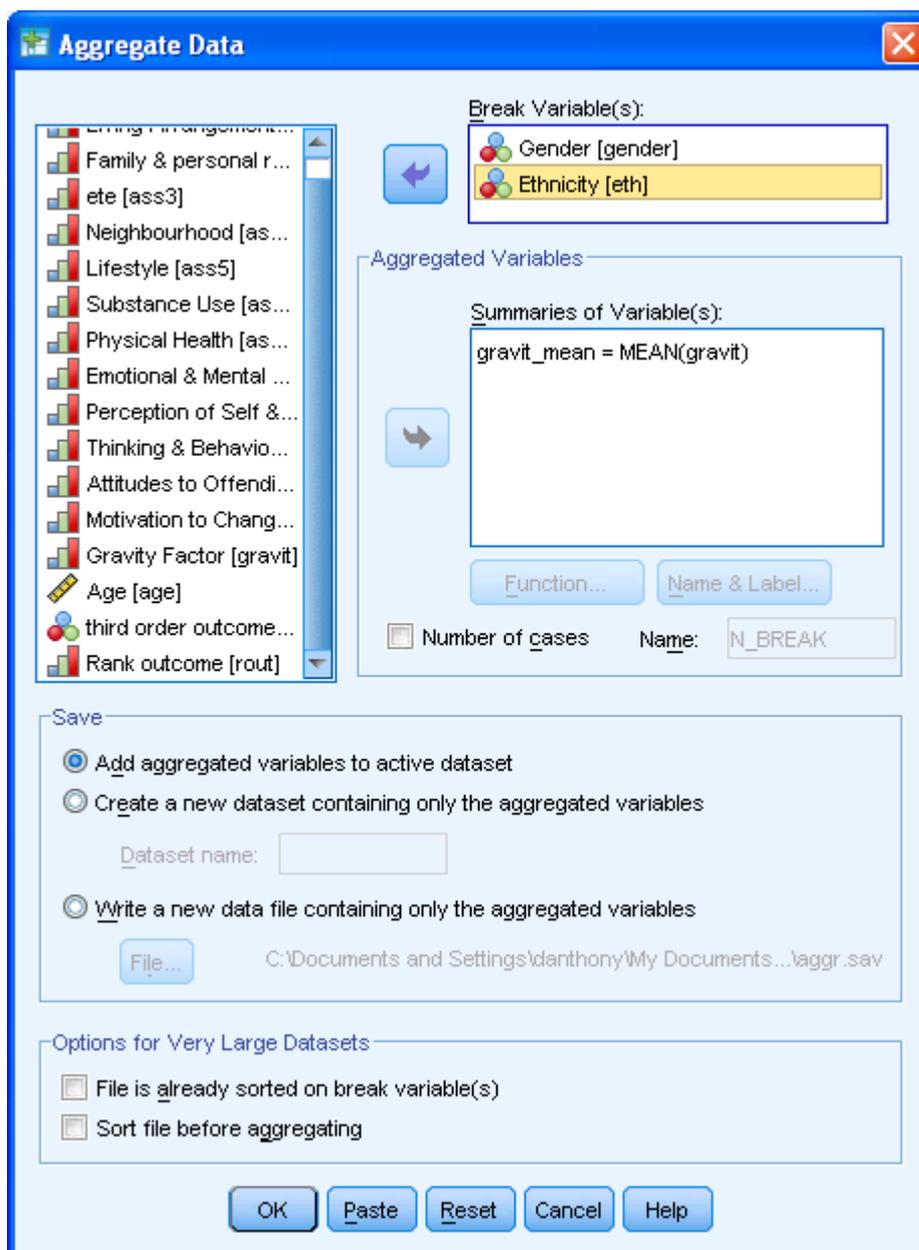
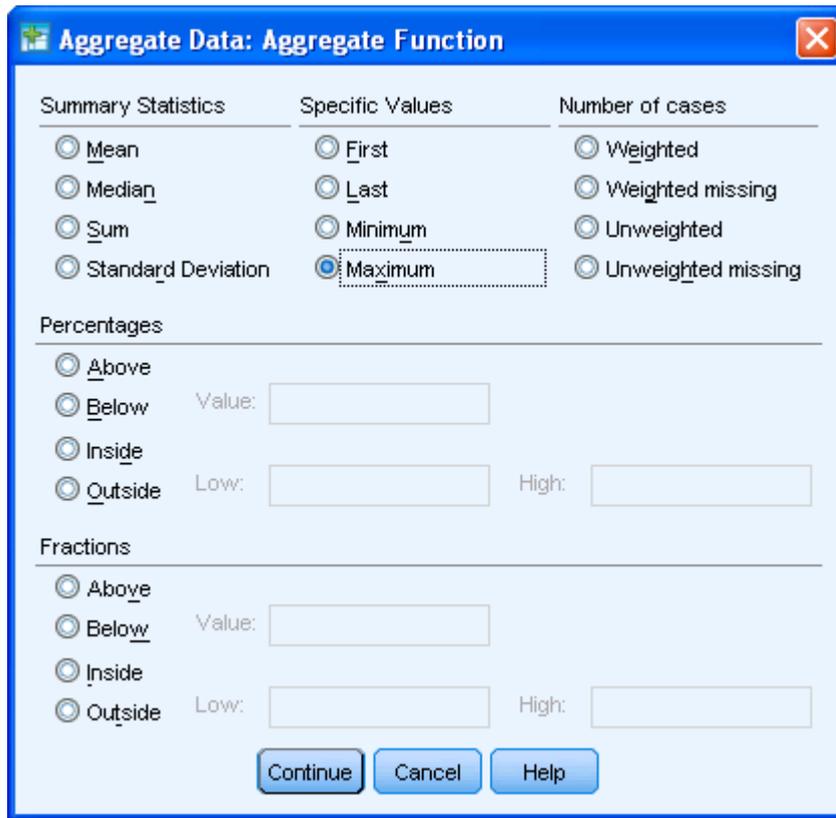


Figure 62: Selecting a different function



**CHALLENGING PERSPECTIVES**

## Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

**We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.**

To find out more and apply, visit [www.jobs.eads.com](http://www.jobs.eads.com). You can also find out more on our **EADS Careers Facebook page**.

**AIRBUS    ASTRIUM    CASSIDIAN    EUROCOPTER**

**EADS**

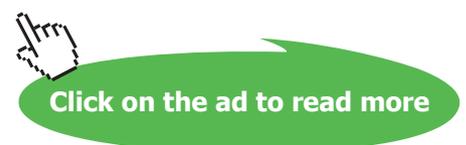
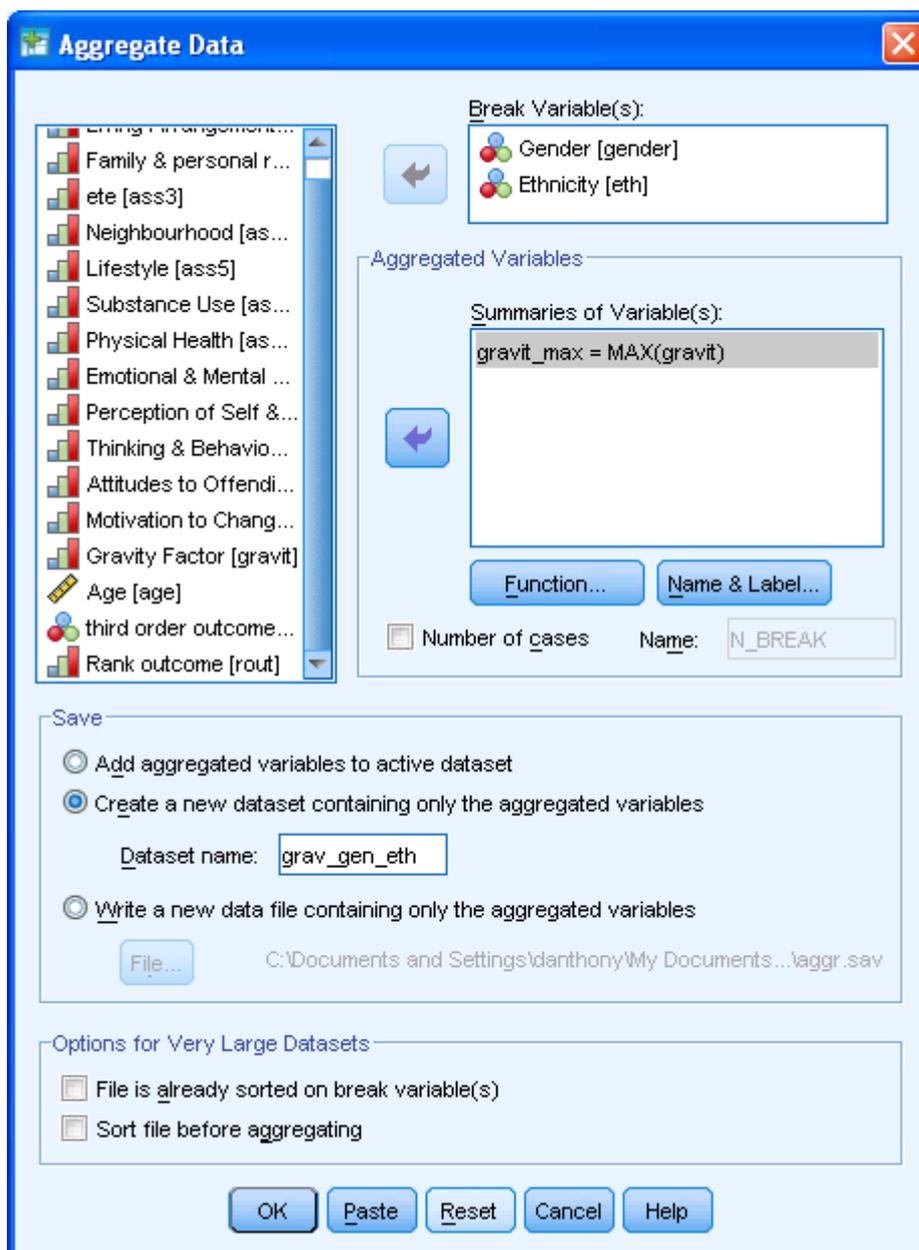


Figure 63: Creating a new dataset



The new dataset is seen in Figure 64 and is much simpler to view. Note it gives females slightly lower maxima. Medians however are more similar, see Figure 65. Note there are figures for no stated gender which reminds us there are missing data for gender.

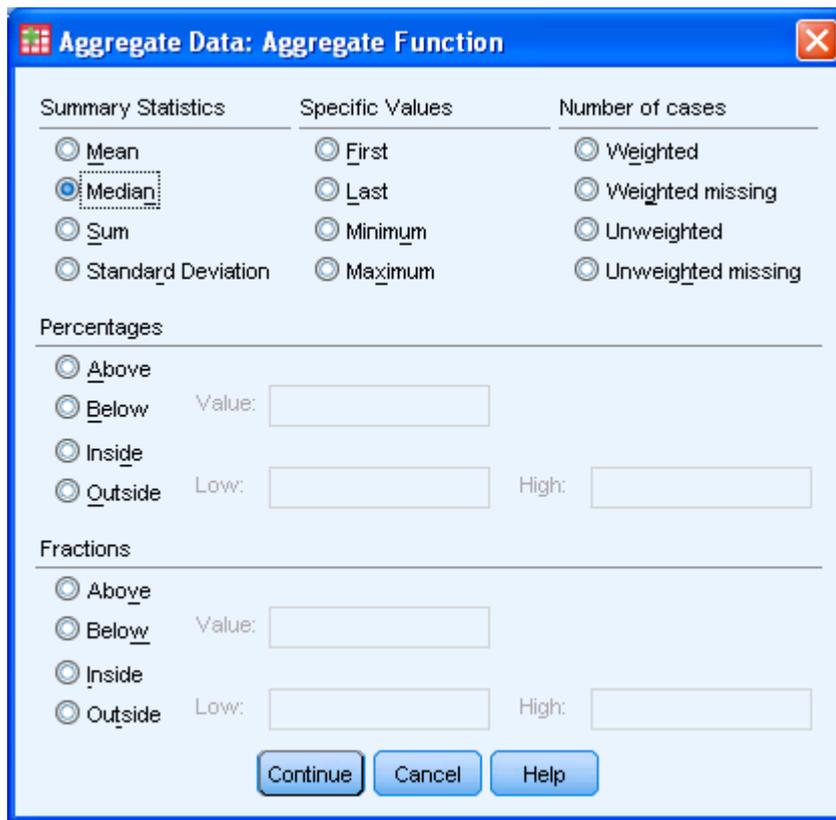
Figure 64: New dataset

The screenshot shows the PASW Statistics Data Editor window titled "Untitled2 [grav\_gen\_eth] - PASW Statistics Data Editor". The window contains a menu bar (File, Edit, View, Data, Transform, Analyze, Direct Marketing, Graphs, Utilities, Add-ons, Window, Help) and a toolbar with various icons. The main area displays a data grid with 12 rows and 13 columns. The first three columns are labeled "gender", "eth", and "gravit\_max". The remaining ten columns are labeled "var". The data is as follows:

	gender	eth	gravit_max	var								
1	.	.	7									
2	.	White	4									
3	Male	.	6									
4	Male	Asian	6									
5	Male	Black or Bl...	8									
6	Male	Mixed	8									
7	Male	White	8									
8	Female	.	3									
9	Female	Asian	5									
10	Female	Black or Bl...	6									
11	Female	Mixed	6									
12	Female	White	7									
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												

The interface also shows a "Data View" button and a "Variable View" button at the bottom left, and a status bar at the bottom right indicating "PASW Statistics Processor is ready".

Figure 65: Medians aggregated

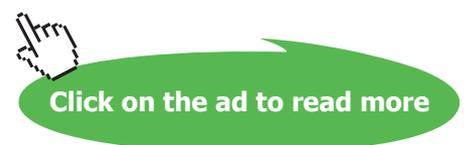


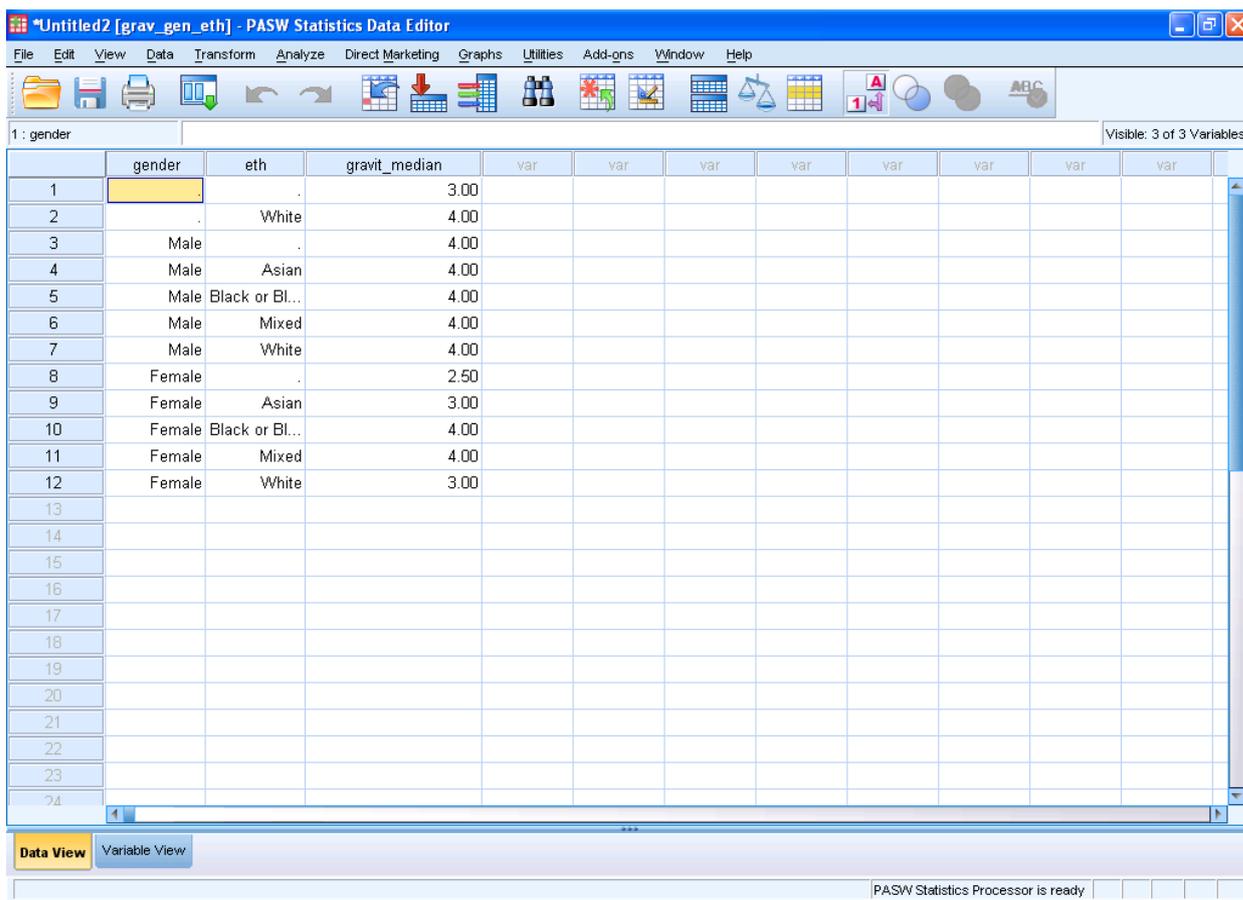
360°  
thinking.

**Deloitte.**

Discover the truth at [www.deloitte.ca/careers](http://www.deloitte.ca/careers)

© Deloitte & Touche LLP and affiliated entities.





### Splitting data

Suppose you wanted a histogram of gravity, but for males separate to females. In the latest version of SPSS you can do this by specifying a row or column variable, see Figure 66 and Figure 67. however the ranges are the same for males and females which gives a less useful graph for females in some senses.

Figure 66: Dialogue for histogram

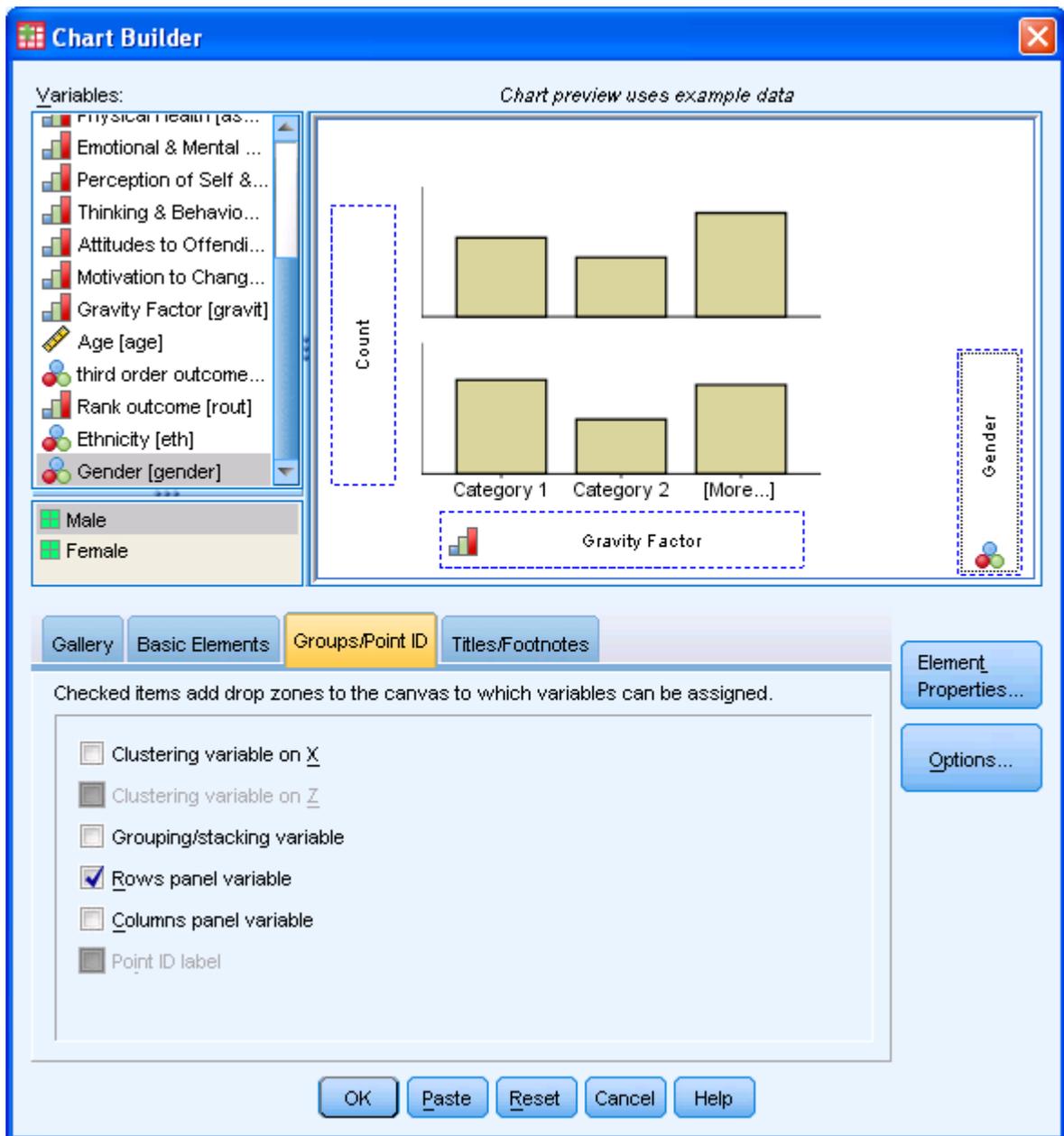
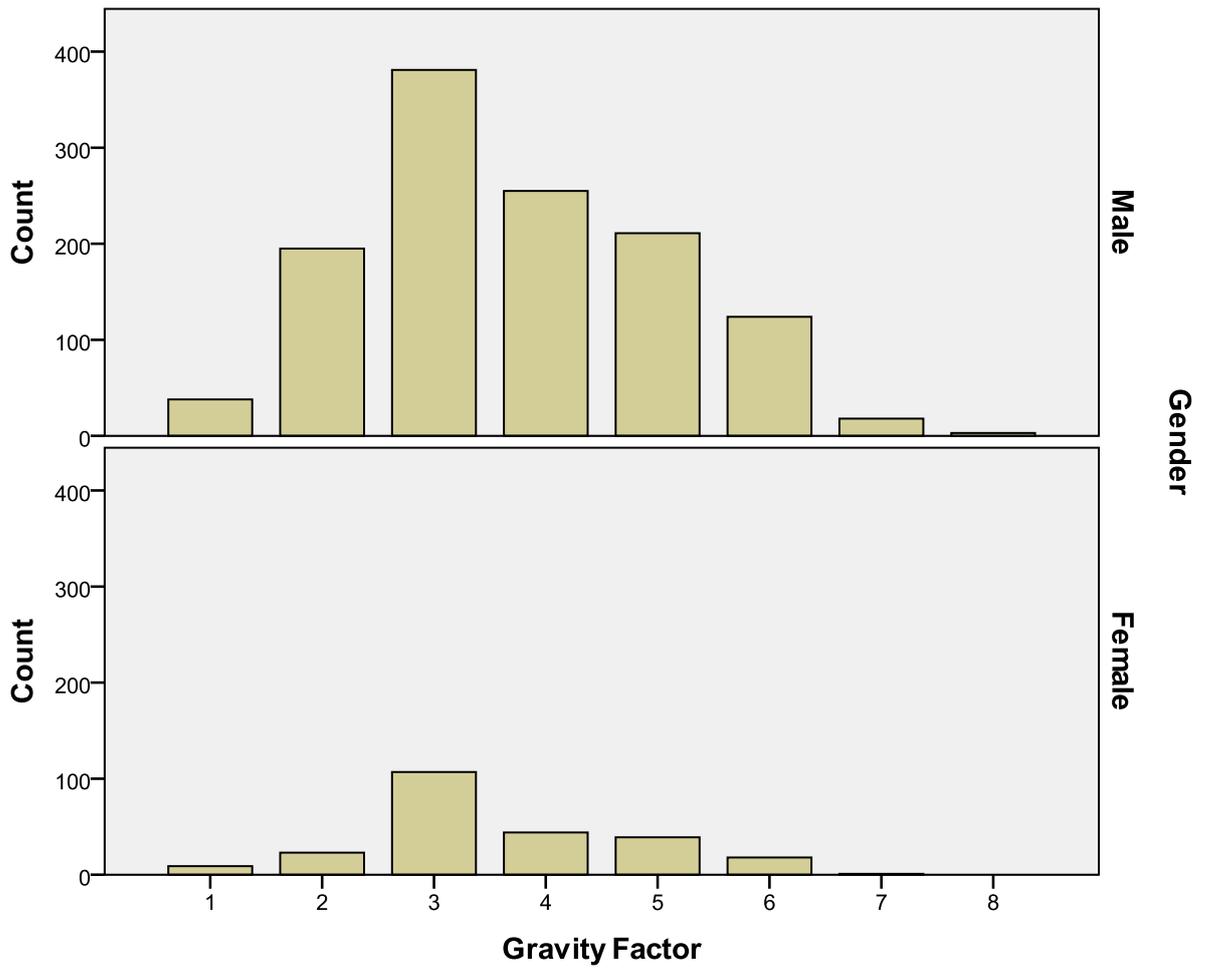
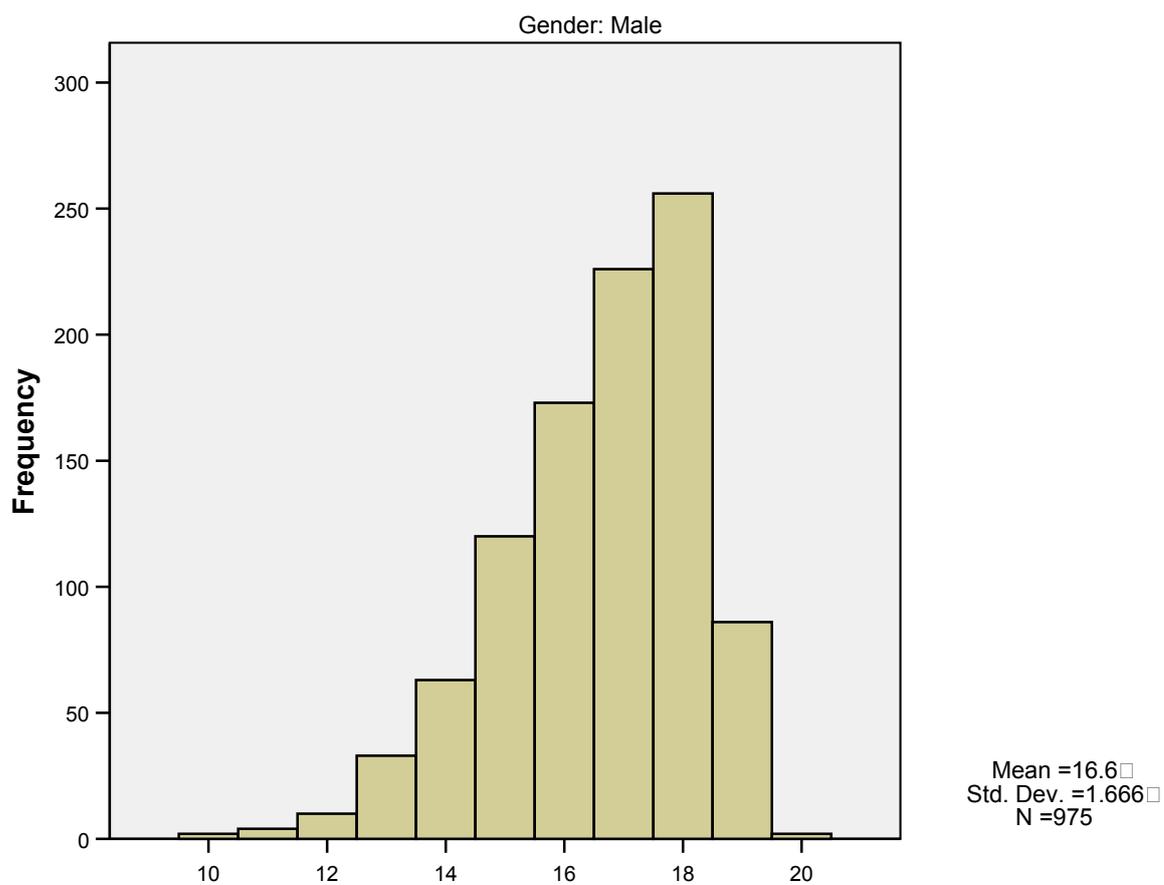


Figure 67: Histogram

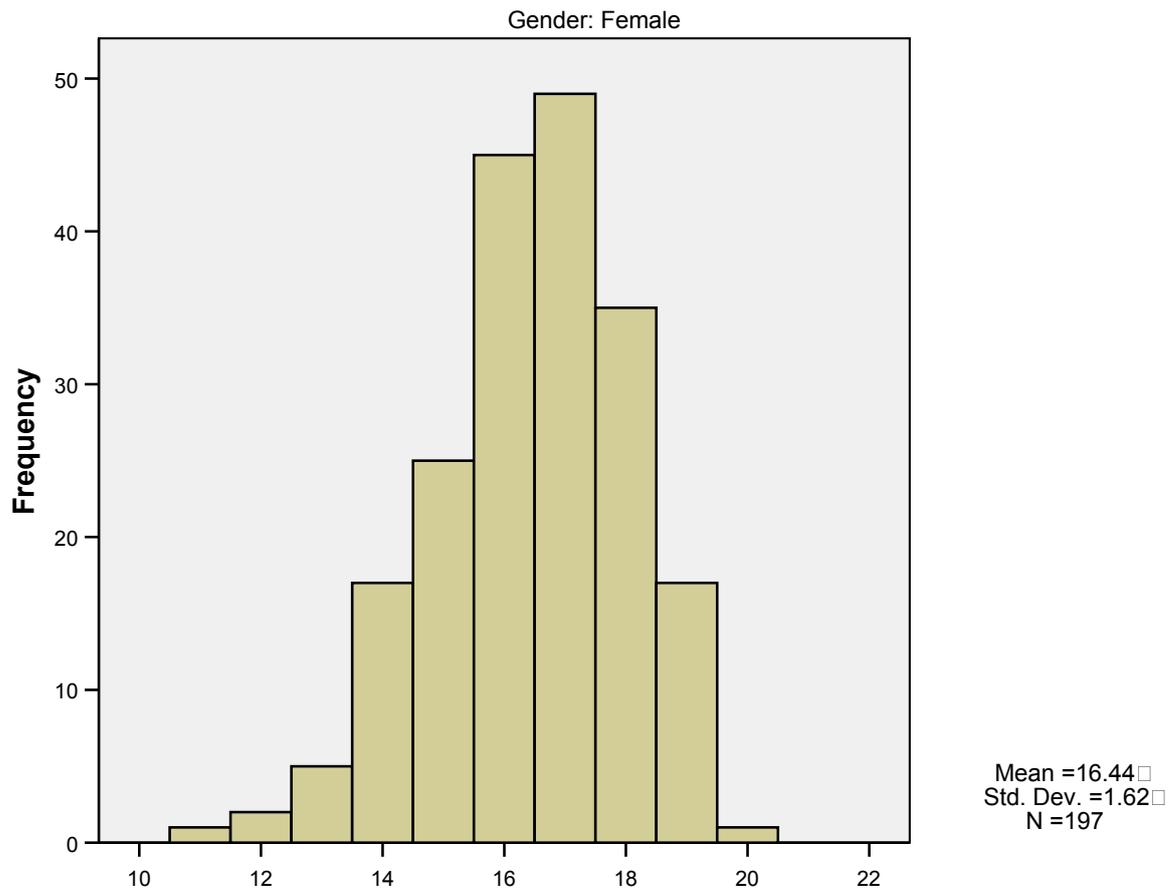


After splitting (**Data** -> **Split File**) the data we get two very different histograms with a “nicer” range for females, see Figure 68 and Figure 69.

**Figure 68:** Histograms after splitting data males

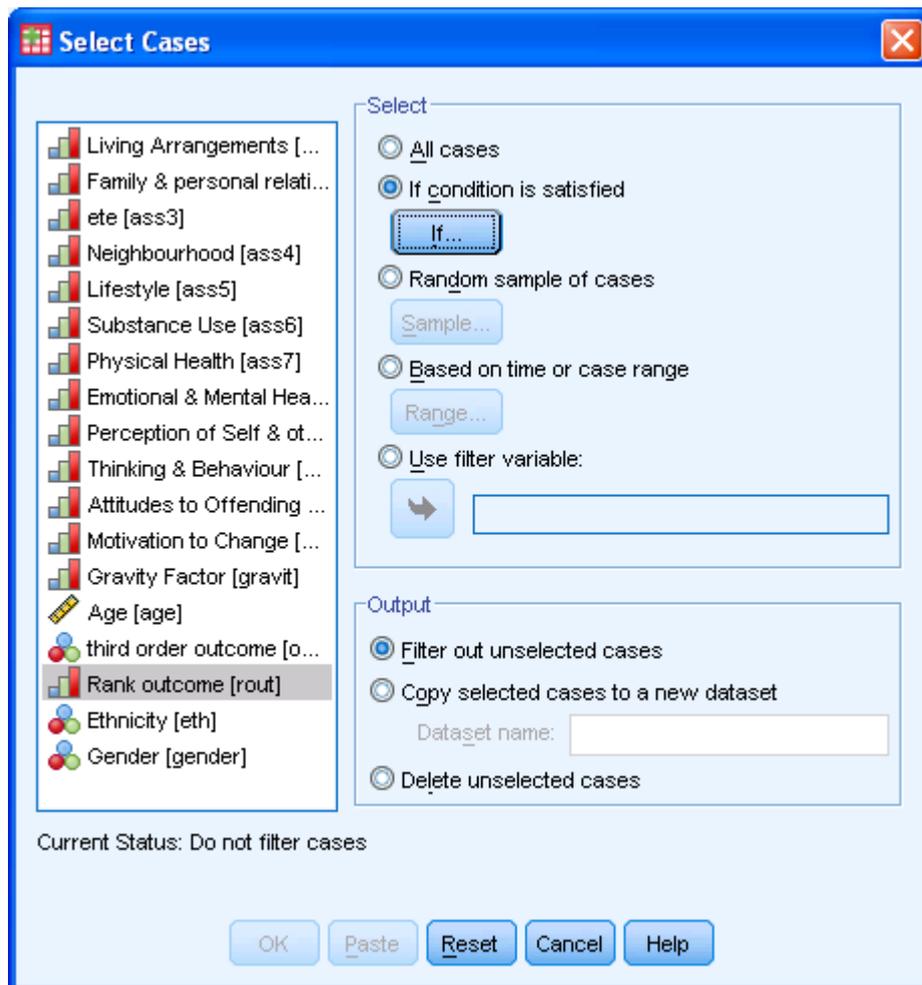


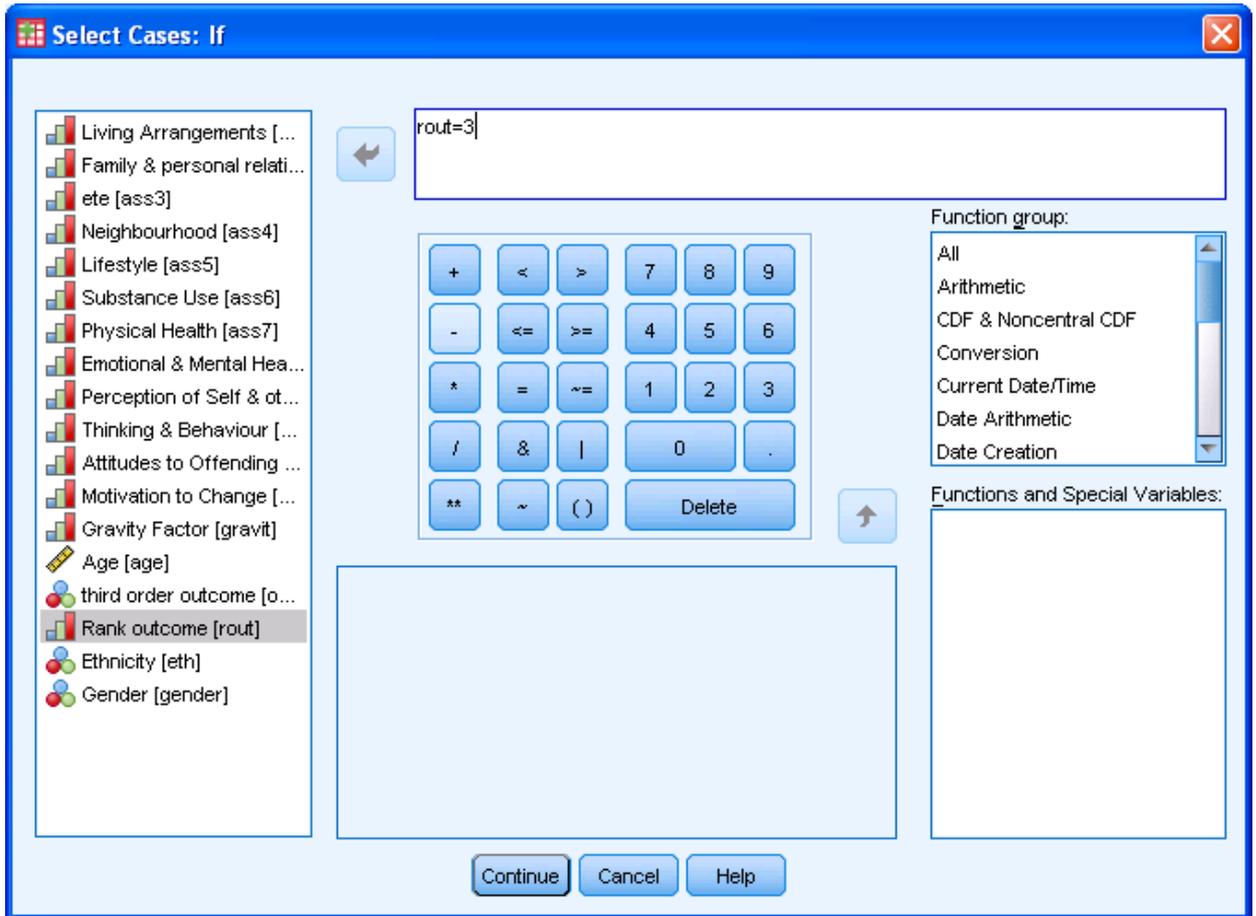
**Figure 69:** females



Sometimes you just want to concentrate on some values. For example if I wanted to exclude all but the most serious cases. I use the select procedure (**Data -> Select Cases**) see Figure 70.

Figure 70: Select dialogue box

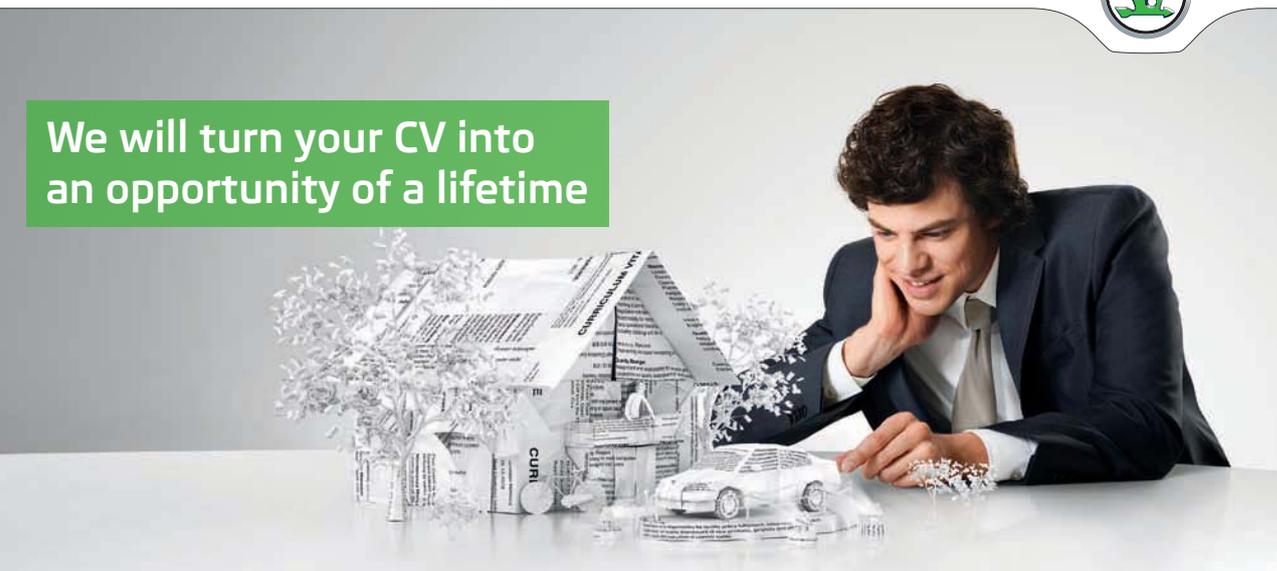




SIMPLY CLEVER

**ŠKODA**  


**We will turn your CV into an opportunity of a lifetime**

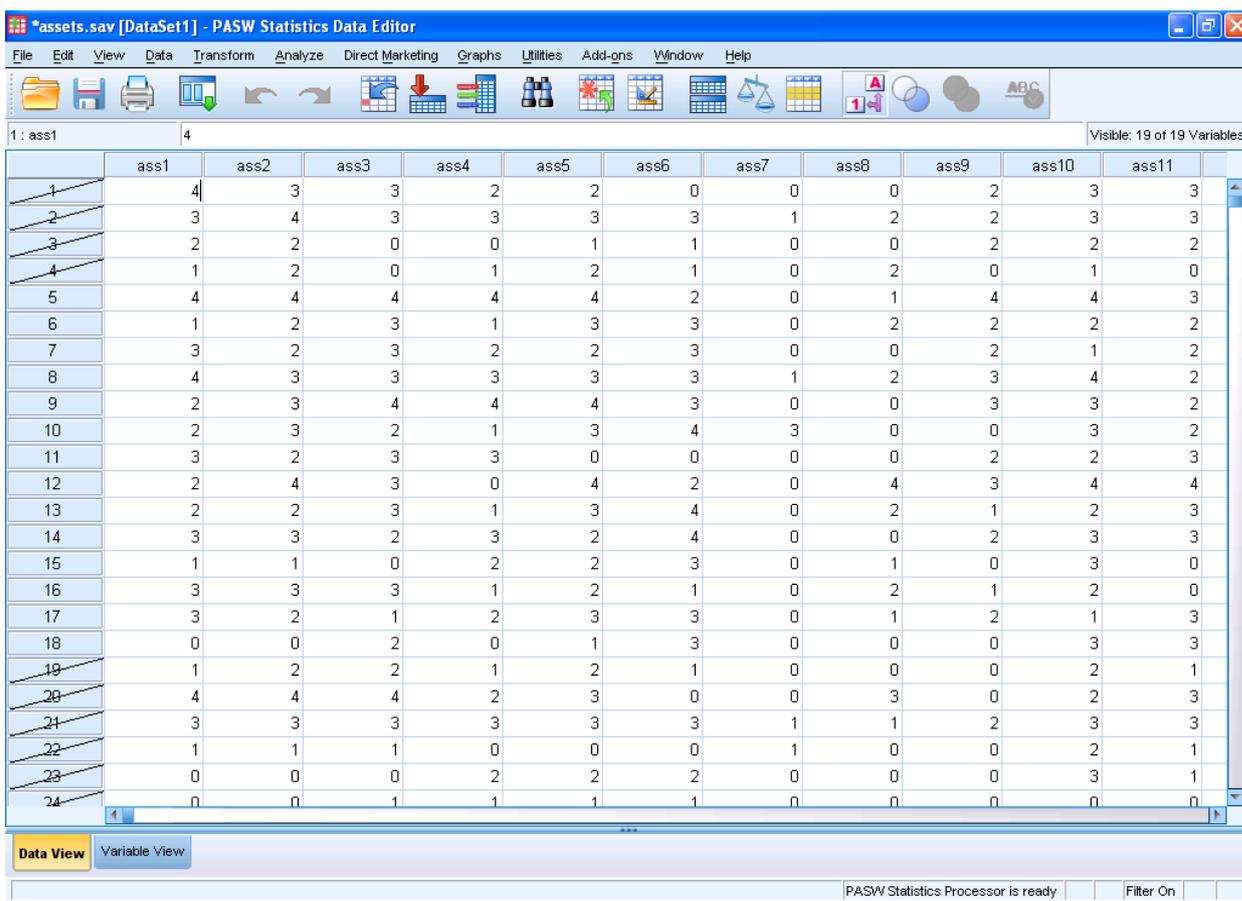


Do you like cars? Would you like to be a part of a successful brand? We will appreciate and reward both your enthusiasm and talent. Send us your CV. You will be surprised where it can take you.

Send us your CV on [www.employerforlife.com](http://www.employerforlife.com)



Figure 71: Data display for selected items



Note in Figure 71 de-selected cases are shown with a diagonal line through their case number. Also note I chose the default option of filtering data, normally you do not want to remove the cases. All analyses are now done on the selected cases only, for example if I keep the data split by gender and only custodial cases are selected I get Table 6.

Table 6: Ethnicity

Gender			Frequency	Percent	Valid Percent	Cumulative Percent
Male	Missing	System	4	100.0		
		Valid				
		Asian	6	4.2	4.2	4.2
		Black or Black British	15	10.4	10.4	14.6
		Mixed	20	13.9	13.9	28.5
Female	Valid	White	103	71.5	71.5	100.0
		Total	144	100.0	100.0	
		Mixed	2	20.0	20.0	20.0
		White	8	80.0	80.0	100.0
		Total	10	100.0	100.0	

But if I select all again, and remove splitting, see Figure 72 and Figure 73 I get all data back for analysis, see Table 7.

**Figure 72:** Select all cases

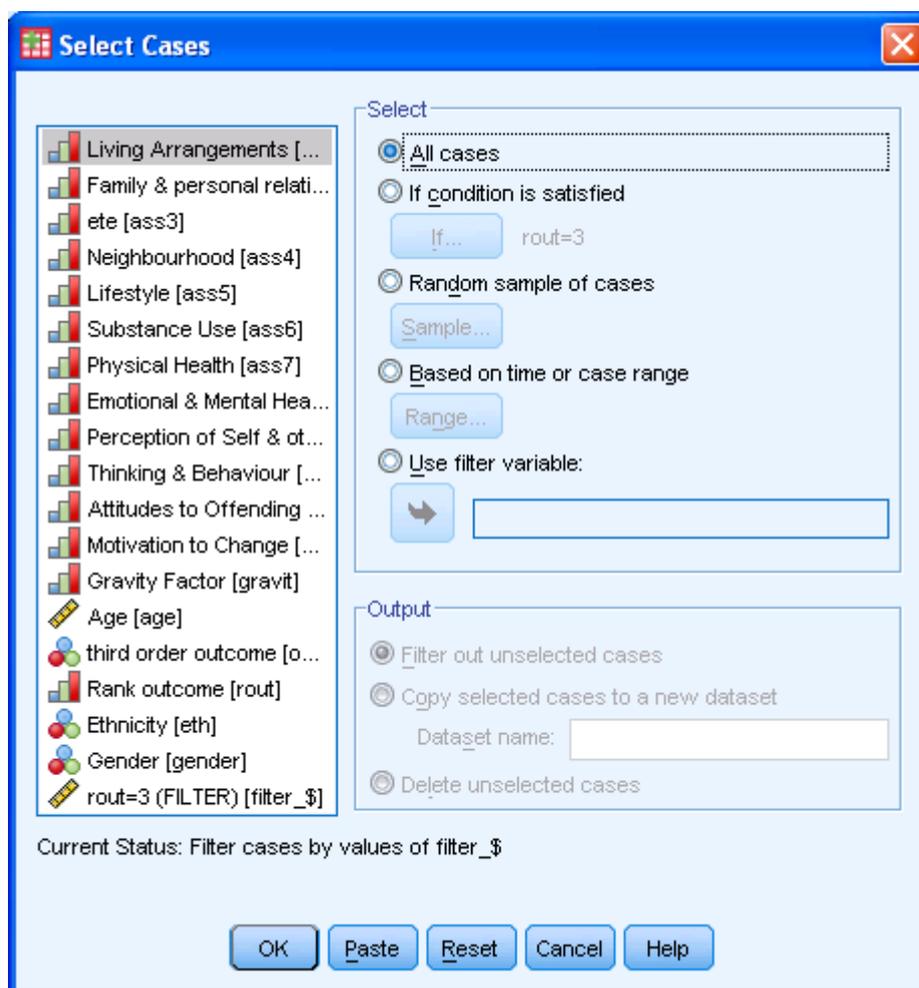


Figure 73: Remove split

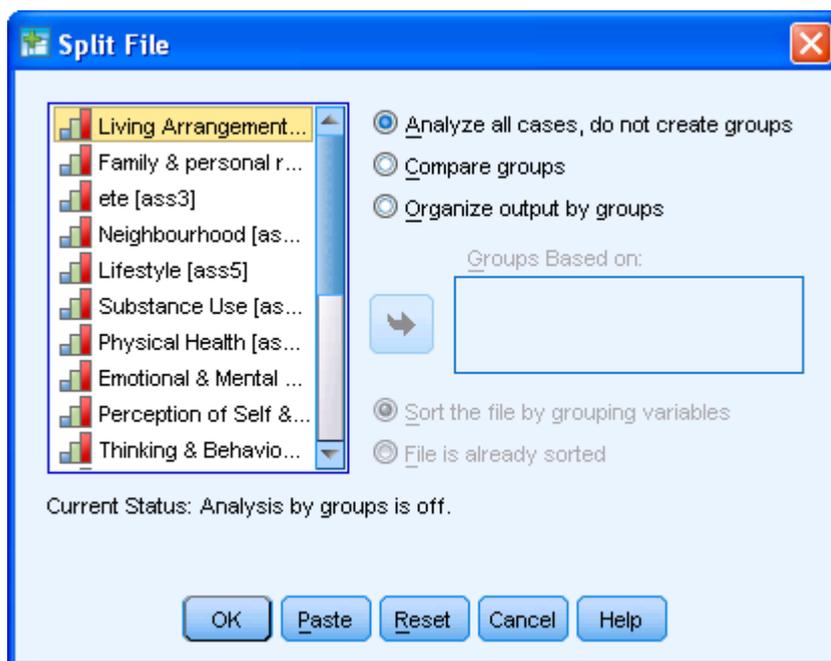


Table 7: All data reported

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Asian	112	4.1	7.1	7.1
	Black or Black British	93	3.4	5.9	13.0
	Mixed	105	3.9	6.6	19.6
	White	1272	47.1	80.4	100.0
	Total	1582	58.6	100.0	
	System	1117	41.4		
Total		2699	100.0		

### Computing new variables

We need to sum the asset sub-scores to create the total score, and SPSS allows us to compute this, see Figure 74, where I have used **Transform** -> **Compute Variable** to create a new variable *totass* that is the sum of the twelve variables *ass1* to *ass12*. In case I need to do this again (if I add new data the variable *totass* will not automatically be re-computed, this is different from a spreadsheet where it would be) I can put the commands into a **syntax** file by clicking on **Paste** in the dialogue box (you can do this in most situations in SPSS) and the commands will appear in their own syntax window (see Figure 75 ) which can be separately saved to a syntax file and re-loaded when needed. A syntax file (or part of it) can be run, see Figure 76. You will note also that any commands are pasted into the Output window (not in earlier versions of SPSS). If you are feeling particularly nerdy you can write your own commands in either the syntax window, or cut and paste from a word processor. I am not going to show you how to write in SPSS syntax in this text though – it is rarely necessary (if ever).

Figure 74: Computing a new variable

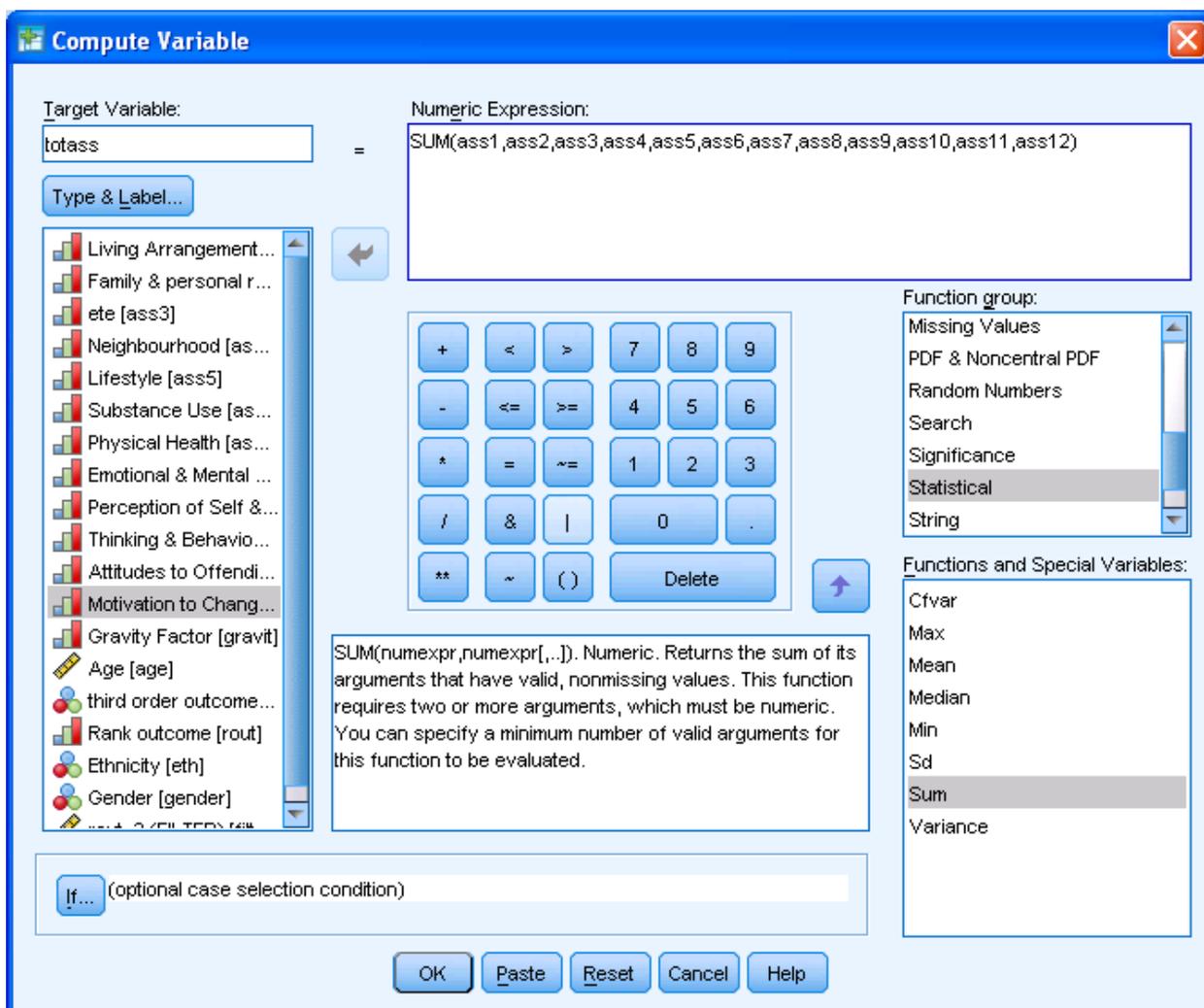
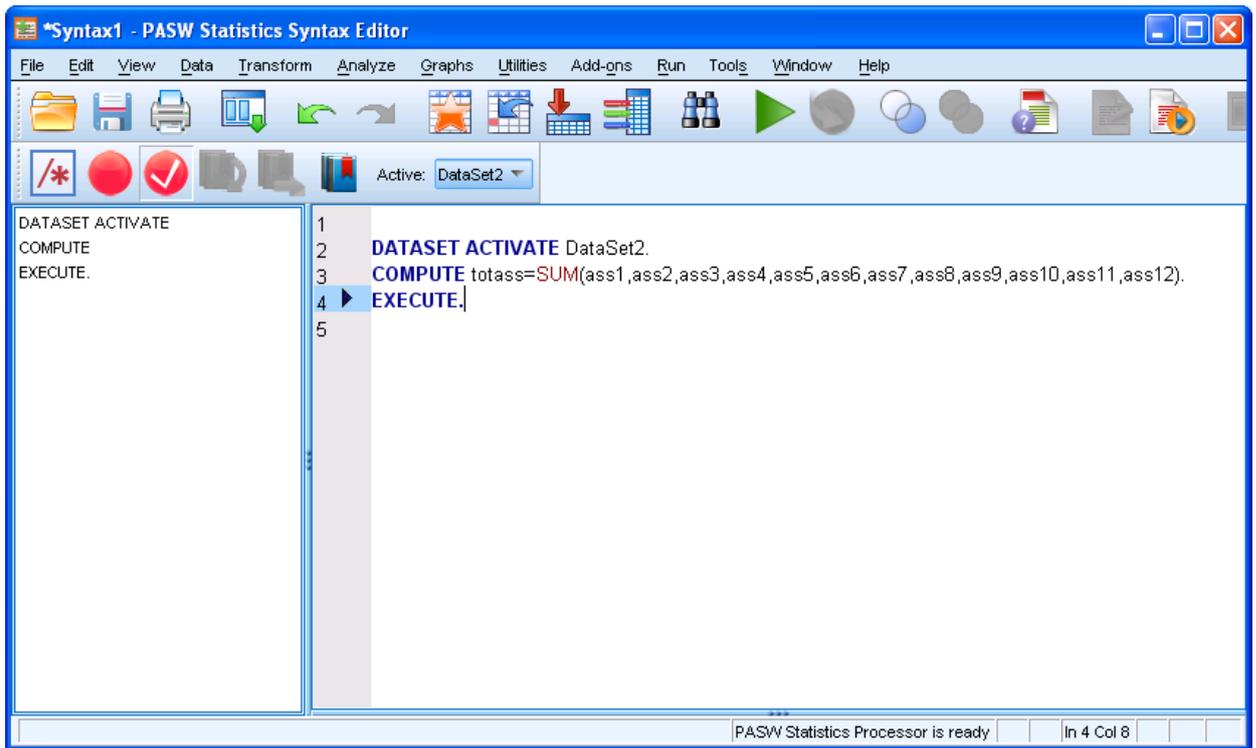


Figure 75: Syntax editor



Sweden Sverige

Linköping University –  
innovative, highly ranked,  
European

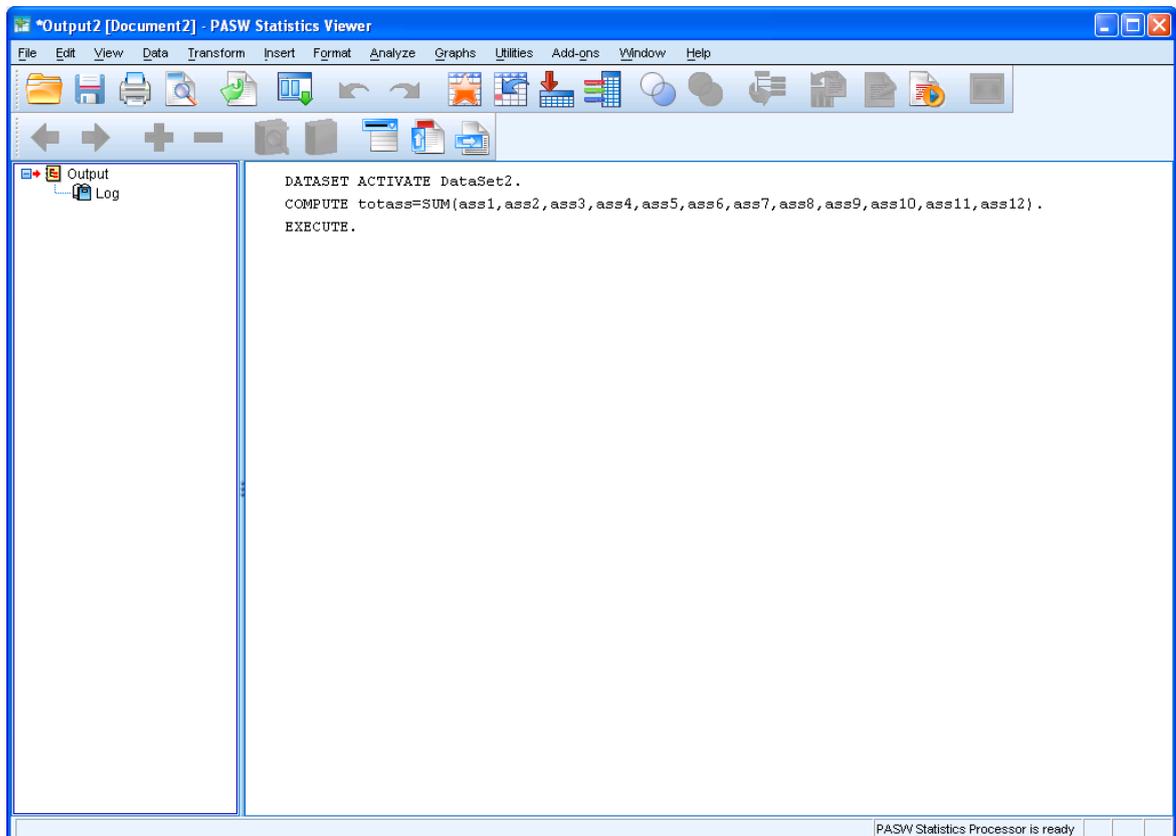
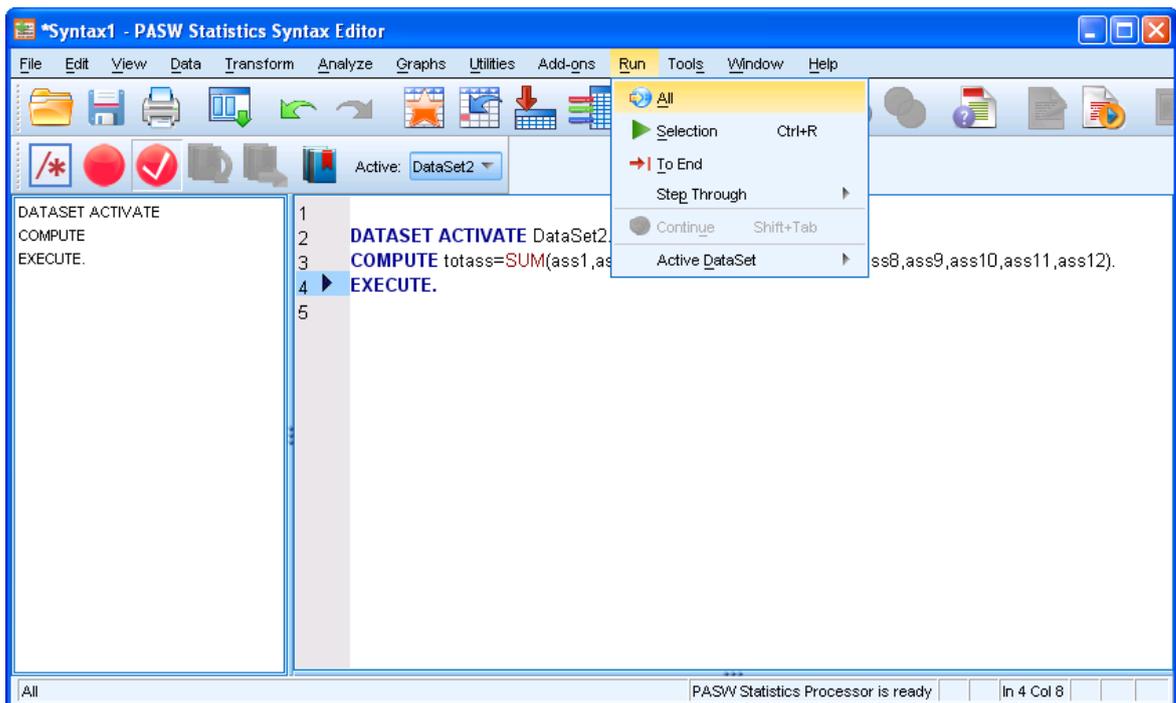
Interested in Computer Science? Kick-start your career  
with an English-taught master's degree.

