LBSS Social Science Statistics I

Student Workbook

Lecture Structure, Lab Exercises, Tutorial Material and Reading

Gwilym Pryce and Julie Clark

(22 September 2009)

SSS1Workbook_2009_2010_v1n

Contents:

Introduction

This document outlines the basic structure of each lecture, provides exercises for lab/home study, and gives details of recommended reading:

Lectures

A very brief outline of the structure of each lecture is presented. More detailed information on the content of each lecture can be found in the powerpoint slides posted on the *Statistics & SPSS* page of the www.gpryce.com website.

Labs

The exercises for the first lab are presented with detailed worked answers. See Pryce (2005)[1] and the syntax files posted on the *Statistics & SPSS* page of www.gpryce.com for answers and/or hints to most of the questions in the remaining labs.

Also, keep an eye out for the *Confidence Intervals* mindmap and the *Hypothesis Testing* mindmap, both of which will be posted up soon (these files will help you identify the right technique for answering questions on confidence intervals and hypothesis tests).

If you are keen to do exercises beyond those provided here, or you would rather not use the Pryce (2005) book, try working through those in the relevant sections of Moore and McCabe (see the notes on reading below). Note that Moore and McCabe provide selected answers to exercises at the back of the book (though these are not worked-answers – just the final result).

NB: it is assumed in this document that you are using a lab PC (as opposed to a your own PC). It is also assumed that the lab administrator has placed the data files in a folder called Q:\QUANTS. If the files have been placed in an alternative folder, you will need to change the syntax below accordingly. In some exercises (such as the last three exercises of Lab 1), you need to be fully logged on (i.e. not using a temporary password). If this is not possible, for these exercises you will need to work with a fellow student who is able to log on, or consult a lab tutor.

Reading

Essential reading is marked with an asterix. All other reading is recommended but not essential. Most of the essential and recommended readings are taken from the following two books:

> Pryce, G. (2005) *Inference and Statistics in SPSS*, Glasgow: Geebeejey Publishing.
>
> Moore, D.S. and McCabe, G.P. (2003) *Introduction to the Practice of Statistics*, 4th Ed., San Francisco: Freeman.

Copies of the 4th edition are available in the library. There is a new edition but while I have not provided detailed reading for the 5th edition, you will probably find that it follows a similar structure to the 4th edition.

[1] Pryce, G. (2005) *Inference and Statistics in SPSS*, Glasgow: Geebeejey Publishing.

L1 Density Functions & CLT

1.1 Structure of Lecture 1

1. **Review of Induction material**
2. **Density Functions**
3. **Normal Distribution**
4. **Central Limit Theorem**

1.2 Lab 1a

1.2.1 Example 1.3.3a How to Create a Histogram in SPSS (Pryce, p.1-24)

Open the **householddata.sav** file (the file is available from the *Statistics & SPSS* page of www.gpryce.com. It may also be available from the Q:\ drive of the lab computers in the Adam Smith Buiding. If you are working from your own PC simply save or copy the file to your hard disk and open in SPSS by clicking File, Open, Data).

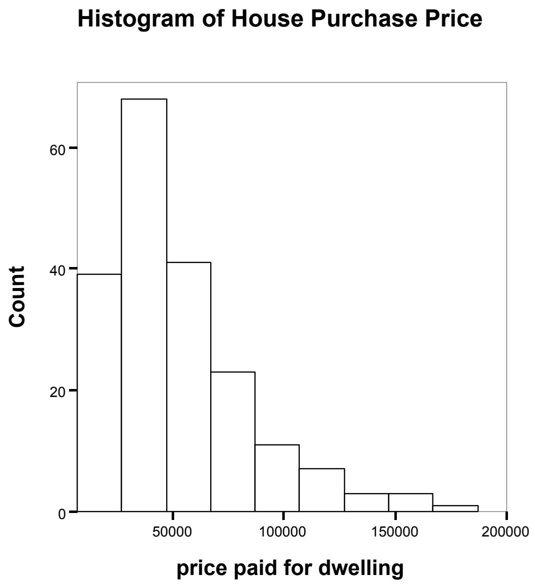
Click Graphs, Interactive, Histogram…, and select the **Dprice** variable (**price paid for the dwelling**), which is the variable for which you want to plot a histogram, and drag it over to box on the right lying on top of the line representing the horizontal axis. Then click the Titles tab and type, 'Histogram of House Purchase Price' as the heading and 'Source: Hypothetical House Price Data' as the caption. Click Paste and run the syntax by highlighting it and clicking ▶:

```
IGRAPH /VIEWNAME='Histogram' /X1 = VAR(Dprice) TYPE = SCALE /Y = $count
 /COORDINATE = VERTICAL  /TITLE='Histogram of House Purchase Price'
 /CAPTION='Source: Hypothetical House Price Data' /X1LENGTH=3.0 /YLENGTH=3.0
 /X2LENGTH=3.0 /CHARTLOOK='NONE' /Histogram  SHAPE = HISTOGRAM CURVE = OFF
  X1INTERVAL AUTO X1START = 0.
EXE.
```

In the Output window, open the Interactive Graph editor by double clicking on the histogram you have just created. In Interactive Graph editor, double click on the horizontal axis to bring up the Scale tab, and untick the Auto box for "Maximum". Change the maximum value for the horizontal axis to 200,000 (without the comma). Click Apply and you should be able to see how the graph has changed (note that you can drag the Scale Axis dialogue box if it is obscuring the graph).

Click OK and then double click on the bars of the graph to open the Histogram dialogue box. Click on the Interval Tool button and click on the down arrow adjacent to the Set options box so that you can select Inteval Size. Enter 20,000 (without the comma) and press Enter on your keyboard. You will see that the bars of the graph have widened and the histogram looks less irregular. Click the **Histogram** button and change the colour of the graph from red to white. Click OK and click outside of the Interactive Graph box (say, in the space to the right) to exit the graph editor.

You should end up with the following graph:



Source: Hypothetical House Price Data

If you want to include this chart in a wordprocessing document, simply right-click on the graph, select Copy Objects, and then paste into your wordprocessing document.

Note that SPSS 15 has more than one way to create a histogram. The simpler GRAPH command produces almost identical and requires rather less convoluted syntax:

```
GRAPH  /HISTOGRAM=Dprice
       /TITLE= 'Histogram of House Purchase Price '
       /FOOTNOTE= 'Source: Hypothetical House Price Data'.
```

You can edit this graph by double clicking on it and working the various menu options of the editor which pops up (unfortunately, the menu system is different to the Interactive Graph editor described above).

1.2.2 Example 1.5.6a Computing the Sample Mean in SPSS (Pryce, p.1-34)

Open the **employees.sav** file. You can do this by clicking File, Open, Data or by using the **GET FILE** command which is more useful if you want to keep a record of which files you have opened (important when you are working with lots of datasets. If the file is located in the Q:\QUANTS directory, then you would type and run the following command in your syntax window:

```
GET FILE='Q:\QUANTS\employees.sav'.
```

The simplest way to compute the mean of a variable is to use the Descriptives function obtained by running the following syntax:

```
DESCRIPTIVES VARIABLES=size
/STATISTICS=MEAN .
```

Alternatively, you can click on Analyze, Descriptive Statistics, Descriptives, select size from the list and paste it into the Variable(s) box by clicking the ▶ button, click on Options to choose which summary statistics you want to include, click Continue, Paste, then highlight the pasted syntax and click ▶.

Both methods will produce the following table which tells us that the average firm size, measured as the number of employees, is 495, based on a sample of 20.

Descriptive Statistics

	N	Mean
Size of firm, measured as the number of employees	20	495.3000
Valid N (listwise)	20	

The question you should immediately ask yourself is the extent to which your sample mean is a good estimate of the population mean. In other words, how close will the average size of ***all*** firms in the UK be to your 495 figure? Even if your sample is truly random (i.e. each firm in the population has an equal chance of being included in your sample), you might have randomly selected an unusual group of firms. This is called **sampling variation**. Later in the course we will explore ways of deriving confidence intervals for the population mean. The confidence intervals will tell us the range of values the population mean is likely to take given a sample mean of 495.

1.2.3 Example 1.5.6b Computing the Sample Standard Deviation in SPSS (Pryce, p.1-35)

Suppose we want to compute the standard deviation of employees who worked in the firms you sampled (i.e. the average deviation of firm size from the mean). Again, the simplest way to compute the standard deviation of a variable is to use the Descriptives function obtained by running the following syntax:

```
GET FILE='Q:\QUANTS\employees.sav'.
DESCRIPTIVES VARIABLES=size  /STATISTICS= STDDEV.
```

If you want to calculate the variance as well, simply add this to the list of statistics you ask SPSS to compute:

```
DESCRIPTIVES VARIABLES=size  /STATISTICS= STDDEV   VARIANCE.
```

This will yield the following table:

Descriptive Statistics

	N	Std. Deviation	Variance
Size of firm, measured as the number of emnployees	20	362.79298	131618.7
Valid N (listwise)	20		

which tells us that the average deviation from the mean firm size is 363 employees. Again your immediate reaction should be to ask how this compares to the standard deviation in the population. Perhaps you have chosen a sample with a lot more variation (or a lot less) in size of firm than in the population of all firms.

Alternatively, you can click on Analyze, Descriptive Statistics, Descriptives, select **size** from the list and paste it into the Variable(s) box by clicking the ▶ button, click on Options to choose which summary statistics you want to include, click Continue, and Paste.

1.2.4 Example 1.5.7 How to obtain a Summary of FIle Info (Pryce, p. 1-37)

Open up the **employees.sav** dataset:

```
GET 'Q:\QUANTS\employees.sav'.
```

Click on File, Display Data File Information, Working File, or simply run the following syntax, DISPLAY DICTIONARY.

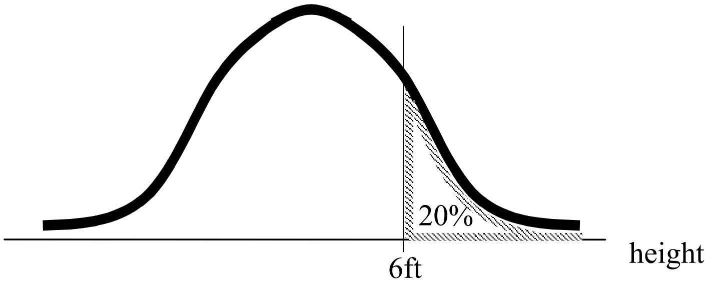
If you have a very long list of variables, you may find that you cannot see all of them listed in the file information that has been pasted to your output file. This is because, in recent versions at least, SPSS pastes the file information as a text box or table. To see all of the text, double click in the text box or table that has been created in the Output window. You will then be able to scroll down the text to view any element.

If the output has been presented as a table (as in SPSS 15), once you have double clicked on the table to open up the Pivot Table edit window, you will be able to change the column widths by dragging the vertical lines. You can also select and copy elements of the table (by right-clicking on the highlighted text and choosing "Copy" from the list) to a word processor. If you want to insert a footnote relating to one of the cells, click on the cell and then choose Insert from the menu bar, then double click on the footnote that has been created to enter the desired text. Press Enter on your keyboard, select File and Close to exit the Pivot Table editor. You can now copy and paste the table to another program – simply right click on the table and choose Copy Objects which will allow you to paste the table as an image. If you would rather paste the table as an item that can be edited in Word, right click on the table in SPSS and select Copy rather than Copy Objects.

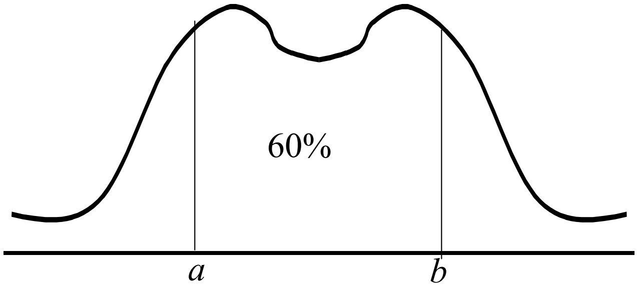
N.B. In older versions of SPSS the menu route for obtaining file information was different. If you have an earlier version of SPSS try clicking Utilities on the menu bar at the top of the screen, and then selecting File Info. Or simply run the DISPLAY DICTIONARY syntax since the syntax has remained unchanged. This raises an important reason for learning and using syntax! Each time a new version of SPSS comes out, there are changes to the menu system, but the syntax commands usually remain unaltered.

1.2.5 Exercise 2.2 Understanding & Calculating Areas under a Density Curve (Pryce, p.2-14)

In the symmetrical bell-shaped distribution below, what is the probability of being less than 6ft tall?



In the following symmetrical distribution, if 60% of the sample falls between *a* and *b*, what % is greater than b? (Assume that a and b are equally spaced from the centre of the distribution). What's the probability of randomly choosing an observation greater than b?



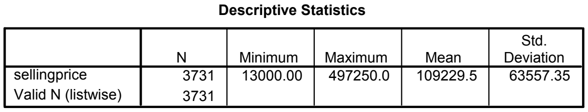
1.2.6 Example 2.6a Sampling Distribution of Means (Pryce p. 2-17)

To illustrate, let's suppose we have access to the selling prices of all properties on the south side of Glasgow area sold in the second half of 2004. Unusually, in this instance, we have information on the population, a total of 3,731 sales. Suppose further that we want to take a random sample of 100 prices from this population. A random sample is one where each observation in the population has an equal chance of entering our sample. How good an approximation of the population average house price will the mean of our random sample of 100 properties provide? And how would the sample mean vary if we were to take another random sample from the population of all house prices? To investigate we shall take repeated random samples of 100 properties, calculate the mean selling price of each sample, and plot a histogram of all the sample means we have calculated.

Open the **Glasgow houseprices_pop_2004q3q4.sav** (the dataset is available from the Q:\ drive of the lab computers; if you are working at home you can download it from the *Statistics & SPSS* page of www.gpryce.com and save it into the **Q:\QUANTS** folder on your hard drive; if this folder does not already exist, you will need to create it using My Computer). The 3,731 observations represent prices from the population of sales on the South Side of Glasgow in the second half of 2004.

