NHS
National Institute for Health Research

The NIHR Research Design Service for the East Midlands

The NIHR Research Design Service for Yorkshire & the Humber

Practical Statistics Using SPSS

Authors

Nicola Spiers

Brad Manktelow

Michael J. Hewitt

This Resource Pack is one of a series produced by The NIHR RDS for the East Midlands / The NIHR RDS for Yorkshire and the Humber. This series has been funded by The NIHR RDS EM / YH.

Reference as:
Spiers, N., Manktelow, B., Hewitt, M. J. The NIHR RDS EM / YH for Research and Development in Primary Health Care: Using SPSS. The NIHR RDS EM / YH, 2007

Last updated: 2009

Nicola Spiers
The NIHR RDS for the East Midlands / Yorkshire & the Humber
Department of Health Sciences
University of Leicester
22-28 Princess Road West
Leicester LE1 6TP

Brad Manktelow
Department of Health Sciences
University of Leicester
22-28 Princess Road West
Leicester LE1 6TP

Michael Hewitt
Evaluation, Audit & Research Dept
Sherwood Forest Hospitals NHS Trust
King's Mill Hospital
Sutton in Ashfield NG17 4JL

The NIHR RDS for the East Midlands
Division of Primary Care,
14th Floor, Tower building
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: 0115 823 0500

www.rds-eastmidlands.nihr.ac.uk
Leicester: enquiries-LNR@rds-eastmidlands.org.uk
Nottingham: enquiries-NDL@rds-eastmidlands.org.uk

The NIHR RDS for Yorkshire & the Humber
ScHARR
The University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA
Tel: 0114 222 0828

www.rds-yh.nihr.ac.uk
Sheffield: rds-yh@sheffield.ac.uk
Leeds: rds-yh@leeds.ac.uk
York: rds-yh@york.ac.uk

Table of Contents

1. Introduction

Who is this Resource Pack for?

This pack has been written for NHS researchers in the Trent region. It is aimed at those who have had some training in statistics and who wish to carry out their own straightforward analyses using SPSS. This pack is not a statistical textbook, although some statistical concepts are necessarily discussed along the way.

For those who are unfamiliar with statistical methods it is intended that the pack will be used in combination with material from other courses and statistical texts such as those listed on Page 51. The pack itself assumes some familiarity with the following statistical concepts: sample, population, distribution, mean, standard deviation, confidence interval, significance test, null hypothesis and p-value.

What are the aims of the pack?

After working through the pack the reader will be able to:

- Carry out a basic statistical analysis using SPSS and obtain correct computer output.
- Interpret the output and communicate the results effectively.

How should the pack be used?

The pack is written as a trainer, with examples to work through and exercises in each section. Suggested answers to the exercises are given at the back of the pack. Experience suggests that most benefit will be gained by using the pack as preparation for, and shortly followed by, analysis of your own data. Sections 2-7 are of most general relevance and you may wish to concentrate on these sections when first using the pack.

Data from a real study are described in Section 4 and used in Sections 5-8. The dataset can be downloaded from:

http://www.trentrdsu.org.uk/uploads/File/spsspackresourcepack.zip

What is the scope of the pack?

The pack covers standard techniques for statistical comparisons of two groups of patients and introduces regression and correlation.

Confidence interval or p-value?

The key question in most statistical comparisons is whether an observed difference between two groups of subjects in a sample is large enough to be evidence of a true difference in the population from which the sample was drawn. There are two standard methods of answering this question:
A **95% confidence interval** gives a plausible range of values that should contain the true population difference. On average, only 1 in 20 of such confidence intervals should fail to capture the true difference. If the 95% confidence interval includes the point of zero difference then, by convention, any difference in the sample cannot be generalised to the population.

A **p-value** is the probability of getting the observed difference, or one more extreme, in the sample purely by chance from a population where the true difference is zero. If the p-value is greater than 0.05 then, by convention, we conclude that the observed

difference could have occurred by chance and there is no statistically significant evidence (at the 5% level) for a difference between the groups in the population.

Confidence intervals and p-values are based upon the same theory and mathematics and will lead to the same conclusion about whether a population difference exists. Confidence intervals are preferable because they give information about the size of any difference in the population, and they also (very usefully) indicate the amount of uncertainty remaining about the size of the difference.

How to choose the appropriate analysis

The choice of an appropriate analysis for a given study is beyond the scope of this pack. Some guidance is given in the SPSS Statistics Coach and in The NIHR RDS EM / YH Resource Pack: '*Using Statistics in Research'*. The choice of the appropriate analysis ideally should be made before the data are collected. If you are in any doubt about the analysis to choose you are strongly recommended to take advice from a statistician, such as that offered by The NIHR RDS EM / YH.

A health warning…

We hope that this pack provides a useful introduction to data analysis with SPSS, using realistic examples for the purpose of learning. Working through the pack will give you experience of the process of analysing a real dataset in SPSS and drawing sound conclusions in context.

However, please remember that the analyses in Chapters 5-8 have been selected for teaching purposes. Our aim is to illustrate the standard elementary statistical analyses for two-samples in SPSS, not to reproduce the published analysis of the Hospital-at-Home trial. The analyses presented lead to valid conclusions about the data, but are not always the optimal analysis (should this exist!). In some cases we introduce more than one analysis, and discuss which is to be preferred. In other cases we indicate that there is an optimal analysis which is beyond the scope of this resource pack.

The application of even 'basic' statistical techniques to real life data requires judgement, careful attention to context and common sense. This pack can be no substitute for experience, or for interaction with statisticians and others practised in data analysis.

We hope that this pack will enthuse you to try out techniques on your own data, and discuss analyses with colleagues, and check out both ideas for analysis (at design stage) and results by consulting a statistician.

Acknowledgements

We would like to express our gratitude to Dr Andrew Wilson of the Department of General Practice and Primary Healthcare, University of Leicester, for permission to use the Hospital-at-Home data. Our thanks also to the patients who provided the information and to the members of the Hospital-at-Home team.

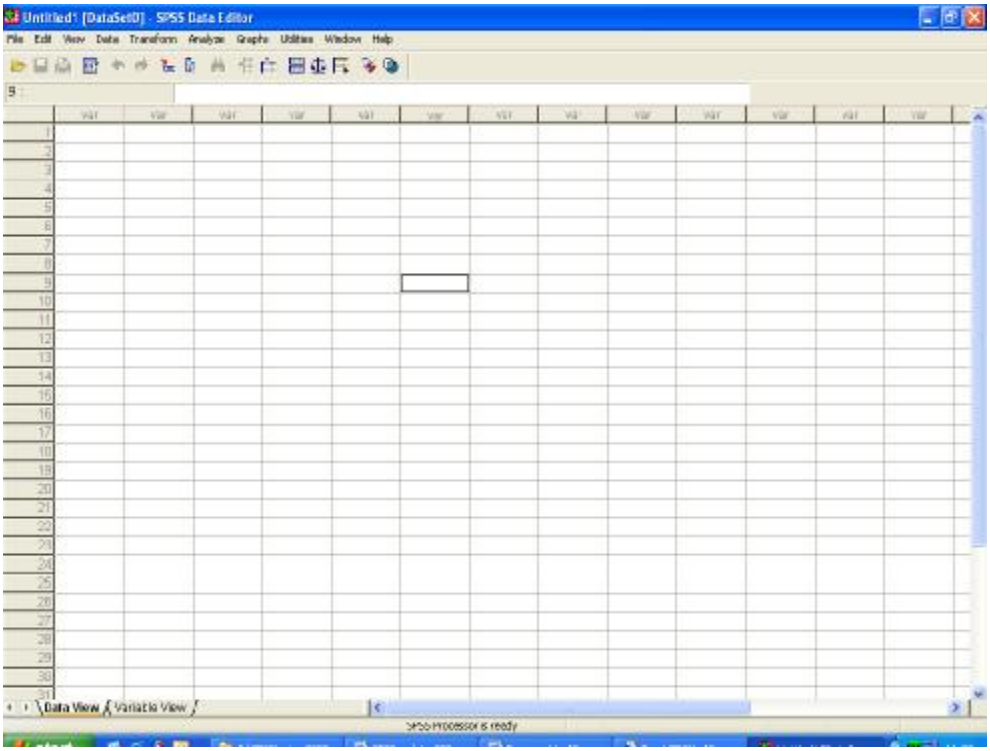
We are grateful to Victoria Owen, Nick Taub, John Bankart and Julia Chernova of The NIHR RDS EM / YH, for helpful comments on the 2007 update.

2. SPSS for Windows: An Introduction

SPSS for Windows is a software package for statistical data analysis. It was originally developed for the Social Sciences, hence the acronym SPSS (**S**tatistical **P**ackage for the **S**ocial **S**ciences), but it is now used in many areas of scientific study. This section serves as an introduction to the package and will show you how to enter and save data. It is written as a step-by-step, interactive exercise, which you can work through if you have access to SPSS for Windows. The data entered will be used in the next Section: Summarising and Presenting Data.

2.1 Starting SPSS for Windows

When the program is opened the **SPSS Data Editor** window will appear



The Menu Bar displays the names of the menus that are available to you. When you click on a menu name, e.g**. File**, a list of commands is displayed. The File menu provides options for opening a file, saving a file, printing etc. Other commonly used menus and options are: Edit (to cut, copy or paste data), Analyse (to analyse data e.g. descriptive statistics, correlation etc.), Graphs (to present data graphically e.g. bar chart, histogram etc.), Window (to move from the Output window to the Data Editor window), and Help. The options on the Menu Bar will depend on which SPSS window is active.

The **Data Editor window** is similar to a spreadsheet. The rows represent individual cases (observations) and the columns represent variables[1]. A single cell is an intersection of a case and a variable eg, the height of person x.

[1] A variable is basically anything that can assume different values that are observed and recorded in research e.g. height, weight, gender.

The **Output window** is where SPSS displays the statistics and reports from the analysis of your data.

2.2 Entering Data

When you start SPSS you are automatically placed in the Data Editor window. The active cell in the window has a heavy black border around it, indicating that any data you type will be entered into that cell. You can move around the Data Editor window to choose an active cell by using the arrow keys (←, ↑, →, ↓), or by clicking on cells with the mouse.

Table 1 presents some patient data that we can enter in to SPSS. For each patient we have collected the following data from their medical notes: gender, age and blood group. The data has already been coded[2] by a researcher.

Table 1: Patient data

Patient ID	Gender	Age	Blood group
1	1	21	2
2	2	39	1
3	2	43	1
4	1	55	1
5	1	26	4
6	1	19	4
7	2	65	2
8	2	41	2
9	1	61	3
10	1	50	1

The codes for Gender are:

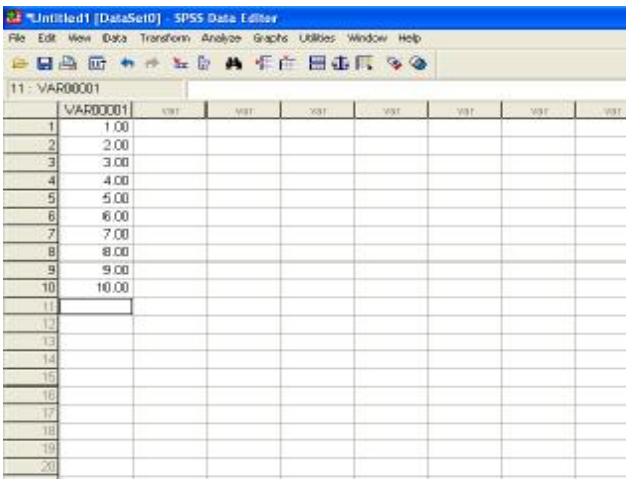
1 = male, 2 = female

and for Blood Group

1 = O, 2 = A, 3 = AB, 4 = B.

Notice the data are organised with individual patients or **cases** in rows. The data can be entered into SPSS as follows:
Starting in the top left cell of the Data Editor window, type in the first value shown under the column heading var0001, ie type 1 and press the **enter** key. This value is then placed in the cell and the black border moves down to the next cell. Type in the other 9 values, remembering to press the **enter** key after each one. Your Data Editor window should look like this

[2] Coding is the process whereby words or categories of a variable are changed into numbers so the computer can make sense of the data (e.g. the variable Gender comprises the categories Female and Male and has been coded so that Male = 1 and Female = 2).

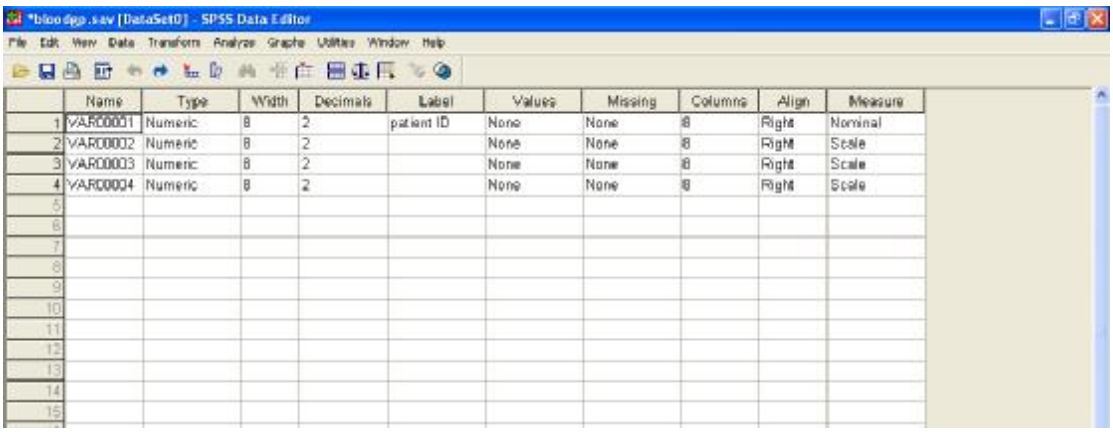


Now move the cursor to the top of the second column and type in the second column of data, gender, and continue in the same way with the third and fourth columns, Age and Blood Group. If you make any mistakes during data entry simply return to the incorrect cell using the arrow keys or mouse, then type over the entry and press the enter key. It is a good idea to recheck your data when entered, as it is easy to miss errors. A simple check is to reread data in a different order from that in which it was entered.

2.3 Defining the variables

At the moment the first column is labelled **var0001**. This can be relabelled so that we can give a name to this variable that more easily reminds us what it is. We can also say more about the type of variable we are dealing with (e.g. how many decimal places we want to show, how we want to record missing values) and define the coding scheme we have used. This process is known in SPSS as **defining the variable** and is described below.

Clicking on the tab labelled **Variable View** at the bottom of the **Data Editor** window brings up a sheet that shows you how each variable is defined and allows you to make changes. Double clicking on any of the column headers on the **Data View** sheet will also bring up this sheet. The **Variable View** sheet is shown below:



We will amend the information about this variable in several ways:

i. Enter a name for the variable

Variable names (column headings) should be unique and meaningful. Current versions of SPSS allow names longer than 8 characters, but this is a reasonable limit in many cases as short names make data easier to handle. Variable names should not contain spaces, hence the name **patient ID** would not be allowed, but Patient_ID is acceptable.

- Type the new name **patID** in the cell in the first row of the **Name** column

ii. Enter labels

To include more detail about a variable in the output, it is possible to give each variable a label. A label can be up to 256 characters long and include spaces. Labelling variables is usually good practice as it makes output easier to read.

To enter labels:

- Click on the appropriate cell in the **Label** column and type **patient ID**

Enter names for the other three variables, gender, age and blood group. Label the age variable 'Age in years'.

We can also enter the coding scheme that we have used for the variable gender. This is important for two reasons. First, it provides us with a permanent record of the scheme. Second it makes interpretation of later analysis much easier as the codes and their meanings are given in full.

- Click on the appropriate cell of the **Values** column for the variable gender, and then on the **…** symbol which appears
- In the **Value Labels** window that appears click on the **Value** box and type **1**
- Click in the **Value label** box and type **Male**
- Click on the **Add** button
- Click in the **Value** box and type **2**
- Click in the **Value label** box and type **Female**, then click on the **Add** button
- Click on **OK**

iii. Change the number of decimal places

By default a column of numeric data will be shown to 2 decimal places and display a maximum of 8 digits for each number. This definition is fine for numbers that require this precision e.g. height of patients in metres such as 1.65, 1.82. However, our data for the gender column requires less precision and can be accommodated in a column with no decimal places ie whole numbers. This is achieved by entering 0 in the decimals column for gender.