→ Click here!

li.u LINKÖPING UNIVERSITY



Figure 76: Running syntax



	ass9	ass10	ass11	ass12	gravit	age	outcome	rout	eth	gender	totass
1	2	3	3	3	.	20	Non custo...	2nd order	White	Male	25.00
2	2	3	3	3	4	20	Non custo...	1st order	Mixed	Male	33.00
3	2	2	2	0	.	20	.	.	.	.	12.00
4	0	1	0	0	.	20	.	.	.	.	10.00
5	4	4	3	4	4	19	Custodial s...	3rd order	Mixed	Male	38.00
6	2	2	2	2	5	19	Custodial s...	3rd order	White	Female	23.00
7	2	1	2	0	6	19	Custodial s...	3rd order	White	Male	20.00
8	3	4	2	3	6	19	Custodial s...	3rd order	White	Female	34.00
9	3	3	2	4	5	19	Custodial s...	3rd order	White	Male	32.00
10	0	3	2	0	5	19	Custodial s...	3rd order	White	Male	23.00
11	2	2	3	3	4	19	Custodial s...	3rd order	White	Male	21.00
12	3	4	4	2	5	19	Custodial s...	3rd order	White	Male	32.00
13	1	2	3	3	2	19	Custodial s...	3rd order	White	Male	26.00
14	2	3	3	3	6	19	Custodial s...	3rd order	White	Male	28.00
15	0	3	0	0	3	19	Custodial s...	3rd order	White	Male	13.00
16	1	2	0	1	7	19	Custodial s...	3rd order	White	Male	19.00
17	2	1	3	2	6	19	Custodial s...	3rd order	White	Male	23.00
18	0	3	3	2	3	19	Custodial s...	3rd order	White	Male	14.00
19	0	2	1	0	1	19	Non custo...	2nd order	.	.	12.00
20	0	2	3	1	1	19	Non custo...	2nd order	.	.	26.00
21	2	3	3	2	3	19	Non custo...	2nd order	.	.	30.00
22	0	2	1	1	6	19	Non custo...	2nd order	.	.	8.00
23	0	3	1	0	.	19	Non custo...	2nd order	.	.	10.00
24	0	0	0	0	3	19	Non custo...	2nd order	Asian	Male	4.00

### Conclusion

I have shown a few simple ways to manipulate data to make various reports simpler or easier to view, or just to get a feel for the data in an exploration of it.

### Exercise

Use the datafile "assets", sort by gravity (descending) and find the most severe case that did not get a custodial sentence. This sort of exploration can identify "odd" cases.

# 6 Chi square

## Key points

- Cross tabulations tell you how many subjects are in each combination of two nominal variables
- Chi square is an inferential test of nominal data

At the end of this chapter you should be able to:

- Create cross tabulation tables
- Compute chi square and interpret the result

## Introduction

An hypothesis is simply a statement that is capable of being tested. For example “there is no significant difference between men and women with respect to repeating self harm by overdose” is an hypothesis because it could be right or wrong. However “what is the effect of gender on self harm” is not, this is a research question. Clearly research questions and hypotheses are related, but not the same thing. In this chapter we will look at the most simple inferential test, chi square. Inferential statistics test whether an hypothesis is right or wrong.

I joined MITAS because  
I wanted **real responsibility**

The Graduate Programme  
for Engineers and Geoscientists  
[www.discovermitas.com](http://www.discovermitas.com)



**Month 16**

I was a construction supervisor in the North Sea advising and helping foremen solve problems

Real work  
International opportunities  
Three work placements







## Significance

In all inferential tests there is one common feature, that of probability value or p value for short, often called significance, and in SPSS often shortened to sig. The p value is the probability of getting some set of data by pure chance. For example if I split a class into two random groups and measure their heights, you would expect the mean value for height to be similar (because they are random). However the two groups probably will not have identical mean values, due to chance and random fluctuation. A small difference in mean values is explained easily by chance, but a large difference may be significant. Clearly it matters how big the sample is, if the two groups consist on one student each, than a large difference means nothing, but two groups of (say) a hundred should only have very small different means. However even a large group can have a large difference by pure chance, so all fifty students in a randomly allocated group could be taller than the other randomly allocated group, but this is very unlikely.

I use a thought experiment to explain p values to my students, and this has always produced the same response over the last fifteen years. I ask you to imagine you are betting with a friend using a coin, heads they win, tails you win. If a head occurs you lose a pound (dollar, euro etc.) otherwise you win one. The first toss is a head. What do you make of this as you pass over your pound? Probably nothing as there is a one in two chance of a head. The second toss is also a head, again no big surprise. There is a one in two chance again, as the two tosses of the coin are independent (i.e. the result of one toss has no effect on the next). You would expect two consecutive heads one time in four. You hand over another pound. I continue with the third and fourth toss, both heads. This is getting annoying as you have now lost four pounds, but a one in sixteen chance of four heads in a row is not that surprising. What usually happens when the next toss is said to be a head is my students start looking a bit surprised, five in a row is getting a bit unbelievable, but it is not actually unbelievable as winning the lottery. By ten in a row they think something strange is going on and that this could mean cheating. But there is roughly one chance in a thousand that this could happen by chance. After losing one hundred tosses all students are convinced there is cheating. But there is a finite chance still that this could simply be a random sequence that just has 100 heads in a row. It is however much less likely than winning the lottery (1 in 14.7 million) and it is less than one in  $10^{-30}$  or roughly 1 in 1,000,000,000,000,000,000,000,000,000,000,000. Do you believe it is by chance? No, but it is as likely as any other sequence of heads and tails, it's just that there are lots of them. Indeed any sequence of 100 heads and/or tails is equally likely, there is nothing special about 100 heads, any more than one head and 99 tails in a row. There is something special in the sense that in 100 throws there are massively more sequences that have roughly fifty head and fifty tails, and only one that has only heads or only tails.

This in essence is probability. You can never be sure that any set of data did not occur by chance, but when the probability is very low (close to zero) you tend not to believe it occurred by chance. In statistics we by convention say something that could occur 5% of the time by chance is significant. P values are in the range of 0 to 1, and a 5% probability is therefore 0.05. You will often see  $p < 0.05$  which is interpreted as being significant. You will often see in reports tables where against some test result is an asterisk (\*) which usually means  $p < 0.05$  or two (\*\*) which is normally  $p < 0.01$  or three (\*\*\*) which usually means  $p < 0.001$ . Note SPSS truncates p values to three decimal places by default. That may give p values of 0.000, which is impossible (because no set of data, however weird, can never occur by chance). You should report this as  $p < 0.001$ . It shows a basic lack of statistical knowledge to state  $p = 0.000$ , and examiners or reviewers tend to dig deep once they see this.

### Example of using chi square

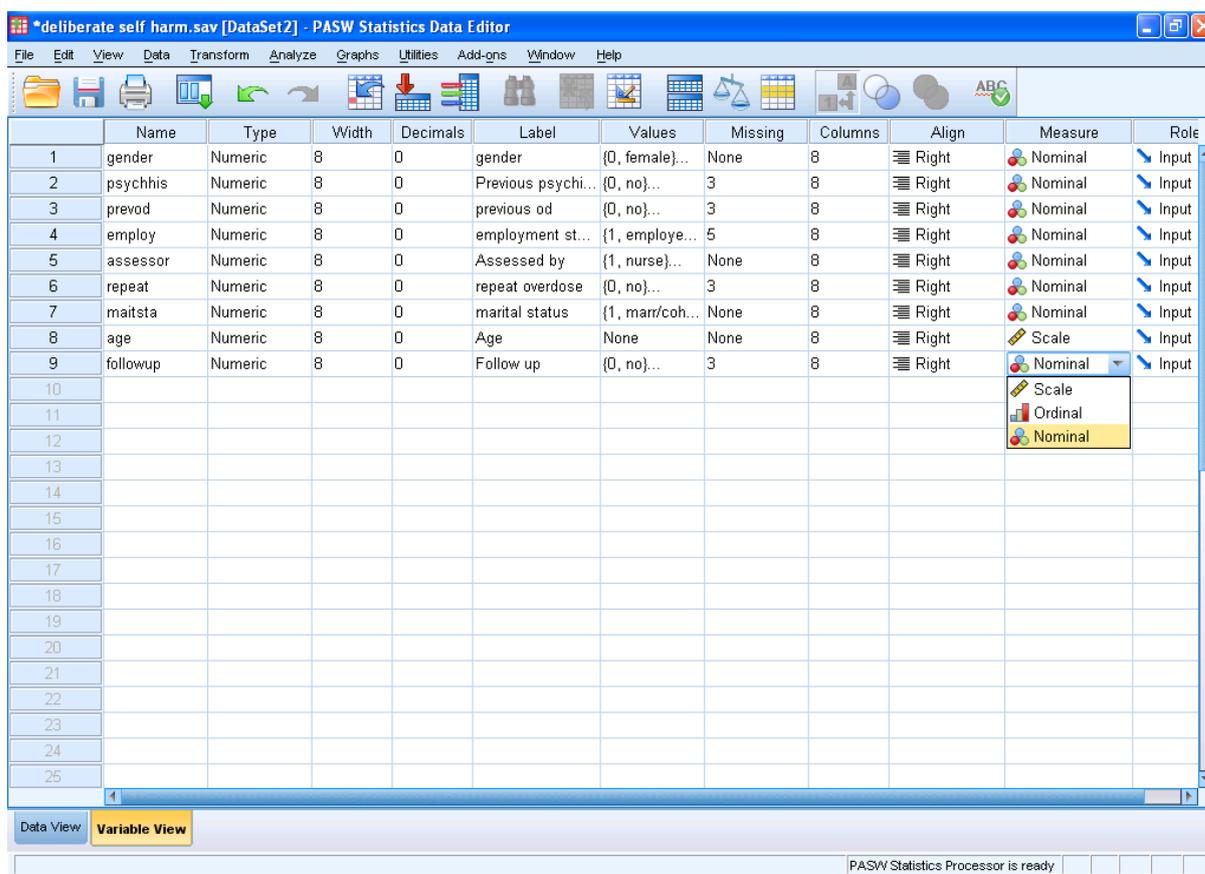
In the late 1990s Jenny Cook (one of my postgraduate students) conducted a study of self harm, it was later published (Cook and Anthony, 1999).

Open the file “*deliberate self harm.sav*” .

If you scroll down the variables you will see that they consist largely of nominal variables, most of which are binary (have two values) that are yes or no (plus a “not recorded” value). Gender is just male/female and assessor is either doctor or nurse, employ is one of four categories and marital status is also in four categories. The only “real” number is age.

Note that to the left of the variable names is an icon that for all bar *Age* is three non overlapping circles, this means the variable is nominal. For *Age* the icon is a rule, which means it is scalar. There are no ordinal variables in this dataset, but if there were it would be an icon with three bars. SPSS does not necessarily know what type of variable you have, and often a number is assumed to be scalar when you build your dataset. You need therefore to define your data types under **Measure** in the **Variable View** as in Figure 77. The reason it is important to set the variables correctly to their type is that SPSS sometimes checks the type, especially in graphs, and will not allow certain actions to be taken on variables that are not of the right type.

Figure 77: Typing variables



Jenny wanted to identify risk factors for people who had taken overdoses. The literature is replete with factors that are thought to be associated with self harm, specifically overdose and repeat overdose in particular. Our paper does summarise this work. For example gender (males more) and previous mental health problems are associated with self harm (though overdose is just one example of self harm which also includes wrist cutting etc.).

### Descriptive statistics

Let us first see if males or females are most likely to take an overdose, I have done a frequency table for gender.

**Table 8:** Gender frequency table

		gender			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	585	51.4	51.4	51.4
	male	553	48.6	48.6	100.0
Total		1138	100.0	100.0	

“I studied English for 16 years but...  
...I finally learned to speak it in just six lessons”  
Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



We see immediately that there is a near 50:50 split between males and females, which is not what the literature led us to believe. It is true there are slightly more women than men, but in any random sample you would not get typically exactly 50% male and 50% female. While there may be a statistically higher number of women (we have not tested this) it is too small to be of practical consequence. You need to ask what would you do if this was statistically significant, would you treat men and women differently in risk assessing for overdose based on a two percent difference (and there may be more women in the population at large!).

What we are interested in is whether the sample goes on to further overdoses. Jenny followed these patients up for two years, and if they repeated she changed this variable from no to yes (0 to 1 as the variable has value labels of 0=no and 1=yes). A frequency table is shown in Table 9.

**Table 9:** Repeat overdose  
**repeat overdose**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	473	41.6	47.1	47.1
	yes	531	46.7	52.9	100.0
	Total	1004	88.2	100.0	
Missing	System	134	11.8		
Total		2699	100.0		

This shows some missing data, unlike gender. It appears that Jenny may have lost patients to follow-up. They could have died, moved or refused to be interviewed. The valid percent, which is the percent of those for whom we have data, is probably more useful, showing just over half (53% rounded to the nearest percent) took a further overdose.

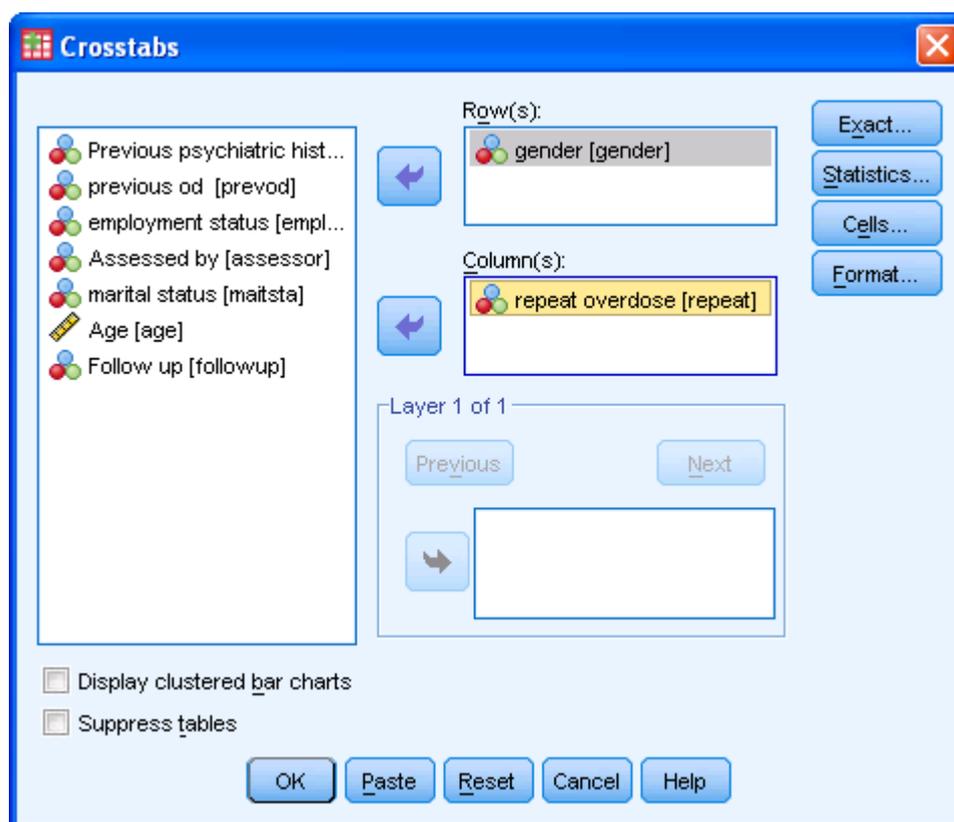
These two tables tell us a lot about the sample. They do not tell us anything about who repeats however. If we want to look at gender, we need to see both genders and whether they repeat or not. We need to cross the table of gender with that of repeat overdose. Such a table with every combination of two variables is called a cross tabulation table (crosstab in SPSS) or contingency table.

### Crosstab of gender and repeat overdose

To get a crosstab use **Analyze -> Descriptive Statistics -> Crosstabs**

You then get the dialogue box as in Figure 78. I have already added gender into row and repeat overdose into column. It makes little difference whether you add to a row or column as long as you have (at least) one variable in each. It will only affect the display of the table. If you have more values in one variable than the other, it looks better if that one is in the rows. Here both variables have two values so it makes no difference at all.

Figure 78: Adding variables in a crosstab



The output is sent to the Output window as in Table 10.

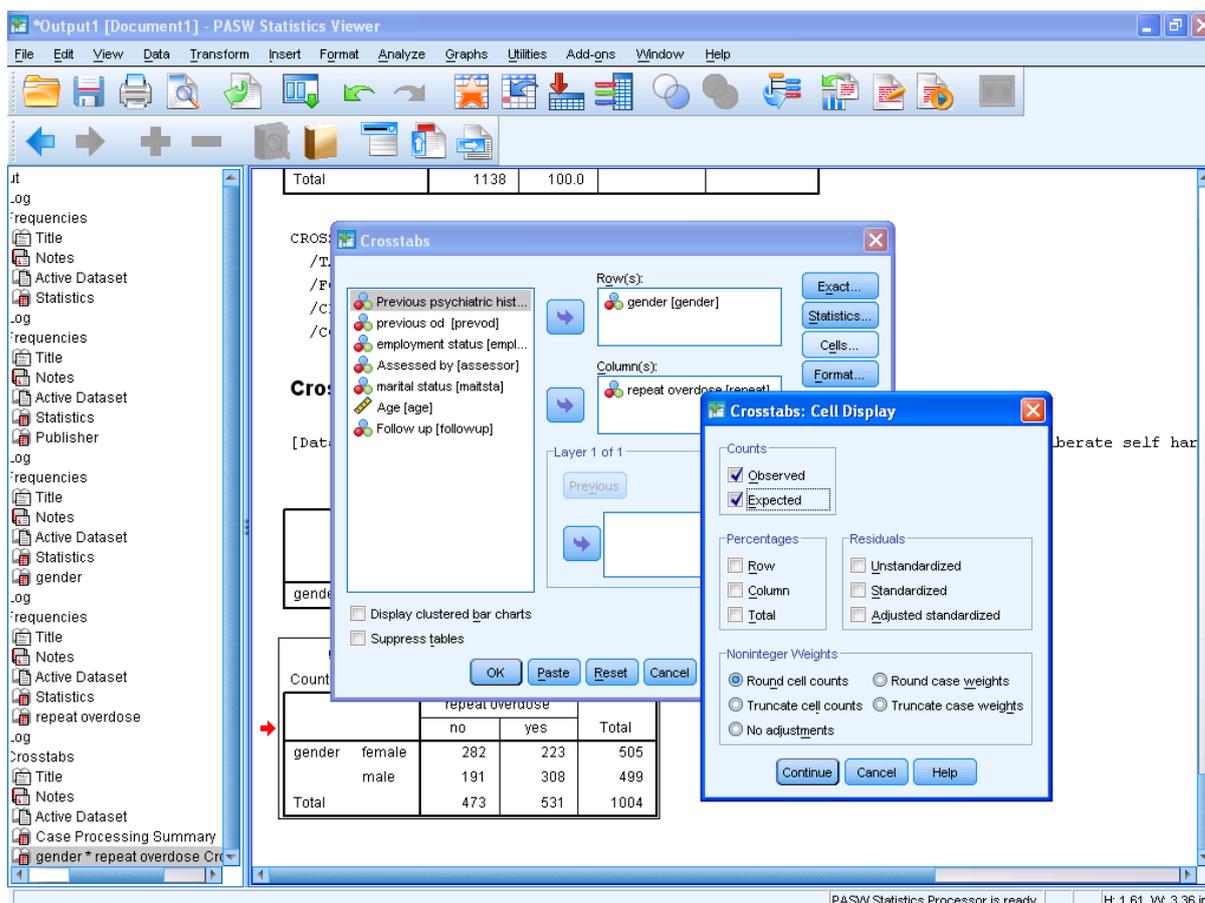
**Table 10:** Crosstab of gender against repeat overdose  
**gender \* repeat overdose Crosstabulation**

Count

		repeat overdose		Total
		no	yes	
gender	female	282	223	505
	male	191	308	499
Total		473	531	1004

This shows many more men repeat than women. Is this what we would expect if gender had nothing to do with repeat overdose? Well if there were more men in the sample then this would be expected, but there were actually fewer. At a rough guess since there were 531 (total for column 2) that repeated, and about half the sample were male, about 265 would be expected to be male, not 308. In fact the expected number would be slightly lower than 265 as there were fewer men. SPSS can compute the expected values for you. If from the dialogue box above I click on **Cells** I get a further dialogue box where I can select **Expected** (the default is just **Observed**) see Figure 79.

**Figure 79:** Getting expected values



Then we get the crosstab but with additional information as in Figure 80

**Figure 80:** Crosstab with expected frequency  
**gender \* repeat overdose Crosstabulation**

			repeat overdose		Total
			no	yes	
gender	female	Count	282	223	505
		Expected Count	237.9	267.1	505.0
	male	Count	191	308	499
		Expected Count	235.1	263.9	499.0
Total		Count	473	531	1004
		Expected Count	473.0	531.0	1004.0

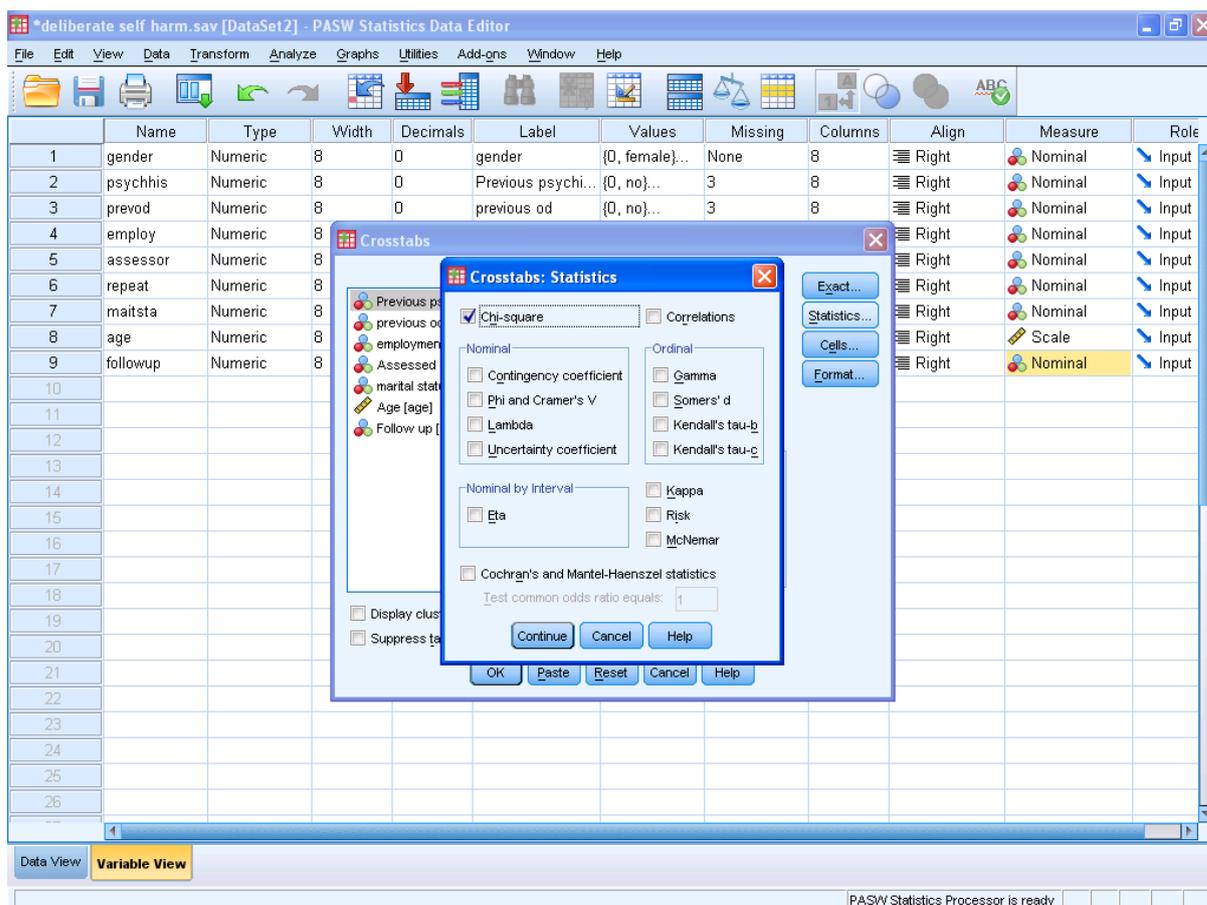
Our mental arithmetic guess was about right. We would have expected 264 (rounded up) males and 267 (rounded down) females to repeat. But could these figures be chance fluctuations? If we repeated this survey with a new 1000 cohort would it be possible that many more females repeat? Well it is always possible. If all men repeated and no woman did, it is *possible* that it was a biased sample and not representative of the population from which it was drawn. But is this *likely*? It does depend on the sample size. Take a ridiculous sample of one, one man, who does repeat. We can say 100% of men repeat, true but trivial. Now a sample of two, one man and one woman. The man repeats. Is this a big deal? No, clearly not. Now a sample of ten, 5 men and 5 women. All the men and none of the women repeat. Interesting, suggestive, but not definitive. A new sample of 1000, 500 of each gender. All the men repeat, no woman does. Could this be a chance finding?

I have been asking this of students for about fifteen years. They all say no. Why is this? Because of the size of the sample. Logically they are wrong, but only in the same way as you are wrong to tell me I will not win the lotto this week, and the next week, and every week for the next year. Yet I could (if I buy a ticket). I have a 1 in 14.7 million chance of getting a win on the UK lotto. Someone does most weeks, so it is not impossible to be me. However planning my finances on this basis would be foolish. If I win this week I still have the same chance next week. It could happen. I think it never has though. And that is two weeks, not 52. However unlikely it may be though it is possible, and winning the lottery twice has a computable value (one chance in 2,160,900,000,000 or one person in about 300-400 worlds of the same population as Earth, but **not** zero).

So how likely is it that the figures we see here are by chance? There is a statistical test called chi square that works this out for you. It looks at the difference between the observed and expected values, and give a high value if these are far apart, and a low one if they are close together. Then it looks at the sample size, and adjusts the value down if the sample is small and up if it is high. Actually it does not do this exactly, but conceptually it does just this.

We can get the chi square value by clicking on **Statistics** from the dialogue box and ticking **Chi square** as in Figure 81.

Figure 81: Adding in chi square



We then get Table 11.

Table 11: Chi square  
Chi-Square Tests

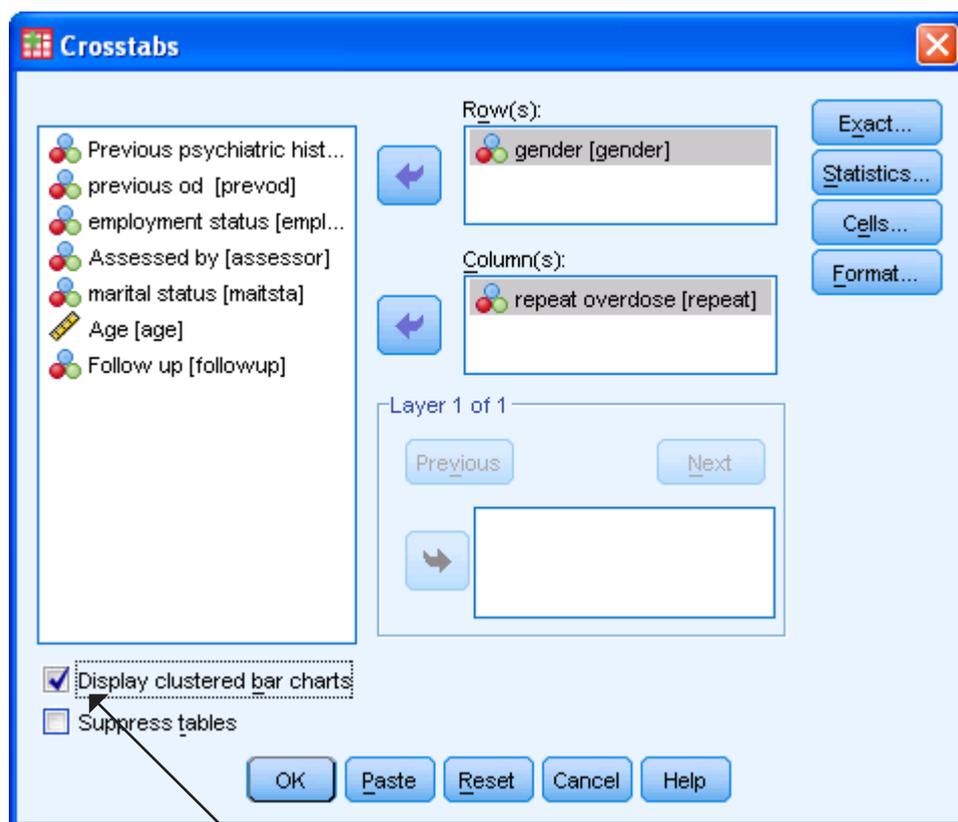
	Value	df	Asymp. Sig (2-sided)	Exact. Sig. (2-sided)	Exact. Sig. (1-sided)
Pearson Chi-Square	31.079 <sup>a</sup>	1	.000		
Continuity Correction <sup>b</sup>	30.378	1	.000		
Likelihood Ratio	31.246	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	31.048	1	.000		
N of Valid Cases	1004				

<sup>a</sup>. 0 cells (.0%) have expected count less than 5. The minimum expected count is 235.09.

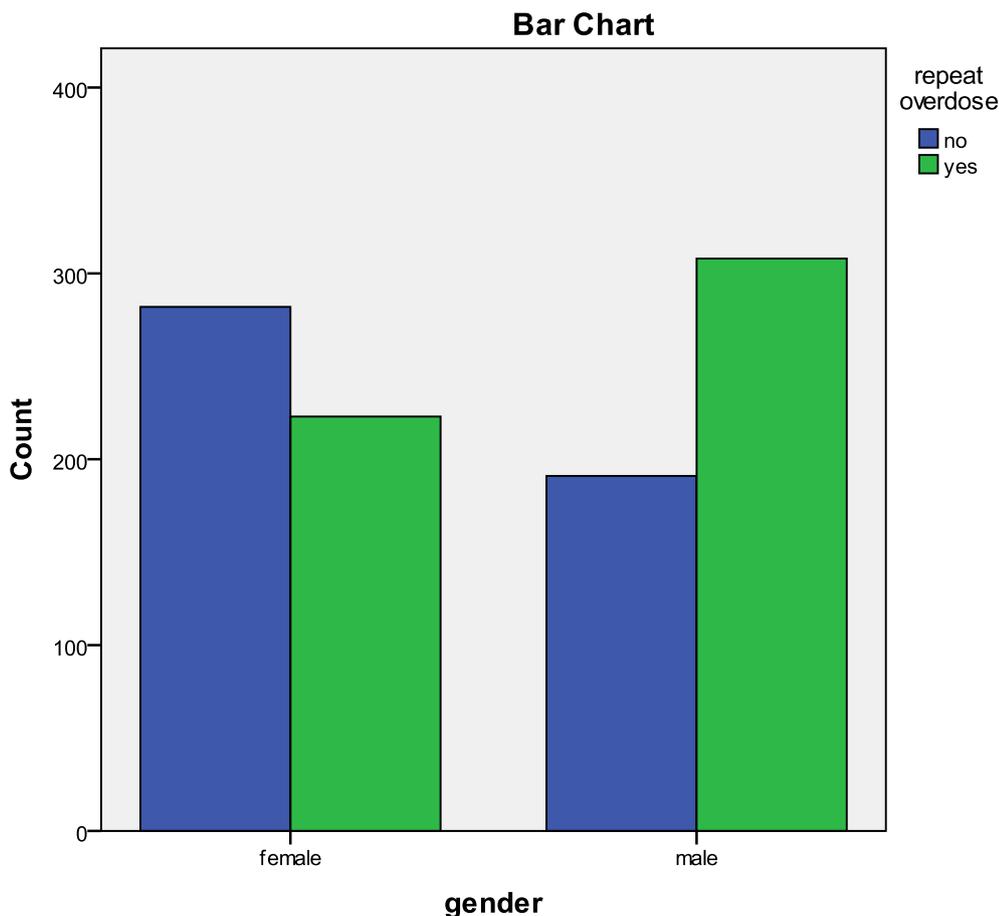
<sup>b</sup>. Computed only for a 2x2 table

What this tells us is that all cells have expected frequencies over five. This is important as it is a condition of chi square that this be so, or at least that 80% of cells do. If this condition is not met then the default chi square value (Pearson Chi-Square, first line in table, not to be confused with Pearson correlation co-efficient which will be covered in a later chapter) should not be used, though the more conservative Continuity Correction can be used (second line in table). Fisher’s exact test is also fine if this condition does not hold. Fischer’s test works out the possibility of each and every possible outcome and adds together all the ones under a certain frequency. It is computationally expensive (takes a lot of computer power) and very difficult to do by hand, hence the continuity correction was often preferred before the days of computers. Pragmatically though I would say if you need to use these variants of chi square you may have too small a sample to be able to say anything very much. Pearson’s chi square give a p value (called Asymp. Sig in the table) of 0.000, which we report as  $p < 0.001$ . We interpret this as meaning there is a significant difference between men and women with respect to repeat self harm – because the p value is less than 0.05. A common error to think high p values are significant, it is low values that are significant. You can get SPSS to give you a clustered bar chart while it does the chi square test by clicking on the relevant item in the crosstabs dialogue box, see Figure 82, which is easier to interpret for the non statistically minded person (almost everyone, and probably your manager). This is useful for showing to (say) the clinical manager , and is seen in Figure 83 which clearly shows the impact of gender.

Figure 82: Getting a clustered bar chart



**Figure 83:** Clustered barchart of gender and repeat overdose



**Crosstab of assessor and repeat overdose**

Patients are seen by either a doctor or nurse. Who would you rather see? It should depend on who gets the better outcome.

Suppose we want to test if patients seen by doctors fare better than those seen by nurse (with respect to repeat self harm). Interestingly when I ask students to hypothesise they are split. Some would prefer to see a doctor on the grounds they are trained in diagnosis and would be more accurate therefore in predicting risk and giving appropriate treatment, and others would prefer to be seen by a nurse as they perceive them to be kinder and have more time to talk (some may disagree) .

Doing the same test with the variable *assessor* gives the following result (see Table 12).

**Table 12:** Chi square of repeat overdose and assessor  
**Assessed by \* repeat overdose Crosstabulation**

			repeat overdose		Total
			no	yes	
Assessed by	nurse	Count	222	254	476
		Expected Count	224.3	251.7	476.0
	doctor	Count	251	277	528
		Expected Count	248.7	279.3	528.0
Total		Count	473	531	1004
		Expected Count	473.0	531.0	1004.0

**Chi-Square Tests**

	Value	df	Asymp. Sig (2-sided)	Exact. Sig. (2-sided)	Exact. Sig. (1-sided)
Pearson Chi-Square	.081 <sup>a</sup>	1	.776		
Continuity Correction <sup>b</sup>	.049	1	.825		
Likelihood Ratio	.081	1	.776		
Fisher’s Exact Test				.800	.412
Linear-by-Linear Association	.081	1	.776		
N of Valid Cases	1004				

<sup>a</sup>. 0 cells (.0%) have expected count less than 5. The minimum expected count is 224.25.

<sup>b</sup>. Computed only for a 2x2 table

In the past four years we have drilled  
**89,000 km**  
 That’s more than **twice** around the world.

**Who are we?**  
 We are the world’s largest oilfield services company<sup>1</sup>. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

**Who are we looking for?**  
 Every year, we need thousands of graduates to begin dynamic careers in the following domains:

- Engineering, Research and Operations
- Geoscience and Petrotechnical
- Commercial and Business

**What will you be?**

**Schlumberger**

careers.slb.com

<sup>1</sup>Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.



This shows a p value of 0.77 (to two decimal places) which is **not** significant. This shows that these data could have occurred by chance 77% of the time. Thus the (real) difference between nurses and doctors means little, and it is highly likely to have occurred by chance. This is the opposite case compared to the former test. The high p value means that it is not significant. So it does not appear to matter whether you are seen by a doctor or nurse, at least with respect to repeat harm by overdose.

### Crosstab of follow-up and repeat overdose

Does follow-up of a patient help? Let's test this with chi square. If I put variable *followup* into a similar chi square analysis as above I get

**Table 13:** Follow-up and repeat overdose  
**Follow up \* repeat overdose Crosstabulation**

			repeat overdose		Total
			no	yes	
Follow up	no	Count	190	81	271
		Expected Count	127.8	143.2	271.0
	yes	Count	283	449	732
		Expected Count	345.2	386.8	732.0
Total		Count	473	530	1003
		Expected Count	473.0	530.0	1003.0

### Chi-Square Tests

	Value	df	Asymp. Sig (2-sided)	Exact. Sig. (2-sided)	Exact. Sig. (1-sided)
Pearson Chi-Square	78.500 <sup>a</sup>	1	.000		
Continuity Correction <sup>b</sup>	77.243	1	.000		
Likelihood Ratio	79.841	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	78.422	1	.000		
N of Valid Cases	1003				

<sup>a</sup>. 0 cells (.0%) have expected count less than 5. The minimum expected count is 127.80.

<sup>b</sup>. Computed only for a 2x2 table

This shows a highly significant result ( $p < 0.001$ , I hope you didn't think  $p = 0.000!$ ). That's great isn't it as it shows follow-up works? Or does it? If you look at the patients who were followed up 387 (rounding up) would be expected to repeat overdose, and only – oh dear – 449 did. This is a disaster, follow-up is actually making things worse! When I ask my students to explain this I get answers like “the patient found the follow-up traumatic, they must have found it difficult to talk about their experiences” etc. If we spoke to the managers and showed them the data maybe they would immediately stop this dangerous practice of seeing patients again.

However every so often a group picks out the obvious. Who do we follow up? The patient who had a clear single reason for overdosing (their partner left them maybe) but they have now come to terms with that and have no particular risk factors that concern us? Possibly not. The unemployed homeless younger male with mental health problems and a history of overdoses? Maybe that one. So it could be (I am not proving here it is of course) that nurses and doctors pick the high risk cases for follow-up. Maybe things would be even worse if they did not (though I have no evidence for this). What we have here is an artefact. It shows the need for good clinical trials to identify what works. Now if we had randomly allocated patients into two groups, one with follow-up, the other not, and got the same results – well that would be different. In our case there was no randomisation. The lesson is you need to be very careful interpreting inferential testing where there is no randomisation. It is possible (epidemiologists do it as a profession) but certainly non-trivial.

## Exercise

Males seem to repeat overdose more than females, according to the chi square analysis. However maybe males have more mental health problems, and it is this, rather than gender, that may be the real issue. Decide how to test this. You will need to create a cross-tabulation table and employ chi square to check this out.

## References and background reading

Websites include:-

[en.wikipedia.org/wiki/Contingency\\_table](http://en.wikipedia.org/wiki/Contingency_table)

[www.physics.csbsju.edu/stats/contingency.html](http://www.physics.csbsju.edu/stats/contingency.html)

[www.jerrydallal.com/LHSP/ctab.htm](http://www.jerrydallal.com/LHSP/ctab.htm)

If you want to explore contingency tables in more detail then Everitt (1992) is a good text, though this does not employ SPSS.

## References

Cook, J. and D. M. Anthony (1999). "Repetition of self harm." *Clinical Effectiveness in Nursing*, 3(4): 181-184.

Everitt, B. (1992). *The analysis of contingency tables*. London. , Chapman & Hall. .

# 7 Differences between two groups

## Key points

- There are tests that assume a normal distribution and those that do not
- For two groups Student's t test for independent groups and Mann Whitney are the appropriate tests

At the end of this unit you should be able to:

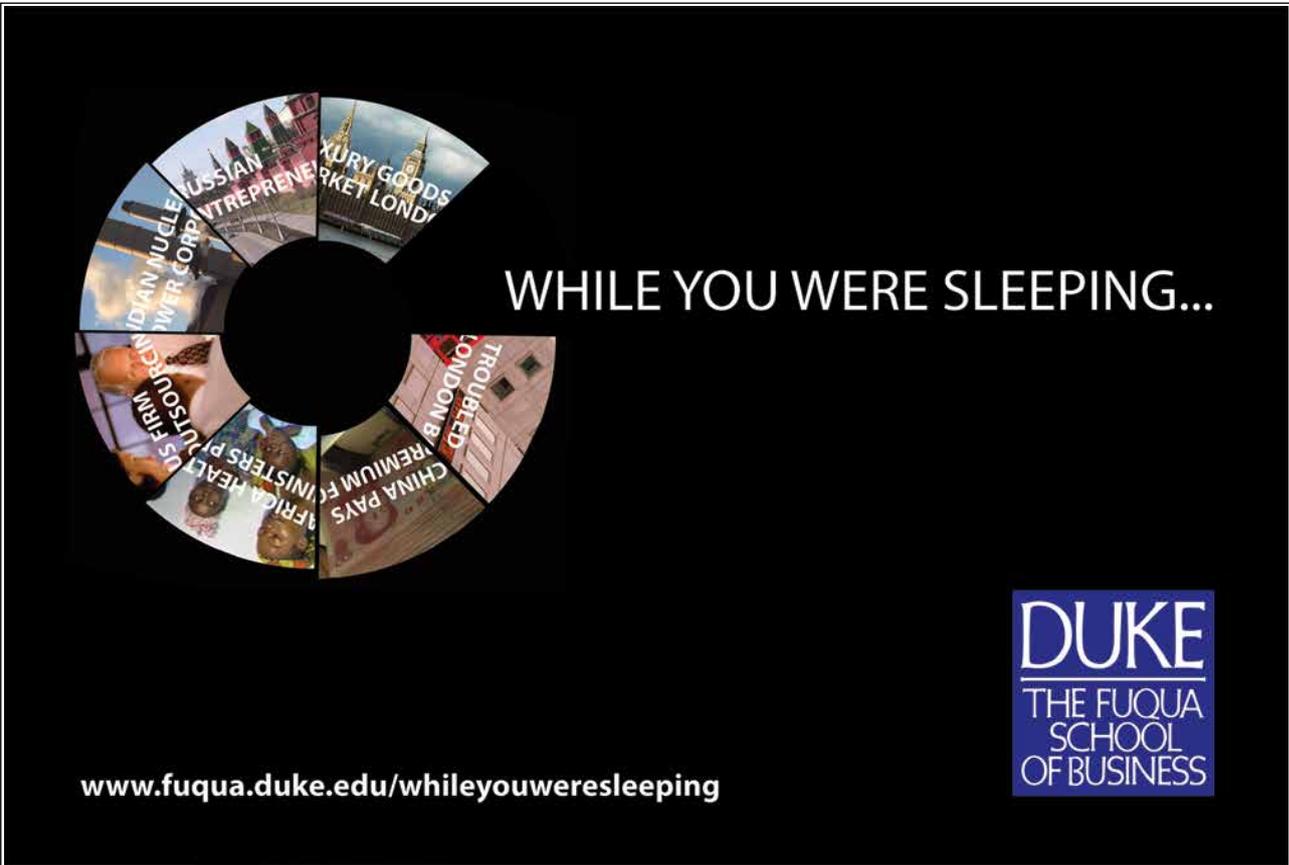
- Decide which test to use for two groups
- Compute the test and interpret the result

## Introduction

In chapter 6 we looked at differences in frequencies for nominal data. Let us consider the online course survey again. We could use chi square on some of these data. An example is:-

Is there a difference between males and females with respect to course they are on? (There are two courses, degree and diploma, so are we recruiting more or less women to the degree programme)?

However what if your data are not nominal? For example *Age*. We *could* use chi square, but this is a very weak test for this type of data. Consider the following:-



WHILE YOU WERE SLEEPING...

[www.fuqua.duke.edu/whileyouweresleeping](http://www.fuqua.duke.edu/whileyouweresleeping)

**DUKE**  
THE FUQUA  
SCHOOL  
OF BUSINESS

**Table 14:** chi square of age against gender

			gender		Total
			Female	Male	
Age	17	Count	1	0	1
		Expected Count	.9	.1	1.0
	18	Count	19	3	22
		Expected Count	19.5	2.5	22.0
	19	Count	24	2	26
		Expected Count	23.0	3.0	26.0
	20	Count	6	0	6
		Expected Count	5.3	.7	6.0
	21	Count	3	0	3
		Expected Count	2.7	.3	3.0
	22	Count	6	0	6
		Expected Count	5.3	.7	6.0
	23	Count	7	0	7
		Expected Count	6.2	.8	7.0
	24	Count	4	1	5
		Expected Count	4.4	.6	5.0
	25	Count	1	1	2
		Expected Count	1.8	.2	2.0
	26	Count	1	2	3
		Expected Count	2.7	.3	3.0
	27	Count	4	0	4
		Expected Count	3.5	.5	4.0
	28	Count	2	0	2
		Expected Count	1.8	.2	2.0
	29	Count	2	1	3
		Expected Count	2.7	.3	3.0
	30	Count	3	0	3
		Expected Count	2.7	.3	3.0
	31	Count	3	0	3
		Expected Count	2.7	.3	3.0
	32	Count	4	1	5
		Expected Count	4.4	.6	5.0
	33	Count	4	0	4
		Expected Count	3.5	.5	4.0
	34	Count	3	3	6
		Expected Count	5.3	.7	6.0
	35	Count	4	0	4
		Expected Count	3.5	.5	4.0
	36	Count	4	0	4
		Expected Count	3.5	.5	4.0

	37	Count	3	1	4
		Expected Count	3.5	.5	4.0
	38	Count	1	0	1
		Expected Count	.9	.1	1.0
	39	Count	2	1	3
		Expected Count	2.7	.3	3.0
	40	Count	1	0	1
		Expected Count	.9	.1	1.0
	41	Count	2	0	2
		Expected Count	1.8	.2	2.0
	43	Count	2	0	2
		Expected Count	1.8	.2	2.0
	44	Count	1	0	1
		Expected Count	.9	.1	1.0
	45	Count	1	0	1
		Expected Count	.9	.1	1.0
	46	Count	4	0	4
		Expected Count	3.5	.5	4.0
	52	Count	1	0	1
		Expected Count	.9	.1	1.0
Total		Count	123	16	139
		Expected Count	123.0	16.0	139.0

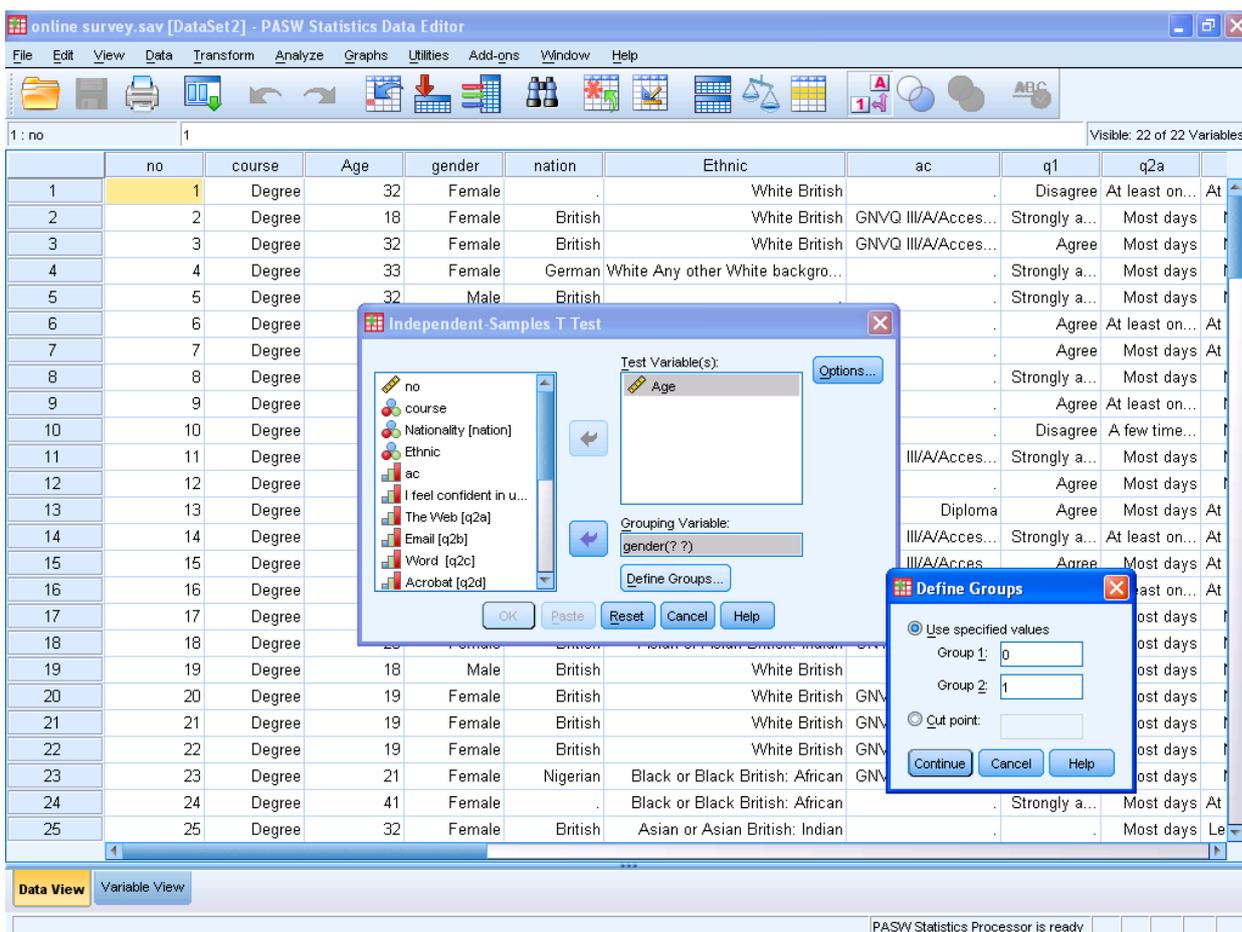
There are several problems here. Firstly chi square assumes expected frequencies in each cell are at least 5, and the test becomes very unreliable if more than 20% of cells have an expected frequency of less than 5, and expected frequencies of less than one are problematic always. In this case 90% of cells have an expected frequency of less than 5, and one is as low as 0.12.

But this is not the main problem. Age is an interval/ratio variable, and chi square does not use this extra information. It assumes each age is different (true) but makes no distinction between them. Chi square for example does not make use of the fact that an age of 20 is more than one of 19 and less than 21. It views these as simply three different categories of age, rather as it would if these were three ethnic groups.

We need a test that makes use of the fact that the data are genuinely numeric and not nominal. One such test is Student's t test for independent groups (there is another Student's t test for paired variables, the paired Student's t test) which tests whether the mean values of two groups are significantly different.

Using the datafile *online survey* you can obtain the Student's t test for independent groups by **Analyze -> Compare Means -> Independent Samples T Test**. Then a dialogue box will appear, see Figure 84.

Figure 84: Dialogue box



You need to put in the variable you are testing (here *age*), the grouping variable (here *gender*) and the values for the two groups using **Define** (here 0 and 1). So we are using variable *gender* to split into two groups (males and females) and the test variable, i.e. the one we are testing to see if it is different between the groups, is *age*. If you complete this you get the output shown in Table 15.

**Table 15:** Student's t test for independent groups

<b>Group Statistics</b>										
	gender	N	Mean	Std. Deviation	Std. Error Mean					
Age	Female	123	26.41	8.758	.790					
	Male	16	27.00	7.330	1.833					

<b>Independent Samples Test</b>										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Age	Equal variances assumed	1.275	.261	-.256	137	.799	-.585	2.289	-5.112	3.941
	Equal variances not assumed			-.293	20.999	.772	-.585	1.995	-4.735	3.564

Excellent Economics and Business programmes at:



**university of  
 groningen**





**“The perfect start  
 of a successful,  
 international career.”**

[www.rug.nl/feb/education](http://www.rug.nl/feb/education)

**CLICK HERE**  
 to discover why both socially and academically the University of Groningen is one of the best places for a student to be

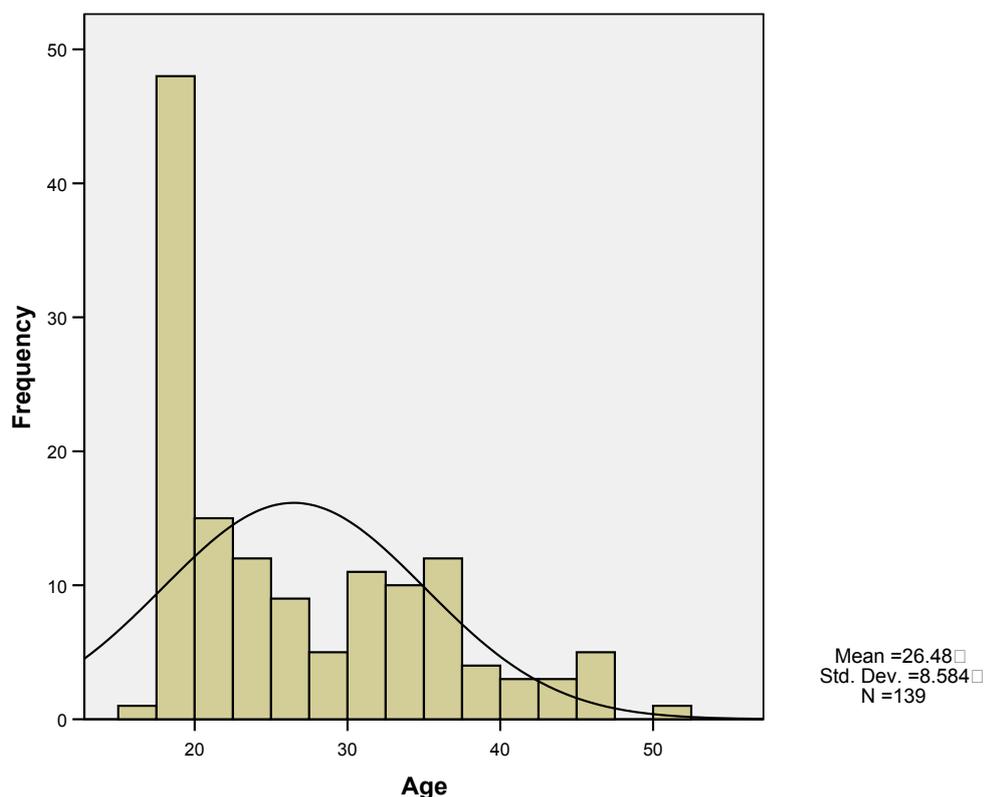


The first part tells you that males are a little older (27.0) compared to females (26.4). But is this significant? The second part tells you this. However this is a complicated table and needs careful examination. One of the requirements of Student's t test for independent groups is that the variance in each group is similar. So a test is done (Levene's test) to check this. The significance here for Levene's test is 0.261, which is not significant. This means the difference in variances between the groups (male and female) is probably caused by chance, so there is no meaningful difference. A really common error is for students to confuse the significance given for Levene (a pre-test) for Student's t test.

This means we can use Student's t test for independent groups. We use the first line of the table (Equal variances assumed). The next bit of the table gives us the significance of the t test, here 0.799. This means that there is no significant difference between the groups, i.e. males and females are similar in terms of age. SPSS allows (makes an allowance for) non-equal variances, in which case you should use the second line of the table (Equal variances not assumed).

There is a further problem though, Student's t test for independent groups assumes the test variable data are normally distributed (it is a parametric test, parametric tests assume normal distribution). But is this true? Look at Figure 85.

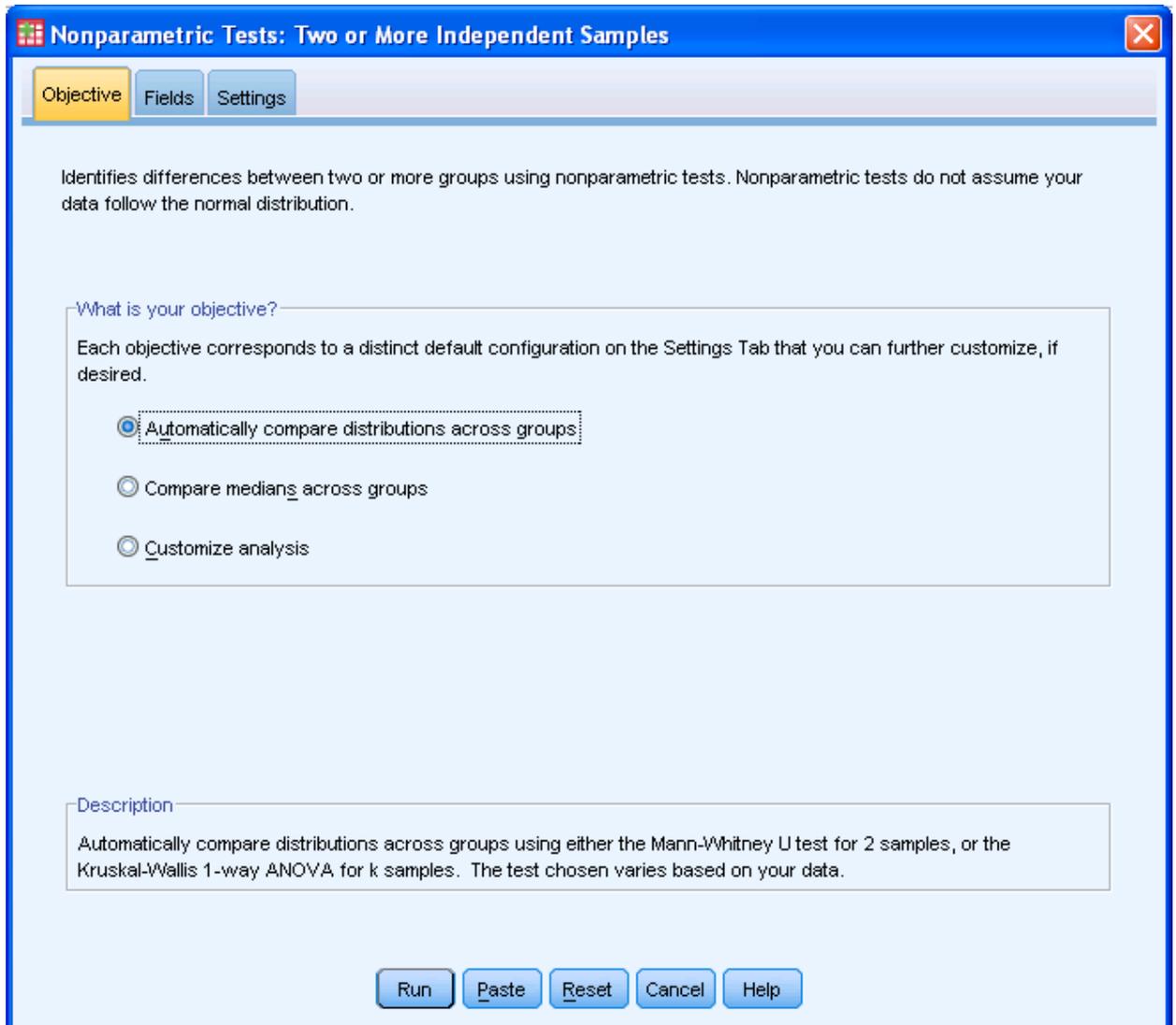
**Figure 85:** Histogram of age

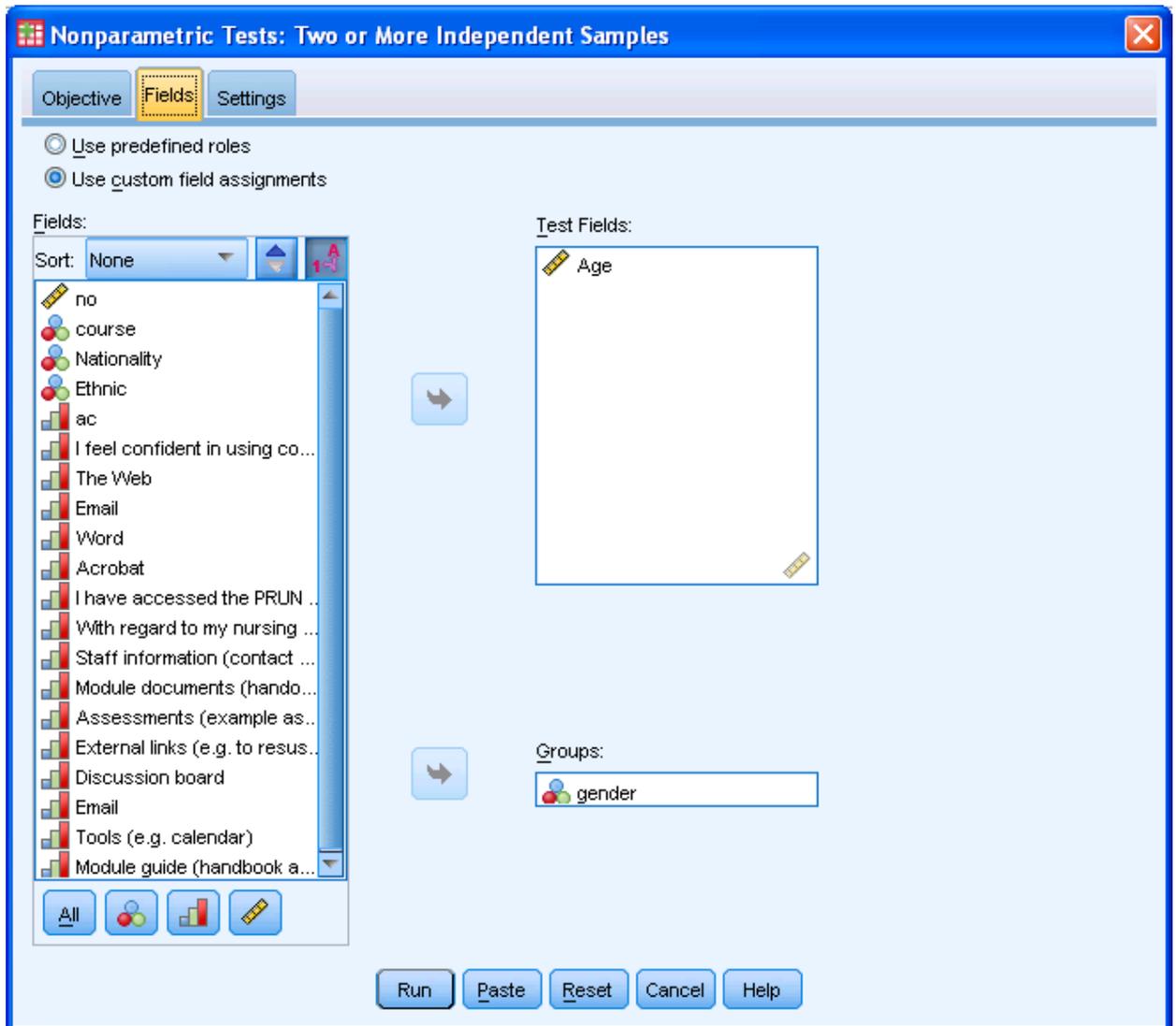


This shows age to be skewed to younger ages. This is because most students are immediately post school, hence a bulge at about 20, but there are many older students across a wide range.

Since the data are not normally distributed we need a test that does not assume normal distribution. These tests are called non-parametric tests. A common one for testing two groups is Mann Whitney.

Mann Whitney is obtained as in a similar manner to Student's t test, but it comes under non-parametric tests. Use **Analyze** -> **Nonparametric Tests** -> **Independent Samples** then you are given a dialogue box asking what of three options you want, it is the first, compare distributions. You then click on the **Fields** and put *Age* into **Test Fields** and *Gender* into **Groups** and then click on **Run**.





The output is shown in Table 16. This shows while the males are older (as they have a higher rank) but this is not significant (see significance called here *Sig*) and is 0.67.

**Table 16:** Mann Whitney test

**Hypothesis Test Summary**

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Age is the same across categories of gender.	Independent-Samples Mann-Whitney U Test	.676	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

The Likert (attitudinal) scores scores for (e.g.) confidence in computer use can also be tested as these are ordinal data, which Mann Whitney can test. Bizarrely SPSS tells me when I try this that I cannot put ordinal variables into the non-parametric tests for independent groups. I feel sure this is in error, however the workaround is to make this variable a scale variable. Again we see there are no significant differences between the two groups, here course types, see Table 17.

**Table 17:** Mann Whitney - confidence tested by gender

**Hypothesis Test Summary**

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of I feel confident in using computers is the same across categories of gender.	Independent-Samples Mann-Whitney U Test	.093	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

**More than two groups**

There is an extension of Student’s t test to more than two groups called one-way analysis of variance. This can compare any number of groups.

