CHAPTER 1
Creating an SPSS data file and preparing to analyse the data

There are two kinds of statistics, the kind you look up and the kind you make up.
REX STOUT

Objectives

The objectives of this chapter are to explain how to:
- create an SPSS data file that will facilitate straightforward statistical analyses
- ensure data quality
- manage missing data points
- move data and output between electronic spreadsheets
- manipulate data files and variables
- devise a data management plan
- select the correct statistical test
- critically appraise the quality of reported data analyses

1.1 Creating an SPSS data file

Creating a data file in SPSS and entering the data is a relatively simple process. In the SPSS window located on the top left-hand side of the screen is a menu bar with headings and drop-down options. A new file can be opened using the *File* → *New* → *Data* commands located on the top left-hand side of the screen. The SPSS IBM Statistics Data Editor has two different screens called the 'Data View' and 'Variable View'. You can easily move between the two views by clicking on the tabs located at the bottom left-hand side of the screen.

1.1.1 Variable View screen

Before entering data in Data View, the features or attributes of each variable need to be defined in Variable View. In this screen, details of the variable names, variable types and labels are stored. Each row in Variable View represents a new variable and each

Medical Statistics: A Guide to SPSS, Data Analysis and Critical Appraisal, Second Edition.
Belinda Barton and Jennifer Peat.

Companion website: www.wiley.com/go/barton/medicalstatistics2e

column represents a feature of the variable such as type (e.g. numeric, dot, string, etc.) and measure (scale, ordinal or nominal). To enter a variable name, simply type the name into the first field and default settings will appear for almost all of the remaining fields, except for *Label* and *Measure*.

The Tab, arrow keys or mouse can be used to move across the fields and change the default settings. In Variable View, the settings can be changed by a single click on the cell and then pulling down the drop box option that appears when you double click on the domino on the right-hand side of the cell. The first variable in a data set is usually a unique identification code or a number for each participant. This variable is invaluable for selecting or tracking particular participants during the data analysis process.

Unlike data in Excel spreadsheets, it is not possible to hide rows or columns in either Variable View or Data View in SPSS and therefore, the order of variables in the spreadsheet should be considered before the data are entered. The default setting for the lists of variables in the drop-down boxes that are used when running the statistical analyses are in the same order as the spreadsheet. It can be more efficient to place variables that are likely to be used most often at the beginning of the spreadsheet and variables that are going to be used less often at the end.

Variable names

Each variable name must be unique and must begin with an alphabetic character. Variable names are entered in the column titled *Name* displayed in Variable View. The names of variables may be up to 64 characters long and may contain letters, numbers and some non-punctuation symbols but should not end in an underscore or a full stop. Variable names cannot contain spaces although words can be separated with an underscore. Some symbols such as @, # or $ can be used in variable names but other symbols such as %, > and punctuation marks are not accepted. SPSS is case sensitive so capital and lower case letters can be used.

Variable type

In medical statistics, the most common types of data are numeric and string. Numeric refers to variables that are recorded as numbers, for example, 1, 115, 2013 and is the default setting in Variable View. String refers to variables that are recorded as a combination of letters and numbers, or just letters such as ‘male’ and ‘female’. However, where possible, variables that are a string type and contain important information that will be used in the data analyses should be coded as categorical variables, for example, by using 1= male and 2 = female. For some analyses in SPSS, only numeric variables can be used so it is best to avoid using string variables where possible.

Other data types are comma or dot. These are used for large numeric variables which are displayed with commas or periods delimiting every three places. Other options for variable type are scientific notation, date, dollar, custom currency and restricted numeric.

Width and decimals

The width of a variable is the number of characters to be entered for the variable. If the variable is numeric with decimal places, the total number of characters needs to include

the numbers, the decimal point and all decimal places. The default setting is 8 characters which is sufficient for numbers up to 100,000 with 2 decimal places.

Decimals refers to the number of decimal places that will be displayed for a numeric variable. The default setting is two decimal places, that is, 51.25. For categorical variables, no decimal places are required. For continuous variables, the number of decimal places must be the same as the number that the measurement was collected in. The decimal setting does not affect the statistical calculations but does influence the number of decimal places displayed in the output.

Labels

Labels can be used to name, describe or identify a variable and any character can be used in creating a label. Labels may assist in remembering information about a variable that is not included in the variable name. When selecting variables for analysis, variables will be listed by their variable label with the variable name in brackets in the dialogue boxes. Also, output from SPSS will list the variable label. Therefore, it is important to keep the length of the variable label short where possible. For example, question one of a questionnaire is 'How many hours of sleep did you have last night?'. The variable name could be entered as q1 (representing question 1) and the label to describe the variable q1 could be 'hrs sleep'. If many questions begin with the same phrase, it is helpful to include the question number in the variable label, for example, 'q1: hrs sleep'.

Values

Values can be used to assign labels to a variable, which makes interpreting the output from SPSS easier. Value labels are most commonly used when the variable is categorical or nominal. For example, a label could be used to code 'Gender' with the label 'male' coded to a value of 1 and the label 'female' coded to a value of 2. The SPSS dialogue box *Value Labels* can be obtained by single clicking on the Values box, then clicking on the grey domino on the right-hand side of the box. Within this box, the buttons *Add*, *Change* and *Remove* can be used to customize and edit the value labels.

Missing

Missing can be used to assign user system missing values for data that are not available for a participant. For example, a participant who did not attend a scheduled clinical appointment would have data values that had not been measured and which are called missing values. Missing values are not included in the data analyses and can sometimes create pervasive problems. The seriousness of the problem depends largely on the pattern of missing data, how much is missing and why it is missing.[1]

For a full stop to be recognized as a system missing value, the variable type must be entered as numeric rather than a string variable. Other approaches to dealing with missing data will be discussed later in this chapter.