The second column, labelled **type**, shows that the variable **gender** is numeric. The column headed **Width** indicates the number of digits for each number. Eight is usually sufficient but can be increased, if required.

Variable Types

Before continuing, examine the range of data types that are available. Clicking on any cell in the **Type** column and then clicking on the **…** symbol will show them.

The Comma, Dot and Scientific Notation options provide for numeric data in different formats. The Date option allows calendar dates to be entered, and gives a choice of formats, e.g. 24-05-95. The Dollar and Custom currency options provide for currency data e.g. £24.99. The String option is often used to put in comments for each observation. For instance you may wish to record the doctor who saw each patient and enter 'Dr Smith' or 'Dr Jones' as text. However, as a rule keep string variables to a minimum, as they cannot be analysed in SPSS and are time-consuming to enter.

The default data format is numeric and most data are best entered in this format, using numeric codes. The use of other types may restrict the statistical techniques available, for example none of the statistical options at all work with data in string format.

2.4 Missing values

(You may wish to skip this section on first reading and continue with Exercise 1 on Page 13)

SPSS is good at handling missing data; there are essentially two options for doing so:

Option 1: System-defined missing values
The system-defined value for missing data in SPSS is '.'. Missing values imported from other software appear on the worksheet as '.', and you can input missing values by entering the full-stop (without quotes) manually.

These missing values are not included in the analysis, and SPSS gives a useful summary of missing values and numbers of patients included (the 'case processing summary') as part of the output for all analyses.

However, you may wish to include missing values in some analyses, and it is also good practice to record reasons for missing values if these are known. In this case option 2 is preferable:

Option 2: User-defined missing values

Here codes defined by the user are declared as missing values in SPSS. For example, in the published analysis of the Hospital-at-Home study, of older people in poor health, the set of missing codes was:

70	too ill to complete
80	too confused
90	refused
99	missing for other/unknown reason

These values were omitted from most analyses, by defining them as missing. To do this, click on the appropriate cell in the missing column in variable view, then click on the … symbol. You are given the options of defining single missing values or ranges. In the study, all non-missing data values were less than 50, so the range 70+ was defined as missing. Patient age was treated as a special case, and missing age was coded as 999. For categorical data, it is sometimes appropriate to include missing values in the analysis, and this can be easily achieved by altering the range of defined missing values.

Including system-defined missing values in the analysis

System-defined missing values can easily be included in any analysis by recoding them to a user-defined code (e.g. 999). However note that this must be done using the **transform** and **recode** menu commands – system-missing is given as an option in the 'old and new values' dialogue. The **transform** and **compute** commands will ignore all missing values.

Two common mistakes to check for with user-defined missing values

Check 1 Choose missing value codes that cannot become confused with data values. For example, the above missing value codes would be obviously inappropriate for blood pressure data.

Check 2 Ensure user-defined missing values are not inadvertently included in analyses. It is usually appropriate to include missing values in initial cross tabulations, so that the rate of missing values can be compared across patient groups. However it can be disastrous, and not always immediately apparent, if missing value codes creep into other analyses, for example patients with BP of 999, and so on. The best way to avoid this is to always check the number of patients included in the analysis, to make sure that you have analysed data from the correct group of patients.

In our example, for gender the data are only likely to be missing because they were not recorded. We will use the code **9** for these observations:

- Click on the **…** symbol in the appropriate cell of the **Missing** column
- Click on **Discrete Missing Values** and type **9** in the first box
- Click on **OK**

EXERCISE 1

Define the remaining variable, Blood group. Use **bloodgp** for the column heading, label the variable as '**Blood group of patient**' and label the values as shown in Table 2.

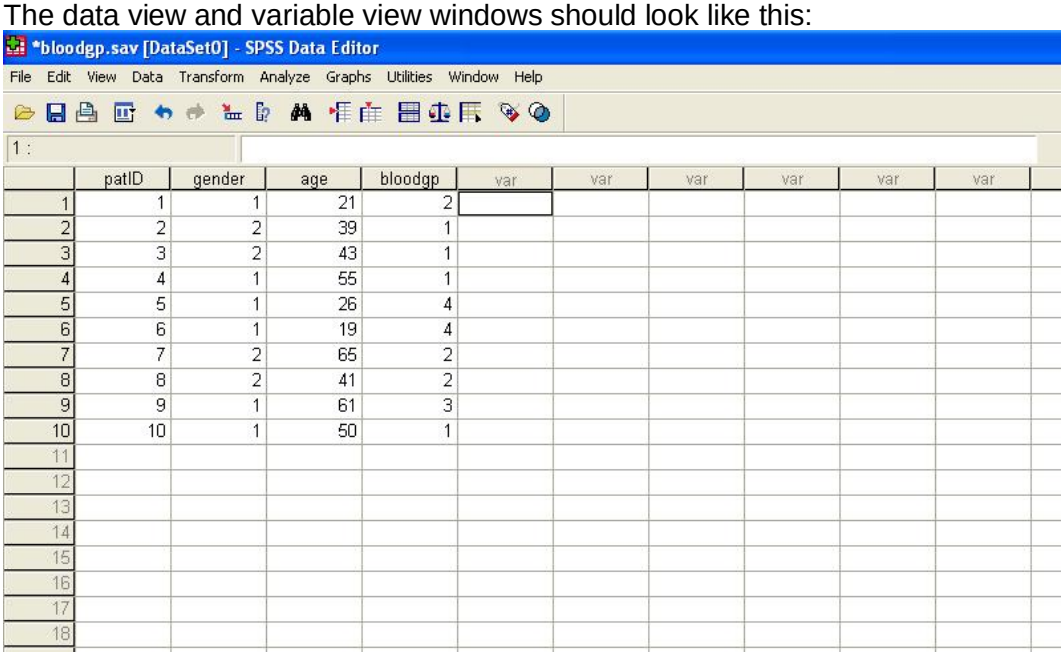
Table 2 Labels for the variable bloodgp

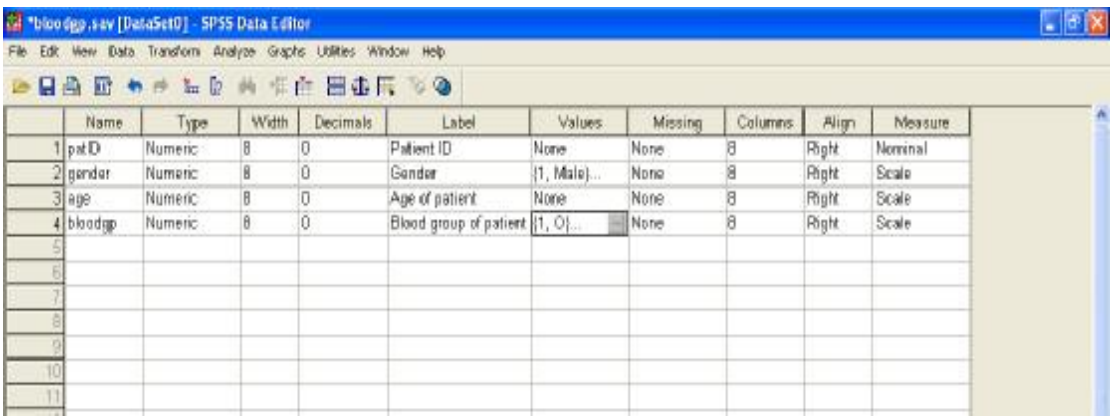
Value	Value label
1	O
2	A
3	AB
4	B

Once all of the variables are correctly defined, clicking on the Data View tab at the bottom of the window will return you to the data sheet.

2.5 Saving your data

The data view and variable view windows should look like this:





Now that we have entered the data and defined the properties of all of the variables it is a good time to save the data.

To save the data

- Click on the **File menu**
- Click on **Save As…**
- Once you have chosen the correct drive and directory, type **pat_data** in the **File name** box
- Click on **OK**

SPSS adds the extension **.sav** to the end of your filename, which helps in recognising the file as an **SPSS data file** for future sessions. The data are now ready to be investigated. Examples of analyses using these and other data will be shown in Sections 4 to 8.

2.6 Help

SPSS includes an extensive online help system. The **Help** menu provides different kinds of help, including **Topics**, **Tutorial** and even a **Statistics Coach**. The topics option allows you to search for help by keywords. The Statistics Coach may be helpful for choosing the appropriate analysis for a particular dataset. The Help system also includes an online version of the SPSS syntax guide, which is useful for more advanced users.

2.7 Exiting SPSS

At any time in your work you may want to exit SPSS. This is achieved either by closing all open data editor windows, or with the following menu command:

- Click on the **File** menu, then on **Exit**.

However, do not forget to save your data first if you want to keep the changes you have made.

Summary

This Section has introduced you to SPSS and has covered the first steps in using the package to analyse data. This has included entering data, defining variables, saving data and using the Help system. The rest of this Resource Pack builds on this Section and introduces you to basic analyses of data using SPSS.

3. Summarising and Presenting Data

The first step in any statistical analysis is summarising and presenting data. A simple summary statistic or graph will give you an immediate impression of the data you have collected. They also give a valuable check on any mistakes and any missing data. This exploration and cleaning of the data is an essential step before rushing into any statistical tests.

Before we consider some of the methods used for summarising and presenting data we need to consider the different types of quantitative data. These are: nominal, ordinal, and scale (which includes interval and ratio). Knowing the different types of data is important for statistical analysis. The type of data determines the type of statistic that is appropriate for analysis.

Nominal data: The observations are classified into categories that are different in character and cannot be measured or ordered. For example, gender, eye colour, blood group. Appropriate statistical techniques are restricted but do include frequency tables and cross-tabulations.

Ordinal data: The observations are classified into categories that can be ordered in an ascending series. For example, severity of an illness may be categorised as mild, moderate or severe. Appropriate statistics and techniques may include medians, interquartile ranges, percentiles, and distribution-free tests.

Scale data (interval and ratio): The observations are scores on a scale where the difference between scores is numerically equal. For example, height or weight. Many statistics and statistical techniques may be appropriate, including means, standard deviations, t-tests and related confidence intervals.

3.1 Summarising data

Summary statistics include measures of location (e.g. mean, mode and median) and measures of dispersion (e.g. range, interquartile range and standard deviation).

Measures of location

These give an idea of the average value on a particular variable.

The **mean** is the sum of all data points (observations or cases) divided by the number of data points; e.g. 7, 3, 11, 12, 9, 14, have a mean of $\frac{56}{6}$ = 9.3

The **median** is the middle value of a set of observations ranked in order (or the mean of the two middle numbers when there is an even number of observations).
For example, when the values 14, 9, 17, 21, 7, 18, 16, 22 are ranked in order (that is 7, 9, 14, 16, 17, 18, 21, 22) because there is an even number of values, 8, the median is the mean of the fourth and fifth scores: $\frac{16+17}{2}$ = 16.5.
By obtaining the median you know that half of the observations have values that are smaller than or equal to the median, and the other half have values larger than or equal to the median.

You can find values that split the sample in other ways. For example, you can find the value below which 25% of the data values fall. Such values are called percentiles, since they tell you the percentage of cases with values below them. For

example, 25% of cases will have values smaller than or equal to the 25th percentile and 75% of cases will have values larger than or equal to the 25th percentile. The median is known as the 50th percentile.

Measures of dispersion

These are several ways of describing the variability of data. These include the range, interquartile range and standard deviation.

The **range** is the difference between the largest and smallest observations. With the observations 19, 21, 22, 22, 25, 27, 28, 42 the range is [42-19] = 23. A large value for the range tells you that the largest and smallest values differ substantially. However the range may be highly influenced by a single very high, or very low, value. To overcome this problem a measure of variability that is often used is the **interquartile range** (IQR). This is the difference between the 75th and 25th percentiles and is usually given as a range of values. In the data above, the lower quartile is [$\frac{1}{2}$ (21+22) = 21.5], and the upper quartile is [$\frac{1}{2}$ (27+28) = 27.5], so that the IQR may be given as (21.5 to 27.5).

One of the most commonly used measures of variability is called the **standard deviation**. It indicates the extent to which the values deviate from the mean of the values. It is defined as the square root of the average squared difference from the mean.

3.2 Presenting data

Data presentation techniques include: bar charts, histograms, and frequency tables.

Frequency Table

A frequency table is a simple and effective way of presenting data. It is particularly suitable when observations fall into one of a number of different categories (nominal or ordinal data). For scale variables such as age or BP where there are a lot of data values, frequency tables are not suitable. An example for the categorical variable blood group is shown in Table 3.

The first column shows the possible categories that the observations could take and the second column shows the frequency with which each category occurs. This is known as a frequency distribution. The total is important for checking that all patients are included.

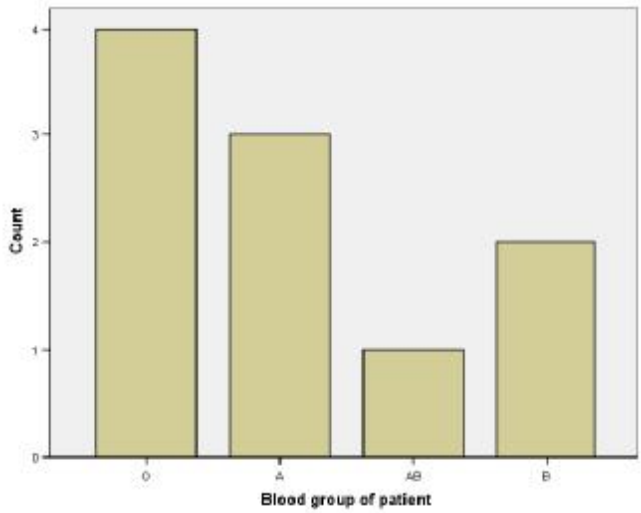
Table 3 **Frequency table of blood groups**

Group	No. of patients
O	56
A	31
B	18
AB	15
Total	**120**

Bar Chart

A bar chart displays the frequency count for each category of a frequency distribution. Figure 1 shows the number of patients in each of the blood groups.

Figure 1: Bar chart of blood groups



Histogram

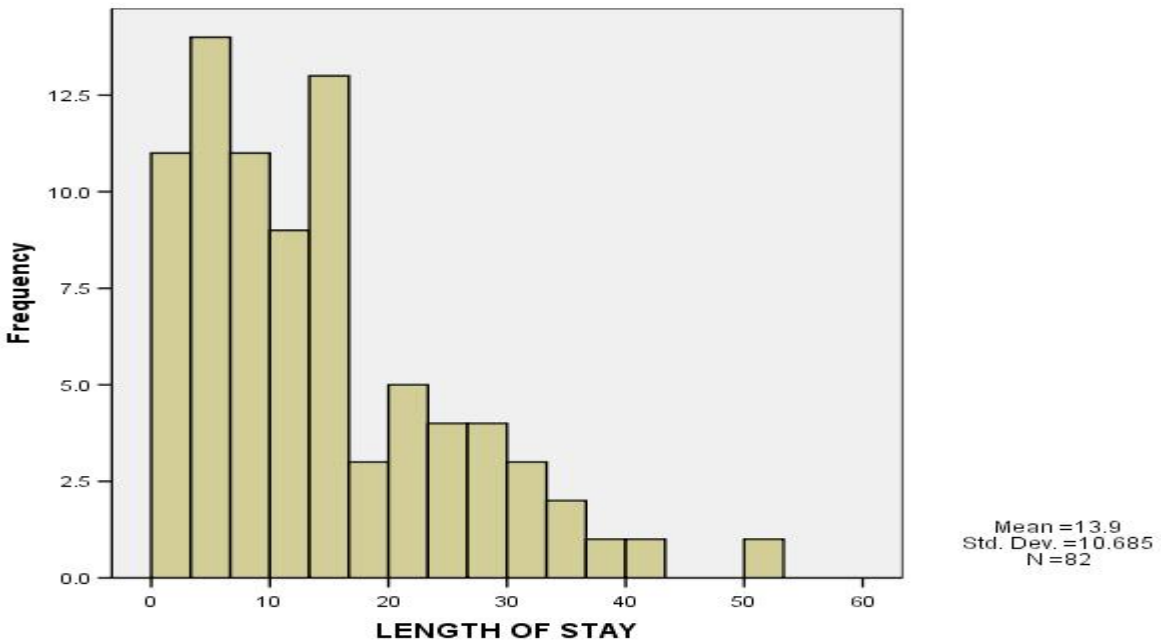
A histogram looks like a bar chart, but each bar represents a range of values or *interval* and is used when your data is continuous. The area of the bar is proportional to the number of patients in the interval, so histograms with bars of unequal width are possible. Table 4 shows the number of patients with various lengths of stay in hospital. To produce a histogram, data are first organised in a grouped frequency table:

Table 4 **Length of stay data**

Length of stay	Frequency (No. of patients)
0-3	11
4-7	18
8-11	10
12-15	18
16-19	4
20-23	5
24-27	5
28-31	4
32-35	4
36-39	1
40-43	1
44-47	0
48-51	0
52-55	1
Total	82

A single bar may, for example, represent all the patients with a length of stay between 4 and 7 days. The SPSS default histogram of the length of stay data is shown in Figure 2.