There is an equivalent to Mann Whitney for more than two groups, this is Kruskal Wallis, this is covered in the next chapter.

**Exercise**

Consider whether diploma or degree students were accessing the Blackboard online course site more frequently. Identify the most appropriate test, run in SPSS and interpret the results

# 8 Differences between more than two groups

## Key points

- One way ANOVA and Kruskal Wallis are the appropriate tests for more than two groups

At the end of this chapter you should be able to:

- Decide which test to use
- Compute and interpret the result

## Introduction

In the last chapter we looked at comparing two groups. Here we extend the case to more than two groups, in fact any number of groups more than two.

## Difference between many groups

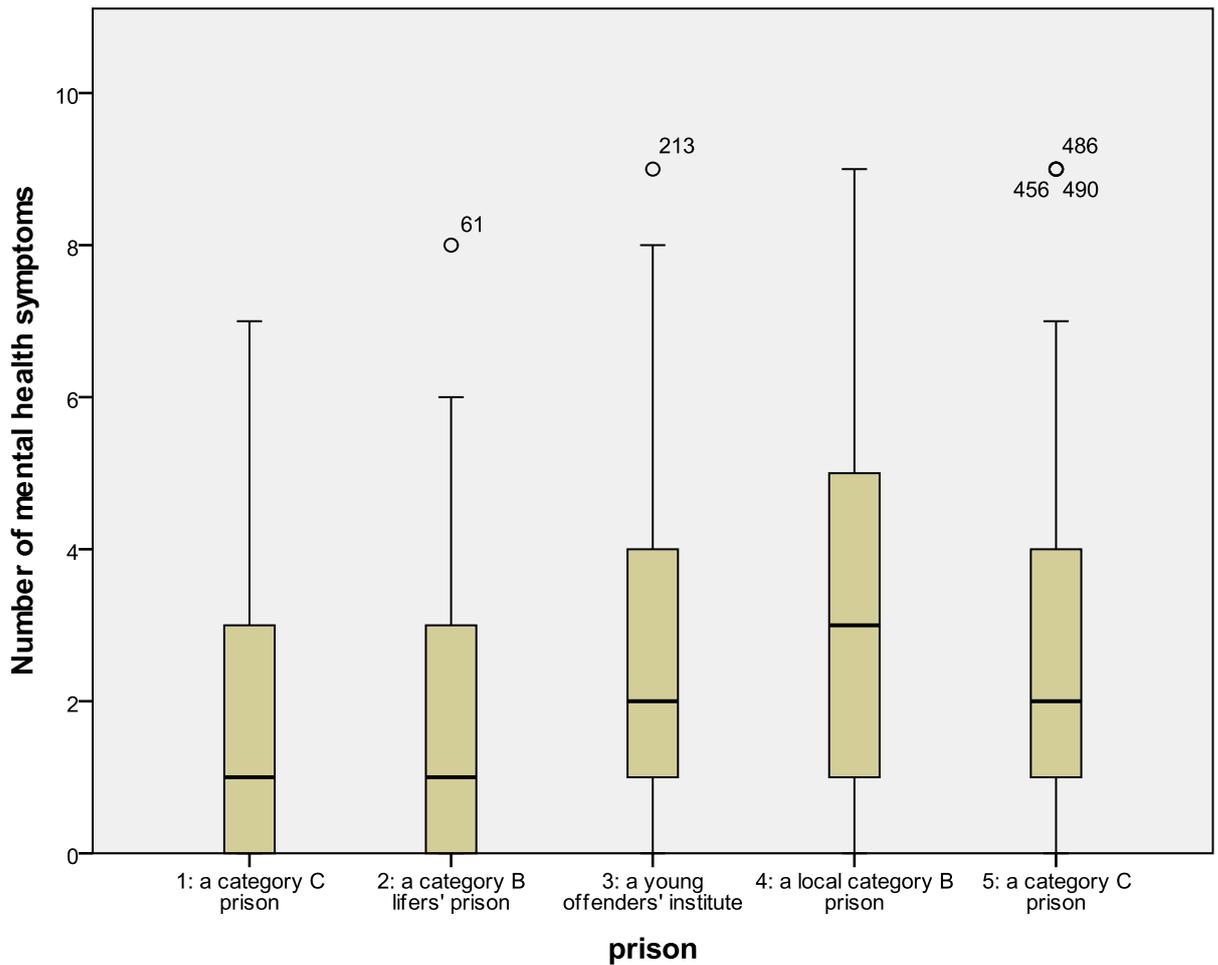
Consider the datafile *mental health in prison.sav*. This has for a number of prisons the responses of prisoners with respect to mental health symptoms. Prisoners were asked about nine symptoms, and a count was made for each prisoner of all the symptoms they had, up to a maximum of nine. There are five prisons. A frequency table is shown in Table 18.

**Table 18:** Prisons

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A a category C prison	50	10.1	10.1	10.1
	B a category B lifers' prison	33	6.7	6.7	16.8
	C a young offenders' institute	281	56.9	56.9	73.7
	D a local category B prison	60	12.1	12.1	85.8
	E A category C prison	70	14.2	14.2	100.0
	Total	494	100.0	100.0	

We would like to see if there is a difference among the five prisons. Figure 86 shows there appears to be more problems in D and fewer in A and B. But are these differences significant? There are five prisons so we cannot use Student's t test or Mann Whitney.

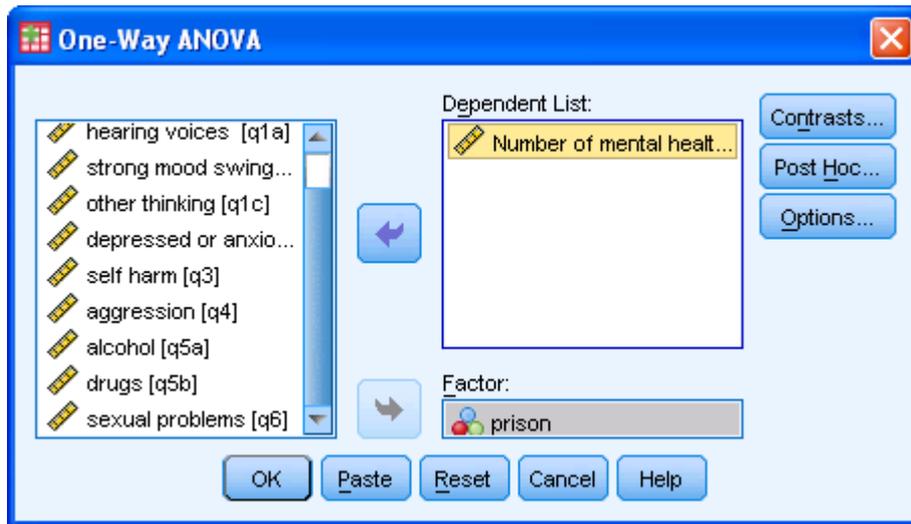
**Figure 86:** Box plot of mental health problems by prison



One way ANOVA (there are other ANOVA tests such as repeated measures ANOVA) may be the relevant test, and this is obtainable as **Analyze -> Compare Means -> One-Way ANOVA**. One way ANOVA is the extension of Student's t test to more than 2 groups. In fact if you do one way ANOVA on two groups it will give identical results as Student's t test for independent groups.

The dialogue box that comes up can be filled in as in Figure 87 and this will give an output seen in Table 19.

Figure 87: Dialogue box for ANOVA



.....Alcatel-Lucent 

[www.alcatel-lucent.com/careers](http://www.alcatel-lucent.com/careers)



What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



**Table 19:** ANOVA output

**ANOVA**

Number of mental health symptom

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	71.942	4	17.985	3.400	.009
Within Groups	2581.308	488	5.290		
Total	2653.249	492			

The relevant thing here is the between groups significance is 0.009.

**Exercise**

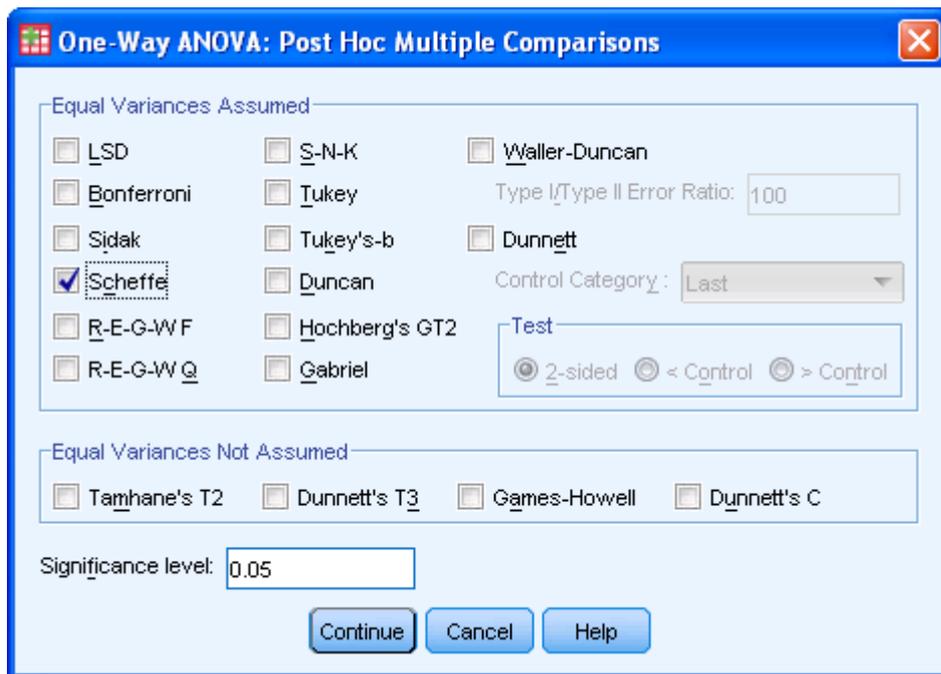
What does it mean to have a  $p=0.009$  for the between groups?

**Which groups are different?**

With two groups, if there is a significant difference then it simply remains to see which group is higher or lower with respect to the test variable.

However with more than two groups, if there is a significant difference, where does it lie? In this case for example it may be that prisons 1 and 2 are not significantly different, but are all different from prisons 3, 4 and 5. What is needed is a post-hoc test. I.e. after finding out whether there is a significant difference, the post -hoc test identifies where the differences lie. In the **One-Way ANOVA** dialogue box if you click on **Post-Hoc** (Multiple Comparisons) you will see there are several such tests, a common one is Sheffe. This is obtained as in Figure 88.

Figure 88: Sheffe test



The output is seen in Table 20.

Table 20: Sheffe output  
Multiple Comparisons

Number of mental health symptoms Scheffe

(I) prison	(J) prison	Mean-Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1: a category C prison	2: a category B lifers' prison	-.028	.516	1.000	-1.62	1.57
	3: a young offenders' institute	-.794	.353	.284	-1.89	.30
	4: a local category B prison	-1.407*	.440	.038	-2.77	-.04
	5: a category C prison	-.811	.426	.459	-2.13	.51
	2: a category B lifers' prison	.028	.516	1.000	-1.57	1.62
2: a category B lifers' prison	1: a category C prison	.028	.516	1.000	-1.57	1.62
	2: a category B lifers' prison	-.766	.423	.514	-2.07	.54
	3: a young offenders' institute	-1.379	.498	.107	-2.92	.16
	5: a category C prison	-.784	.486	.627	-2.29	.72
3: a young offenders' institute	1: a category C prison	.794	.353	.284	-.30	1.89
	2: a category B lifers' prison	.766	.423	.514	-.54	2.07
	4: a local category B prison	-.613	.327	.477	-1.62	.40
	5: a category C prison	-.018	.307	1.000	-.97	.93
	1: a category C prison	1.407*	.440	.038	.04	2.77
4: a local category B prison	2: a category B lifers' prison	1.379	.498	.107	-.16	2.92
	4: a local category B prison	.613	.327	.477	-.40	1.62
	5: a category C prison	.595	.405	.706	-.66	1.85
	1: a category C prison	.811	.426	.459	-.51	2.13
5: a category C prison	2: a category B lifers' prison	.784	.486	.627	-.72	2.29
	3: a young offenders' institute	.018	.307	1.000	-.93	.97
	4: a local category B prison	-.595	.405	.706	-1.85	.66

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

This shows (see row 1) that prison 1 is not significantly different from prison 3 as although the mean difference is 0.794, it is not significant ( $p=0.284$ ).

### Exercise

Go through the remainder of the table and decide which prisons are the same (not significantly different from each other) and which are significantly different.

However are the data (mental health symptoms) normally distributed? If not then we have a problem as ANOVA is a parametric test.

### Exercise

Determine if the data for number of mental health symptoms are normally distributed.

### Non-parametric testing for more than two groups

If data are not normally distributed we can use the non-parametric test, which for more than two groups is Kruskal Wallis. If you use the same method as in the last chapter for Mann Whitney (which Kruskal Wallis is the equivalent test for more than two groups), i.e. Analyse -> Nonparametric Tests -> Independent Samples using the inputs in Figure 89.

**Maastricht University** *Leading in Learning!*

**Join the best at the Maastricht University School of Business and Economics!**

**Top master's programmes**

- 33<sup>rd</sup> place Financial Times worldwide ranking: MSc International Business
- 1<sup>st</sup> place: MSc International Business
- 1<sup>st</sup> place: MSc Financial Economics
- 2<sup>nd</sup> place: MSc Management of Learning
- 2<sup>nd</sup> place: MSc Economics
- 2<sup>nd</sup> place: MSc Econometrics and Operations Research
- 2<sup>nd</sup> place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

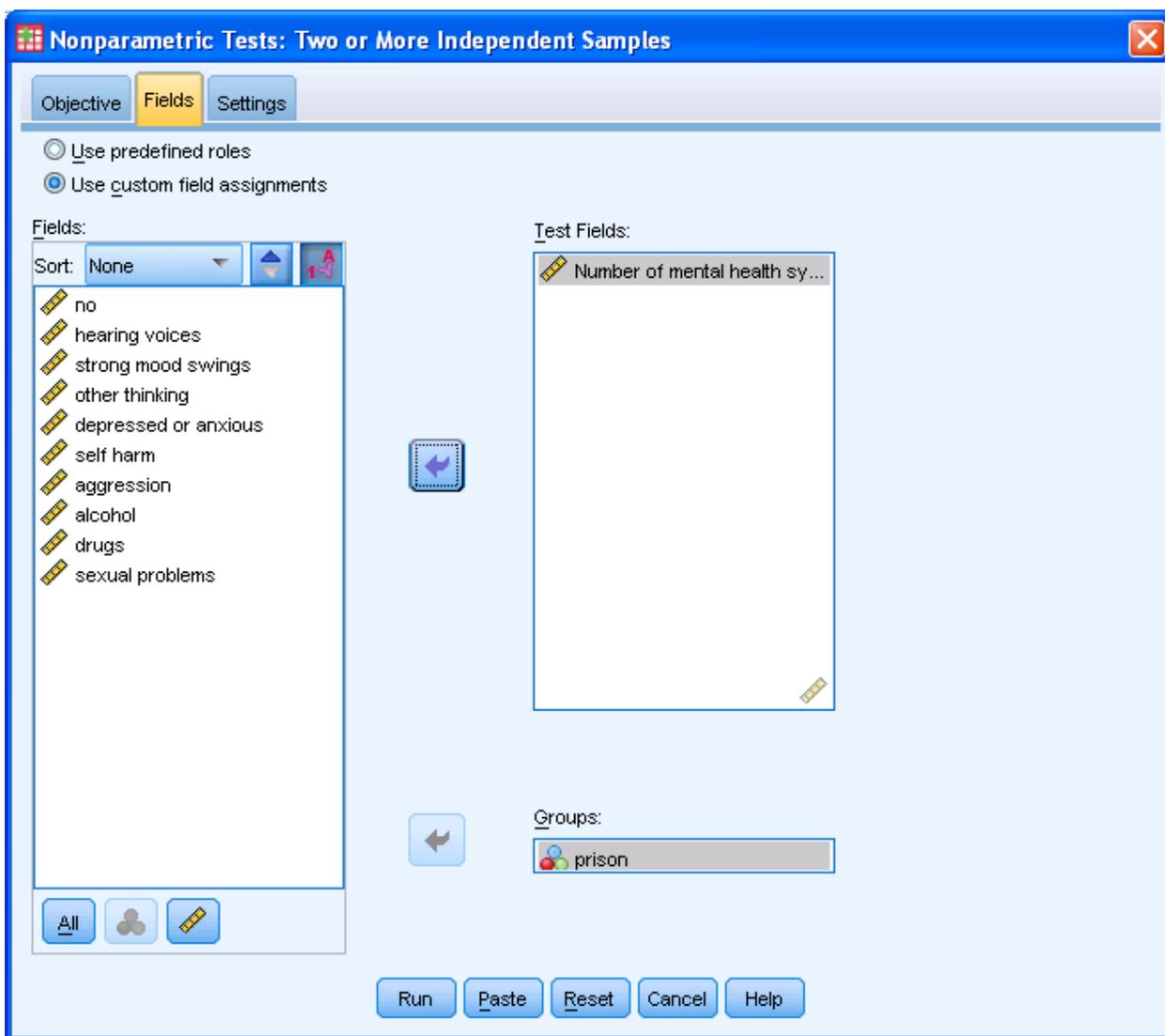
**Maastricht University is the best specialist university in the Netherlands (Elsevier)**

**Visit us and find out why we are the best!**  
**Master's Open Day: 22 February 2014**

[www.mastersopenday.nl](http://www.mastersopenday.nl)



**Figure 89:** Dialogue box for k groups non-parametric test



The output is seen in Table 21.

**Table 21:** Kruskal Wallis output

**Hypothesis Test Summary**

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Number of mental health symptoms is the same across categories of prison.	Independent-Samples Kruskal-Wallis Test	.008	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

This shows that the difference between prisons is significant,  $p < 0.05$  ( $p = 0.008$ ). Thus the null hypothesis (i.e. there is no difference among the prisons with respect to number of symptoms) is rejected, or there are differences among the prisons. Unfortunately there are no post hoc tests available in the non parametric tests. What we could do is a Mann Whitney between all possible groups, but this is increasingly unhelpful as the number of groups increases, and with five prisons there are ten different possible combinations of two prisons out of the five. It is also problematic as with ten pair-wise tests you would expect a  $p < 0.05$  not one time in twenty, but roughly four times in ten. You might think one time in two, but to have a situation where no  $p < 0.05$  for ten independent tests, assuming purely random fluctuations, would mean  $0.95^{10} = 0.60$  to 2 decimal places. Thus to have at least one false positive would have a  $p$  value of  $1 - 0.6 = 0.4$ . There are methods of working out post hoc comparisons with Kruskal Wallis, but not in SPSS, and the methods are not standard or universally accepted.

### Conclusion

We have looked at differences among many groups, and introduced a parametric test (one-way ANOVA) and non-parametric test (Kruskal Wallis).

**Empowering People. Improving Business.**

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

[www.bi.edu/master](http://www.bi.edu/master)

**BI NORWEGIAN BUSINESS SCHOOL**

EFMD EQUIS ACCREDITED



# 9 Correlation

## Key points

- Pearson and Spearman rank are the appropriate tests for correlation of two variables
- Data must be at least ordinal for correlation to make sense

## At the end of this chapter you should be able to:

- Decide which test to use
- Compute and interpret the result

## Introduction

If two variables are correlated, this means as one goes up the other goes up (positive correlation) OR as one goes up the other goes down (negative correlation). However complete correlation is rare, it is seldom the case that when one variable goes up by  $n$  units, the other goes up (or down) by  $m$  units.

An example might make this clearer. Older patients tend to have more complex problems than younger ones. Thus you might expect older patients to stay in hospital longer. This, if true, would be described as there being a positive correlation between length of stay and age of patient. But clearly some very elderly patients come in for a cataract operation and are in for (say) one day, and a nineteen year old man might be in hospital months following a serious motor accident. However other things being equal, there is a trend for older patients to have longer lengths of stay.

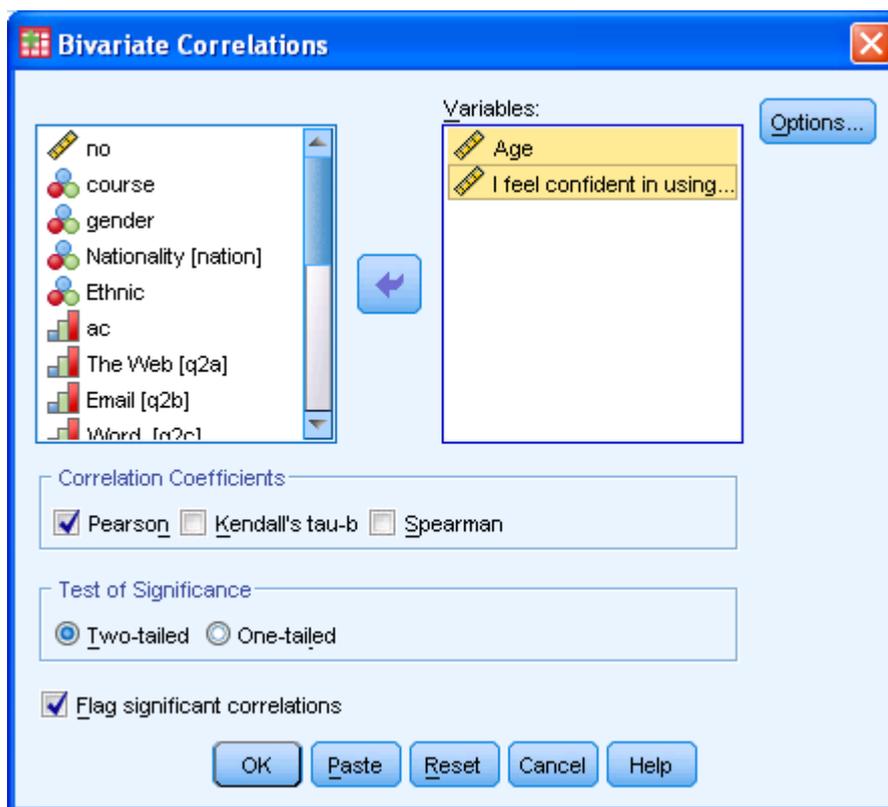
If (for example) it were true that for every year after eighteen, each year of life added one day to length of stay, then length of stay would be entirely predictable from your age. This is total 100% complete correlation, and we would say  $r$  (the correlation co-efficient) is unity ( $+1.0$ ). It is positive as length of stay increases when age increases.

For very young babies however the length of stay might be more, not less, than for older children. If it were true that for every month born pre-term one day were added to length of stay then the length of stay is still entirely predictable, but this is negative correlation, because as length of stay increases age decreases. The correlation co-efficient is unity, but negative ( $r=-1.0$ ).

But as you cannot completely predict length of stay from age, it might be that  $r=0.3$  or  $r=-0.4$  (say). An  $r=0.0$  would mean that as age varied, length of stay did not change in any particular way, it might go up or down or do nothing at all.

It is time to consider a real world example. There is a general assumption that younger people are more confident in using IT. Thus in our online course example we might expect the older students to express less confidence in using computers. To conduct a correlation analysis you use **Analyze** -> **Correlate** -> **Bivariate** and enter the variables as in Figure 90.

Figure 90: Correlation dialogue box



Then you would get the output as in Table 22.

Table 22: Output for correlation

<b>Correlations</b>			
		Age	I feel confident in using computers
Age	Pearson Correlation	1	.106
	Sig. (2-tailed)		.221
	N	139	136
I feel confident in using computers	Pearson Correlation	.106	1
	Sig. (2-tailed)	.221	
	N	136	145

The first thing to note is the significance [Sig (2 tailed)]. If the correlation is not significant it is not worth interpreting the co-efficient. It is like the difference of means in Student’s t test, if the difference between the means is not significant you would not try to interpret the difference in the means (as there probably is none).

In this case there is no significant correlation, so we interpret this as age is not correlated to confidence in using computers. Thus the view that older students would be less confident is not supported. In other words the null hypothesis (there is no significant correlation between age and confidence) is accepted.

There is a limitation with the correlation co-efficient we have chosen. We are using Pearson’s correlation co-efficient (usually denoted as  $r$ ) which is a parametric test. But a histogram showed clearly that age is not normally distributed (neither is computer confidence). We should have used a different non-parametric test, Spearman rank. This does the same thing as Pearson, does not assume data are normally distributed, and is usually denoted as  $r_s$ . To get this in SPSS, click on Spearman rather than Pearson in Figure 90. If we did this we get the output seen in Table 23.

**Table 23:** Spearman rank output

			Age	I feel confident in using computers
Spearman’s rho	Age	Correlation Coefficient	1.000	.138
		Sig. (2-tailed)	.	.110
		N	139	.136
	I feel confident in using computers	Correlation Coefficient	.138	1.000
		Sig. (2-tailed)	.110	.
		N	.136	145

**Exercise**

Interpret the output of Table 23.

**Exercise**

Using the data in the RAE datafile consider the following hypotheses:-

- There is no significant correlation between full time equivalent (FTE) doctoral students and RAE rating
- There is no significant correlation between funding and RAE rating

Remember you will need to check whether the data (both variables) are normally distributed

# 10 Paired tests

## Key points

- Paired Student's t test and Wilcoxon are the appropriate tests for two groups
- Repeated ANOVA and Friedman is for more than two groups

At the end of this chapter you should be able to:

- Decide which test to use
- Compute and interpret the result

## Paired Student's t test

Use the dataset "*anxiety 2.sav*"

Firstly we will see if there is a difference between trial1 and trial2. We cannot use Student's t test for independent groups as this is not a comparison of two groups of one test variable. Rather it is a comparison of two variables in one group. Each subject is recorded twice, once in trial1 and again in trial2. The data are paired, so it is the paired Student's t test we should use. To obtain this follow the steps **Analyze -> Compare Means -> Paired Samples T Test** and Figure 91.

## Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

Get Help Now



Go to [www.helpmyassignment.co.uk](http://www.helpmyassignment.co.uk) for more info

 **Helpmyassignment**

Figure 91: Dialogue box paired t test

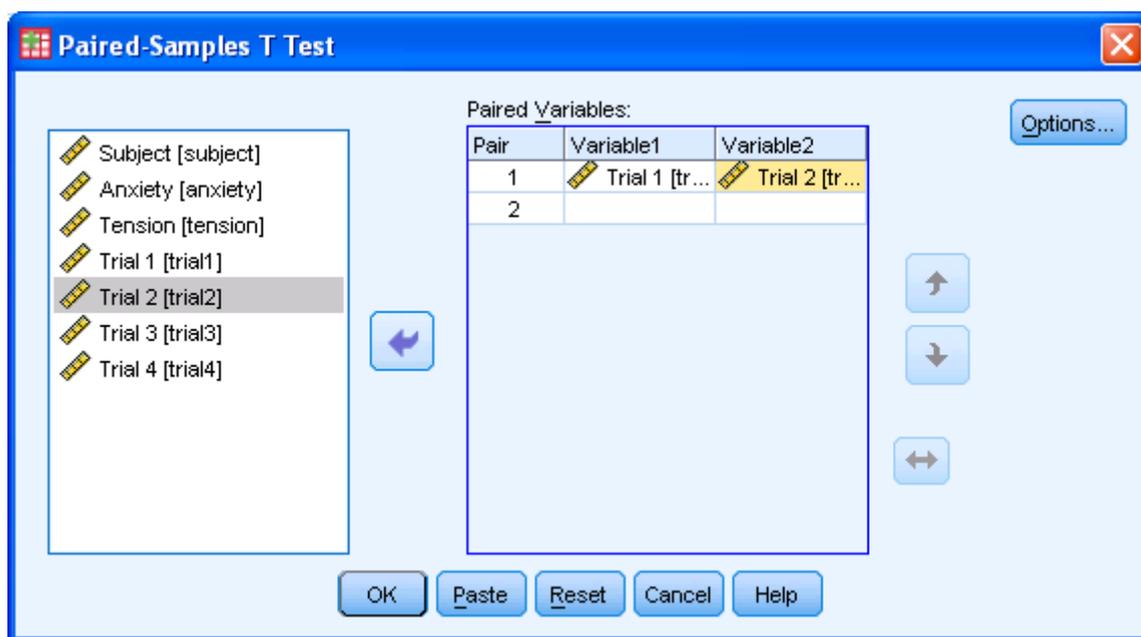


Table 24 shows the mean for trial2 is lower. But is it significantly lower? This is seen in Table 25 and Table 26. There is no significant correlation between trial1 and trial2 but there is a difference in the means ( $p < 0.001$ ).

Table 24: Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Trial 1	16.50	12	2.067	.597
	Trial 2	11.50	12	2.431	.702

Table 25: Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Trial 1 & Trial 2	12	.488	.107

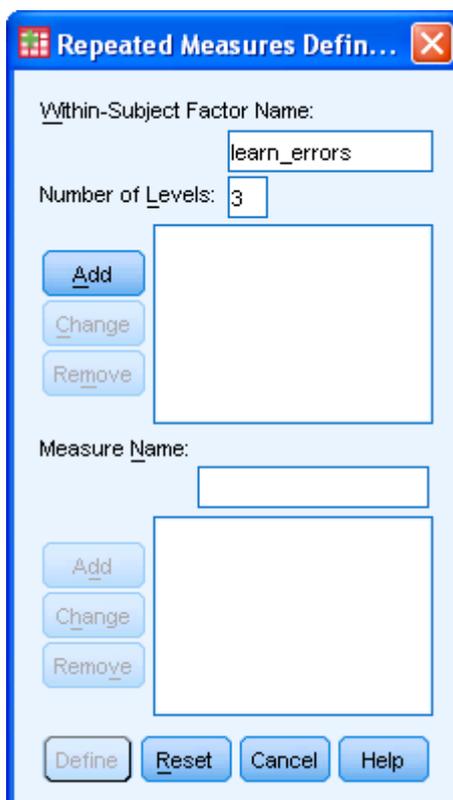
Table 26: Paired Samples Test

		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
					Pair 1	Trial 1 - Trial 2			

## Repeated ANOVA

However we have more than two measurements, how would we (for example) test the differences between trial1, trial2 and trial3. We cannot use the paired test as that only allows for two variables. Here we need to use ANOVA repeated measures. Follow **Analyze** -> **General Linear Model** -> **Repeated Measures** and Figure 92. Here we create a factor, which contains the variables we want to test. You can call this factor pretty much anything, and I have called it learn\_errors. Note you cannot have spaces in factor names so I have used an underscore. We have three variables for testing so the number of levels is set to three. I next clicked on **add** and then **define** to obtain Figure 93. I have also added **Descriptive statistics** (Figure 94) which is obtainable from **Options** in Figure 93.

**Figure 92:** Creating factor



**Figure 93:** Defining factor

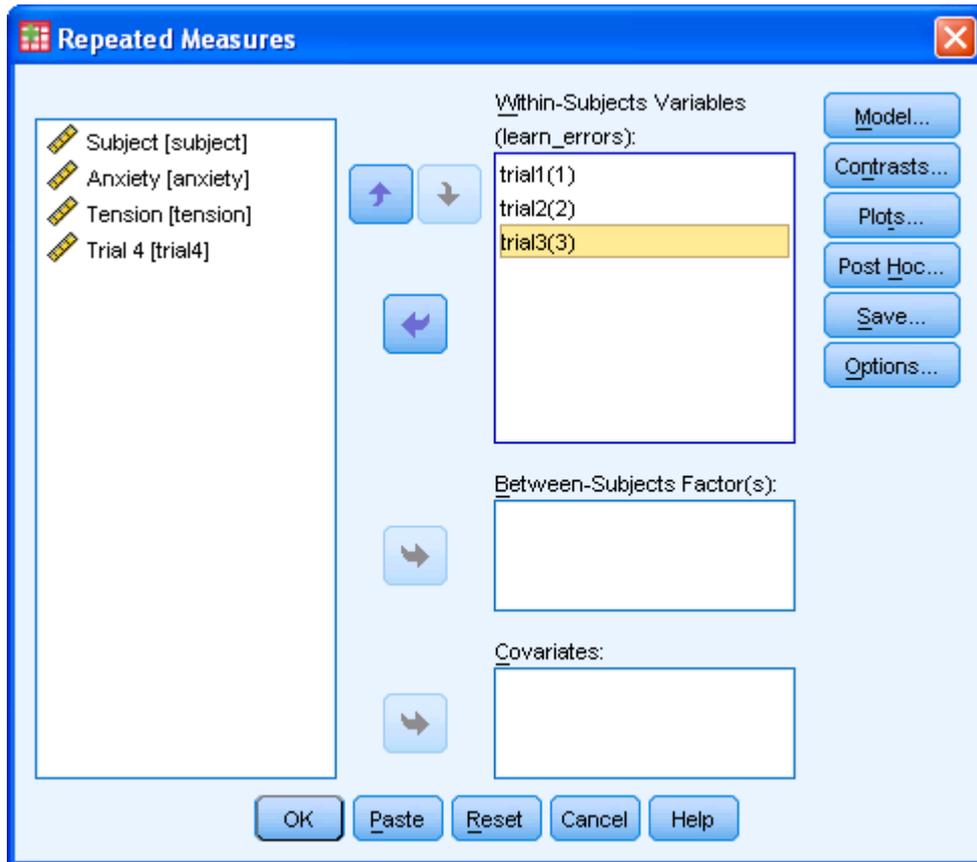


Figure 94: Adding descriptive statistics

# Brain power



Plug into The Power of Knowledge Engineering.  
Visit us at [www.skf.com/knowledge](http://www.skf.com/knowledge)

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

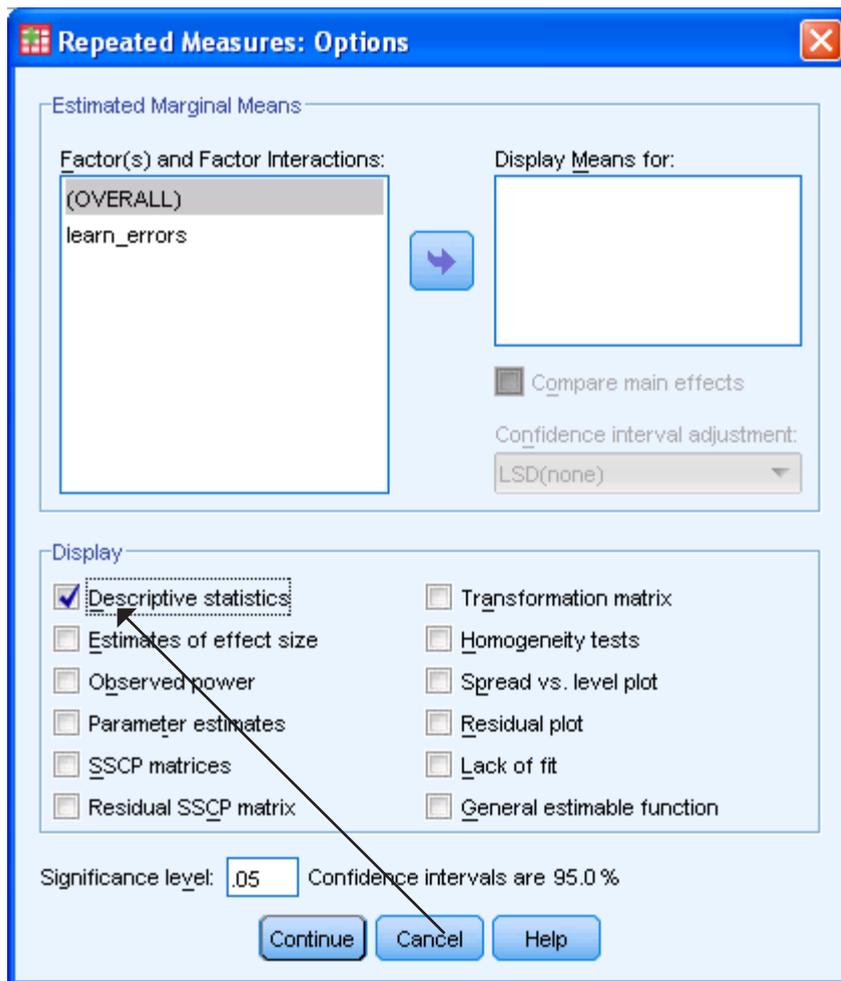
Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

The Power of Knowledge Engineering







There is a lot of output, but most of it is not important now. Table 27 tells you which variables are in the factor for testing. Table 28 shows you that the mean value is going down between trial1 and trial2 and also again between trial2 and trial3. Here we have one variable being tested, anxiety, but in principle there could be many. If that were the case then we would be doing a multivariate analysis (because there are more than one dependent variable of interest). Table 29 tells us that for multivariate analysis there is a significant result. This table is seemingly pointless in this case, however it may be of use (see below). Table 30 tests for sphericity (sphericity means that data are uncorrelated, a low p value indicates a problem here) which shows it is not significant, so it is not a problem. Table 31 gives results if sphericity is assumed which is the case here. The last two tables show the between subjects (Table 32) which explores the differences between each of the subjects (not of interest here) and within subjects (Table 31) that tells us of the difference of successive results on each individual (which we are interested in here). ANOVA assumes sphericity, and this is not a problem here. However multivariate tests do not assume sphericity, hence if this were a problem we would use the p value in the multivariate test. Each of the various tests (Pillai, Wilks, Hotelling, Roy) have their advantages and limitations (see (Field, 2009) but here it makes little difference which one you consider, they are all highly significant. So there is a significant difference between the three variables.

**Table 27:** Within-Subjects Factors

Measure: MEASURE\_1

learn_errors	Dependent Variable
1	trial1
2	trial2
3	trial3

**Table 28:** Descriptive Statistic

	Mean	Std. Deviation	N
Trial 1	16.50	2.067	12
Trial 2	11.50	2.431	12
Trial 3	7.75	2.417	12

**Table 29:** Multivariate Tests(b)

Effect		Value	F	Hypothesis df	Error df	Sig.
learn_errors	Pillai's Trace	.922	59.098(a)	2.000	10.000	.000
	Wilks' Lambda	.078	59.098(a)	2.000	10.000	.000
	Hotelling's Trace	11.820	59.098(a)	2.000	10.000	.000
	Roy's Largest Root	11.820	59.098(a)	2.000	10.000	.000

a Exact statistic

b Design: Intercept

Within Subjects Design: learn\_errors

**Table 30:** Mauchly's Test of Sphericity

**Mauchly's Test of Sphericity<sup>b</sup>**

Measure: MEASURE\_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon <sup>a</sup>		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
learn_errors	.608	4.969	2	.083	.719	.797	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept

Within Subjects Design: learn\_errors

**Table 31:** Tests of Within-Subjects Effects

Measure: MEASURE\_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Learn_errors	Sphericity Assumed	462.500	2	231.250	91.667	.000
	Greenhouse-Geisser	462.500	1.437	321.809	91.667	.000
	Huynh-Feldt	462.500	1.594	290.091	91.667	.000
	Lower-bound	462.500	1.000	462.500	91.667	.000
Error(learn_errors)	Sphericity Assumed	55.500	22	2.523		
	Greenhouse-Geisser	55.500	15.809	3.511		
	Huynh-Feldt	55.500	17.538	3.165		
	Lower-bound	55.500	11.000	5.045		

**Table 32:** Tests of Between-Subjects Effects

Measure: MEASURE\_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	5112.250	1	5112.250	465.712	.000
Error	120.750	11	10.977		

“I studied English for 16 years but...  
...I finally learned to speak it in just six lessons”  
Jane, Chinese architect

ENGLISH OUT THERE

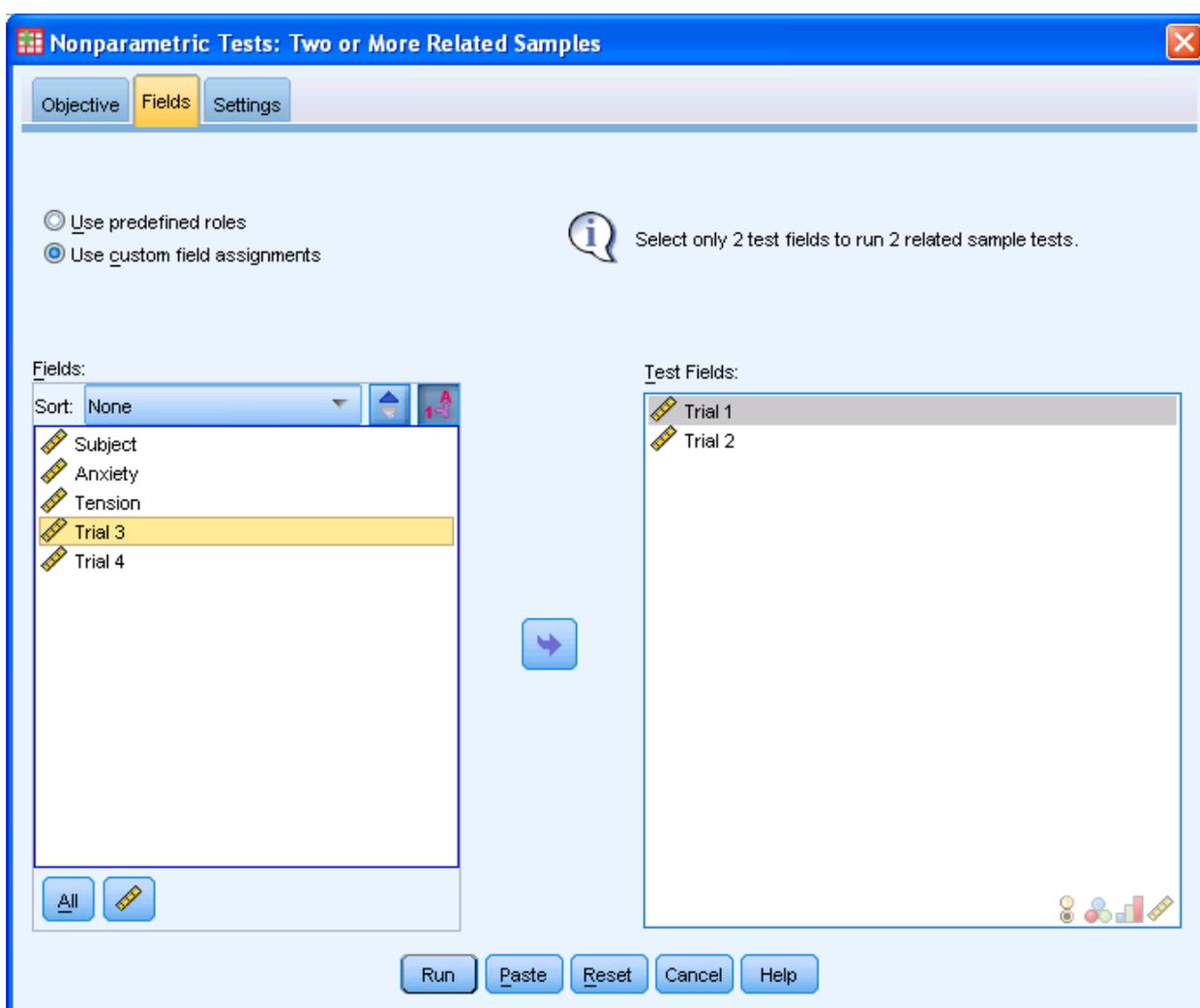
Click to hear me talking before and after my unique course download



## Wilcoxon

Student's t test and repeated ANOVA are parametric tests. What if our data are not normally distributed? Then we need to use equivalent non-parametric tests. For paired variables we can use Wilcoxon. Follow **Analyze** -> **Nonparametric** -> **Related Samples** and Figure 95 to set this up for trial1 and trial2. Table 33 shows the difference is significant at  $p=0.002$ . This does not show us however which is higher, to show this use the error bars option of the bar chart in the **Graph builder**, see Figure 96, where I have selected two variables (click on one, then hold down the control key when selecting the second) and moved them across to the **y-axis** which will give the output in Figure 97, which shows clearly that the values have become lower in trial2, in fact there is no overlap at all for the 95% confidence interval (i.e. the range where we expect 95% of values to occur). Indeed all twelve values are lower in trial2 than trial1. Note this is a (possibly artificially constructed) dataset for tutorial purposes, such clear differences are rare in real data.

**Figure 95:** Dialogue box Wilcoxon



**Table 33:** Test Statistics

**Hypothesis Test Summary**

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Trial 1 and Trial 2 equals 0.	Related-Samples Wilcoxon Signed Ranks Test	.002	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

**What do you want to do?**

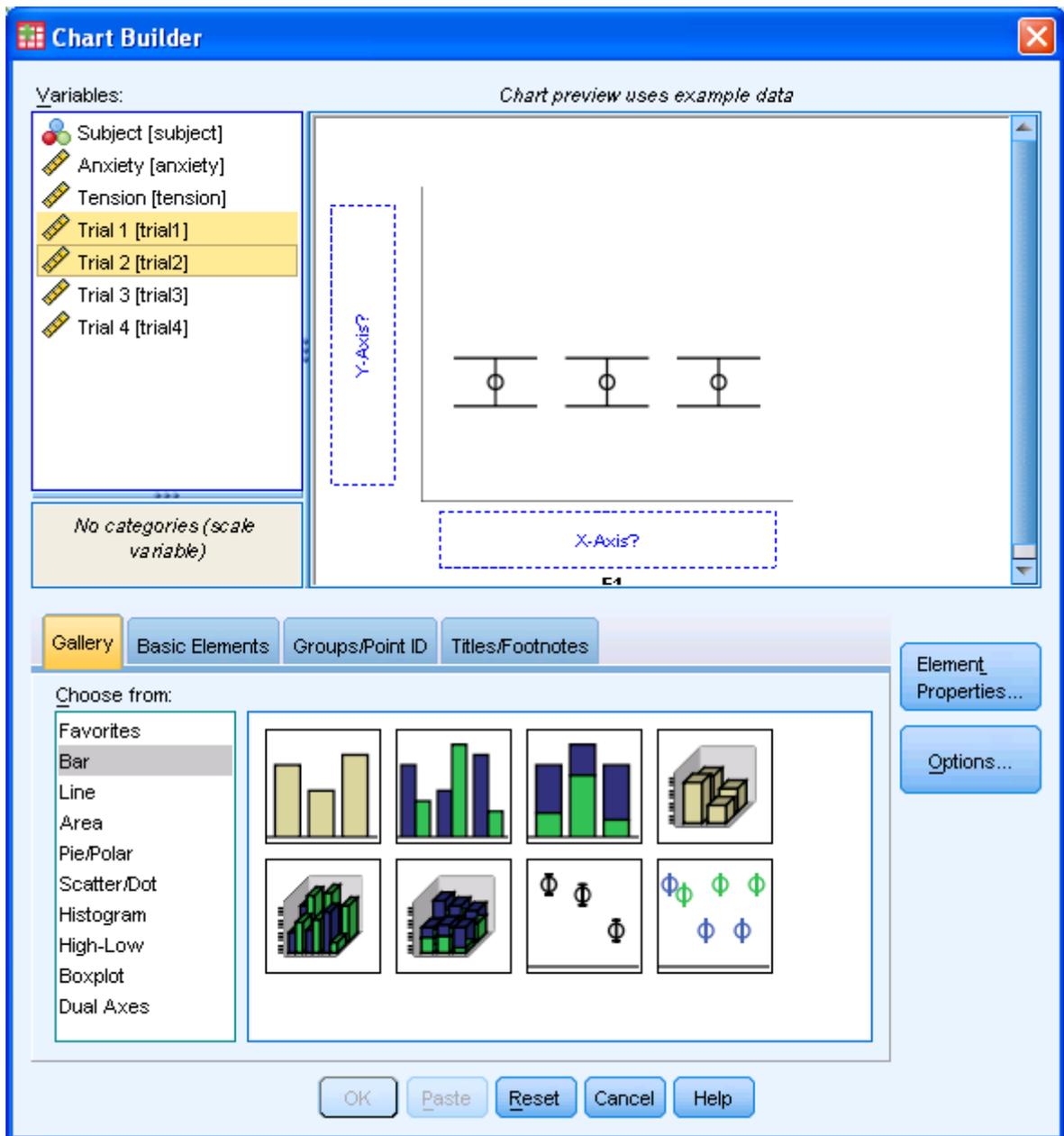
No matter what you want out of your future career, an employer with a broad range of operations in a load of countries will always be the ticket. Working within the Volvo Group means more than 100,000 friends and colleagues in more than 185 countries all over the world. We offer graduates great career opportunities – check out the Career section at our web site [www.volvogroup.com](http://www.volvogroup.com). We look forward to getting to know you!

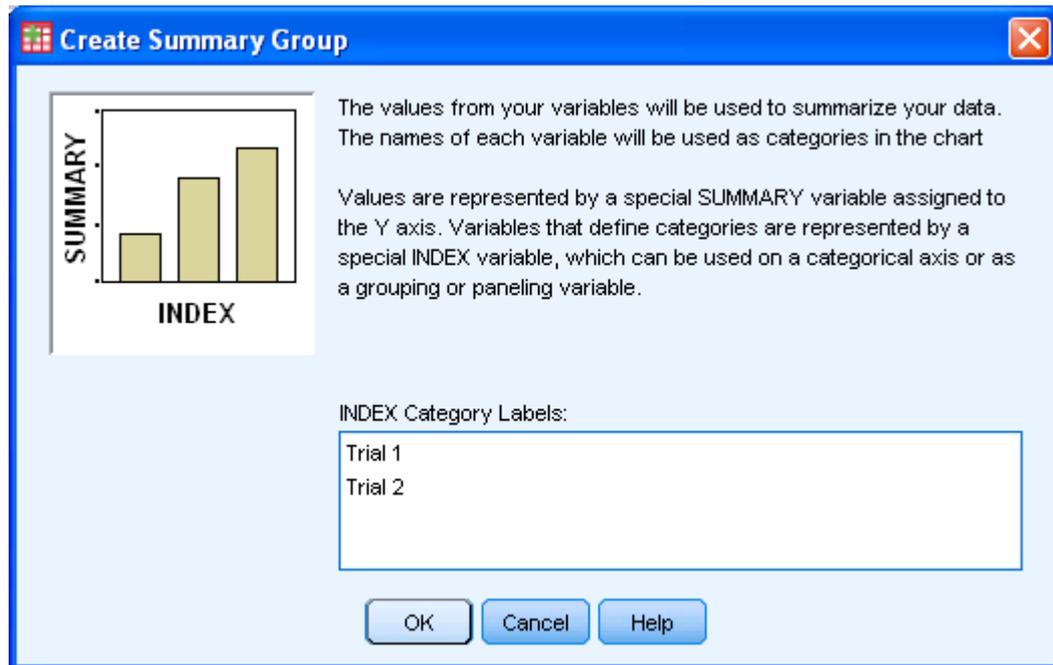
**VOLVO**  
 AB Volvo (publ)  
[www.volvogroup.com](http://www.volvogroup.com)

VOLVO TRUCKS | RENAULT TRUCKS | MACK TRUCKS | VOLVO BUSES | VOLVO CONSTRUCTION EQUIPMENT | VOLVO PENTA | VOLVO AERO | VOLVO IT  
 VOLVO FINANCIAL SERVICES | VOLVO 3P | VOLVO POWERTRAIN | VOLVO PARTS | VOLVO TECHNOLOGY | VOLVO LOGISTICS | BUSINESS AREA ASIA

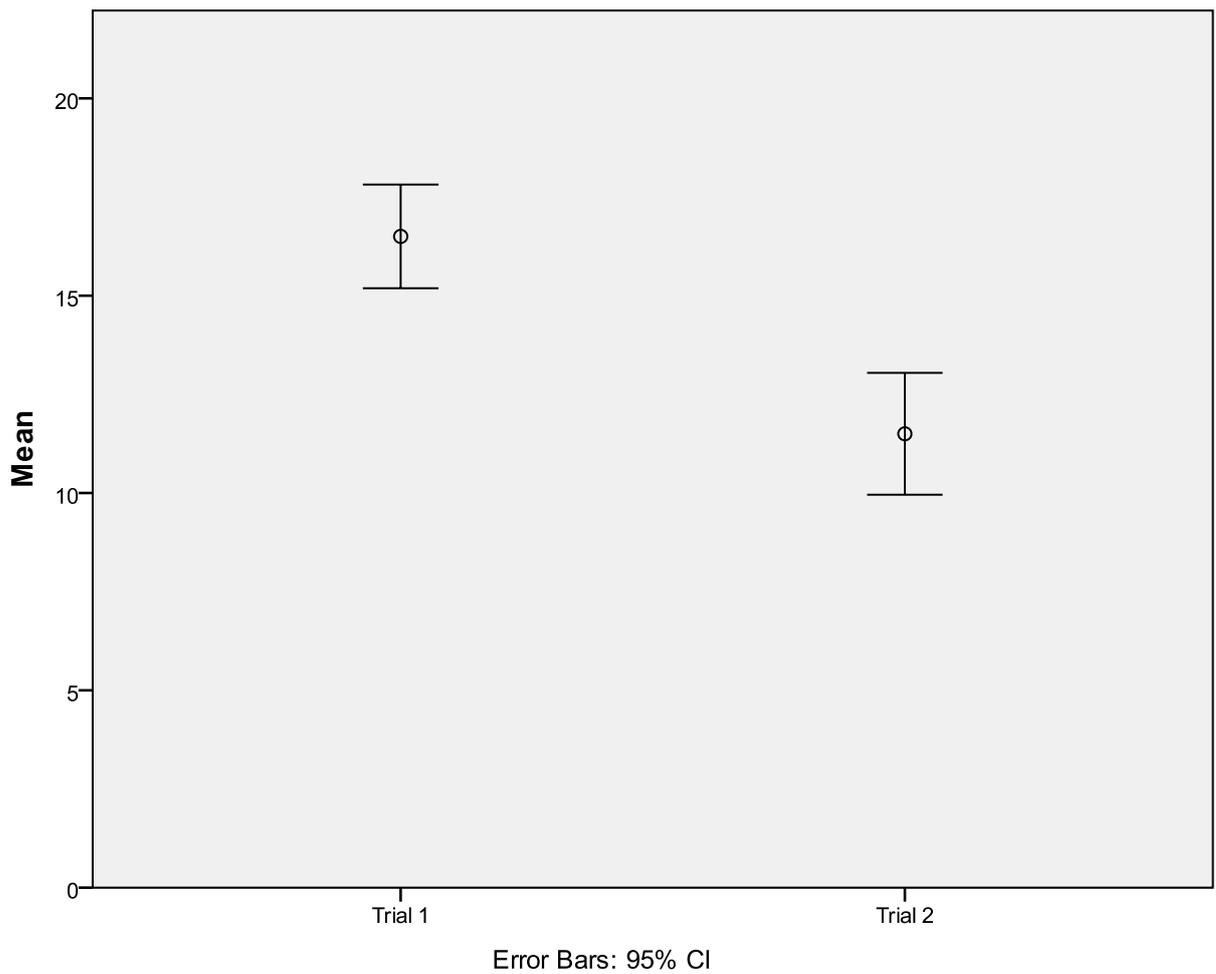


Figure 96: Error bars for anxiety scores





**Figure 97:** error bar chart



### Friedman

For more than two related variables there are several tests, one being Friedman. Follow Figure 98 (obtainable from *Non parametric tests* then *related samples*). Table 34 shows this is significant at the  $p < 0.001$  level. Again an error bar shows this is going down from trial1 to trial2 and then further to trial3, see Figure 99 (the chances of you getting such clear results in small samples are pretty minimal, I only use these files as I have no studies using paired data).

Figure 98: Dialogue box Friedman

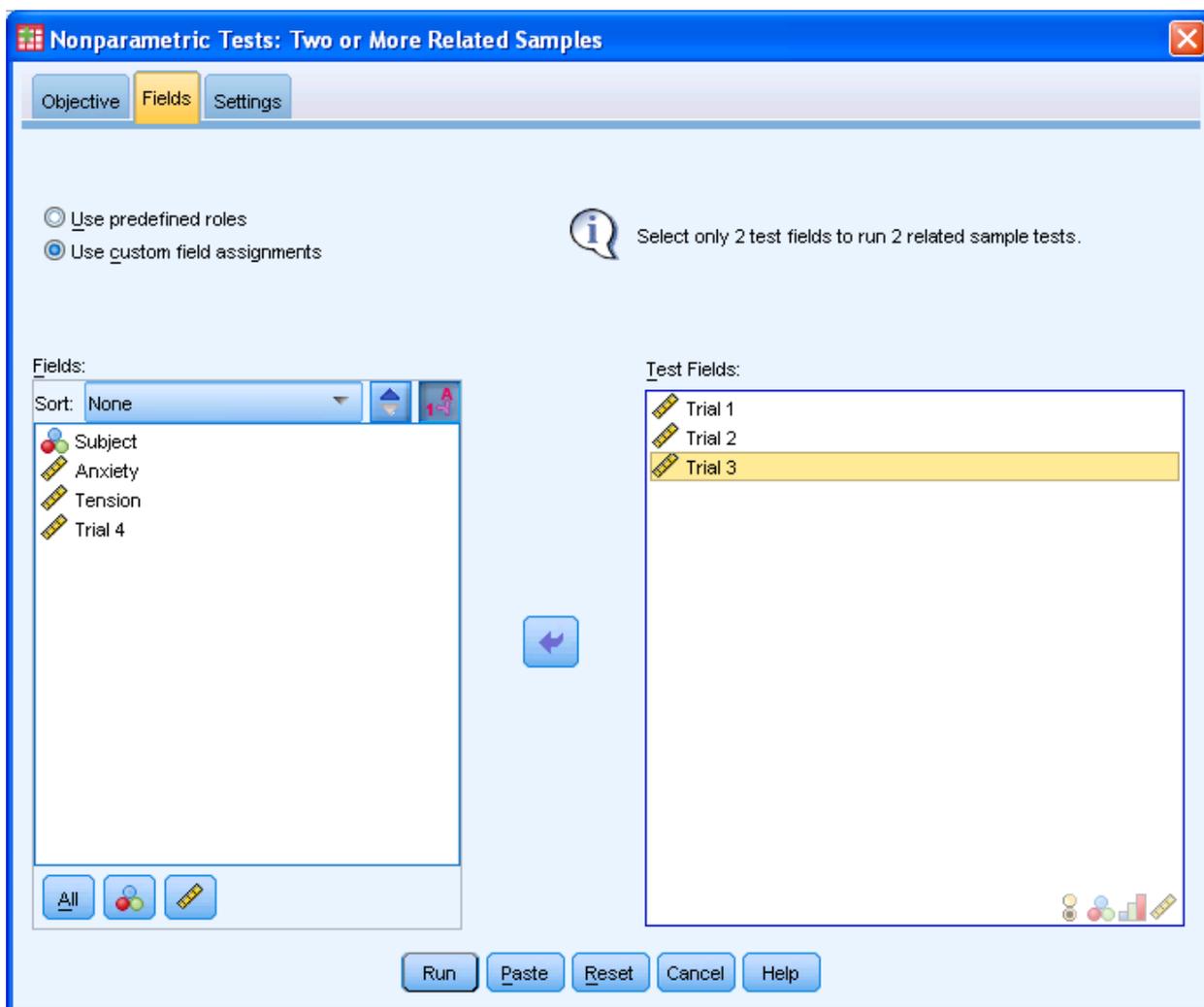


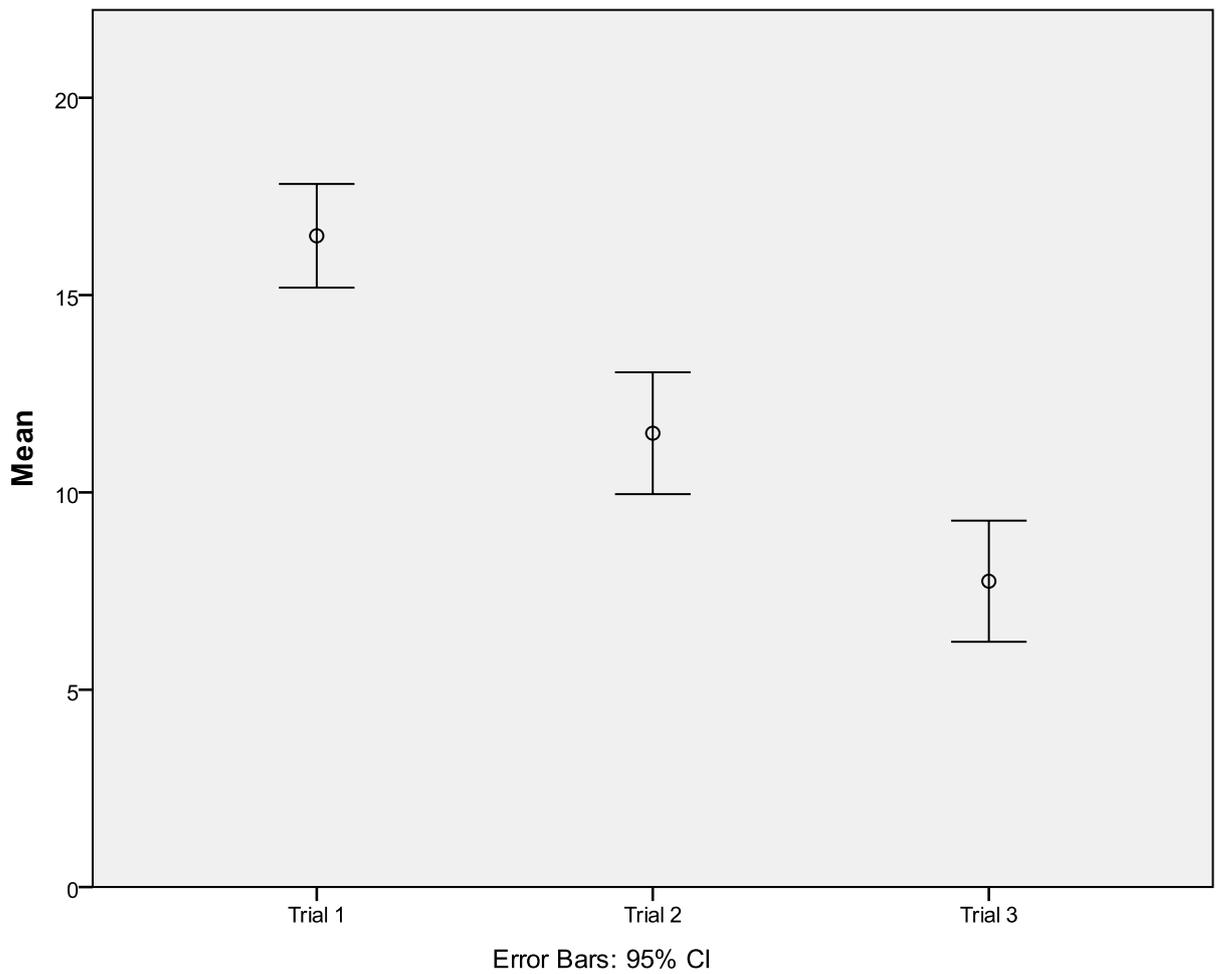
Table 34: Test Statistics

### Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Trial 1, Trial 2 and Trial 3 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

**Figure 99:** Error bars for three trials



### Exercise

Test for the difference between trial2 and trial3. Decide the appropriate test, run it and interpret the results.

Now test for the differences between all four trials, trial1 trial2 trial3 and trial4. Again decide the appropriate test, run it and interpret the results

### Further Reading

<http://academic.reed.edu/psychology/RDDAwebsite/spssguide/spsshome.html> has some general information on SPSS, but a specific introduction to paired tests and repeated measures.

<http://www.utexas.edu/cc/docs/stat38.html> has a detailed discussion of repeated measures ANOVA.

### References

FIELD, A. 2009. *Discovering statistics using SPSS (and sex and drugs and rock 'n' roll)*, London, Sage.

# 11 Measures of prediction: sensitivity, specificity and predictive values

## Key points

1. Sensitivity is a measure of detection of abnormal cases
2. Specificity measures the detection of normal cases
3. For a given risk assessment scale, specificity is inversely related to sensitivity
4. Both sensitivity and specificity are used in receiver operating characteristic (ROC) which is covered in the next chapter.