Columns and align

Columns can be used to define the width of the column in which the variable is displayed in the Data View screen. The default setting is 8 and this is generally sufficient to view

the name in the Variable View and Data View screens. Align can be used to specify the alignment of the data information in Data View as either right, left or centre justified within cells.

Measure

In SPSS, the measurement level of the variable can be classified as nominal, ordinal or scale under the *Measure* option. The measurement scales used which are described below determine each of these classifications.

Nominal variables

Nominal scales have no order and are generally categories with labels that have been assigned to classify items or information. For example, variables with categories such as male or female, religious status or place of birth are nominal scales. Nominal scales can be string (alphanumeric) values or numeric values that have been assigned to represent categories, for example 1 = male and 2 = female.

Ordinal variables

Values on an ordinal scale have a logical or ordered relationship across the values and it is possible to measure some degree of difference between categories. However, it is usually not possible to measure a specific amount of difference between categories. For example, participants may be asked to rate their overall level of stress on a five-point scale that ranges from no stress, mild, moderate, severe or extreme stress. Using this scale, participants with severe stress will have a more serious condition than participants with mild stress, although recognizing that self-reported perception of stress may be subjective and is unlikely to be standardized between participants. With this type of scale, it is not possible to say that the difference between mild and moderate stress is the same as the difference between moderate and severe stress. Thus, information from these types of variables has to be interpreted with care.

Scale variables

Variables with numeric values that are measured by an interval or ratio scale are classified as scale variables. On an interval scale, one unit on the scale represents the same magnitude across the whole scale. For example, Fahrenheit is an interval scale because the difference in temperature between 10 °F and 20 °F is the same as the difference in temperature between 40 °F and 50 °F. However, interval scales have no true zero point. For example, 0 °F does not indicate that there is no temperature. Because interval scales have an arbitrary rather than a true zero point, it is not possible to compare ratios.

A ratio scale has the same properties as ordinal and interval scales, but has a true zero point and therefore ratio comparisons are valid. For example, it is possible to say that a person who is 40 years old is twice as old as a person who is 20 years old and that a person is 0 years old at birth. Other common ratio scales are length, weight and income.

Role

Role can be used with some SPSS statistical procedures to select variables that will be automatically assigned a role such as input or target. In Data View, when a statistical procedure is selected from *Analyze* a dialogue box opens up and variables to be analysed must be selected such as an independent or dependent variable. If the role of the variables has been defined in Variable View, the variables will be automatically displayed in the destination list of the dialogue box. Role options for a variable are input (independent variable), target (dependent variable), both (can be an input or an output variable), none (no role assignment), partition (to divide the data into separate samples) and split (this option is only used in SPSS Modeler). The default setting for *Role* is input.

1.1.2 Saving the SPSS file

After the information for each variable has been defined, the variable details entered in the Variable View screen can be saved using the commands shown in Box 1.1. When the file is saved, the name of the file will replace the word *Untitled* at the top left-hand side of the Data View screen. The data can then be entered in the Data View screen and also saved using the commands shown in Box 1.1. The data file extension is *.sav*. When there is only one data file open in the Data Editor, the file can only be closed by exiting the SPSS program. When there is more than one data file open, the SPSS commands *File* → *Close* can be used to close a data file.

Box 1.1 SPSS commands for saving a file

SPSS Commands

Untitled – SPSS IBM Statistics Data Editor
File → Save As
Save Data As
Enter the name of the file in File name
Click on Save

1.1.3 Data View screen

The Data View screen displays the data values and is similar to many other spreadsheet packages. In general, the data for each participant should occupy one row only in the spreadsheet. Thus, if follow-up data have been collected from the participants on one or more occasions, the participants' data should be an extension of their baseline data row and not a new row in the spreadsheet. However, this does not apply for studies in which controls are matched to cases by characteristics such as gender or age or are selected as the unaffected sibling or a nominated friend of the case and therefore the data are paired. The data from matched case–control studies are used as pairs in the statistical analyses and therefore it is important that matched controls are not entered on a separate row

but are entered into the same row in the spreadsheet as their matched case. This method will inherently ensure that paired or matched data are analysed correctly and that the assumptions of independence that are required by many statistical tests are not violated. Thus, in Data View, each column represents a separate variable and each row represents a single participant, or a single pair of participants in a matched case–control study, or a single participant with follow-up data. This data format is called 'wide format'. For some longitudinal modelling analyses, the data may need to be changed to 'long format', that is, each time a point is represented on a separate row. This is discussed in Chapter 6.

In Data View, data can be entered and the mouse, tab, enter or cursor keys can be used to move to another cell of the data sheet. In Data View, the value labels button which is displayed at the top of the spreadsheet (17th icon from the left-hand side), with an arrow pointing to '1' and another arrow pointing to 'A' can be used to switch between displaying the values or the value labels that have been entered.

1.2 Opening data from Excel in SPSS

Data can be entered in other programs such as Excel and then imported into the SPSS Data View sheet. Many researchers use Excel or Access for ease of entering and managing the data. However, statistical analyses are best executed in a specialist statistical package such as SPSS in which the integrity and accuracy of the statistics are guaranteed.

Opening an Excel spreadsheet in SPSS can be achieved using the commands shown in Box 1.2. In addition, specialized programs are available for transferring data between different data entry and statistics packages (see Section Useful Websites).