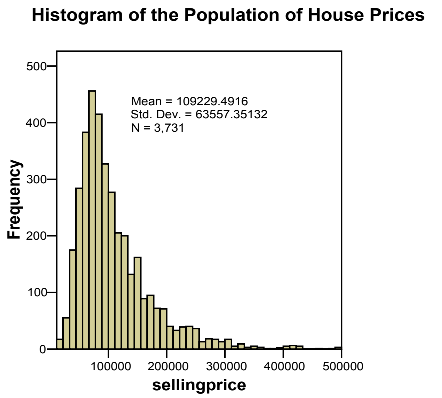
Run descriptives on the **sellingprice** variable to calculate the population mean, and run a histogram to verify that house prices do indeed have a *non*-normal distribution (remember that you can edit the chart by double clicking on it in the Output window – see Chapter 1 for more details):

```
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.

DESCRIPTIVES   VARIABLES=sellingprice .

GRAPH /HISTOGRAM=sellingprice /TITLE= 'Histogram of the Population of House Prices'.
```



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
sellingprice	3731	13000.00	497250.0	109229.5	63557.35
Valid N (listwise)	3731				



Histogram of the Population of House Prices

It is clear from the histogram that house prices are positively skewed. That is, the mean is pulled up by extreme values of selling price in the right tail of the distribution. Since properties cannot sell for negative values, there is a lower bound to the possible range of values and a short left tail to the distribution. The distribution is definitely not normal.

Now suppose that this data represents the population of house prices. Imagine taking repeated samples from this population (each with a sample size of 100 observations). What would the histogram of sample means look like if you took enough samples (e.g. 500 samples)? What would be the value of the mean of all the sample means?

1.3 Lab 1b The Sampling Distribution of the Mean

1. Open the **Glasgow_houseprices_pop_2004q3q4.sav** file which contains house price values in 2004 in Glasgow. You should be able to open the file directly from the www.gwilympryce.co.uk website directly by using the following command:

   ```
   GET FILE='http://www.gwilympryce.co.uk/STATISTICS/Glasgow_houseprices_pop_2004q3q4.sav'.
   ```

   If not, download the file from the *Statistics and SPSS* page of www.gwilympryce.co.uk to your harddisk, then open using the File, Open, Data option on the Menu bar.

2. Let's assume that this is the population of house prices. So to compute the population mean (£109,229) you would run the following command:

   ```
   DESCRIPTIVES VARIABLES=sellingprice   /STATISTICS=MEAN .
   ```

3. Now take a random sample of 100 observations from the population of 3,731 and compute the sample mean. You can do this by running the following lines all in one go (i.e. highlight all three lines and then press Ctrl+R):

   ```
   TEMPORARY.
   SAMPLE 100 FROM 3731.
   DESCRIPTIVES VARIABLES=sellingprice   /STATISTICS=MEAN .
   ```

4. Repeat this process ten times. I.e. repeatedly extract random samples of 100 and computing the sample mean. (You can do this by highlighting and running the above three lines of syntax ten times). You should now be able to see the results from ten Descriptive Statistics tables in your Output window, each reporting the mean house price for a particular sample. Look carefully at how these means vary. You should find that the sample means vary from sample to sample, but not by much -- they tend to be slightly higher or slightly lower than the population mean (£109,229). I.e. the sample means are symmetrically distributed around the population mean. You might want to verify that this is indeed the case by repeating the above process and computing more sample means.

Example of 120 samples:

5. What we would like to do now is see what the histogram of sample means looks like if we extracted a large number of samples of equal size (n = 100).

   The '**x_bar__120sample_means__n_eq_100.sav**' dataset lists sample means from 120 random samples of size *n* = 100 from the sellingprice data. If you open this file you will see a list of 120 sample means under the variable name X_BAR.

   Now run a histogram on the X_BAR variable to get an idea of the shape of the sampling distribution of the mean:

   ```
     GRAPH  /HISTOGRAM=X_BAR
    /TITLE= 'Histogram of 120 Sample Means of sample size n = 100'
         /FOOTNOTE= 'Source: Based on 120 means of size n = 100 extracted from the sellingprice dataset'.
   ```

   You will see that the histogram is approximately normal as the Central Limit Theorem would predict.

   There are a number of important things to note from these results:

   First, the *distribution of sample means is approximately normal* even though the original variable is not normal. This result is universally true and is called the *Central Limit Theorem.* The theorem holds true providing the sample size is large. It is a fabulous discovery because it allows us to make probability statements about how close our sample results are to those we would obtain if we computed them for the population. The sampling distribution we observe here is not quite normal because our selection of samples is itself random and we have only chosen 120 samples. If we were to extract a very large number of large samples we should end up with a histogram of sample means that is very close to normal

Second, *the average of sample means* – the mean of means which is *almost exactly equal to the population mean* of £109,230. This is not a coincidence – the laws of mathematics dictate that if we take an infinite number of random samples, the mean-of-sample-means will exactly equal the population mean. This is a very important finding because it implies that the sample mean offers us an *unbiased estimate* of the population mean. In other words, if we take enough samples, the mean of means will converge to the population mean. Any variation from sample to sample evens itself out in the long run.

Third, as well as calculating the mean of means, we can also compute *the standard deviation of the mean*. This standard deviation has a special name: *the standard error of the mean*. It is given this unique name to distinguish it from the standard deviation of the original variable (usually calculated using a single sample, or occasionally – as in the second step of this example – on the whole population). The standard error of the mean is very important because it tells us how much the sample mean varies from sample to sample. If the standard error is large, it suggests that sample means jump around a lot from sample to sample, and this makes it less likely that any single sample will give us a precise estimate of the population mean. If, on the other hand, the standard error is very small, it suggests that the mean computed from a single sample is likely to give a good estimate of the population mean.

Unfortunately, *we rarely know the true standard error of the mean*. Even in our experiment of 120 random samples, the standard error we have calculated is only an estimate of the true standard error. We would need to take an infinite number of samples to deduce the exact value. In reality, we almost always have just one sample which, on the face of it, means we have no idea at all how the mean varies between samples. Thankfully, however, statisticians have found a clever way of approximating what the standard error is likely to be. Their solution rests on the fact that the standard deviation of a single sample is not unrelated to the standard error of the mean. Other things being equal, if the standard deviation of the original variable (as computed from a single sample) is large, then the standard error is also likely to be large.

1.4 Additional (Optional) Exercises for Lab 1b: Using Macros

The "x_bar__120sample_means__n_eq_100.sav" file was created using an automated routine or program called a macro. Two versions of this macro are provided below, one for use in the LBSS labs and one for use on your home PC or laptop. SPSS is a bit temperamental re whether it will allow such macros to run properly, so **if you can't get the macro to work, don't spend any further time on it.**

The idea behind using the macro is to allow you to extract your own series of random samples, varying the sample size and number of samples, and then seeing what the sampling distribution of the mean looks like.

A "macro" is a series of operations tied to a single command name that the user has created. One such programmed series of commands available from www.gpryce.com is the **CLT** macro. This allows you to draw multiple random samples, and plots a histogram of the means of those samples. The macro works by taking a random sample from the file in memory, computing the mean and saving the mean of that sample as a separate single-cell data file in your current folder. It repeats this until the desired number of samples have been extracted and then combines all the single-cell data files into a single column called X_BAR and saves it as a new dataset. It then runs a histogram on the column of sample means contained in the X_BAR variable, and finally computes a table of descriptives.

To run the macro you need to highlight the following CLT program and run as one command. Rather than typing in this macro,

- copy and paste it into Notepad then cut and paste from Notepad into SPSS (using Notepad helps to get rid of the hidden codes and characters that are bundled when you copy from Word and which can cause the program to falter when pasted into SPSS);

```
*.............................................................................................................................................................
*CLT macro for Lab use.
*.............................................................................................................................................................
DEFINE CLT (variable = !ENCLOSE('(',')') /nsample = !ENCLOSE('(',')') /Npop = !ENCLOSE('(',')') /reps = !ENCLOSE('(',')') ).
!DO !L = 1 !TO !reps.
- TITLE !reps Repeated Samples of size !nsample .
- temporary.
- sample !nsample from !Npop.
- MATRIX.
- GET VARIABLE / VARIABLES = !variable.
- COMPUTE N = NROW(VARIABLE).
- COMPUTE I = MAKE(n,1,1).
- COMPUTE X_BAR = (1/N)*(TRANSPOS(I) * VARIABLE).
- SAVE {X_BAR} / OUTFILE =!CONCAT('"H:\CLT__', !variable, '_sample', !L, '.sav"')  /VARIABLES = X_BAR.
- END MATRIX.
!DOEND.
GET FILE= !CONCAT('"H:\CLT__', !variable, '_sample', '1.sav"').
!DO !J = 2 !TO !reps.
- ADD FILES /FILE=*
/FILE=!CONCAT('"H:\CLT__', !variable, '_sample', !J, '.sav"').
- EXECUTE.
!DOEND.
SAVE / OUTFILE =!CONCAT('"H:\CLT__n', !nsample, !variable, '_sample', 'ALL', !reps, '.sav"') .
TITLE !reps Repeated Samples of size !nsample .
GRAPH /HISTOGRAM=X_BAR /TITLE= 'Histogram of Sample Means from Repeated Samples'.
TITLE !reps Repeated Samples of size !nsample .
DESCRIPTIVES VARIABLES=X_BAR /STATISTICS=MEAN STDDEV MIN MAX .
!ENDDEFINE.
*.............................................................................................................................................................
```

Once you have run the above syntax for lab use, or the home PC version,

```
*.............................................................................................................................................................
*CLT macro for HOME use.
*.............................................................................................................................................................
DEFINE CLT (variable = !ENCLOSE('(',')') /nsample = !ENCLOSE('(',')') /Npop = !ENCLOSE('(',')') /reps = !ENCLOSE('(',')') ).
!DO !L = 1 !TO !reps.
- TITLE !reps Repeated Samples of size !nsample .
- temporary.
- sample !nsample from !Npop.
- MATRIX.
- GET VARIABLE / VARIABLES = !variable.
- COMPUTE N = NROW(VARIABLE).
- COMPUTE I = MAKE(n,1,1).
- COMPUTE X_BAR = (1/N)*(TRANSPOS(I) * VARIABLE).
- SAVE {X_BAR} / OUTFILE =!CONCAT('CLT__', !variable, '_sample', !L, '.sav')  /VARIABLES = X_BAR.
- END MATRIX.
!DOEND.
GET FILE= !CONCAT('CLT__', !variable, '_sample', '1.sav').
!DO !J = 2 !TO !reps.
- ADD FILES /FILE=*
/FILE=!CONCAT('CLT__', !variable, '_sample', !J, '.sav').
- EXECUTE.
!DOEND.
SAVE / OUTFILE =!CONCAT('CLT__n', !nsample, !variable, '_sample', 'ALL', !reps, '.sav') .
TITLE !reps Repeated Samples of size !nsample .
GRAPH /HISTOGRAM=X_BAR /TITLE= 'Histogram of Sample Means from Repeated Samples'.
TITLE !reps Repeated Samples of size !nsample .
DESCRIPTIVES VARIABLES=X_BAR /STATISTICS=MEAN STDDEV MIN MAX .
!ENDDEFINE.
*.............................................................................................................................................................
```

enter the following syntax:

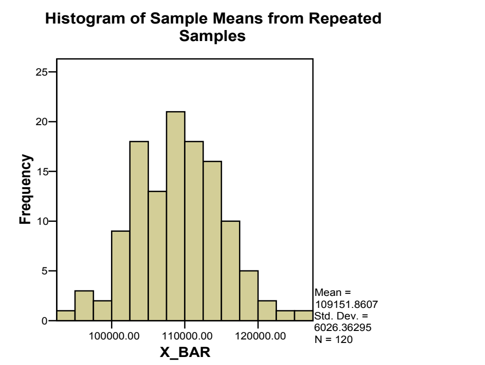
```
CLT variable=(sellingprice) nsample=(100) Npop=(3731) reps=(120).
```

where the items in parenthesis can be changed according to the desired specification. In this case, the variable we are interested in is sellingprice. So this is the name of the variable we have entered in the first set of brackets.

We want to extract samples of size 100, so we enter ‘100’ in the second set of brackets. We need to tell SPSS how large our population is. In this case it is 3,731, which we enter in the third set of brackets. Finally, we enter the number of repetitions – the number of samples we want to draw – in the fourth set of brackets. Let’s go for 120.

If you select and run the above syntax you will be telling SPSS to extract 120 random samples on the sellingprice variable (each sample having 100 observations), and to compute the mean selling price in each sample; and finally to construct a histogram and table of descriptives of those means.

Each time you run this routine you will get slightly different answers, but you should end up with a histogram of means that looks something like the following graph. (Note that if you do decide to run the macro more than once, you will need to re-open the original house price data file each time prior to running the CLT macro).



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
X_BAR	120	94191.71	125245.5	109151.9	6026.363
Valid N (listwise)	120				

1.4.1 Exercise 2.6b Impact on CLT of Reducing Sample Size (Pryce, p.2-20)

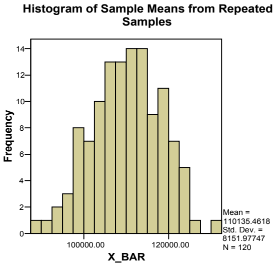
Open up the **Glasgow_houseprices_pop_2004q3q4.sav** file and re-run the CLT command with 60 observations. Copy and paste the histogram of means into a word processor, then re-run the CLT command again with 30 observations. Repeat for 20 observations, 10 and 5.

Your syntax should look like this:

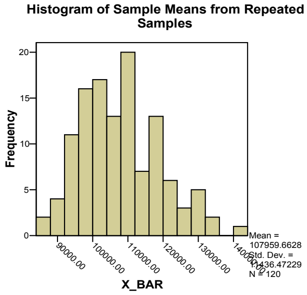
```
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.
CLT variable=(sellingprice) nsample=(60) Npop=(3731) reps=(120).
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.
CLT variable=(sellingprice) nsample=(30) Npop=(3731) reps=(120).
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.
CLT variable=(sellingprice) nsample=(20) Npop=(3731) reps=(120).
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.
CLT variable=(sellingprice) nsample=(10) Npop=(3731) reps=(120).
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.
CLT variable=(sellingprice) nsample=(5) Npop=(3731) reps=(120).
```

and your histograms should look something like the following:

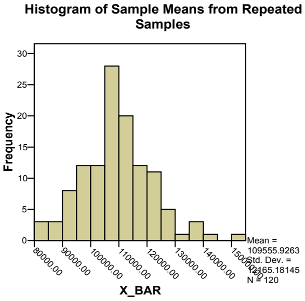
Sample size = 60



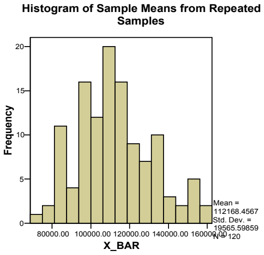
Sample size = 30



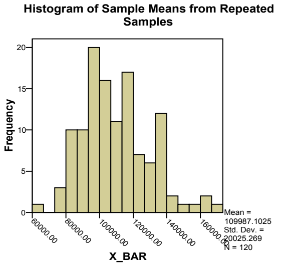
Sample size = 20



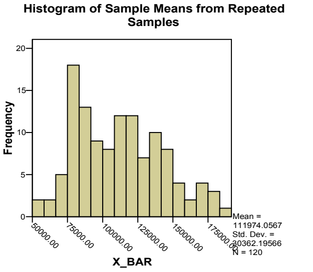
Sample size = 10



Sample size = 10



Sample size = 5



Note that two histograms have been presented for sample size = 10. This is not a mistake but included to demonstrate that if you run the CLT command twice with exactly the same sample sizes etc., you will obtain different results each time. This is because the selection of samples is random and the actual observations in each sample will vary each time.