At the end of this chapter you should be able to:

1. Calculate sensitivity and specificity

**gaiTeye**<sup>®</sup>  
*Challenge the way we run*

**EXPERIENCE THE POWER OF  
FULL ENGAGEMENT...**

**RUN FASTER.  
RUN LONGER..  
RUN EASIER...**

**READ MORE & PRE-ORDER TODAY  
WWW.GAITEYE.COM**

## Introduction

In this chapter we will be considering material that is required to understand receiver operating characteristic (ROC). ROC will be covered in the next chapter.

Diagnosis of fractures is a relatively new area for nursing, and is part of the advanced practitioner role for nurses in some accident and emergency departments since at least the mid 1990s. Classification involves placing objects that are similar to each other (e.g. diseases) into a group. A person with a given set of signs and symptoms is given a specific diagnosis. In one of the examples given in this chapter, radiographs are used to split those who have a fracture from those who do not.

But how accurate are nurses at diagnosing fractures? Diagnosis from a radiograph can be complex, and you can make a mistake. The nurse can be perfect at picking up fractures by always saying that there is a fracture, as then she<sup>1</sup> will miss none, but nor will she ever declare a radiograph clear. The technical description of such a classification is that it is very sensitive, in that no fracture is missed; but it is not specific at all, in that all radiographs are said to show a fracture. These two terms will be discussed and defined below, but sensitivity measures how good we are at picking up the disease, and specificity how good we are at identifying the normal. The above is clearly a very extreme case, but shows that a very sensitive measure is not necessarily useful. At the other extreme a nurse may never say a radiograph shows a fracture, and by missing all fractures that do occur shows a low sensitivity (in fact zero). However, she does not ever misinterpret a normal radiograph as showing a fracture, and is said to have a high specificity. Clearly this form of classification is useless also. Thus ensuring that you never make a mistake in misinterpreting a fracture as normal, or never making the opposite mistake of misinterpreting a normal radiograph as showing a fracture will not typically give a useful classification.

Similarly, you might want to predict those patients who will develop pressure sores by using a system that splits patients who are high risk for developing a pressure sore, according to some scoring system. Again the system is most unlikely to be 100% accurate; many patients stated to be at high risk will not develop sores, and some who at not considered at risk may do so. The concept of a threshold is inherent in such a system, as the patient is given a score indicating risk. But at which score do we consider them at risk? We can choose any score as the one that defines risk, but we will typically want the threshold score to split the patients into those likely to develop sores and those unlikely to develop sores, which is not necessarily (or usually) the same thing as splitting them into those that definitely will not from those that definitely will develop sores. If the score indicates increasing risk, raising the threshold will result in a less sensitive tool, and lowering the threshold will give a more sensitive one. Unfortunately, as shown above, the more sensitive measure is not necessarily a better one.

## Sensitivity and specificity

Let us now define these two concepts. When making a binary decision, as for example deciding on the diagnosis of a fracture or no fracture (it is or it is not a fracture, there are two cases), one may make a true diagnosis where the fracture is present or absent, or a false diagnosis, also where the fracture is present or absent. There are four cases in total:

1. True-positive decision: the patient has a fracture (disease, will develop a sore etc.) and we state they do
2. True-negative decision: the patient does not have a fracture etc. and we state this correctly
3. False-positive decision: the patient does not have a fracture etc. but we state they do
4. False-negative decision: the patient has a fracture etc., but we say they do not.

True positive and true negative rates can be used together to give measures of accuracy.

Sensitivity is identical to a true positive ratio. It is defined as the proportion of all positive cases that are correctly identified as positive. Sensitivity measures how well we perform in identifying those subjects with a fracture (or other condition). As all the positive cases comprise those that we got right (true positives) and those we got wrong (false negatives), the ratio of true positive to all positives is given by

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN})$$

Specificity is defined as the proportion of all negative cases that are correctly identified as negative. Specificity measures how well we perform in identifying those subjects who are normal (i.e. do not have a fracture or some other condition of interest). As all the negative cases comprise those that we got right (true negative) and those we got wrong (false positive), the ratio of true negative to all negatives is given by.

$$\text{Specificity} = \text{TN}/(\text{TN}+\text{FP})$$

Specificity is the same as the true negative ratio. In practice the false positive ratio is often used. The true negatives and false positives are all the negative cases, and it follows that the true negative ratio and false positive ratios add to unity. Using the false positive ratio is therefore identical to using (1- specificity).

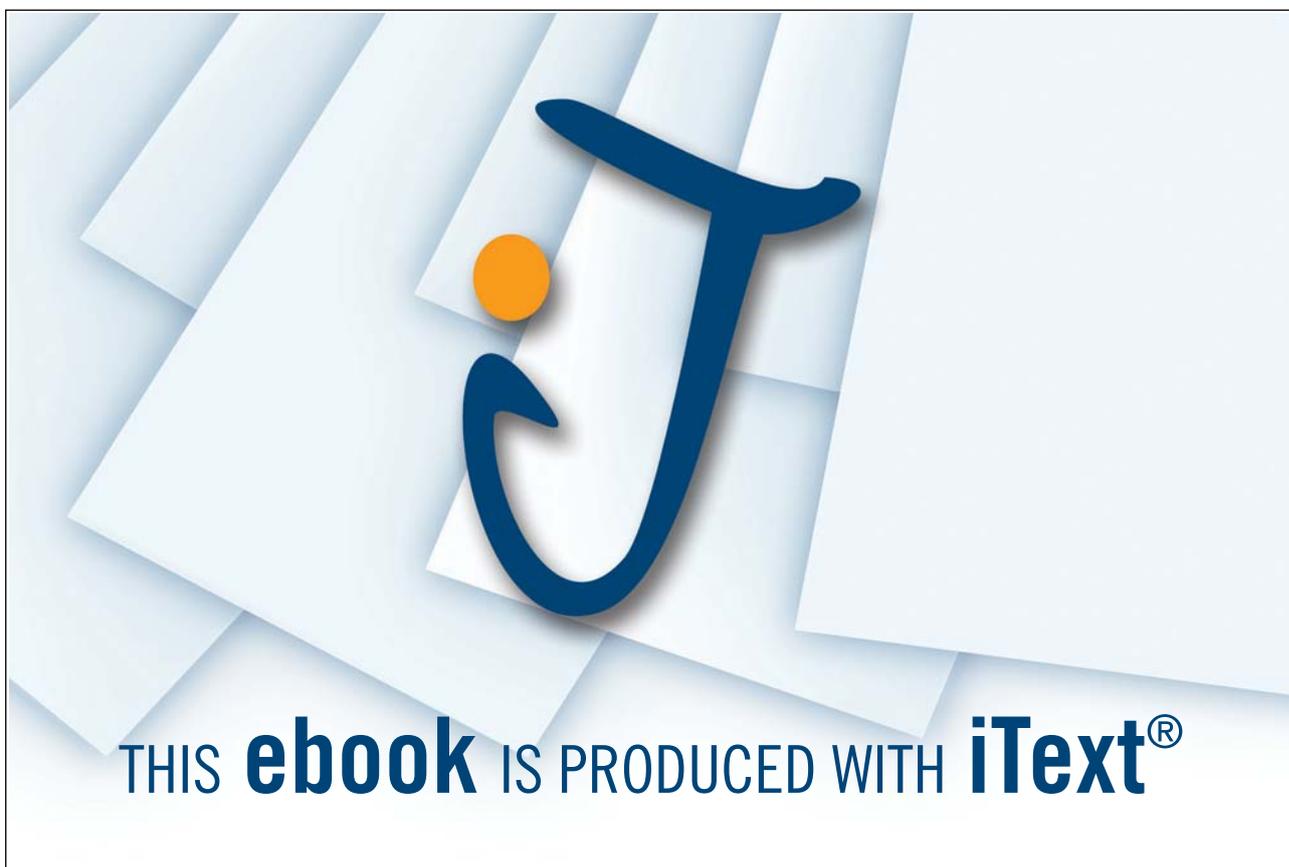
Sensitivity and specificity are simply the number of correct cases as a fraction of the totals for those patients who have/ do not have the condition.

## Thresholds

One does not typically state with absolute certainty that, for example, a patient has a fracture or that a patient will develop a pressure sore, or that they do not have a fracture or will not develop a sore. A more natural statement would be that one is very sure, or fairly confident, or that it is quite unlikely, etc.

However, sometimes a straightforward choice of where to allocate resources requires a decision to be made, even if we cannot be 100% sure of what the right decision is; so we need a best guess. For example, the patient is at high risk of pressure sore formation and therefore needs a special mattress, or they are not and the mattress can be allocated to someone else. Thus one needs to identify a threshold above which the patient is at high risk and below which they are not. For example, the Waterlow score above which a patient is said to be at high risk could be stated to be 10. In this case any value above 10 (the threshold) is said to indicate risk. It is also possible to have a scoring system such as the Braden score, which works in the opposite direction; i.e. low scores are indicative of risk, in which case values below a given threshold would indicate risk.

If it is important to get the diagnosis right and making a false positive diagnosis is not problematic, then one can set the threshold low (or, if increasing score implies less risk, we would set the threshold high, as the scale is the other way round, where decreases in the scale imply more risk). If a patient monitor gives an alarm when there is nothing wrong with the patient, it causes some irritation. However, if the alarm fails to go off and some injury or death could ensue then clearly this is far more important to avoid than the irritation of false alarms. There is a penalty for a false positive alarm, i.e. resetting the alarm, irritation and noise. However, as the penalty for a false negative alarm could be death, we set the monitor to be very sensitive, or to have a low threshold. Thus we accept a poor specificity in this situation in order to ensure a very high sensitivity.



While in health care it is usually important to not miss a true case (a diagnosis of some disease, for example), i.e. a false negative is almost always a problem, thus setting an arbitrarily low threshold is not advised. There may be a significant penalty in getting the decision wrong either way; failure to accept true positive, and reject true negative may both be dangerous. For example, if a test shows a patient to have a pulmonary embolism (PE) we would treat with anticoagulants, as PE can be fatal. The penalty for a false-negative diagnosis would therefore be a failure to treat the PE, resulting in death from respiratory arrest. However, unnecessary treatment with anticoagulants can cause dangerous bleeding in post-surgical patients, and therefore a highly sensitive test that always gets the PE cases but also wrongly diagnosis many who do not have the disease could also cause danger. The benefit from a true-positive diagnosis is that the potentially life-threatening condition of PE is treated, and the benefit from a true-negative diagnosis is that iatrogenic bleeding of a patient who does not have PE is prevented.

There is thus a trade-off in setting the threshold level. For example, in mass screening for human immunodeficiency virus (HIV) one might have a blood test where high values indicate a high probability of having the virus and lower values indicate less probability of having the virus. If one sets the threshold too high, many of the population with the condition will not be picked up, while if one sets the threshold too low many of the population will be diagnosed HIV positive who do not in fact have the virus, causing distress.

The level at which one sets the threshold depends on the relative merits of specificity and sensitivity in the given domain. This may be, and often is, a subjective assessment.

## Confidence levels

One method of obtaining several threshold levels simultaneously is to allocate confidence levels to the decisions. For each level of confidence, one measures true positives and false positives as a percentage of the maximum true positives and false positives possible. Thus one could ask clinicians to state how sure, using a score range of 1-5, they were in making a given decision (e.g. a diagnostic decision). Rather than use numerical confidence levels, which could be meaningless to a clinician, the levels could be obtained from a Likert scale, for example a five-point scale such as the one shown below:

5 Very confident that disease is present

4 Quite confident that disease is present

3 Unsure if disease is present or not

2 Quite confident that disease is not present

1 Very confident that disease is not present

In this example, if the level was allocated a number as shown above, then one might set a threshold of 3, whereby any value more than or equal to 3 would be considered a positive diagnosis, which would include many cases of 'unsure'. Choosing 4 as the threshold value would remove the 'unsure' cases, with only 'very confident' or 'quite confident' being now considered a positive diagnosis. Thus less false-positive but more false negative diagnoses would probably ensue.

## Gold standard

If one is to test the classification ability of a person, scoring system, etc., then one needs to know what the correct classification should be. This is not always easy. If one uses a specialist in classifying as the 'gold standard' for determining how well an inexperienced trainee is performing, one implicitly trusts the specialist to always be right - but she may make mistakes. Where possible the 'gold standard' should be objective and demonstrably correct; where this is impossible it should be a classifier that one has much more confidence in than the one being assessed.

In a prospective study the gold standard may be whether the patient ultimately contracts the disease or condition. This may be very clear in some cases (e.g. from post mortem examination).

## Sensitivity, specificity and predictive values

You should now be familiar with the concepts of sensitivity and specificity. However there are two other concepts that are useful. While sensitivity and specificity are good research measures to assess an instrument, clinically the measures of predictive values, positive predictive value (PPV) and negative predictive values (NPV) are more useful. These (and sensitivity and specificity) are discussed well in Altman & Bland (1994b, 1994a, 1994c). Wikipedia has a clear and (when I last looked anyway) accurate description.

$$PPV = TP/(TP+FP)$$

$$NPV = TN/(TN+FN)$$

This looks rather similar to sensitivity and specificity:-

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$

But where in sensitivity and specificity the denominator is the number **found** to be positive (sensitivity) or negative (specificity) for predictive values have denominators of number of cases that **test** positive (PPV) or negative (NPV).

So when would one use one pair of tests rather than the other? It depends on your interest. As a researcher evaluating a test you want to know if it distinguishes between cases (patients with HIV for example) from controls (patients who do not have HIV). But what if you are a patient who has just had a test? You want to know, given a positive result, how likely is it that you have the condition, in this example HIV. You want the PPV which tells you this. If you had a negative result the NPV would tell you how likely it is you do not have the condition.

PPV and NPV are highly dependent on the prevalence of a disease. You could have an almost perfectly sensitive test, say 0.999 which is almost perfectly specific, say also 0.999. You get a positive test for HIV. Surely you must almost certainly have HIV? If the prevalence of a disease is very low then PPV will be much lower than if prevalence is high, even where sensitivity and specificity are both great. If prevalence were 1.5% you would almost certainly be HIV positive, but if prevalence were 0.01% you would have only a 50:50 chance of being HIV positive. If you find this confusing you should read Ben Goldacre excellent article which makes this apparent contradiction very clear (Goldacre, 2006).

### Example

In youth offending teams (YOTs) young people are assessed using the asset score, a risk assessment tool. Asset scores (see Appendix) are considered to indicate risk of reconviction according to the numerical total score:-

**Table 35:** Risk bands for asset scores

Score band	Risk level
0-4	Low
5-9	Low-medium
10-16	Medium
17-24	Medium-high
25-48	High

Suppose we now wanted to see if this worked as a risk tool for custodial sentence (it was not designed to do this, but as an illustration it will suffice). If we took anyone who is at high risk (>24) then we can see (Table 35) that 65 of the 137 custodial sentences were in the high risk group, and 102 of the 958 non-custodial sentences were in the high risk group.

www.sylvania.com

**We do not reinvent the wheel we reinvent light.**

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM **OSRAM SYLVANIA**



**Table 36:** Asset total

Risk band	Non custodial sentence	Custodial sentence	Total
Low	140	0	140
low-medium	229	4	233
Medium	297	29	326
medium-high	190	39	229
High	102	65	167
Total	958	137	1095

We can reformat this as Table 37.

**Table 37:** TN, FN, FP, TP values

	Non custodial sentence	Custodial sentence	Total
Not high risk	856 (TN)	72 (FN)	928
High risk	102 (FP)	65 (TP)	167
Total	958	137	1095

$$\text{Sensitivity} = 65/137$$

$$= 0.47$$

$$\text{Specificity} = 856/958$$

$$= 0.89$$

This we have a high specificity but a low sensitivity.

For predictive values we have:-

$$\text{PPV} = 65/167$$

$$= 0.39$$

$$\text{NPV} = 856/928$$

$$= 0.92$$

Thus for the young person in the high risk group, the chance of getting a custodial sentence is a little lower than sensitivity would suggest, and a young person not in the high risk group is more likely to not get a custodial sentence than specificity would suggest. I.e. they are less likely on both counts to get a custodial sentence, and this is because the chance of getting a custodial sentence is only about 12%. What if it were 1.2%?

**Table 38:** TN, FN, FP, TP values

	Non custodial sentence	Custodial sentence	Total
Not high risk	856 (TN)	7 (FN)	863
High risk	102 (FP)	6 (TP)	108
Total	958	13	971

$$\text{Sensitivity} = 6/13$$

$$= 0.46$$

$$\text{Specificity} = 856/958$$

$$= 0.89$$

So sensitivity and specificity are unchanged by the prevalence (here custodial sentence). This is the strength of these measures, they are invariant to prevalence. (Note the slight change in sensitivity is simply a rounding error, I have changed 72 to 7 and 65 to 6. In other words the prevalence is not exactly 1.2% but as near as I can get. If I had kept it at exactly 1.2% the sensitivity would not have changed at all.)

$$\text{PPV} = 6/108$$

$$= 0.06$$

$$\text{FPV} = 856/863$$

$$= 0.99$$

So the PPV is massively worse when the prevalence reduces by a factor of ten. So if custodial sentencing were to be rare in a population, the predictive value is weak even when the sensitivity remains the same.

## Exercise

Calculate the sensitivity, specificity, PPV and NPV for asset score against custodial outcome where medium-high risk is considered the threshold. I have started the table for you:-

	Non custodial sentence	Custodial sentence	Total
Less than medium-high risk	(TN)	33 (FN)	
At least medium-high risk	292 (FP)	(TP)	
Total			1095

How has changing the threshold changed the four measures?

## Conclusion

We have looked at four measures of the outcomes of an instrument at a given threshold. In the next chapter we will extend this to many thresholds.

## Resources<sup>2</sup>

[http://en.wikipedia.org/wiki/Sensitivity\\_%28tests%29](http://en.wikipedia.org/wiki/Sensitivity_%28tests%29)

[http://en.wikipedia.org/wiki/Specificity\\_%28tests%29](http://en.wikipedia.org/wiki/Specificity_%28tests%29)

[http://en.wikipedia.org/wiki/Positive\\_predictive\\_value](http://en.wikipedia.org/wiki/Positive_predictive_value)

[http://en.wikipedia.org/wiki/Negative\\_predictive\\_value](http://en.wikipedia.org/wiki/Negative_predictive_value)

## References

ALTMAN, D., G & BLAND, J. M. 1994a. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ*, 308.

ALTMAN, D., G & BLAND, J. M. 1994b. Statistics Notes: Diagnostic tests 2: predictive values *BMJ*, 309.

ALTMAN, D., G & BLAND, J. M. 1994c. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots *BMJ*, 309.

GOLDACRE, B. 2006. Crystal Balls... and Positive Predictive Values. *The Guardian*, December 9th.

# 12 Receiver operating characteristic

## Key Points

- The area under a ROC curve can be used to directly compare the classification by two or more scales, techniques, human assessors, etc.

At the end of this chapter the student should be able to:

- Plot a ROC curve
- Compare two assessment methods using ROC

## Introduction

In the last chapter we looked at sensitivity and specificity for a given threshold. Here we consider a technique that looks at many thresholds at once. This is receiver operating characteristic (ROC), which is a technique used in classification analysis. ROC analysis is well known in psychology, medicine, medical physics, chemistry and medical imaging, but is less well known in many branches of health care. For a good introduction to its use in psychology see Rose (1995) and for medical imaging see Swets (1979). A summary for medicine is given in Altman & Bland (1994a, 1994c, 1994b) and a detailed description can be found in Zweig & Campbell (1993). The accuracy of diagnosis of distal fractures and the prediction of pressure ulcers show the worth of ROC in areas such as nursing (Overton-Brown and Anthony, 1998) or physiotherapy.

## CHALLENGING PERSPECTIVES

# Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of



learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

**We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.**

To find out more and apply, visit [www.jobs.eads.com](http://www.jobs.eads.com). You can also find out more on our **EADS Careers Facebook page**.










The potential use for ROC will be any area where decisions are made, where you are not certain when you are right, and where the decisions can be made with various levels of confidence. This is typical of most health care decisions, and all of the interesting ones (i.e. the decisions that require some expertise).

One can use ROC to compare different classification systems directly. Examples might include:

- Two or more interpreters of a radiograph rating the presence of a fracture on a Likert scale
- Two or more pressure ulcer risk scales
- Two or more clinicians diagnosing a disease
- Two or more automatic image analysis computer programs

### Assessment at many different thresholds: ROC

The sensitivity and specificity is, as shown in chapter 11, dependent on the threshold used. However, one does not need to insist on using only one threshold. One can use several; ROC is the term applied to the analysis and measurement of sensitivity and specificity at many thresholds simultaneously. The advantage of ROC is that one can decide which threshold is optimal for a given decision-making problem. One can also directly compare different classification systems by looking at the full spectrum of thresholds possible. Without such an approach it is always difficult to compare two systems, as system A may show greater sensitivity than system B simply because it was assessed at a threshold that allowed great sensitivity, while in general system B might actually perform better than system A.

Considering distal fractures again, by looking at the sensitivity and the specificity of the diagnosis together, ROC can assess the nurse and compare her ability with other professionals. This will clearly be necessary if nurses are to be accepted as 'safe' in this new domain for nursing practice. ROC was used to compare nurses with accident and emergency medical staff (casualty officers) with regard to the assessment of distal fractures, and no significant differences between the two groups were found. This study gives quantitative evidence that nurses are competent in this new area (Overton-Brown and Anthony, 1998).

## DRAWING THE ROC CURVE

An ROC curve is simply a plot of the true positive rate against the false positive rate for given thresholds. Note some workers use sensitivity against specificity, but most use  $(1 - \text{specificity})$ , which is the false positive rate. The two approaches are entirely analogous. A system that classifies randomly produces a straight line, with the diagonal running from bottom left to top right on the ROC plot. Some ROC curves may be shown as percentages from 0% to 100% rather than as numbers 0 to 1; this is simply achieved by multiplying all the sensitivities and specificity figures by 100.

The better the classification the further the curve is from the diagonal. In a perfect classification all the patients below a certain threshold will have no fractures, and above it all the patients will have fractures. Ideally, data that are just above the threshold will give a point on the graph that is close to zero for the false positive rate and close to 100% for the true positive rate, i.e. a point that lies in the top left-hand corner of the graph and as far from the diagonal as possible.

In the real world there will usually be some incorrect classifications, but a good classification will have a threshold above which few false positives will be present but many true positives will be seen. Above the optimal threshold, the number of false positives will start to reduce, but less quickly than the number of true positives, and below the threshold the opposite will be true. Thus a curve is seen when plotting true against false positive. Two or more curves may be compared directly. If one curve lies above and to the left of the other it is a more accurate classifier.

### MEASURE OF CLASSIFICATION

The area under the curve gives some measure of the ability of the system to classify. An area of one-half of the maximum would be random (the area under the diagonal, which exactly splits the graph in two equal parts), while the nearer the value is to the maximum area the better the classification.

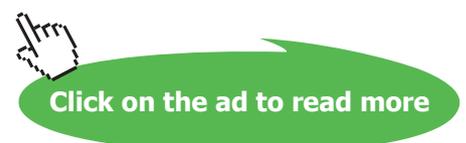


360°  
thinking.

Discover the truth at [www.deloitte.ca/careers](http://www.deloitte.ca/careers)

**Deloitte.**

© Deloitte & Touche LLP and affiliated entities.



## ADVANTAGES OF ROC

ROC has several advantages:

1. ROC analysis allows you to assess a classification at several sensitivities and specificities, not merely one pair.
2. Two or more ROC plots can be compared visually. A plot lying above and to the left of another plot shows greater accuracy.
3. Sensitivity and specificity are calculated totally separately, using results from different groups. Sensitivity considers only those that have the condition, while specificity considers only those that do not. Thus the ROC plot is independent of prevalence of the disease (or condition under scrutiny, in the above cases fractures or ulcers).

## PROBLEMS WITH ROC

ROC is not without its difficulties.

1. Two ROC curves can have the same area but have very different characteristics. For example, one plot may have a much better specificity at low sensitivity, and the other better specificity at higher sensitivity. The area under an ROC curve is a single measure, and information is necessarily lost by collapsing data down to one value. The whole plot should be examined.
2. Relative costs of false classification (false positive and false negative) can be complex, so identifying the optimal threshold from an ROC plot is not necessarily easy. However, techniques do exist to achieve this end.
3. Comparison of ROC plots statistically can be difficult, especially where the tests are correlated (e.g. where two tests are performed on the same patient).

## Creating ROC curves in SPSS

In SPSS 14.0 (and any versions since about 2000) you can easily get an ROC. Let us consider the dataset of young offenders. This dataset, “young offenders” has the asset score, which is a measure of risk of offending, and whether the young person got a custodial sentence or not, where they were tried, gravity of offence and some demographic details.

To get an ROC curve use **Analyze -> ROC Curve**. In Figure 100 I have put total asset score as the test variable, which is the variable used to test whether, in this case, the young person will get a custodial sentence. The state variable is the one that shows whether the young person did or did not get a custodial sentence. I have set this to one. Finally I have ticked to have an optional diagonal line. The output is shown in Figure 101 and Table 39.

Figure 100: ROC dialogue box

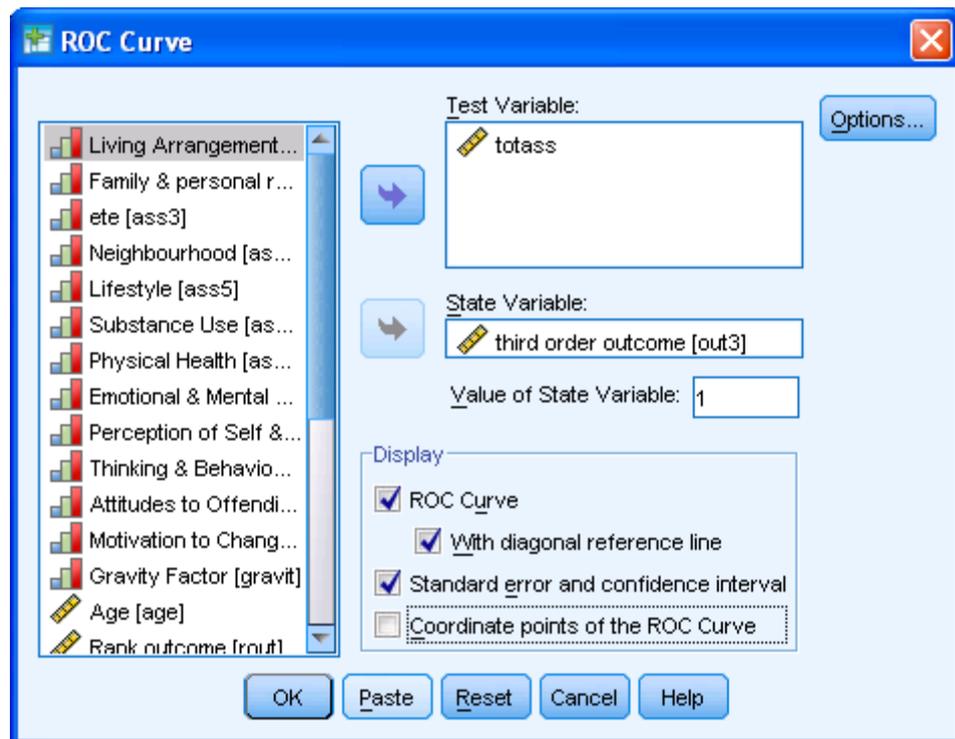
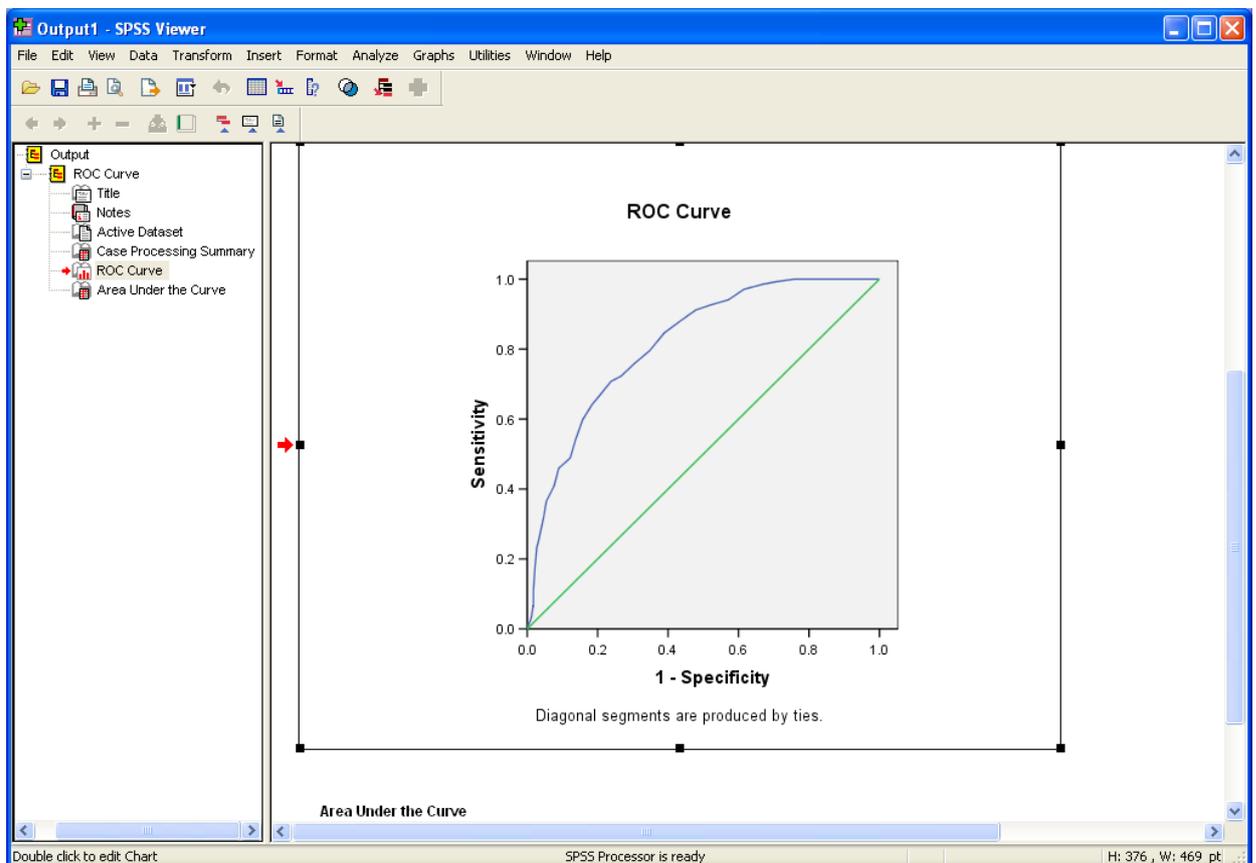


Figure 101: Output for ROC



**Table 39:** Area output  
**Area Under the Curve**  
 Test Result Variable(s):totass

Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval	
				Upper Bound
.818	.017	.000	.784	.852

The test result variable(s): totass has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

- a. Under the nonparametric assumption
- b. Null hypothesis: true area = 0.5

You can do ROC curves for one than one variable and get an immediate measure of the various classification abilities. See Figure 102 that shows, unsurprisingly, that age is a poorer classifier than asset score. You might ponder why age works at all.

SIMPLY CLEVER

**ŠKODA**  


We will turn your CV into an opportunity of a lifetime



Do you like cars? Would you like to be a part of a successful brand? We will appreciate and reward both your enthusiasm and talent. Send us your CV. You will be surprised where it can take you.

Send us your CV on [www.employerforlife.com](http://www.employerforlife.com)

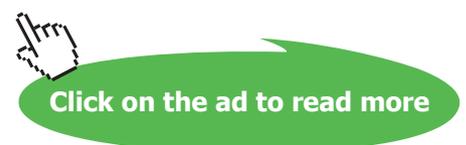
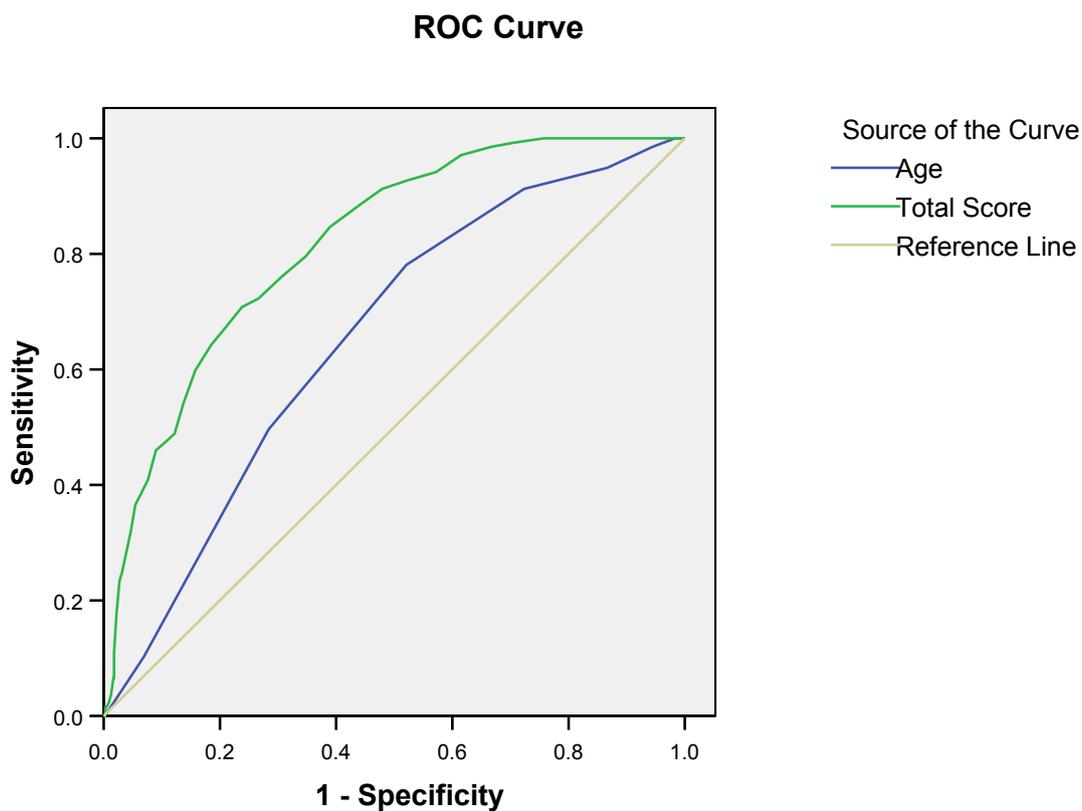


Figure 102: ROC for age and asset



### Exercise

Using the “*asset*” datafile, conduct an ROC for gravity and total asset score. Interpret the result.

### CONCLUSION

ROC is a useful method for assessing classification systems, and may be used to compare different classifiers quantitatively. It should be considered in audit and assessment tools that are numerical in nature and provide a clear output, but which are subject to probability rather than absolute certainty in the decision.

### Reading

Wikipedia has a good introduction on [http://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](http://en.wikipedia.org/wiki/Receiver_operating_characteristic)

## References

- ALTMAN, D., G & BLAND, J. M. 1994a. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ*, 308.
- ALTMAN, D., G & BLAND, J. M. 1994b. Statistics Notes: Diagnostic tests 2: predictive values *BMJ*, 309.
- ALTMAN, D., G & BLAND, J. M. 1994c. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots *BMJ*, 309.
- OVERTON-BROWN, P. & ANTHONY, D. M. 1998. Towards a partnership in care: nurses' and doctors' interpretation of extremity trauma radiology. *Journal of Advanced Nursing*, 27, 890-6.
- ROSE, D. 1995. Psychophysical methods. In: BREAKWELL, G. M., HAMMOND, S. & FIFE-SCHAW, C. (eds.) *Research methods in psychology*. London: Sage.
- SWETS, J. A. 1979. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14, 109-121.
- ZWEIG, M. H. & CAMPBELL, G. 1993. Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 39, 561-577.

# 13 Reliability

## Introduction

If a measure is repeatable then we should get the same result when we measure the same thing twice (or more). Reliability measures give a quantitative value for repeatability. For example, they may be used to assess how repeatable a machine is, or to what extent two humans give the same value.

## Stability: test-retest

This tests whether a measuring device (which could be a human) gives the same result when the measurement is repeated. Typically, the measurements could be physical measurements, and the correlation between one score and the other is calculated. A high correlation is a necessary, but not sufficient condition for stability. For example, a high correlation could be seen if the second measure were exactly half that of the first occasion. There are problems with this technique, especially where the testing of humans using, say, some form of skills test. In such a case if too little time is left between measurements the subject may simply repeat what they remember they said rather than re-compute the score, but if a large time lag is allowed it could be argued that any changes noted are real (i.e. not an artefact or error).

## Equivalence: comparing two forms of test or two observers

Reliability is a measure of how repeatable the data are that are captured; e.g. if I repeat a serum glucose assay will I get the same value for blood sugar?

It is possible to get highly repeatable results that are inaccurate, for example if my weighing scale is weighing consistently light it may always report me a kilogram below my real weight, although it may be always within 0.1 kg of the previous reading. A less reliable scale (say consecutive readings within 0.2 kg of each other) may be more accurate, with a mean value close to my real weight. So reliability is not the same as accuracy. However, if you always use the same instrument, the fact that it is reliable may be more useful than its accuracy. If I want to lose weight the scales that are always one kilogram too low will still allow me to see if I am losing or gaining. Paradoxically, these inaccurate scales are more useful than the more accurate but less reliable ones, provided I always use the same scales.

A highly reliable measuring device may not be a valid one. A micrometer is capable of measuring very tiny distances (thousandths of an inch), but this is of no use if you want to measure the length of a football pitch. Investigators sometimes choose highly accurate and/or highly reliable measures (remembering that accuracy and reliability are separate features) rather than less precise and/or less reliable measures that are actually relevant. For example you could measure poverty by measuring income, but you could be on a relatively high income and still be poor if you live in an expensive city and have a lot of commitments. For these and other reasons poverty measures are not always based on income, or at least not solely on income.

If two observers are involved in a study (inter-rater reliability), then we would want them to give the same score. However, in practice there will always be some discrepancies. For example, two lecturers will almost always give different marks to the same piece of work, although there may still be an overall trend to agreement.

One way of assessing equivalence is to calculate the number of agreements and divide by the number of possible agreements. Suppose two student nurses assess a group of patients for grade of sore, where the sore could be described as superficial or deep. To simplify the example let us assume that there are the same number of patients with superficial as with deep sores - as assessed by an expert in tissue viability (wound care). If the nurses agreed on the type of sore all the time their assessment would be said to be reliable, but in practice they will sometimes disagree. Suppose they agreed half of the time, then one would say that the assessment was 50% reliable. But in fact there are only two possibilities (superficial or deep sore) and so if the two nurses were guessing all the time you would expect them to agree at about this level.

Now suppose that the sores were graded on a scale of 1-4, where 1 indicates a reddened area and 4 indicates a deep cavity sore, with scores of 2 and 3 indicating intermediate sores. Assume there are equal numbers of sores of each level. If the two nurses agreed about half the time on the grading of the sores they are in fact doing rather better, for by chance they would agree only about 25% of the time (one time in four).

So the number of possible classifications is important - clearly it is easier to guess correctly a yes/no answer than an answer to a multiple choice question with four possible answers.

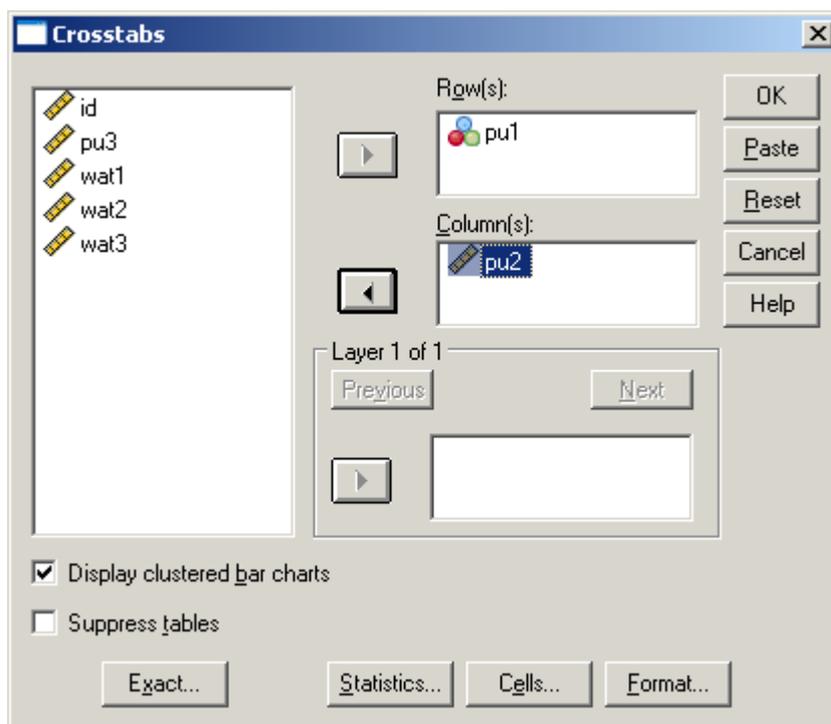
### Kappa: a test of reliability for nominal data

The presence or absence of an ulcer is nominal data, and is best measured in terms of reliability by some test like kappa. Kappa only looks at whether the two observers say the same thing.

Load dataset “*pu*”. To run a kappa analysis you select the pull-downs, **Analyze** -> **Descriptive Statistics** -> **Crosstabs** which are the same as for crosstabs, and those used for conducting chi square analysis.

You then get a dialogue box in which you put the two variables that contain the results of each assessor, which is identical again to any other cross tabulation.

Figure 103: Dialogue box for kappa analysis



Sweden Sverige

Linköping University –  
innovative, highly ranked,  
European

Interested in Computer Science? Kick-start your career  
with an English-taught master's degree.

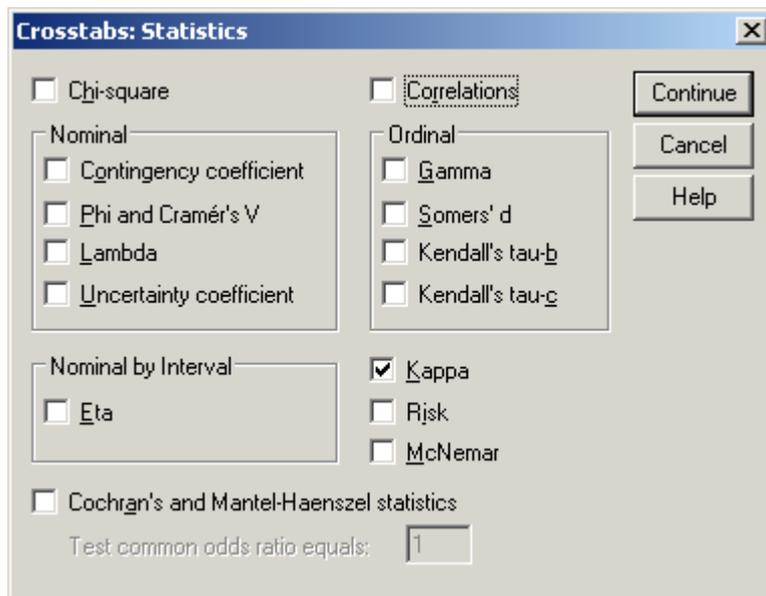
→ Click here!

li.u LINKÖPING  
UNIVERSITY



Then selecting the statistics box, instead of putting chi square, tick Kappa as in Figure 104.

**Figure 104:** Statistics dialogue box



pu1 and pu2 are presence (1) or absence (0) of pressure ulcer. The cross tabulation shows that on 36 occasions both nurses say there is no ulcer, on 10 occasions both say there is one, and in 4 cases they disagree, see Table 40. Thus the straight agreement is 40 out of 50 or 80%, which looks good.

**Table 40:** cross tabulation for the two observers of pressure ulcer presence

		pu2		Total
		0	1	
pu1	0	36	1	37
	1	3	10	13
Total		39	11	50

However suppose the results were as in Table 41. We can call this case B, and we could call the earlier case, case A. Nurse 1 says 13 of the patients have an ulcer, and nurse 2 only 1. But the raw agreement is 72% which also looks pretty good. Note however that the one occasion nurse 2 says there is an ulcer, nurse 1 does not. Therefore not once do the nurses agree on when there is an ulcer, but on many occasions they agree there is none.

**Table 41:** Cross tabulation where nurse 2 rarely picks up an ulcer

		pu2		Total
		0	1	
pu1	0	36	1	37
	1	13	0	13
Total		49	1	50

If we use kappa for the first case the value (see Table 42) the kappa is high (nearly 0.8)

**Table 42:** Kappa for “real” case, case A

	Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Measure of Agreement Kappa	.781	.104	5.557	.000
N of Valid Cases	50			

Now if case B has kappa computed the result is terrible, close to zero, indicating no agreement, despite the raw score seeming so good. This is entirely appropriate, showing case B where the few cases of ulcers identified by nurse 1 are never picked up by nurse 2.

**Table 43:** Kappa for case B

	Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Measure of Agreement Kappa	-.039	.037	-.599	.549
N of Valid Cases	50			

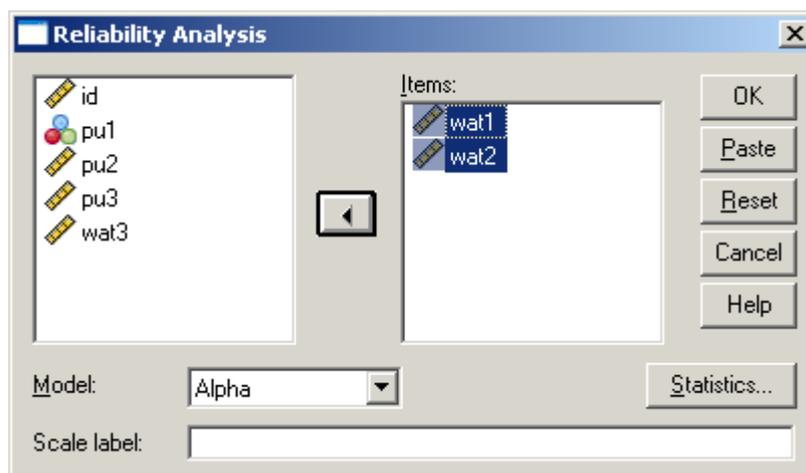
### Reliability for ordinal or interval data

Kappa is good for nominal data, but not for ordinal or interval scores. For it would be possible (e.g.) for a nurse to always score very close to a colleague and yet have a kappa of close to zero since they rarely or never score exactly the same. A different test, the intra class correlation coefficient (ICCC) is more appropriate. Having said that if there are a very few possible values in an ordinal scale, say three values, then kappa is acceptable. But for something like the Waterlow score, that can take values from 1 to over 30, it is certainly not an appropriate test.

You might think we could use correlation, but there is a problem with this, if nurse 1 always scores ten points lower than nurse 2 then there is total correlation, yet they always disagree by a large amount. Of course if you knew nurse 1 marked lower, you could account for this. The ICCC however does not suffer from this problem as it measures how close the two markers are.

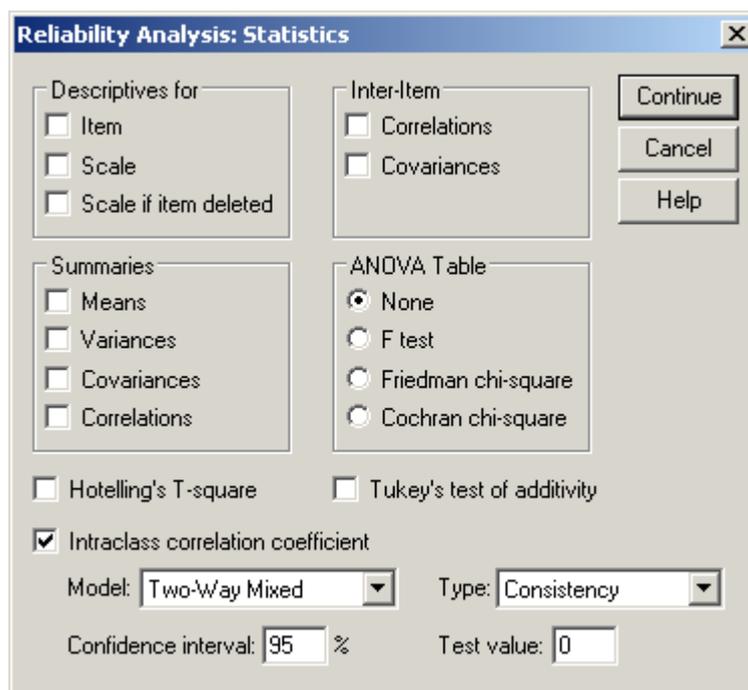
Use **Analyze -> Scale -> Reliability Analysis** then put in the two variables for the scores for each nurse, see Figure 105.

Figure 105: Dialogue box for ICC



Then you need to say which statistical analysis by clicking on the “Statistics” button when you will see a dialogue box as in Figure 106

Figure 106: Selecting ICC



ICCC can take any value between -1.0 and +1.0. A value close to +1.0 means high agreement. This example thus shows a very high ICC, see Table 44.

**Table 44:** Intraclass Correlation Coefficient

	Intraclass Correlation(a)	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.986(b)	.976	.992	144.264	49.0	49	.000
Average Measures	.993(c)	.988	.996	144.264	49.0	49	.000

The nice additional aspect of the ICC is it can be used to compare the outcomes of more than two subjects. In this way it is similar to ANOVA, and indeed is based on ANOVA in the way it is computed. It can also be used to evaluate test-retest performance. I.e. it does not need to be two separate evaluators, but one evaluator who measures more than once.

### Conclusion

Reliability can be evaluated for nominal data using Cohen’s Kappa, and for ordinal/interval data using the ICC. ICC can also be used for more than two assessors.

### Exercise

Use the “*pu*” dataset. Do a kappa and ICC on these variables to compare nurse 1 with nurse 3.

### Reading

Some website I found useful include:-

<http://core.ecu.edu/psyc/wuenschk/docs30/InterRater.doc>

[http://www.uvm.edu/~dhowell/StatPages/More\\_Stuff/icc/icc.html](http://www.uvm.edu/~dhowell/StatPages/More_Stuff/icc/icc.html)

<http://www2.chass.ncsu.edu/garson/pa765/correl.htm>

<http://www2.chass.ncsu.edu/garson/pa765/reliab.htm#intraclass>

<http://www.mega.nu/ampp/rummel/uc.htm>

<http://www.bmj.com/cgi/content/full/316/7142/1455>

# 14 Internal reliability

## Keypoints

- Internal reliability measures the extent to which components of a research tool (typically a questionnaire) are measuring the same thing

At the end of this chapter you should be able to:

- Identify methods of ensuring validity of measurements or research tools used
- Set up SPSS to measure internal reliability
- Interpret the output from SPSS for Cronbach  $\alpha$

I joined MITAS because  
I wanted **real responsibility**

The Graduate Programme  
for Engineers and Geoscientists  
[www.discovermitas.com](http://www.discovermitas.com)



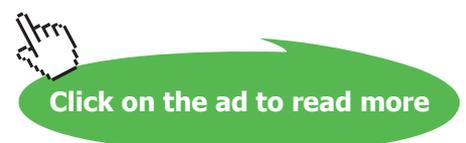


**Month 16**  
I was a construction  
supervisor in  
the North Sea  
advising and  
helping foremen  
solve problems

Real work  
International opportunities  
Three work placements







## Introduction

Many questionnaires are designed to measure a construct or constructs. For example the SF-36 is a 36 item questionnaire that measures eight health constructs:-

- limitations in physical activities because of health problems
- limitations in social activities because of physical or emotional problems
- limitations in usual role activities because of physical health problems
- bodily pain
- general mental health (psychological distress and well-being)
- limitations in usual role activities because of emotional problems
- vitality (energy and fatigue)
- general health perceptions.

How would you show that the tool actually measures the constructs? There are many aspects to this, for example you could show the questions to experts and ask if they consider the questions at face value to be valid (face validity). You could also see if the various questions that are allegedly measuring the same construct actually agree with each other. This is a different measure of validity.

Let us consider an example. If four questions purport to measure some construct, then it would be strange if three questions were highly correlated with each other, and one of them was not at all correlated with the others. That would be more consistent with the proposition that three of the questions measure the construct (or some construct) and the other is measuring something else altogether. We do need to be careful, as it is entirely possible for different things to be correlated to each other yet not measure a single construct. For example age and visual acuity are correlated, in the sense that as one ages, vision generally deteriorates. However you would be a bit bothered if your optician tested your eyes by asking your age. Clearly while related, age and visual acuity are not the same construct. But while not sufficient, it is necessary for variables to be correlated if they are said to be measuring the same construct.

Here we will look at a method of assessing whether a tool is likely to be measuring a given construct.

## Validity

Validity is a highly important, though often overlooked aspect of any study, whether quantitative (inferential or descriptive) or qualitative. Validity refers to whether you are measuring, recording, observing or assessing the appropriate items; in other words, are the data relevant?

*“I would rather be vaguely right, than precisely wrong”* - John Maynard Keynes (Cairncross, 1996).

If you wanted to measure poverty, an economist might suggest you could collect data on income, which may be available and be both reliable and accurate. However, while highly relevant to poverty, income is not the full story. Two people on the same income could be in very different positions. One might be in an area that has much cheaper living costs for example. A sociologist might suggest you consider how many items, considered by a set of focus groups to be indicative of poverty a group of people possesses. Various such alternative measures of poverty that are not purely based on raw income have been suggested (another is a subsistence measure, i.e. whether a person can afford a basket of necessary foodstuffs; and another is a relative measure based on half or some fraction of the median income). All these measures have problems but, arguably, most or all of them are more valid than simply counting income.

Validity is the determination of the extent to which an instrument actually reflects the (often abstract) construct being examined. There are several 'common sense' methods of dealing with validity. The way in which validity is defined varies between the texts.

Bryman & Cramer (Bryman and Cramer, 1997) discuss validity in the following terms.

- Face validity: where an item appears a priori to be sensible for inclusion in, say, a questionnaire.
- Construct validity. This is where the ability to measure some trait is used. If, for example, the intelligence of students studying for a doctorate compared with that of students who failed to obtain university entry did not differ according to some new scale of intelligence, then you might suspect that the scale was not valid.
- Concurrent validity. This applies when the researcher employs a test on groups known to differ and in a way relevant to the concept under study.
- Predictive validity. This applies where a test is applied and the subject group is followed up to determine whether the test can predict those subjects who will develop some condition.
- Convergent validity. This is a more general term. It includes the above measures of validity and refers to the process whereby the measure is tested for harmonization with some pre-existing measure. However, Bryman & Cramer note that using different methods is more useful than the more common application of convergent validity testing where similar methods (e.g. two questionnaires) are employed.
- Divergent validity. The idea here is that measures of different concepts should not correspond with each other; in other words, the measures discriminate.

All the above descriptions of validity are open to some question, but it is nonetheless useful to consider them. For example, it is possible for a group of experts to be wrong about the relevance of an item (content validity), as there may be systematic errors in the field. Consider the field of medicine, where a normal pineal or adrenal gland was taken to be a symptom of disease (due to the practise of doing post-mortems on the poor, who had enlarged adrenal and pineal glands). Construct validity is not necessarily disproven by doctoral students being shown to be of similar intellect to students unable to gain entry to university; perhaps the doctoral students were from a privileged background rather than inherently more intelligent. A new highly superior measuring device may not show much similarity with an inferior older device; furthermore, perhaps the older device was itself not valid. However, the least that can be said if experts consider your questionnaire items irrelevant or badly formed, or if a pilot sample cannot understand your questionnaire, or if there is no difference found between two groups where you expect a difference, or if your instrument is not even similar to a well tried and trusted device, is that you should seriously consider the validity of your instrument (be it a questionnaire, a physical measurement device or an interview question).

### Internal reliability

The validity measures described above are examples of external reliability. However, in a questionnaire, for example, you might want to see if the items within the questionnaire are measuring the same thing, i.e. whether they show internal reliability.

"I studied English for 16 years but...  
...I finally learned to speak it in just six lessons"  
Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



If you want to measure an attitude, for example, then you might have a questionnaire with several questions that aim to address this particular aspect. The reason for asking more than one question would typically be to gain a more accurate measure by averaging responses over several items, thus reducing the possibility of single idiosyncratic responses unduly affecting the measure. This is similar in concept to asking many questions in an examination (say a multiple choice examination), where all the questions concern the same subject; an average of these forms a better basis for assessment than asking one question.

You would want the questions in the attitude (or knowledge) questionnaire to show some correlation with each other, for if they did not then they are probably not measuring the same thing. A new questionnaire should be checked for internal reliability, or homogeneity. Various methods exist, of which the most common are split-half and Cronbach's  $\alpha$ . In either case a value close to zero indicates that the questions are not addressing the same item, and a value near 1 means that they are. Values over about 0.7-0.8 are generally considered adequate. If you obtain low values for internal reliability then you should probably remove some items.

Note: You might consider that what is described here as internal reliability is really internal validity, as we are exploring the relevance of items to be included in a cluster of items that are to measure some construct. I have some sympathy with this view, and in some texts you may see the measures described below classified as measures of validity. However, I am following the terminology used in both SPSS and several texts e.g. (Bryman and Cramer, 2009, Bryman and Cramer, 1997).

## Cronbach $\alpha$

Cronbach  $\alpha$  considers all the correlations among all the variables and reports the average of these. Thus it is not sensitive to the ordering of variables, and may be considered a robust measure. It is a very commonly seen test for internal reliability in the literature, and is the default in SPSS.

A coefficient of 1.0 indicates perfect agreement among the items (i.e. they all measure exactly the same thing) and logically, therefore, all but one could be removed from the instrument. A coefficient of zero indicates no agreement at all among the items. Thus the items are not measuring the same thing, and are therefore not useful if being used as a battery of variables to measure some attribute. A high coefficient (typically above 0.7) shows that the variables are probably measuring the same thing, and are plausible for use as a cluster of variables to measure an attribute.

Note that these measures of internal reliability do not prove or disprove that collections of variables are useful measures of an attribute. It would be possible to use a set of variables that are internally reliable but measure something completely different to what you wanted to measure. A set of questions meant to measure depression could, for example, be measuring anxiety.

### Example

In a survey of online course students the following questions were asked:-

**How confident are you using computers?**

**How often have you used the Web?**

**How often have you used email?**

**How often have you used Word?**

**How often have you accessed the online course?**

The question is, if this one question that could be summarised as how computer literate is this student, using several way to find out? Or is this more than one question, and if so how many?

We could do a correlation analysis of each variable with each other variable. see Table 45. What is clear is that use of the Web, email and Word are correlated with each other. But while there is a significant correlation between use of each of the web and email with confidence in using computers, the value of the correlation coefficient is small. The correlation between confidence and use of Word is not even significant. Access to the online module is highly correlated to confidence in using computers, but not to any of the other variables.

**Table 45:** Correlation table of questions on online courses

			I feel confident in using computers	The Web	Email	Word	I have accessed the PRUN 1100 online module (Blackboard site)
Spearman's rho	I feel confident in using computers	Correlation Coefficient	1.000	-.193*	-.305**	-.092	-.229**
		Sig. (2-tailed)	.	.020	.000	.272	.007
		N	145	145	143	143	140
	The Web	Correlation Coefficient	-.193*	1.000	.607**	.487**	.086
		Sig. (2-tailed)	.020	.	.000	.000	.311
		N	145	147	145	145	142
	Email	Correlation Coefficient	-.305**	.607**	1.000	.324**	.089
		Sig. (2-tailed)	.000	.000	.	.000	.298
		N	143	145	145	143	140
	Word	Correlation Coefficient	-.092	.487**	.324**	1.000	.058
		Sig. (2-tailed)	.272	.000	.000	.	.498
		N	143	145	143	146	141
	I have accessed the PRUN 1100 online module (Blackboard site)	Correlation Coefficient	-.229**	.086	.089	.058	1.000
		Sig. (2-tailed)	.007	.311	.298	.498	.
		N	140	142	140	141	145

\*. Correlation is significant at the 0.05 level (2-tailed).

\*\*. Correlation is significant at the 0.01 level (2-tailed).

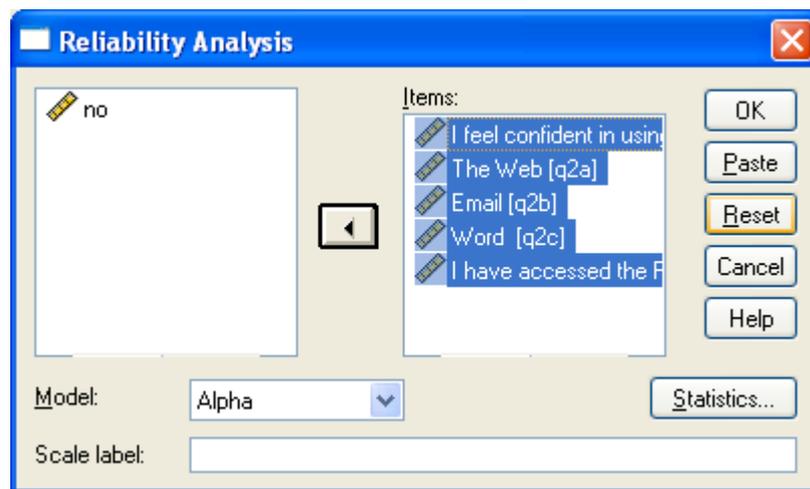
## Exercise

### Why did I use Spearman and not Pearson's correlation?

In a Cronbach  $\alpha$  analysis each variable is correlated with each other and an average correlation is computed. To set up an  $\alpha$  analysis use **Analyze** -> **Scale** -> **Reliability Analysis**.

Now we need to decide which variables to include, this is shown in Figure 107.

**Figure 107:** Dialogue box for internal reliability analysis



The output of this test is seen in Table 46. A value of 0.8 is considered acceptable for Cronbach  $\alpha$ , but here we see much lower figure of 0.5 (rounded up from 0.455). Thus by conventional standards these five questions are not measuring the same construct.

**Table 46:**Reliability Statistic

Cronbach's Alpha	N of Items
.455	5

We saw that two items were less correlated with the other variables. What happens if we remove them. Table 47 show Cronbach  $\alpha$  for three variables:-

- How often have you used the Web?
- How often have you used email?
- How often have you used Word?

**Table 47:** Reliability Statistics

Cronbach's Alpha	N of Items
.793	3

Note that by ditching two variables the value is now roughly at the level needed, 0.8.

## Conclusion

Validity assesses the relevance of the items in your research tool, i.e. whether they measure what they are supposed to measure. Internal reliability is concerned with the correct clustering of items designed to measure the same construct.

## Exercise

Load the datafile “*assets*”. Use the twelve asset subscores, gravity factor and age to compute a Cronbach  $\alpha$ . Repeat this analysis but with age added, and then again with gravity added and age removed. What do you infer from these analyses.

## References

BRYMAN, A. & CRAMER, D. 1997. *Quantitative data analysis*, London, Routledge.

BRYMAN, A. & CRAMER, D. 2009. *Quantitative data analysis with SPSS 14, 15 and 16 : a guide for social scientists / Alan Bryman and Duncan Cramer.*, Hove Routledge.

CAIRNCROSS, A. 1996. *Economist* 20 April.

# 15 Factor analysis

## Key points

- Principal components analysis (PCA) is a special case of factor analysis (FA)
- PCA/FA is a method of reducing the number of variables or dimensions used to describe a dataset
- PCA/FA is typically used to explore data

At the end of this chapter you should be able to:

- Set up a dialogue box in SPSS to generate a PCA analysis
- Interpret the output of an SPSS session for PCA/FA

## Introduction

Measurements taken on a subject may be correlated. This means that there is a certain amount of redundancy in the data. In most real world problems there comes a point where adding a new variable tells us very little further about the subject. In extreme cases a variable may add no new information at all, in which case it is not necessary or even useful. However it is not always obvious which variables contain the most information. This unit discusses methods of reducing data to the minimum number of variables, while keeping the maximum amount of information.

In chapter 14 we evaluated whether an instrument measured a construct, and decided that in at least one case the instrument measured more than one construct, and removing certain items made the internal reliability (Cronbach's  $\alpha$ ) higher. But suppose you have a set of data and want to know *how many* constructs it contains, rather than measuring how well it measures one construct?

One way is using factor analysis, or a particular case of factor analysis, principal components analysis.

PCA and the related method of FA are methods of data reduction. A simple and straightforward account is given by Hammond (1995) and a more detailed introduction to its use in SPSS is given by Field (2009). PCA is used to remove data that are either irrelevant or contain little useful information for the task at hand. It is best illustrated by an example.