Box 1.2 SPSS commands for opening an Excel data file

SPSS Commands

Untitled – SPSS IBM Statistics Data Editor
File → Open → Data
Open Data
    Click on Files of type to show Excel (.xls, *.xlsx,*xlsm)*
    Look in: find and click on your Excel data file
    Click Open
Opening Excel Data Source
    Check that the correct Worksheet within the file is selected
    Tick 'Read variable names from the first row of data' (default setting)
    Click OK

If data are entered in Excel or another database before being exported into SPSS, it is a good idea to use variable names that are accepted by SPSS to avoid having to rename the variables. For numeric values, blank cells in Excel are converted to the system missing value, that is a full stop, in SPSS.

Once in the SPSS spreadsheet, features of the variables can be adjusted in Variable View, for example, by changing column widths, entering the labels and values for categorical variables and checking that the number of decimal places is appropriate for each

variable. Once data quality is ensured, a back-up copy of the database should be archived at a remote site for safety. Few researchers need to resort to their archived copies but, when they do, they are an invaluable resource.

The spreadsheet that is used for data analyses should not contain any information that would contravene ethics guidelines by identifying individual participants. In the working data file, names, addresses and any other identifying information that will not be used in data analyses should be removed. Identifying information that is required can be recoded and de-identified, for example, by using a unique numerical value that is assigned to each participant.

1.3 Categorical and continuous variables

While variables in SPSS can be classified as scale, ordinal or nominal values, a more useful classification for variables when deciding how to analyse data is as categorical variables (ordered or non-ordered) or continuous variables (scale variables). These classifications are essential for selecting the correct statistical test to analyse the data and are not provided in Variable View by SPSS. Categorical variables have discrete categories, such as male and female, and continuous variables are measured on a scale, such as height which is measured in centimetres.

Categorical values can be non-ordered or ordered. For example, gender which is coded as 1 = male and 2 = female and place of birth which is coded as 1 = local, 2 = regional and 3 = overseas are non-ordered variables. Categorical variables can also be ordered, for example, if the continuous variable length of stay was recoded into categories of 1 = 1–10 days, 2 = 11–20 days, 3 = 21–30 days and 4 = >31 days, there is a progression in magnitude of length of stay. A categorical variable with only two possible outcomes such as yes/no or disease present/disease absent is referred to as a binary variable.

1.4 Classifying variables for analyses

Before conducting any statistical tests, a formal, documented plan that includes a list of hypotheses to be tested and identifies the variables that will be used should be drawn up. For each question, a decision on how each variable will be used in the analyses, for example, as a continuous or categorical variable or as an outcome or explanatory variable, should be made.

Table 1.1 shows a classification system for variables and how the classification influences the presentation of results. An outcome or dependent variable is a variable is generally the outcome of interest in the study that has been measured, for example, cholesterol levels or blood pressure may be measured in a study to reduce cardiovascular risk. An outcome variable is proposed to be changed or influenced by an explanatory variable. An explanatory or independent variable is hypothesized to affect the outcome variable and is generally manipulated or controlled experimentally. For example, treatment status defined as whether participants receive the active drug treatment or inactive treatment (placebo) is an independent variable.

A common error in statistical analyses is to misclassify the outcome variable as an explanatory variable or to misclassify an intervening variable as an explanatory variable. It is important that an intervening variable, which links the explanatory and outcome

Table 1.1 Names used to identify variables

Variable name	Alternative name/s	Axis for plots, data analysis and tables
Outcome variables	Dependent variables (DVs)	*y*-axis, columns
Intervening variables	Secondary or alternative outcome variables	*y*-axis, columns
Explanatory variables	Independent variables (IVs) Risk factors Exposure variables Predictors	*x*-axis, rows

variable because it is directly on the pathway to the outcome variable, is not treated as an independent explanatory variable in the analyses.[2] It is also important that an alternative outcome variable is not treated as an independent risk factor. For example, hay fever cannot be treated as an independent risk factor for asthma because it is a symptom that is a consequence of the same allergic developmental pathway.

In part, the classification of variables depends on the study design. In a case–control study in which disease status is used as the selection criterion, the explanatory variable will be the presence or absence of disease and the outcome variable will be the exposure. However, in most other observational and experimental studies such as clinical trials, cross-sectional and cohort studies, the disease will be the outcome and the exposure or the experimental group will be an explanatory variable.

1.5 Hypothesis testing and *P* values

Most medical statistics are based on the concept of hypothesis testing and therefore an associated *P* value is usually reported. In hypothesis testing, a 'null hypothesis' is first specified, that is a hypothesis stating that there is no difference, for example, there is no difference in the summary statistics of the study groups (placebo and treatment). The null hypothesis assumes that the groups that are being compared are drawn from the same population. An alternative hypothesis, which states that there is a difference between groups, can also be specified. The *P* value is then calculated, that is, the probability of obtaining a difference as large as or larger than the one observed between the groups, assuming the null hypothesis is true (i.e. no difference between groups).

A *P* value of less than 0.05, that is a probability of less than 1 chance in 20, is usually accepted as being statistically significant. If a *P* value is less than 0.05, we accept that it is unlikely that a difference between groups has occurred by chance if the null hypothesis was true. In this situation, we reject the null hypothesis and accept the alternative hypothesis, and therefore conclude that there is a statistically significant difference between the groups. On the other hand, if the *P* value is greater than or equal to 0.05 and therefore the probability with which the test statistic occurs is greater than 1 chance in 20, we accept that the difference between groups has occurred by chance. In this case, we accept the null hypothesis and conclude that the difference is not attributed to sampling.