Figure 2 **SPSS Default histogram of length of stay**

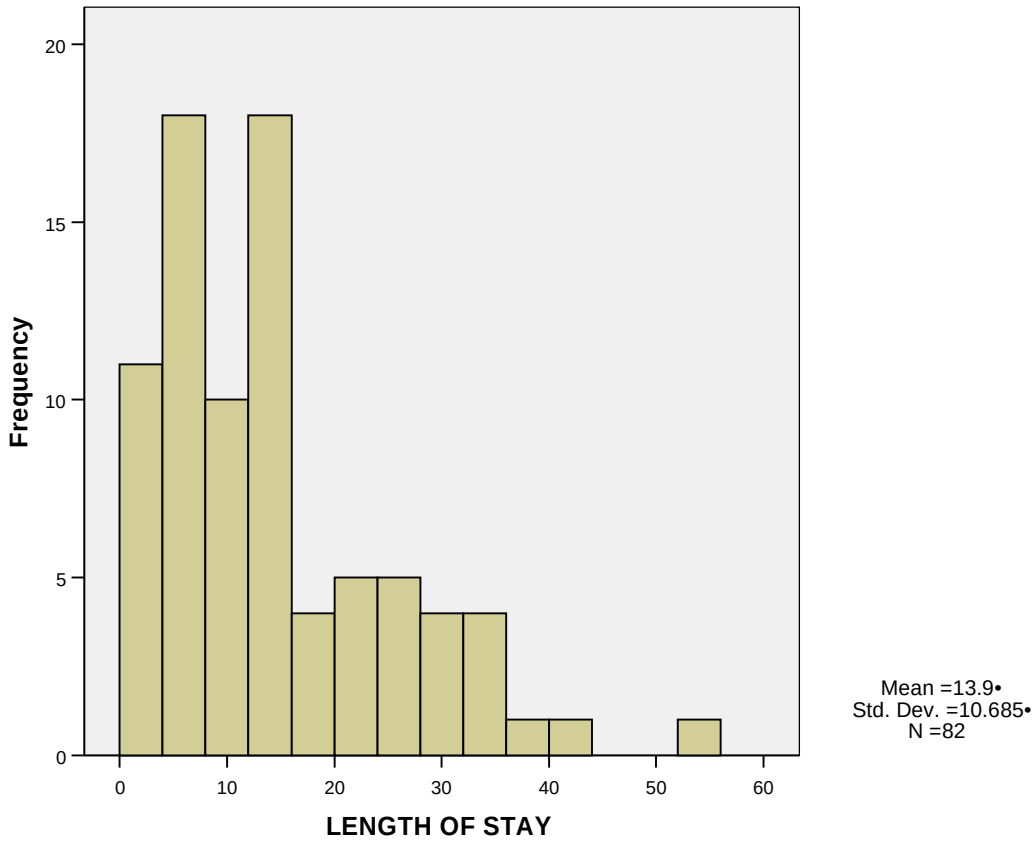


Presenting histograms in published documents can be fiddly, particularly where continuous data have been rounded. Here continuous data has been rounded to whole days, yet SPSS uses an interval width of 3.333 days. While the default

histogram in Figure 2 is fine for getting a feel for the shape of the distribution (ie clearly not symmetrical), it is, strictly speaking, incorrect, as the interval widths are shown as equal when in fact they are not.

For data rounded to whole days, it is best to use interval widths of whole days, as in Table 4. The histogram in Figure 3 correctly represents the data in Table 4, with equal interval widths of 4 days. (Figure 3 was created by double-clicking on the default histogram to enter the chart editor, then choosing the X tab for the x-axis and choosing the histogram Options tab. The bin size (interval width) was changed to a custom value of 4.)

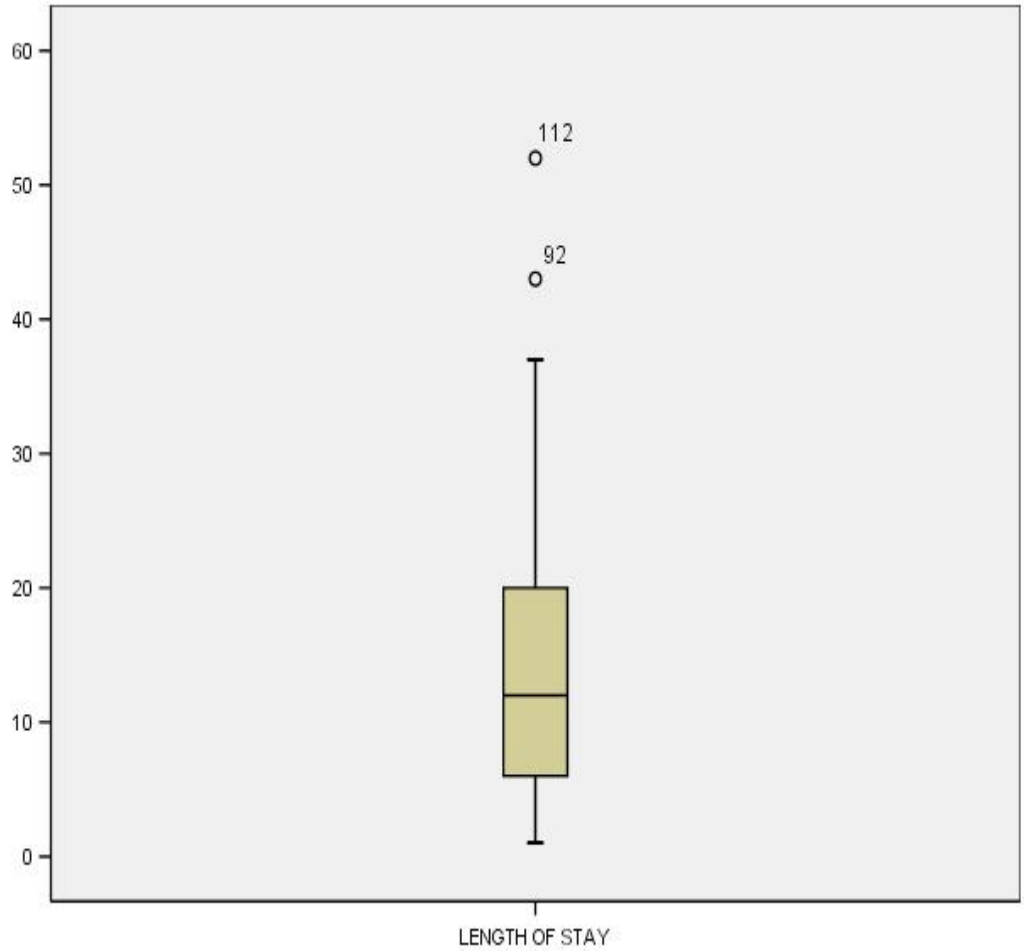
Figure 3: Corrected Histogram of Length of Stay



Boxplot

An alternative graphic for displaying continuous data is a **Boxplot**, (also known as a Box and Whisker Plot), such as Figure 4. Here the individual data values are used, and there is no need to be concerned about grouped intervals.

Figure 4 **Boxplot of length of stay**



Values of length of stay are shown on the left-hand axis. The upper and lower limits of the box show the limits of the interquartile range between 25th and 75th percentiles (6 to 20), and the middle horizontal line represents the median value (approx 12). The thin lines, often known as whiskers, extend to the extreme values of the sample (1 and 52). However if a point is more than 1.5 times box lengths from the upper or lower edge of the box it will be shown by SPSS as an individual point (known as an **outlier**). Here there are two outliers, with stays of 92 and 112 days.

A box plot can tell you, amongst other things, how symmetrical about its median a distribution is (skewness) and how long its tails are (pointedness or kurtosis). These data show a pattern typical of length of stay data. The distribution has 'positive skew', with a long positive tail of high values from a few patients with unusually long stays in hospital.

There are several reasons why an outlier may stand out from the majority of the other observations:

- It may be a genuine extreme observation, for example somebody particularly tall or with very high blood pressure. It is worth remembering that unusual things do happen. In this case you should include such observations in your analysis.
- On the other hand an outlier may be the result of a recording or transcribing error: for example a systolic blood pressure measurement being recorded as

416 instead of 146. If you expect that this is the cause of your outlier you should try to correct the record if possible. If this is not possible and you are sure that this is the cause of the problem then you should recode the observation to missing.

- Another possible explanation for outliers is that the extreme points may relate to individuals from a different population from the rest of the observations. In this situation it is a little trickier to know what to do. Your decision will depend on the question you are investigating and the population you want to investigate. However, the general recommendation is that you should include all observations unless you can present a good justification for not doing so.

Both histograms and boxplots can be useful in checking assumptions, such as whether your data are likely to have come from a Normal distribution, and for checking for outliers.

Boxplots are particularly useful, as we shall see later, if you wish to look at more than one group.

Example
This example explains how to summarise and present data using SPSS. The file created in Section 2, **pat_data.sav,** will be used.

If the file is not already open it can be loaded into SPSS as follows:

- Menu commands: **File** ⇒ **Open=>Data**
- Choose the drive and directory containing the data file
- Click on **pat_data.sav** to put the file name in the **File name** box
- Click on **Open**

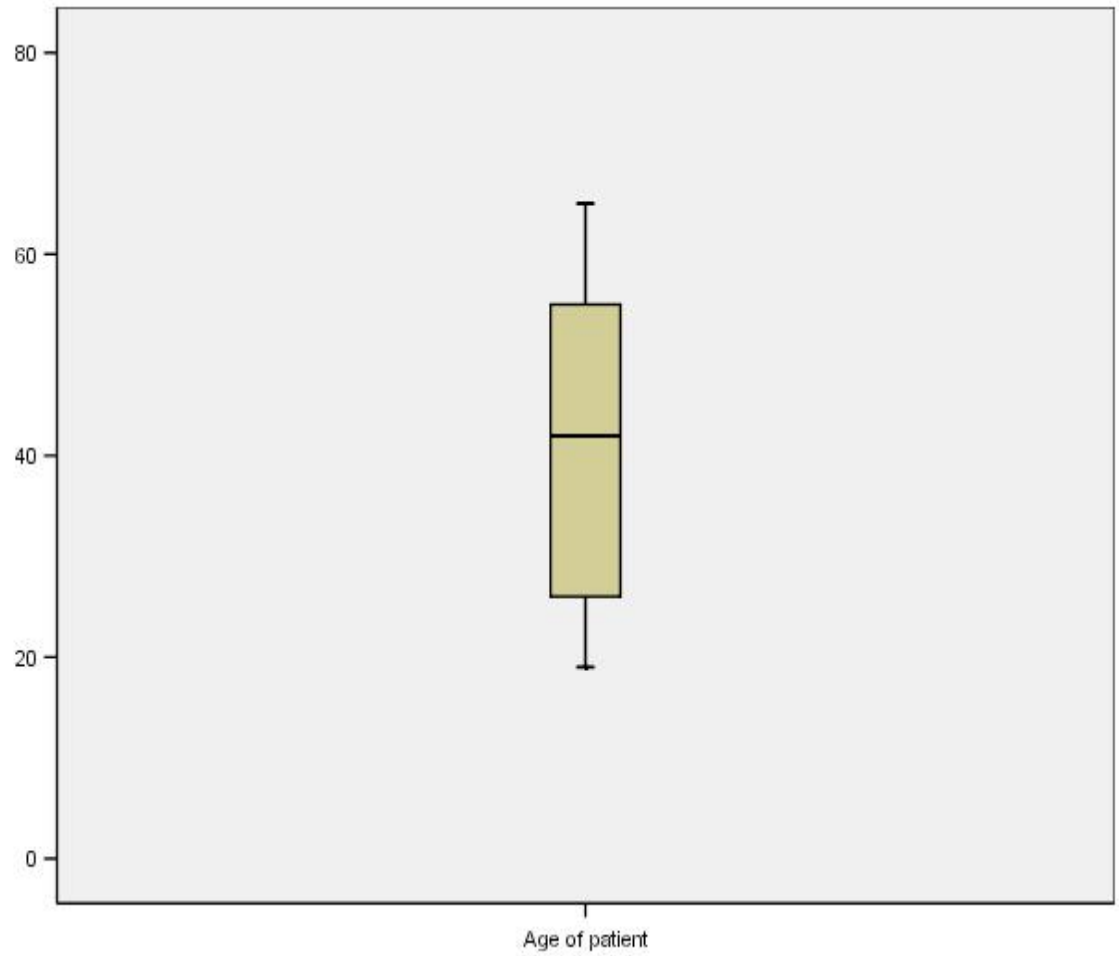
Now the dataset is loaded into SPSS we can start looking at it.
(i) Generate a boxplot for **age**

- Menu commands: **Graphs** ⇒ **Boxplot...**
- Ensure that the box labelled **Simple** is highlighted
- Ensure that **Summaries of separate variables** is checked
- Click on **Define**
- Click on the variable **age**
- Click on the arrow ▶ next to the label **Boxes represent:** to move the variable into the box
- Click on **OK**

The boxplot appears in a new window, named Output1. Results of any analysis that you request from SPSS will appear in this window. You can switch between Data and Output windows via the Window menu on the SPSS menu bar**.**

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age of patient	10	100.0%	0	.0%	10	100.0%



The boxplot shows minimum, 25th percentile, median, 75th percentile and maximum, it is also easy to see the interquartile range (length of box) and the range. From the boxplot, the distribution of age is reasonably symmetrical about the median.