Consider collecting data to describe the size of a human body. One could take a (potentially infinitely) large number of cross-sections, ending up with a huge data set that defines the individual body, but this is impractical in most situations. Thus one might choose anatomical lengths, or other data such as

1. left arm
2. right arm
3. left inside leg
4. right inside leg
5. head circumference

- 6. waist
- 7. chest
- 8. hips
- 9. weight
- 10. age
- 11. sex

But do we need all these items? Data may be:

- Redundant. The left and right sides of the body are virtually the same, so why measure both?
- Correlated. Chest and waist size are correlated, so if you know one a plausible guess can be made for the other
- Irrelevant. Possibly age is not useful if we only consider adults.

If irrelevant redundant data are removed then no loss of accuracy occurs. If correlated data are removed some accuracy is lost, but this may be acceptable. PCA is a method by which data that are not working for us are removed.

There is some confusion in the texts between PCA and FA. The latter is a method of obtaining meaningful factors from a data set. These factors are produced by combining the variables of the original dataset, and are typically fewer in number than the original variables. PCA is one specific method of obtaining factors. It is the method favoured by most researchers using FA, but there are other methods. This chapter will concentrate on FA using principal components and, unless stated to the contrary, in this chapter the terms FA and PCA will be considered as synonymous (as is often the case in the literature).

In the past four years we have drilled  
**89,000 km**  
 That's more than **twice** around the world.

**Who are we?**  
 We are the world's largest oilfield services company<sup>1</sup>. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

**Who are we looking for?**  
 Every year, we need thousands of graduates to begin dynamic careers in the following domains:

- Engineering, Research and Operations
- Geoscience and Petrotechnical
- Commercial and Business

**What will you be?**

**Schlumberger**

careers.slb.com

<sup>1</sup>Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.



PCA is related to correlation. In fact the analysis starts by producing a correlation matrix of all the variables. If two variables are highly correlated then one can probably be removed and still keep most of the information. By this I mean that, as one variable seems to be largely predictable from another, you only need to keep one. For example, if a particular serum value A was almost always about twice another serum value B, then reporting both is largely unnecessary, because if you know B then  $A = 2B$ . You will only be able to get back all the lost data if the variables are totally correlated; however, a small loss of data is often acceptable, and an absolute correlation of less than 1.0 (i.e.  $> -1$  and  $< +1$ ) that is close to unity (i.e. much nearer to 1.0 or -1.0 than 0.0) may allow you to remove a variable with little loss of information. For example, a measurement of the left arm is likely to be very slightly different from one of the right arm, but we would accept either one, or the mean value of both as an accurate measure of arm length for most purposes.

One variable may be partially correlated with several other variables. Thus a variable may not be strongly correlated with one variable, but is predictable from several taken together.

We usually view the variables in PCA as dimensions. This is analogous with spatial dimensions. Indeed the analogy is quite helpful in understanding PCA. I will now give some examples of physical dimensional systems (i.e. three-dimensional space, two-dimensional areas, one-dimensional lines), not because they are realistic representations of what we will perform PCA on, but because conceptually they are easy to understand. I will then extend the concept to other systems where the dimensions are not physical, but relate to any variables.

Suppose I am flying in a plane. To know my location you would need three dimensions: longitude, latitude and altitude. However, if I am sailing you would not need my altitude as it will always be (by the definition of sea level) zero. Thus my position is predicted from two dimensions: longitude and latitude. Now suppose I am driving along the M1 (a motorway linking London and the North of England, which passes near to my home Leicester) and I ask my passenger how far we are from home, they would only need one figure: the miles or kilometres to (in my case) Leicester. I still have three dimensions of travel, but only one is needed in this case, as the other two are constrained to be along the route of the motorway.

The above are pretty straightforward examples of dimension. However what if I were walking over the hilly park near my home. Assume you do not have a map with contours, if you knew my longitude and latitude very accurately you still do not know exactly where I am, as I may be up a hill or down a valley. However, your estimate of my exact position would be very close, as Leicestershire is not mountainous and, at most, I will be a couple of hundred metres away.

A further example is if you only had the time I had been gone to estimate where on a walk I might be. Since, from experience, you know I usually walk at a brisk rate of 3 miles per hour, you could use this to predict where I could be. However, you could be wrong for a variety of reasons, as I may walk a bit slower or faster, depending on the weather, my mood, etc. So time is correlated with distance, but the correlation is not entirely reliable. One could nevertheless use such a one-dimensional data set (time gone) to give an approximation of where I am. If I were known to stop for a pint in the village pub on the way home if I were thirsty, then knowing how thirsty I was to start with may add some useful information to the estimate of when I will return. However, it is possible that this information is not an accurate or sensitive measure, as it may be more dependent on who I see in the pub when I get there, which determines whether or not I stay.

The above examples are contrived and possibly rather silly, but are used to show:

1. A system may be under-defined. If I want to describe a student's academic ability I may have examination and course marks, but is it clear that I have enough? Suppose I only have one mark. It may not be representative of the student's work. This mark could be seen as one dimension of the student's work. This is a similar problem to only knowing when I left the house for my walk. However, even though the data are not ideal, they probably serve as the best estimate I have, and so are not useless.
2. A system may be over-defined. Even if three dimensions are available, they are not necessarily all needed. This could be extended to any three (or any number of) variables, not just physical dimensions. In the student assessment problem, I may have several marks from different courses and over many years. Therefore, adding another mark from a new assessed work may not make much difference to my assessment of the student, so why am I collecting such data?
3. The dimensions should be independent. In working out a location in space, three properly chosen dimensions are needed. These are the three coordinates (length, breadth and height; or longitude, latitude and altitude). It is not possible to work out longitude from latitude or altitude, or any combination of these. If it were, then at least one dimension would be redundant, and could be disposed of.

In the sailing example, being given my altitude is useless as it is redundant as it will always be zero; a two-dimensional system (longitude and latitude) is equally as good as one that adds altitude. So if a parameter is constant, for example all the sample are of a similar age (e.g. a sample of first-year undergraduates), then the data on age are not likely to tell us anything extra. In the walking example, a two-dimensional system (longitude and latitude) is almost as good as the three-dimensional one. In the example where you only know how long I have been gone, a one-dimensional system (time) is a lot better than nothing, adding more data (thirst) helps, but may not be enough to be very useful. Thus, if we were paying for information, we would buy the time data, and only purchase the thirst data if it were very cheap.

A real-world example of data that is amenable to PCA analysis is compression of image data. There are many ways of reducing the amount of image data. You could simply send only half of the available data, which will give a more grainy image. However rather than simply throwing away half the data, you might try to work out which data contain the most information, and send the half (or some proportion) that contains the most. Images are divided into pixels, or points of light intensity. If an image is 100 pixels high and 100 pixels wide it contains 10 000 pixels, each of which can be viewed as a dimension. So we have 10 000 dimensions, where each dimension can be any value between (say) zero and 255. You could remove alternate rows and columns of pixels to reduce the data set, in this case to a quarter of the original size. You would probably get a better picture if you took some form of average of four adjacent points (in fact the median is usually used as it preserves lines better than the mean, this is known as median filtering). In this case we have reduced the dimensions from 10,000 to 2500, thus preserving (we hope) more information by combining the variables rather than simply throwing three-quarters of the data away. However, is median filtering optimal?

PCA would approach this problem by looking at the 10 000 dimensions of a sample of many pictures and work out what the correlations are among the pixels. It would then work out the best combination of variables to form a smaller number of new variables from the original data set that contain as much information as possible. In practice there are other more efficient algorithms used to compress image data, but PCA *could* be used.

PCA is indicated when:

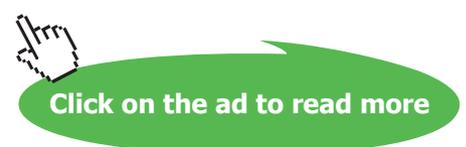
- there are many variables that are correlated
- the aim is to reduce the number of variables used to describe the data (i.e. when data compression is useful)

### Technique

A correlation matrix is computed for all the variables that form the data set. It can be shown that a set of new variables, called eigenvectors, can be created that completely describe the correlation matrix, but these new variables are independent of each other (i.e. they are not correlated with each other). In a data set where some of the variables are totally redundant, the eigenvectors will be fewer in number than the original variables, but will contain exactly the same information. In general, however, for a data set with  $N$  variables, there will be, at most,  $N$  eigenvectors. The amount of information in each eigenvector, as measured by the amount of variance in the data set it describes, is not in general the same; some eigenvectors explain more variance than others, and may be said to contain more information.

**WHILE YOU WERE SLEEPING...**

[www.fuqua.duke.edu/whileyouweresleeping](http://www.fuqua.duke.edu/whileyouweresleeping)



Each eigenvector has an associated eigenvalue, which is the measure of the amount of variance in the data set it describes. If all the eigenvalues are summed, then the ratio of each eigenvalue to this sum is the proportion of the variance for which it accounts. Thus a sensible strategy for reducing the number of the variables in the data set is to take the first few eigenvectors with the highest eigenvalues. The algorithm for PCA can be described as follows:

1. Obtain eigenvectors of the correlation matrix
2. Arrange eigenvectors in decreasing order of their associated eigenvalues
3. Take the most important eigenvectors to describe the data.

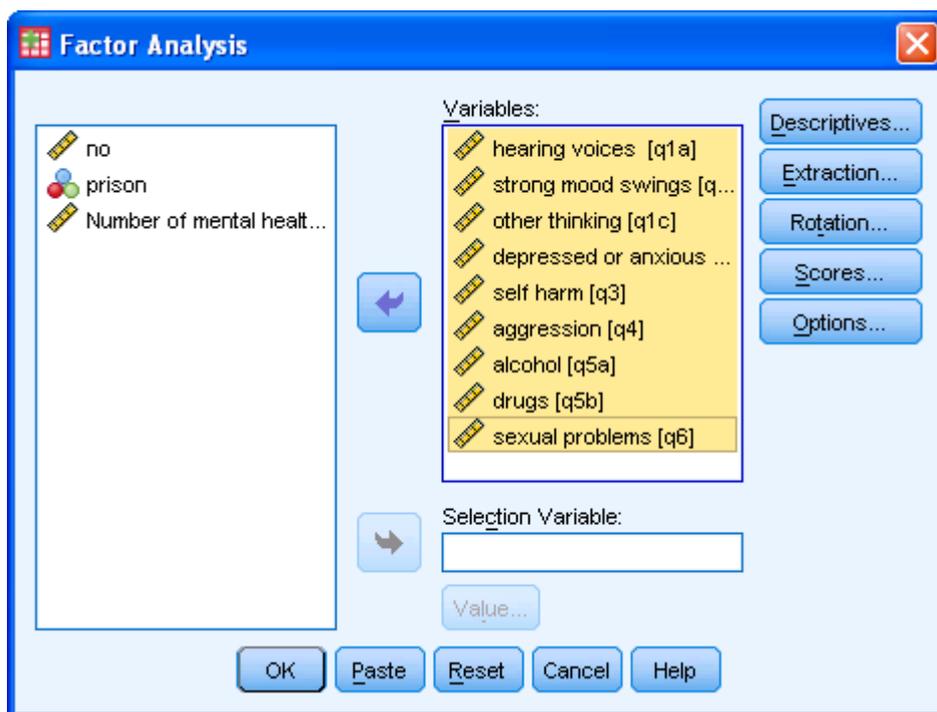
While these steps can be achieved quite simply using mathematical packages such as Matlab, SPSS will do the analysis automatically for you.

### An example of factor analysis

Using the prison database, to conduct a factor analysis we use **Analyze -> Dimension Reduction -> Factor**.

This will give us a dialogue box as in Figure 108. Here I have asked for all the symptoms to be included.

Figure 108: Dialogue box for factor analysis



This results in a table (Table 48). By default SPSS considers all factors with an eigenvalue more than 1.0 (this is called Kaiser’s criterion). Here there are two, between them accounting for more than half of the variance, with less than half accounted for by the other seven.

**Table 48:** Total Variance Explained  
Extraction Method: Principal Component Analysis.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.931	43.682	43.682	3.931	43.682	43.682
2	1.114	12.383	56.065	1.114	12.383	56.065
3	.847	9.416	65.481			
4	.729	8.099	73.581			
5	.668	7.425	81.005			
6	.545	6.052	87.057			
7	.432	4.796	91.853			
8	.413	4.593	96.446			
9	.320	3.554	100.000			

Extraction Method: Principal Component Analysis.

The loadings of the first two factors are shown in Table 49. I would interpret this as showing drugs and alcohol (more loaded on factor 2) as being separate from the other symptoms. However all symptoms load on factor 1, which seems to act as an average of all symptoms.

**Table 49:** Component Matrix

	Component	
	1	2
hearing voices	.717	-.138
strong mood swings	.742	-.006
other thinking	.777	-.205
depressed or anxious	.711	-.249
self harm	.704	-.202
Aggression	.665	.120
Alcohol	.432	.704
Drugs	.493	.637
sexual problems	.626	-.187

Extraction Method: Principal Component Analysis.

a 2 components extracted.

Some researchers prefer to use a scree plot to see how many factors there are. The screen plot requested in Figure 109 and is shown in Figure 109: Dialogue box for extraction

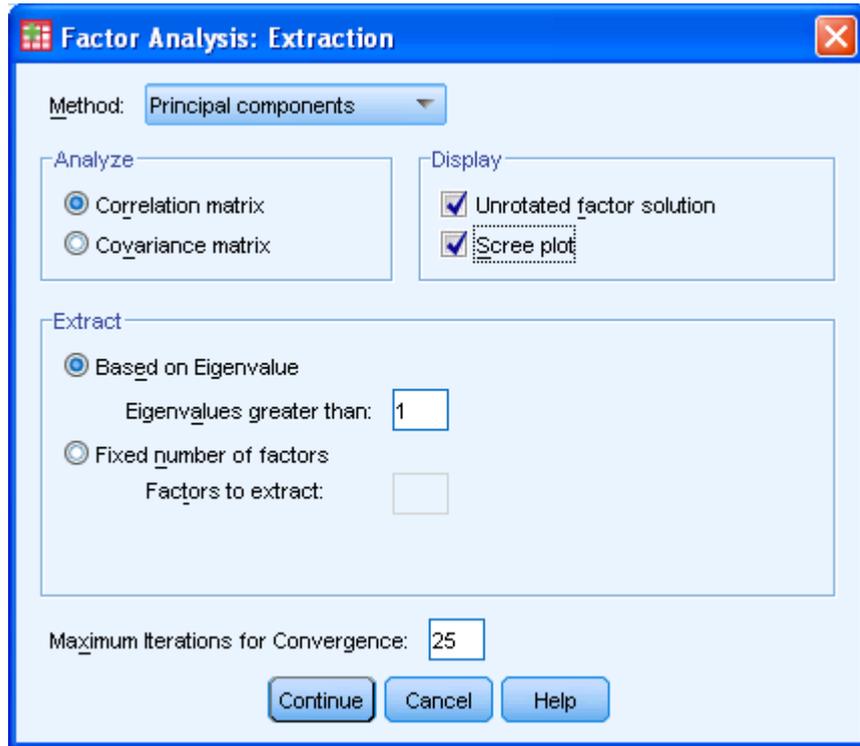


Figure 110. This shows that factor 1 is by far the most important, and there may arguably be another factor, factor 2.

Excellent Economics and Business programmes at:



**university of  
 groningen**





**“The perfect start  
 of a successful,  
 international career.”**

[www.rug.nl/feb/education](http://www.rug.nl/feb/education)

**CLICK HERE**  
 to discover why both socially  
 and academically the University  
 of Groningen is one of the best  
 places for a student to be



Figure 109: Dialogue box for extraction

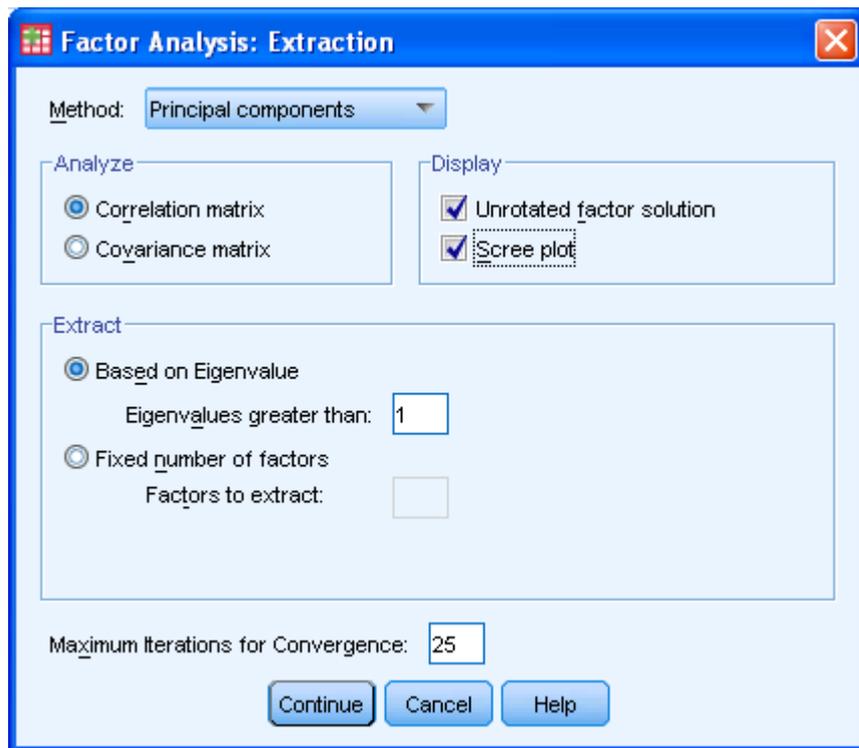
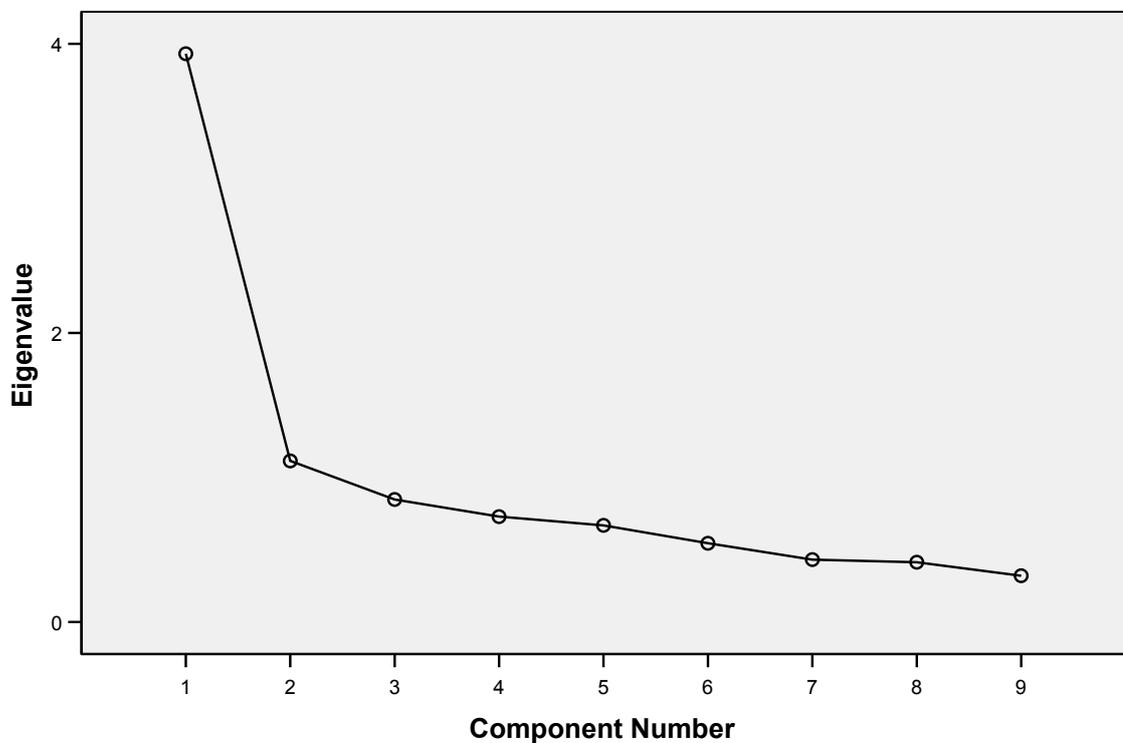


Figure 110: Scree plot

### Scree Plot



However while a scree plot is probably better than Kaiser, both tend to give too many factors. I would advise the method proposed by Pallant (2010) of parallel analysis.

Why do Kaiser's criterion and scree plots keep too many factors. How do we know this? If you create a load of random data and compute the factors in it, then all the factors should be rubbish, since factors of random data should not be meaningful. If you look at the highest eigenvalue then this should be where rubbish factors start. Just to be sure, let us take many random sets and take the average highest eigenvalue. You have now conducted parallel analysis. Parallel analysis usually returns fewer factors than the other two methods.

The Monte Carlo ([www.allenandunwin.com/spss3/Files/MonteCarloPA.zip](http://www.allenandunwin.com/spss3/Files/MonteCarloPA.zip)) program asks you for the number of variables, subjects and replications. Pallant suggests the replications be set to 1000 (the number of random sets) and number of variables here is nine, with 494 subjects (see Figure 111), we get Figure 112. This indicates any factor with an eigenvalue below 1.21 is likely to be "noise". This means that my interpretation of factor two might be pointless, since it is quite likely to be simply random nonsense. Pity as it seemed to make sense! But that's life.

Figure 111: Monte Carlo Program

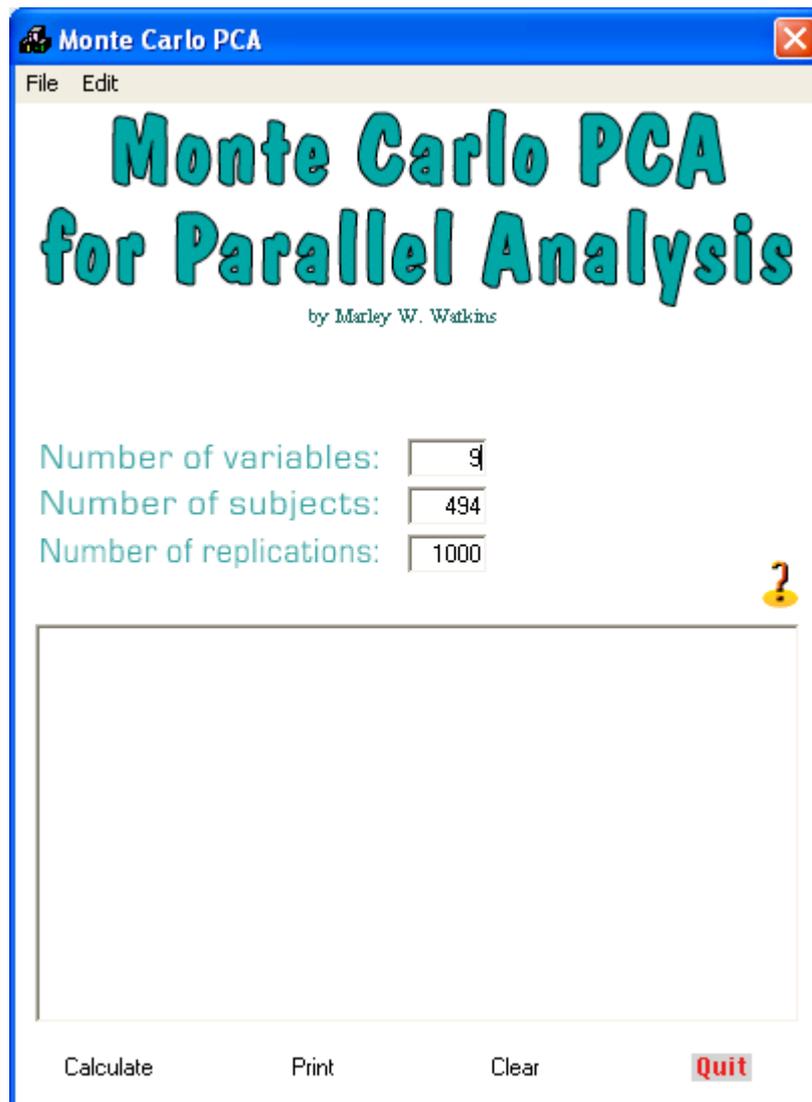
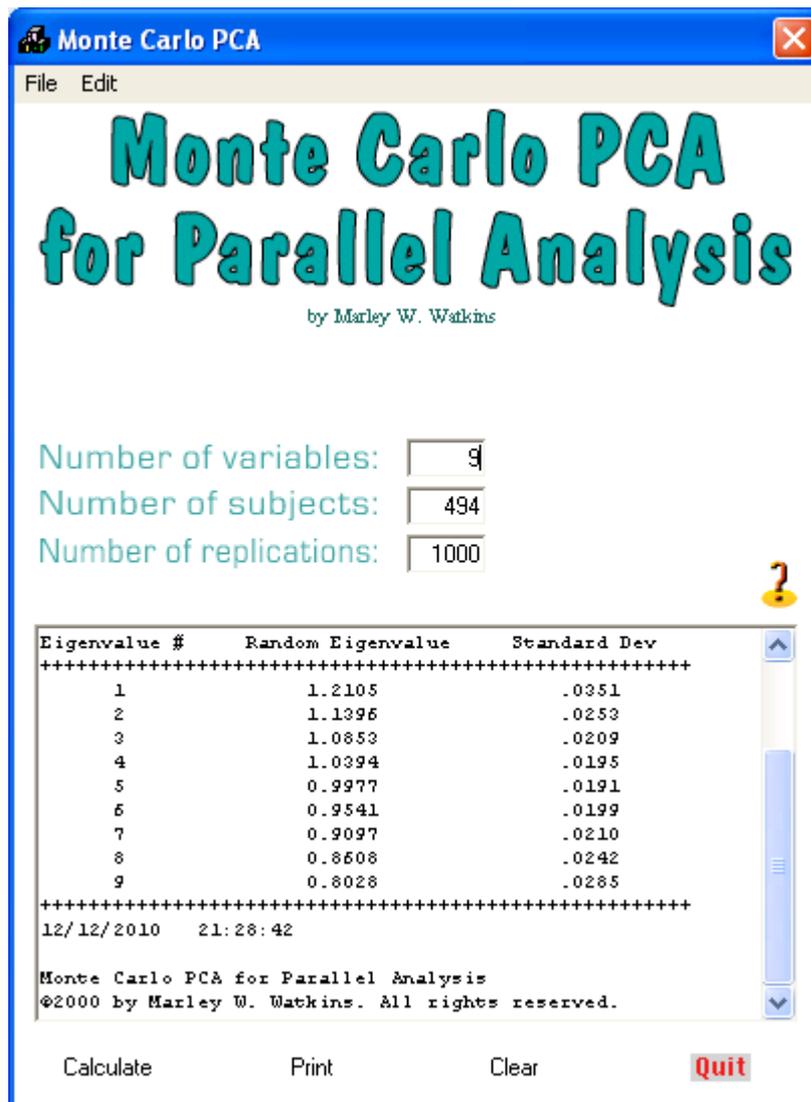
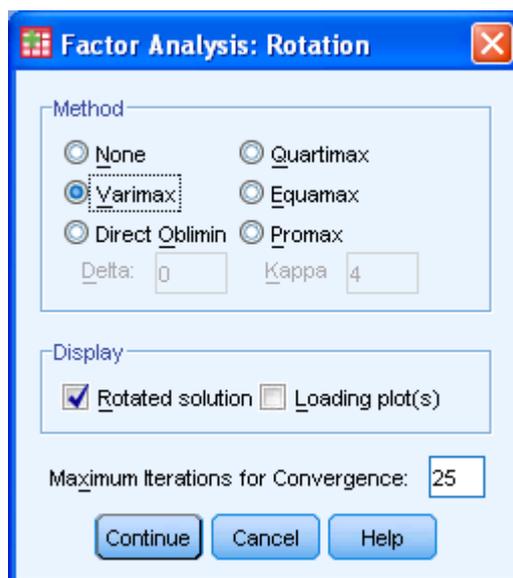


Figure 112: Results from Monte Carlo program



The default position for SPSS is straightforward principal components analysis (PCA) whereby the derived factors are all orthogonal to each other, i.e. each factor is independent of each other factor. However sometimes these principal components do not have obvious interpretations, and we may relax the strict requirement for orthogonality to create more meaningful factors. These *rotations* attempt to create factors that have variables load or not load more obviously than with strict PCA. A common method is varimax, and this is obtained as in Figure 109.

**Figure 113:** Dialogue box for rotation



This in addition to the matrix gives a rotated matrix, this is shown in Table 50. This shows an even more clear distinction between alcohol and drugs, and the other variables.

**Table 50:** Rotated Component Matrix

	Component	
	1	2
Hearing voices	.708	.176
strong mood swings	.675	.307
other thinking	.791	.141
depressed or anxious	.750	.073
self harm	.724	.113
Aggression	.552	.389
Alcohol	.096	.820
Drugs	.180	.786
sexual problems	.646	.093

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 3 iterations.

## Conclusion

Factor analysis is used to reduce variables in a dataset to a smaller number of variables. While sometimes the factors derived are difficult to interpret, and the exact number of factors can be difficult to ascertain, it remains a useful method to explore data. It may also be used in a confirmatory analysis, to determine if the factors identified in one dataset remain those in a new dataset. For example an instrument used in one culture may not work in the same way in another

### Exercise

Consider the datafile “assets”. Conduct a factor analysis of the twelve asset sub-scores using PCA. Decide how many factors there are. Interpret the factors.

### References

FIELD, A. 2009. *Discovering statistics using SPSS (and sex and drugs and rock ‘n’ roll)*, London, Sage.

HAMMOND, S. 1995. Introduction to multivariate data analysis. *In*: BRAKWELL, G., HARNMOND, S. & FIFE-SCHAW, C. (eds.) *Research methods in psychology*. London: Sage.

PALLANT, J. 2007. *Title SPSS survival manual : a step by step guide to data analysis using SPSS for Windows* Maidenhead Open University Press.

.....Alcatel-Lucent 

[www.alcatel-lucent.com/careers](http://www.alcatel-lucent.com/careers)

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



# 16 Linear Regression

## Key points

- Regression is based on correlation
- Regression assumes linearity
- Regression gives the formula for a linear relationship between two or more variables
- Regression may be used to predict one variable from another

At the end of this unit you should be able to:-

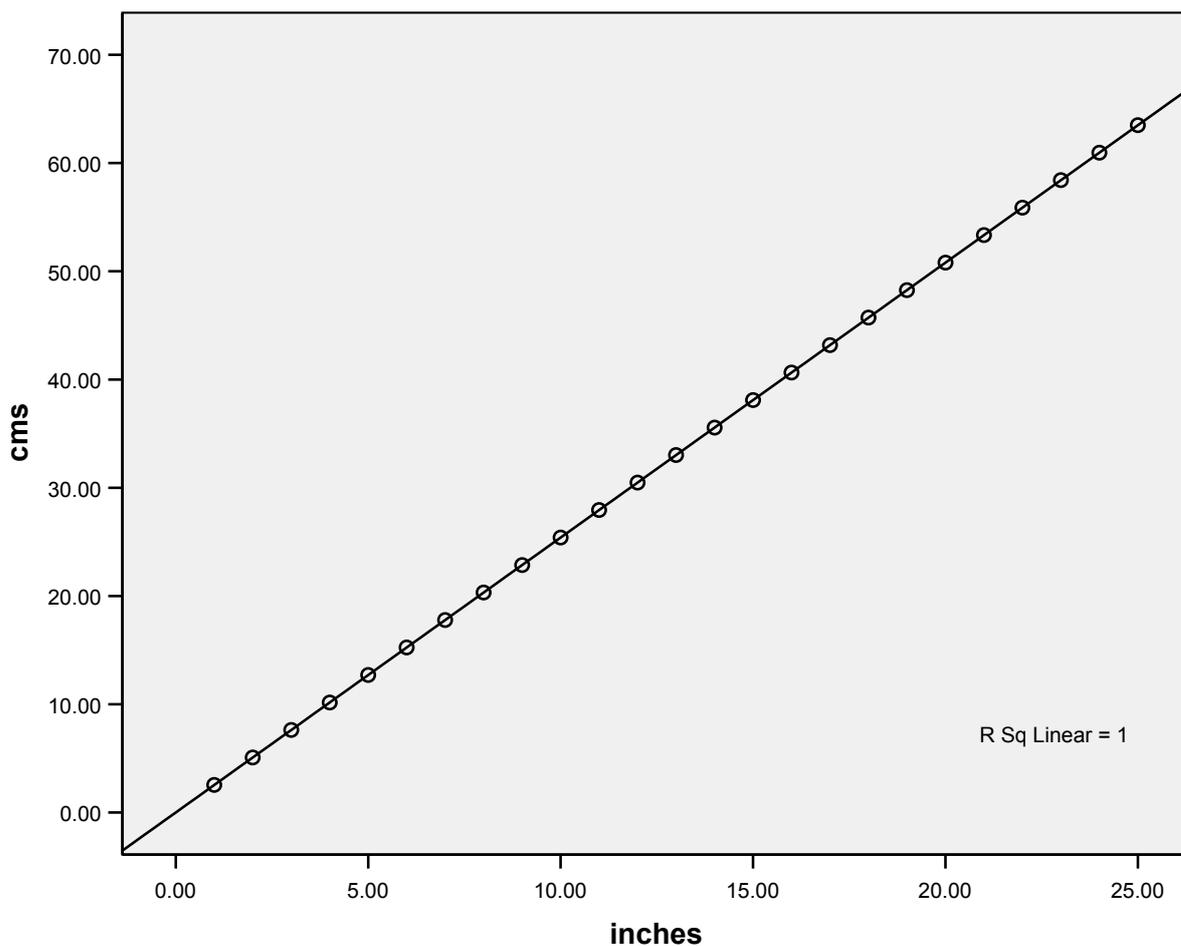
- State the conditions for regression analysis
- Set up SPSS for linear regression
- Interpret the results from the SPSS output for linear regression

## Introduction

Suppose you want to be able to predict one variable given one or more other variables. For example the severity of sentencing we know has something to do with age, asset score and gravity of the offence. Could we predict the likely sentence given these values? Linear regression is a technique that attempts to give the best guess for one variable given several others. For example an older offender with multiple problems (drug dependency, difficulty in dealing with aggression) and having committed a serious offence is clearly more likely to receive a custodial sentence than a small child found shoplifting on the first occasion.

Simple linear regression explores possible relationships among data of a linear nature, i.e. where the relationship can be modelled as a straight line. As an example of a linear relationship consider Figure 114 which shows a plot of inches against centimetres. For any value in inches, you can read off the corresponding value in centimetres, and vice versa,.

**Figure 114:** Plot of inches against centimetres



You can see that 10 inches (inches are on the x axis) corresponds to about 25 centimetres (on the y axis). In fact we do not even need the graph, as the following formula gives the exact relationship:

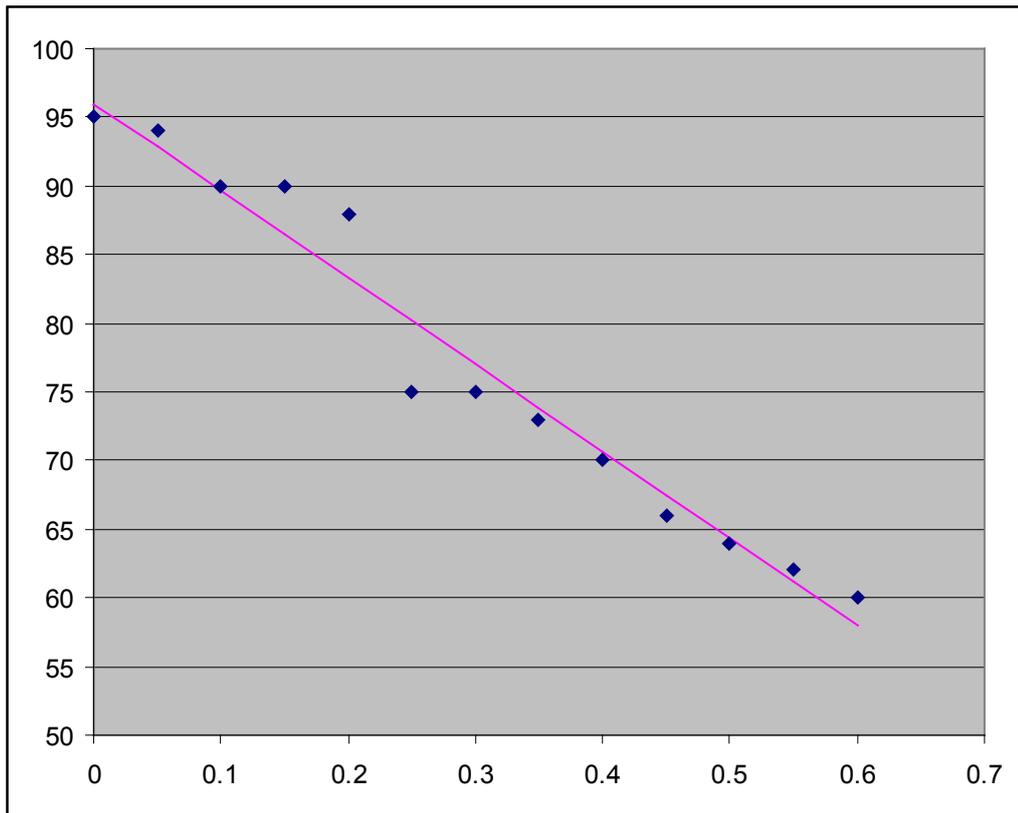
$$\text{Centimetres} = 2.54 \times \text{inches}$$

Note the following aspects of this graph:

- The relationship is exact, if you know the value in inches then you know the precise value in centimetres
- Zero inches is also zero centimetres, so the line goes through the origin of the graph ( $x = 0, y = 0$ ).
- The relationship is a straight line; another word for this is linear.
- The relationship is positive, i.e. as the number of inches increases so does the number of centimetres.

It is possible to construct relationships that obey none of the above. For example, consider the relationship between the dose of a drug that slows the heart beat and pulse rate, see Figure 115.

**Figure 115:** Dose of (fictitious) drug (x axis) against pulse (y axis)



- The relationship is not exact, you cannot predict with certainty that a given dose will result in a particular pulse rate. This is because a patient's pulse depends on other factors (exercise, stress, time of day, etc.). However, the drug does, on average, exert a measurable effect; it is simply that we cannot be absolutely sure of its precise effect.
- If no drug is given the heart still beats, so if the relationship could be modelled as a straight line it would not go through the origin, as zero drug is not associated with zero heart beat (fortunately).
- The relationship need not be linear, and definitely will not be outside a certain range. For if the dose takes the heart beat too low the patient may die, resulting in a pulse rate of zero, but a still higher dose cannot give a lower pulse rate.
- If the drug dose is increased the heart beat slows down more. The relationship is negative.

The relationship is shown graphically in Figure 115. Note that three of the above features are evident by inspection:

- The relationship is not exact. While there is an overall trend observable (as the drug dose is increased the heart rate decreases (shown by the straight line)), the actual drug dose and pulse values (the small squares) do not lie on the line, but are scattered around it.
- The line that best fits the relationship (the straight line) does not go through the origin.
- The slope of the line is negative, i.e. as the dose is increased the pulse rate decreases.

You will probably recognize the graph in Figure 115 as a scatterplot. It shows that the two variables are correlated. Regression is based on correlation, and uses the correlation coefficient to produce the straight line that gives the best guess of what one variable will be, given the value of the other. For example, if no drug is given we would expect the heart rate to be 96 beats/min; in fact we record it was 95. We would guess that a dose of 0.4 mg would result in a heart rate of about 70 beats/min, but we may record a different pulse rate. However, it is clear that 70 beats/min would be a better estimate than (say) 110 beats/min, as none of the patients with such high pulses are receiving doses as high as 0.4 mg: the three patients with a pulse rate of 100-110 beats/min are receiving doses of 0.05-0.15 mg, whereas patients with a pulse rate of 60-70 beats/min are receiving doses above 0.3 mg. Therefore, while the prediction is inexact, it is still helpful.

## Linear regression

In the above examples you may have wondered how we decided where to draw the straight line through the points. You may also wonder whether it is even appropriate to draw such a line, in other words are we deluding ourselves that there is a relationship between the drug dose and pulse rate? It is the technique of linear regression that answers both these questions.

Linear regression has three main outputs:

- The probability that the straight line that describes the relationship between the two variables is purely down to chance. I.e. a low  $p$  value shows it is probably a real effect
- The slope of the line
- The point where the line crosses the  $x$  axis

If you know where the line crosses the x axis, and how steep (the slope) the line is, you can draw the line. There is a very simple equation that defines any straight line:

$$y = (\text{Slope times } x) + \text{Constant}$$

where the constant is the value of y at which the line crosses the y axis (Le. the value of y when  $x = 0$ ); this is also called the intercept. You will often see this equation written as

$$y = mx + c,$$

where c is the constant and m is the slope.

In the inches and centimetres example (Fig. 7.1), the equation is

$$\text{Centimetres} = 2.54 \times \text{Inches} + 0$$

i.e. the slope of the line is 2.54, and the constant is zero. While you may find reading the value off the graph easier than working it out from the formula, in fact the equation and the graph are identical in the information they contain.

## Conditions

Regression makes the following requirements of the data:

- Normality. Regression is a parametric test, and assumes that the data are normally distributed.
- Linearity. A linear relation exists between the dependent and independent variables. Thus it is sensible to model the relationship as a straight line.
- Independence. The data should have been drawn independently. In the drug dose versus pulse rate example above, several pulse readings from the same person should not have been used.
- Unexplained variance. The portion of the variance of the dependent variable not explained by the independent variable is due to random error and follows a normal distribution. The regression module in SPSS allows the option of saving the residuals (the difference between the predicted values according to the regression equation and the actual values). These residuals can be checked for normality using the Kolmogorov-Smirnov test (for example) or can be shown as a histogram. The data should not exhibit heteroscedasticity, where the scatter around the line of best fit varies a lot as one proceeds along the line. Figure 115 shows no apparent change in scatter as you move along the best fit, and thus shows little or no heteroscedasticity.

## LIMITATIONS TO REGRESSION

Regression is a powerful technique for prediction. There are, however, some limitations to the technique:

- The prediction based on linear regression is not necessarily correct, or even guaranteed to be close. It is simply the best guess on the information available.
- The fact that two or more variables are correlated does not necessarily imply causation. You should have a theoretical rationale for how one variable might predict another before using regression to calculate a predicted value.
- The regression equation may fail outside of the range of values from which it was constructed. For example, based on the regression analysis of for the heart drug where the values of the dose ranged from 0 to 0.5 mg, a very high drug dose of (say) 5 mg may not produce 10 times the reduction in pulse over the pulse that a dose of 0.5 mg produces (just as well).

## Example predicting sentence outcome

Using the asset data we have considered now a few times, let us see if we can predict the sentence outcome from the asset sub-scores alone. After **Regression** -> **Linear** Figure 116 show you how to choose linear regression and pick the variable you are trying to predict (dependent variable) and the variables used to predict it (independent variables).

**Maastricht University** *Leading in Learning!*

**Join the best at the Maastricht University School of Business and Economics!**

**Top master's programmes**

- 33<sup>rd</sup> place Financial Times worldwide ranking: MSc International Business
- 1<sup>st</sup> place: MSc International Business
- 1<sup>st</sup> place: MSc Financial Economics
- 2<sup>nd</sup> place: MSc Management of Learning
- 2<sup>nd</sup> place: MSc Economics
- 2<sup>nd</sup> place: MSc Econometrics and Operations Research
- 2<sup>nd</sup> place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

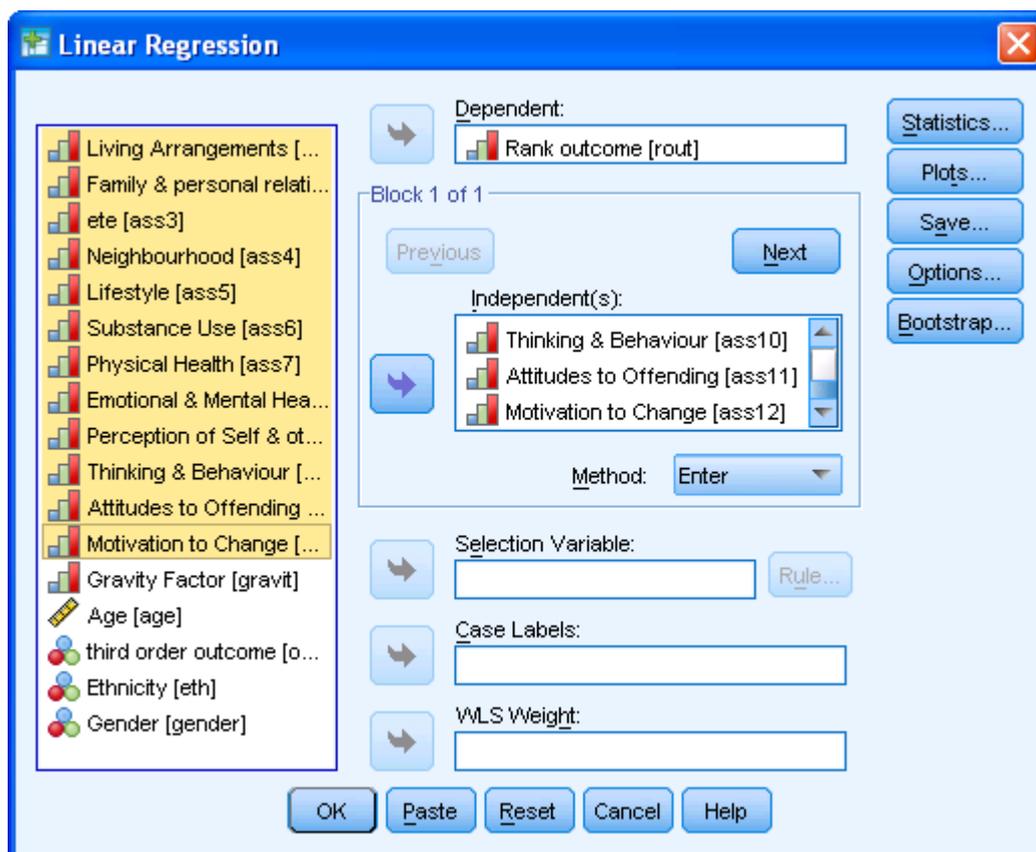
**Maastricht University is the best specialist university in the Netherlands (Elsevier)**

**Visit us and find out why we are the best!**  
**Master's Open Day: 22 February 2014**

[www.mastersopenday.nl](http://www.mastersopenday.nl)



**Figure 116:** Dialogue box for linear regression



Having entered the variables we get the output as seen in Table 51

**Table 51:** Selection of variables for entry  
**Variables Entered/Removed<sup>b</sup>**

Model	Variables Entered	Variables Removed	Method
1	Motivation to Change, Physical Health, Emotional & Mental Health, Neighbourhood, Substance Use, Thinking & Behaviour, ete, Perception of Self & others, Living Arrangements, Attitudes to Offending, Lifestyle, Family & personal relationships <sup>a</sup>	.	Enter

a. All requested variables entered.

b. Dependent Variable: Rank outcome

The R Square ( $R^2$ ) figure gives an estimate of the variability accounted for by the regression. Here 29.7% of the variance is accounted for by the regression equation, so the remaining 70.3% are not explained by these variables. See Table 52

**Table 52: Model Summary**

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.545a	.297	.290	.593

a. Predictors: (Constant), Motivation to Change, Physical Health, Emotional & Mental Health, Neighbourhood, Substance Use, Thinking & Behaviour, etc, Perception of Self & others, Living Arrangements, Attitudes to Offending, Lifestyle, Family & personal relationships

The ANOVA analysis shows whether the overall regression equation is significant, clearly in this case it is, see Table 53.

**Table 53: ANOVA**

ANOVA <sup>b</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	159.703	12	13.309	37.865	.000 <sup>a</sup>
	Residual	377.133	1073	.351		
	Total	536.836	1085			

a. Predictors: (Constant), Motivation to Change, Physical Health, Emotional & Mental Health, Neighbourhood, Substance Use, Thinking & Behaviour, etc, Perception of Self & others, Living Arrangements, Attitudes to Offending, Lifestyle, Family & personal relationships

b. Dependent Variable: Rank outcome



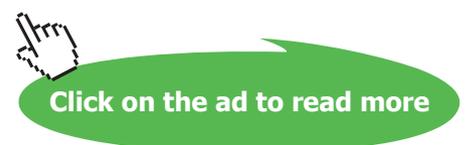
## Empowering People. Improving Business.

BI Norwegian Business School is one of Europe's largest business schools welcoming more than 20,000 students. Our programmes provide a stimulating and multi-cultural learning environment with an international outlook ultimately providing students with professional skills to meet the increasing needs of businesses.

BI offers four different two-year, full-time Master of Science (MSc) programmes that are taught entirely in English and have been designed to provide professional skills to meet the increasing need of businesses. The MSc programmes provide a stimulating and multi-cultural learning environment to give you the best platform to launch into your career.

- MSc in Business
- MSc in Financial Economics
- MSc in Strategic Marketing Management
- MSc in Leadership and Organisational Psychology

[www.bi.edu/master](http://www.bi.edu/master)



Finally the co-efficients of the equation are given. These are called Beta ( $\beta$ ) values in most texts, and in SPSS are labelled B. The output can be interpreted as meaning that

$$\text{Rank outcome} = 0.984 + 0.068 * \text{Living Arrangements} + 0.052 * \text{Family \& personal relationships} + 0.043 * \text{Ete} + 0.025 * \text{Neighbourhood} + 0.036 * \text{Lifestyle} + 0.100 * \text{Substance Use} - 0.015 * \text{Physical Health} - 0.043 * \text{Emotional \& Mental Health} - 0.009 * \text{Perception of Self \& others} + 0.047 * \text{Thinking \& Behaviour} + 0.068 * \text{Attitudes to Offending} + 0.059 * \text{Motivation to Change}$$

**Table 54: Coefficients(a)**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.984	.042		23.177	.000
	Living Arrangements	.068	.022	.114	3.102	.002
	Family & personal relationships	.052	.022	.090	2.346	.019
	Ete	.043	.019	.076	2.260	.024
	Neighbourhood	.025	.021	.036	1.170	.242
	Lifestyle	.036	.023	.061	1.593	.112
	Substance Use	.100	.018	.170	5.498	.000
	Physical Health	-.015	.030	-.014	-.508	.612
	Emotional & Mental Health	-.043	.020	-.067	-2.139	.033
	Perception of Self & others	-.009	.022	-.014	-.413	.680
	Thinking & Behaviour	.047	.022	.070	2.096	.036
	Attitudes to Offending	.068	.022	.115	3.053	.002
	Motivation to Change	.059	.023	.096	2.562	.011

a Dependent Variable: Rank outcome

However note that many of the above are not significant, for example Physical health has a p value of 0.612. Why would we include a variable that is not even significant? We can redo the regression using the “Stepwise” method (select this option instead of “Enter” in Figure 116).

What happens here is that each variable is entered but only kept if it is significant in the sense that it adds to the equation in terms of prediction over and above what is already in the equation. Here in Table 55 you can see each variable entered, but in Table 56 you can see the R<sup>2</sup> for each variable added. Table 57 and Table 58 show the ANOVA and beta values for each entry of the new variable. Table 59 shows the final excluded variables and Table 60 the final variables kept in the equation.

**Table 55:** Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	Lifestyle	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	Motivation to Change	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	Substance Use	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	Living Arrangements	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
5	Attitudes to Offending	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
6	ete	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
7	Family & personal relationships	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
8	Emotional & Mental Health	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
9	Thinking & Behaviour	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a Dependent Variable: Rank outcome

**Table 56:** Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.419(a)	.175	.175	.63901
2	.475(b)	.225	.224	.61971
3	.506(c)	.256	.254	.60773
4	.521(d)	.272	.269	.60131
5	.531(e)	.282	.279	.59736
6	.536(f)	.288	.284	.59535
7	.539(g)	.290	.286	.59450
8	.542(h)	.294	.288	.59338
9	.544(i)	.296	.290	.59250

a Predictors: (Constant), Lifestyle

b Predictors: (Constant), Lifestyle, Motivation to Change

c Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use

d Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements

e Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending

f Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, ete

g Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, ete, Family & personal relationships

h Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships, Emotional & Mental Health

i Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships, Emotional & Mental Health, Thinking & Behaviour

**Table 57: ANOVA(j)**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	94.199	1	94.199	230.689	.000(a)
	Residual	442.637	1084	.408		
	Total	536.836	1085			
2	Regression	120.915	2	60.457	157.423	.000(b)
	Residual	415.921	1083	.384		
	Total	536.836	1085			
3	Regression	137.220	3	45.740	123.846	.000(c)
	Residual	399.616	1082	.369		
	Total	536.836	1085			
4	Regression	145.972	4	36.493	100.927	.000(d)
	Residual	390.864	1081	.362		
	Total	536.836	1085			
5	Regression	151.446	5	30.289	84.881	.000(e)
	Residual	385.390	1080	.357		
	Total	536.836	1085			
6	Regression	154.389	6	25.731	72.596	.000(f)
	Residual	382.447	1079	.354		
	Total	536.836	1085			
7	Regression	155.839	7	22.263	62.991	.000(g)
	Residual	380.997	1078	.353		
	Total	536.836	1085			
8	Regression	157.621	8	19.703	55.957	.000(h)
	Residual	379.215	1077	.352		
	Total	536.836	1085			
9	Regression	159.099	9	17.678	50.356	.000(i)
	Residual	377.737	1076	.351		
	Total	536.836	1085			

a Predictors: (Constant), Lifestyle

b Predictors: (Constant), Lifestyle, Motivation to Change

c Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use

d Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements

e Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending

f Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc

- g Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships
- h Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships, Emotional & Mental Health
- i Predictors: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships, Emotional & Mental Health, Thinking & Behaviour
- j Dependent Variable: Rank outcome

**Table 58:** Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.186	.033		36.034	.000
	Lifestyle	.247	.016	.419	15.188	.000
2	(Constant)	1.157	.032		36.072	.000
	Lifestyle	.166	.018	.282	8.994	.000
	Motivation to Change	.161	.019	.262	8.341	.000
3	(Constant)	1.124	.032		35.288	.000
	Lifestyle	.122	.019	.207	6.317	.000
	Motivation to Change	.142	.019	.230	7.396	.000
	Substance Use	.117	.018	.199	6.644	.000
4	(Constant)	1.119	.032		35.478	.000
	Lifestyle	.094	.020	.159	4.700	.000
	Motivation to Change	.118	.020	.192	6.040	.000
	Substance Use	.099	.018	.168	5.575	.000
	Living Arrangements	.094	.019	.158	4.920	.000
5	(Constant)	1.074	.033		32.216	.000
	Lifestyle	.077	.020	.130	3.785	.000
	Motivation to Change	.071	.023	.115	3.076	.002
	Substance Use	.099	.018	.169	5.617	.000
	Living Arrangements	.092	.019	.155	4.843	.000
	Attitudes to Offending	.083	.021	.140	3.917	.000
6	(Constant)	1.044	.035		29.904	.000
	Lifestyle	.054	.022	.091	2.478	.013
	Motivation to Change	.065	.023	.105	2.813	.005
	Substance Use	.099	.018	.169	5.647	.000
	Living Arrangements	.086	.019	.145	4.524	.000
	Attitudes to Offending	.078	.021	.131	3.664	.000
	Ete	.053	.018	.093	2.881	.004
7	(Constant)	1.032	.035		29.225	.000
	Lifestyle	.051	.022	.086	2.339	.020
	Motivation to Change	.063	.023	.102	2.751	.006
	Substance Use	.096	.018	.163	5.427	.000

	Living Arrangements	.065	.022	.110	3.017	.003
	Attitudes to Offending	.075	.021	.128	3.564	.000
	Ete	.045	.019	.079	2.389	.017
	Family & personal relationships	.043	.021	.074	2.026	.043
8	(Constant)	1.038	.035		29.367	.000
	Lifestyle	.049	.022	.083	2.260	.024
	Motivation to Change	.064	.023	.104	2.808	.005
	Substance Use	.099	.018	.168	5.600	.000
	Living Arrangements	.068	.022	.114	3.121	.002
	Attitudes to Offending	.077	.021	.130	3.635	.000
	Ete	.049	.019	.086	2.602	.009
	Family & personal relationships	.058	.022	.099	2.607	.009
	Emotional & Mental Health	-.044	.019	-.068	-2.250	.025
9	(Constant)	.992	.042		23.725	.000
	Lifestyle	.040	.022	.069	1.827	.068
	Motivation to Change	.059	.023	.096	2.573	.010
	Substance Use	.100	.018	.170	5.649	.000
	Living Arrangements	.070	.022	.118	3.243	.001
	Attitudes to Offending	.066	.022	.112	3.033	.002
	Ete	.045	.019	.079	2.383	.017
	Family & personal relationships	.055	.022	.094	2.456	.014
	Emotional & Mental Health	-.049	.020	-.075	-2.494	.013
	Thinking & Behaviour	.045	.022	.068	2.052	.040

a Dependent Variable: Rank outcome

Table 59 shows which variables are not in the equation at each stage.

**Table 59:** Excluded Variables(j)

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics	
						Tolerance	
1	Living Arrangements	.251(a)	8.018	.000	.237	.733	
	Family & personal relationships	.236(a)	7.587	.000	.225	.746	
	Ete	.172(a)	5.158	.000	.155	.670	
	Neighbourhood	.142(a)	4.519	.000	.136	.756	
	Substance Use	.232(a)	7.672	.000	.227	.788	
	Physical Health	.080(a)	2.796	.005	.085	.922	
	Emotional & Mental Health	.073(a)	2.538	.011	.077	.906	
	Perception of Self & others	.149(a)	4.821	.000	.145	.779	
	Thinking & Behaviour	.178(a)	5.613	.000	.168	.739	
	Attitudes to Offending	.244(a)	7.851	.000	.232	.745	
	Motivation to Change	.262(a)	8.341	.000	.246	.727	
	2	Living Arrangements	.195(b)	6.098	.000	.182	.679
		Family & personal relationships	.186(b)	5.919	.000	.177	.704
ete		.124(b)	3.730	.000	.113	.645	
Neighbourhood		.110(b)	3.568	.000	.108	.743	
Substance Use		.199(b)	6.644	.000	.198	.770	
Physical Health		.052(b)	1.853	.064	.056	.907	
Emotional & Mental Health		.035(b)	1.238	.216	.038	.881	
Perception of Self & others		.074(b)	2.305	.021	.070	.694	
Thinking & Behaviour		.108(b)	3.305	.001	.100	.669	
Attitudes to Offending		.146(b)	3.959	.000	.119	.518	
3		Living Arrangements	.158(c)	4.920	.000	.148	.651
		Family & personal relationships	.153(c)	4.855	.000	.146	.681
		ete	.120(c)	3.680	.000	.111	.645
	Neighbourhood	.086(c)	2.812	.005	.085	.731	
	Physical Health	.006(c)	.212	.832	.006	.849	
	Emotional & Mental Health	.008(c)	.277	.782	.008	.862	
	Perception of Self & others	.060(c)	1.908	.057	.058	.691	
	Thinking & Behaviour	.107(c)	3.338	.001	.101	.669	
	Attitudes to Offending	.145(c)	4.009	.000	.121	.518	
	4	Family & personal relationships	.102(d)	2.828	.005	.086	.520
		ete	.103(d)	3.194	.001	.097	.637
		Neighbourhood	.060(d)	1.933	.053	.059	.705
		Physical Health	-.008(d)	-.277	.782	-.008	.841
Emotional & Mental Health		-.024(d)	-.832	.406	-.025	.819	
Perception of Self & others		.037(d)	1.176	.240	.036	.674	
Thinking & Behaviour		.103(d)	3.263	.001	.099	.668	

	Attitudes to Offending	.140(d)	3.917	.000	.118	.518
5	Family & personal relationships	.093(e)	2.587	.010	.079	.517
	ete	.093(e)	2.881	.004	.087	.632
	Neighbourhood	.058(e)	1.877	.061	.057	.704
	Physical Health	-.013(e)	-.468	.640	-.014	.839
	Emotional & Mental Health	-.030(e)	-1.058	.290	-.032	.817
	Perception of Self & others	.005(e)	.159	.874	.005	.628
	Thinking & Behaviour	.076(e)	2.343	.019	.071	.624
6	Family & personal relationships	.074(f)	2.026	.043	.062	.494
	Neighbourhood	.046(f)	1.503	.133	.046	.691
	Physical Health	-.018(f)	-.651	.515	-.020	.835
	Emotional & Mental Health	-.044(f)	-1.540	.124	-.047	.796
	Perception of Self & others	-.011(f)	-.326	.744	-.010	.610
	Thinking & Behaviour	.064(f)	1.964	.050	.060	.611
7	Neighbourhood	.043(g)	1.388	.166	.042	.689
	Physical Health	-.019(g)	-.683	.495	-.021	.835
	Emotional & Mental Health	-.068(g)	-2.250	.025	-.068	.726
	Perception of Self & others	-.020(g)	-.602	.547	-.018	.599
	Thinking & Behaviour	.058(g)	1.747	.081	.053	.604
8	Neighbourhood	.036(h)	1.152	.250	.035	.681
	Physical Health	-.010(h)	-.358	.721	-.011	.817
	Perception of Self & others	-.005(h)	-.160	.873	-.005	.576
	Thinking & Behaviour	.068(h)	2.052	.040	.062	.594
9	Neighbourhood	.035(i)	1.138	.255	.035	.681
	Physical Health	-.013(i)	-.476	.634	-.015	.815
	Perception of Self & others	-.012(i)	-.348	.728	-.011	.571

a Predictors in the Model: (Constant), Lifestyle

b Predictors in the Model: (Constant), Lifestyle, Motivation to Change

c Predictors in the Model: (Constant), Lifestyle, Motivation to Change, Substance Use

d Predictors in the Model: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements

e Predictors in the Model: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending

f Predictors in the Model: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc

g Predictors in the Model: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships

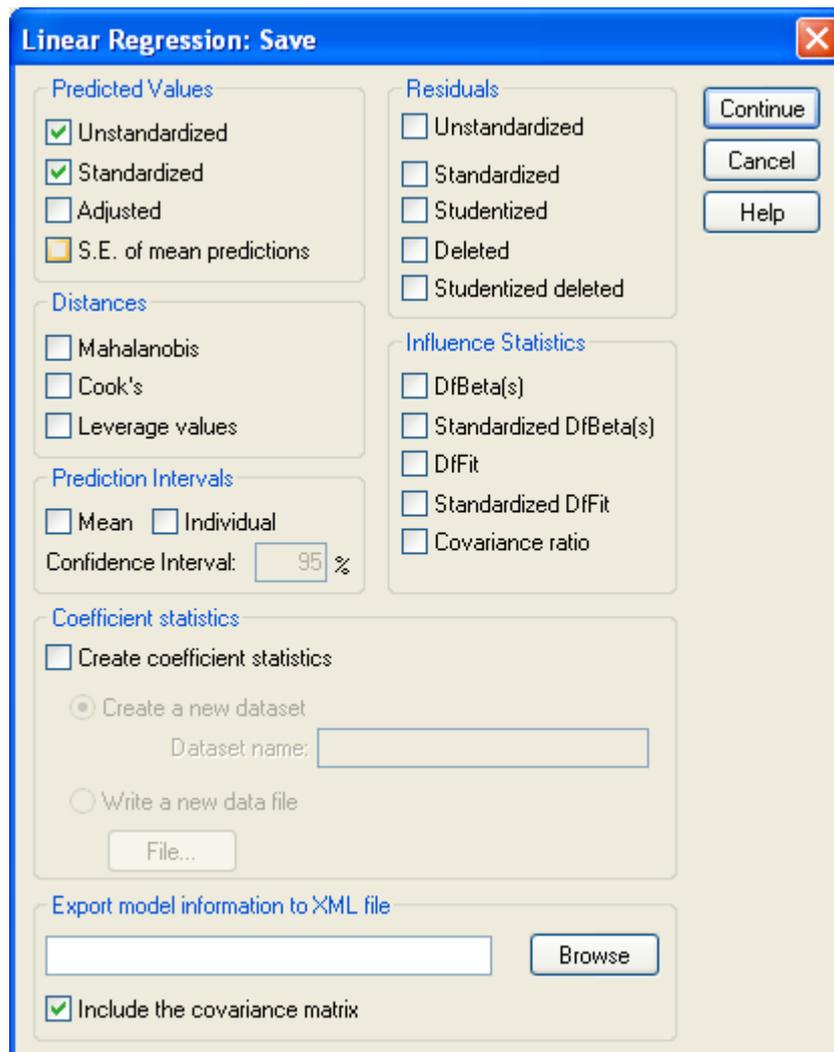
h Predictors in the Model: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships, Emotional & Mental Health

i Predictors in the Model: (Constant), Lifestyle, Motivation to Change, Substance Use, Living Arrangements, Attitudes to Offending, etc, Family & personal relationships, Emotional & Mental Health, Thinking & Behaviour

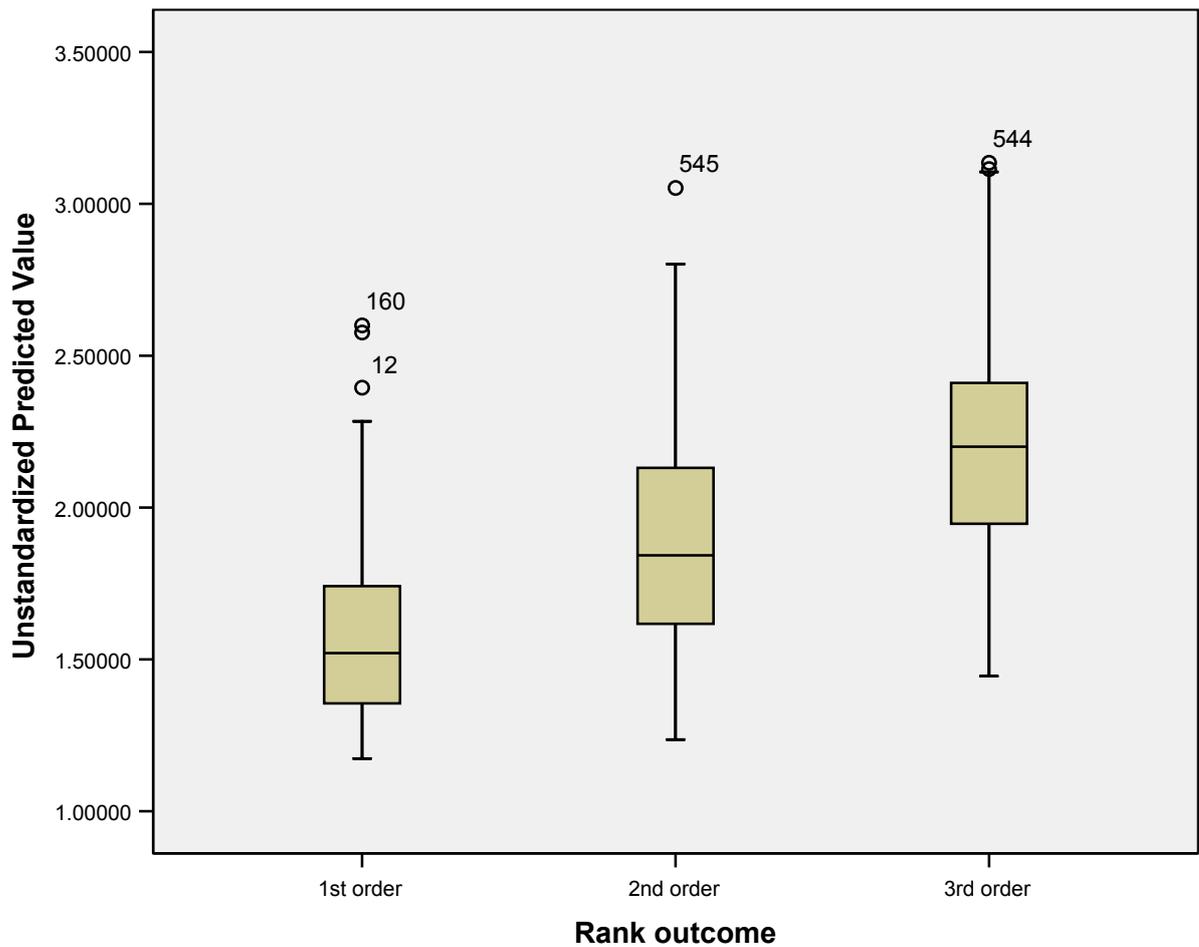
j Dependent Variable: Rank outcome



Figure 117: Predicted values



We can see that this gives a reasonable prediction as in Figure 118 the predicted ranks are broadly in agreement with the actual ones. But note that there are a wide range within each actual outcome of the predicted outcome. I.e. the predictions are not perfect, and (e.g.) for third (worst) outcome the range of predictions goes down to 1.5 (somewhere between first and second rank) and just over 3 (third rank).



