Medical Statistics

A Guide to SPSS, Data Analysis and Critical Appraisal

Second Edition

Belinda Barton
Children's Hospital Education Research Institute, The Children's Hospital at Westmead, Sydney, Australia

Jennifer Peat
Honorary Professor, Australian Catholic University and Research Consultant, Sydney, Australia

WILEY Blackwell

BMJ Books

Contents

Introduction

Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write.
H.G. WELLS

Anyone who is involved in medical research should always keep in mind that science is a search for the truth and that, in doing so, there is no room for bias or inaccuracy in statistical analyses or interpretation. Analyzing the data and interpreting the results are the most exciting stages of a research project because these provide the answers to the study questions. However, data analyses must be undertaken in a careful and considered way by people who have an inherent knowledge of the nature of the data and of their interpretation. Any errors in statistical analyses will mean that the conclusions of the study may be incorrect.[1] As a result, many journals may require reviewers to scrutinize the statistical aspects of submitted articles, and many research groups include statisticians who direct the data analyses. Analyzing data correctly and including detailed documentation so that others can reach the same conclusions are established markers of scientific integrity. Research studies that are conducted with integrity bring personal pride, contribute to a successful track record and foster a better research culture, advancing the scientific community.

In this book, we provide a step-by-step guide to the complete process of analyzing and reporting your data – from creating a file to entering your data to how to report your results for publication. We provide a guide to conducting and interpreting statistics in the context of how the participants were recruited, how the study was designed, the types of variables used, and the interpretation of effect sizes and *P* values. We also guide researchers, through the processes of selecting the correct statistic, and show how to report results for publication. Each chapter includes worked research examples with real data sets that can be downloaded and used by readers to work through the examples.

We have included the SPSS commands for methods of statistical analysis, commonly found in the health care literature. We have not included all of the tables from the SPSS output but only the most relevant SPSS output information that is to be interpreted. We have also included the commands for obtaining graphs using SigmaPlot, a graphing software package that is frequently used. In this book, we use SPSS version 21 and SigmaPlot version 12.5, but the messages apply equally well to other versions and other statistical packages.

We have written this book as a guide from the first principles with explanations of assumptions and how to interpret results. We hope that both novice statisticians and seasoned researchers will find this book a helpful guide.

In this era of evidence-based health care, both clinicians and researchers need to critically appraise the statistical aspects of published articles in order to judge the implications and reliability of reported results. Although the peer review process goes a long way to improving the standard of research literature, it is essential to have the skills to decide whether published results are credible and therefore have implications for

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