In accepting or rejecting a null hypothesis, it is important to remember that the *P* value only provides a probability value and does not provide absolute proof that the null hypothesis is true or false. A *P* value obtained from a test of significance should only be interpreted as a measure of the strength of evidence against the null hypothesis. The smaller the *P* value the stronger the evidence against the null hypothesis.

1.6 Choosing the correct statistical test

Selecting the correct test to analyse data depends not only on the study design but also on the nature of the variables collected. Tables 1.2–1.5 show the types of tests that can be selected based on the nature of variables. It is of paramount importance that the correct test is used to generate *P* values and to estimate a size of effect. Using an incorrect test will inviolate the statistical assumptions of the test and may lead to inaccurate or biased *P* values.

Table 1.2 Choosing a statistic when there is one outcome variable only

Type of variable	Number of times measured in each participant	Statistic	SPSS menu
Binary	Once	Incidence or prevalence and 95% confidence interval (95% CI)	Descriptive statistics; Frequencies
	Twice	McNemar's chi-square; Kappa	Descriptive statistics; Crosstabs
Continuous	Once	Tests for normality	Non-parametric tests; One sample; Kolmogorov–Smirnov Descriptive statistics; Explore; Plots; Normality plots with tests
		One-sample *t*-test	Compare means; One-sample *t*-test
		Mean, standard deviation (SD) and 95% CI	Descriptive statistics; Explore
		Median (Mdn) and inter-quartile (IQR) range	Descriptive statistics; Explore
	Twice	Paired *t*-test	Compare means; Paired-samples *t*-test
		Mean difference and 95% CI	Compare means; Paired-samples *t*-test
		Measurement error	Compare means; Paired-samples *t*-test
		Mean-versus-differences plot	Graphs; Legacy Dialogs; Scatter/Dot
		Intra-class correlation coefficient	Scale; Reliability Analysis
	Three or more	Repeated measures ANOVA	General linear model; Repeated measures

Table 1.3 Choosing a statistic when there is one outcome variable and one explanatory variable

Type of outcome variable	Type of explanatory variable	Number of levels of the categorical variable	Statistic	SPSS menu
Categorical	Categorical	Both variables are binary	Chi-square	Descriptive statistics; Crosstabs
			Odds ratio or relative risk	Descriptive statistics; Crosstabs
			Logistic regression	Regression; Binary logistic
			Sensitivity and specificity	Descriptive statistics; Crosstabs
			Likelihood ratio	Descriptive statistics; Crosstabs
		At least one of the variables has more than two levels	Chi-square	Descriptive statistics; Crosstabs
			Chi-square trend	Descriptive statistics; Crosstabs
			Kendall's correlation	Correlate; Bivariate
Categorical	Continuous	Categorical variable is binary	ROC curve	ROC curve
			Survival analyses	Survival; Kaplan–Meier
		Categorical variable is multi-level and ordered	Spearman's correlation coefficient	Correlate; Bivariate
Continuous	Categorical	Explanatory variable is binary	Independent samples *t*-test	Compare means; Independent samples *t*-test
			Mean difference and 95% CI	Compare means; Independent samples *t*-test
		Explanatory variable has three or more categories	Analysis of variance	Compare means; One-way ANOVA
Continuous	Continuous	No categorical variables	Regression	Regression; Linear
			Pearson's correlation	Correlate; Bivariate

1.7 Sample size requirements

The sample size is one of the most critical issues in designing a research study because it affects all aspects of interpreting the results. The sample size needs to be large enough so that a definitive answer to the research question is obtained. This will help to ensure

Table 1.4 Choosing a statistic for one or more outcome variables and more than one explanatory variable

Type of outcome variable/s	Type of explanatory variable/s	Number of levels of categorical variable	Statistic	SPSS menu
Continuous – only one outcome	Both continuous and categorical	Categorical variables are binary	Multiple regression	Regression; Linear
	Categorical	At least one of the explanatory variables has three or more categories	Two-way analysis of variance	General linear model; Univariate
	Both continuous and categorical	One categorical variable has two or more levels	Analysis of covariance	General linear model; Univariate
Continuous – outcome measured more than once	Both continuous and categorical	Categorical variables can have two or more levels	Repeated measures analysis of variance	General linear model; Repeated measures Mixed Models; Linear
No outcome variable	Both continuous and categorical	Categorical variables can have two or more levels	Factor analysis	Dimension reduction: Factor

Table 1.5 Parametric and non-parametric equivalents

Parametric test	Non-parametric equivalent	SPSS menu
Mean and standard deviation	Median and inter-quartile range	Descriptive statistics; Explore
Pearson's correlation coefficient	Spearman's or Kendall's correlation coefficient	Correlate; Bivariate
One-sample sign test	Wilcoxon signed rank test	Non-parametric tests; One Sample
Two sample *t*-test	Sign test or Wilcoxon matched pair signed rank test	Non-parametric tests; Related samples
Independent *t*-test	Mann–Whitney U	Non-parametric tests; Independent samples
Analysis of variance	Kruskall–Wallis one-way test	Non-parametric tests; Independent samples
Repeated measures analysis of variance	Friedman's two-way ANOVA test	Non-parametric tests; Related samples

generalizability of the results and precision around estimates of effect. However, the sample has to be small enough so that the study is practical to conduct. In general, studies with a small sample size, say with less than 30 participants, can usually only provide imprecise and unreliable estimates.

P values are strongly influenced by the sample size. The larger the sample size the more likely a difference between study groups will be statistically significant. Box 1.3 provides a definition of type I and type II errors and shows how the size of the sample can contribute to these errors, both of which have a profound influence on the interpretation of the results. In addition, type I and II error rates are inversely related because both are influenced by sample size – when the risk of a type I error is reduced, the risk of a type II error is increased. Therefore, it is important to carefully calculate the sample size required prior to the study commencing and also consider the sample size when interpreting the results of the statistical tests.