(ii) Compute summary statistics for **age**

- Menu commands: **Analyse** ⇒ **Descriptive Statistics** ⇒ **Descriptives…**
- Click on the variable **Age**
- Click on the arrow ▶ to move the variable to the right-hand box
- Click on **OK**
- Inspect the output window:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age of patient	10	19	65	42.00	16.19
Valid N (listwise)	10				

(ii) Generate a frequency table for **bloodgp**

- Menu commands: **Analyze** ⇒ **Descriptive Statistics** ⇒ **Frequencies...**
- Click on the variable **bloodgp** in the left-hand box

- Click on the arrow ▶ to move the variable to the right-hand box
- Click on the button labelled **Statistics...**
- Locate the statistics grouped under the heading **Central Tendency** and click on **Mode** (to find the most common blood group in our sample). The box should be checked with a tick.
- Click on **Continue**
- Click on **OK**
- Inspect the output window:

Statistics

Blood group of patient

N	Valid	10
	Missing	0
Mode		1

Blood group of patient

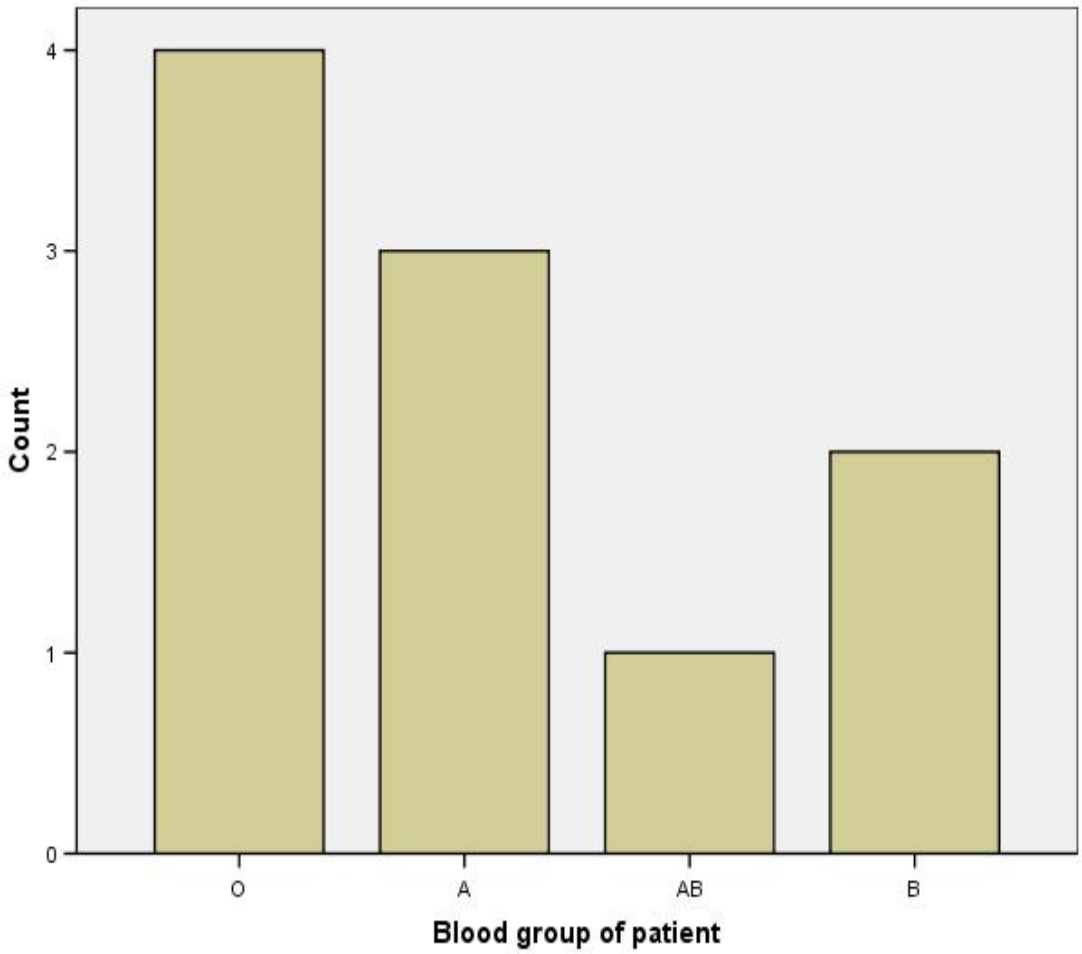
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	O	4	40.0	40.0	40.0
	A	3	30.0	30.0	70.0
	AB	1	10.0	10.0	80.0
	B	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

The first table above shows the number of valid observations and the number of missing observations. This table also shows the mode as being the group coded 1 and we know from Table 2, on Page 13 or from the variable view, that this is blood group O. The lower table is a frequency table showing group counts and percentages and verifies that the mode is type O.

(iii) Generate a bar chart for **bloodgp**

- Menu commands: **Graphs** ⇒ **Bar...**
- Click on **Simple** and ensure that 'summaries for groups of cases' is checked
- Click on the button labelled **Define**
- Click on the variable **bloodgp**
- Click on the arrow ▶ next to the label **Category Axis:** to move the variable into the box. Ensure that **Bars Represent N of cases** is checked
- Click on **OK**

Figure 5 **Bar chart for blood group**



A note on editing and saving SPSS output

Usually in SPSS sessions, a large amount of computer output is produced, some of which is repeated, and some may be just plain wrong. Usefully, the SPSS output editor includes an outline view, to help with editing output, which appears on the left-hand side of the output view:

Here, the outline view shows that there are two identical barcharts in this output window. To suppress the first barchart when printing or exporting output, click on the appropriate minus symbol in the outline view. To permanently delete the first barchart, first select by clicking on the coloured icon, then use **Edit → Delete.**

The output window can be completely cleared by using **Edit → Select all→ Delete.** Output may also be saved for editing later in SPSS, using **File** and **Save as** commands. SPSS output files have a **.spo** extension. Output may be exported to other packages (e.g. Word, PowerPoint) using the **File** and **Export** commands.

SPSS allows the user considerable command over the format of tables (see Options in the Edit menu bar). Double clicking on the graphics in the **Output Window** will open the **Chart Editor Window** and allow you to customise your graphics. It is possible, for example, to change symbols, colours and line types as well as to define and label axes. The chart and table options are quite detailed and are worth exploring once you have some experience of basic analyses in SPSS.

EXERCISE 2

Generate a histogram for **age.**

Summary

In this Section we have looked at a few ways of summarising and displaying data. This is an important part of any analysis but is often overlooked in the rush to apply more complex methods. However time spent understanding your data and checking for potential problems is likely to save you time and difficulties later.

4. Introducing an example: A Randomised Controlled Trial of Care in Hospital-at-Home v Care in a Hospital Ward

In Sections 5-8 we will analyse data from a clinical trial that took place in Leicestershire in 1995-97[3]. The objective of the trial was to compare effectiveness of patient care in the Hospital-at-Home scheme with hospital care. There were 199 patients assessed by their GP as suitable for care in Hospital-at-Home for a range of conditions including heart failure, chest infection, falls, and urinary tract infections, entered into the trial. Patients were randomised to receive either standard care in a hospital ward, or care in Hospital-at-Home, with care at home provided by nurses under the supervision of the patient's GP. Patient data was collected at admission (three days from randomisation), at discharge (two weeks from randomisation) and at three months.

The pack takes you through a somewhat simplified analysis of these data, in order to illustrate standard statistical analyses of two-sample data using SPSS. The follow-up data at three months are not considered here, and the dataset has been truncated.

The variables used for example analyses in this pack are, Barthel index of function, Sickness Impact Profile (SIP), satisfaction score and length of stay.

Variable *(label on SPSS dataset)*	*Description*
abarthel (admission barthel) dbarthel (discharge Barthel)	Observer rated score of physical function, based on ability to perform basic daily activities of living. Barthel scores range from 0 to 21, with 21 indicating ability to perform all the basic activities unaided.
Asip (admission SIP) dsip (discharge SIP)	SIP68, a shortened version of the sickness impact profile, measuring individuals' perceptions of the impact of sickness on usual daily activities, behaviours and feelings. Interviewer administered, with a maximum score of 68.
Satisf2 (Satisfaction questionnaire)	Patient self-completed questionnaire Satisfaction scores have maximum value of 24 (complete satisfaction).
Stay (length of stay)	Length of stay in Hospital-at-Home or hospital ward (days).

We will analyse patient outcome data from the trial, concentrating on patient outcomes at discharge to determine which group did better, on average, those in hospital wards, or those in Hospital-at-Home.

The data are in the form of an SPSS data file 'SPSS pack.sav' dataset and can be downloaded from:

http://www.trentrdsu.org.uk/uploads/File/spsspackresourcepack.zip

[3] Wilson A, Parker H, Wynn A, Jagger C, Spiers N, Jones J, Parker G. Randomised controlled trial of effectiveness of Leicester Hospital-at-Home scheme compared with hospital care. British Medical Journal 1999; 319:1542-1546.

Once you have this dataset it can be loaded into SPSS using the commands **File→Open→Data.**

When the data are loaded into SPSS you should spend a little time in data view getting a feel for the data, and some time in variable view investigating how the variables have been defined and, in particular, seeing how the missing values have been coded. It is good practice to explore data thoroughly before carrying out statistical tests, and for most tests we start with an appropriate plot of the data.

5. Parametric methods: Student's T-test and Related Confidence Intervals

One of the most common study designs that researchers find they need to analyse, is that of two groups, where they wish to compare the average value of some variable of interest. One way of doing this is by using Student's t-test, to obtain a p-value. Student's t-test requires that some assumptions are made for the test to be valid:

- First, the data should be scale (See Page 16).

- Second, the test assumes that the sampling means are Normally distributed. If the distribution is Normal, the distribution is symmetrical about the mean (so the mean is a good measure of centre), and approximately 95% of all data lies in the interval mean plus or minus 2 standard deviations. Normality of sampling means can usually be assumed if the samples are large enough, say at least 50 subjects in each group, or if the samples come from populations that are themselves Normally distributed.

- One further assumption often made is that the variances in each of the groups are approximately equal. There are test procedures to overcome this and, as we shall see, SPSS reports t-tests assuming both equal and unequal variances.

5.1 Confidence interval for a difference in means and t-test for two independent samples

When the two groups are independent the t-test for independent samples can be used to investigate the difference between the mean values of the two groups. The Hospital-at-Home Trial is an example of a straightforward "parallel groups" clinical trial where patients are randomised to one of two treatments. The two treatment groups are independent with no pairing or matching of patients, so techniques for two independent samples are appropriate.

Example

Suppose we want to compare the discharge Barthel Scores between the two groups (Hospital-at-Home and hospital ward). Before carrying out any statistical test it is advisable to think about and investigate the data, to satisfy yourself that the assumptions you are making are reasonable. We have produced a boxplot showing the discharge Barthel Score of the two treatment groups. (Figure 6.).

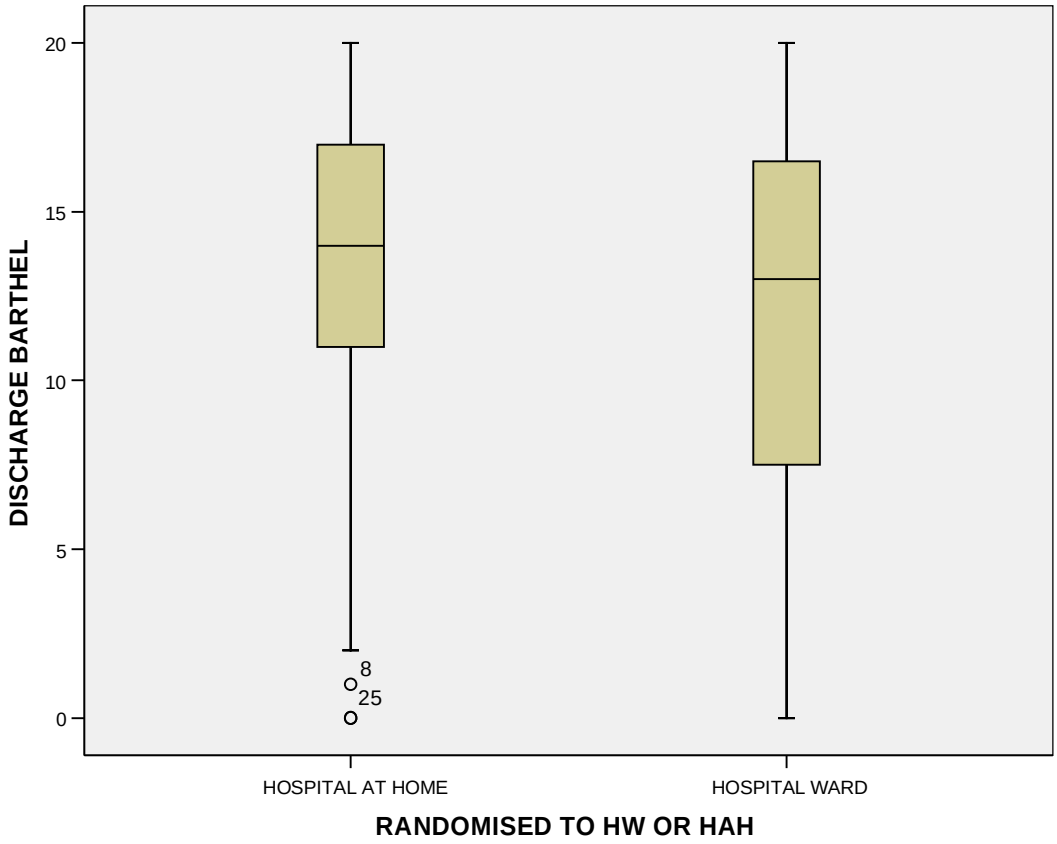
The following menu commands are used to produce a boxplot for different groups of patients.

Menu commands

- **Graphs** ⇒ **Boxplot...**
- Ensure that the box labelled **Simple** is highlighted
- Ensure that **Summaries for groups of cases** is checked
- Click on **Define**
- Click on the variable *dbarthel*

- Click on the arrow ▶ next to the box labelled label **Variable:** to move the variable into the box
- Click on the variable *treat*
- Click on the arrow ▶ next to the box labelled label **Category axis:**
- Click on **OK**

Figure 6 Boxplot of Discharge Barthel Score by Treatment Group



Although, strictly speaking, the Barthel score is ordinal, where there are many categories, and distances between categories are not obviously unequal, a score can be treated approximately as a scale measure.

We can use the boxplot above to investigate the assumptions we need to make in order to use a t-test. The boxplot does provide some evidence that the distribution of discharge Barthel Scores may be skewed towards the lower values. This is especially noticeable in the Hospital-at-Home group. One assumption of the t-test is that the observations come from a population that has a Normal distribution. Given that the sample size is not large, a Mann-Whitney U-test is the first choice test here. However, for teaching purposes we will proceed with the t-test. Although the interquartile range is larger in the hospital ward group, there is no clear difference in spread between the groups.

We can now use SPSS to carry out the significance test and to estimate a 95% confidence interval for the difference in mean discharge Barthel Score between the two groups.

Menu commands:

- **Analyze** ⇒ **Compare Means** ⇒ **Independent-Samples T Test**;
- Using the arrow ▶ put the variable *dbarthel* in the **Test Variable(s)** box
- Similarly, put the variable *treat* in the **Grouping Variable** box
- Click on the **Define Groups…** button
- Enter **1** in the **Group 1** box
- Enter **2** in the **Group 2** box
- Click on the **Continue** button
- Click on the **OK** button

The results from this procedure will now be in the output window.

T-Test

Group Statistics

	RANDOMISED TO HW OR HAH	N	Mean	Std. Deviation	Std. Error Mean
DISCHARGE BARTHEL	HOSPITAL AT HOME	33	13.21	5.95	1.04
	HOSPITAL WARD	36	12.00	5.59	.93

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DISCHARGE BARTHEL	Equal variances assumed	.040	.841	.873	67	.386	1.21	1.39	-1.56	3.98
	Equal variances not assumed			.871	65.514	.387	1.21	1.39	-1.57	3.99

The difference in mean discharge Barthel score between the groups is 1.2, indicating that the group randomised to Hospital-at-Home was, on average, slightly more physically able. However, we are interested in whether this difference is likely to have arisen by chance.

As we mentioned earlier, SPSS uses two procedures to calculate the t-test. In the top row of the lower table, an assumption has been made that the variances in the two groups are equal whereas in the bottom row no such assumption has been made. We will concentrate on the top row, as we have no reason to suspect that the variances are not equal. In fact, SPSS tests for equality of variances and finds no evidence of a difference (p = 0.841: columns two and three) but as usual it is worthwhile using some clinical knowledge and common sense too.

If we look first at the last column we can see that the 95% confidence interval for the difference between the groups is (–1.56 to 3.98). The fourth column gives the test statistic t = 0.873 and the two-sided[4] p-value, p = 0.386, is given in column six.

From this we conclude that there is no evidence of a statistically significant difference in mean Barthel Scores at discharge between the two groups, although the true difference is likely to lie somewhere between 4.0 points higher in the Hospital-at-Home group and 1.6 points higher in the hospital ward group. Note that in rounding,

[4] This pack uses two-sided confidence intervals and significance tests. These are usually appropriate. For a discussion of the rare situations in which a one-sided test may be appropriate see:
Bland JM, Altman DG. One and two sided tests of significance. British Medical Journal 1994: 309:248.

it is usual to make the confidence interval wider ie round the lower limit down and the upper limit up.

EXERCISE 3

Using similar methods, compare the Barthel Scores for the two groups at admission. Obtain a boxplot to check whether the assumptions for a t-test appear to be met. Arguably a t-test is more appropriate here. Why?

EXERCISE 4

Again using similar techniques, compare the Sickness Impact Profiles (SIP) at discharge.