Although the results vary from experiment to experiment, generally you will find that the histogram of means will become less normal in shape the smaller the sample size, particularly if the underlying variable is non-normal. In fact, the smaller the sample size, the more the distribution of means will resemble the histogram of the original variable. So you will notice that the histogram of means for sample size = 5 has a similarly skewed distribution to the house price histogram, whereas the sample size of 60 or 100 histograms are fairly symmetrical.

The increasing skew as sample size is reduced is worrying because it suggests that our estimate of the population mean will no longer be unbiased. Remember that the population mean equals £109,230. As the sample size gets smaller, the mean-of-means tends to systematically overestimate the population mean. If the skew of the original variable were the other way – a negative skew – we would find that the mean-of-sample-means would tend to underestimate the population mean. We then have to ask ourselves whether a biased estimator is of any use to us at all…

You will also find that the standard error rises significantly the smaller the sample size you use. In the last histogram (sample size = 5), the standard error of the mean equals £30,362, whereas it was just £6,026 in our first experiment (sample size = 100). This is a very important finding because it implies that the smaller the sample size, the greater the variation from sample to sample in the sample mean. So if our sample is small, we can be far less certain that the sample mean will be close to the population mean.

1.4.2 Exercise 2.8 Proportions and the CLT (Pryce, p.2-23)

Open the **Glasgow_houseprices_pop_2004q3q4.sav** file and create a new variable called 'over100k' which equals zero if the house price is less than or equal to £100,000, and equals one if the house price is greater than £100,000. If you calculate the average of this variable it will give you the proportion of properties over £100K, so you can treat it as a proportion.
Run a histogram on 'over100k'. Does it look as you expected? Now obtain the sampling distribution of the over100k variable using the CLT syntax (let the sample size = 100, and use two hundred repetitions).

Answer:

The full set of syntax needed to complete this exercise is as follows:

```
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.
COMPUTE over100k = 0.
IF(sellingprice > 100000) over100k = 1.
EXECUTE.
DESCRIPTIVES   VARIABLES=over100k  .
GRAPH /HISTOGRAM=over100k /TITLE= 'Histogram: Proportion of House Prices > £100k'.
CLT   variable=(over100k)   nsample=(100)   Npop=(3731)   reps=(200).
```

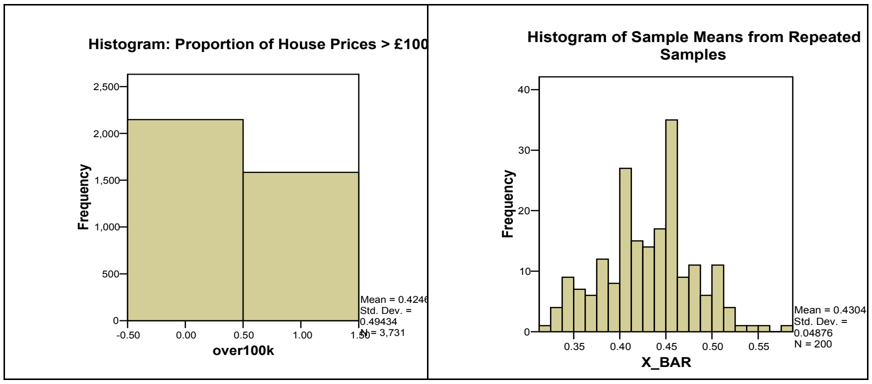
We have used the COMPUTE syntax to calculate the over100k variable above. You can do this using the point-n-click method by selecting Transform, Compute, type in over100k in the **Target Variable** box, then enter 0 in the **Numeric Expression** box, then click OK (or Paste); then Transform, Compute, select over100k as the target variable, enter 1 in the **Numeric Expression** box, then click the If button, click "Include cases if satisfies condition", then enter "sellingprice > 100000" in the box, click Continue and OK (or Paste).

You could alternatively have used the RECODE syntax,

```
GET  FILE='Q:\QUANTS\Glasgow_houseprices_pop_2004q3q4.sav'.
RECODE  sellingprice  (Lowest thru 100000=0)  (100000.01 thru Highest=1)  INTO  over100k .
EXECUTE .
```

You can also access the RECODE facility using the point-n-click method by selecting Transform, Recode, Into Different Variables, select **sellingprice** and paste it into the **Input Variable -> Output Variable** box, then under the **Output Variable** box on the right, enter "over100k" in the **Name** box, enter a description of the new variable in the **Label** box (such as "Selling price over £100k"), then click the Change button. Then click the Old and New Values button, then select the Range lowest through option, and enter 100000 in the box, then under the **New Value** box (top right) enter 0, and click the Add button. Now select the Range through highest option, and enter 100000.01 in the box, then under the **New Value** box (top right) enter 1, and click the Add button. Then click Continue, and OK (or Paste).

The histogram of over100k should look like the graph below on the left, where as the sampling distribution of the proportion should look like the histogram on the right:



1.4.3 Additional Exercise: Sampling Distribution of Mean Landvalue .

(NB: this exercise uses one of the standard SPSS datafiles which come with recent versions of SPSS).

1. Open the **Home sales [by neighborhood].sav** file, which is in the **C:\Program Files\SPSS\** folder on the lab computers. Compare the histogram of the population land value with the sampling distribution of mean land value.

2. Compare the population mean land value with the mean-of-means (i.e. the average value of the sampling distribution of mean land value).

3. Attempt the exercise using sample sizes of 100, 30, and 5; drawing 140 samples in each case.

Answer:

*1. and 2.: First compute the population mean and run the population histogram:

```
GET  FILE='C:\Program Files\SPSS\Home sales [by neighborhood].sav'.
DESCRIPTIVES VARIABLES=landval  /STATISTICS=MEAN STDDEV.
GRAPH /HISTOGRAM=landval /TITLE= 'Histogram: landval'.
```

*3. Then run the CLT syntax for the required sample sizes, based on 140 repetitions in each instance:

```
*Sample size of 100.
GET  FILE='C:\Program Files\SPSS\Home sales [by neighborhood].sav'.
DESCRIPTIVES VARIABLES=landval  /STATISTICS=MEAN STDDEV.
CLT variable=(landval) nsample=(100) Npop=(2440) reps=(140).

*Sample size of 30.
GET  FILE='C:\Program Files\SPSS\Home sales [by neighborhood].sav'.
DESCRIPTIVES VARIABLES=landval  /STATISTICS=MEAN STDDEV.
CLT variable=(landval) nsample=(30) Npop=(2440) reps=(140).

*Sample size of 5.
GET  FILE='C:\Program Files\SPSS\Home sales [by neighborhood].sav'.
DESCRIPTIVES VARIABLES=landval  /STATISTICS=MEAN STDDEV.
CLT variable=(landval) nsample=(5) Npop=(2440) reps=(140).
```

1.5 Tutorial 1: Density Functions, CLT and Z-Scores1

Types of Variable

1. What are the different types of variable?

2. How would you use Graphs and Summary Statistics to summarise variables of each type?

3. Compute the mean, standard deviation and coefficient of variation for X:

X
6
2
4
7

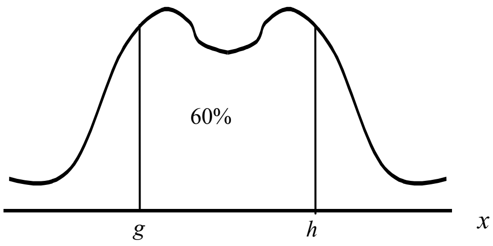
Histograms, Density Functions, and the Normal Distribution

1. In what sense do histograms tell us more about a variable than the mean or the standard deviation?

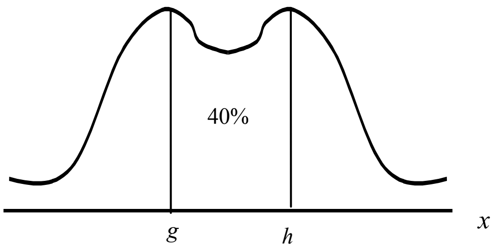
2. What is a density function? How is it related to a histogram?

4. In each of the graphs below, (i) what % of values of *x* are greater than *h*, and (ii) what is the probability of randomly choosing an observation with a value of *x* less than *g*?

(*a*)



(*b*)



5. What do we mean when we say that a variable is normally distributed?

6. Why is the normal distribution so important in statistics?

The Standard Normal Curve

1. What is the z-distribution and how is it useful?

2. How does one convert a value of a normally distributed variable *x* to a point on the standard normal curve? E.g. if the mean value of *x* is £30,000, and the standard deviation of *x* = £10,000, what is the z-score associated with *x* = £40,000

Additional Activities:

If there is time, go over the standard normal table (attached) and talk about how one goes about:

1. Calculating the probabilities associated with z using tables

2. Calculating the probabilities associated with z using SPSS

3. Finding the z value associated with a given probability

4. Finding z-scores that bound Central Probabilities

1.6 Reading 1

1.6.1 Pryce I&S in SPSS

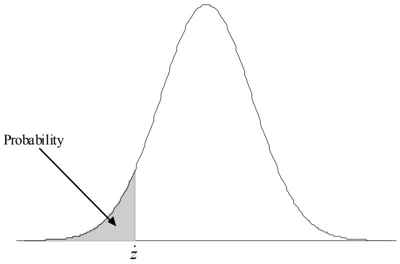
- *Pryce, Sections 1.3, 1.5, 2.4
- *Pryce, Section 2.6
- *Pryce, Section 2.5
- Pryce, rest of Chapter 1.

1.6.2 M&M 4th Ed.

- Section 1.3; Chapter 5.

1.7 Table 1A Standard Normal Probabilities

Each entry in the body of the table is the area under the standard normal curve to the left of *z*.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.40	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.30	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.20	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.10	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.00	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.90	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.80	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.70	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.60	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.50	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.40	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.30	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.20	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.10	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.00	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.90	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.80	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.70	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.60	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.50	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.40	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.30	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.20	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.10	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.00	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.90	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.80	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.70	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.60	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.50	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.40	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.30	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-.20	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.10	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

L2 Calculating z-scores

2.1 Structure of Lecture 2

1. **Calculating the probabilities associated with z using tables**
2. **Calculating the probabilities associated with z using SPSS**
3. **Finding the z value associated with a given probability**
4. **Finding z-scores that bound Central Probabilities**
5. **Calculating z for a normally distributed variable**
6. **Calculating z scores for Sampling Distributions**
7. **Student's t-distribution**

2.2 Lab 2a and 2b

2.2.1 Examples 3.2.1a, b, c, d Calculate the Probability that z is less than zi (Pryce, p. 3-3)

In each of the exercises below, first draw a diagram to help you understand the problem.

Example 3.2.1(a) Calculate the probability that z is less than –3.4
Run your finger down the column of figures under z in Table A (see page 13 below). You won't have to go very far since –3.4 is the very first row. The second decimal place of –3.4 is 0, so we need the probability listed in the column headed ".00", which is 0.0003. So we can say Pr(z < –3.4) = 0.0003, which is a very small probability indeed (there is less than a tenth of 1 percent chance that z will be less than –3.4).

Example 3.2.1(b) Calculate the probability that z is less than –1.23
We want the value identified by the row of –1.2 and the column .03 which is .1093. That is, there is a 11% chance that z is less than –1.23.

Example 3.2.1(c) Calculate the probability that z is greater than 2.34
Table A only lists the probabilities associated with negative values of z. However, because the distribution is symmetrical around z = 0, the probability that z is less than a negative value is equal to the probability that z is greater than the equivalent positive value. So, Pr(z > 2.34) = Pr(z < –2.34), which is identified by running your finger down the column of z values until you get to –2.3, then moving along that row until you get to the column headed .04. The value you should arrive at is .0096. In other words, Pr(z > 2.34) = Pr(z < –2.34) = .96%, which is a hair's-breadth short of one percent.

Example 3.2.1(d) Calculate the probability that z lies between –1.96 and 1.96
We know that the area between z = –1.96 and z = 1.96 is equal to 1 minus the sum of two probabilities: the probability that z is less than –1.96 and the probability that z is greater than 1.96. The probability that z is less than –1.96 can be found from Table A by going along the z = –1.90 row until you get to the column headed ".06". The figure you arrive at is .0250. Because the distribution is symmetrical, this means that the probability that z is greater than the positive value of 1.96 is also equal to .0250. So the sum of the probabilities in the two tails associated with ±1.96 is .0250 + .0250 = .05. So to compute Pr(–1.96 < z < 1.96) we must simply calculate 1 – 0.05 = 0.95. So, Pr(–1.96 < z < 1.96) = 0.95, which is the same as saying that 95% of z-values lie between ±1.96.

2.2.2 Example 3.2.2 Find Pr(z<-0.95) using CDFNORM (Pryce, p. 3-4).

Type the following syntax:

```
COMPUTE Pr_z_lt_zi = CDFNORM(–0.95).
FORMATS Pr_z_lt_zi (F5.4).
EXECUTE.
```

The values of your Pr_z_lt_zi variable should now all equal 0.1711. In other words, the probability that z is less than –0.95 equals 0.1711, or 17.11%.

Now repeat the exercises set out in examples 3.2.1a to 3.2.1d above, using SPSS commands and compare your answers with those obtained using Table A.

2.2.3 Exercise 3.2 Calculate the probabilities associated with z-scores (Pryce, p.3-6)

For each of the following exercises, draw an appropriate diagram of the z-distribution to help you understand the problem, then compute the probabilities using both Table A and the relevant macro command.

1. Find the proportion of z-scores that are less than 1.8.
2. Find the proportion of z-scores that are less than 3.
3. Find the probability that z is less than 2.