Box 1.3 Type I and type II errors

Type I errors

- are false positive results
- occur when a statistical significant difference between groups is found but no clinically important difference exists
- the null hypothesis is rejected in error
- usually occur when the sample size is very large

Type II errors

- are false negative results
- a clinical important difference between groups does exist but does not reach statistical significance
- the null hypothesis is accepted in error
- usually occur when the sample size is small

1.8 Study handbook and data analysis plan

The study handbook should be a formal documentation of all of the study details that is updated continuously with any changes to protocols, management decisions, minutes of meetings and so on. This handbook should be available for anyone in the team to refer to at any time to facilitate considered data collection and data analysis practices. Suggested contents of data analysis log sheets that could be kept in the study handbook are shown in Box 1.4.

Box 1.4 Data analysis log sheets

Data analysis log sheets should contain the following information:

- Title of proposed paper, report or abstract
- Author list and author responsible for data analyses and documentation
- Specific research questions to be answered or hypotheses to be tested
- Outcome and explanatory variables to be used
- Statistical methods
- Details of database location and file storage names
- Journals and/or scientific meetings where results will be presented

Data analyses must be planned and executed in a logical and considered sequence to avoid errors or misinterpretation of results. In this, it is important that data are treated carefully and analysed by people who are familiar with their content, their meaning and the interrelationship between variables.

Before beginning any statistical analyses, a data analysis plan should be agreed upon in consultation with the study team. The plan can include the research questions or hypotheses that will be tested, the outcome and explanatory variables that will be used, the journal where the results will be published and/or the scientific meeting where the findings will be presented.

A good way to handle data analyses is to create a log sheet for each proposed paper, abstract or report. The log sheets should be formal documents that are agreed to by all stakeholders and that are formally archived in the study handbook. When a research team is managed efficiently, a study handbook is maintained that has up-to-date documentation of all details of the study protocol and the study processes.

1.9 Documentation

Documentation of data analyses, which allows anyone to track how the results were obtained from the data set collected, is an important aspect of the scientific process. This is especially important when the data set will be accessed in the future by researchers who are not familiar with all aspects of data collection or the coding and recoding of the variables.

Data management and documentation are relatively mundane processes compared to the excitement of statistical analyses but are essential. Laboratory researchers document every detail of their work as a matter of course by maintaining accurate laboratory books. All researchers undertaking clinical and epidemiological studies should be equally diligent and document all of the steps taken to reach their conclusions.

Documentation can be easily achieved by maintaining a data management book with a log sheet for each data analysis. In this, all steps in the data management processes are recorded together with the information of names and contents of files, the coding and names of variables and the results of the statistical analyses. Many funding bodies and ethics committees require that all steps in data analyses are documented and that in addition to archiving the data, the data sheets, the output files and the participant records are kept for 5 years or up to 15 years after the results are published.

1.10 Checking the data

Prior to beginning statistical analysis, it is essential to have a thorough working knowledge of the nature, ranges and distributions of each variable. Although it may be tempting to jump straight into the analyses that will answer the study questions rather than spend time obtaining descriptive statistics, a working knowledge of the descriptive statistics often saves time by avoiding analyses having to be repeated for example because outliers, missing values or duplicates have not been addressed or groups with small numbers are not identified.

When entering data, it is important to crosscheck the data file with the original records to ensure that data has been entered correctly. It is important to have a high standard of

data quality in research databases at all times because good data management practice is a hallmark of scientific integrity. The steps outlined in Box 1.5 will help to achieve this.

Box 1.5 Data organization

The following steps ensure good data management practices:
- Crosscheck data with the original records
- Use numeric codes for categorical data where possible
- Choose appropriate variable names and labels to avoid confusion across variables
- Check for duplicate records and implausible data values
- Make corrections
- Archive a back-up copy of the data set for safe keeping
- Limit access to sensitive data such as names and addresses in working files

It is especially important to know the range and distribution of each variable and whether there are any outliers or extreme values (see Chapter 2) so that the statistics that are generated can be explained and interpreted correctly. Describing the characteristics of the sample also allows other researchers to judge the generalizability of the results. A considered pathway for data management is shown in Box 1.6.

Box 1.6 Pathway for data management before beginning statistical analysis

The following steps are essential for efficient data management:
- Obtain the minimum and maximum values and the range of each variable
- Conduct frequency analyses for categorical variables
- Use box plots, histograms and other tests to ascertain normality of continuous variables
- Identify and deal with missing values and outliers
- Recode or transform variables where necessary
- Rerun frequency and/or distribution checks
- Document all steps in a study handbook

1.11 Avoiding and replacing missing values

Missing values must be omitted from the analyses and not inadvertently included as data points. This can be achieved by proper coding that is recognized by SPSS as a system missing value. The default character to indicate a missing value is a full stop. This is preferable to using an implausible value such as 9 or 999 which was commonly used in the past. If these values are not accurately defined as discrete missing values in Missing column displayed in Variable View, they are easily incorporated into the analyses, thus producing erroneous results. Although these values can be predefined as system missing, this coding scheme is discouraged because it is inefficient, requires data analysts to be familiar with the coding scheme and has the potential for error. If missing values are

Table 1.6 Classification of variables in the file surgery.sav

Variable label	Type	SPSS measure	Classification for analysis decisions
ID	Numeric	Scale	Not used in analyses
Gender	String	Nominal	Categorical/non-ordered
Place of birth	String	Nominal	Categorical/non-ordered
Birth weight	Numeric	Scale	Continuous
Gestational age	Numeric	Ordinal	Continuous
Length of stay	Numeric	Scale	Continuous
Infection	Numeric	Scale	Categorical/non-ordered
Prematurity	Numeric	Scale	Categorical/non-ordered
Procedure performed	Numeric	Nominal	Categorical/non-ordered

required in an analysis, for example to determine if people with missing data have the characteristics as people with complete data, the missing values can easily be recoded to a numeric value in a new variable.