**Figure 118:** rank outcome predictions

We can also see there is a significant correlation between our predictions and the actual values, see Table 61.

**Table 61:** Correlation

			Unstandardized Predicted Value	Rank outcome
Spearman's rho	Unstandardized Predicted Value	Correlation Coefficient	1.000	.557(**)
		Sig. (2-tailed)	.	.000
		N	471	468
Rank outcome	Rank outcome	Correlation Coefficient	.557(**)	1.000
		Sig. (2-tailed)	.000	.
		N	468	1622

\*\* Correlation is significant at the 0.01 level (2-tailed).

## Conclusion

Linear regression allows you to predict one variable (outcome) from one or more other variables (in this case risk factors).

## Exercise

Using gravity as your outcome to be predicted, and asset scores as variables, compute a regression analysis, then perform it and interpret it.

# Need help with your dissertation?

Get in-depth feedback & advice from experts in your topic area. Find out what you can do to improve the quality of your dissertation!

[Get Help Now](#)



Go to [www.helpmyassignment.co.uk](http://www.helpmyassignment.co.uk) for more info



# 17 Logistic regression

## Introduction

In the last chapter we looked at linear regression, where we tried to predict one outcome (variable) from one or more parameters (other variables). This works if the outcome is at least ordinal. What if the outcome is nominal, in particular a binary outcome such as yes or no? Then we need to use another type of regression called logistic regression.

Sometimes what we wish to predict is the occurrence or non-occurrence of some event. Certain types of patients suffer from pressure sores much more than other groups (e.g. those with poor nutrition). However we cannot do a regression for nutritional status (as measured by serum albumin, for example) against pressure sore, because a patient can have a sore or not, but cannot have 0.5 of a sore. In other words, the presence or absence of a sore is nominal data and thus cannot be normally distributed (this concept being meaningless for nominal data). Therefore, as it stands, regression is inappropriate.

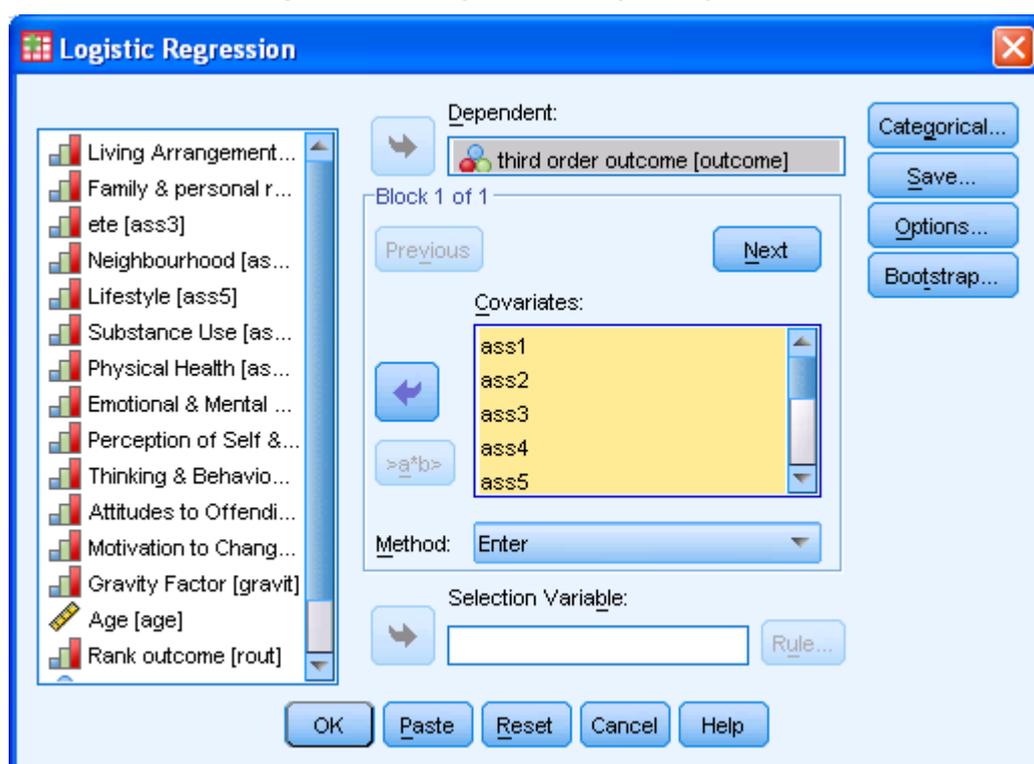
There is a form of regression, called logistic regression that can be used where the dependent variable is dichotomous, i.e. can take two values. In the pressure sore case, a sore can exist (we could label this 1) or not exist (we could label this 0). Logistic regression returns a probability, in this case of whether the sore will occur or not.

## Example asset scores (again)

In the last chapter we regressed asset sub-scores against the level of sentence severity. Suppose however all we are interested in is whether the young person received a custodial sentence or not. We coded this as a third order outcome, which was called outcome.

I am going to regress sentence against asset sub-scores by first using **Regression** -> **Binary Logistic**. I enter third order outcome as dependent and all the sub-scores as covariates, see Figure 119.

Figure 119: Dialogue box for logistic regression



As in linear regression the output (after several steps and a lot of tables) shows me the significance (p value) of each variable (co-variate) see Table 62. But many of these are not significant, so why use them?

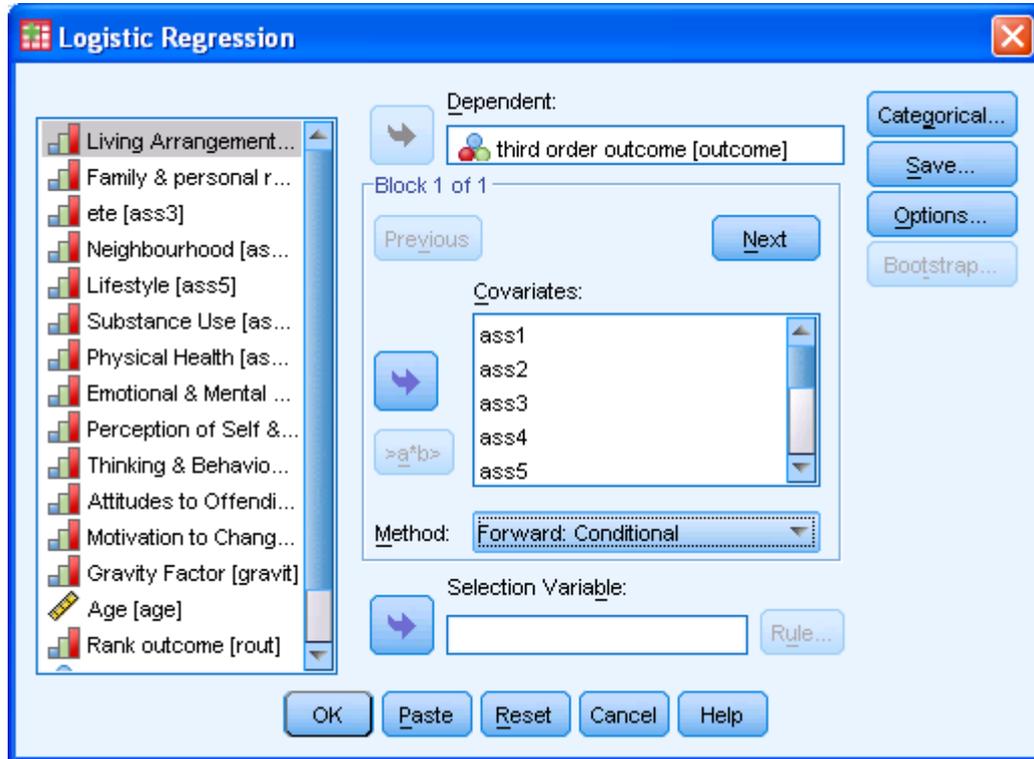
Table 62: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	ass0_1	.179	.107	2.830	1	.092	1.197
	ass1_1	.142	.124	1.317	1	.251	1.152
	ass2_1	.084	.110	.587	1	.444	1.088
	ass3_1	.052	.106	.237	1	.626	1.053
	ass4_1	.128	.129	.985	1	.321	1.137
	ass5_1	.457	.095	23.127	1	.000	1.580
	ass6_1	-.127	.136	.871	1	.351	.881
	ass7_1	-.354	.108	10.843	1	.001	.702
	ass8_1	.039	.112	.119	1	.730	1.039
	ass9_1	.258	.139	3.426	1	.064	1.294
	ass10_1	.300	.129	5.401	1	.020	1.350
	ass11_1	.190	.113	2.816	1	.093	1.209
	Constant	-4.688	.346	183.412	1	.000	.009

a Variable(s) entered on step 1: ass0\_1, ass1\_1, ass2\_1, ass3\_1, ass4\_1, ass5\_1, ass6\_1, ass7\_1, ass8\_1, ass9\_1, ass10\_1, ass11\_1.

If instead of the *Enter* we choose *Forward Conditional* (see Figure 120) then only variables that significantly predict the outcome are kept, rather as happened in linear regression with the stepwise method.

Figure 120: Dialogue box to change method



# Brain power



Plug into The Power of Knowledge Engineering.  
Visit us at [www.skf.com/knowledge](http://www.skf.com/knowledge)

By 2020, wind could provide one-tenth of our planet's electricity needs. Already today, SKF's innovative know-how is crucial to running a large proportion of the world's wind turbines.

Up to 25 % of the generating costs relate to maintenance. These can be reduced dramatically thanks to our systems for on-line condition monitoring and automatic lubrication. We help make it more economical to create cleaner, cheaper energy out of thin air.

By sharing our experience, expertise, and creativity, industries can boost performance beyond expectations. Therefore we need the best employees who can meet this challenge!

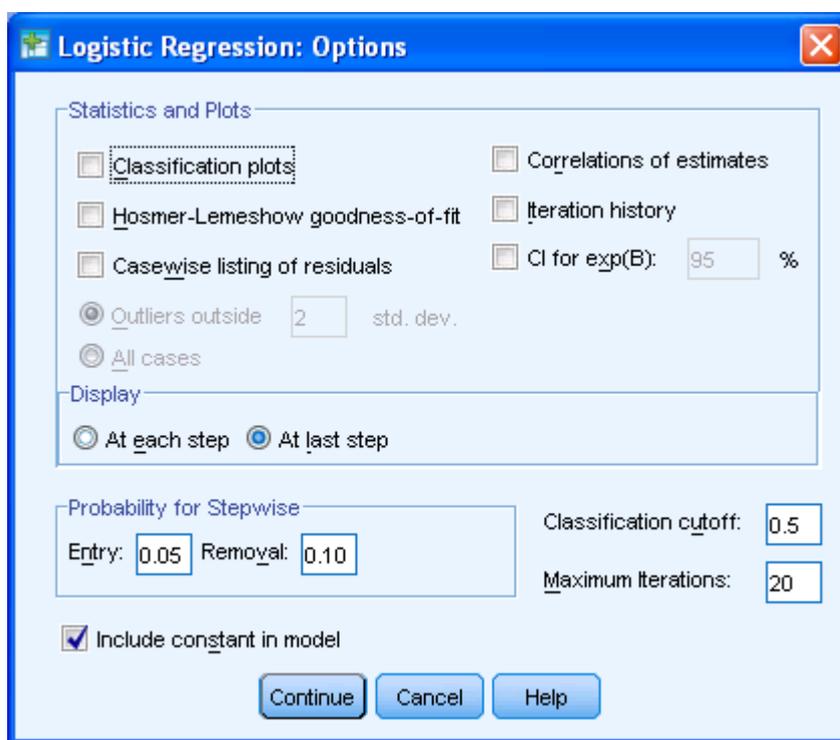
The Power of Knowledge Engineering





We can also elect to have just the last step of the regression analysis shown, using **Options** see Figure 121.

**Figure 121:** Options for logistic regression



We are given a classification table showing how many are correctly predicted, see Table 63.

**Table 63:** Classification table

**Classification Table<sup>a</sup>**

Observed			Predicted		
			third order outcome		Percentage Correct
			Non custodial sentence	Custodial sentence	
Step 6	third order outcome	Non custodial sentence	934	24	97.5
		Custodial sentence	107	30	21.9
	Overall Percentage				88.0

The variables that remain significant are shown at the last step (Table 64), and those that are excluded (Table 65).

**Table 64:** Variables in the equation

<b>Variables in the Equation</b>							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 6 <sup>a</sup>	ass1	.293	.091	10.286	1	.001	1.340
	ass6	.491	.088	31.250	1	.000	1.634
	ass8	-.321	.101	10.184	1	.001	.725
	ass10	.319	.136	5.490	1	.019	1.376
	ass11	.354	.123	8.248	1	.004	1.425
	ass12	.226	.112	4.084	1	.043	1.254
	Constant	-4.471	.324	190.066	1	.000	.011

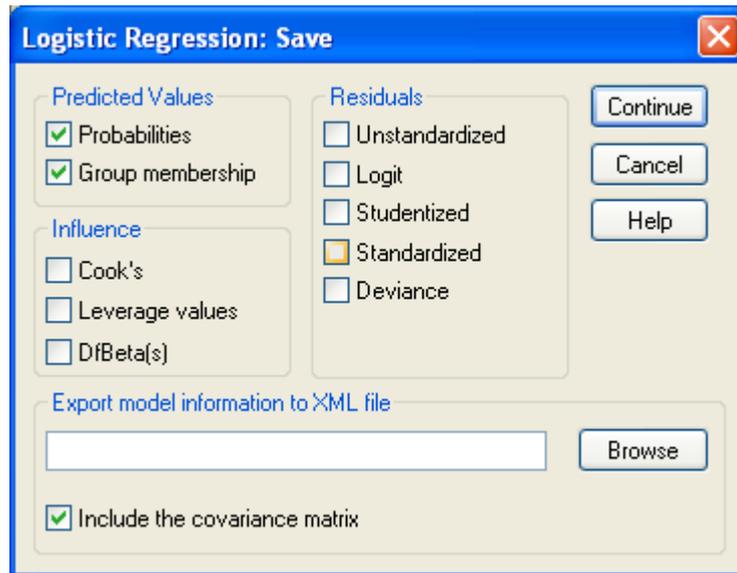
a. Variable(s) entered on step 6: ass10.

**Table 65:** Variables not in the Equation

<b>Variables not in the Equation</b>					
			Score	df	Sig.
Step 6	Variables	ass2	2.715	1	.099
		ass3	2.296	1	.130
		ass4	1.197	1	.274
		ass5	2.710	1	.100
		ass7	.958	1	.328
		ass9	.766	1	.381
	Overall Statistics		6.765	6	.343

Linear regression predicts the numerical value of an outcome. But with binary outcomes we can state how many are correct, and additionally we can give a probability of getting the binary outcome. Here we may have a probability close to zero which would suggest the young person will not get a custodial sentence, and close to one meaning s/he will. We can get both these predictions by choosing to in a dialogue box under *Save* (Figure 122).

Figure 122: Saving probabilities and group membership



We can graphically see the prediction in a bar chart, this shows we are getting most predictions correct (Figure 123).

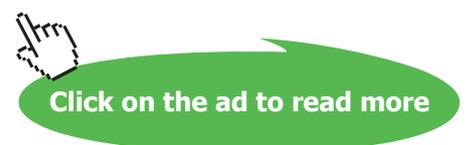
“I studied English for 16 years but...  
...I finally learned to speak it in just six lessons”

Jane, Chinese architect

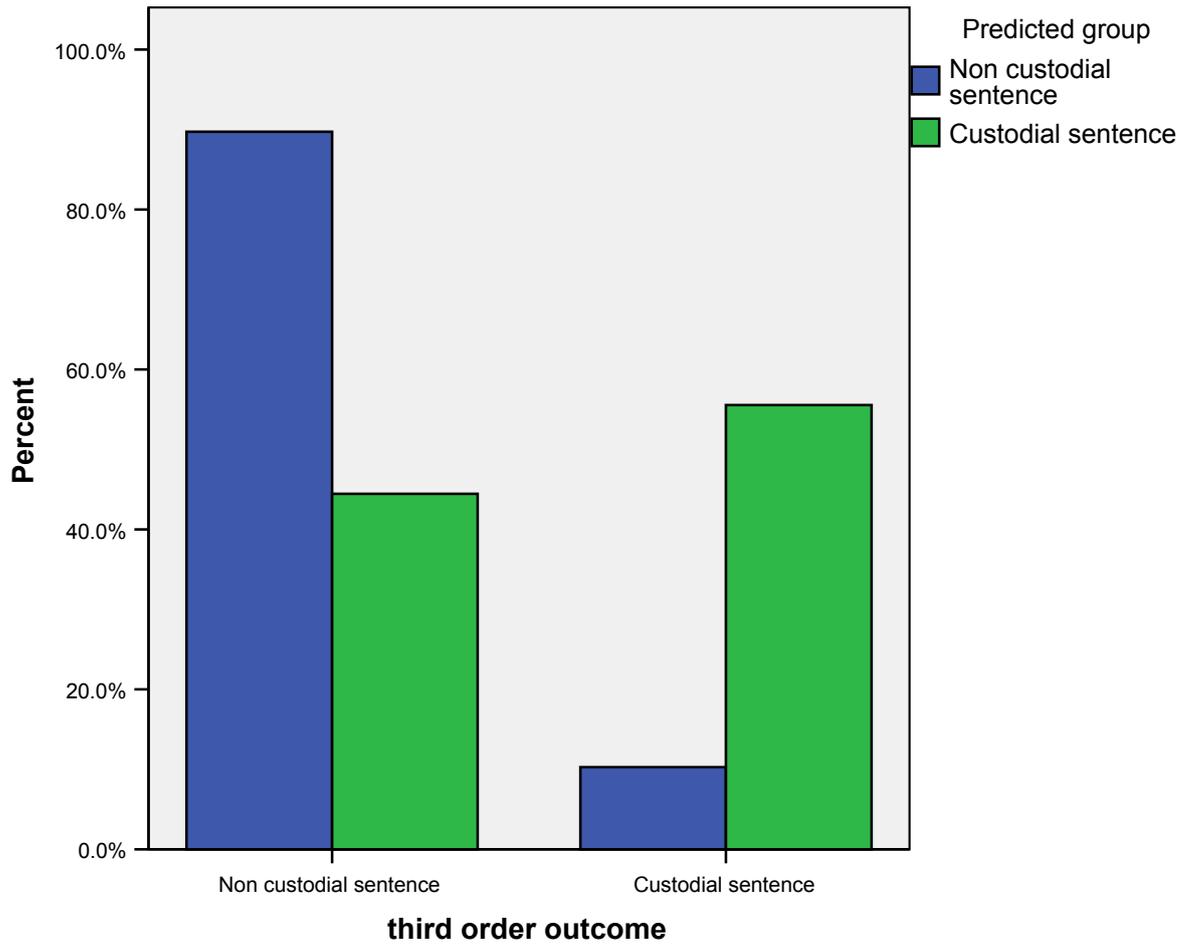
ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



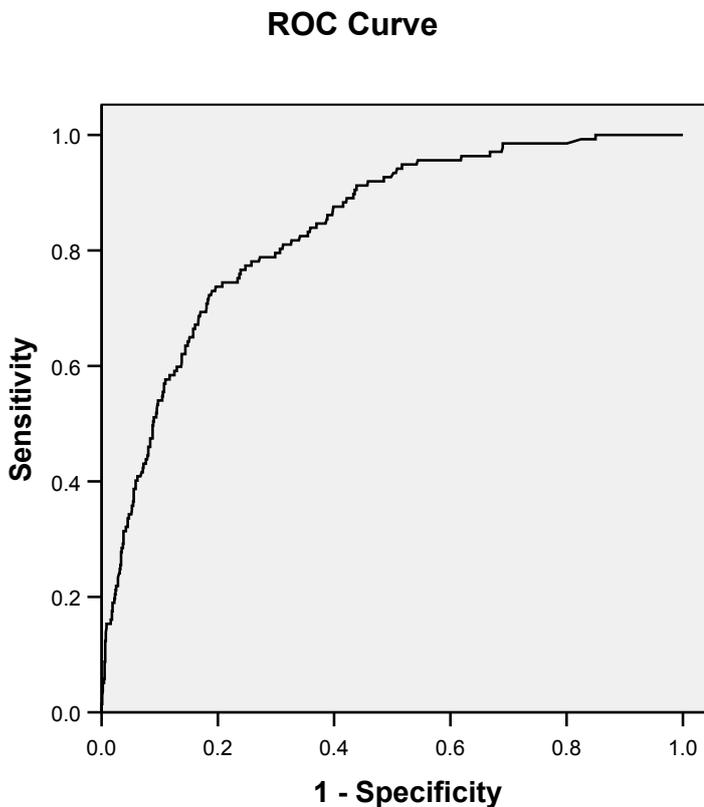


**Figure 123:** Bar chart of predicted and actual third custodial sentence



We can use a technique from an earlier chapter to quantify the prediction probability, the ROC curve, which shows an area under the curve (AUC) of 0.837 (Figure 124 and Table 66).

**Figure 124:** ROC curve for prediction



Diagonal segments are produced by ties.

**Table 66:** Area Under the Curve

Test Result Variable(s): Predicted probability

Area
.837

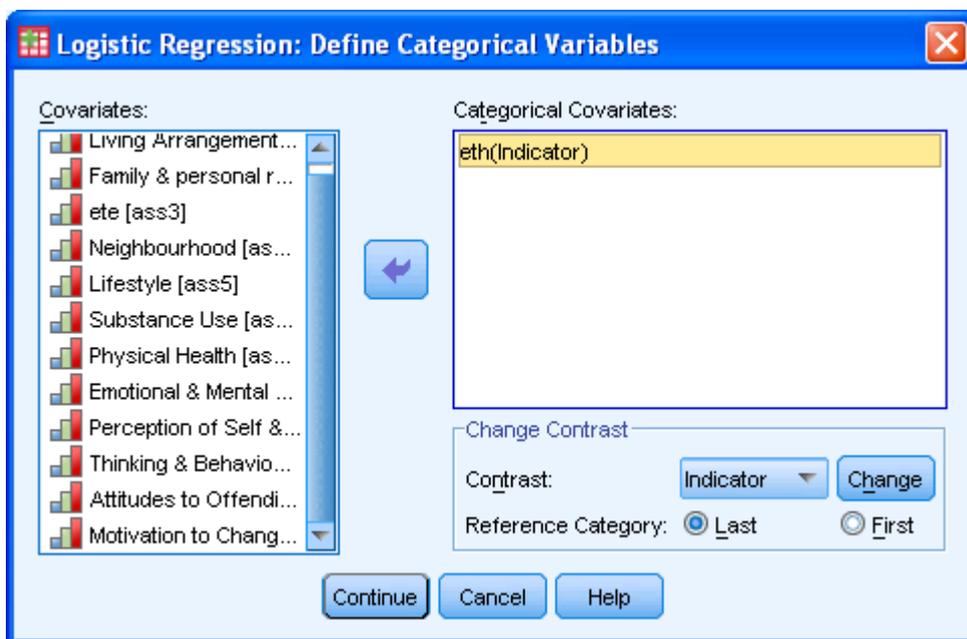
The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

What if we add some more variables, say ethnic group. Ethnicity is a nominal variable, so we cannot simply put it in the equation. For if we code Asian as 1, White as 4, then it does not follow that Black (2) or Mixed race (3) are in any meaningful way between White and Asian. However we can make “dummy variables” for each ethnic group.

To create a dummy variable first let us consider if we had two groups, say Asian and Black, then one variable is needed, it could be coded 1 if the subject is Asian, otherwise 0 if Black. So two groups one variable. If we had three groups, say Asian, Black and Mixed, then you might think three new variables are needed, Asian (0 or 1), Mixed (0 or 1) or Black (0 or 1). But if you had two variables Asian (0/1) and Mixed (0/1) then if both were zero then you know the subject is not Asian or Black, so must be Mixed. Similarly if there are four groups, now including White, you need three variables – Asian, Black and Mixed, for if all three were zero, then the subject must be White.

While you can do this manually SPSS has a method for logistic regressions (but not linear regression) which is to state the covariate is categorical (a synonym for nominal). If I put variable *eth* into the covariate and then click on **Categorical** I get the dialogue box as in Figure 125. Normally you would choose the reference category to be the most common. Here it is White, and this is the last (1=Asian, 2=Black, 3 Mixed and 4 White).

Figure 125: Making a co-variate categorical



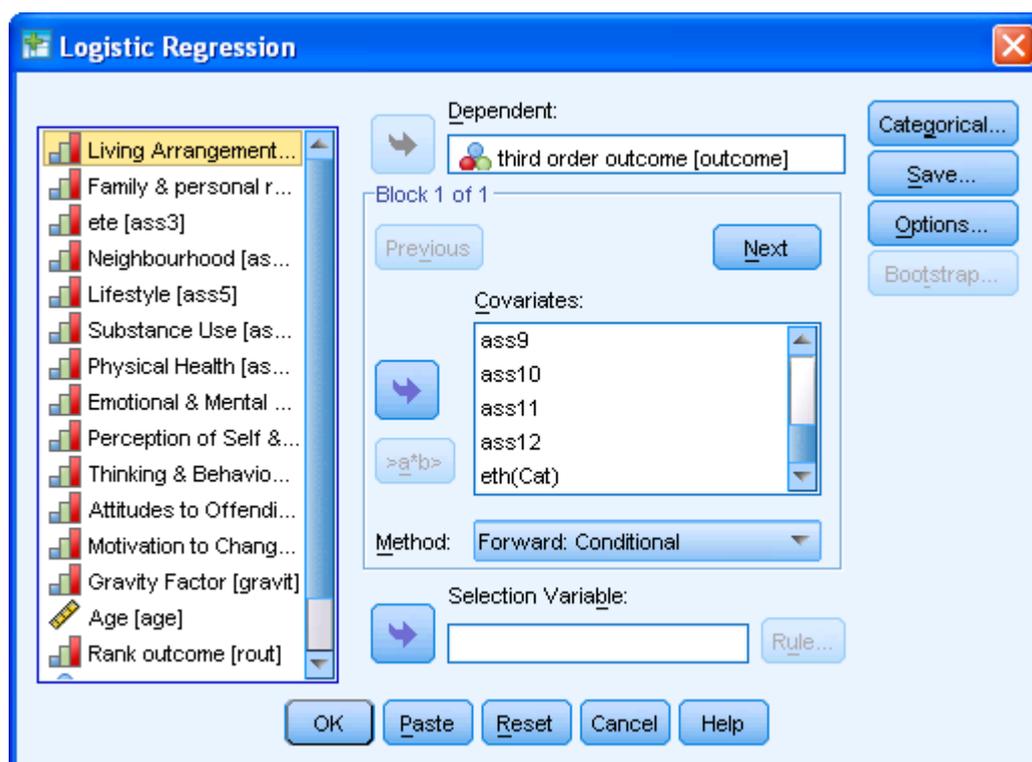
What do you want to do?

No matter what you want out of your future career, an employer with a broad range of operations in a load of countries will always be the ticket. Working within the Volvo Group means more than 100,000 friends and colleagues in more than 185 countries all over the world. We offer graduates great career opportunities – check out the Career section at our web site [www.volvogroup.com](http://www.volvogroup.com). We look forward to getting to know you!

**VOLVO**  
 AB Volvo (publ)  
[www.volvogroup.com](http://www.volvogroup.com)

VOLVO TRUCKS | RENAULT TRUCKS | MACK TRUCKS | VOLVO BUSES | VOLVO CONSTRUCTION EQUIPMENT | VOLVO PENTA | VOLVO AERO | VOLVO IT  
 VOLVO FINANCIAL SERVICES | VOLVO 3P | VOLVO POWERTRAIN | VOLVO PARTS | VOLVO TECHNOLOGY | VOLVO LOGISTICS | BUSINESS AREA ASIA





Note none of the ethnic groups appear to be significant predictors of custody; they all appear in the variables not in the equation, see.

**Table 67: Adding ethnicity  
Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 5 <sup>a</sup>	ass1	.317	.092	11.782	1	.001	1.372
	ass6	.503	.089	31.610	1	.000	1.653
	ass8	-.353	.104	11.556	1	.001	.703
	ass10	.355	.136	6.803	1	.009	1.427
	ass11	.484	.113	18.193	1	.000	1.623
	Constant	-4.372	.328	177.273	1	.000	.013

a. Variable(s) entered on step 5: ass10.

**Variables not in the Equation**

			Score	df	Sig.
Step 5	Variables	ass2	2.271	1	.132
		ass3	2.396	1	.122
		ass4	1.076	1	.300
		ass5	2.622	1	.105
		ass7	.921	1	.337
		ass9	.349	1	.555
		ass12	2.266	1	.132
		eth	6.435	3	.092
		eth(1)	.184	1	.668
		eth(2)	3.579	1	.059
		eth(3)	2.290	1	.130
		Overall Statistics			14.047

### Conclusion

Regression is a technique for predicting one dependent variable from one (simple linear regression) or more (multiple regression) independent variables. It makes assumptions about the data (e.g. normality and linearity), and the results may be nonsensical if there is no clear reason why a causal relationship should exist. If the dependent variable is binary (e.g. yes/no, on/off, male/female) then logistic regression may be employed to determine the probability of one or other value of the variable being obtained.

Logistic regression allows us to predict the likelihood of a binary outcome based on many variables, including other binary variables (here we used ethnic group).

### Exercise

Run a logistic regression using all the above variables and in addition gravity and gender. Interpret the output.

# 18 Cluster analysis

## KEY POINTS

Cluster analysis is a method of exploring data

- A metric is a measure of how far apart two items are
- Clustering can simply split items into clusters at the same level or place them in a hierarchical structure

At the end of this chapter the student should be able to:

- Use SPSS to undertake a cluster analysis
- Interpret dendrograms and icicle plots

**gaiTeye**<sup>®</sup>  
Challenge the way we run

**EXPERIENCE THE POWER OF FULL ENGAGEMENT...**

.....

**RUN FASTER.  
RUN LONGER..  
RUN EASIER...**

**READ MORE & PRE-ORDER TODAY**  
**WWW.GAITEYE.COM**



## Introduction

A method of splitting data into natural clusters is named cluster analysis. The aim of cluster analysis is to investigate the underlying structure of the data. For a fuller treatment on cluster analysis see Everitt (1980).

As an illustration life on Earth could be split into the animal kingdom, the plant kingdom, fungi, and microorganisms (bacteria, viruses). Animals can be further split into chordates (having a backbone) and several other phyla, for example echinoderms (starfish, etc.). These large categories can be split into smaller ones, for example mammals are part of the chordates and reptiles are another group.

There are two ways of performing a hierarchical cluster analysis. One method is to group variables, the idea here being to decide which variables are close to each other and how the variables group together. In the second method, individuals in the sample are placed into clusters. Later in this chapter an example is given of each of these ways of doing a cluster analysis.

Individuals can be allocated into groups (or clusters) on the basis of how 'close' they are to each other. We would naturally say that pigs are closer to humans than, say, crocodiles. But apes are closer to humans than pigs. But on what basis is this measure of 'closeness' made? One measure would be the percentage of DNA each species has in common. If, for example, on this basis chimpanzees and humans were 95% the same, and some other ape was 90% the same as humans, then we would say that chimpanzees were closer to humans than some other ape (say orang-utans). It is noted that a different form of measuring closeness might give a totally different taxonomy. Suppose we have features that measure how an animal feeds, humans might then be seen as very close to pigs (both omnivores) and more different from gorillas (strict vegetarians). Thus the taxonomy is highly dependent on our measures (called metrics) which should be chosen to identify the feature of interest, which itself may not be straightforward or obvious.

We could put humans, chimps and other apes in one cluster (primates), and dogs, pigs, cats and horses in other clusters. These all fit into a supercluster (mammals) and all mammals along with reptiles and most fish go into chordates, which themselves go with elasmobranchs (sharks) and echinoderms (starfish etc.) into the animal kingdom, which is quite separate from, say, bacteria. Finally, even bacteria and humans are related, as they have a common ancestor (something like algae about 3.5 billion years ago). If life were to be found on Mars, it would probably be even further removed from us than bacteria. This form of clustering is hierarchical, in the sense that not only are there different groups (clusters) but there is also a natural ordering and depth of clusters and relationships between clusters.

Cluster analysis places cases into classes. It can be used in similar applications as principal component analysis (PCA) and factor analysis (FA) to identify the groupings within a data set. Cluster analysis can be used in exploratory data analysis.

Cluster analysis can be used to group individual cases into discrete groups. It uses some method of determining how close two individuals are using the metric or distance (see below) and individuals that are close to each other based on this metric are said to be in same group. Grouping can be:-

- In a cascade, or hierarchy, or taxonomy. All items belong in one super-ordinate or general group, which is broken down into smaller and smaller subgroups. The investigator does not state how many groups there are.
- Non-hierarchical. The investigator states how many groups there are and the program assigns items to groups depending on how similar they are to each other.

While the investigator may need to state the number of clusters that will be found, this should be done on a theoretical basis. The investigator will also need to be able to interpret the clusters once they have been located, and this can be difficult.

Cluster analysis is typically used for exploration of data, and tasks suitable to cluster analysis include:

- finding a typology
- model fitting
- prediction based on groups
- hypothesis testing
- hypothesis generation
- data reduction.

## Hierarchical methods

Data are not partitioned into groups in one go, but are grouped into broad classes. The method typically used is agglomerative, whereby cases that are near to each other are placed in the same cluster. The distances between the clusters may then be calculated, and nearer clusters are placed in super-clusters, etc. The algorithm is as follows:

1. Compute differences between items
2. Fuse items that are nearest
3. Go to step 1 until all items or subgroups are in one large group.

There are several options about how to merge items or clusters. In nearest neighbour clustering, groups are fused according to the distance between nearest neighbours, i.e. close cases or clusters are put together. In centroid clustering, groups are depicted as lying in Euclidean space, and a group is said to lie where the centroid of the group lies. Median clustering is where groups are merged, and the new position is the median of the values of the old groups.

## Metrics

If clusters are formed by merging close members, there has to be a measure of what closeness is. A metric is a measure of how close (or far apart) two individuals are. Metrics have the following features, where  $d(x,y)$  is the distance between  $x$  and  $y$ :

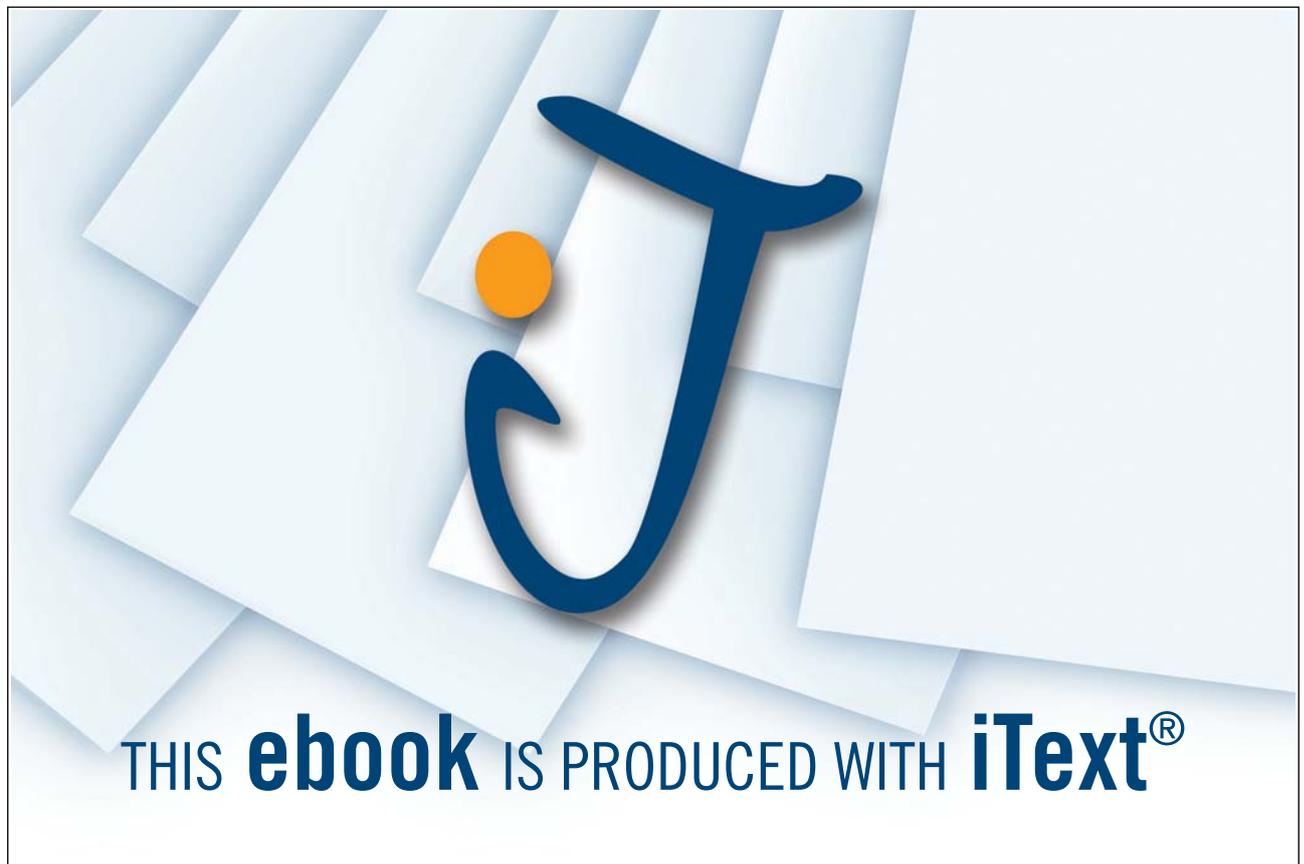
- $d(x,y) \geq 0$ : the distance between  $x$  and  $y$  is either zero or positive (like ordinary distances); in particular,  $d(x,y) = 0$  implies  $x = y$ , and vice versa. If the distance between two objects is zero, they must be in the same place.
- $d(x,y) = d(y,x)$ : the distance between  $x$  and  $y$  is the same as the distance between  $y$  and  $x$ .
- $d(x,y) + d(y,z) \geq d(x,z)$ : the distance between  $x$  and  $y$  added to the distance between  $y$  to  $z$  can never be less than the distance between  $x$  and  $z$ .

There are many metrics, but typically the simple Euclidean distance coefficient is used. For one dimension (variable), with two values  $x_1$  and  $x_2$ , this is just the distance

$$d = x_1 - x_2$$

For two dimensions, with values  $(x_{11}, x_{12})$  and  $(x_{21}, x_{22})$ , this is the pythagorean distance

$$d = \sqrt{((x_{11}-x_{21})^2 + (x_{12}-x_{22})^2)}$$



In general two items can be measured in N dimensions, That is, N variables are used to represent each item.

$$d = \sqrt{((x_{11}-x_{21})^2 + (x_{12}-x_{22})^2 + (x_{13}-x_{23})^2) + \dots + (x_{1N}-x_{2N})^2}$$

This is difficult to visualise for more than two or three dimensions, but the arithmetic is simple.

### Non-hierarchical methods

In this form of clustering only one level of clusters is allowed, and the technique splits data into several clusters that do not overlap. Often you need to specify how many clusters you expect in the beginning. An example of non-hierarchical clustering is k-means clustering

### Optimization techniques

These techniques are usually non-hierarchical. Again the number of groups is typically determined by the investigator and data are partitioned so as to optimize some predefined numerical measure (metric), as above. However, these techniques are typically iterative and involve large amounts of computing. Examples of optimization are unsupervised neural networks (e.g. Kohonen nets). Optimization is not necessarily going to produce a solution, and may give slightly different results on each run, and in some techniques the results can be totally different in some runs. Neural networks, for example, can get into ‘local minima’ and fail to reach a solution.

### Density search techniques

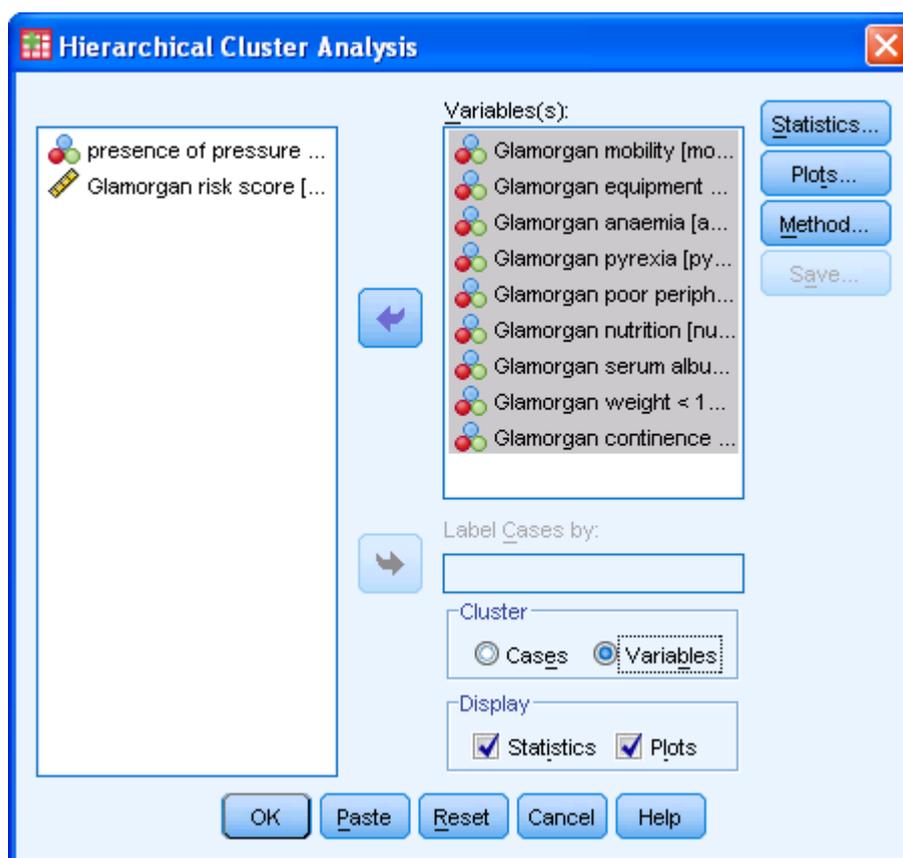
In this technique cases are displayed on a graph, and areas of high density that are surrounded by more diffuse areas are considered to be clusters. This can be done by visual inspection.

### Example 1: Hierarchical clustering

Using a new paediatric risk assessment score for, we would like to see if the variables form clusters. Use datafile “*glam*” we can bring up a dialogue box with **Analyze -> Classify -> Hierarchical Cluster**.

Then we have entered the variables of the risk score. By default, SPSS assumes that we want to cluster by cases (members). This may indeed be the case, and then we could look at the members of each cluster and interpret the meaning of the clusters. In the present example, clustering by cases would give such a large number of clusters (336) in the final stage of the analysis, that it would be very difficult to make much sense of it. Here we are better advised to see how the variables cluster, so we have elected to cluster by variable, which will give us at most nine clusters (the original nine variables), see Figure 126.

**Figure 126:** Selecting variables for cluster analysis



SPSS gives a choice of plots: icicle plots (so named as they look like icicles hanging from the ceiling), see Figure 127 or dendrograms, see Figure 128. The icicle plot starts with all the variables (or cases, if that option is selected) in the one cluster, this is row one. In the second row two clusters are formed, where the closest variables (or cases) are put into a cluster.

Figure 127: Icicle plot

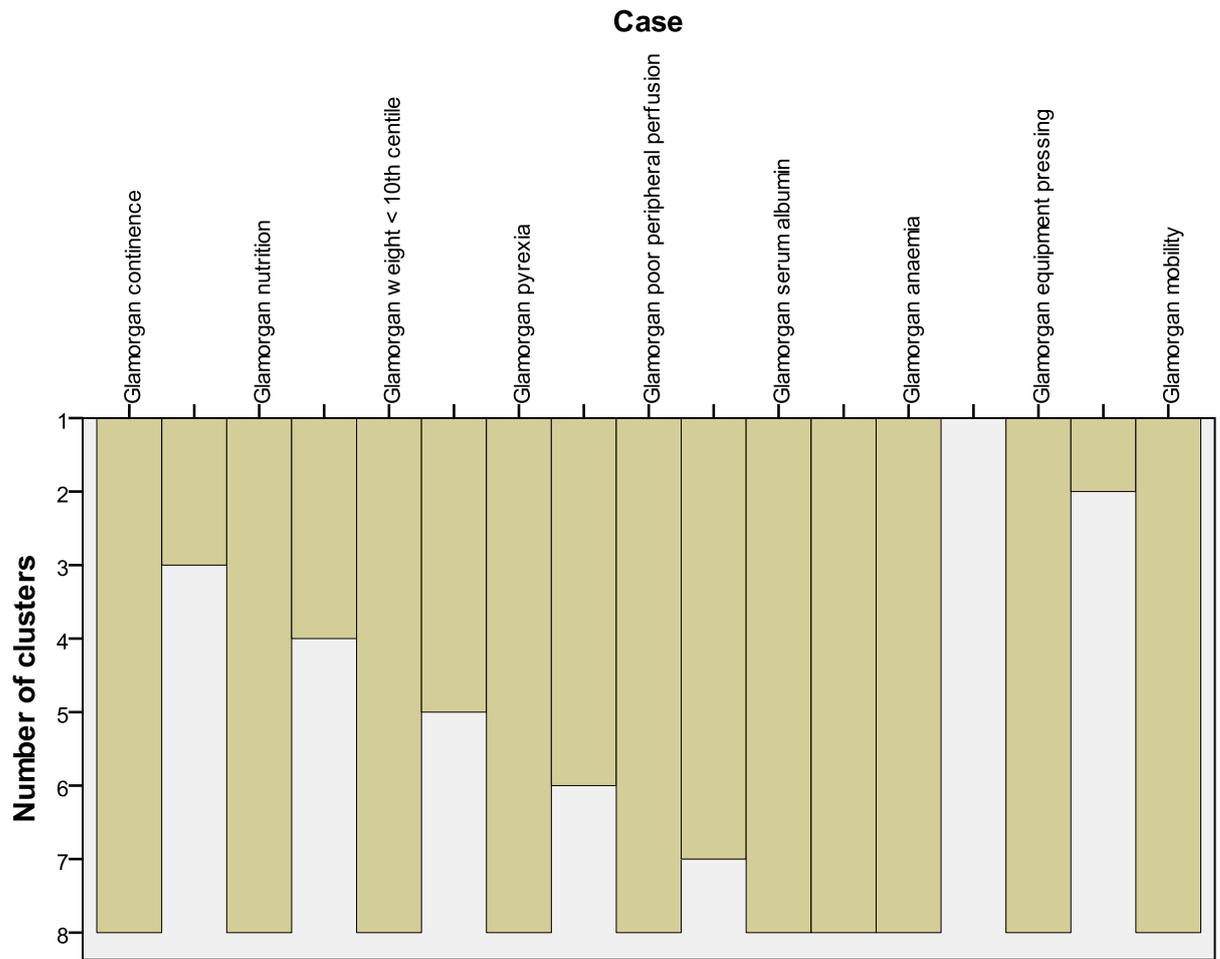
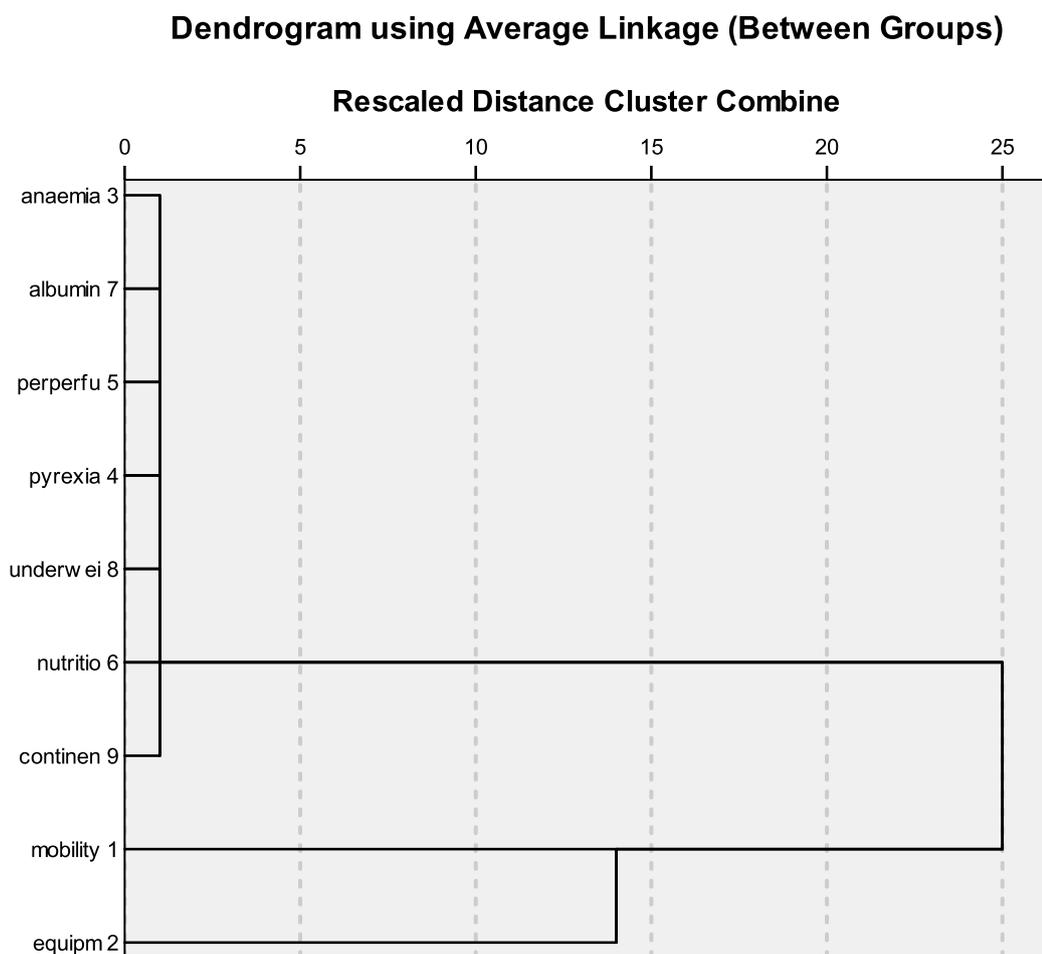


Figure 128: Dendrogram



I find the icicle plot difficult to follow, and prefer the dendrogram . This can be read from right to left, in which case equipment and mobility and nutrition seem linked in one cluster with all other items in another.

There are several ways of interpreting this diagram (one of the difficulties of cluster analysis), but it seems to me that it could be seen conceptually as consisting of one cluster that are largely intrinsic (fever, blood values e.g.) and the other extrinsic (nutrition and use of equipment e.g.).

### Example: Infant, perinatal and neonatal mortality (hierarchical clustering)

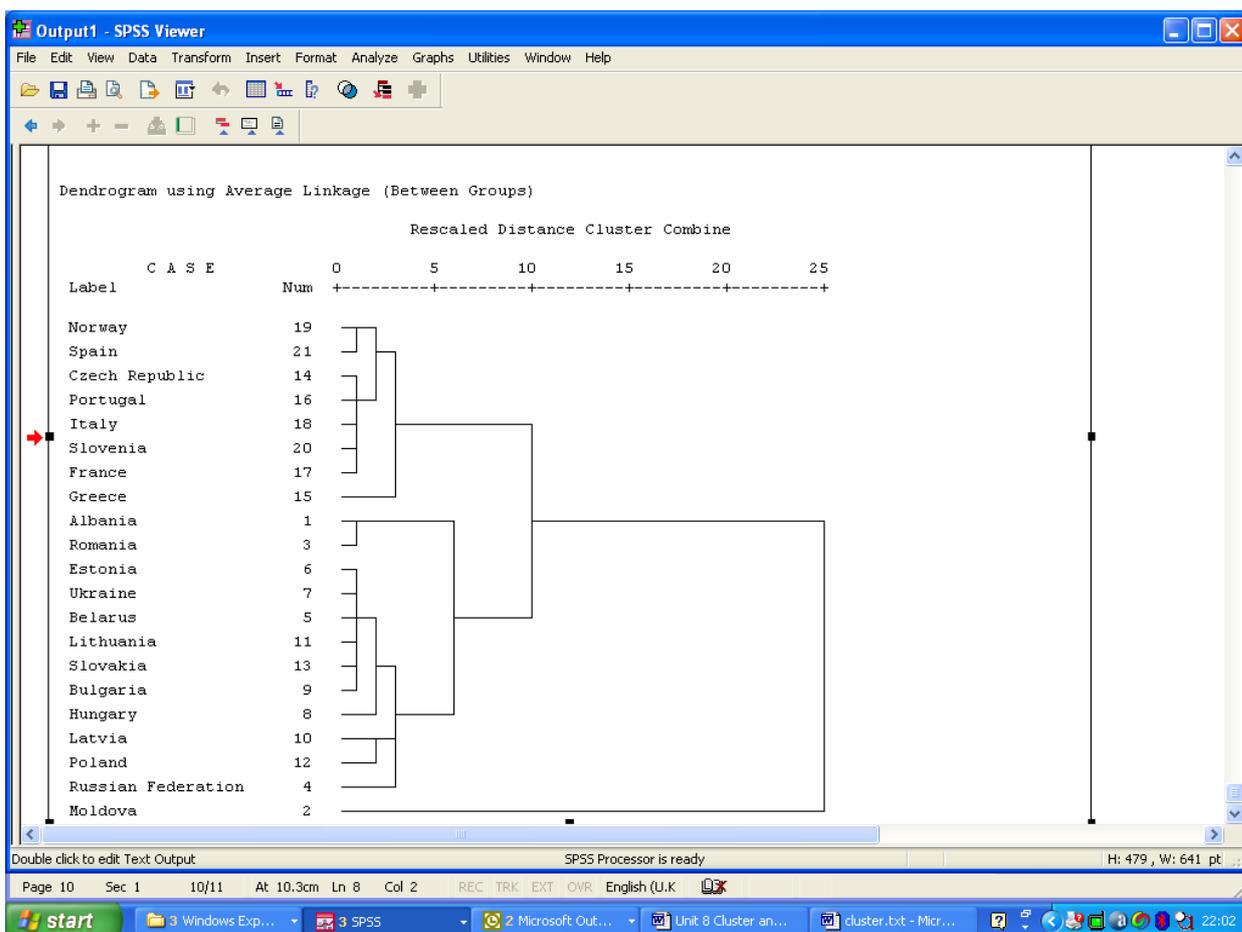
Let us look at the data in datafile “*perinatal*” (see Table 68) where we observe trends (some of the data from the smaller countries have been removed for clarity, but the overall picture is not altered by this). However, the three measures are not in the same rank order for all values. For example, Albania has the worst infant mortality, but Moldova has the worst perinatal mortality and Albania the fifth worst.

**Table 68:** Perinatal, neonatal and infant mortality around 1995 (data from WHO (WHO, 1996))

Country	Infant mortality	Perinatal mortality	Neonatal mortality rate
Albania	26	15	11
Moldova	25	30	15
Romania	23	15	10
Russian Federation	19	20	15
Belarus	17	15	10
Estonia	16	15	10
Ukraine	16	15	10
Hungary	15	10	10
Bulgaria	14	15	10
Latvia	14	20	10
Lithuania	13	15	10
Poland	13	20	15
Slovakia	12	15	10
Czech Republic	9	10	5
Greece	9	15	5
Portugal	9	10	5
France	7	10	6
Italy	7	10	5
Norway	7	5	5
Slovenia	7	10	5
Spain	7	5	5

We could use all three measures together to see which countries cluster together, see Figure 129.

Figure 129: Dendrogram for mortality data



Working from the right to the left of the dendrogram shown in the figure, Moldova appears as a cluster on its own. It has, in general, the worst figures of all, although not on every measure, as noted above. The dendrogram splits into two, with an almost complete split into western Europe and eastern Europe. In the eastern countries, Russia, Poland and Latvia cluster closely, and most of the former state socialist countries that were outside the borders of the USSR (Slovakia, Lithuania, Estonia, etc.) are also close to each other. You will note that the western countries have very few levels, and are thus much more similar to each other than are then eastern countries. The only former state-socialist countries that are in the western cluster are Slovenia and the Czech Republic. These data can be interpreted as a massive difference between eastern and western Europe with respect to infant, perinatal and neonatal mortality, but with much more diversity in the east. Note that the clustering algorithm had no information on the geography of the countries, merely the mortality data.

Further interpretation is possible given extra information, For example, the two former state-socialist countries (Slovenia and the Czech Republic) were among the richer of the former eastern bloc countries. They also had much more liberal economic policies, and therefore would have less difficulty adjusting to the post-Communist transition, and they are geographically close to rich western countries (e.g. Austria and Germany) with whom they could more easily trade.

## Problems with cluster analysis

Optimization techniques typically require large amounts of computing time, and in the case of neural networks scale badly, are not guaranteed to give a solution, and any solution found may be difficult to analyse.

It is difficult to choose the optimal metric in hierarchical clustering, the solution is sensitive to the chosen metric, and the clustering may be difficult to interpret.

## References

EVERITT, B. 1980. *Cluster analysis*, London, Heinemann.

WHO 1996. *Perinatal mortality*. , Geneva, World Health Organization.



www.sylvania.com

We do not reinvent  
the wheel we reinvent  
light.

Fascinating lighting offers an infinite spectrum of possibilities: Innovative technologies and new markets provide both opportunities and challenges. An environment in which your expertise is in high demand. Enjoy the supportive working atmosphere within our global group and benefit from international career paths. Implement sustainable ideas in close cooperation with other specialists and contribute to influencing our future. Come and join us in reinventing light every day.

Light is OSRAM

**OSRAM**  
**SYLVANIA**



Download free eBooks at [bookboon.com](http://bookboon.com)



# 19 Introduction to power analysis

## Key points

- You need to consider power to decide how big a sample you need

At the end of this chapter you should be able to:

- Work out your sample size

## Introduction

Power analysis is the science of finding out how big your sample needs to be. This, for any given test (chi square, Student's t test for independent groups etc.) depends on four basic factors:-

- $\alpha$ : the alpha level, conventionally set to 0.05
- $1-\beta$ : the power, conventionally set to 0.8
- $n$ : the sample size
- $e$  effect: the size of difference in means, the strength of correlation, or some other measure of the strength of the effect

This is introducing a new term. We know that  $\alpha$  is the probability of a Type I error, or the probability of a result suggesting an effect is real, when there is no effect. An alternative way to say this is that  $\alpha$  is the probability of incorrectly rejecting the null hypothesis. The probability of getting a Type II error is  $\beta$ , or the probability of missing a real effect, i.e. not reaching the alpha level where a real effect exists. Power is the probability of correctly rejecting the null hypothesis, and clearly that is the inverse of  $\beta$ , as  $\beta$  is the probability of incorrectly accepting the null hypothesis. If you find all this confusing then either:-

The null hypothesis is true (there is no effect) in which case there are two possibilities.

- a) The p value returned is less than  $\alpha$ , indicating the null hypothesis should not be accepted, which is wrong, this is a type I error.
- b) The p value returned is more than (or equal to)  $\alpha$ , indicating the null hypothesis should not be accepted, which is correct.

The null hypothesis is not true (there is an effect) in which case there are two possibilities.

- c) The p value returned is less than  $\alpha$ , indicating the null hypothesis should not be accepted, which is correct.
- d) The p value returned is more than (or equal to)  $\alpha$ , indicating the null hypothesis should not be accepted, which is incorrect, this is a Type II error.