Find the following probabilities using the standard normal table:

a) Prob($0 < z < 1.87$) i.e. the probability that z lies between 0 and 1.87
b) Prob($-1.11 < z < 0$) i.e. the probability that z lies between –1.11 and 0.
c) Prob($-1.18 < z < 0$)
d) Prob($-1.26 \leq z < 2.11$)
e) Prob($-2.06 < z \leq 0.63$)
f) Prob($z \geq 1.47$)
g) Prob($z < 2.05$)
h) Prob($z < -0.99$)
i) Prob($z > -1.05$)
j) Prob($0.86 \leq z \leq 1.72$)
k) Prob($z < 1.58$)
l) Prob($-1.96 < z < -1.28$)
m) Prob($0 < z < 1.83$)
n) Prob($-2.08 < z < 0.63$)

2.2.4 Exercise 3.2.5 Find the values of z associated with given probabilities (Pryce, p. 3-13)

1. Find a value z_i such that:
   a) Prob($z \geq z_i$) = 0.06 i.e. the probability that z is greater than z_i is 0.06.
   b) Prob($z > z_i$) = 0.95
   c) Prob($z < z_i$) = 0.80
2. Find z_i such that 95% of z-scores are above z_i. That is, Prob($z > z_i$) = 95%.
3. Find z_i such that Prob($-z_i < z < z_i$) = 95%.

2.3 Reading 2

2.3.1 Pryce I&S in SPSS

- *Pryce, Chapter 3

2.3.2 M&M 4th Ed.

- M&M section 1.3 and chapter 5.

L3 Introduction to Confidence Intervals

3.1 Structure of Lecture 3

1. **Inuition Behind CIs**
2. **Three steps of confidence interval estimation**
3. **Large Sample Confidence Interval for the Mean**
4. **Small Sample Confidence Interval for the Mean**

3.2 Labs 3a and 3b

3.2.1 Exercise 4.2 Large Sample Confidence Interval for One Mean (Pryce, p.4-3)

Find z_i such that Prob($-z_i < z < z_i$) = 95%.

3.2.2 Exercise 4.2.1 The Three Steps of CIs (Pryce, p. 4-4)

Suppose you are interested in the disappearance of thousands of civil servants and other workers during Joseph Stalin's Great Purge in Soviet Russia 1936-38. One of the questions you are interested in is the average age of the workers when they disappeared. Your thesis is that Stalin felt most threatened by older, more established 'enemies', and so you anticipate their average age to be over 50. Unfortunately, you only have access to 506 records on the age of individuals when they disappeared.

You have calculated the average age in this sample to be 56.2 years, which would appear to confirm your thesis. The standard deviation of your sample was found to be 14.7 years. Assuming that your 506 records constitute a random sample from the population of those who disappeared (a questionable assumption?), calculate the 95% confidence interval for the population mean age. Does your expected value for the population average age fall below the interval? Compute also the 99% confidence interval and reconsider whether your theorised average age still falls below the range of possible values for the population mean.

3.2.3 Exercise 4.2.1 Calculating the Confidence Interval for the Mean

In the following questions, work out the solution by hand and then run the syntax to check your results:

1. As an economic historian, you are interested in the average age of the heads of household who had their homes repossessed in Scotland during the Great Depression. Records are kept at local sheriff court archives, and from an examination of a sample of 200 records from across Scotland, you find that the average age of head of household was 34.5 years with standard deviation of 20 years. What is the 90% confidence interval for the mean age?

2. For your dissertation, you interview 78 nurses and ask them their weight. The average weight within your sample turns out to be 70kg with a standard deviation of 20kg. What is the 95% confidence interval for the mean weight of all nurses?

3. For your dissertation, you interview 90 teenage girls, randomly selected from across Glasgow about how they would rate the Government's sexual health advice service. You find that, on an ascending scale of zero to 100, the average rating given is 45.3 with a standard deviation of 35.2. What is the 99% confidence interval for the mean rating of official sexual health advice by all Glaswegian teenage girls?

4. As an Occupational Therapist, you are interested in how long it takes amputees to recover from their operation and reach a reasonable level of proficiency with their prosthetic replacement. You decide to take a sample of 150 amputees and find an average recovery time of 97.5 days, with a

standard deviation of 60. What is the 90% confidence interval for the population mean recovery time?

5. For your dissertation you attempt to estimate the exposure to toxic substances of workers in the chemical industry in the post war period. You have examined a random sample of 354 records from the archives of random medical tests from a range of firms from that period. This sample has been provided by the industry with a press release stating that the average exposure score is 194.35. The government had previously stated that retrospective collective compensation claims from workers could only proceed if there was evidence that the average exposure of all workers in the industry exceeded the threshold of 200 set down in European health and safety legislation. You examine the dataset and calculate the standard deviation of the sample scores to be 142.98. Do you think the compensation claim has grounds to proceed?

3.2.4 Exercise 4.3 Computing Small Sample CIs (Pryce, p. 4-13)

1. As a war historian, you are interested in the average survival time of soldiers on the front line in the First World War. There are over 1,000 records available for examination at the Ministry of Defence, but because of the considerable time and cost of accessing these records, you decide to examine a random sample of just 16 records. Your sample reveals an average of 18.7 weeks, with a standard deviation of 2.8 weeks. Find the 90% confidence interval estimate of the mean survival time of all soldiers on the front line.

2. For your PhD, you want to estimate the number of times an American phrase or pronunciation occurs in a typical 5 minute conversation between teenage youths in Liverpool. Because of the time taken to build up sufficient rapport with such youths for them to speak in a relaxed and typical way, you only manage to observe 23 such conversations. The average number of Americanisms in 5 minute conversations amongst your small sample is 137.71, with a standard deviation of 69.56. What is the 95% confidence interval of the average incidence of Americanisms for all youths in Liverpool?

3. As a political scientist you have been researching the political perspectives of students from working class backgrounds. You have surveyed 36 students, asking them to rate on an ascending scale of 0 to 100 their preference for the public ownership of the means of production. The average rating in your sample is 59.75 with a standard deviation of 42.93. What can you say from your small sample about the political perspectives of working class students as whole?

*3.3 Tutorial 2: Confidence Intervals**

*See also *T2 Info Sheet: Summary Guide to Confidence Intervals*

1. *Need-to-Know* Background

- List different kinds of variable? What are their characteristics?
- Why is distinguishing between variables types important for calculating confidence intervals?
- What is a distribution curve?
- What is the central limit theorem? What are its limitations?
- How are different distribution curves important in calculating confidence intervals?

2. Confidence intervals

a) What are the key components of a confidence interval?

b) What is it useful for?

c) We need four pieces of information to calculate a confidence interval. What are they? What notation is used to represent them in formulae?

i. .
ii. .
iii. .
iv. .

d) What is variance and when does it matter to calculating confidence intervals?

3. Working Examples

Confidence intervals can be calculated manually, using formulae, or using SPSS syntax (macros).

- **Look over *T2 Info Sheet* to review hand calculation and macros; if you are already comfortable with this, there is also background on a basic hypothesis test which can sometimes be used in the calculation of confidence intervals.**

3.3.1 Hand Calculation Using Formulae

Note: the equation editor can be found in Microsoft by opening Word and looking under Insert/ Object. It can be temperamental – save often and split large documents into smaller sections!

- **What do these symbols mean?**

$\bar{x}$	n
s	c
μ	σ
H_o	H_1
F	z
t	α
P	

Example 1:

As a war historian, you are interested in the average survival time of soldiers on the front line in the First World War. There are over 1,000 records available for examination at the Ministry of Defence, but because of the considerable time and cost of accessing these records, you decide to examine a random sample of just 16 records. Your sample reveals an average of 18.7 weeks, with a standard deviation of 2.8 weeks. Find the 90% confidence interval estimate of the mean survival time of all soldiers on the front line.

- **What value goes next to these symbols?**

$\bar{x} =$

$n =$

$s =$

$c =$

Example 1 Using Hand Calculation:

Use the Three Steps:

STEP 1: Choose the appropriate test statistic

a) Are the variable(s) continuous or discrete
b) Are we working with 1 or 2 samples?**
c) For 2 samples only: is the variance equal (pooled) or unequal?
d) Are we working with large or small samples?

- Why *this* formula?
- How would you 'translate' it into English?

$$\mu = \bar{x}_i \pm t_i \frac{s}{\sqrt{n}}$$

STEP 2: Establish the value of t or z

As with the z distribution, each value of t has an associated statistic, based on the degrees of freedom (df) of that particular sample the specified confidence interval. In this case, we want to be 90% confident that the population parameter (e.g. the mean or the proportion) falls within the range defined by the interval estimate.
First calculate the degrees of freedom, base upon the sample size:

df = (n-1)
df = (?-1)
df = (?)

Next calculate the area in each of the tails of the distribution curve (based on our desired confidence level of a 90% confidence interval):

Prob (-t*<t<t*) = 0.90

Area of each tail = $\left(\frac{0.10}{2}\right)$

Area of each tail = 0.05

Finally: look up a t table to find the ± t value associated with the specified df and an upper/ lower tail probability of 0.05. This gives a result of -1.753

So:
Prob (-t*<t<t*) = 0.90
Prob (-1.753 < t < 1.753) = 0.90

STEP 3: Calculate the confidence interval

Using the values calculated above, substitute the figures into the formula and calculate manually:

$$\mu = \bar{x}_i \pm t_i \frac{s}{\sqrt{n}}$$

$$\mu = ? \pm ? \frac{?}{\sqrt{?}}$$

The final result calculates parameters for the population mean equal to the sample mean, plus and minus the margin of error, giving the upper and lower boundary of the confidence interval.

e.g.
μ = 18.7 ± 1.2271
= 18.7 + 1.2271 AND 18.7 – 1.2271
= 19.9271 AND 17.4729

Therefore we can say with 90% confidence that the population mean survival time of all soldiers on the front line lies between 17.4729 weeks and 19.9271 weeks when using this sample as an estimate

Example 1 Using Syntax:

- **Write out the macro, including the appropriate values:**

```
Small sample confidence interval for the population mean:
___________?   n = (  ) x_bar = (   ) s = (   ) c = (   ).
```

- **What does this output mean?**

```
       n        x_bar       TiL          SE       err        Lower        Upper.
  16.00000   18.70000   -1.75305    .70000    1.22714   17.47286   19.92714.
```

Example 2:

Suppose the mean height of girls, in your sample of ten, equals 100 cm (standard deviation = 30cm), and the mean height of 12 boys is 94cm (s.d. = 31cm). Calculate the 95% confidence interval for the difference in population means assuming homogenous[2] variances.

$n_1 =$	$s_1 =$
$n_2 =$	$s_2 =$
$x_1 =$	$c =$
$x_2 =$	

Example 2 Using Syntax:

- **Write out the macro, including the appropriate values:**

```
 ________?  n1=(  )   n2=(  )   x_bar1=(  )   x_bar2=(  )   s1=(  )   s2=(  )   c=(  ).
```

- **What does this output mean?**

```
SAMPDIFF        SP        TiL        SE          err           Lower          Upper.
 6.00000   30.55405   -2.08596   13.08246   27.28954   -21.28954   33.28954.
```

[2] If this was not specified, we would check for equality of variances – see other sheet

Example 2 Using Hand Calculation:

Three main steps:

STEP 1: Choose the appropriate test statistic

a. Are the variable(s) continuous or discrete
b. Are we working with 1 or 2 samples?**
c. For 2 samples only: is the variance equal (pooled) or unequal?
d. Are we working with large or small samples?

STEP 2: Establish the value of t or z

First calculate the degrees of freedom, base upon the sample size:

df = (?)

Next calculate the area in each of the tails of the distribution curve (based on our required confidence level):

Prob (-t*<t<t*) =___?

Area of each tail = $\left(\frac{?}{?}\right)$

Area of each tail = ___?

Finally: look up a t table:
Prob (___ < t <___) = 0.95

STEP 3: Calculate the confidence interval

Because we are working with two samples, at this stage we would normally calculate whether the samples had equal or unequal variance. This involves a hypothesis test – more of which later in the course. For now, we have been told that the samples have equal variance. The appropriate formula is the small independent samples confidence interval for the difference between two means:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

However, this formula uses a pooled estimator of variance (s_p). This must be calculated first so that the result can be inserted into the main formula.

We compute s_p as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- **What values would you substitute in as A to F?**

$$s_p = \sqrt{\frac{(A_1 - 1)B_1^2 + (C_2 - 1)D_2^2}{E_1 + F_2 - 2}}$$

The final value of s_p comes out as

$s_p = \sqrt{\frac{18671}{20}}$ = 30.55 This value can be substituted into the formula for the small independent samples confidence interval for the difference between two means.

- **Which other values are substituted in as G to K?**

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t * s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\mu_1 - \mu_2 = (G_1 - H_2) \pm Ix30.55 \sqrt{\left(\frac{1}{J_1} + \frac{1}{K_2}\right)}$$

The final confidence interval is:

$\mu_1 - \mu_2 = 6 \pm 27.27$

- **What are the upper and lower bounds of the confidence interval?**
- **Why is the interval so wide?**

3.4 T2 Info Sheet: Summary Guide to Confidence Intervals

1. The Three Steps: Calculating Confidence Intervals

STEP 1: Choose the appropriate test statistic

i. Are the variable(s) continuous or discrete
ii. Are we working with 1 or 2 samples?**
iii. For 2 samples only: is the variance equal (pooled) or unequal?
iv. Are we working with large or small samples?

STEP 2: Establish the value of t or z

STEP 3: Calculate the confidence interval

**NB: When you are working with two samples – to choose the correct test statistic you must first CALCULATE (do not guess!) whether the samples have Equal or Unequal Variances. See below for the details of F tests

2. Using Macros

Test Statistic	Macro
Large sample CI for 1 mean	CI_L1M n=( ) x_bar = ( ) s = c = ( ) .
Small sample CI for 1 mean	CI_S1M n=( ) x_bar = ( ) s = c = ( ) .
Small independent samples CI for difference between 2 means (pooled variance)	CI_S2MP n1=( ) n2 = () x_bar1 = ( ) x_bar2 = ( ) s1 = ( ) s2 =( ) c=( ).
Small independent samples CI for difference between 2 means (different variance)	CI_S2MD n1=( ) n2 = () x_bar1 = ( ) x_bar2 = ( ) s1 = ( ) s2 =( ) c=( ).
Large sample CI for one proportion (both traditional and Wilson methods of calculation)	CI_L1P n=( ) x = ( ) c = ( ).
Large sample CI for comparing two proportions (both traditional and Wilson methods of calculation)	CI_L2P n1=( ) n2 = ( ) x1 = ( ) x2 = ( ) c = ( ).
Sample size for desired margin of error for the mean	N_L1M e ( ) c = ( ) s = ( ).