The file **surgery.sav**, which contains the data from 141 babies who underwent surgery at a paediatric hospital, can be opened using the *File* → *Open* → *Data* commands. The classification of the variables as shown by SPSS and the classifications that are needed for statistical analysis are shown in Table 1.6.

In the spreadsheet, the variable for 'place of birth' is coded as a string variable. The command sequences shown in Box 1.7 can be used to obtain frequency information of this variable, where L = local, O = overseas and R = regional.

Box 1.7 SPSS commands for obtaining frequencies

SPSS Commands

surgery.sav – SPSS IBM Statistics Data Editor
Analyze → Descriptive Statistics → Frequencies
Frequencies
Highlight Place of birth and click into Variable(s)
Click OK

Frequency table

Place of Birth

		Frequency	Per cent	Valid per cent	Cumulative per cent
Valid		9	6.4	6.4	6.4
	L	90	63.8	63.8	70.2
	0	9	6.4	6.4	76.6
	R	33	23.4	23.4	100.0
	Total	141	100.0	100.0	

The nine children with no information of birthplace are included in the valid and cumulative percentages shown in the Frequency table. If the variable had been defined as numeric, the missing values would have been omitted.

When collecting data in any study, it is essential to have methods in place to prevent missing values in, say, at least 95% of the data set. Methods such as restructuring questionnaires in which participants decline to provide sensitive information or training research staff to check that all fields are complete at the point of data collection are invaluable in this process. In large epidemiological and longitudinal data sets, some missing data may be unavoidable. However, in clinical trials, it may be unethical to collect insufficient information about some participants so that they have to be excluded from the final analyses.

If the number of missing values is small and the missing values occur randomly throughout the data set, the cases with missing values can be omitted from the analyses. This is the default option in most statistical packages and the main effect of this process is to reduce statistical power, that is the ability to show a statistically significant difference between groups when a clinically important difference exists. Missing values that are scattered randomly throughout the data are less of a problem than non-random missing values that can affect both the power of the study and the generalizability of the results. For example, if people in higher income groups selectively decline to answer questions about income, the distribution of income in the population will not be known and analyses that include income will not be generalizable to people in higher income groups. When analysing data, it is important to determine whether data is missing completely at random, missing at random, or missing not at random.[3]

In some situations, it may be important to replace a missing value with an estimated value that can be included in analyses. In longitudinal clinical trials, it has become common practice to use the last score obtained from the participant and carry it forward for all subsequent missing values – this is commonly called last observation carried forward (LOCF) or last value carried forward (LVCF). In other studies, a mean value (if the variable is normally distributed) or a median value (if the variable is non-normal distributed) may be used to replace missing values. This can be undertaken in SPSS using the commands *Transform* ⟶ *Replace Missing Values.*

These simple imputation solutions are not ideal – for example, LOCF can result in reducing the variance of the sample and can also lead to biased results.[3] Other more complicated methods for replacing missing values have been described and should be considered.[4]

1.12 SPSS data management capabilities

SPSS has many data management capabilities which are listed in the *Data* and *Transform* menus. A summary of some of the commands that are widely used in routine data analyses are shown in Table 1.7. How to use the commands to select a subset of variables and recode variables is shown below. Also, the use of these commands is demonstrated in more detail in the following chapters.

Table 1.7 Data management capabilities in SPSS

Menu	Command	Purpose
Data menu	Identify duplicate cases	Labels primary and duplicate cases as defined by specified variables.
	Sort cases	Sorts the data set into ascending or descending order using one or more variables.
	Merge files	Allows the merge of one or more data files using a variable common to both sets. Additional cases or variables can be added.
	Restructure	Changes data sets from wide format (one line per subject) to long format (multiple lines per subject, for example, when there are multiple time points), and vice versa.
	Split file	Separates the file into separate groups for analysis.
	Select cases	By using conditional expressions, a subgroup of cases can be selected for analysis based on one or more variables. Arithmetic operators and functions can also be used in the expression.
Transform menu	Compute	Creates a new variable based on transformation of existing variables or by using mathematical functions.
	Recode	Reassigns values or collapses ranges into the same or a new variable.
	Visual binning	A categorical variable can be created from a scale variable. Variables can be grouped into 'bins' or interval cut off points such as quantiles or tertiles.
	Date and time wizard	Used to create a date/time variable, to add or subtract dates and times. The result is presented in a new variable in units of seconds and needs to be back converted to hours or days.
	Replace missing values	Replaces missing values with an option to use various methods of imputation.

1.12.1 Using subsets of variables

When conducting an analysis, it is common to want to use only a few of the variables. A smaller subset of variables to be used in analyses can be selected as shown in Box 1.8. When using a subset, only the variables selected are displayed in the data analysis dialogue boxes. This can make analyses more efficient by avoiding having to search up and down a long list of variables for the ones required. Using a subset is especially useful when working with a large data set and only a few variables are needed for a current analysis. To see the whole data set again simply use the SPSS commands *Utilities → Show All Variables*.