5.2 Confidence interval for a difference in means and t-test for paired data

Sometimes the two groups are not independent. The study design may be that each patient in one group is **matched** to a patient in the other by variables that it is believed could influence the outcome but are not of interest in this particular study. For example, patients might be matched by age and smoking habit. It could also be the case that two groups are not independent as they consist of **repeated measurements** from the same individual, such as 'before and after' measurements. It is inappropriate to use the t-test for independent samples in these circumstances.

The paired-sample t-test and related confidence interval work by calculating the difference between the measurements for each patient and then calculating the mean of these differences. For studies where the patients have been individually matched, the difference within each of the matched pairs is calculated and then these differences are investigated. If there were really no difference between the two sets of measurements we would expect the average of differences from all pairs to be around zero. The assumption that this procedure makes is that the differences are approximately Normally distributed. Once again this should be examined but it is worth remembering that differences between measurements are much more likely to follow a Normal distribution than the raw measurements themselves.

Example

Suppose that we are interested in the overall change in Barthel Score from admission to discharge for all patients. In other words, is the mean change in Barthel Score from admission to discharge equal to zero?

Once again we should start by checking that any assumptions we are making are reasonable in this case. Here we are interested in the distribution of the differences calculated by subtracting each patient's discharge Barthel score from their admission score (see Figure 7). It is necessary to calculate these differences to be able to plot them. In SPSS this can be done using the commands below:

Menu commands:

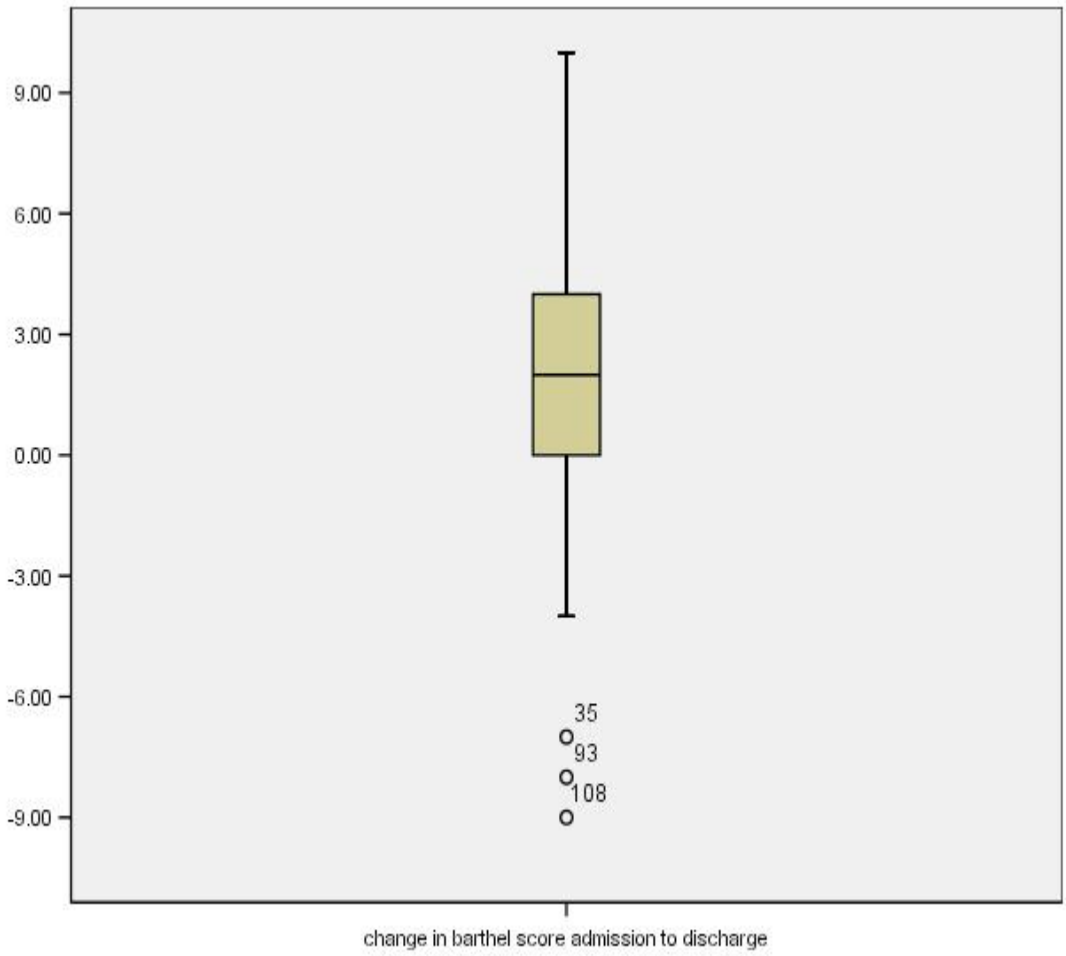
- **Transform ⇒ Compute…**

- Type the name of your new variable (*chbart*) in the box labelled **Target variable:**
- Click on the **Type & Label…** button to add a label (*Change in Barthel Score admission to discharge*) and to check the variable is of the correct format, in this case numeric, then click **Continue**
- Select the variable *dbarthel* then clicking the arrow ▶ to place this variable in the **Numeric Expression:** box
- Next type or select the minus '-' symbol
- Select the variable *abarthel* then clicking the arrow ▶ to place this variable in the **Numeric Expression:** box to give the expression '**dbarthel-abarthel'**
- Click on the **OK** button

In data view, examine the new variable chbart, which appears on the far right hand side of the dataset in Data View, and at the bottom in Variable View.

This new variable can now be plotted to check the assumptions of the paired t-test (details of how to obtain a boxplot for a single group are given on Page 21).

Figure 7 Boxplot of Change in Barthel Score



The boxplot supports the assumption that the differences come from a Normal distribution. SPSS identifies three people who declined noticeably in physical function as outliers, but none is excessively far from the median and they seem to

represent clinically plausible changes in Barthel Score, so we will decide to retain these scores.

Menu commands:

- **Analyze** ⇒ **Compare Means** ⇒ **Paired-Samples T Test**
- Select the variable *abarthel* (it will be shown in the **Current Selections** box against **Variable 1**)
- Then select the variable *dbarthel* (this will be shown in the **Current Selections** box against **Variable 2**)
- Clicking the arrow ▶ will now place these variables in the **Paired Variables** box indicating that we are going to analyse the difference *abarthel* minus *dbarthel*
- Click on the **OK** button

(Note, although analysing the change score dbarthel-abarthel is more intuitive, SPSS will not allow this, as it sorts the variables alphabetically.)

T-Test

[DataSet1] F:\SPSStrainer2002\SPSSpack.sav

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	ADMISSION BARTHEL	11.31	61	5.188	.664
	DISCHARGE BARTHEL	13.03	61	5.674	.727

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	ADMISSION BARTHEL & DISCHARGE BARTHEL	61	.767	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	ADMISSION BARTHEL - DISCHARGE BARTHEL	-1.721	3.738	.479	-2.679	-.764	-3.597	60	.001

From the last table we can see that the Barthel Scores at admission were, on average, 1.7 points lower than at discharge. A 95% confidence interval for this difference is given as (-2.68 to –0.76) and does not include zero. The test statistic is t = -3.597 giving a p-value of p = 0.001. We can conclude, therefore, that there is strong evidence for an improvement in Barthel Score from admission to discharge and that the average improvement is estimated to be between 0.7 and 2.7 points.

Although the paired-sample t-test is appropriate for these data, in a case such as this where data at baseline and outcome are available, a slightly more sophisticated analysis (Analysis of Covariance) is preferred, as it is more likely to detect a real difference between the treatment groups. For details of Analysis of Covariance, which is beyond the scope of this guide, please refer to the texts listed in Section 10.

EXERCISE 5

Using similar techniques investigate whether there is evidence of a change in SIP score between admission and discharge.

Summary

In this section we have looked at the difference in mean response in situations where we have two groups. These methods can be extended to situations with more than two groups by using Analysis of Variance (ANOVA), see the books listed in Section 10 for an introduction to ANOVA.

The tests presented in this section made the assumption that the Barthel score is Normally distributed. In reporting the actual study it was felt that this assumption was not defensible, and non-parametric analyses were reported. The next chapter presents distribution-free analyses that may be appropriate for ordinal data and scale data when Normality cannot be assumed.

6. Distribution-Free methods: Mann-Whitney U Test and Wilcoxon Matched-Pairs Test.

Often, the assumptions for the t-test are thought not to hold. It may be, for example, that the dataset is small and the variable has a skewed distribution or that the outcome variable has been measured on an ordinal scale with few categories, rather than an interval scale. In these circumstances it may be more appropriate to use distribution-free (or non-parametric) methods. This generally applies to data where the median and interquartile ranges are more useful statistics than the mean and standard deviation. Distribution-free tests are less powerful than the t-test (ie less likely to detect a real difference between groups) if the data really do follow a Normal distribution, although this is only likely to be noticeable for small samples. Non-parametric tests are likely to be more powerful than t-tests if the population distributions are skewed.

6.1 Mann-Whitney U-test for two-independent samples

The Mann-Whitney U-test is used to investigate whether there is a tendency for values in one group to be higher, or lower, than in another group. If the distributions in the two groups are of similar shape this test can be interpreted as a test of differences in median.

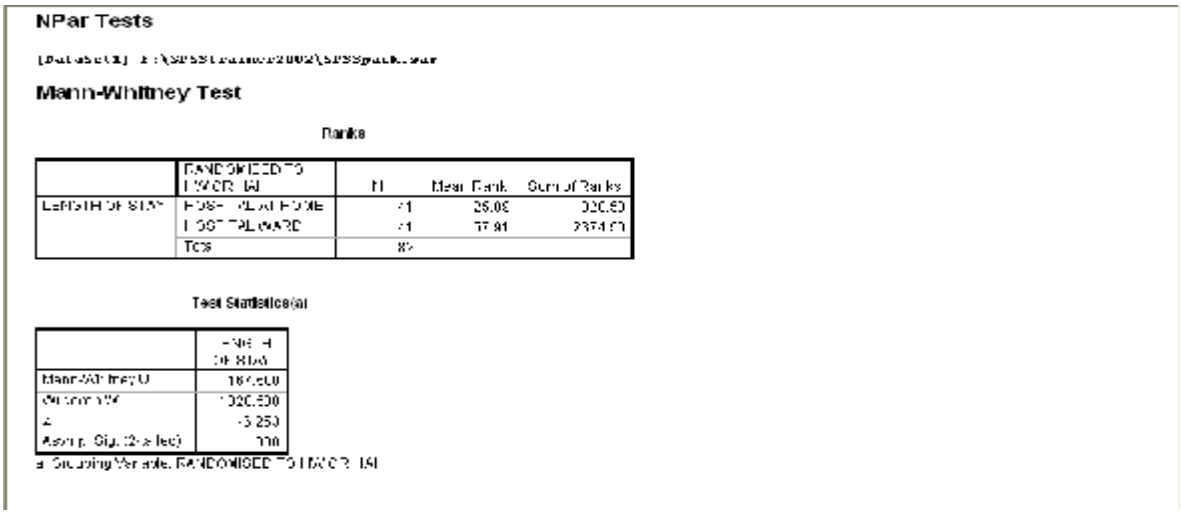
Example

We are interested in how the length of stay of patients varies between the two treatment groups. However, it may be unwise to use a t-test as the data appear to be positively skewed. (You should check this by producing a histogram or boxplot.) As there are not a very large number of observations, only 41 with length of stay data in each group, it may be unwise to use a parametric test. We will, therefore, use the Mann-Whitney U-test.

Menu commands:

- **Analyze** ⇒ **Nonparametric Tests** ⇒ **2 Independent Samples**
- Put the variable *stay* in the **Test Variable List** box
- Put the variable *treat* in the **Grouping Variable** box
- Click on the **Define Groups…** box and enter *1* in the box next to **Group 1** and *2* in the box next to **Group 2**
- Click on the **Continue** button

- Make sure **Mann-Whitney U** is selected in the **Test Type** box
- Click on the **OK** button



NPar Tests

[DataSet1] F:\SPSS Training\2002\SPSSpack.sav

Mann-Whitney Test

Ranks

	RANDOMISED TO HWORD [W]	N	Mean Rank	Sum of Ranks
LENGTH OF STAY	POST-TRIAL HOME	41	25.05	1027.00
	HOSPITAL WARD	41	57.95	2374.00
	Total	82		

Test Statistics(a)

	LENGTH OF STAY
Mann-Whitney U	167.500
Wilcoxon W	1027.500
Z	-5.250
Asymp. Sig. (2-tailed)	.000

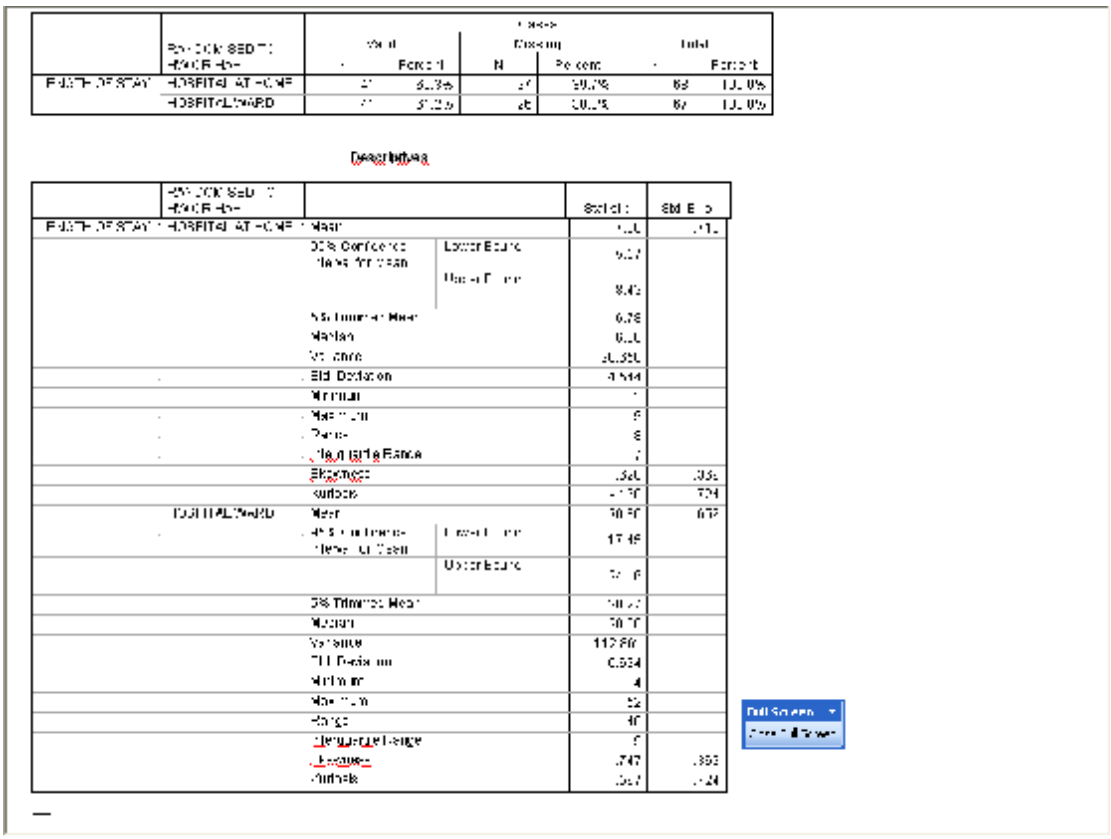
a Grouping Variable: RANDOMISED TO HWORD [W]

The test works by ranking the observations from 1 for the shortest stay to 82 for the longest stay. From the mean ranks it can be seen that patients in the hospital ward group generally had higher ranks (ie longer stays).

The lower table gives the test statistic U = 167.5. The p-value associated with this is given as p = 0.000 which means $p < 0.001$. This indicates that there is very strong evidence that length of stay in one group is longer than the other.