3. F Tests: A Four Step Process

The F test is used to calculate whether two samples have equal or unequal variance. The null hypothesis (H_o) is that the variance in each case is equal; the alternative hypothesis (H_1), that the variance is not equal.

1. Specify the null and alternative hypotheses:

i.e. $H_o : \sigma_1 = \sigma_2$

$H_1 : \sigma_1 \neq \sigma_2$

2. Specify the level of significance (α) and the test statistic for the F test

$\alpha = P$ (probability of wrongly rejecting H_o [3]) = 0.05[4]

[3] Rejecting the null hypothesis of equal variance when it is in fact true is referred to as a “type 1 error”

$F = \frac{s_1^2}{s_2^2}$ The F test will test for equality of variance between the two samples

3. Specify the decision rule

Reject H_o (the hypothesis that the variances are equal) if (and only if

P (the calculated level of type 1 error) is not greater than the tolerated level of significance specified.

i.e. Reject H_o if $P \leq \alpha$

so Reject H_o if $P \leq 0.05$

4. Compute P using the F statistic (see step 2: square the standard deviation of each sample and divide one by the other) and state your conclusion.

The associated significance of the F statistic can be found by looking up a table of the F-distribution or using syntax:

CDF.F. CDF.F (q, df1, df2).

Substitute the value of F you have calculated for q and when you run the function it will return the cumulative probability that a value from the F distribution, with degrees of freedom df1 and df2, will be less than the figure specified.

OR

```
COMPUTE Fsig = 1 – CDF.F(q,df1,df2).
FORMATS Fsig (F5.4).
EXECUTE.
```

OR

Using the macro

```
H_S2VF  n1 =( ) n2 = ( ) s1=( ) s2=( ).
```

A significance level of less than 5% would lead us to reject the null hypothesis of equal variance (with 95% confidence).

The decision rule (specified in step three) is the criterion for accepting or rejecting the H_o of equal variance.

[4] This would give a confidence level of 95% ; for a confidence level of e.g. 99%, you would want to specify a significance of 0.01

3.5 Reading 3

3.5.1 Pryce I&S in SPSS

- *Pryce, Chapter 4

3.5.2 M&M 4th Ed.

- Sections 6.1 (p. 415-429); 7.1 and 7.2. Chapter 8.

L4 Confidence Intervals for All Occasions

4.1 Structure of Lecture 4

1. CI for two indep means

- Pooled Variances
- Different Variance

2. CI for two paired means

3. CI for one proportion

4. CI for two proportions

5. Sample size determination

6. FAQs about CIs

4.2 Labs 4a and 4b

4.2.1 Exercise 4.4.1 Pooled Variance CI for the difference between 2 Means (Pryce, p.4-16)

Suppose the mean height of girls, in your sample of ten, equals 100 cm (standard deviation = 30cm), and the mean height of 12 boys is 94cm (s.d. = 31cm). Calculate the 95% confidence interval for the difference in population means assumimg homogenous variances.

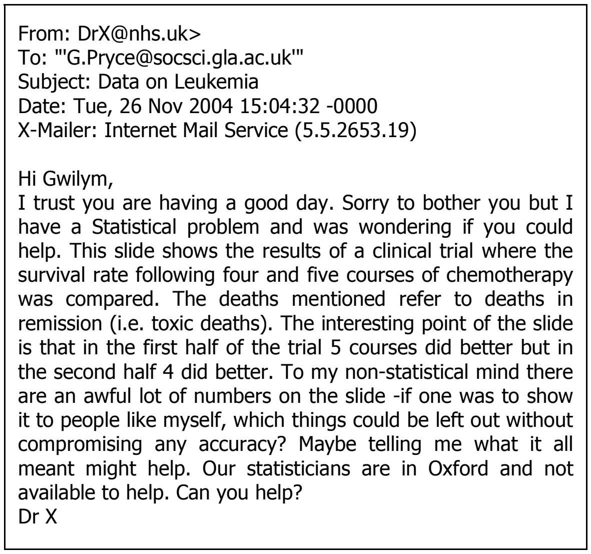
4.2.2 Exercise 4.4.2 Heterog. Variance CI for the difference between 2 Means (Pryce, p.4-17)

Run the heterogeneous variance CI method on the child height example above.

4.2.3 (a) Exercise 4.6 Calculating Confidence Intervals for Proportions (Pryce, Chapter 4):

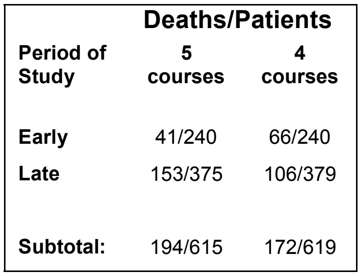
Consider the following email from a Consultant Paediatrician and the associated Powerpoint slide (based on a real-life statistical problem). The data refer to the number of deaths of children with Leukaemia (numerator) over the total number in the sample (denominator). There are four samples: two in the "early" period (samples taken of children with Leukaemia in 1995-98) and two in the "late" period (samples taken of children with Leukaemia in 1998-2002). The first sample in each period is of children given 5 courses of chemotherapy. The second sample in each period is of children given 4 courses of chemotherapy. You would think that 5 courses of chemotherapy would result in fewer deaths than 4 courses, but the picture seems less straightforward and so rather puzzling. To solve Dr X's problem, answer the following questions:

a) Derive a confidence interval for the proportion of deaths of children who underwent 5 courses in the Early period.
b) Derive a confidence interval for the proportion of deaths of children who underwent 4 courses in the Early period and compare with (i).
c) Derive a confidence interval for the proportion of deaths of children who underwent 5 courses in the Late period.
d) Derive a confidence interval for the proportion of deaths of children who underwent 4 courses in the Late period and compare with (iii).
e) Derive a confidence interval for the proportion of deaths of children who underwent 5 courses in the combined periods.
f) Derive a confidence interval for the proportion of deaths of children who underwent 4 courses in the combined periods and compare with (v).



From: DrX@nhs.uk>
To: "'G.Pryce@socsci.gla.ac.uk'"
Subject: Data on Leukemia
Date: Tue, 26 Nov 2004 15:04:32 -0000
X-Mailer: Internet Mail Service (5.5.2653.19)

Hi Gwilym,
I trust you are having a good day. Sorry to bother you but I have a Statistical problem and was wondering if you could help. This slide shows the results of a clinical trial where the survival rate following four and five courses of chemotherapy was compared. The deaths mentioned refer to deaths in remission (i.e. toxic deaths). The interesting point of the slide is that in the first half of the trial 5 courses did better but in the second half 4 did better. To my non-statistical mind there are an awful lot of numbers on the slide -if one was to show it to people like myself, which things could be left out without compromising any accuracy? Maybe telling me what it all meant might help. Our statisticians are in Oxford and not available to help. Can you help?
Dr X



	Deaths/Patients	
Period of Study	**5 courses**	**4 courses**
Early	41/240	66/240
Late	153/375	106/379
Subtotal:	194/615	172/619

(b) Additional Exercise for those with a medical interest: Differences Among Outcome Measures in Occupational Low Back Pain

Ferguson et al (2005) note that, "Low back pain recurrence rates have been reported as high as 70%; however, these rates vary greatly depending on the definition of recurrence (1–6). The high rate of recurrent low back pain as well as variability suggests that we do not have good understanding of low back pain recovery. Examining the various outcome measures that have been used in the past and developing our understanding of the relationship among them may provide insight as to why recurrence rates are so high." They construct a cross-sectional survey of 208 workers who have returned to work after a work-related episode of low back pain, and compare different outcome measures of recurring lower back pain after returning to work.

They find apparent differences in the percentage of subjects recovered for different outcome measures, as summarised in the table below (x refers to the number of subjects who have recovered according to each criteria, and p gives x as a proportion of the sample size, n).

	Criteria used to define "recovery"	x	p
(a)	full duty return to work	206	0.99
(b)	activities of daily living	52	0.25
(c)	symptoms	35	0.17
(d)	functional performance probability	26	0.125
(e)	motion	123	0.59
(f)	velocity	27	0.13
(g)	acceleration	21	0.10
	n	208	

Source: Ferguson, S.A. et al (2005) Differences Among Outcome Measures in Occupational Low Back Pain, Journal of Occupational Rehabilitation, 15(3) 329 - 341.

a) Compare proportions (a) and (e) at the 95% confidence level
b) Compare proportions (b) and (c) at the 95% confidence level
c) Compare proportions (c) and (d) at the 95% confidence level
d) Compare proportions (f) and (g) at the 95% confidence level
e) Comment on your results in each case.

4.2.4 Example 4.7 Using SPSS to calculate CIs when you hav the original data (Pryce, p.4-23)

1. Open up the data on house prices and run the following EXAMINE syntax on purchase price (note the /PLOT NONE /CINTERVAL 95 qualifiers which suppress the graph output and specify the confidence level respectively). Now check how this interval compares to that obtained using our macro method.

2. You have a sample of 83 observations on income as listed in the **income.sav** data file (which you can obtain from the Downloads page of www.geebeejey.co.uk or from the Q: drive of the Faculty labs). Compute (a) a 90% confidence interval for income; and (b) a 90% confidence interval for the proportion of households with incomes of £80,000 and over.

4.2.5 Exercise 4.7 Using the GRAPH/ERRORBAR command to compare CIs (Pryce, p.4-25)

1. Open the **auction.sav** data set. The file records the estimated and final sale prices of 100 items entered for auction at venues in Cumberland and Durham. Use the EXAMINE VARIABLES command to compare the 95% confidence intervals of the auctioneer's estimated value (value) and the actual purchase price (purchase). Now compare the two variables using the GRAPH /ERRORBAR command. Does the graph confirm what you found from using the 'Explore' function? What happens if you change the confidence level to 80%? Explain your findings.

2. Compare the 95% confidence intervals for the purchase price of lots entered in Durham vs those entered in Cumberland. What do your results tell you?

4.2.6 Exercise 4.8.2 Sample Size Determination

For your PhD, you want to estimate the mean hourly wage rate of unskilled labour in Easterhouse within ±£0.10 at the 99% confidence level. A 1987 study (large sample size) by the Department of Employment resulted in a standard deviation of £0.85. Using this as an approximation for σ, compute the necessary sample size to arrive at the desired level of accuracy.

4.3 Reading 4

4.3.1 Pryce I&S in SPSS

- *Pryce, Chapter 4

4.3.2 M&M 4th Ed.

- M&M section 6.3 and exercises for 6.3
- Sections 6.1 (p. 415-429); 7.1 and 7.2. Chapter 8.

L5 Introduction to Hypothesis Tests

5.1 Structure of Lecture 5

1. **Significance**
2. **Four steps of hypothesis testing**
3. **Hypotheses about the population mean**
    - large samples
    - small samples

5.2 Labs 5a and 5b

5.2.1 Exercise 5.4 Large Sample Hypothesis Tests on One Mean (Pryce, p.5-8)

Again, suppose your area of research is the disappearance of thousands of civil servants and other workers during Joseph Stalin's Great Purge in Soviet Russia 1936-38. One of the questions you are interested in is the average age of the workers when they disappeared. Your thesis is that Stalin felt most threatened by older, more established 'enemies', and so you anticipate their average age to be over 50. Unfortunately, you only have access to 506 records on the age of individuals when they disappeared.

You have calculated the average age in this sample to be 56.2 years, which would appear to confirm your thesis. The standard deviation of your sample was found to be 14.7 years. Assuming that your 506 records constitute a random sample from the population of those who disappeared (a questionable assumption?), use hypothesis testing to investigate your theory about the average age of the Disappeared.

5.2.2 Exercise 5.5 Small Sample Hypothesis Tests on One Mean (Pryce, p.5-9)

1. A machine used to fill pre-packaged emergency insulin injections is correctly adjusted if the average net weight of insulin is 11 ounces per syringe. A random sample of 107 syringes had an average fill of 10.92 ounces and standard deviation of 0.28 ounces. Is the machine adjusted properly? Test at the 5% significance level.

2. A newspaper has claimed that the average time GPs spend with patients in a particular session has fallen to an all-time low of 3.6 minutes. A report from the Scottish Executive disagrees, contending that the health service has met New Labour's manifesto target of 4 minutes average consultation time, and that the data used by the newspaper was based on a freak sample. The newspaper's survey was based on a random sample of 80 doctors with a mean of 3.60 minutes and a standard deviation of 1.80 minutes. Is the government bluffing or is there a good chance that this is indeed a freak sample? Perform the appropriate hypothesis test using a significance level of 0.05.

3. For your PhD, you want to estimate the number of times an American phrase or pronunciation occurs in a typical 5 minute conversation between teenage youths in Liverpool. Because of the time taken to build up sufficient rapport with such youths for them to speak in a relaxed way, you only manage to observe 23 such conversations. The average number of Americanisms in 5 minute conversations amongst your small sample is 137.71, with a standard deviation of 69.56. The last study that was done demonstrated that the average was 128.2 words per 5 minute conversation and this has become the accepted wisdom in the literature. Do a hypothesis test to establish whether the average has in fact increased at the 5% significance level. Also do a two-tail test for whether there has been any change at all. Compare your answer with the 95% confidence interval estimated in the previous set of lab exercises.

4. As part of your research you seek to analyse the policy of using the planning system to encourage the building of residential properties on brownfield sites (i.e. recycled industrial land), rather than greenfield sites (i.e. former agricultural or park land). One of your concerns is that brownfield land is more likely to be contaminated, and the methods for surveying the level of contamination do not preclude the possibility that a site may be declared safe when it is not. Surveys gauge contamination by taking bore-extracts every 100m. The land is declared safe for residential construction if the average level of toxicity is no more than 1g per extract. In your case study area, the former steelworks site in Cambuslang, Glasgow, you find that a random sample of 64 bores has been taken yielding an average of 0.88g with standard deviation of 0.79g. The average is below the safety threshold but do you think the Local Authority should grant this site residential planning permission?

5. As part of your research into the contamination levels experienced by workers in the Cambuslang steel industry in the first half of the twentieth century, you examine 273 random medical checks on workers which reveal an average contamination level of 92.7 units with a standard deviation of 39.7 units. The legal threshold for exposure is that the average should not exceed 94 units and so the steel industry always claimed on the basis of this sample that their workers were safe. How sure can you be that this conclusion is valid?