The probability of these four outcomes (there are none other) are:-

- a)  $\alpha$
- b)  $1-\alpha$
- c)  $1-\beta$
- d)  $\beta$

This must be true since the two possibilities (null hypothesis is true or not true) in each case must add to unity.

If we had data available we could estimate what effect size we could expect. However sometimes there are no data and in these cases Cohen (1989) gives conventions for different tests for what constitute small, medium and large effects. For example the effect size for Student's t test for independent groups is the number of standard deviations that the two means differ by, or  $d=(\mu_1-\mu_2)/\sigma$  where  $d$  is the effect size,  $\mu_1, \mu_2$  are the means and  $\sigma$  is the standard deviation. Cohen gives small as 0.2, medium as 0.5 and large as 0.8. Large effects are easy to detect, and therefore cheap to detect as they need fewer subjects. Small effects are conversely difficult and expensive as we need larger sample sizes. In any given situation there is an effect size that we probably don't care about, or in other words it is practically (or clinically) not significant. For example a drug that reduces blood pressure by 5 mm Hg is useful, one that reduces it by 0.5 mm Hg is probably not. What effect size is important is a subjective judgment and uses knowledge of the subject under scrutiny. The cost of conducting the study will clearly be an important issue.

The test is always a given as it is whatever test is appropriate to analyse your data. If you set three of four factors, the fourth can be determined. The alpha level is normally set to 0.05. This leaves sample size effect size and power. The most common forms of power analysis are:-

A priori: conducted before the study. Most commonly given  $\alpha$ ,  $\beta$  and effect size, the sample needed to detect that effect size is computed.

Post hoc: conducted after the study. Most commonly given  $\alpha$ ,  $e$  and  $n$  determine how much power the study had.

## Power analysis with tables

Books such as Cohen (1989) have tables giving a range of power against sample size for specific tests, and for given  $\alpha$  and effect size. You can read off the sample size for a given power, effect and alpha level (say) by reading across in the relevant table.

## Free power analysis software

In addition to several packages that you can buy, these two are completely free. PS from [biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize](http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize) and Gpower from [www.psych.uni-duesseldorf.de/aap/projects/gpower/](http://www.psych.uni-duesseldorf.de/aap/projects/gpower/), the latest version is a Windows package at [www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/](http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/).

I have been using gpower for a decade. It works fine, I still use the MSDOS program. I have just started using PS. In some respects I prefer gpower, even the non-Windows version. But PS seems to work well and has good graphics. Both cover the basic tests only, though this is usually all you need. If you need power analysis of exotic tests you will need to buy one of the proprietary packages.

### Example: chi square

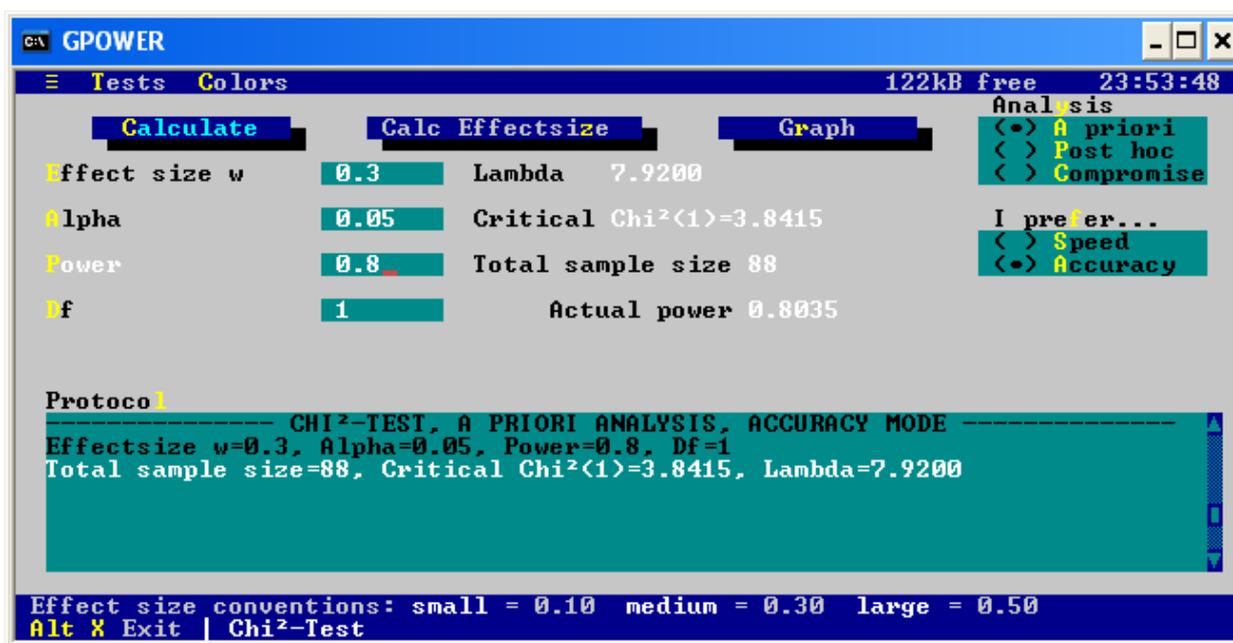
$p(O_i)$  is the probability of the observed value of the  $i$ th cell. This is clearly closely related to the chi square statistic. Do not worry if you do not immediately comprehend the following section. The important bit is that when observed and expected values are far apart the chi square statistic is big, and so is  $w$ .

$w$  is defined as:-

$$W = \sum (p(O_i) - p(E_i))^2 / p(O_i)$$

Suppose we wanted to plan how many patients we should recruit to see if males are different to females with respect to repetition of self harm. This is the a priori question. Using gpower (I am using the DOS version, but the Windows version is basically the same) I find if we wanted to only detect a large effect then a sample of 32 would suffice. However that might mean that a medium effect is missed, and given the seriousness of the subject (people may kill themselves) we probably want to detect smaller effects. A medium effect can be detected with 88 subjects (see Figure 130) and a small effect with 785. So our study with more than 1000 subjects is unlikely to miss even a small effect.

Figure 130: gpower screen for computing sample size for medium effect in chi square



PS is an alternative power analysis package, also free, and deals with effects a little differently. For example for Student’s t test you enter the difference between means and the standard deviation (which will be used to determine effect size) and also the ratio of control to experimental patients. Clearly the authors have clinical trials in mind, but where the numbers are equal the ratio is simply 1.0.

What effect size might our study have missed? This is the post hoc question. We know from our descriptive statistics that 61.7% of males repeated compared to only 44.2% of females. The ratio of males to females was roughly 1.0, see Table 69.

Table 69: Self harm repetition and gender  
gender \* repeat overdose Crosstabulation

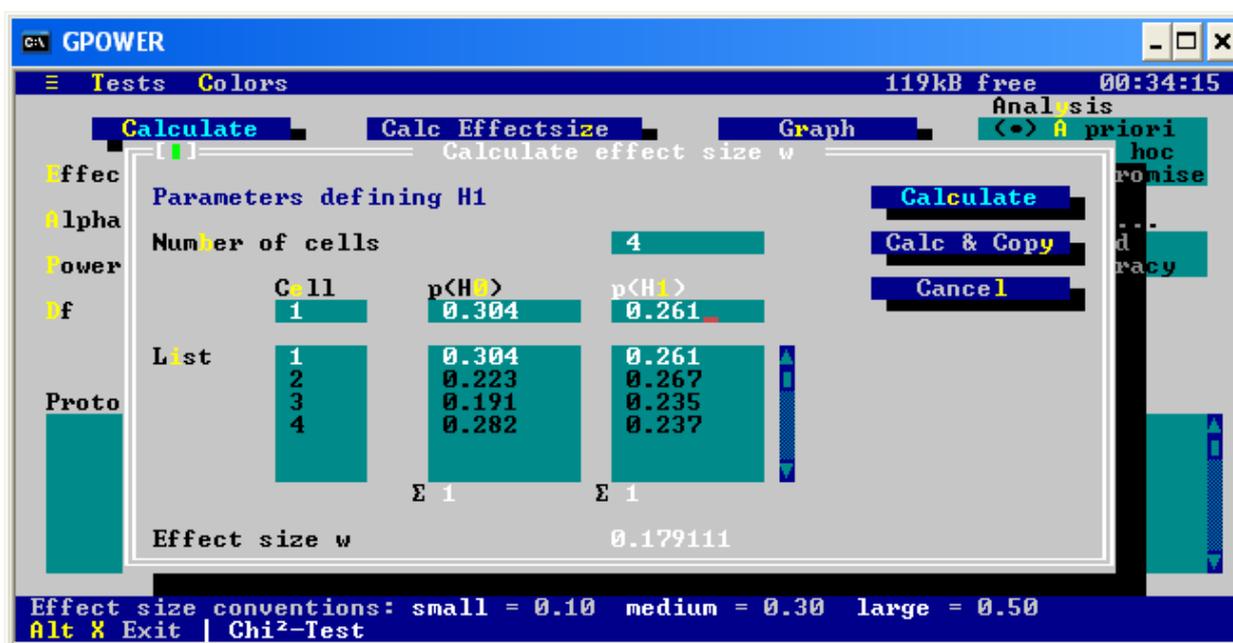
			repeat overdose		Total
			yes	no	
gender	male	Count	308	191	499
		Expected Count	263.9	235.1	499.0
	female	Count	223	282	505
		Expected Count	267.1	237.9	505.0
Total		Count	531	473	1004
		Expected Count	531.0	473.0	1004.0

Since the overall total is almost exactly 1000, we can create the probabilities for observed and expected by taking all values back 3 decimal places:-

			repeat overdose	
			yes	no
gender	male	Prob	0.308	0.191
		Expected prob	0.263	0.235
	female	Prob	0.223	0.282
		Expected prob	0.267	0.237

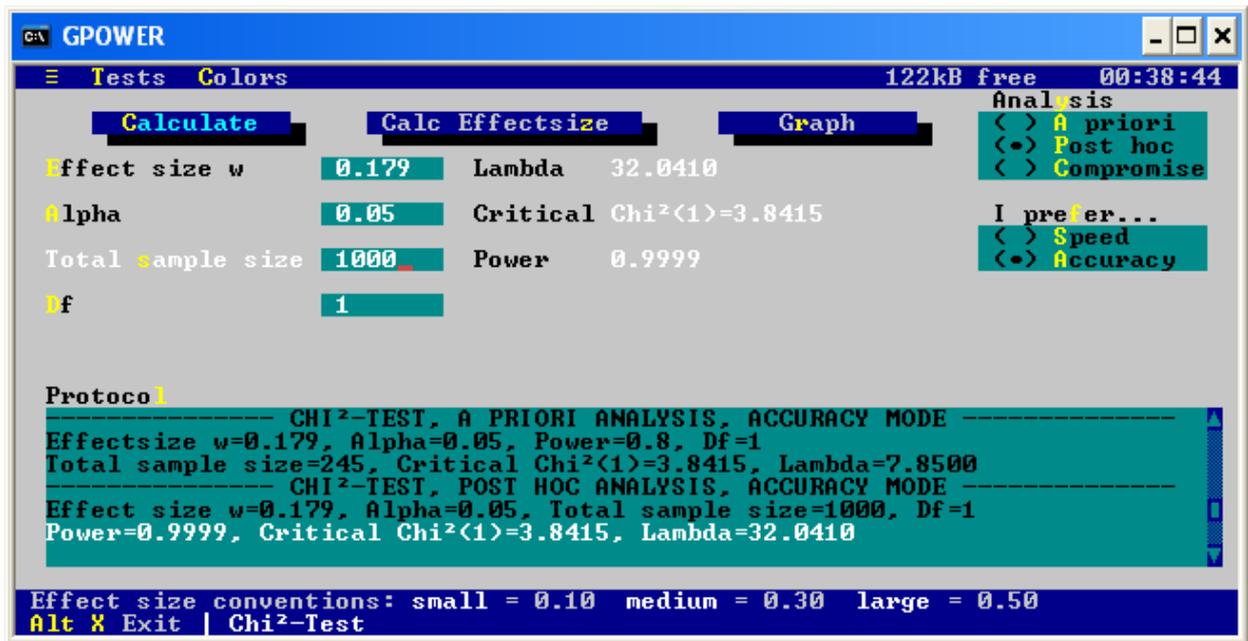
I can use these values to compute the effect size, see Figure 131.

Figure 131: calculating observed effect



Note I have cheated a little here, as the numbers I approximated above add up to a little over 1.0 for each column, and the program does not allow this, so I moved down a fraction the top two numbers to allow a calculation to proceed. This does not materially change anything. I can now calculate the power (a post hoc examination) given this effect size and the sample size I had, see Figure 132.

Figure 132: Calculating power



What we see is that even with the rather small effect we noted (between small and medium), there is a vanishingly small chance of missing this effect with a sample of 1000. In fact power is given as 0.999. Alternatively to see this effect with a conventional power we only needed  $n=245$ , see Figure 133.

**CHALLENGING PERSPECTIVES**

## Internship opportunities

EADS unites a leading aircraft manufacturer, the world's largest helicopter supplier, a global leader in space programmes and a worldwide leader in global security solutions and systems to form Europe's largest defence and aerospace group. More than 140,000 people work at Airbus, Astrium, Cassidian and Eurocopter, in 90 locations globally, to deliver some of the industry's most exciting projects.

An **EADS internship** offers the chance to use your theoretical knowledge and apply it first-hand to real situations and assignments during your studies. Given a high level of responsibility, plenty of learning and development opportunities, and all the support you need, you will tackle interesting challenges on state-of-the-art products.

We welcome more than 5,000 interns every year across disciplines ranging from engineering, IT, procurement and finance, to strategy, customer support, marketing and sales. Positions are available in France, Germany, Spain and the UK.

To find out more and apply, visit [www.jobs.eads.com](http://www.jobs.eads.com). You can also find out more on our **EADS Careers Facebook page**.

**AIRBUS** **ASTRIUM** **CASSIDIAN** **EUROCOPTER**

**EADS**

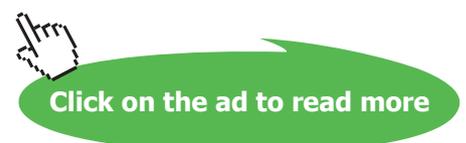
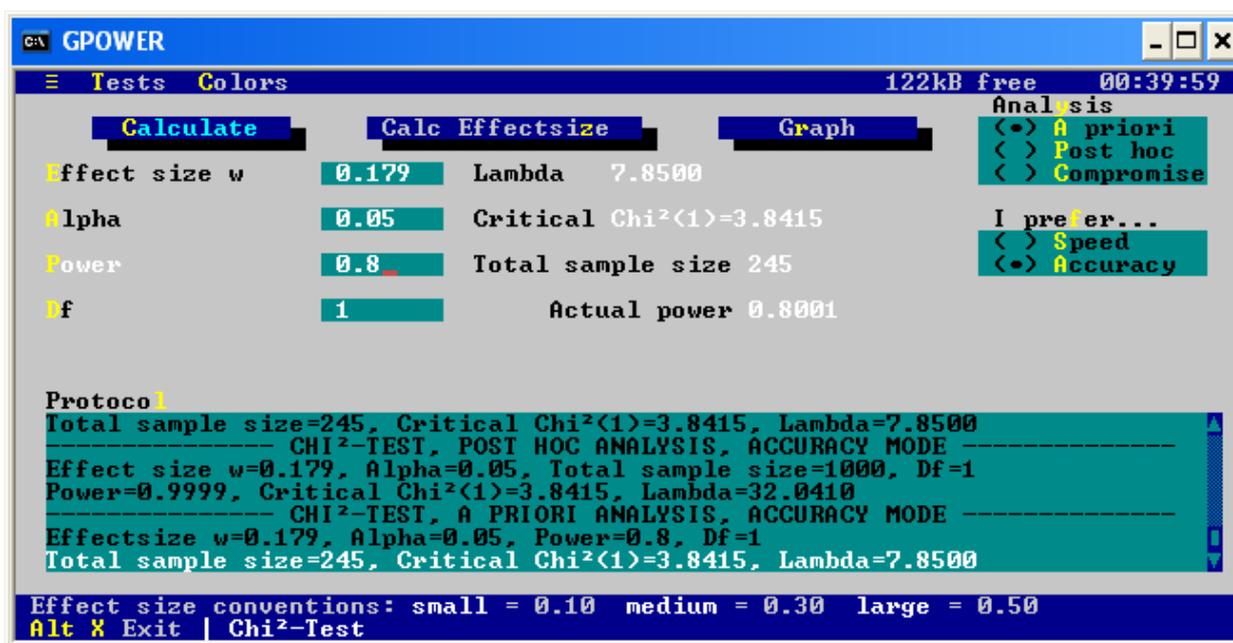


Figure 133: calculating sample size



What if we wanted to detect a really small effect. The situation is very different; the power for an effect of 0.05 and a sample of 1000 is now only 0.35. Alternatively we could say that we are only about 35% confident of locating such an effect. However to detect this we would need a sample of 3140. Therefore we would need to collect about three times the data we have. But is it worth it? That is a question of resources. What would we do if we knew there was a tiny difference in repeated self harm in males and females? If the answer is nothing then it is not worth collecting the large sample.

### Exercise

You want to conduct a Student's t test for independent groups. You are testing whether a drug reduces blood pressure compared to a control. A clinically significant difference has been stated to be a reduction of 5 mm Hg in the diastolic pressure. The standard deviation in previous groups, where you have access to the data, is 20. Look at the graph in Figure 134 that I created from PS. What power (roughly) would you have with a sample of 100 subjects? What sample size would you need (roughly) to detect the clinically significant difference with a power of 0.8? Now look at Figure 135. How big a sample (roughly) would you need to detect a fall of 10 mm Hg, and how many (roughly) to detect a 2.5 mm Hg.

Figure 134: sample size against power

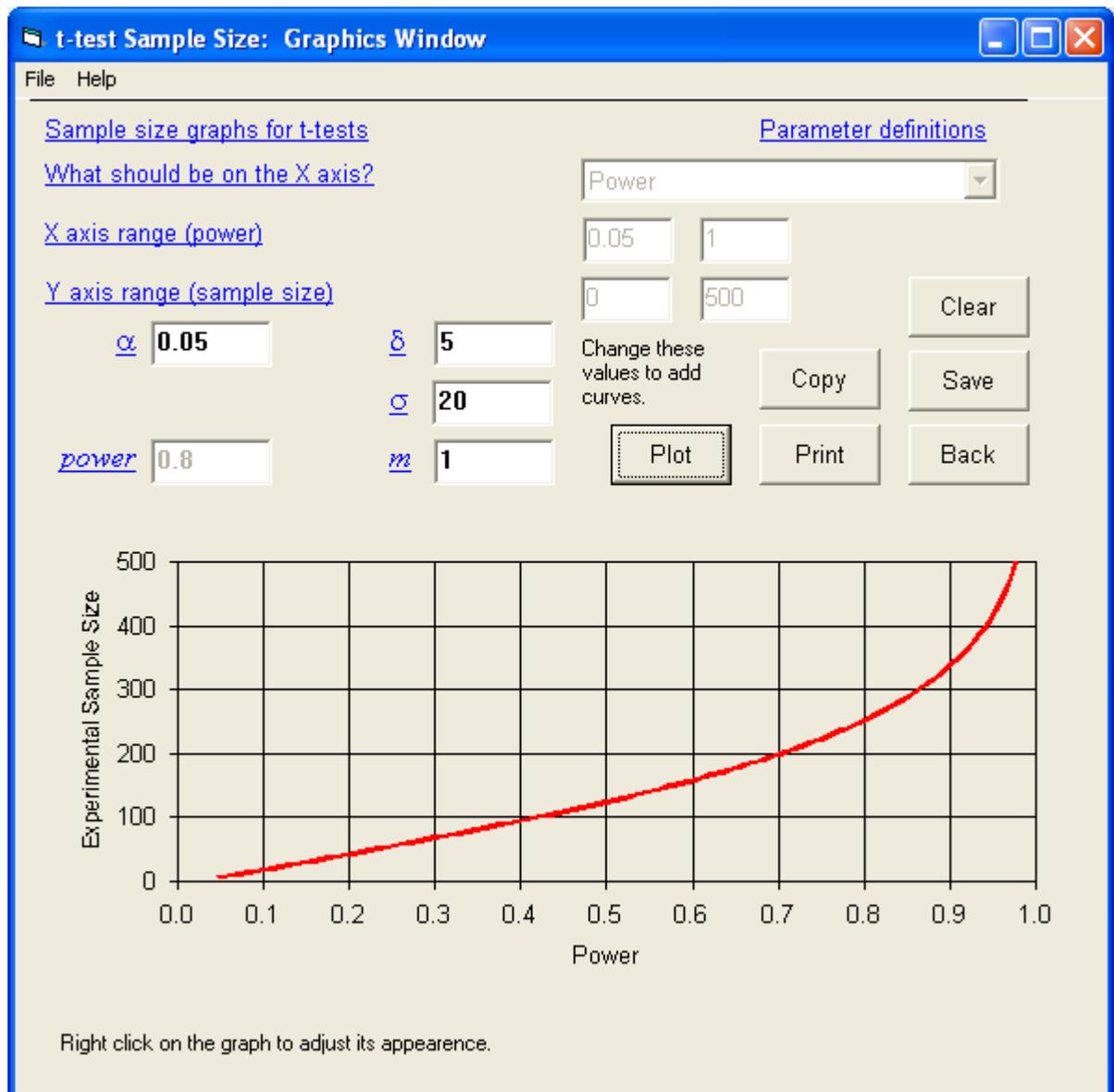
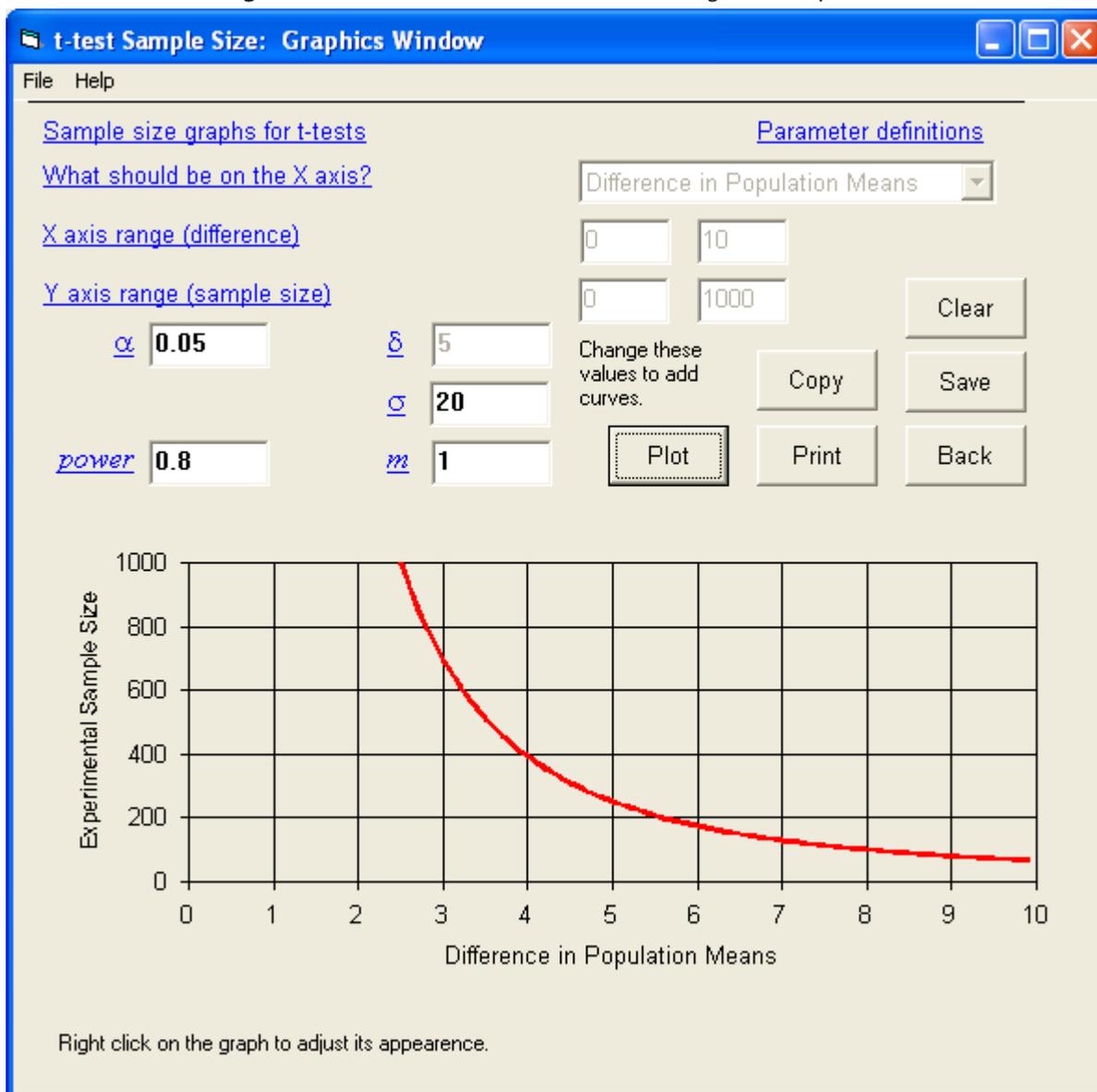


Figure 135: Effect sizes (difference in means) against sample size



### Additional tests

Some tests are not given in the free programs. But all is not lost. For some tests like Mann Whitney, that do a similar job to Student's t test may be calculated from the output of (for Mann Whitney) the Student's t test for independent groups. This is because of the asymptotic relative efficiency (ARE) which is the ratio of power of one test over the other. The ARE for Mann Whitney relative to the Student's t test for independent groups is 0.955. Thus as the power of the latter for a given scenario is  $(1-\beta)$  it will be  $0.955(1-\beta)$  for Mann Whitney. Thus Mann Whitney is less powerful than the parametric test, but not much less.

Some additional AREs are (<http://www.tufts.edu/~gdallal/npar.htm>):-

sign test	$2/\pi = 0.637$
Wilcoxon signed-rank test	$3/\pi = 0.955$
median test	$2/\pi = 0.637$
Wilcoxon-Mann-Whitney U test	$3/\pi = 0.955$
Spearman correlation coefficient	0.91

Exercise

If you wanted to use Mann Whitney for a medium effect and standard  $\alpha$  and power, using Figure 136 what power (roughly) would you get with a sample of 100?

360° thinking.

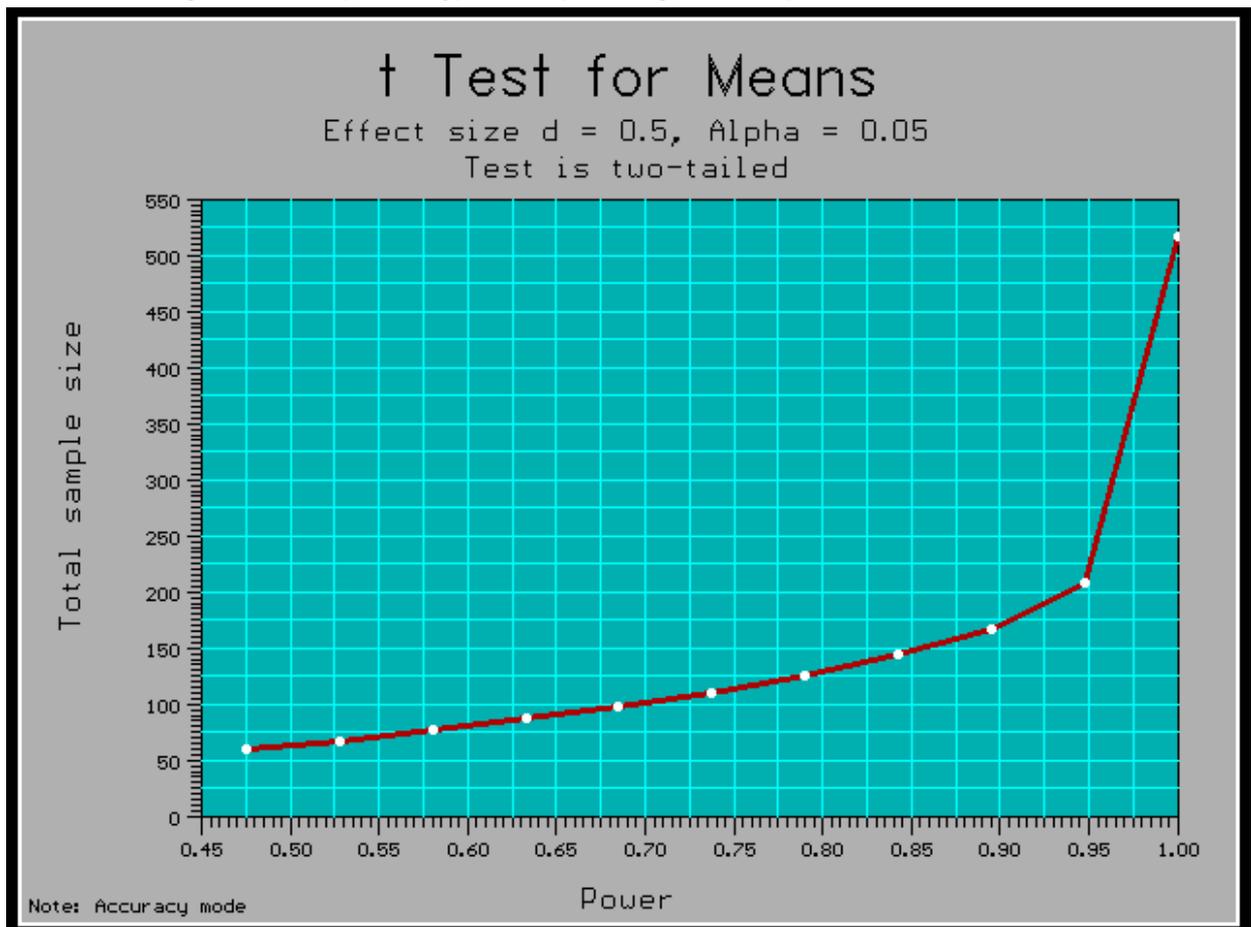
**Deloitte.**

Discover the truth at [www.deloitte.ca/careers](http://www.deloitte.ca/careers)

© Deloitte & Touche LLP and affiliated entities.



**Figure 136:** Graph from gpower of power against sample size for medium effect



### Further Reading

Cohen’s large book (Cohen, 1989) is the bible of power analysis. It contains tables on many tests, and is the basis of many computer programs. Miles (n.d.) has a nice webpage explaining power analysis.

### References

COHEN, J. 1989. *Statistical Power Analysis for the Behavioural Sciences. 2nd Ed.* , Hillsdale NJ, Erlbaum.

MILES, J. n.d. *Getting the Sample Size Right: A Brief Introduction to Power Analysis* [Online]. Available: [www.jeremymiles.co.uk/misc/power/index.html](http://www.jeremymiles.co.uk/misc/power/index.html) [Accessed].

# 20 Which statistical tests to use

## Introduction

A common question is “what test should I use” and here I detail for all the common ones, what test is appropriate. These are the tests covered in chapters six to ten.

If you need to test an hypothesis, which one should you use? The question really should go one stage back, what is your research question? This will determine whether you even need an hypothesis, and then you can look to see what test is most appropriate. Some research questions are answered by descriptive statistics, others that are qualitative in nature do not need a statistical treatment at all (e.g. grounded theory, phenomenology, narrative biography)

The sorts of research designs we have looked at in this module are:-

Cross tabulation. E.g. you want to test whether males are more likely to be promoted than females in universities.

- Null hypothesis: The proportions of males and females promoted are not significantly different.

Independent groups. You want to test whether two or more groups are different with respect to a test variable. E.g. you want to test if males access on online course more, less or the same as females.

- Null hypothesis: There is no significant difference in number of accesses to an online course between the two genders.

Repeated measures. You want to see if two or more measurements on the same subject are different. E.g. you want to see if an online course improves the knowledge of statistics in a cohort of students.

- Null hypothesis: The marks in a test to measure statistics knowledge pre-course are not significantly different from those obtained post course.

Correlation. E.g. you want to see if marks in a course are correlated to pre-course educational attainment.

- Null hypothesis: There is no significant correlation between pre-course educational attainment (degree classification) and grade awarded in a postgraduate module.

Note we are not saying in the null hypotheses that there are no differences, correlations etc. We are saying in the null hypotheses that these effects are not significant, ie. Any effects are explicable by random error. Also we are not saying we believe the null hypothesis, we are using the null hypothesis as something we can test.

Let us summarise the tests we have encountered so far

- Chi square. This is the only test that looks at frequencies of nominal data. It employs cross tabulation. It is a weak test, but the only one we have discussed that deals with nominal data (there are others, but by far this is the most common). So if you have nominal data only, think chi square.
- Student's t test for independent groups. This tests differences of data that are at least ordinal and split into two groups. It is a parametric test (it assumes a normal distribution).
- Mann Whitney. This also tests differences of data that are at least ordinal and split into two groups. It is a non-parametric test (it does not assume a normal distribution).
- One way analysis of variance (one way ANOVA). This tests differences of data that are at least ordinal and split into more than two groups. It is a parametric test.
- Kruskal Wallis. This also tests differences of data that are at least ordinal and split into more than two groups. It is a non-parametric test.
- Student's t test for paired data. This tests if two measures that are at least ordinal on the same subject are different. It is a parametric test.

SIMPLY CLEVER **ŠKODA**  


**We will turn your CV into an opportunity of a lifetime**



Do you like cars? Would you like to be a part of a successful brand? We will appreciate and reward both your enthusiasm and talent. Send us your CV. You will be surprised where it can take you.

Send us your CV on [www.employerforlife.com](http://www.employerforlife.com)





- Wilcoxon test. This also tests if two measures that are at least ordinal on the same subject are different. It is a non-parametric test.
- Repeated ANOVA. This tests if more than two measures that are at least ordinal on the same subject are different. It is a parametric test.
- Friedman. This also tests if more than two measures that are at least ordinal on the same subject are different. It is a non-parametric test.
- Pearson's correlation co-efficient. This tests if two variables that are at least ordinal are correlated. It is a parametric test.
- Spearman rank. This also tests if two variables that are at least ordinal are correlated. It is a non-parametric test.

So let us construct a cookbook of tests. You will find many textbooks show this as a table, or flowchart. I will employ a text based algorithm.

## Algorithm for choosing test

### Check type of data

Do you have two variables? Are data from both variables nominal? Do you want to test if frequencies in a cross tabulation of the two variables are what you would expect if the variables are independent of each other? Use chi square.

### Testing difference in ordinal/interval/ratio data in different groups

Two groups? Are data normally distributed? Use Student's t test for independent groups, otherwise use Mann Whitney.

More than two groups? Are data normally distributed? Use one way ANOVA, otherwise use Kruskal Wallis

### Testing changes in a variable in same or related subjects

Two measures? Are data normally distributed? Use paired Student's t test, otherwise use Wilcoxon.

More than two measures? Are data normally distributed? Use repeated ANOVA, otherwise use Friedman.

### Correlation of two variables

Are data normally distributed? Use Pearson's correlation, otherwise use Spearman rank.

### Exercise

Match each of the studies in Table 70 to tests in Table 71.

You will need to make some assumptions to determine whether data are normally distributed and the type of data employed. Use common sense to make reasonable judgements. What is important is that you can make a reasoned assessment of what test is likely to be optimal. I suggest you make a precise research question, state the null hypothesis, state what type of data is needed, whether it is normally distributed, and then use these to pick the appropriate test.

**Table 70:** Study designs

Study design
a) Are days in hospital (length of stay) correlated to age?
b) Are four ethnic groups 3 different with respect to the percentage mark on a course?
c) Are four ethnic groups 3 different with respect to their attitude to online learning on a four point scale from 1 (like it a lot) to 4 (dislike it a lot)?
d) Are four ethnic groups different with respect to whether they pass or fail a course?
e) Are males and females different with respect to the percentage mark on a course?
f) Are males and females different with respect to their attitude to online learning on a four point scale from 1 (like it a lot) to 4 (dislike it a lot)?
g) Is attitude to online learning on a four point scale from 1 (like it a lot) to 4 (dislike it a lot) different before and after delivery of an online module?
h) Is blood pressure correlated to weight?
i) Is blood pressure reduced in a group of patients after relaxation exercises?
j) Is commitment on a four point scale (very committed, committed, not very committed, not committed at all) changed in a series of six monthly weight checks in a weight reduction programme.
k) Is weight reduced in a series of six monthly weight checks in a weight reduction programme.

**Table 71:** Tests

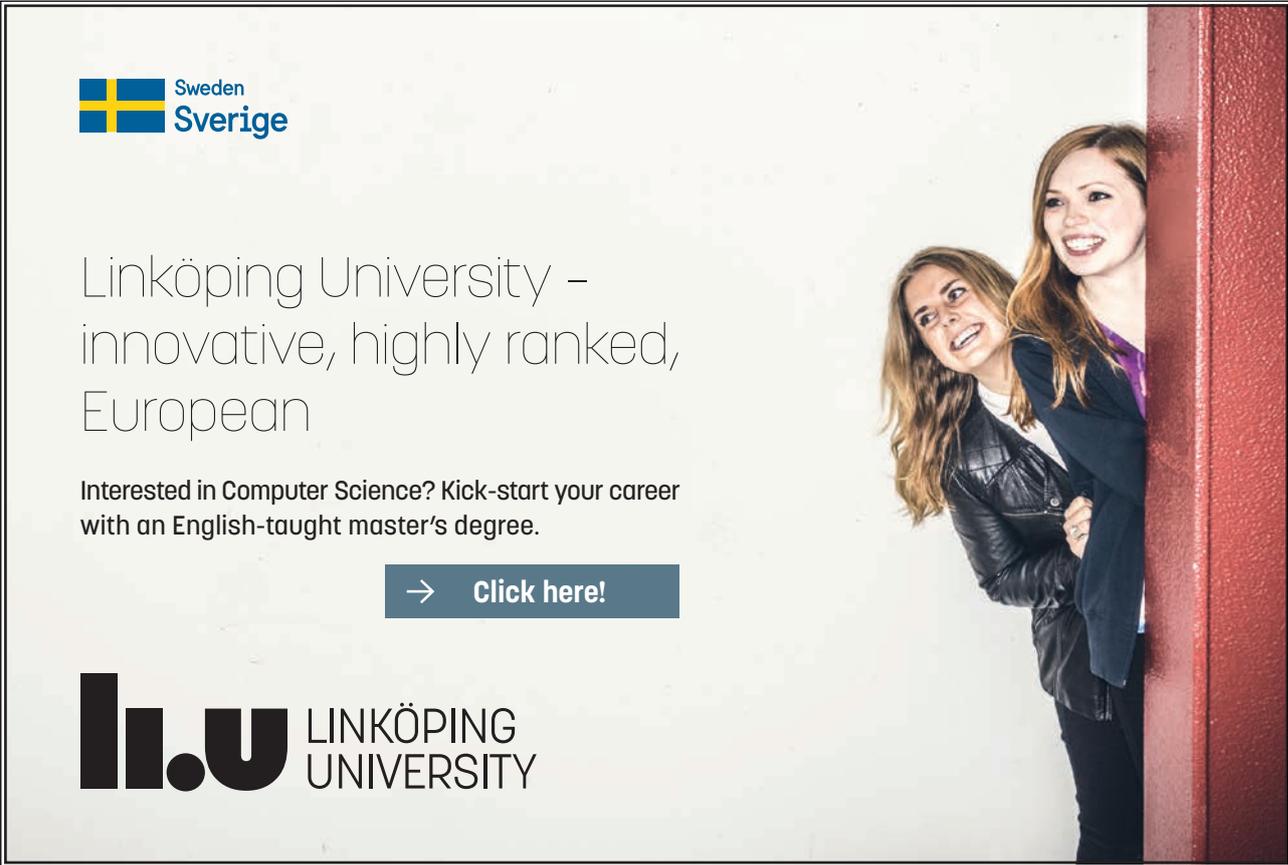
Test
a) Chi square
b) Student's t test for independent groups
c) Mann Whitney
d) Student's t test for paired data
e) Wilcoxon
f) One way ANOVA
g) Kruskal Wallis
h) Repeated ANOVA
i) Friedman
j) Pearson's correlation co-efficient
k) Spearman rank correlation co-efficient

# Endnotes

<sup>1</sup> I am using female for nurses and doctors in this unit. I realise that men do both jobs, but I do not like the use of “s/he”.

<sup>2</sup> Wikipedia is not to be relied upon totally as it is not peer reviewed. However while using a collaborative group to write material may not start with perfection, it is unlikely to end up with rubbish. I have found its pages on statistics very good and to date completely accurate. For obvious reasons more controversial subjects may be less free from bias or subjectivity. I tested this by looking pages on “abortion” and “Saddam Hussein”. You may check yourself how biased these pages are in your opinion. Pages in Wikipedia are not static, and while very useful should be used as a reference with caution, as the reader will be accessing a different text than the one I quote, in all likelihood, as the pages are dynamic.

<sup>3</sup> UK white, UK black, African Black, UK Asian



 Sweden  
Sverige

Linköping University –  
innovative, highly ranked,  
European

Interested in Computer Science? Kick-start your career  
with an English-taught master's degree.

→ Click here!

**li.u** LINKÖPING  
UNIVERSITY



# Answers to exercises

## Chapter 4

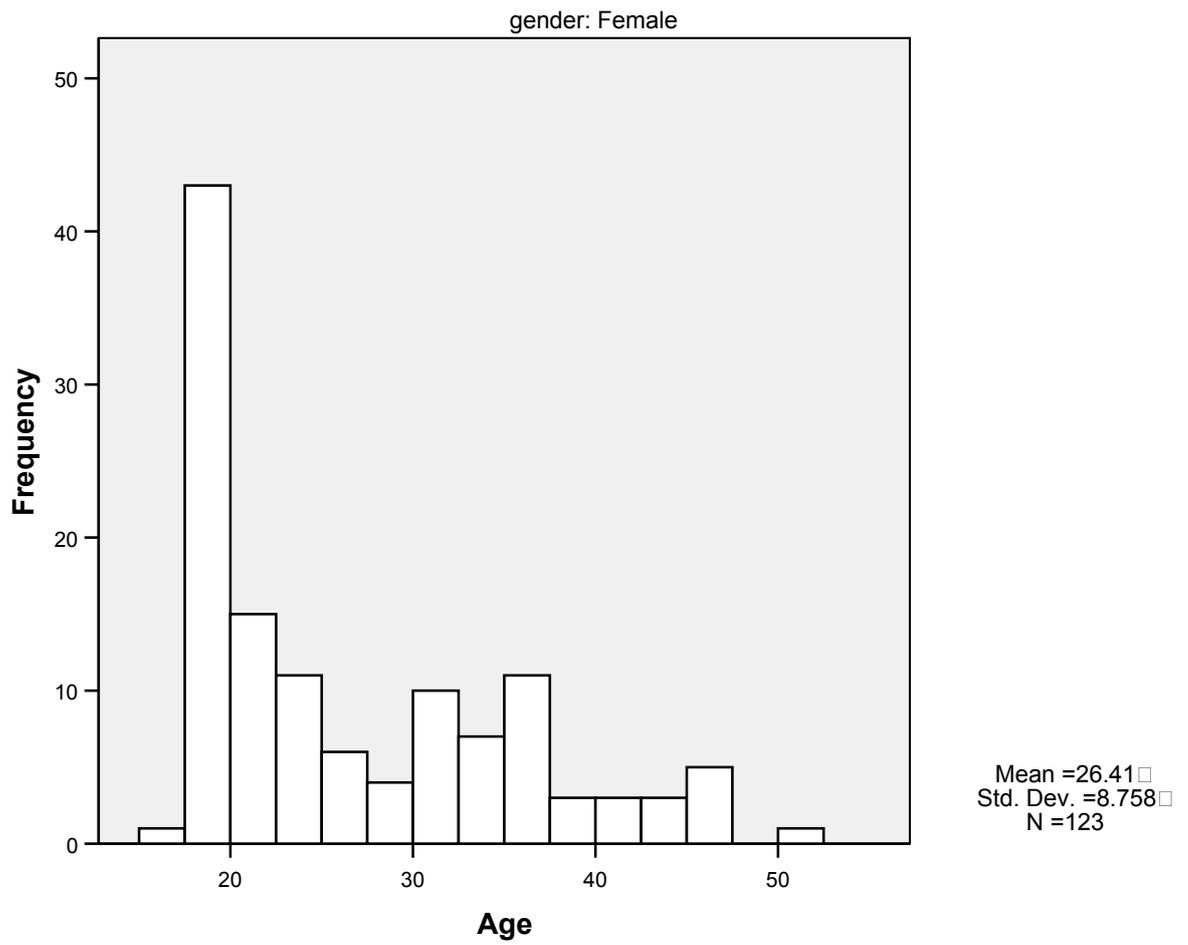
I collected data on ethnic groups (this is different from nationality) e.g. White British, Black or Black British, Asian or Asian British. Separately I obtained data on attitude to the online course ranging from very useful to useless on a four point scale. How would I best show these data to a teaching colleague to show them the general picture of ethnicity in this course, and also describe their attitudes to the relevance of online courses? What type of data is ethnic group, and what type of data is the attitudinal variable?

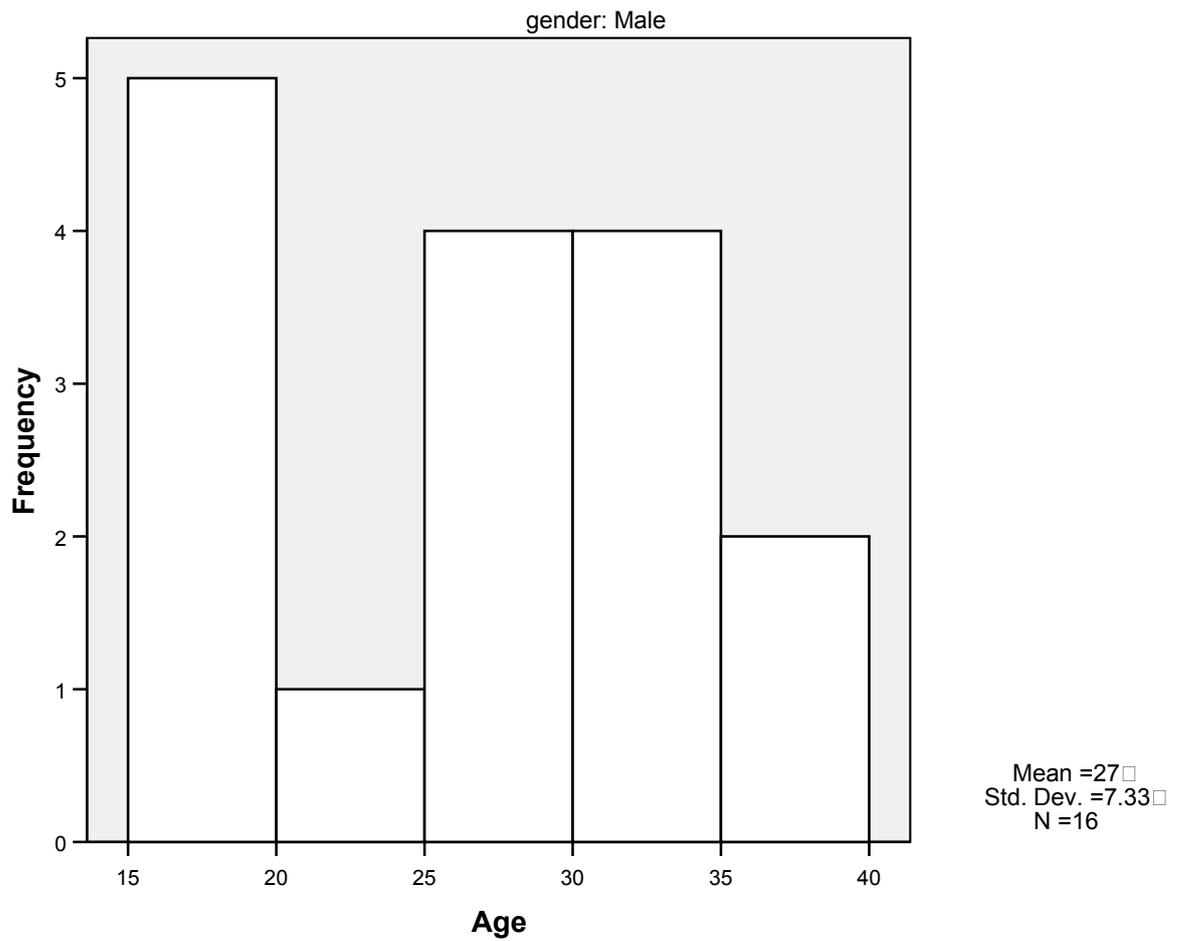
You could use a table or pie chart or bar chart to show either attitude or ethnicity. The former though is better shown as a bar chart than pie chart as the data are ordinal and the bar chart shows the bars starting at one end of the scale and moving to the other.

Ethnic group is nominal

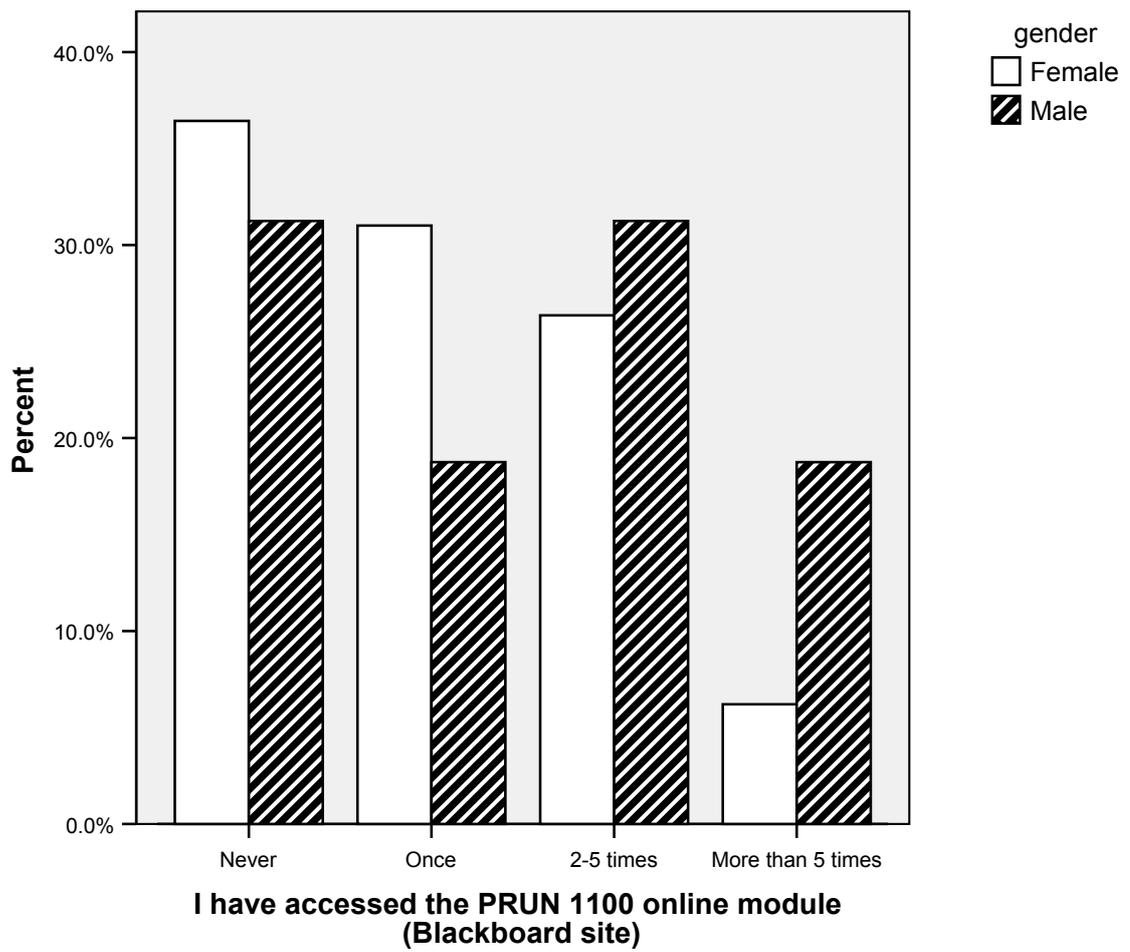
Attitude is ordinal

Split data by gender and do a histogram of age for each gender. Unsplit the data and perform a clustered bar chart using percentages of gender against frequency of accessing the course. Interpret the results of both operations.





There are fewer males than females. Also the “traditional” 18-21 year old undergraduate is more common in females with a much larger spike in this age band.



Males are proportionately more likely to access the web site more often.

## Chapter 5

Use the datafile “assets”, sort by gravity (descending) and find the most severe case that did not get a custodial sentence. This sort of exploration can identify “odd” cases.

All gravity scores of eight (the highest) had custodial sentences, but some of the scores of seven did not.

## Chapter 6

Males seem to repeat overdose more than females, according to the chi square analysis. However maybe males have more mental health problems, and it is this, rather than gender, that may be the real issue. Decide how to test this. You will need to create a cross-tabulation table and employ chi square to check this out.

**gender \* Previous psychiatric history Crosstabulation**

			Previous psychiatric history		Total
			no	yes	
gender	male	Count	330	218	548
		Expected Count	275.9	272.1	548.0
	female	Count	237	341	578
		Expected Count	291.1	286.9	578.0
Total		Count	567	559	1126
		Expected Count	567.0	559.0	1126.0

So more males have a mental health history than expected (as we got 330 and expected 276).

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	41.549(b)	1	.000		
Continuity Correction(a)	40.783	1	.000		
Likelihood Ratio	41.811	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	41.512	1	.000		
N of Valid Cases	1126				

a Computed only for a 2x2 table

b 0 cells (.0%) have expected count less than 5. The minimum expected count is 272.05.

and it is significant as Pearson chi square gives  $p < 0.001$

So it is plausible that males have higher levels of mental illness leading to more repetition of self harm in this sample.

## Chapter 7

Consider whether diploma or degree students were accessing the Blackboard online course site more frequently. Identify the most appropriate test, run in SPSS and comment to the discussion board with your interpretation.

The right test is Mann Whitney (since the Likert score will most likely not be normally distributed, and even if it were it would be difficult to be sure with only four levels).

### Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of I have accessed the PRUN 1100 online module (Blackboard site) is the same across categories of course.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Thus  $p < 0.001$  (not 0.000 as SPSS indicates) and therefore there is a significant difference. Since degree students have the higher mean rank, it is they who use the site more. N.B. you will need to change the test variable to be Scale as the designers in SPSS v18 for some reason do not want us to use ordinal data in these tests. I hope they are better programmers than statisticians! The clustered bar chart shows this:-

I joined MITAS because  
I wanted **real responsibility**

The Graduate Programme  
for Engineers and Geoscientists  
[www.discovermitas.com](http://www.discovermitas.com)



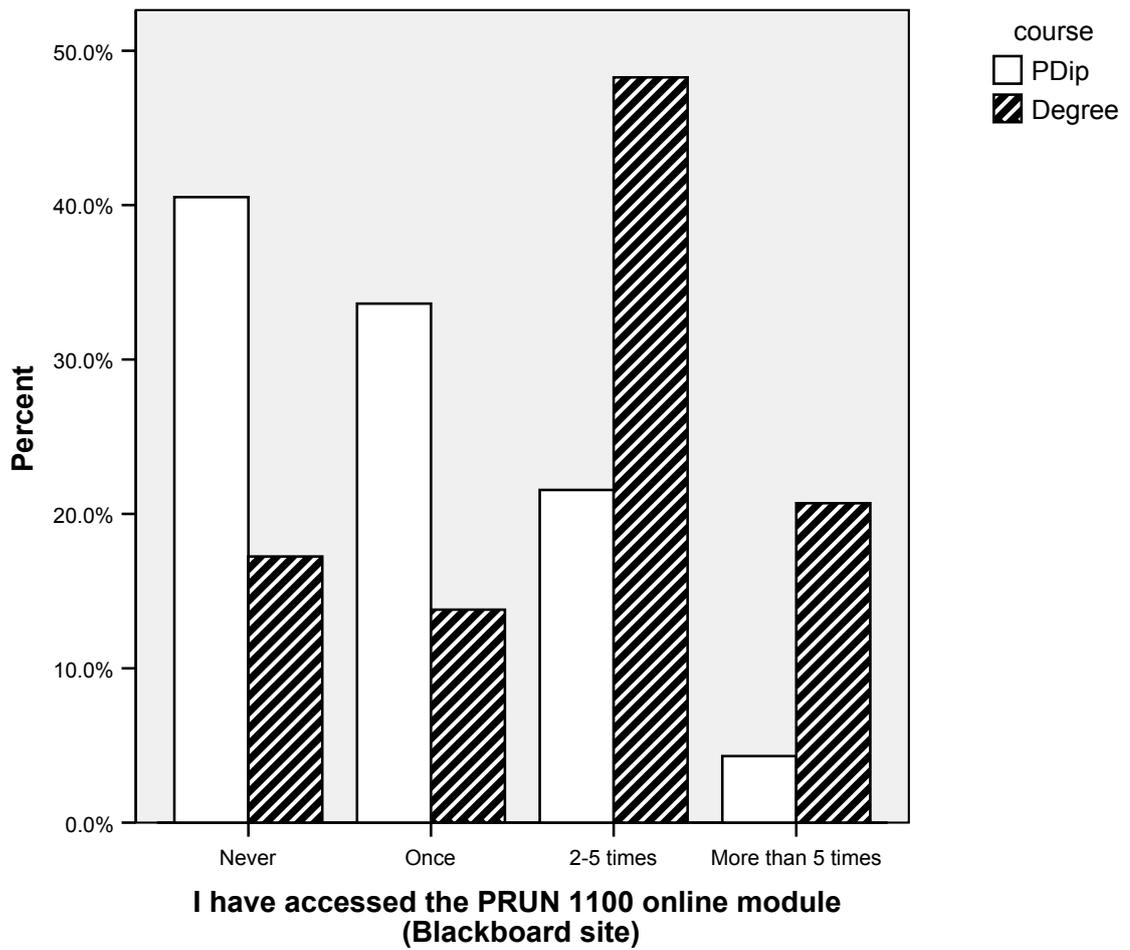
Real work  
International opportunities  
Three work placements



**Month 16**  
I was a construction  
supervisor in  
the North Sea  
advising and  
helping foremen  
solve problems







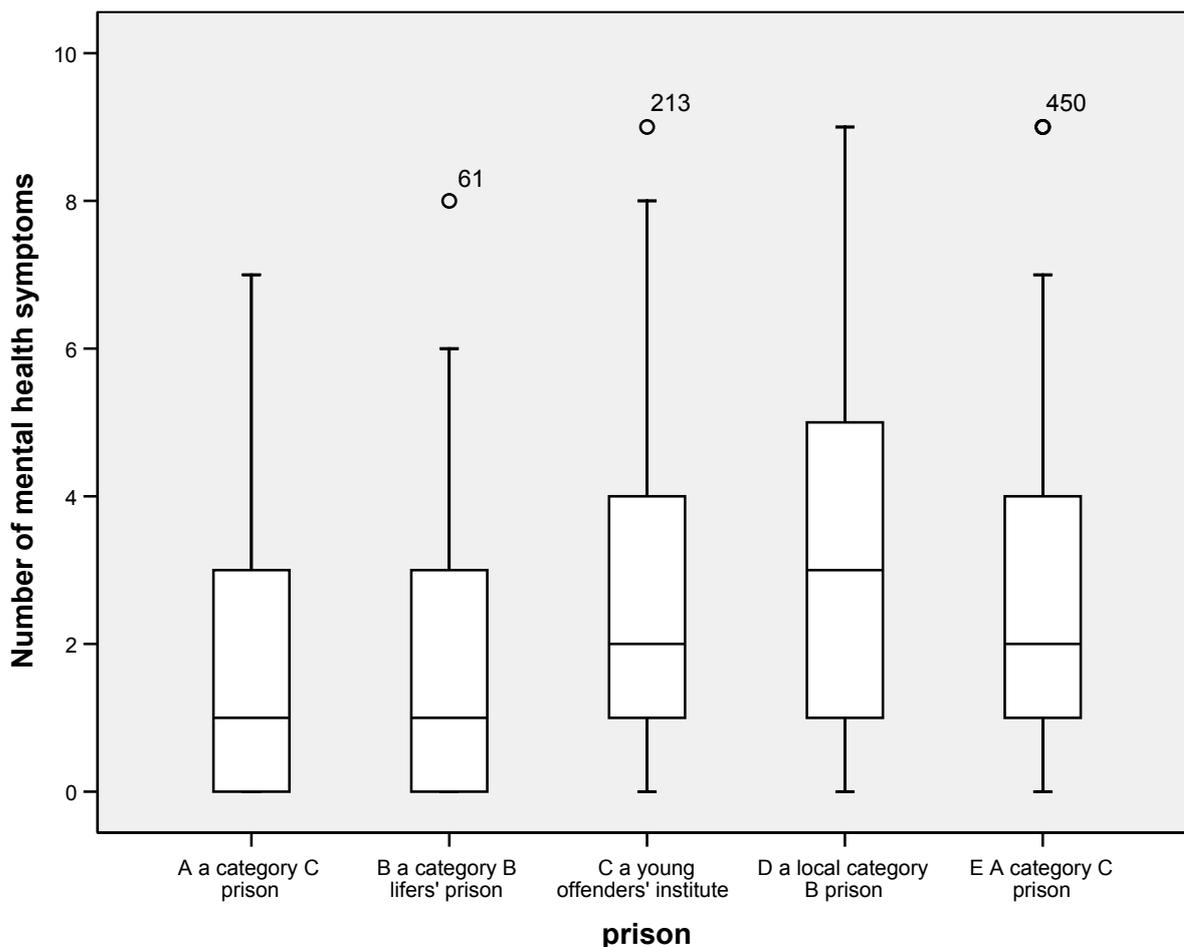
## Chapter 8

What does it mean to have a  $p=0.009$  for the between groups?

This is a  $p$  value  $< 0.05$  so there is a significant difference among the prisons with respect to number of symptoms

Go through the remainder of the table and decide which prisons are the same (not significantly different from each other) and which are significantly different.

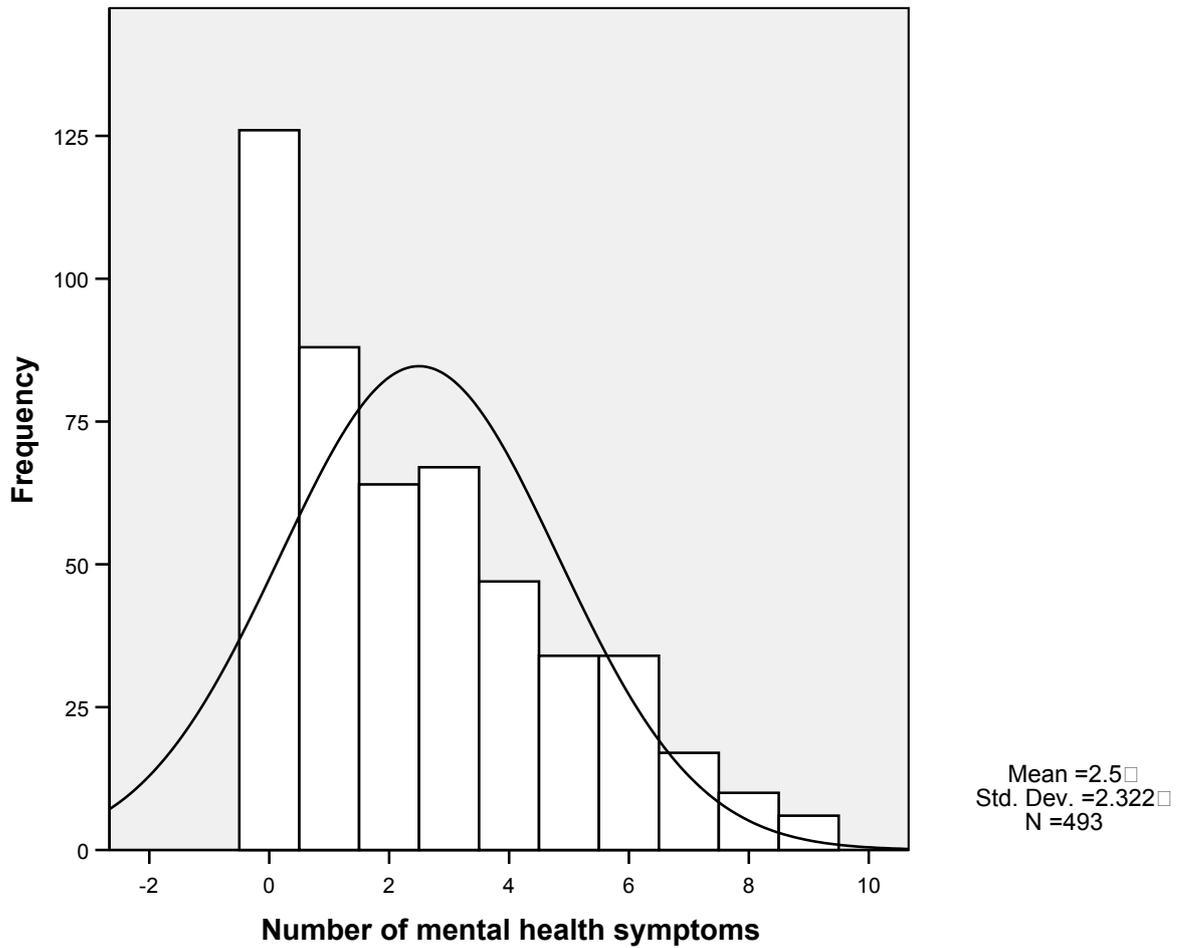
SPSS gives asterisks against those that are significantly different. So A and D are different from each other, all other prisons are not significantly different. This can be illustrated in a box plot:-



Oddly this seems to show A and B are much the same. Why therefore is B and D not significantly different? There are more respondents from A (50) than B (33) so this is probably due to the small number of responses from B. If we get more responses from B the p value might go down, but it could also go up. All we can say now is that although B is not significantly different to A, and A is significantly different to D, B is not.