This procedure in SPSS does not give us any estimate of the size of the difference or a corresponding confidence interval. If we make the assumption that the distributions of the outcomes in each group are of roughly the same shape and of equal spread the Mann-Whitney test becomes a test of the equality of medians. However this assumption is not reasonable for these data. The median length of stay can be displayed for each group using the explore command:

- **Analyze** ⇒ **Descriptive Statistics** ⇒ **Explore**
- Put the variable *stay* in the **Dependent List** box
- Put the variable *treat* in the **Factor List** box
- Click on the **Statistics** box and ensure that **Descriptives** is checked
- Click on the **Continue** button
- Click on the **OK** button



If the distributions in the two groups are similar it is appropriate to report a confidence interval for the difference in medians. Such confidence intervals are produced by some alternative statistical packages, such as MINITAB. The method is described in standard texts[5]

EXERCISE 6

The variable *satisf2* is a score of patient satisfaction with their treatment with a maximum score of 24. Using the Mann-Whitney test, investigate whether there is evidence of a difference in satisfaction level between the two groups.

6.2 Wilcoxon matched-pairs test

As with the matched-pairs t-test the Wilcoxon matched-pairs test should be used when the two groups are not independent. The test works with the differences between the matched observations. This means that although the test is suitable for situations when the differences do not follow a Normal distribution, the original observations should be on an interval scale to enable the differences to be ranked according to size. The test is applicable so long as the differences have a symmetrical distribution. If not, the sign test can be used.

Example

[5] Altman, D.G, Machin, D., Bryant, T., Gardner, M. J. (Eds) Statistics with Confidence (2nd Ed.) BMJ 2000. (Includes software for calculating Confidence Intervals)

To illustrate this test we will once again look at the change in Barthel Score from admission to discharge. We saw on Page 35 that these differences appear to come from a Normal distribution.

Menu commands:

- **Analyze** ⇒ **Nonparametric Tests** ⇒ **2 Related Samples**;
- Select the variable *abarthel* (it will be shown in the **Current Selections** box against **Variable 1**);
- Then select the variable *dbarthel* (this will be shown in the **Current Selections** box against **Variable 2**);
- Clicking the arrow ▶ will now place these variables in the **Test Pair(s) List** box indicating that we are going to analyse the difference *dbarthel* minus *abarthel*;
- Make sure **Wilcoxon** is selected in the **Test Type** box;
- Click on the **OK** button.

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
DISCHARGE BARTHEL - ADMISSION BARTHEL	Negative Ranks	11[a]	24.09	265.00
	Positive Ranks	39[b]	25.90	1010.00
	Ties	11[c]		
	Total	61		

a. DISCHARGE BARTHEL < ADMISSION BARTHEL

b. DISCHARGE BARTHEL > ADMISSION BARTHEL

c. ADMISSION BARTHEL = DISCHARGE BARTHEL

Test Statistics[b]

	DISCHARGE BARTHEL - ADMISSION BARTHEL
Z	-3.611[a]
Asymp. Sig. (2-tailed)	.000

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

The lower box gives us the test statistic z = -3.611 and the corresponding p-value, p < 0.001. There is, therefore, very strong evidence for a change in Barthel Score from admission to discharge.

As expected, this agrees with the result from the paired t-test in Section 5.

EXERCISE 7

Carry out a Wilcoxon matched-pairs test to investigate whether there is evidence of a change in SIP scores from admission to discharge. Check that the result is consistent with the results from the paired t-test on the same data (Exercise 5).

Summary

In this section we have looked at alternative methods of analysis that do not require the same assumptions about the data as the t-test and related confidence interval. These tests offer the attraction of being valid in a much wider range of circumstances and it may be tempting to use them, all of the time.

However there are two disadvantages to these distribution-free methods. First, they are not as powerful as t-tests when the latter are valid. Secondly it is more difficult to obtain confidence intervals for differences, they are not available as options in SPSS.

You should examine and think about your data, and if you believe that all of the assumptions are met the t-test is recommended as the most powerful option, giving the best chance of detecting any difference between groups.

7. Methods for Categorical Data: Chi-squared Test and Relative Risk

When the data is presented as a cross-tabulation of frequencies (known as a contingency table) the methods described so far are unsuitable. Here we will look at standard analyses to detect an association between a categorical explanatory variable (such as treatment group) and a categorical outcome variable (such as discharge destination or survival). In other words, did the pattern of discharge destinations differ between Hospital-at-Home and hospital ward patients? And did the risk of death differ between Hospital-at-Home and hospital ward patients?

7.1 Chi-Squared Test

This statistical test is used to investigate whether there is an association between two categorical variables.

Example

In this example we are interested in whether there is a difference between the two treatment groups in discharge destination (home; hospital/residential care; died). The hypothesis tested is a null hypothesis of no association between the explanatory variable and the outcome. In other words, that the pattern of discharge destinations is the same in each group.

Menu commands:

- **Analyze** ⇒ **Descriptive Statistics** ⇒ **Crosstabs**
- Put the variable *treat* in the **Column(s)** box
- Put the variable *disdest* in the **Row(s)** box
- Click on the **Statistics…** box and make sure that the box next to **Chi-square** is ticked
- Click on the **Continue** button
- Click on the **Cells**…box and check the box next to **Column** under the **Percentages** title
- Click on the **Continue** button
- Click on the **OK** button

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DISCHARGE DESTINATION * RANDOMISED TO HV OR HAH	85	63.0%	50	37.0%	135	100.0%

DISCHARGE DESTINATION * RANDOMISED TO HV OR HAH

Crosstabulation

			RANDOMISED TO HV OR HAH		Total
			HOSPITAL AT HOME	HOSPITAL WARD	
DISCHARGE DESTINATION	home	Count	23	29	52
		% within RANDOMISED TO HV OR HAH	51.1%	72.5%	61.2%
	hospital/res dential	Count	16	5	21
		% within RANDOMISED TO HV OR HAH	35.6%	12.5%	24.7%
	died	Count	6	6	12
		% within RANDOMISED TO HV OR HAH	13.3%	15.0%	14.1%
Total		Count	45	40	85
		% within RANDOMISED TO HV OR HAH	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.181[a]	2	.045
Likelihood Ratio	6.469	2	.040
Linear-by-Linear Association	6.109	1	.434
N of Valid Cases	85		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.65.

The bottom table gives the result from the chi-squared test, p = 0.045. The null hypothesis of no association is rejected (at the 5% significance level) indicating that there is some evidence of differences in discharge destinations between the two treatment groups. The middle table is the contingency table. We can see from this table that those treated by Hospital-at-Home are less likely to be discharged home (51.1% compared with 72.5%) and are more likely to be discharged to hospital or residential care (35.6% compared with 12.5%) than those treated in hospital wards.

The chi-squared test is unreliable for small samples, so if one of the cells has an expected frequency of 5 or less, Fisher's Exact Test, should be reported. For 2x2 tables, SPSS computes Fisher's Exact Test automatically. For other tables, Fisher's Exact Test can be requested by clicking on the Exact button in the Crosstabs dialogue box and then checking the Exact option.

7.2 Relative Risk With Confidence Interval

Although the chi-squared test is very useful it does not give an estimate of the size of any difference between the groups. One measure that is often used is the **Relative Risk**.

The **Relative Risk** is the ratio of the risk (probability) of the outcome of interest occurring in one group compared to the risk of it happening in the other group. For example, the 'risk' of discharge to home in the Hospital-at-Home group is 51.1%, compared with 72.5% in the hospital ward, so the relative risk (51.1 ÷ 72.5) is 0.705. A relative risk of 1.0 indicates no difference between the groups, but here the relative risk is somewhat less than 1 as those in Hospital-at-Home were less likely to be discharged home.

One point to remember is that the Relative Risk is appropriate for clinical trials and cohort studies, but unsuitable for the analysis of case-control studies. For these types of studies an alternative measure, the **Odds Ratio**, should be used. Details of the Odds Ratio are beyond the scope of this resource pack, please see the further resources listed in Section 10.

Example

To illustrate the calculation of the Relative Risk and confidence interval we will look at the relative risk of death between the two treatment groups for the Hospital-at-Home Trial. Note that, to get the correct relative risk, the treatment or grouping variable must be in the rows of the table.

Menu commands:

- **Analyze** ⇒ **Descriptive Statistics** ⇒ **Crosstabs**
- Click on the **Reset** button to clear the dialog box
- Put the variable *treat* in the **Row(s)** box
- Put the variable *died* in the **Column(s)** box
- Click on the **Statistics…** box and make sure that the box next to **Risk** is ticked
- Click on the **Continue** button
- Click on the **Cells…** box and make sure that the box labelled **Row** under the heading **Percentages** is ticked
- Click on the **Continue** button
- Click on the **OK** button

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
RANDOMISED TO HW OR HAH * DIED	85	63.0%	50	37.0%	135	100.0%

RANDOMISED TO HW OR HAH * DIED Crosstabulation

			DIED		
			.00	1.00	Total
RANDOMISED TO HW OR HAH	HOSPITAL AT HOME	Count	39	6	45
		% within RANDOMISED TO HW OR HAH	86.7%	13.3%	100.0%
	HOSPITAL WARD	Count	34	6	40
		% within RANDOMISED TO HW OR HAH	85.0%	15.0%	100.0%
Total		Count	73	12	85
		% within RANDOMISED TO HW OR HAH	85.9%	14.1%	100.0%

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for RANDOMISED TO HW OR HAH (HOSPITAL AT HOME / HOSPITAL WARD)	1.147	.338	3.891
For cohort DIED = .00	1.020	.857	1.213
For cohort DIED = 1.00	.889	.312	2.536
N of Valid Cases	85		

As before, the middle table is the contingency table, with those who were alive at discharge shown under column .00 and those who died before the end of the trial under column 1.00.

The bottom table gives the estimates for the differences between the two treatment groups. You will notice that there are three estimates of the difference between the two groups.

Starting from the top, the first figure, 1.147, is an estimate of the Odds Ratio for mortality. We will not consider this further.

From the second row we can see that the relative risk for survival is 1.020, and from the third row the relative risk for mortality is 0.889. 95% confidence intervals are given for both of these estimates. The relative risk can be estimated directly from the cross tabulation, $\frac{86.7}{85.0} = 1.02$, although in some cases the rounding of the percentages presented in the table will introduce small errors.

The relative risk of survival in the Hospital-at-Home group compared to the hospital ward group of 1.02 indicates that the risk of survival was similar in the two groups. The probability of surviving was 2% greater in the Hospital-at-Home group (ie 86.7% is 85.0% multiplied by 1.02) but, as we shall see, this very small difference is not statistically significant.

The value 1 for a relative risk corresponds to no difference between the groups. In our example both of the 95% confidence intervals contain the value 1 and we can conclude, therefore, that there is no evidence (at the 5% significance level) for any difference in mortality between the two treatment groups. However, of more interest is likely to be the confidence interval, suggesting that the true risk of mortality in the

hospital ward treated group is likely to lie between 31% and 254% of the risk of mortality in the Hospital-at-Home group. The width of the confidence interval indicates that the study is too small to tell us much about mortality differences between the groups.

EXERCISE 8

Using a calculator, and referring to the cross-tabulation, verify that the relative risk for mortality is 0.889.

EXERCISE 9

Is there an association between gender and mortality? Investigate this question using:

a) The chi-squared test;
b) Relative risk and confidence interval.

Write your conclusions in a sentence or two, as you would for a report.

Summary

We have seen that the Chi-Squared test allows us to test for an association between two categorical variables. This test can be extended to allow for more than two categories for each variable and also when the categories can be ordered (ordinal data).

While the Chi-Squared test is useful it does not quantify the difference between groups. This can be done using the relative risk.

8. Examining Relationships Between Two Continuous Variables: The Correlation Coefficient

All of the methods covered so far have concerned the situation where you have two groups that you wish to compare. Often, however, researchers are interested in investigating the relationship between two continuous measures, for example age and height among children or blood pressure and age amongst adults. We will look at using the **correlation coefficient** to establish whether such a relationship exists. The related technique, **linear regression** that allows you to describe a straight line relationship between two variables is beyond the scope of this pack.

Correlation

Correlation is a measure of the strength of *linear* relationship between two variables. The statistical measure of linear association is known as the correlation coefficient, denoted by the symbol ***r,*** and measures how close the points lie to a straight line. Its value always lies between –1 and +1. The value +1 indicates a perfect positive relationship between the two variables (Figure 8) and the value –1 indicates a perfect negative relationship. More often, however, you will find that the value of the correlation coefficient lies somewhere away from these two extremes (for example Figure 9). If the correlation coefficient takes the value zero then there is no linear relationship between the two variables (Figure 10), although you should always remember that there might be a non-linear relationship (Figure 11) so always plot the data first!

Figure 8 ***r* = +1**

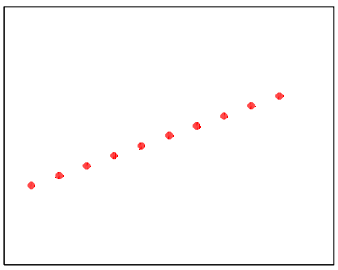


Figure 9 ***r* = -0.81**

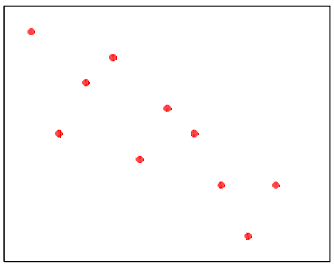


Figure 10 ***r* = 0**

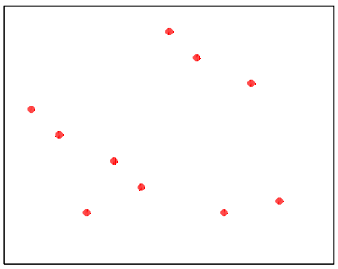
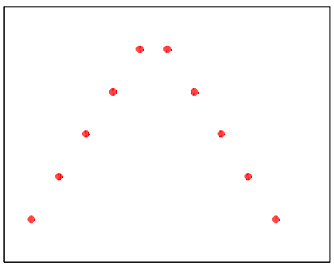


Figure 11 ***r* = 0**



Example

We would expect to find an association between the admission Barthel Score and the discharge Barthel Score, as these scores are repeated measures on the same patients. It is likely that patients with better function at admission have better function

at discharge and vice versa. We will use the correlation coefficient to test the hypothesis that there is no such relationship

First we will produce a scatterplot to be able to visualise the relationship.

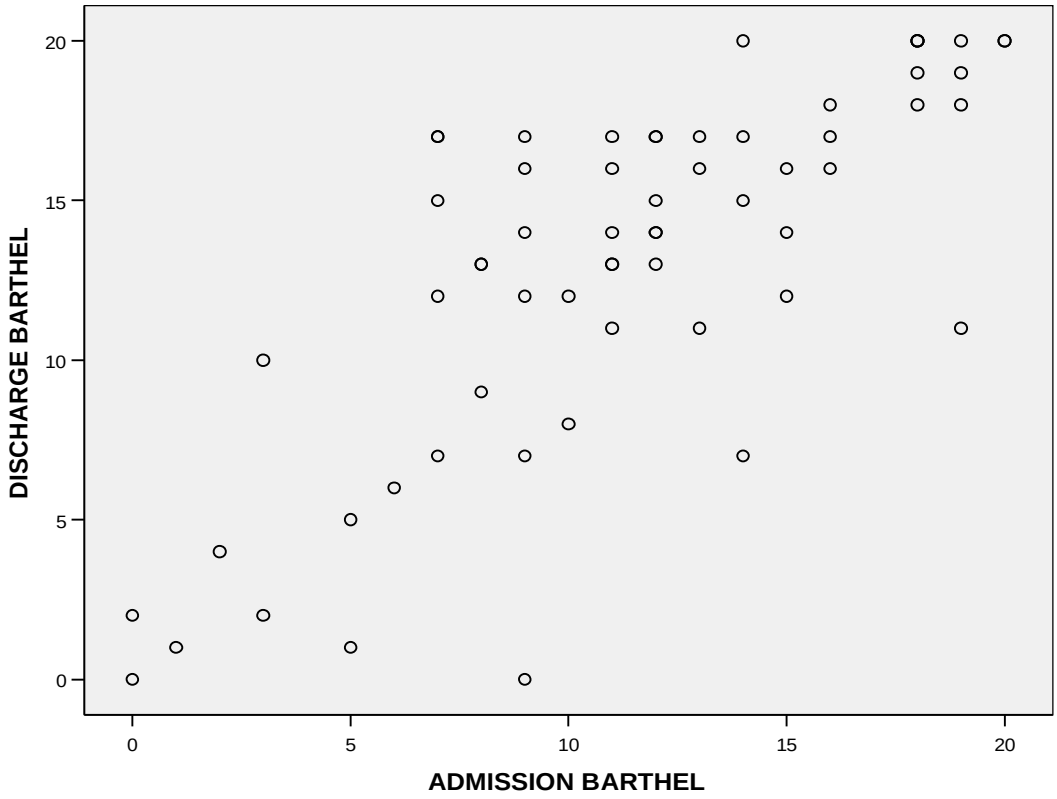
Menu commands:

- **Graphs** ⇒ **Scatter…**;
- Click on **Simple** and then on the button labelled **Define**
- Put the variable *dbarthel* in the **Y Axis** box;
- Put the variable *abarthel* in the **X Axis** box;
- Click on the **OK** button.