Box 1.8 SPSS commands for defining variable sets

SPSS Commands

surgery.sav – IBM SPSS Statistics Data Editor
Utilities → Define Variable Sets
Define variable sets
Enter Set Name e.g. Set_1
Highlight Variables required in the analyses and click them into Variables in Set
Click on Add Set
Click Close
surgery.sav – IBM SPSS Statistics Data Editor
Utilities → Use Variable Sets
Click Uncheck all
Tick Set 1
Click OK

1.12.2 Recoding variables and using syntax

In the frequency statistics conducted above, place of birth was coded as a string variable and the missing values were treated as valid values and included in the summary statistics. To remedy this, the syntax shown in Box 1.9 can be used to recode place of birth from a string variable into a numeric variable.

Box 1.9 Recoding a variable into a different variable

SPSS Commands

surgery.sav – SPSS IBM Statistics Data Editor
Transform → Recode → Into Different Variables
Recode into Different Variables
Highlight Place of birth and click into Input Variable → Output Variable
Enter Output Variable Name as place2,
Enter Output Variable Label as Place of birth recoded/ Click Change
Click Old and New Values
Recode into Different Variables: Old and New Values
Old Value → Value=L, New Value → Value=1/Click Add
Old Value → Value =R, New Value → Value =2/Click Add
Old Value → Value=0, New Value → Value=3/Click Add
Click Continue
Recode into Different Variables
Click OK (or Paste/Run → All)

The paste command is a useful tool to provide automatic documentation of any changes that are made. The paste commands writes the recoding to a Syntax screen

that can be saved or printed for documentation and future reference. Using the *Paste* command for the above recode provides the following documentation.

```
RECODE place ('L'=1) ('R'=2) ('O'=3) INTO place2.
VARIABLE LABELS place2 'Place of birth recoded'.
EXECUTE.
```

After recoding, the value labels for the three new categories of place2 that have been created can be added in the Variable View window. In this case, place of birth needs to be defined as 1 = Local, 2 = Regional and 3 = Overseas. This can be added by clicking on the Values cell and then double clicking on the grey domino box on the right of the cell to add the value labels. Similarly, gender which is also a string variable can be recoded into a numeric variable (gender2) with Male = 1 and Female = 2. After recoding variables, it is important to check that the number of decimal places is correct. This process will ensure that the number of decimal places is appropriate on the SPSS data analysis output.

1.12.3 Dialog recall

A useful function in SPSS to repeat recently conducted commands is the *Dialog Recall* button. This button recalls the most recently used SPSS commands. The *Dialog Recall* button is the fourth icon at the top left-hand side of the Data or Variable View screen or the sixth icon in the top left-hand side of the SPSS Output Viewer screen.

Using the *Dialog Recall* button to obtain *Frequencies* for place 2, which is labelled 'Place of birth recoded', the following output is produced.

Frequencies

Place of Birth Recoded

		Frequency	Per cent	Valid per cent	Cumulative per cent
Valid	Local	90	63.8	68.2	68.2
	Regional	33	23.4	25.0	93.2
	Overseas	9	6.4	6.8	100.0
	Total	132	93.6	100.0	
Missing	System	9	6.4		
Total		141	100.0		

The frequency of place of birth shown in the Frequencies table show that the recoding sequence was executed correctly. When the data are recoded as numeric, the nine babies who have missing data for birthplace are correctly omitted from the valid and cumulative percentages.

1.12.4 Displaying names or labels

There are options to change how variables are viewed when running the analyses or how to display the variable names on SPSS output. By using the command *Edit* → *Options* → *General* you can select whether variables will be displayed by their variable names or

their labels in the dialog command boxes. There is also an option to select whether variables are presented in alphabetical order, in the order they are entered in the file or in measurement level. Under the command *Edit* → *Options* → *Output*, there are options to select whether the variable and variable names will be displayed as labels, values or both on the output.

1.13 Managing SPSS output

1.13.1 Formatting SPSS output

There are many output formats in SPSS. The format of the frequencies table obtained previously can easily be changed by double clicking on the table and using the commands *Format* → *TableLooks*. To obtain the output in the format below, which is a classical academic format with no vertical lines and minimal horizontal lines that is used by many journals, highlight *Academic* under *TableLooks*. The column widths, font and other features can also be changed using the commands *Format* → *Table Properties*. By clicking on the table and using the commands *Edit* → *Copy*, or by clicking on the table and right clicking the mouse and selecting '*Copy*', the table can be copied and pasted into a word file.

Place of birth (recoded)

		Frequency	Per cent	Valid per cent	Cumulative per cent
Valid	Local	90	63.8	68.2	68.2
	Regional	33	23.4	25.0	93.2
	Overseas	9	6.4	6.8	100.0
	Total	132	93.6	100.0	
Missing	System	9	6.4		
Total		141	100.0		

1.13.2 Exporting output and data from SPSS

Output can be saved in SPSS and printed directly from SPSS. However, output sheets can also be exported from SPSS into many different programs including Excel, Word or .pdf format using the commands shown in Box 1.10. A data file can also be exported to Excel using the *File* → *Save as* → *Save as type: Excel* commands.

Box 1.10 Exporting SPSS output into an Excel, Word, PowerPoint or PDF document

SPSS Commands

Output – SPSS IBM Statistics Viewer
File → *Export*
Export Output

Objects to Export: Click All, All visible or Selected to indicate part of output to export
Document: Click on Type to select output file type; File Name: enter the file name;
click Browse to select the directory to save the file
Click OK

1.14 SPSS help commands

SPSS has two levels of extensive help commands. By using the commands *Help* → *Topics* → *Index*, the index of help topics appears in alphabetical order. By typing in a keyword, followed by enter, a topic can be displayed. At the end of the *Help* options, the next major heading is *Tutorial*, which is a step by step guide how to perform certain SPSS procedures such as 'using the data editor', 'working with output', or 'working with syntax'. There is also the 'Statistics Coach', which displays the SPSS commands to be used for some statistical tests and corresponding example outputs.