6. You are a research assistant on a project investigating sectarian attitudes amongst Rangers supporters, you interview 24 supporters in pubs after a football match at Ibrox. You find that the average age of supporters with sectarian attitudes is 22.7 years (s.d. = 9.2 years) which is well below the most recent estimate published five years ago which said that the average age of this group was around 30 years. Run a lower-tail t-test to see if this difference between estimates is statistically significant and reflect on the robustness of your method.

*5.3 Tutorial 3: Hypothesis Testing Part I**

* See also T3 Info Sheet

1. **The Basics – *in your own words explain…***

- **What is a *statistical* hypothesis?**
- **What is the end product of a hypothesis test?**
- **What do we mean by *statistically* significant?**
- **What does it mean when we say that:**

$$\alpha = P = 0.10$$

- **Independent samples can have either pooled or different (i.e. equal or unequal) variance. What do we mean by 'independent samples'? Can you think of an example of matched samples?**
- **What is an F test? What is Levene's test? What does the null hypothesis in each test mean? When would you use each?**

Setting Up Hypotheses: One and Two-tailed Tests

It might help to quickly sketch a z distribution curve and shade off which region(s) of the curve would lead to a rejection of our hypothesis.

- **Decide whether to carry out an upper, lower or two-tailed test?**
- **Explain your choice.**
- **How would you write out the null and alternative hypotheses?**

a) The manufacturer claims that seven out of ten cats prefer new Mouse-flavoured cat food to their usual brand. The advertising authority wants to use a hypothesis test to statistically verify that these claims are accurate.

H_0: μ=0.7

H_1: μ ____.

b) An ethnographic study caused alarm when it suggested that the 15 students studying statistics were drinking far more heavily than their 26 colleagues on a faculty night out. Design research hypotheses to test whether the alcohol consumption of the statistics group is really different from the other group in a follow-up statistical investigation.

H_0: ______________.

H_1: ______________.

2. Working Through the Four Steps

See T3 Info Sheet for the four steps of hypothesis testing

Note: when testing for equal/ unequal variance is required, although this is a 'step 2' consideration, it is simplest to carry out the equality of variance test first!

- **Let's review a couple of examples.**

Example 1

Using Hand Calculation

For your PhD, you want to estimate the number of times an American phrase or pronunciation occurs in a typical 5 minute conversation between teenage youths in Liverpool. Because of the time taken to build up sufficient rapport with such youths for them to speak in a relaxed way, you only manage to observe 23 such conversations. The average number of Americanisms in 5 minute conversations amongst your small sample is 137.71, with a standard deviation of 69.56. The last study that was done demonstrated that the average was 128.2 words per 5 minute conversation and this has become the accepted wisdom in the literature. Do a hypothesis test to establish whether the average has in fact increased at the 5% significance level. Also do a two-tail test for whether there has been any change at all. Compare your answer with the 95% confidence interval estimated in the previous set of lab exercises.

STEP 1: Specify the Null (H_o) and Alternative (H_1) hypotheses

Start with the assumption about the population average:

$H_0 : \mu =$ ________.

$H_1 : \mu >$ ________.

STEP 2: Specify the threshold level of significance (α) and the test statistic

- **Firstly, state the specified significance level:**

α = P (probability of wrongly rejecting H_o [5]) =________.

Then select the appropriate test statistic based on:

a) What kind of variable(s) are we working with?
b) Do we have one or two samples?
c) Between two samples: is there pooled or different variance?
d) Do we have large or small samples?

Here you can work with hand calculations or with syntax…or both to show off!

[5] Rejecting the null hypothesis of equal variance when it is in fact true is referred to as a "type 1 error"

- **Hand calculation - why *this* formula?**

$$t_i = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}; df = n - 1$$

Note: for this test statistic we also need to calculate the degrees of freedom figure

STEP 3: Specify the decision rule

- **State at what level of probability (P) we will reject our null hypothesis**

i.e. Reject H_o (the hypothesis that the variances are equal) if (and only if) P (the calculated level of type 1 error) is not greater than the tolerated level of significance specified.

Reject H_o iff[6] $P \le \alpha$

so Reject H_o iff P ________.

STEP 4: Compute P and state your conclusion

- **Go back to the Step 2 for the formula you selected and check you can fill in the appropriate values so that you could carry out the calculation:**

$$t_i = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}; df = n - 1$$

- **What do each of these symbols mean and what figures would you fill in to the equation for:**

$\bar{x}$ =

μ =

n =

s =

- **How many degrees of freedom are we working with?**

df = n -1

df = ____-1

df = _____.

- **State your conclusion**

Our calculations will leave us with two figures: a result of 0.6556 for the t statistic formula and a result for degrees of freedom. Using a t table, we can look up the probability associated with a t value of 0.6556 and the relevant degrees of freedom. The closest value on the table is 0.686 (there will rarely be an *exact* match; just take the probability figure which is the best approximation of your t statistic).

Thus we can say that:

[6] 6 "iff" is shorthand for "if and *only* if"

P = Prob ($t > 0.6556$) $\approx$ 0.25 See note[7] below

- **Check the decision rule in step 3: based on this hypothesis test result can we reject our null hypothesis and accept the alternative? Why?**

Example 1 Cont:

Using SPSS Syntax

As with confidence intervals, both SPSS syntax and macros can be used to compute P in a hypothesis test.

- **For Example 1 (see above), which Macro would you choose? Write it out and fill in the appropriate values**

________ n=(   ) x_bar = (   ) m = (   ) s= (   ).

- **This yields the SPSS output below. What does it mean?**

```
    n          x_bar           SE          ti         SIGt_2TL   SIGt_LTL   SIGt_UTL
23.00000   0137.71000   14.50426   .65567         .51884       .74058       .25942
```

[7] This sign "$\approx$" means "approximately equal to…"

5.4 T3 Info Sheet: Summary Guide to Hypothesis Testing

1. Hypothesis Testing: the four step process

These four steps underpin every hypothesis test. However, as checking for equality of variance can be confusing, since this involves a hypothesis test within a hypothesis test, if you have two samples it can be simpler to carry out the Levene's test *before* going on to the main hypothesis test.

STEP 1: Specify the Null (H_o) and Alternative (H_1) hypotheses

a) What assumptions about the population parameter do we want to test:

i. Upper tail?

ii. Lower tail?

iii. Two-tails?

H_o : ….

H_1 : ….

STEP 2: Specify the threshold level of significance (α) and the test statistic

b) What kind of variable(s) are we working with?

c) Do we have one or two samples?

d) Between two samples: is there equal or unequal (pooled or different) variance?

e) Large or small samples?

α = ….

STEP 3: Specify the decision rule

- State at what level of probability (P) we will reject our null hypothesis

Reject H_o iff[8] ….

STEP 4: Compute P and state your conclusion

- As with confidence intervals, both SPSS syntax and macros can be used to compute P

We can confidently accept/ reject the null hypothesis that …. In favour of the alternative that ……

[8] iff = "*if and only if*"

2. Hypothesis Testing Macros

To select the appropriate syntax think about:

a) The type of variable (continuous or discrete?)
b) How many samples are we dealing with?
c) For two samples – pooled or different variance?
d) Large or small sample size?

Test Statistic	Macro
Large sample significance test on 1 mean	H_L1M n=( ) x_bar = ( ) m = ( ) s = ( ) .
Small sample significance test on 1 mean	H_S1M n=( ) x_bar = ( ) m = ( ) s = ( ) .
Small independent samples significance test for the equality of 2 means (pooled variance)*	H_S2Mp n1=( ) n2 = () x_bar1 = ( ) x_bar2 = ( ) s1 = ( ) s2 =( ) .
Small independent samples significance test for the equality of 2 means (different variance)*[*]	H_S2Md n1=( ) n2 = () x_bar1 = ( ) x_bar2 = ( ) s1 = ( ) s2 =( ) .
Large sample significance test on 1 proportion	H_L1P n=( ) x = ( ) pi[9] = ( ).
Large sample significance test on 2 proportions	H_L2P n1=( ) n2 = ( ) x1 = ( ) x2 = ( ) .
Simple small sample F test on equality of variances (this can be used if **only** summary statistics are available; where there is original data, Levene's test provides a more accurate alternative to test for homogenous variances)	H_S2VF n1=( ) n2 = () s1 = ( ) s2 =( ) .

3. Levene's Test

This can be carried out automatically by SPSS using the commands:

Analyze/ Compare Means/ Independent Samples T Test

As always, you can employ the **Paste** function to see and then run the syntax equivalent of the automated commands.

NB: using SPSS automated functions ***always*** gives a two-tailed significance level for the t-tests! Using the macros provided will give you the upper and lower tailed significance levels.

[*] When you are working with two samples – to choose the correct test statistic you must first CALCULATE (do not guess!) whether the samples have Equal or Unequal Variances.

[9] This stands for pi (π) the population (as opposed to the sample) proportion

5.5 Reading 5

5.5.1 Pryce I&S in SPSS

- *Pryce, Chapter 5

5.5.2 M&M 4th Ed.

- M&M section 6.2 and exercises for 6.2

L6 Hypothesis Tests for All Occasions

6.1 Structure of Lecture 6

1. **Review of Significance**
2. **Review of 1 sample test on the mean**
3. **Hypothesis test on 2 means**
    - Pooled variance
    - Different variance
4. **Deciding on whether variances are equal**
5. **Hypothesis tests about proportions**
    - One population
    - Two populations

6.2 Lab 6a and 6b

6.2.1 Example 6.3.2 Pooled Variance Independent Samples t-Test (Pryce, p.6-3)

Suppose you are interested in testing the view put forward by some feminist sociologists, that in the porn movie-making world, women play a controlling rather than subordinate role. This claim is apparently substantiated by the higher salaries of female porn stars as compared to their male counterparts. You aim to test this hypothesis by sending an anonymous postal survey to a sample of male and female porn stars. Your 39 male respondents report a mean income of £156,452 per annum (s.d. = £46,198) whereas your 42 female respondents report a mean income of £179,231 per annum (s.d. = £46,125). What can you conclude from your survey results?

6.2.2 Example 6.4 Testing for Equality of Means using Original Data (Pryce, p.6-7)

Use the **auction.sav** data to compare the mean purchase price of auction sales in Cumberland and Durham.

6.2.3 Exercise 6.4 Testing for Equality of Means (Pryce, p.6-9)

As part of your PhD research, you want to test whether the new "Fun Phonics" reading method is better than the "Letterland" method. You examine the reading ability of six year old children from two similar schools. The first used the FP method and you found that this produced an average reading proficiency score of 53.7 (based on a sample of 22 children; s.d. = 11.5). The second school used the Letterland method and you found that this produced an average reading proficiency score of 42.51 (sample = 24; s.d. = 16.9). Test whether the FP method produces higher results at the 1% significance level.

6.2.4 Example 6.6 Large Sample Hypothesis Tests on One Proportion (Pryce, p.6-13)

As a historian, you want to find the proportion of citizens in medieval Scotland that contracted the plague. From a sample of 400 parish records, you find that 22 died of the plague. Until your study was done, the assumption in the literature was that 10% of the population had died. Test whether this assumption was valid using a 2% tolerance level for Type I errors and use both one- and two-tail tests.

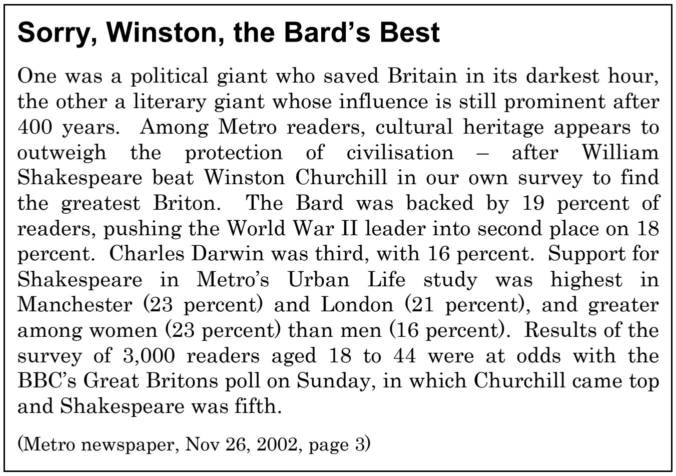
6.2.5 Exercise 6.6.1a, b and c Hypothesis Tests

Exercise 6.6.1a Hypothesis Tests:

In your dissertation sample of 59 nurses, 29 had performed a manual bowel evacuation. You want to test the hypothesis that, as part of the general broadening in the range of tasks performed by nurses, more nurses are performing manual bowel evacuations compared with 10 years ago when only 40% of nurses had performed the procedure. Also test whether there has been any change since 5 years ago when the proportion was 45%. Compare your answer with the 90% confidence interval.

Exercise 6.6.1b Hypothesis Tests:
For your PhD you interview 102 drug users in Glasgow and find that 70 had shoplifted to support their habit. Before the advent of “Big Issue” magazine, it was estimated that 78% of drug users in Glasgow shoplifted to support their habit. Test the hypothesis that the shoplifting rate is now lower amongst drug users. What would be the effect on your results if you had a smaller sample (e.g. 35, of whom 25 had shoplifted)? Compare your answer with the 95% confidence interval.

Exercise 6.6.1c Hypothesis Tests:
From the newspaper article below, compute a confidence interval for the proportion of the Metro survey who voted for The Bard. Also, conduct a hypothesis test at the 5% level of significance that The Bard won more than 18% of the vote. Do a second hypothesis test that he gained more than 16% of the vote.



Sorry, Winston, the Bard’s Best

One was a political giant who saved Britain in its darkest hour, the other a literary giant whose influence is still prominent after 400 years. Among Metro readers, cultural heritage appears to outweigh the protection of civilisation – after William Shakespeare beat Winston Churchill in our own survey to find the greatest Briton. The Bard was backed by 19 percent of readers, pushing the World War II leader into second place on 18 percent. Charles Darwin was third, with 16 percent. Support for Shakespeare in Metro’s Urban Life study was highest in Manchester (23 percent) and London (21 percent), and greater among women (23 percent) than men (16 percent). Results of the survey of 3,000 readers aged 18 to 44 were at odds with the BBC’s Great Britons poll on Sunday, in which Churchill came top and Shakespeare was fifth.

(Metro newspaper, Nov 26, 2002, page 3)

6.2.6 Example 6.7 Hypothesis Tests on Two Proportions (Pryce, p.6-17)

Suppose 1,630 out of a random sample of 7,180 black males have experienced stop-and-search procedures from the police in London, compared with 1,684 out of a sample of 9,916 white males. Test whether the difference in the proportions was due to sampling variation or due to a genuine difference in the incidence of stop-and-search across the two populations.