Menu commands to calculate the correlation coefficient:

- **Analyze** ⇒ **Correlate** ⇒ **Bivariate**;
- Put the variables *abarthel* and *dbarthel* in the **Variables(s)** box;
- Make sure that the box next to **Pearson** is ticked;
- Click on the **OK** button.

Figure 11 Scatterplot of Discharge Barthel by Admission Barthel



Correlations

		abarthel	dbarthel
abarthel	Pearson Correlation	1	.767(**)
	Sig. (2-tailed)		.000
	N	85	61
dbarthel	Pearson Correlation	.767(**)	1

	Sig. (2-tailed)	.000	
	N	61	69

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

From the scatterplot we can see that, as expected, there is a clear positive relationship between the Barthel Scores at admission and discharge.
This is supported by the strongly positive value of the correlation coefficient of 0.767. SPSS also carries out a significance test with the null hypothesis of $r = 0$, that is, of no linear relationship between the two variables. In this case the null hypothesis is rejected with $p < 0.001$, so the positive correlation is statistically significant. A confidence interval for the correlation coefficient in the population would be more useful here, but is not available in SPSS.

Caution when using the Correlation Coefficient

If the correlation coefficient shows no linear relationship remember that a non-linear relationship between the two variables is not ruled out.

Conversely, if the correlation coefficient shows a linear relationship, be careful in interpreting this. The relationship may be due to measurement error, or to repeated measures on the same subject, or to measuring the same underlying trait by two different methods, rather than anything more interesting!

In short, do not report a correlation coefficient without first examining the corresponding scatterplot, and understanding the variables (ie what was measured and how.)

EXERCISE 10

Using similar techniques, investigate the relationship between the change in SIP observed during the trial and admission SIP. Would you expect the two variables to be related?

Summary

In this section we have looked at using the correlation coefficient to assess whether there is a linear relationship between two continuous variables. Although this technique is useful, correlation coefficients must be interpreted with care. Another commonly used method of analysis is linear regression. This is used to describe the relation between two continuous variables and examines how much a variable (dependent, response or outcome variable) changes when the other variable (independent, predictor or explanatory variable) changes by a certain amount. This method may also be extended to look at the effect of several variables on a response variable simultaneously. Details of these methods can be found in the books listed in Section 10.

9. Concluding Remarks

The pack was written with the SPSS beginner in mind and to encourage use of the package rather than produce an expert user. After working through the pack it is likely that you will have more questions than when you started. Fortunately help is at hand from a number of sources. First, SPSS for Windows contains a comprehensive on-line Help system that can be consulted should a question arise. Second, other manuals and textbooks are available, examples of which are listed in Section 10. Third, you may wish to attend a short course in data analysis and statistics, such as those provided by The NIHR RDS EM / YH. Finally, there are many experienced

users of SPSS for Windows in the field of health services research who may be willing and able to help.

When using a statistical package to analyse data we should not forget the wider context of the research process. This includes an appreciation of the research design, data collection, data analysis and interpretation of results. These aspects have not been discussed at length in this pack. Further information can be found in other resource packs in this series and textbooks of statistics, again examples of which are listed in Section 10.

This resource pack has provided a practical step-by-step guide to SPSS for Windows for those new to the package. It is hoped that it has provided you with a firm foundation and confidence in using SPSS for Windows for basic data analysis.

10. Further Reading and Resources

There are many good textbooks available on basic statistical methods. Some of the most widely available are listed below.

Altman, D.G. (1991), *Practical Statistics for Medical Research*, London, Chapman and Hall.

Bland, M. (2000), *An Introduction to Medical Statistics*, 3rd ed., Oxford, Oxford University Press.

Altman, D.G, Machin, D., Bryant, T., Gardner, M. J. (Eds) Statistics with Confidence (2nd Ed.) BMJ 2000.

Pallant, J. (2004) *SPSS Survival Manual*, Open University Press.

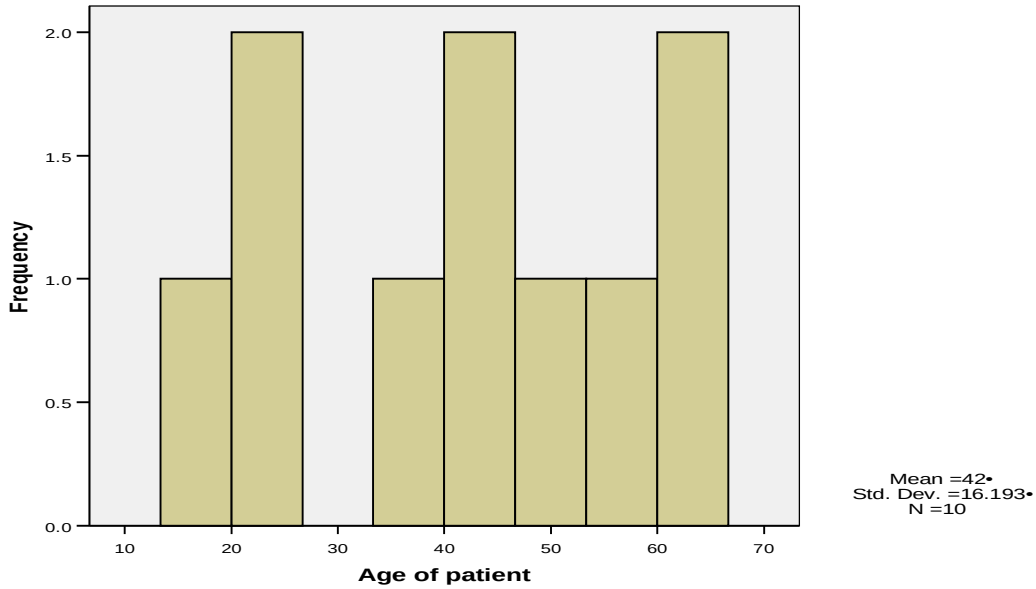
Rowntree, D. (1991), *Statistics without Tears*, London, Penguin.

Swinscow, T.D.V. (2002), *Statistics at Square One*, 10th ed., London, BMJ.
Also available at: http://www.bmj.com/collections/statsbk/index.dtl

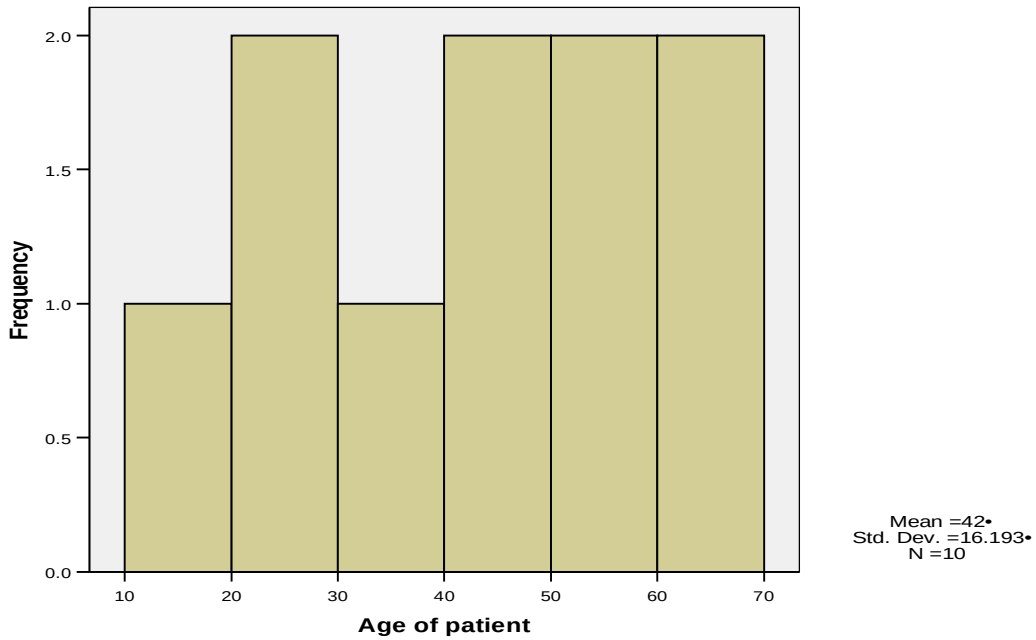
Answers to exercises

Exercise 2

The default histogram for ‘age’ is shown below:

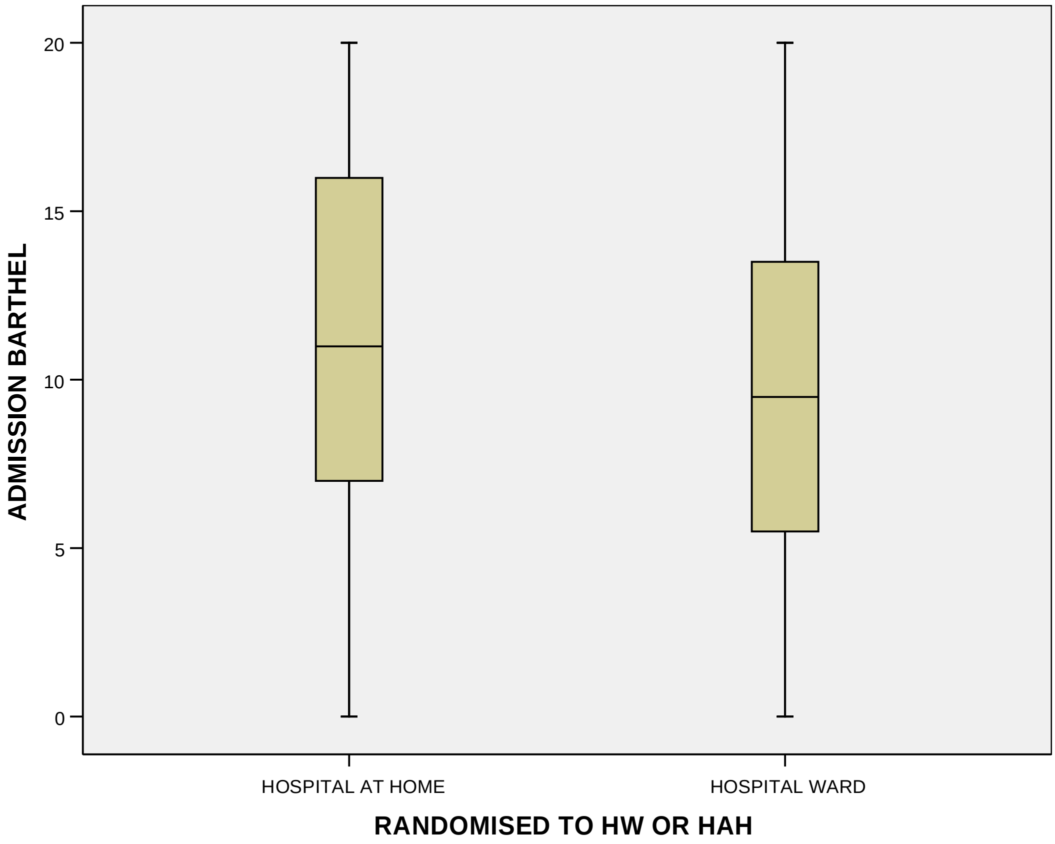


However, using the graphics options available by double-clicking on the plot can give you an improved plot of the same data. For example:



This example illustrates that care should be taken in interpreting the shape of histograms based upon small samples of data. With larger datasets, changing the number and width of bins in the histogram (as here from 5 years to 10 years) has less effect.

Exercise 3



The Barthel scores at admission have symmetric distributions, so assumptions of the t-test are reasonable. There is no evidence of any difference in spread.

T-Test

[DataSet1] [illegible]\SPSStrainer2002\SPSSpack.sav

Group Statistics

	RANDOMISED TO HW OR HAH	N	Mean	Std. Deviation	Std. Error Mean
ADMISSION BARTHEL	HOSPITAL AT HOME	[illegible]	10.9[illegible]	[illegible].580	[illegible]
	HOSPITAL WARD	11	9.73	[illegible]	.025

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
ADMISSION BARTHEL	Equal variances assumed	1.933	.168	[illegible]	[illegible]	[illegible]	1.190	[illegible]	-1.40[illegible]	[illegible]
	Equal variances not assumed			.909	73.[illegible]	[illegible]	1.198	[illegible]	[illegible]	[illegible]

The difference in mean admission Barthel score between the groups is 1.2, indicating that the group randomised to Hospital-at-Home was, on average, slightly more physically able.
SPSS tests for equality of variances and finds no evidence of a difference (p = 0.168)
The 95% confidence interval for the difference between the groups is (–1.41 to 3.80).
and t=1.933 , with p = 0.363.

From this we conclude that there is no evidence of a statistically significant difference in mean Barthel Scores at admission between the two groups although the true difference is likely to lie somewhere between 3.8 points higher in the Hospital-at-Home group and 1.5 points higher in the hospital ward group.

Note on Statistical Tests for Baseline Comparisons

The lack of difference between the treatment groups at baseline is not surprising, as patients were randomised to two groups that should have approximately equal characteristics. Significance tests comparing groups at baseline are sometimes used to test whether randomisation has been successful in producing balanced groups. However, there is increasing agreement between statisticians that this is not the best approach. We feel that the decision on whether there is an imbalance between the two treatment groups is better taken by inspecting for any clinically significant imbalance and considering the likely impact of this on the outcome.

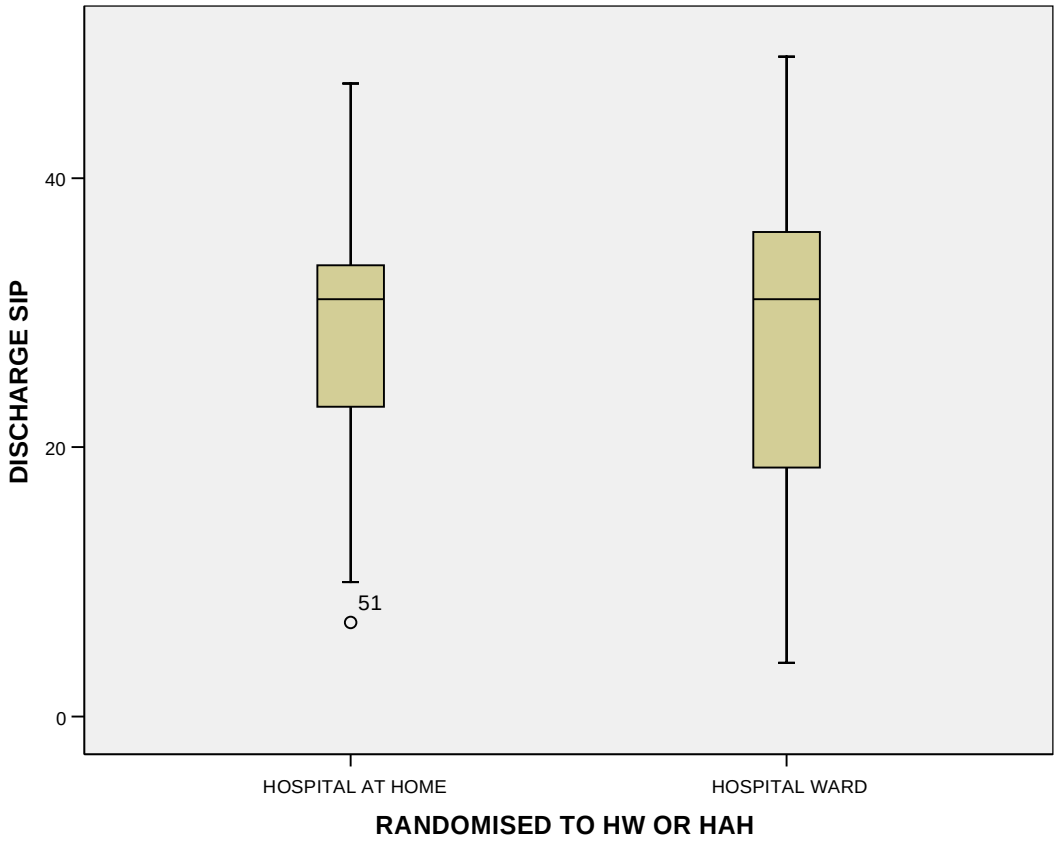
Exercise 4

Explore

RANDOMISED TO HW OR HAH

Case Processing Summary

	RANDOMISED TO HW OR HAH	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
DISCHARGE SIP	HOSPITAL AT HOME	36	52.9%	32	47.1%	68	100.0%
	HOSPITAL WARD	31	46.3%	36	53.7%	67	100.0%



Again there is some evidence of negative skew. There may also be evidence that the Hospital-at-Home group has a smaller variance than the hospital ward group.