There is also another level of help that explains the meaning of the statistics shown in the output. For example, help can be obtained for the above frequencies table by doubling clicking on the left-hand mouse button to outline the table with a hatched border and then single clicking on the right-hand mouse button on any of the statistics labels. This produces a dialog box with *What's This?* at the top. Clicking on *What's This?* provides an explanation of the highlighted statistical term. Clicking on *Cumulative Percent* opens up a dialog box providing the explanation that this is 'The percentage of cases with non-missing data that have values less than or equal to a particular value'.

1.15 Golden rules for reporting numbers

Throughout this book the results are presented using the rules that are recommended for reporting statistical analyses in the literature.[5–7] Numbers are usually presented as digits except in a few special circumstances as indicated in Table 1.8. When reporting data, it is important not to imply more precision than actually exists, for example, by using too many decimal places. Results should be reported with the same number of decimal places as the measurement, and summary statistics should have no more than one extra decimal place. A summary of the rules for reporting numbers and summary statistics is shown in Table 1.8.

1.16 Notes for critical appraisal

When critically appraising statistical analyses reported in the literature, that is when applying the rules of science to assess the validity of the results from a study, it is important to ask the questions shown in Box 1.11. Results from studies in which outliers are treated inappropriately, in which the quality of the data is poor or in which an incorrect statistical test has been used are likely to be biased and to lack scientific merit. The CONSORT, which stands for Consolidated Standards of Reporting Trials

Table 1.8 Golden rules for reporting numbers

Rule	Correct expression
In a sentence, numbers less than 10 are words	In the study group, eight participants did not complete the intervention
In a sentence, numbers 10 or more are numbers	There were 120 participants in the study
Use words to express any number that begins a sentence, title or heading. Try and avoid starting a sentence with a number	Twenty per cent of participants had diabetes
Numbers that represent statistical or mathematical functions should be expressed in numbers	Raw scores were multiplied by 3 and then converted to standard scores
In a sentence, numbers below 10 that are listed with numbers 10 and above should be written as a number	In the sample, 15 boys and 4 girls had diabetes
Use a zero before the decimal point when numbers are less than 1	The *P* value was 0.013
Do not use a space between a number and its per cent sign	In total, 35% of participants had diabetes
Use one space between a number and its unit	The mean height of the group was 170 cm
Report percentages to only one decimal place if the sample size is larger than 100	In the sample of 212 children, 10.4% had diabetes
Report percentages with no decimal places if the sample size is less than 100	In the sample of 44 children, 11% had diabetes
Do not use percentages if the sample size is less than 20	In the sample of 18 children, 2 had diabetes
Do not imply greater precision than the measurement instrument	Only use one decimal place more than the basic unit of measurement when reporting statistics (means, medians, standard deviations, 95% confidence interval, inter-quartile ranges, etc.), for example, mean height was 143.2 cm
For ranges use 'to' or a comma but not '-' to avoid confusion with a minus sign. Also use the same number of decimal places as the summary statistic	The mean height was 162 cm (95% CI 156 to 168) The mean height was 162 cm (95% CI 156, 168) The median was 0.5 mm (inter-quartile range −0.1 to 0.7) The range of height was 145–170 cm
P values >0.05 should be reported with two decimal places	There was no significant difference between groups ($P = 0.35$)
P values between 0.001 and 0.05 should be reported to three decimal places	There was a significant difference in blood pressure between the two groups ($t = 3.0$, $df = 45$, $P = 0.004$)
P values shown on output as 0.000 should be reported as <0.0001	Children with diabetes had significantly lower levels of insulin than control children without diabetes ($t = 5.47$, $df = 78$, $P < 0.0001$)

developed by the CONSORT Group also provides a useful checklist to determine whether appropriate study information for a randomized controlled trial (RCT) has been reported (see http://www.consort-statement.org/home/).

Box 1.11 Questions for critical appraisal

Answers to the following questions are useful for checking the integrity of statistical analyses:

- Have details of the methods and statistical packages used to analyse the data been reported?
- Are the variables classified correctly as outcome and explanatory variables?
- Are any intervening or alternative outcome variables mistakenly treated as explanatory variables?
- Are missing values and outliers treated appropriately?
- Is the sample size large enough to avoid type II errors?

References

1. Tabachnick BG, Fidell LS. *Using multivariate statistics*, 4th edn. Allyn and Bacon: Boston, MA, 2001.
2. Peat JK, Mellis CM, Williams K, Xuan W. *Health science research. A handbook of quantitative methods.* Allen and Unwin: Crows Nest, Australia, 2001.
3. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Stat Methods Med Res*. Published online before print February 19, 2013, doi: 10.1177/0962280213476378.
4. Fairclough DL. *Design and analysis of quality of life studies in clinical trials*, 2nd edn. Chapman & Hall/CRC: Boca Raton, FL, 2010.
5. Stevens J. *Applied multivariate statistics for the social sciences*, 3rd edn. Lawrence Erlbaum Associates: Mahwah, NJ, 1996.
6. Peat JK, Elliott E, Baur L, Keena V. *Scientific writing: easy when you know how*. BMJ Books: London, 2002.
7. Lang TA, Secic M. *How to report statistics*. American College of Physicians: Philadelphia, PA, 1977.