6.2.7 Exercise 6.7 Hypothesis Tests on Two Proportions (Pryce, p.6-18)

1. Re-do the Metro survey question from the previous set of exercises using the two-proportions t-test. How do your results compare with your previous answer?

2. Two surveys of mortgage payment protection insurance (MPPI) are carried out, one on single parents with 1 child and one on single parents with 3 children. Amongst the first group, 67 out of a sample of 300 were found to have taken out MPPI, compared with 15 out of a sample of 101 in the second group. Is take-up significantly lower amongst the households with three children?

Example 2

- ***Look over the four steps of hypothesis testing. For the next example, there are two samples involved so we will need to be able to specify whether they have equal or unequal variance. Before we proceed to the main question, we will look over the principles of Levene's test for equality of variance. This is itself a hypothesis test with four steps.***

FIRST: Levene's Test for Equality of Variance

Levene's Test STEP 1: Specify the Null (H_o) and Alternative (H_1) hypotheses

For Levene's test, out hypothesis is about equality of variance so:

$H_0 = \sigma_1 = \sigma_2$

$H_1 = \sigma_1 \neq \sigma_2$

- **What does this mean?**

Levene's Test STEP 2: Specify the threshold level of significance (α) and the test statistic

- **Firstly, state the specified significance level:**

α = P (probability of type 1 error) = **0.05**

- **The test statistic will be Levene's test.**

The hand calculation for this is laborious but we can carry out the calculation using an option programmed in SPSS. See sheet 2b

Levene's Test STEP 3: Specify the decision rule

- **State at what level of probability (P) we will reject our null hypothesis.**

Reject H_0 iff $P \leq \alpha$

Reject H_0 iff $P \leq 0.05$

__Levene’s Test STEP 4:__ **Compute P and state your conclusion**

- **The SPSS computation for Levene’s test in Example 2 yields the output on the next page (p 8). What does it mean? Do we accept or reject the null hypothesis?**

		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 Difference	
									Lower	Upper
Actual final purchase price	Equal variances assumed	**4.757**	**.032**	1.068	98	.288	518.24357	485.20255	-444.62509	1481.11224
	Equal variances not assumed			.967	53.366	.338	518.24357	535.66702	-555.99654	1592.48369

- **Having made this calculation, we can now return to Example 2 and begin the hypothesis test process.**

Example 2 -- *Using Hand Calculation*

A par t of your research into the housing market you develop a theory that house prices in Durham are more expensive than house prices in Cumberland. Data you have collected from a sample of 43 houses in Durham show that the mean purchase price of a house in Durham is 1165.3488, with a standard deviation of 3297.2699. A sample of 57 houses in Cumberland shows that the mean price of a house in Cumberland is 647.1053, with a standard deviation 1394.2116. Test the hypothesis that house purchase prices in Durham are higher than house purchase prices in Durham at the 5% significance level.

STEP 1: Specify the Null (H_o) and Alternative (H_1) hypotheses

- **Here our hypotheses are about the population means. How would you write then out?**

$H_0:\ \mu_{DURHAM} =$ ______________.

$H_1:$ ______________________.

STEP 2: Specify the threshold level of significance (α) and the test statistic

- **Firstly, state the specified significance level:**

α = P __________________________.

- **Why are we using the test statistic below?**

$$T_C = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2}} \quad ; \; df = \min\,[\, n_1 - 1, n_2 - 1\,]$$

where s_p , the pooled standard deviation is calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

STEP 3: Specify the decision rule

- **State at what level of probability (P) we will reject our null hypothesis.**

Reject H_0 iff

- **What other probability level might we use if we wanted a more stringent criterion?**

STEP 4: **Compute P and state your conclusion**

- **Returning to out test statistic:**

$$T_C = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2} - 2}} \quad ; \; df = \min\,[\,n_1 - 1, n_2 - 1\,]$$

where s_p , the pooled standard deviation is calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- **What values would you use for:**

$n_1 =$
$n_2 =$

$\bar{x}_1 =$

$\bar{x}_2 =$

$(\mu_1 - \mu_2) =$

This yields the result t_c =0.9675

- **What is our figure for degrees of freedom?**

$df = \min\,[\,n_1 - 1, n_2 - 1\,]$
df = min [_____ -1, _____-1]
df = min [____ , ____]
df = _________. [i.e. the lower of the two values]

So, finally:

- **What do we do with t statistic and the degrees of freedom figures we have calculated?**

- **In this case, we do not have an exact probability as the t value we calculated lies between 0.15 (t=0.851) and 0.20 (t=1.050). However, we can then say that P lies between 0.15 and 0.20. Do we accept or reject the null hypothesis? Why?**

Example 2 -- *Using SPSS Syntax*

- **For Example 2 (on page 9), which Macro would you choose? Write it out and fill in the appropriate values**

________ n1=(   ) n2 = (   ) x_bar1 = (   ) x_bar2 = (   ) s1 = (   )   s2 =(   ) .

- **This yields the output below. What does it mean?**

```
    df         x1b_x2b        SE                  ti             SIGt_2TL    SIGt_LTL   SIGt_UTL
2.000000    518.243500   535.667016      .967473    .338846       .830577         .169423
```

6.4 Reading 6

6.4.1 Pryce I&S in SPSS

- *Pryce, Chapter 6

6.4.2 M&M 4th Ed.

- M&M section 6.2 and exercises for 6.2
- Chapters 7 & 8
- Optional: sections 6.3 and 6.4

L7 Relationships between Categorical Variables

7.1 Structure of Lecture 7

1. Independent Events

2. Contingent Events

3. Chi-Square

4. Further Study

7.2 Labs 7a and 7b

7.2.1 Example 2.3a The Two Sided Die (Pryce, p.2-7)

7.2.2 Example 2.3.5 Contingent Probability of Going to University (Pryce, p.2-9)

7.2.3 Example 2.3c Relationship Between Class and Higher Education (Pryce, p.2-10)

7.2.4 Exercise 2.1 Understanding Randomness and Probability (Pryce, p. 2-11)

7.2.5 Example 7.2.1 Relationship Between Social Class and Voting Preference (Pryce, p.7-2)

Suppose you have collected data on voting preferences and social class. Your data were collected using a simple telephone questionnaire which achieved 556 responses. Your survey consisted of two simple questions: (1) Do you consider yourself to be working class? and (2) Did you vote New Labour at the last general election?

Everyone who responded answered either yes or no to both of these questions. You entered your data into an SPSS file and saved it as **votes.sav**. How would you establish whether there was a relationship between whether someone described themselves as working class and whether they voted New Labour?

7.2.6 Example 7.2.3a Relationship Between Class and HE (Pryce, p.7-5)

Example 7.2.3a Relationship Between Class and Higher Education
Suppose you are interested in whether there is a relationship between social class and access to higher education. You have 300 observations in your data distributed across the categories of each variable as demonstrated in the following table. Test whether there is a relationship between class-background and higher education.

		A	B
		Working Class	Middle Class
	Go to university	18	84
	Do not go to university	162	36

7.2.7 Example 7.2.3b Relationship Between First Time Buyer and Location (Pryce, p.7-8)

7.3 Reading 7

7.3.1 Pryce I&S in SPSS

- *Pryce, Sections 2.1, 2.2, 2.3; 7.1, 7.2

7.3.2 M&M 4th Ed.

- Chapter 9.

L8 Regression

8.1 Structure of Lecture 8

1. **Linear & Non-Linear Relationships**
2. **Fitting a line using OLS**
3. **Inference in Regression**
4. **Ommitted Variables & R2**
5. **Types of Regression Analysis**
6. **Properties of OLS**
7. **Assumptions of OLS**
8. **Doing Regression in SPSS**

8.2 Labs 8a and 8b

8.2.1 Example 1.4.1 How to Create a Scatter Plot in SPSS (Pryce, p. 1-25)

8.2.2 Example 7.3.2 Building an Automatic Valuation Model (Pryce, p.7-12)

Suppose you have been commissioned by Her Majesty's Valuation Office to construct an automatic valuation model (AVM) of residential dwellings. The model will be used to help with the forthcoming Council Tax revaluation when a reliable and up-to-date valuation will be needed of every residential property in England and Wales. Sending out Chartered Surveyors to each and every property was deemed infeasible, and so a cheaper computer-based system is being sought. You have been given a sample of house price data – **avmdata.sav** – with which to build your model. You start off by exploring the relationship between purchase price and the floor area of a house. You reason that the larger the floor area, the more valuable the property, which suggests the following relationship (assumed to be linear) between house price and floor area:

House price = $\alpha + \beta$ Floor area + ε

If we run this regression in SPSS using the data provided,

```
GET  FILE='Q:\QUANTS\avmdata.sav'.
REGRESSION  /DEPENDENT purchase  /METHOD=ENTER floorare  .
```

we would get the following Model Summary output:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.721[a]	.519	.519	26925.177

a. Predictors: (Constant), Floor Area (sq meters)

The Model Summary table includes the R Square, which tells you the proportion of the dependent variable that your independent variables explain (in this case = 0.519 = 51.9%). If you have more than one explanatory variable you should use the Adjusted R Square which controls for the fact that the R-square will rise each time you add an additional variable, even if there is no real gain in the explanatory power of the model.

The second table produced by the regression procedure is the Analysis of Variance (ANOVA) table:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.3E+11	1	4.341E+11	598.724	.000[a]
	Residual	4.0E+11	554	724965158.6		
	Total	8.4E+11	555			

a. Predictors: (Constant), Floor Area (sq meters)

b. Dependent Variable: Purchase Price

The most useful information in the ANOVA table at this stage is the F-statistic. This tests the null hypothesis H_0 that all slope coefficients are jointly equal to zero. It is another summary test of the whole model and is related to the R Square measure. If "Sig" is small, you can confidently reject the null. Also useful is the total degrees of freedom (df) which tells you at a glance how many observations were included in the model (in this case 555+1).

The third and most important table in the output is the Coefficients table:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-10029.0	3242.545		-3.093	.002
	Floor Area (sq meters)	638.825	26.108	.721	24.469	.000

a. Dependent Variable: Purchase Price

This table tells you the estimated value of the intercept – the point where the y axis is crossed (labelled Constant in SPSS output) – and of the slope coefficient – how much the value of *y* increases for each increase in the value of *x*. In this case we only have one independent variable and so only one slope coefficient is produced (for Floor Area). The estimated values of the intercept and slope coefficient are listed under the column headed B in the table. The t statistic tests the hypothesis that B = zero and is calculated by dividing B by the Standard Error of B, which is listed in column three. If Sig. is small – less than 0.05 say – then you can confidently reject the null hypothesis that B = 0.

While the Constant term is occasionally of interest, it is usually the slope coefficient(s) that we are most interested in. In the above example, the coefficient on Floor Area equals 638.8, which tells us that as the floor area of a house rises by one square metre, the value of the house rises by £638.8. So the slope coefficient is measured and interpreted in terms of the scale used for the dependent variable (which in this case is pounds sterling). Because the t value is large for the slope coefficient (resulting in a small significance level = 0.000), we can reject the null hypothesis that the slope coefficient = 0 (i.e. that there is no relationship between house price and floor area). Similarly, the t value is large for the intercept term, and so we can reject the null hypothesis that the slope coefficient is zero.

8.2.3 Exercise 7.3.2 Regression Analysis (Pryce, p. 7-13)

Using data from **avmdata.sav**, do a scatter plot of the relationship between purchase price and floor area. Comment on the plot and then insert a linear line of best-fit. How well do you think the regression line fits the data? Use the REGRESSION /DEPENDENT purchase /METHOD=ENTER floorare. syntax to run a linear regression to obtain the numerical values of the relationship. What does the statistical output tell you about the relationship?
Run a 3-D Scatter plot with floor spikes using the following menu sequence: Graphs, Scatter, Simple, and Define. Place purchase price on the Y axis, floorarea on the X axis and number of bathrooms on the Z axis, then click OK or Paste. Alternatively you can use the following syntax: GRAPH /SCATTERPLOT(XYZ)=floorare WITH purchase WITH bathroom.

To include the floor spikes, double click on the graph, then right-click on the data points in the body of the graph, select Properties Window, select Spikes, Floor, Apply and Close.

Now try viewing the graph from different angles by using the 3-D rotation facility. To rotate the graph in, right-click on the data points in the graph, and select Properties Window, and 3-D Rotation. Try changing the horizontal view (move the slider and click Apply). Comment on the graph and run a regression equation with the second explanatory variable included. Has the inclusion of the extra variable added anything to the explanatory power of the model? Try replacing it with number of bedrooms and comment on your results.
Experiment with a number of 2-explanatory variable models, comparing the 3-D scatter plot with the regression output. Then experiment with more than 2 explanatory variables.

8.2.4 Example 7.3.3 Applying the Rule of Thumb for CIs (Pryce, p.7-15)

Below is the output from a regression of floor area on number of bathrooms:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.928	4.700		8.708	.000
	Number of Bathrooms	64.622	3.819	.584	16.920	.000

a. Dependent Variable: Floor Area (sq meters)

We can calculate the confidence interval on the slope coefficient using our rule of thumb as follows:

$$\begin{aligned} \beta &= 64.6 \pm 2 \times 3.8 \\ &= 64.6 \pm 7.6 \end{aligned}$$

In other words, the 95% confidence interval for the population slope coefficient is (57, 72).

8.2.5 Exercise 7.3.4 Comparing SPSS CIs with those derived from the Rule of Thumb

1. Use the rule of thumb to calculate the 95% confidence interval for the slope and intercept terms in the regression of purchase price on floor area. Confirm your results by comparing them with the SPSS estimates of the confidence intervals.
2. Run a regression of purchase price on two explanatory variables: floor area and number of bathrooms. Comment on the meaning of the intercept and slope estimates and calculate the 95% confidence intervals.