Determine if the data for number of mental health symptoms are normally distributed.

The histogram shows it is clearly skewed to lower values. It is not normally distributed, note the normal curve superimposed is nothing like what we found.



There is a significant difference among the prisons, with D highest, A lowest and the other prisons somewhere in between. There is no post-hoc test available in Kruskal Wallis, so this is as far as we can go. However common sense suggests a similar interpretation as when we used the (inappropriate) parametric test.

## Chapter 9

Interpret the output of Table 23.

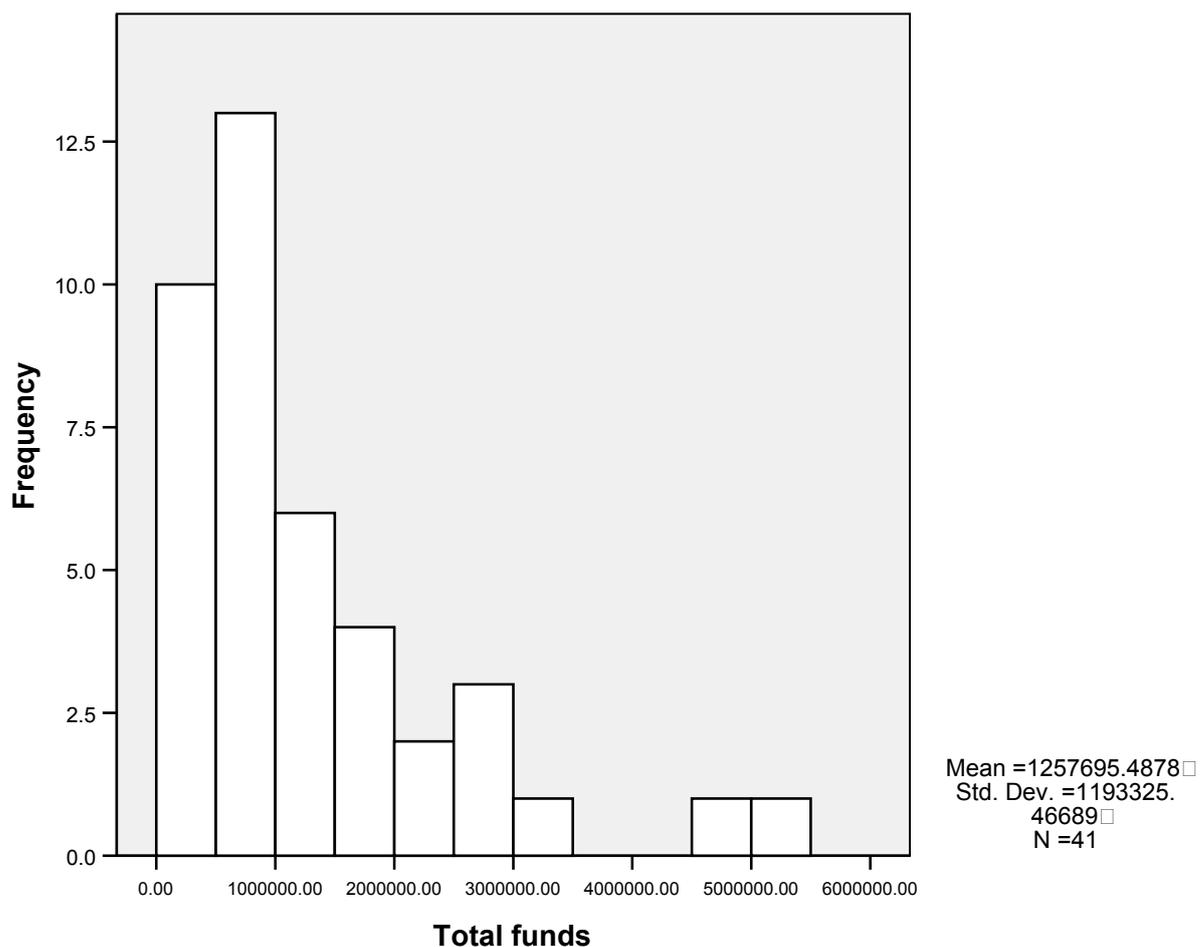
There is no significant correlation between age and (self-assessed) confidence in using computers. Therefore it is wrong to interpret the positive correlation as meaning older students are more confident.

Using the data in the RAE datafile consider the following hypotheses:-

- There is no significant correlation between full time equivalent (FTE) doctoral students and RAE rating
- There is no significant correlation between funding and RAE rating

Remember you will need to check whether the data (both variables) are normally distributed

Looking at the RAE data, funding is not normally distributed:-



So non-parametric testing is appropriate.

**Correlations**

			Full time doctoral students	Total funds	RAE rating
Spearman's rho	Full time doctoral students	Correlation Coefficient	1.000	.271	.298
		Sig. (2-tailed)	.	.087	.058
		N	41	41	41
	Total funds	Correlation Coefficient	.271	1.000	.640(**)
		Sig. (2-tailed)	.087	.	.000
		N	41	41	41
	RAE rating	Correlation Coefficient	.298	.640(**)	1.000
		Sig. (2-tailed)	.058	.000	.
		N	41	41	42

\*\* Correlation is significant at the 0.01 level (2-tailed).

Spearman shows RAE ratings are highly significantly correlated with income, but not doctoral students (though it is approaching significance). The correlation with funding is positive (unsurprisingly). If you took the view that close to 0.05 is worth interpreting then doctoral students are also positively correlated, but the effect is much smaller. A correlation of less than 0.3 is usually said to be medium, and one of 0.5 large. Thus there is a large significant correlation with funding and medium (but only approaching significance) one for doctoral students. Of course you would want to be careful using this information in practice. You could get rid of all doctoral students and concentrate on funding, but the quality of the research institute is in part measured by its doctoral students, and thus funding could then go down.

### Chapter 10

Test for the difference between trial2 and trial3. Decide the appropriate test, run it and interpret the results.

The histogram for trial3 looked decidedly not normally distributed, and in any event with such a small sample non-parametric testing is preferred.

“I studied English for 16 years but...  
...I finally learned to speak it in just six lessons”  
Jane, Chinese architect

ENGLISH OUT THERE

Click to hear me talking before and after my unique course download



So I use Wilcoxon for the two variable comparison and Friedman for the four variable comparison.

### Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Trial 2 and Trial 3 equals 0.	Related-Samples Wilcoxon Signed Ranks Test	.002	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

So there is a significant difference and trial3.

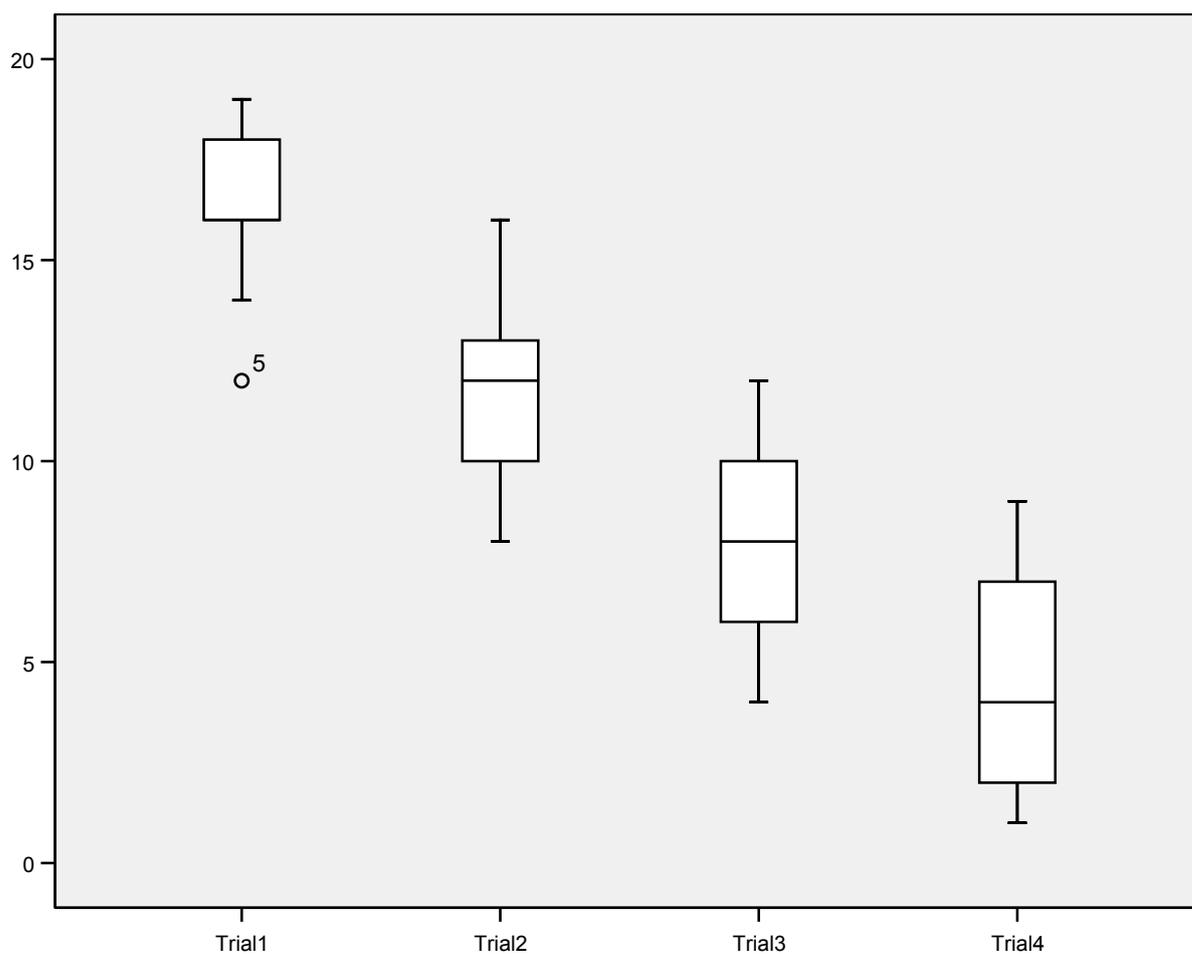
Now test for the differences between all four trials, trial1 trial2 trial3 and trial4. Again decide the appropriate test, run it and interpret the results

### Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Trial 2, Trial 3, Trial 1 and Trial 4 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

So there is a significant difference among the four trials with the mean rank going down from the first to last trial. This is shown nicely in a boxplot.



### Chapter 11

Calculate the sensitivity, specificity, PPV and NPV for asset score against custodial outcome where medium-high risk is considered the threshold. I have started the table for you:-

	Non custodial sentence	Custodial sentence	Total
Less than medium-high risk	666 (TN)	33 (FN)	699
At least medium-high risk	292 (FP)	104 (TP)	396
Total	958	137	1095

How has changing the threshold changed the four measures?

Sensitivity =  $104/137$  = 0.76

Specificity =  $666/958$  = 0.70

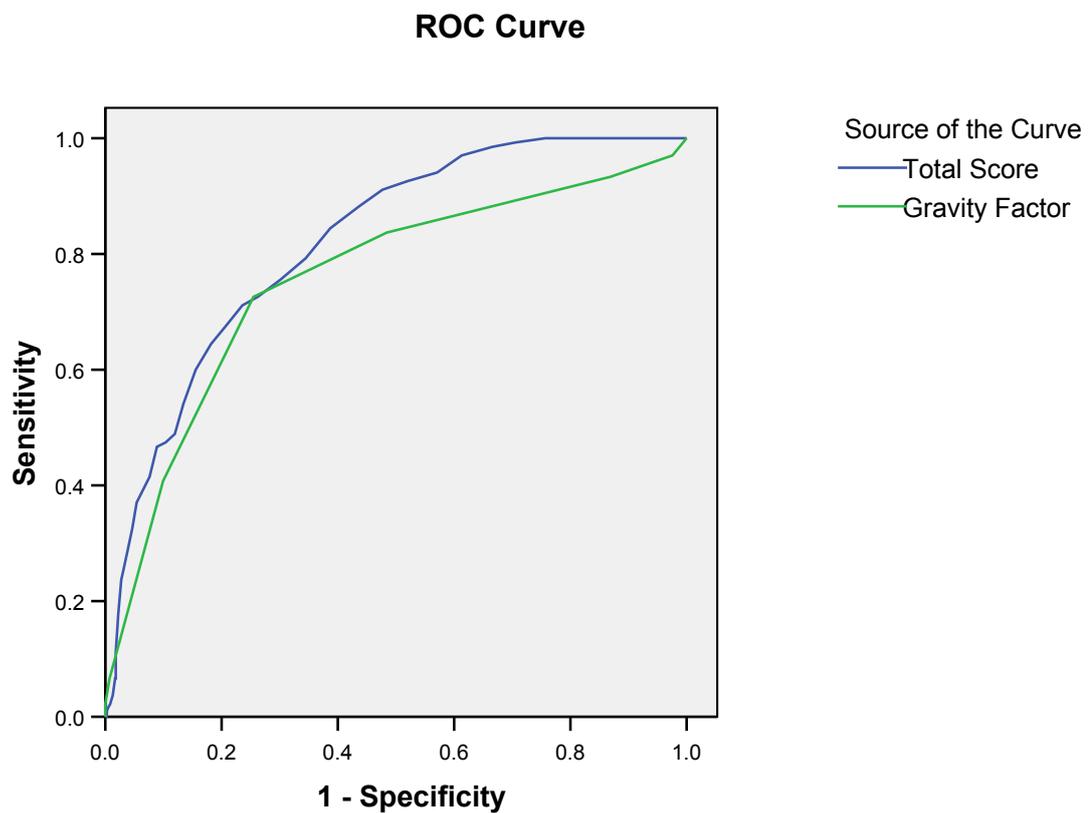
PPV =  $104/396$  = 0.26

NPV =  $666/699$  = 0.95

So sensitivity has increased, specificity decreased (as expected) and PPV has reduced and NPV increased.

### Chapter 12

Using the “asset” datafile, conduct an ROC for gravity and total asset score. Interpret the result.



Diagonal segments are produced by ties.

**Area Under the Curve**

Test Result Variable(s)	Area
Total Score	.819
Gravity Factor	.756

The test result variable(s): Total Score, Gravity Factor has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

The ROC plot shows total asset score is a better classifier of third order outcome, and the areas under the curve is seen to be also greater, confirming this

**Chapter 13**

Use the “pu” dataset. Do a kappa and ICC on these variables to compare nurse 1 with nurse 3.

**pu1 \* pu3 Crosstabulation**

Count

		pu3		Total
		0	1	
pu1	0	33	4	37
	1	2	11	13
Total		35	15	50

In the past four years we have drilled  
**89,000 km**  
That's more than **twice** around the world.

**Who are we?**  
We are the world's largest oilfield services company<sup>1</sup>. Working globally—often in remote and challenging locations—we invent, design, engineer, and apply technology to help our customers find and produce oil and gas safely.

**Who are we looking for?**  
Every year, we need thousands of graduates to begin dynamic careers in the following domains:

- Engineering, Research and Operations
- Geoscience and Petrotechnical
- Commercial and Business

**What will you be?**

**Schlumberger**

careers.slb.com

<sup>1</sup>Based on Fortune 500 ranking 2011. Copyright © 2015 Schlumberger. All rights reserved.



**Symmetric Measures**

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Measure of Agreement	Kappa	.703	.112	4.995	.000
N of Valid Cases		50			

- a Not assuming the null hypothesis.
- b Using the asymptotic standard error assuming the null hypothesis.

So the nurses agree reasonably well and significantly on PU presence.

**Intraclass Correlation Coefficient**

	Intraclass Correlation(a)	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.942(b)	.900	.967	33.333	49.0	49	.000
Average Measures	.970(c)	.947	.983	33.333	49.0	49	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

- a Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
- b The estimator is the same, whether the interaction effect is present or not.
- c This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

So there is high (and significant) agreement with respect to Waterlow scores.

NB. Kappa would have been a poor test for Waterlow scores as the variables have many possible values which are ordinal, indeed SPSS fails to even compute on these data as:-

**Symmetric Measures**

		Value
Measure of Agreement	Kappa	.(a)
N of Valid Cases		50

- a Kappa statistics cannot be computed.They require a symmetric 2-way table in which the values of the first variable match the values of the second variable.

**Chapter 14**

**Why did I use Spearman and not Pearson's correlation?**

Because of non normal distributions of data.

Load the datafile "assets". Use the twelve asset subscores, gravity factor and age to compute a Cronbach  $\alpha$ . Repeat this analysis but with age removed, and then again with gravity and age removed. What do you infer from these analyses.

Use the twelve asset sub-scores, age and gravity factor to compute a Cronbach  $\alpha$ . Gives

**Reliability Statistics**

Cronbach's Alpha	N of Items
.843	14

Repeat this analysis but with age removed, gives

**Reliability Statistics**

Cronbach's Alpha	N of Items
.878	13

and then age and with gravity removed, gives

**Reliability Statistics**

Cronbach's Alpha	N of Items
.888	12

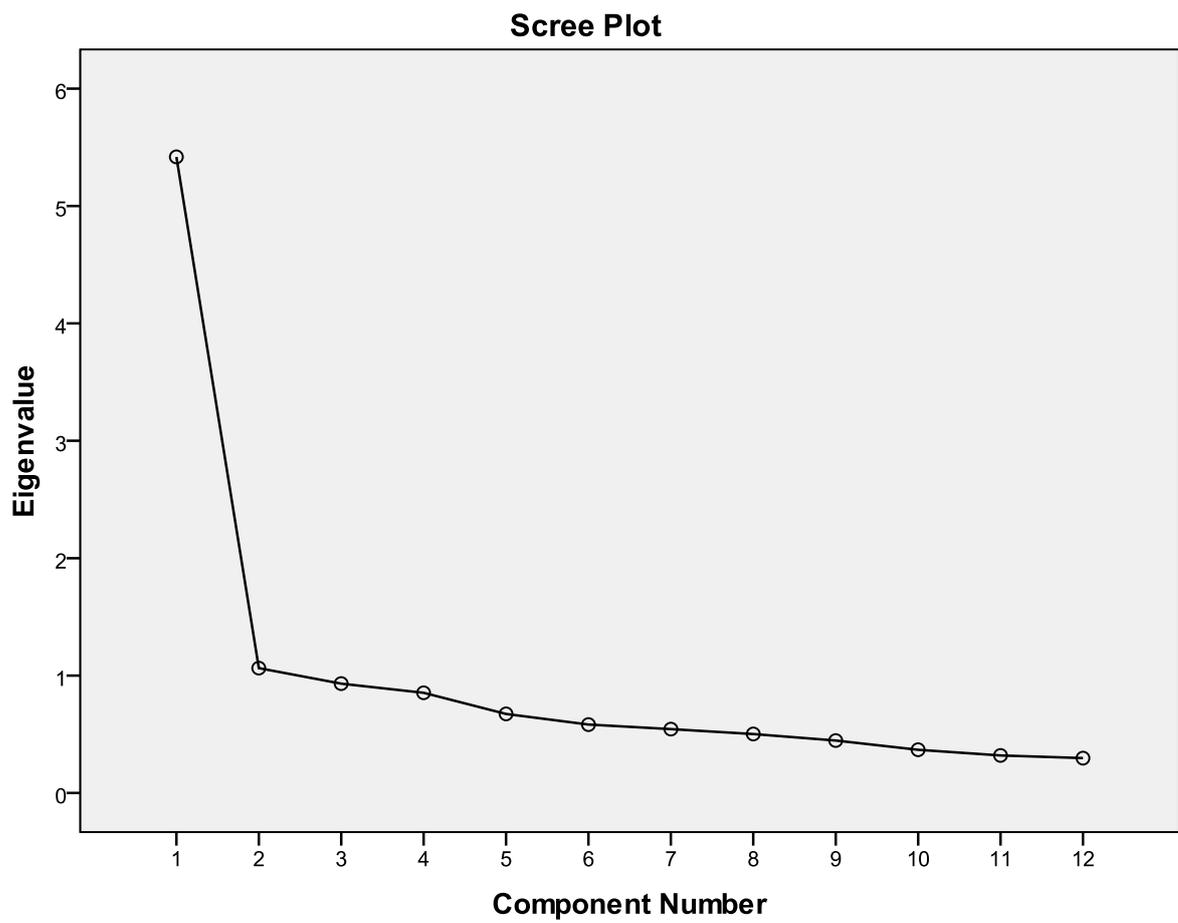
What do you infer from these analyses.

The Cronbach is acceptable with all sub-scores and age and gravity. But removing age and then gravity removed the Cronbach improves, showing (unsurprisingly) that the age and gravity are probably more different from the sub-scores of the asset than the sub-scores are to each other. Nevertheless they do seem to be measuring much the same thing as sub-scores, which is also unsurprising.

## Chapter 15

Consider the datafile “*assets*”. Conduct a factor analysis of the twelve asset sub-scores using PCA. Decide how many factors there are. Interpret the factors.

Putting all variables in except total asset score (since we have all the sub-scores) gives a scree plot, that I would interpret as having possibly only one factor:-



although using Kaiser's criterion it gives two.

**Total Variance Explained**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
Total						
1	5.418	45.153	45.153	5.418	45.153	45.153
2	1.063	8.862	54.016	1.063	8.862	54.016
3	.931	7.762	61.777			
4	.853	7.111	68.889			
5	.674	5.613	74.501			
6	.582	4.854	79.355			
7	.544	4.532	83.887			
8	.502	4.186	88.073			
9	.447	3.721	91.794			
10	.368	3.068	94.863			
11	.320	2.664	97.527			
12	.297	2.473	100.000			

Extraction Method: Principal Component Analysis.

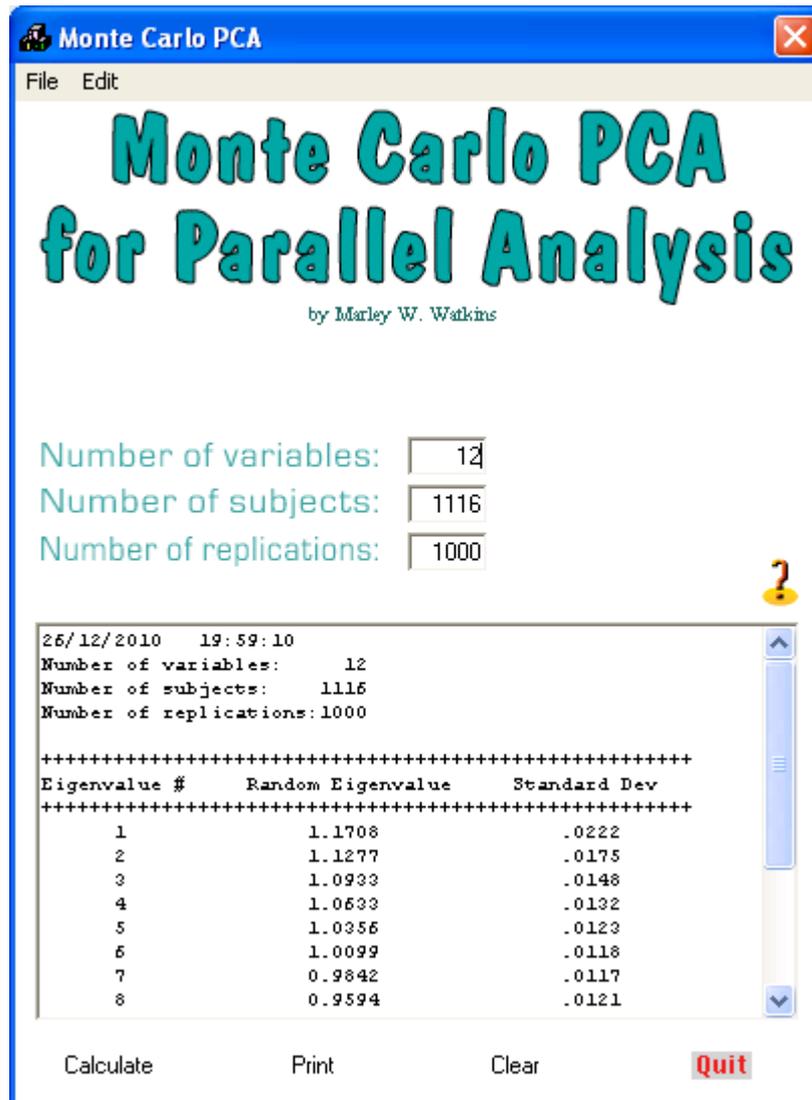
**Component Matrix<sup>a</sup>**

	Component	
	1	2
Living Arrangements	.725	.261
Family & personal relationships	.747	.233
ete	.698	-.086
Neighbourhood	.595	.099
Lifestyle	.778	-.048
Substance Use	.582	.408
Physical Health	.460	.456
Emotional & Mental Health	.565	.286
Perception of Self & others	.708	-.203
Thinking & Behaviour	.681	-.328
Attitudes to Offending	.710	-.457
Motivation to Change	.740	-.327

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

The first component seems to be the twelve sub-scores. The second one loads on some variables but is difficult to interpret. Parallel analysis indicates only one factor as the first (random) eigenvalue is higher than the second one above.



## Chapter 16

Using gravity as your outcome to be predicted, and asset scores as variables, compute a regression analysis, then perform it and interpret it.

**Coefficients<sup>a</sup>**

Model	B	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		Std. Error	Beta			
1	(Constant)	3.283	.086		38.034	.000
	Thinking & Behaviour	.275	.037	.218	7.357	.000
2	(Constant)	3.231	.087		37.113	.000
	Thinking & Behaviour	.214	.041	.170	5.260	.000
	Family & personal relationships	.129	.035	.118	3.641	.000
3	(Constant)	3.255	.087		37.300	.000
	Thinking & Behaviour	.168	.044	.134	3.826	.000
	Family & personal relationships	.100	.037	.091	2.706	.007
	Motivation to Change	.111	.041	.095	2.690	.007
4	(Constant)	3.230	.088		36.698	.000
	Thinking & Behaviour	.164	.044	.130	3.722	.000
	Family & personal relationships	.079	.038	.072	2.057	.040
	Motivation to Change	.095	.042	.081	2.252	.025
	Substance Use	.075	.037	.068	2.054	.040

a. Dependent Variable: Gravity Factor

So four of the twelve sub-scores seem to predict gravity.

## Chapter 17

Run a logistic regression using all the above variables and in addition gravity and gender. Interpret the output.

Looking at the last step we get

**Variables in the Equation**

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 6 <sup>a</sup>	ass1	.354	.098	13.015	1	.000	1.425
	ass6	.480	.095	25.382	1	.000	1.616
	ass8	-.296	.111	7.125	1	.008	.744
	ass11	.540	.105	26.232	1	.000	1.716
	gender	-1.538	.433	12.625	1	.000	.215
	gravit	.633	.091	48.413	1	.000	1.884
	Constant	-6.277	.492	163.015	1	.000	.002

a. Variable(s) entered on step 6: ass8.

Indicating gravity and gender over and above the sub-scores (some of them) affect third order outcomes, with higher gravity (positive B) and more males (B is negative and gender is 0 for males and 1 for females).

## Chapter 19

You want to conduct a Student's t test for independent groups. You are testing whether a drug reduces blood pressure compared to a control. A clinically significant difference has been stated to be a reduction of 5 mm Hg in the diastolic pressure. The standard deviation in previous groups, where you have access to the data, is 20. Look at the graph in **Figure 134** that I created from PS. What power (roughly) would you have with a sample of 100 subjects? What sample size would you need (roughly) to detect the clinically significant difference with a power of 0.8? Now look at **Figure 135**. How big a sample (roughly) would you need to detect a fall of 10 mm Hg, and how many (roughly) to detect a 2.5 mm Hg.

What power (roughly) would you have with a sample of 100 subjects?

**Answer 0.4**

What sample size would you need (roughly) to detect the clinically significant difference with a power of 0.8?

**Answer 250**

How big a sample (roughly) would you need to detect a fall of 10 mm Hg

**Answer 50**

and how many (roughly) to detect a 2.5 mm Hg.

**Answer 1000**

What power (roughly) would you get with a sample of 100?

**Answer power about  $0.67 * 0.955$  or about 0.64**

\*

If you wanted to use Mann Whitney for a medium effect and standard  $\alpha$  and power, using **Figure 136** what sample size (roughly) would you need. Using **Figure 136** what power (roughly) would you get with a sample of 100?

## Chapter 20

Match each of the studies in **Table 70** to tests in **Table 71**

Study design	
e) Are days in hospital (length of stay) correlated to age?	Spearman rank correlation co-efficient
f) Are four ethnic groups <sup>3</sup> different with respect to the percentage mark on a course?	One way ANOVA
g) Are four ethnic groups <sup>3</sup> different with respect to their attitude to online learning on a four point scale from 1 (like it a lot) to 4 (dislike it a lot)?	Kruskall Wallis
h) Are four ethnic groups <sup>1</sup> different with respect to whether they pass or fail a course?	Chi square
i) Are males and females different with respect to the percentage mark on a course?	Student's t test for independent groups
j) Are males and females different with respect to their attitude to online learning on a four point scale from 1 (like it a lot) to 4 (dislike it a lot)?	Mann Whitney
k) Is attitude to online learning on a four point scale from 1 (like it a lot) to 4 (dislike it a lot) different before and after delivery of an online module?	Wilcoxon
l) Is blood pressure correlated to weight?	Pearson's correlation co-efficient
m) Is blood pressure reduced in a group of patients after relaxation exercises?	Repeated ANOVA
n) Is commitment on a four point scale (very committed, committed, not very committed, not committed at all) changed in a series of six monthly weight checks in a weight reduction programme.	Friedman
o) Is weight reduced in a series of six monthly weight checks in a weight reduction programme.	Student's t test for paired data

<sup>1</sup> UK white, UK black, African Black, UK Asian

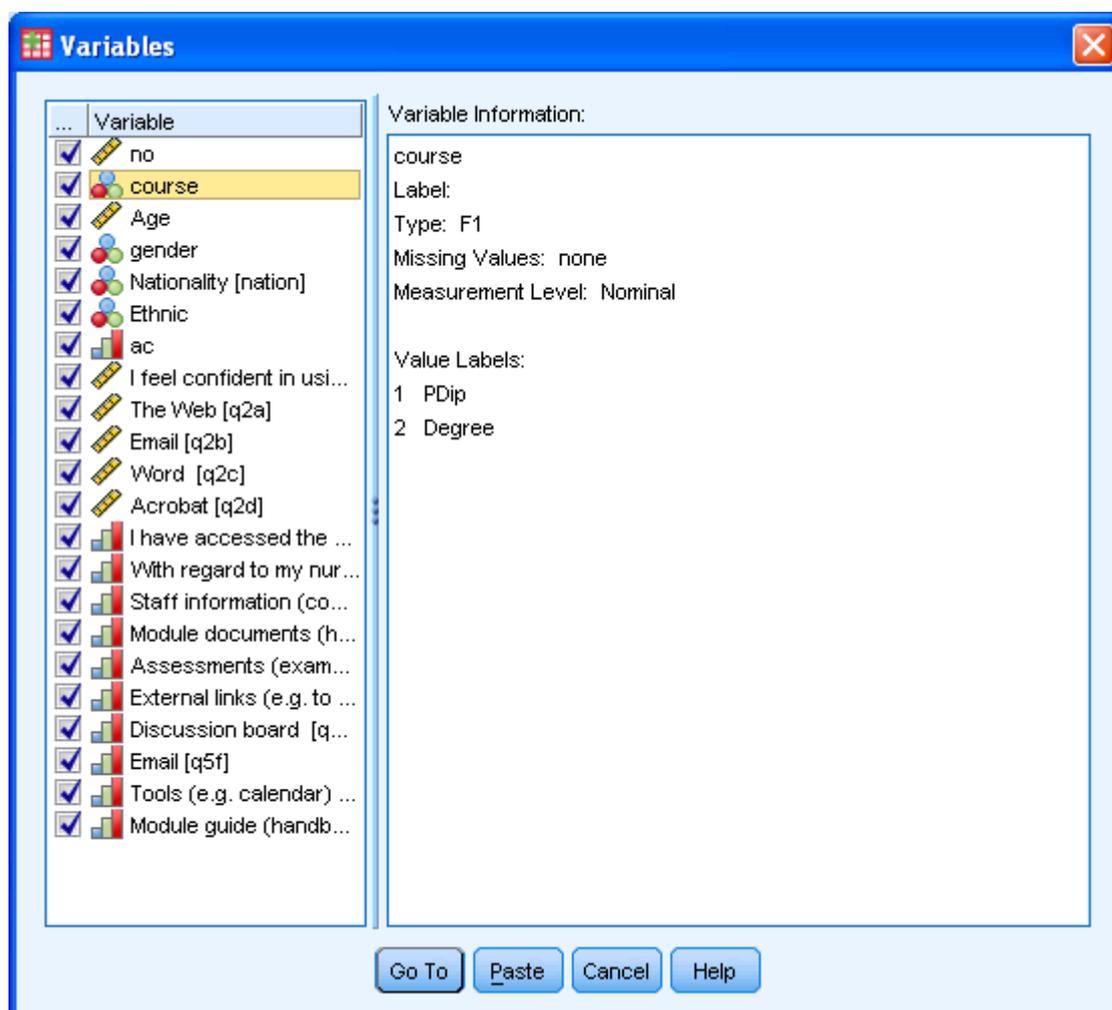
# Appendix Datasets used in this text

The text uses several datasets, and some datasets are used in more than one chapter. Here I describe the variables in each dataset.

## Online course survey

Used in chapter 3, 4, 7, 9, 14

This dataset was from an evaluation of an online course. There are demographic variables (age, gender etc.) and several Likert scores measuring attitudes to various aspects of the course. You will note some of these are Ordinal level (which they should be) and some are Scale. I only assigned variables to Scale due to SPSS (wrongly in my view) assuming some tests need Scale variables.

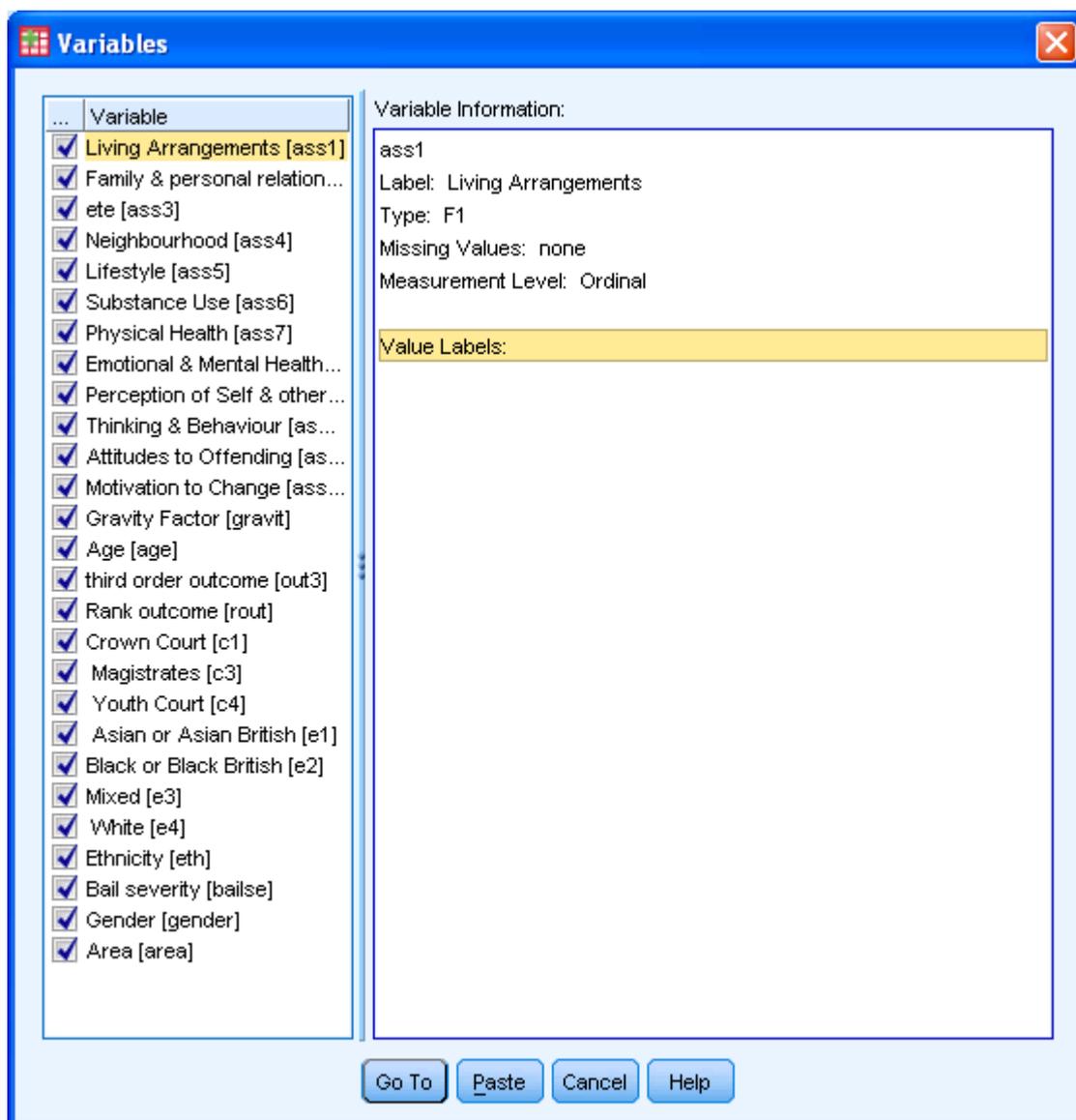


I did not publish from these data, though I did write a report (Anthony, 2006).

## Asset data

Used in chapter 5, 11, 12, 15, 16, 17

This is a database of risk factors in young offenders which has some additional demographic data (age, gender, ethnic group etc) and outcomes (court tried in, sentence etc.).



The asset score is composed of twelve sub-scores (here ass1 to ass12) which are added together to form the total score. These are shown below

1. Living arrangements risk
2. Family and personal relationships risk
3. Statutory education or ETE risk
4. Neighbourhood risk
5. Lifestyle risk
6. Substance use risk
7. Physical health risk
8. Emotional and mental health risk
9. Perception of self and others risk
10. Thinking and behaviour risk
11. Attitudes to offending risk
12. Motivation to change problem

*Gravity* is a measure of how serious the offence was, with high numbers meaning very serious offences (typically violent) and low numbers more trivial.

The advertisement features a circular collage of news headlines on a black background. The headlines include: 'RUSSIAN ENTREPRENEUR MARKET LONDON', 'LUXURY GOODS MARKET LONDON', 'THOUGHTED LONDON B...', 'CHINA PAYS PREMIUM FIRM', 'AFRICA HEALS US FIRM', 'US FIRM PLUTSOU', 'INDIAN NUCLE', and 'POWER CORP'. To the right of the collage, the text 'WHILE YOU WERE SLEEPING...' is written in large white capital letters. At the bottom left, the URL 'www.fuqua.duke.edu/whileyouweresleeping' is displayed. At the bottom right is the Duke University logo, which reads 'DUKE THE FUQUA SCHOOL OF BUSINESS'.



There are over thirty outcomes in terms of sentencing, and these were grouped into three ranked outcomes, of which the most serious, 3<sup>rd</sup> order, were custodial sentences. While not necessary for understanding this dataset, the outcomes are:-

first order outcome, any of

1. Absolute Discharge
2. Bound Over
3. Compensation Order
4. Conditional Discharge
5. Costs
6. Compensation Order
7. Conditional Discharge
8. Costs
9. Referral Order
10. Referral Order Extension
11. Reparation Order
12. Sentence Deferred

Second order outcome, any of

1. Action Plan Order
2. Anti Social Behaviour Order (Crime)
3. Attendance Centre Order
4. Community Punishment and Rehabilitation Order
5. Community Punishment Order
6. Community Rehabilitation Order
7. Community Rehabilitation Order + Conditions
8. Community Rehabilitation Order + ISSP
9. Curfew Order
10. Order Extended
11. Order To Continue
12. Supervision Order
13. Supervision Order (Conditions)

Third order outcome, any of

1. Detention and Training Order
2. Licence Recall
3. Section 90–92
4. Young Offenders Institute

A separate binary variable *out3* splits data into either the most serious 3<sup>rd</sup> order outcome or either of the less serious ones.

Ethnicity is coded in *eth*

1. Asian
2. Black
3. Mixed
4. White

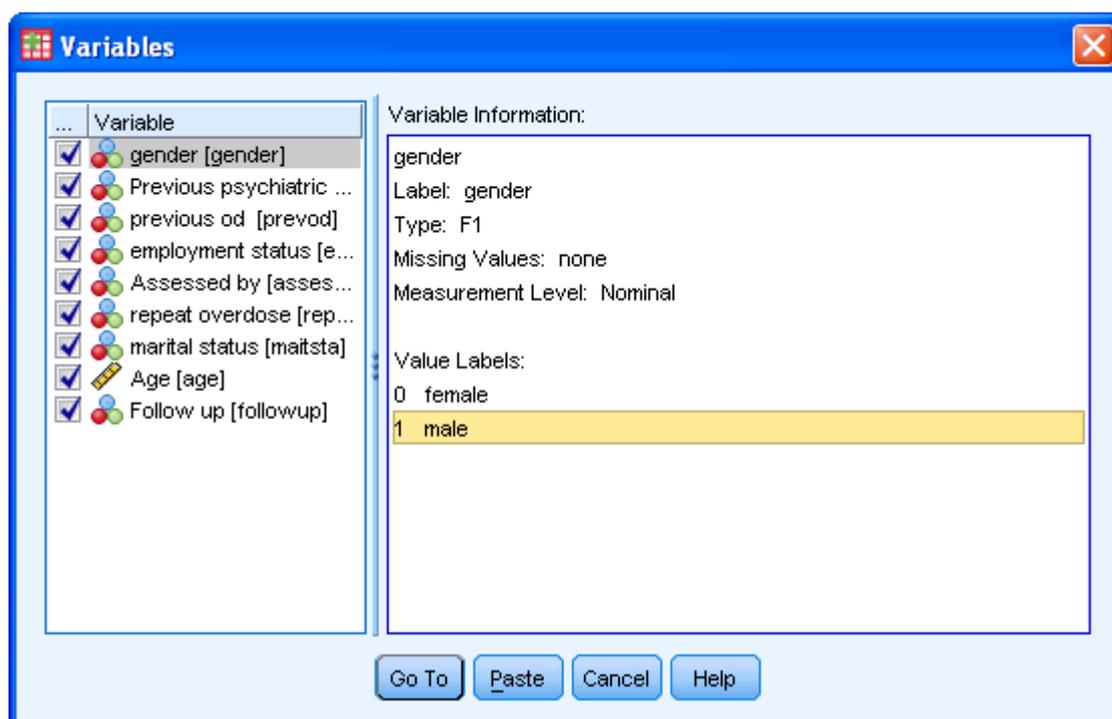
*Gender* is self explanatory

I did not publish from this paper, but did write a report (Williams et al., 2006).

## Deliberate self harm

Used in chapter 6

There are nine variables; all bar Age (Scale) are nominal



This dataset was collected by a psychiatric nurse in her masters degree. She collected demographic data (gender, employment, age, marital status) and clinical data (previous overdose, previous psychiatric history, whether overdose repeated) and who assessed the patient (nurse or doctor). Jenny was trying to identify variables that are associated with repeating overdose.

We did publish from this (Cook and Anthony, 1999).

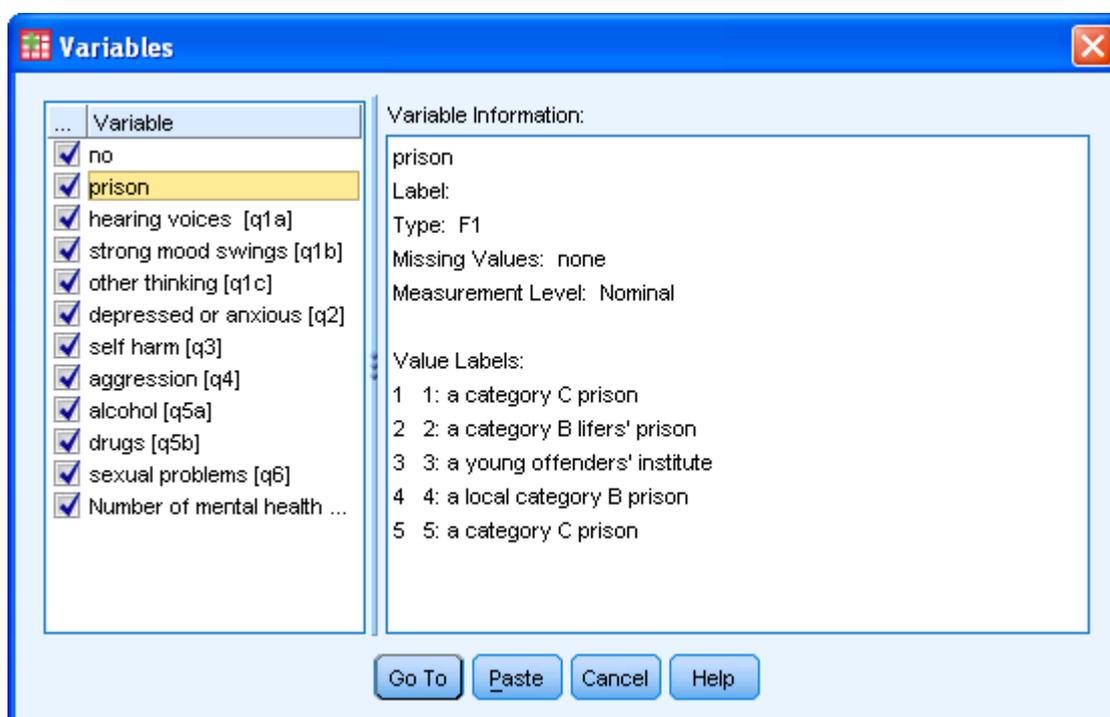
Download free eBooks at [bookboon.com](http://bookboon.com)

## Mental health in prison

Used in chapter 8, 15, we wrote a report (Anthony and Collins, 2004) and published from this (Anthony, 2005a).

This has for a number of prisons the responses of prisoners with respect to mental health symptoms. Prisoners were asked about nine symptoms, and a count was made for each prisoner of all the symptoms they had, up to a maximum of nine. There are five prisons. Note category C are prisons for low risk prisoners (but are not open prisons), category B are higher security prisons, the highest security prison is category A. This sample contains only category B and C prisons, and a young offenders institute.

This dataset consists mainly of yes/no answers to specific mental health problems, but also has a variable *prison* identifying the prison the prisoner is in, and a computed variable “Number of mental health symptoms” which adds up the number of symptoms.



The survey instrument is below:-

**Q1 Do you have any problems with any of the following now or in the last 6 months (please tick box with a ✓)?**

	No	Yes
hearing voices (when there is no-one there)	<input type="checkbox"/>	<input type="checkbox"/>
strong mood swings	<input type="checkbox"/>	<input type="checkbox"/>
other problems with thinking (Please state below)	<input type="checkbox"/>	<input type="checkbox"/>

*Examples: difficulty in concentrating, seeing things that are not there, feeling that the television or radio is talking about you, paranoia*

**Q2 Do you feel depressed, very sad, or very anxious now or in the last 6 months (please tick box with a ✓)**

*Examples: feeling that live is not worth living, worrying greatly about the future*

No	Yes
<input type="checkbox"/>	<input type="checkbox"/>

**Q3 Do you sometimes think you want to harm yourself, or have you ever actually harmed yourself (please tick box with a ✓)?**

No	Yes
<input type="checkbox"/>	<input type="checkbox"/>

**Q4 Do you ever lose your temper and hit someone, or otherwise physically harm or abuse other people (please tick box with a ✓)?**

*Example: you get frustrated or annoyed easily and sometimes hit people when this happens*

No	Yes
<input type="checkbox"/>	<input type="checkbox"/>

**Q5 Did you drink too much alcohol or have a drug problem before you came into prison (please tick box with a ✓)?**

*Examples: you drink until you are drunk most days, you are addicted to heroin or crack cocaine*

	No	Yes
I drank too much alcohol	<input type="checkbox"/>	<input type="checkbox"/>
I had a drugs problem	<input type="checkbox"/>	<input type="checkbox"/>

**Q6 Did you have any mental health problems that caused problems with your sex life before you came to prison (please tick box with a ✓)?**

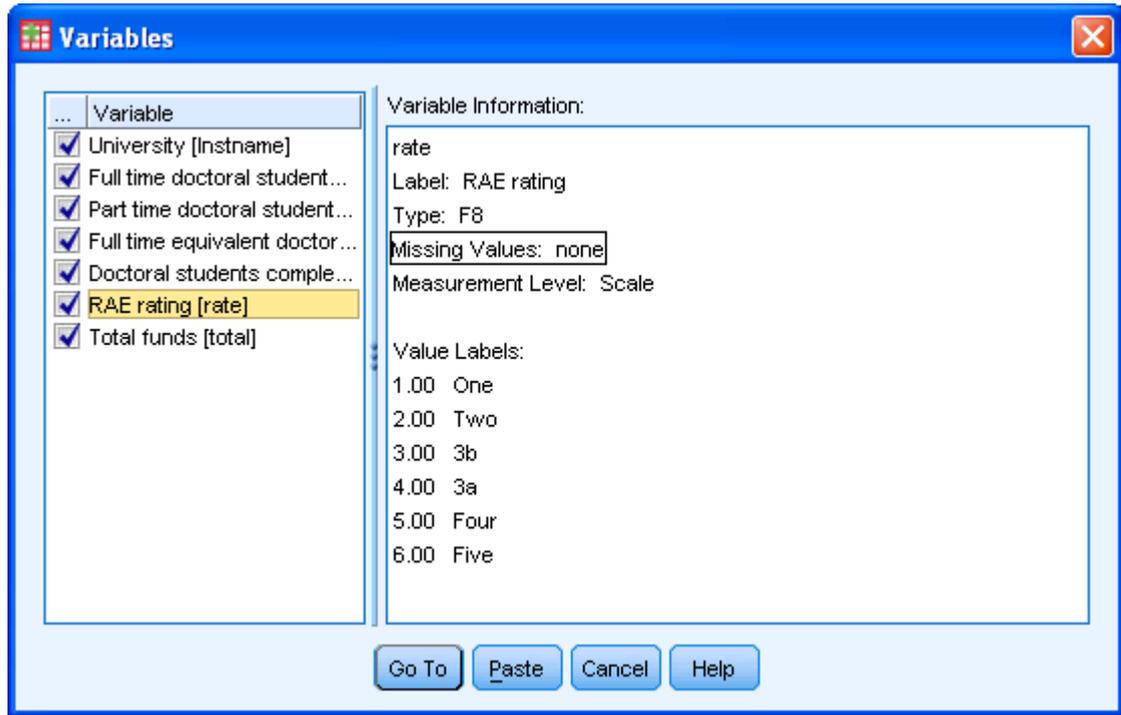
*Examples: you have side effects from your medication that cause problems, you are too depressed or anxious to make love.*

No	Yes
<input type="checkbox"/>	<input type="checkbox"/>

## RAE example

Used in chapter 9

Universities in the UK are evaluated at regular intervals for their research outputs. The last one was 2008. The RAE gave scores 1, 2 3b, 3a, 4, 5 and 5\* where 1 was least good and 5\* best. Each subject was entered and evaluated separately, these were called units of assessment (UoAs). I looked at the RAE score for one submission, nursing. I have created the datafile “research assessment exercise” which I used in a paper (Anthony, 2005b).



There are variables on the full and part time research student numbers, the funds obtained by research applications and the rating given to the university.

Excellent Economics and Business programmes at:



**university of  
 groningen**

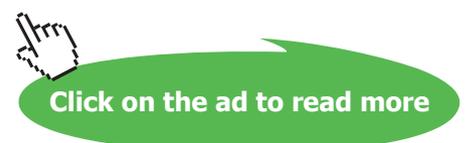




**“The perfect start  
 of a successful,  
 international career.”**

[www.rug.nl/feb/education](http://www.rug.nl/feb/education)

**CLICK HERE**  
 to discover why both socially  
 and academically the University  
 of Groningen is one of the best  
 places for a student to be



# Anxiety

Used in Chapter 10.

In the SPSS directory there are several test data files, one is called *anxiety 2* (in SPSS v18 this is in subdirectory “Samples/English”. Load this file, or if it is not available enter the data in Table 72 (if you loaded the file from SPSS there are two additional variables, anxiety and tension, but we are not using these).

**Table 72:** Anxiety 2 data

Subject	Trial1	Trial2	Trial3	Trial4
1	18	14	12	6
2	19	12	8	4
3	14	10	6	2
4	16	12	10	4
5	12	8	6	2
6	18	10	5	1
7	16	10	8	4
8	18	8	4	1
9	16	12	6	2
10	19	16	10	8
11	16	14	10	9
12	16	12	8	8

This data set is taken from twelve students. The students are each given four trials on a learning task, and the number of errors for each trial is recorded. The errors for each trial are recorded in separate variables trial1 to trial4.

# Pressure ulcer dataset

Used in chapter 13.

In dataset “*pu*”

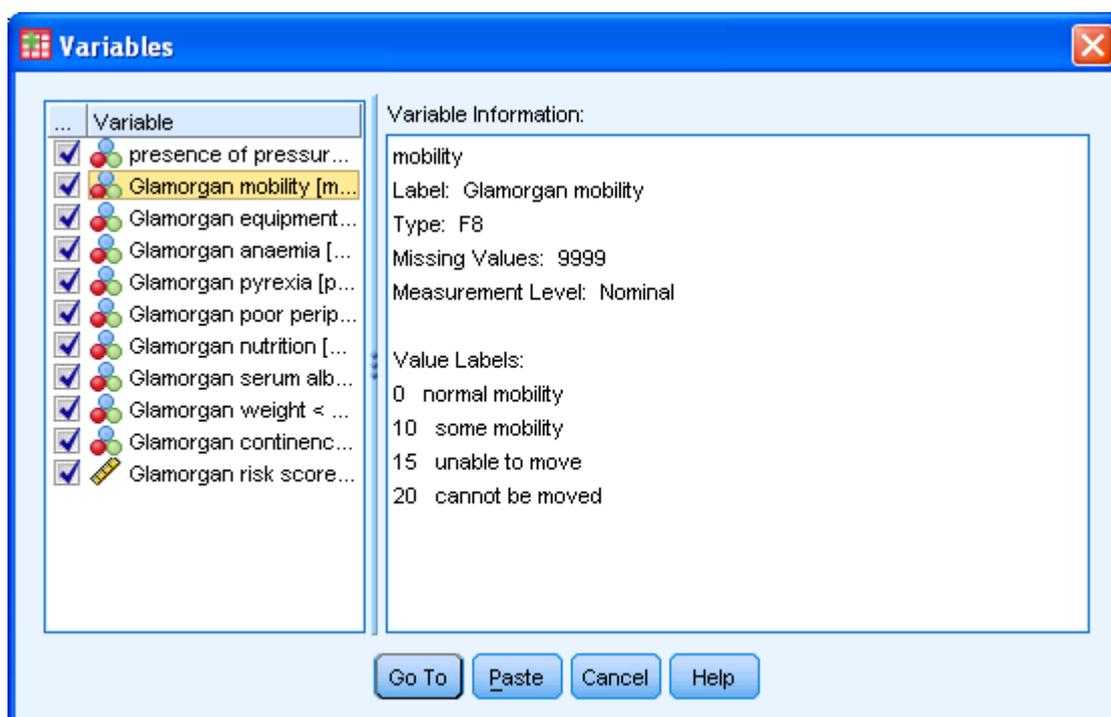
In this fictional dataset you will find three variables (*pu1*, *pu2* and *pu3*) for presence or absence of pressure ulcers recorded by three different nurses, and three others (*wat1*, *wat2*, and *wat3*) for Waterlow scores from the same three nurses.

# Glamorgan risk assessment scale for pressure ulcers

Used in chapter 18.

In dataset “*glam*”. This consists of several risk factors for children of pressure ulcers, see Figure 137. The first variable is whether the child has a pressure ulcer and the last is the total score.

**Figure 137:** Glamorgan score



# Infant mortality

Used in chapter 18.

Three measures of mortality for countries in Europe; infant mortality (number of infant deaths - one year of age or younger - per 1000 live births), perinatal mortality (number of stillbirths and deaths in the first week of life per 1,000 live births) and neonatal mortality (number of deaths during the first 28 completed days of life per 1,000 live births in a given year or period).

.....Alcatel-Lucent 

[www.alcatel-lucent.com/careers](http://www.alcatel-lucent.com/careers)

What if you could build your future and create the future?

One generation's transformation is the next's status quo. In the near future, people may soon think it's strange that devices ever had to be "plugged in." To obtain that status, there needs to be "The Shift".



# References

- ALTMAN, D., G & BLAND, J. M. 1994a. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ*, 308.
- ALTMAN, D., G & BLAND, J. M. 1994b. Statistics Notes: Diagnostic tests 2: predictive values *BMJ*, 309.
- ALTMAN, D., G & BLAND, J. M. 1994c. Statistics Notes: Diagnostic tests 3: receiver operating characteristic plots *BMJ*, 309.
- ANTHONY, D. 2006. Final report to LNR Healthcare Workforce Deanery of project Health Care Assistant Online Course, Leicester, De Montfort University.
- ANTHONY, D. M. 2005a. Mental health needs assessment of prisoners. *Clinical Effectiveness in Nursing*, 9.
- ANTHONY, D. M. 2005b. The Nursing Research Assessment Exercise 2001: An analysis. *Clinical Effectiveness in Nursing*, 9, 4-12.
- ANTHONY, D. M. & COLLINS, G. 2004. Mental Health Needs Assessment: HMP Ashwell, HMP Gartree, HMYOI Glen Parva, HMP Leicester, HMP Stocken, Leicester, Mary Seacole Research Centre.
- BRYMAN, A. & CRAMER, D. 1997. *Quantitative data analysis*, London, Routledge.
- BRYMAN, A. & CRAMER, D. 2005. *Quantitative data analysis with SPSS 12 and 13 : a guide for social scientists*, Hove, Routledge.
- BRYMAN, A. & CRAMER, D. 2009. *Quantitative data analysis with SPSS 14, 15 and 16 : a guide for social scientists / Alan Bryman and Duncan Cramer.*, Hove Routledge.
- BUBL. Available: [bubl.ac.uk/LINK/r/researchmethods.htm](http://bubl.ac.uk/LINK/r/researchmethods.htm) [Accessed 11 Nov 2010].
- CAIRNCROSS, A. 1996. *Economist* 20 April.
- COHEN, J. 1989. *Statistical Power Analysis for the Behavioural Sciences*. 2nd Ed. , Hillsdale NJ, Erlbaum.
- COOK, J. & ANTHONY, D. M. 1999. Repetition of self harm. *Clinical Effectiveness in Nursing.*, 3, 181-4.
- EVERITT, B. 1980. *Cluster analysis*, London, Heinemann.
- EVERITT, B. 1992. *The analysis of contingency tables*, London. , Chapman & Hall. .
- FIELD, A. 2009. *Discovering statistics using SPSS (and sex and drugs and rock 'n' roll)*, London, Sage.

- GALLOWAY, A. 1997. Questionnaire Design & Analysis [Online]. Available: [www.tardis.ed.ac.uk/~kate/qmcweb/qcont.htm](http://www.tardis.ed.ac.uk/~kate/qmcweb/qcont.htm) [Accessed 11 Nov 2010].
- GOLDACRE, B. 2006. Crystal Balls... and Positive Predictive Values. The Guardian, December 9th.
- HAMMOND, S. 1995. Introduction to multivariate data analysis. In: BRAKWELL, G., HARNMOND, S. & FIFE-SCHAW, C. (eds.) Research methods in psychology. London: Sage.
- LIKERT, R. 1932. A Technique for the Measurement of Attitudes. Archives of Psychology, 140, 1-55.
- MARSTON, L. 2010. Introductory statistics for health and nursing using SPSS, London, Sage.
- MILES, J. n.d. Getting the Sample Size Right: A Brief Introduction to Power Analysis [Online]. Available: [www.jeremymiles.co.uk/misc/power/index.html](http://www.jeremymiles.co.uk/misc/power/index.html) [Accessed].
- OVERTON-BROWN, P. & ANTHONY, D. M. 1998. Towards a partnership in care: nurses' and doctors' interpretation of extremity trauma radiology. Journal of Advanced Nursing, 27, 890-6.
- PALLANT, J. 2010. Title SPSS survival manual, Maidenhead Open University Press.
- ROSE, D. 1995. Psychophysical methods. In: BREAKWELL, G. M., HAMMOND, S. & FIFE-SCHAW, C. (eds.) Research methods in psychology. London: Sage.
- STATPAC. 2010. Survey & Questionnaire Design [Online]. Available: [www.statpac.com/surveys/](http://www.statpac.com/surveys/) [Accessed 11 Nov 2010].
- SWETS, J. A. 1979. ROC analysis applied to the evaluation of medical imaging techniques. Investigative Radiology, 14, 109-121.
- THE UNIVERSITY OF AUCKLAND. Developing questionnaires websites [Online]. Available: <http://www.fmhs.auckland.ac.nz/soph/centres/hrmas/resources/questionnaire.aspx> [Accessed 2010 11 Nov ].
- TYRELL, S. 2009. SPSS: Stats practically short and simple, Copenhagen, Ventus.
- WHO 196. Perinatal mortality. , Geneva, World Health Organization.
- WILLIAMS, B., YATES, J. & ANTHONY, D. M. 2006. The use of remands and custodial sentences for juveniles in Leicester and Leicestershire and Rutland.
- ZWEIG, M. H. & CAMPBELL, G. 1993. Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clinical Chemistry 39, 561-577.



**Maastricht University** *Leading in Learning!*

**Join the best at the Maastricht University School of Business and Economics!**

**Top master's programmes**

- 33<sup>rd</sup> place Financial Times worldwide ranking: MSc International Business
- 1<sup>st</sup> place: MSc International Business
- 1<sup>st</sup> place: MSc Financial Economics
- 2<sup>nd</sup> place: MSc Management of Learning
- 2<sup>nd</sup> place: MSc Economics
- 2<sup>nd</sup> place: MSc Econometrics and Operations Research
- 2<sup>nd</sup> place: MSc Global Supply Chain Management and Change

Sources: Keuzegids Master ranking 2013; Elsevier 'Beste Studies' ranking 2012; Financial Times Global Masters in Management ranking 2012

**Maastricht University is the best specialist university in the Netherlands (Elsevier)**

**Visit us and find out why we are the best!**  
**Master's Open Day: 22 February 2014**

[www.mastersopenday.nl](http://www.mastersopenday.nl)