T-Test

Group Statistics

	RANDOMISED TO HW OR HAH	N	Mean	Std. Deviation	Std. Error Mean
DISCHARGE SIP	HOSPITAL AT HOME	36	28.83	9.23	1.54
	HOSPITAL WARD	31	28.10	12.43	2.23

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
DISCHARGE SIP	Equal variances assumed	4.954	.030	.278	65	.782	.74	2.65	-4.56	6.03
	Equal variances not assumed			.272	54.667	.787	.74	2.71	-4.70	6.17

The confidence interval and p-value you report will depend on whether you assume that the two groups come from populations with equal variances or not.

As you will see Levene's test provides evidence that the variances are not equal, p = 0.03. However, you need to be careful when interpreting this test as, like all hypothesis tests, it is dependent on the size of your sample as well as the difference between the two groups.

In this case the choice of test makes no difference to the conclusion of no evidence for a difference between the groups, and little difference to the size of the confidence interval.

Equal variances: Difference in means = 0.74 (95% CI: -4.56 to 6.03), p = 0.78

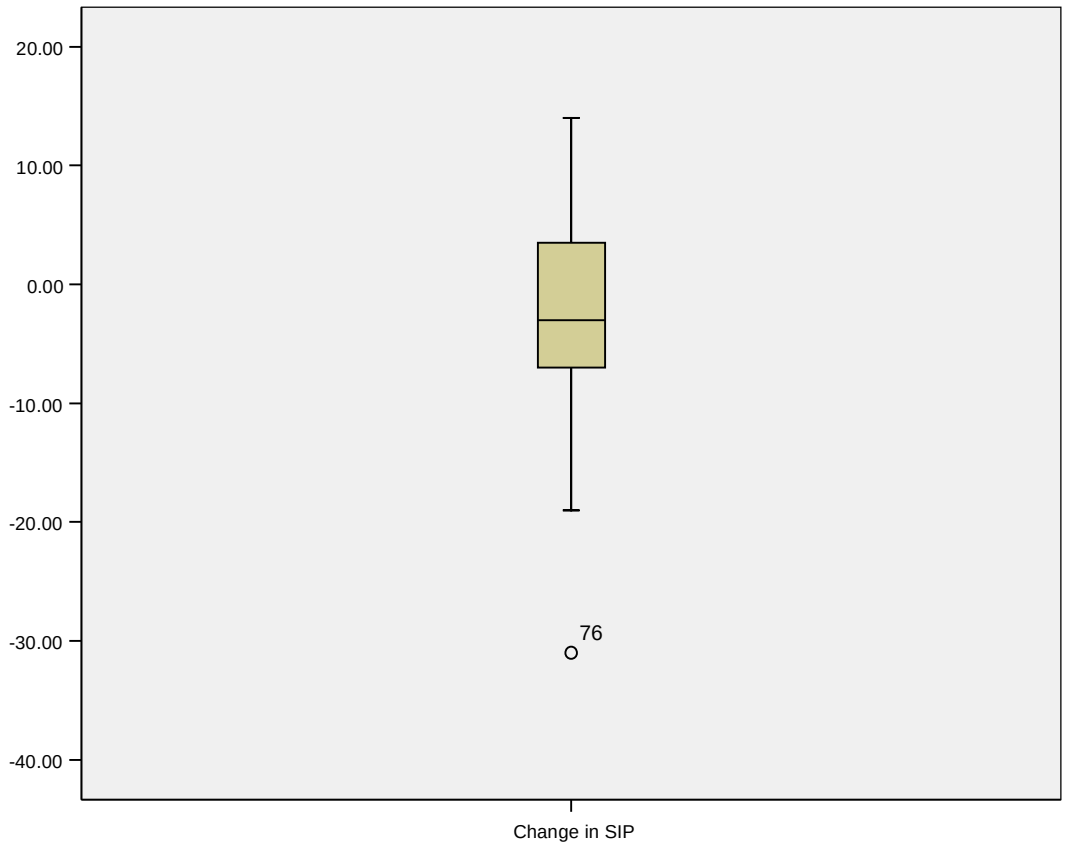
Unequal variances: Difference in means = 0.74 (95% CI: -4.70 to 6.17), p = 0.79

Exercise 5

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Change in SIP	60	44.4%	75	55.6%	135	100.0%



The boxplot appears to support the assumption that the differences are approximately Normally distributed, although the one patient with extreme improvement may give some cause for query. (See Page 21 for a discussion of outliers.)

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	ADMISSION SIP	30.42	60	9.99	1.29
	DISCHARGE SIP	27.97	60	10.86	1.40

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	ADMISSION SIP & DISCHARGE SIP	60	.664	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	ADMISSION SIP - DISCHARGE SIP	2.45	8.59	1.11	.23	4.67	2.210	59	.031

There is evidence at the 5% significance level that SIP fell during the study.

Mean Difference [Admission-Discharge] = 2.45 (95% CI: 0.23 to 4.67), p = 0.031

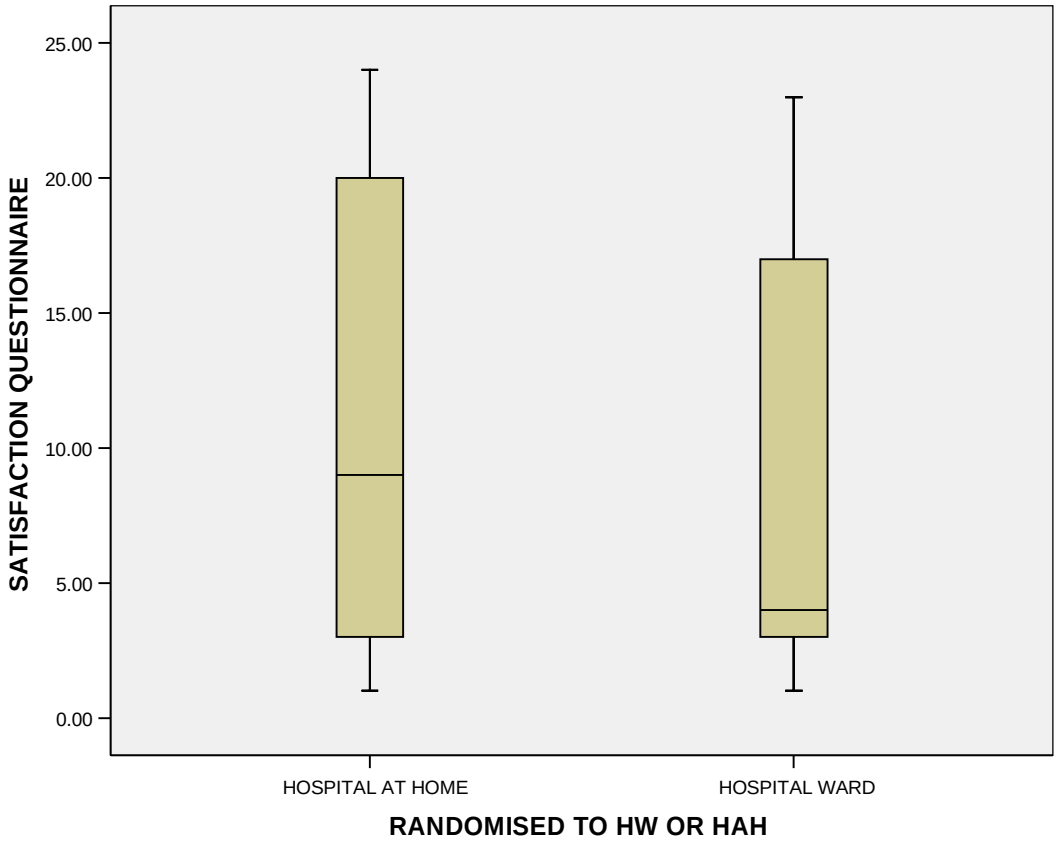
Exercise 6

Explore

RANDOMISED TO HW OR HAH

Case Processing Summary

	RANDOMISED TO HW OR HAH	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
satisf2	HOSPITAL AT HOME	63	92.6%	5	7.4%	68	100.0%
	HOSPITAL WARD	57	85.1%	10	14.9%	67	100.0%



The boxplots show that the patient satisfaction scores are positively skewed. This is particularly obvious in the hospital ward group. In this case a non-parametric test is clearly the most suitable choice.

NPar Tests

Mann-Whitney Test

Ranks

	RANDOMISED TO HW OR HAH	N	Mean Rank	Sum of Ranks
satisf2	HOSPITAL AT HOME	63	65.83	4147.50
	HOSPITAL WARD	57	54.61	3112.50
	Total	120		

Test Statistics[a]

	satisf2
Mann-Whitney U	1459.500
Wilcoxon W	3112.500
Z	-1.775
Asymp. Sig. (2-tailed)	.076

a. Grouping Variable: RANDOMISED TO HW OR HAH

The evidence for a difference between the two groups is marginal: p = 0.076. There may be some very weak evidence that the patients in the Hospital-at-Home group have higher satisfaction scores than the hospital ward patients.

Exercise 7

NPar Tests

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
DISCHARGE SIP - ADMISSION SIP	Negative Ranks	37[a]	28.36	1049.50
	Positive Ranks	19[b]	28.76	546.50
	Ties	4[c]		
	Total	60		

a. DISCHARGE SIP < ADMISSION SIP

b. DISCHARGE SIP > ADMISSION SIP

c. ADMISSION SIP = DISCHARGE SIP

Test Statistics[b]

	DISCHARGE SIP - ADMISSION SIP
Z	-2.054[a]
Asymp. Sig. (2-tailed)	.040

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

There is evidence that there is a change in SIP over the course of the trial p = 0.040.

This is consistent with the result from Exercise 5, although the p-value is slightly larger for the distribution-free test.

Exercise 8

RANDOMISED TO HW OR HAH * DIED Crosstabulation

			DIED		
			.00	1.00	Total
RANDOMISED TO HW OR HAH	HOSPITAL AT HOME	Count	39	6	45
		% within RANDOMISED TO HW OR HAH	86.7%	13.3%	100.0%
	HOSPITAL WARD	Count	34	6	40
		% within RANDOMISED TO HW OR HAH	85.0%	15.0%	100.0%
Total		Count	73	12	85
		% within RANDOMISED TO HW OR HAH	85.9%	14.1%	100.0%

Relative Risk = $\frac{6/45}{6/40} = \frac{6\times 40}{6\times 45} = \frac{40}{45} = \frac{8}{9} = 0.889$

(Or approximately: **Relative Risk =** $\frac{13.3}{15.0} = 0.887 \approx 0.889$**)**

Exercise 9

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEX * DIED	85	63.0%	50	37.0%	135	100.0%

SEX * DIED Crosstabulation

			DIED		
			.00	1.00	Total
SEX	MALE	Count	26	3	29
		% within SEX	89.7%	10.3%	100.0%
	FEMALE	Count	47	9	56
		% within SEX	83.9%	16.1%	100.0%
Total		Count	73	12	85
		% within SEX	85.9%	14.1%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.517[b]	1	.472		
Continuity Correction[a]	.152	1	.696		
Likelihood Ratio	.540	1	.462		
Fisher's Exact Test				.744	.357
Linear-by-Linear Association	.511	1	.475		
N of Valid Cases	85				

a. Computed only for a 2x2 table

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.09.

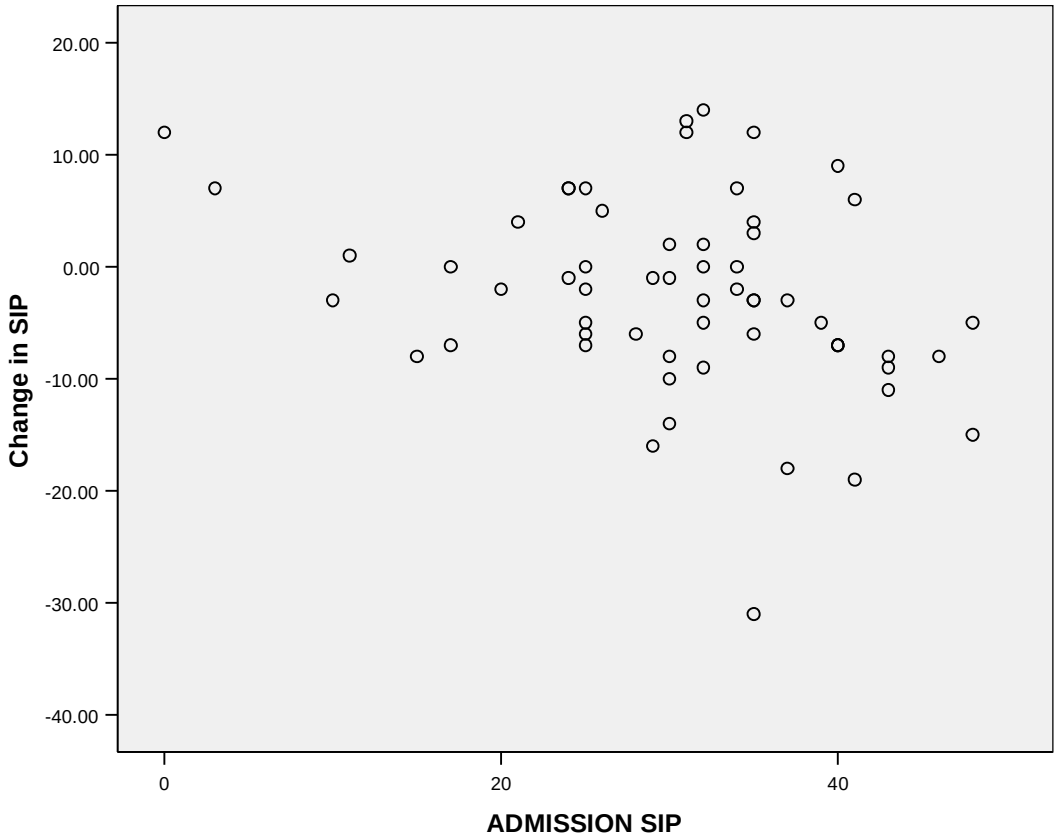
Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for SEX (MALE / FEMALE)	1.660	.413	6.674
For cohort DIED = .00	1.068	.903	1.264
For cohort DIED = 1.00	.644	.189	2.196
N of Valid Cases	85		

A possible solution is:

There was no evidence for a difference in mortality between male and female patients. The estimated relative risk of mortality in males compared to females was 0.64 (95% CI: 0.18 to 2.20). As only a small number of deaths were observed, the estimated confidence interval is wide. Lack of association between sex and mortality is confirmed by Fisher's Exact test (p=0.74).

Exercise 10



The scatter plot suggests that there may be evidence of a negative linear relationship between SIP at admission and the change observed during the trial. The patient with a change of -30 appears unusual and it may be worth checking this value. In general, people experiencing greater sickness impact at admission improved more. The correlation coefficient of -0.324 is statistically significant (p=0.012), confirming that there is a negative correlation between change in SIP and SIP at admission.

As with the previous example, this correlation is expected, as these are measures on the same patient, and change scores are always correlated with scores at baseline. In addition, there is probably a ceiling effect with the Barthel scores, meaning that those with higher function at baseline are likely to have a smaller improvement. This example is a good illustration of how interpreting a correlation coefficient may not be straightforward.