Part 1: Multiple Regression – The Basics

1. In your own words…

- **What is a linear relationship?**

- **What do we mean by a statistical model?**

- **What does the slope of the regression line tell us?**

- **Why does it matter that the dependent variable is continuous in OLS?**

2. This is the regression equation for Ordinary Least Squares (OLS):

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- **What do each of these terms represent?**

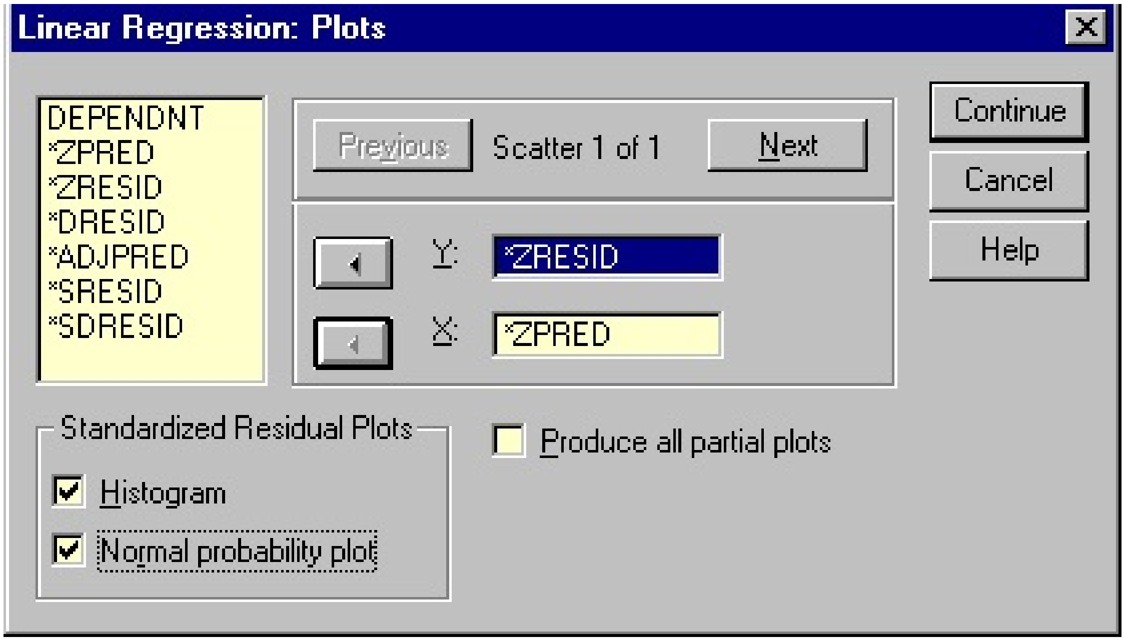
X =
Y =
β =
α =
ε =

3. Residuals

- **What do we mean by the residuals in a regression analysis? Why do we care about them?**

One way of examining the residuals is using plots generated by SPSS.

Once we have entered the dependent and independent variables we want to analyze into the linear regression dialogue box (**ANALYZE/ REGRESSION/ LINEAR**), click on the **PLOTS** dialogue box:



Y: *ZRESID

X: *ZPRED

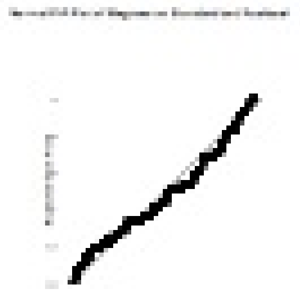
☑ Histogram

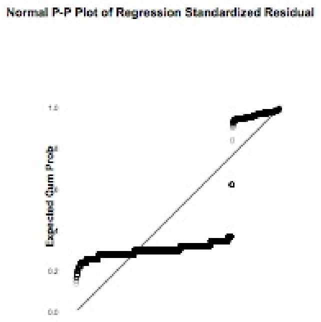
☑ Normal probability plot

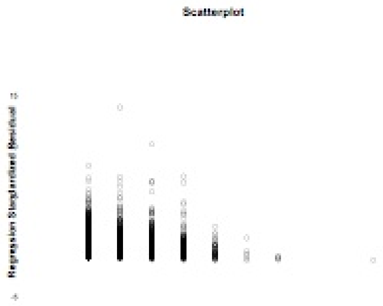
□ Produce all partial plots

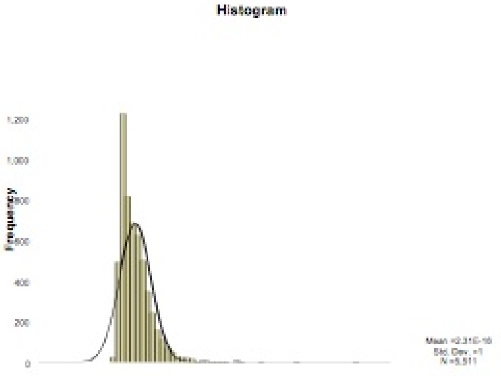
To plot the residual values against the predicted values from the model, enter ZRESID onto the Y axis and ZPRED onto the X axis and then check the boxes for the histogram and normal probability plots. When we analyze the residuals we are looking for a normal distribution, indicating linear relationship between the actual and predicted values.

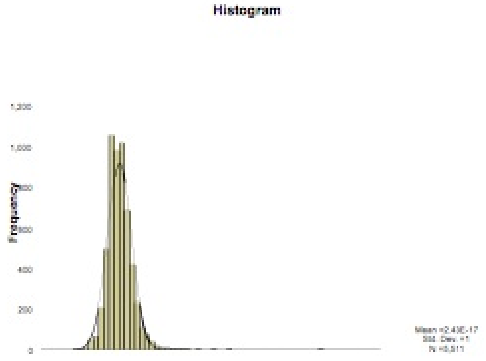
- **Comment on the following plots:**











Part 2: Worked Examples & Understanding SPSS Output

Example

Suppose you have been commissioned by Her Majesty's Valuation Office to construct an automatic valuation model (AVM) of residential dwellings. The model will be used to help with the forthcoming council tax revaluation when a reliable and up to date valuation will be needed of every residential property in England and Wales. Sending out chartered surveyors to each and every property was deemed unfeasible, and so a cheaper computer-based system is being sought. You have been given a sample of house price data – avmdata.sav – with which to build your model.

1. Your theory: first decide on the rationale behind how you treat each variable in the regression.

- **The variables in the dataset are below. Which variable will you select as the dependent variable?**
- **Which do you think might be useful as independent variables?**
- **Why – what is your hypothesis about the relationship between the dependent variable and each independent variable you have selected?**

Number of Bathrooms

Number of Bedrooms

Central Heating

Date Built

Floor Area (sq metres)

First Time Buyer

Garage

New Property

Parliamentary Constituency Code

Purchase Price

Type of Property

2. Preparing the dataset

- **Why are we concerned about missing variables in the dataset and how do we deal with them?**

- **Remember the type of variable!**

Some of the variables you identified as potentially relevant will be categorical. With the continuous (or even ranked) variables, the regression will tell us that *for every unit increase (or decrease) in an independent variable, the dependent variable will also increase (or decrease) by a specified amount*.

However, although SPSS will run the categorical variables in a regression, it treats them as though their values made numeric sense: but it makes no sense to say afterwards that, for example, *for every unit increase in the central heating variable, the dependent variable will increase (or decrease) by a particular amount.*

If we check the variable view of SPSS, we would see that the central heating variable is coded as:
0 = none
1 = full gas
2= full electric
3= full oil
4 = full solid
5 = part gas
6 = part electric
7 = part oil
8 = part solid

- **One way of dealing with categorical variables is using syntax to turn them into new 'dummy' (or 'design') variables. Explain the following syntax. What will happen when we run it?**

```
COMPUTE fullgas = 0.
IF (central = 2)  fullgas = 1.
IF (central = -9) fullgas = -9.
EXECUTE.
```

- **The current central heating variable contains nine categories but we would only use eight dummy variables. Why?**

- **Which other variables would need to be modified before we could run a valid regression?**

We can also use the **TRANSFORM/ COMPUTE VARIABLE** options in SPSS to manipulate numerical variables.

3. Carrying Out the Preliminary Regression

SPSS Commands:

ANALYZE/ REGRESSION/ LINEAR

After this, insert the dependent and independent variables into the appropriate boxes, select **PLOTS** and include the histogram and the normal probability plots for the residuals (see the section on Residuals above). As always, you can use the **PASTE** command in the linear regression dialogue box to copy the underlying syntax into a file and then run it from there.

4. Understanding SPSS Output

Your initial regression will generate four tables:
a) Variables Entered (Removed)
b) Model Summary
c) ANOVA
d) Coefficients

NB: Remember whenever you present tables in your work you must always explain them: it is not sufficient to copy a table and assume it speaks for itself!

a) Variables Entered (Removed)

This table lists all the variables which have been entered into the regression and the method of analysis. Check this to make sure that it contains everything you want but you are very unlikely to include this in your work because you will *already* have explained which variables you are using and why.

b) Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.835(a)	.698	.686	21748.051

a Predictors: (Constant), Parking Space, Part Oil Central Heating, Part Electric Central Heating, New Build Property, Full Electric Central Heating, Semi-detached Bungalow, Full Oil Central Heating, No Central Heating, Full Solid Central Heating, Converted Flat, Part Gas Central Heating, Semi-detached Bungalow, Single Garage, Terraced House, Floor Area (sq metres)

- **What does the model summary tell us?**

- **Would we report the R squared or the adjusted R squared figure? Why?**

So:

- **We can conclude from this model summary that _____% of the variation in the dependent variable, _______________ is explained by the other variables,**
  :

c) The ANOVA table

This table gives us the F value along with its associated significance figure (Sig). These figures are the results of a hypothesis test, testing the hypothesis that all the slope coefficients related to all the independent variables in the model are jointly equal to zero (i.e. if the slope coefficients were equal to zero, this would suggest that there was NO relationship between the dependent and the independent variables - think about a graph of the x and y axis if you are having trouble with this). If the significance value of the F statistic is less than 0.05, we reject the null hypothesis that the slope coefficients related to all the independent variables in the model are jointly equal to zero (and therefore have no relationship with the dependent variable) and accept the alternative hypothesis, that all the slope coefficients are NOT jointly equal to zero (and therefore that there is a linear relationship between the dependent and independent variables).

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	583114800576.830	21	27767371456.040	58.708	.000(a)
	Residual	252570093684.370	534	472977703.529		
	Total	835684894261.200	555			

a Predictors: (Constant), Parking Space, Part Oil Central Heating, Part Electric Central Heating, New Build Property, Full Electric Central Heating, Semi-detached Bungalow, Full Oil Central Heating, No Central Heating, Full Solid Central Heating, Converted Flat, Part Gas Central Heating, Semi-detached Bungalow, Single Garage, Terraced House, Floor Area (sq metres)
b Dependent Variable: Purchase Price

- **What are the results of the hypothesis test shown in the ANOVA table? What does this tell us?**

d) Coefficients

Coefficients(a)

Model	Unstandardized Coefficients			Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	91577.787	46749.151		1.959	.051
	Number of Bathrooms	10468.620	3260.457	.107	3.211	.001
	Number of Bedrooms	-364.911	1795.190	-.008	-.203	.839
	Date Built	-44.738	23.739	-.061	-1.885	.060
	Floor Area (sq metres)	442.130	41.721	.499	10.597	.000
	No Central Heating	-8621.527	3451.720	-.062	-2.498	.013
	Full Electric Central Heating	3268.344	4912.007	.017	.665	.506
	Full Oil Central Heating	-7417.077	6474.508	-.028	-1.146	.252
	Full Solid Central Heating	-14426.625	5853.797	-.060	-2.464	.014
	Part Gas Central Heating	-3224.774	4384.097	-.018	-.736	.462
	Part Electric Central Heating	8269.021	21878.031	.009	.378	.706
	Part Oil Central Heating	-9562.089	9307.432	-.025	-1.027	.305
	New Build Property	13704.322	4509.021	.080	3.039	.002
	Semi-detached	-15282.313	2934.357	-.183	-5.208	.000
	Terrace House	-23297.369	3234.803	-.287	-7.202	.000
	Detached Bungalow	8636.113	4523.452	.053	1.909	.057
	Semi-detached Bungalow	-8420.246	6370.512	-.035	-1.322	.187
	Purpose Built Flat	-6837.623	6194.811	-.031	-1.104	.270
	Converted Flat	-17510.682	9070.671	-.050	-1.930	.054
	Single Garage	13088.673	2776.911	.168	4.713	.000
	Double Garage	25639.372	4170.525	.208	6.148	.000
	Parking Space	10052.499	2936.540	.100	3.427	.001

a Dependent Variable: Purchase Price

- **How does this table relate to the regression equation?**

- **At this stage we can begin the process of refining the model by removing variables from the regression <u>ONE BY ONE</u> before re-running the regression again and making a new selection. What are our criteria for removing a variable? Which would you remove first? Why not just take a group out at a time?**

Here are the model summary and ANOVA tables from the final model:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.832(a)	.692	.685	21762.936

a Predictors: (Constant), Parking Space, No Central Heating, Semi-detached Bungalow, New Build Property, Full Solid Central Heating, Detached Bungalow, Floor Area (sq metres), Double Garage, Terraced House, Number of Bathrooms, Single Garage

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	578032695113.398	11	52548426828.491	110.949	.000(a)
	Residual	257652199147.802	544	473635366.081		
	Total	835684894261.200	555			

a Predictors: (Constant), Parking Space, No Central Heating, Semi-detached Bungalow, New Build Property, Full Solid Central Heating, Detached Bungalow, Floor Area (sq metres), Double Garage, Terraced House, Number of Bathrooms, Single Garage
b Dependent Variable: Purchase Price

- **What do they tell us?**

Here is the coefficients table from the final model:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2943.375	4574.817		-.643	.520
	Number of Bathrooms	9761.531	3195.347	.099	3.055	.002
	Floor Area (sq metres)	478.217	28.139	.540	16.995	.000
	No Central Heating	-7264.780	3397.691	-.052	-2.138	.033
	Full Solid Central Heating	-15564.789	5804.013	-0.65	-2.682	.008
	New Build Property	12888.105	4378.742	.075	2.943	.003
	Semi-detached	-12047.682	2518.946	-.144	-4.783	.000
	Terrace House	-18835.472	2718.418	-.232	-6.929	.000
	Detached Bungalow	11562.748	4228.130	.070	2.735	.006
	Single Garage	12769.713	2672.174	.164	4.779	.000
	Double Garage	35328.240	4059.577	.206	6.239	.000
	Parking Space	9476.761	2908.527	.095	3.258	.001

a Dependent Variable: Purchase Price

- **What does it tell us?**

8.4 Reading 8

8.4.1 Pryce I&S in SPSS

- *Pryce, Sections 1.4, 1.5; 7.3, 7.4

8.4.2 Other reading

- Field, A. chapters on regression.
- M&M 4th Ed. Chapters 2, 10 and 11; see Chapter 15 for Logistic regression.
- Kennedy, P. ‘A Guide to Econometrics’
- Bryman, Alan, and Cramer, Duncan (1999) “Quantitative Data Analysis with SPSS for Windows: A Guide for Social Scientists”, Chapters 9 and 10.
- Achen, Christopher H. Interpreting and Using Regression (London: Sage, 1982).
- Pryce, G. Advanced Regression in SPSS