Laboratory Tutorial 1-3: Preparing data for analysis

In this laboratory tutorial you will:

1. Import some raw data from Excel
2. Define variables using a codebook
3. Check for errors and missing values within the data table
4. Save your data file in both SPSS and dBase 4 formats
5. Load your database into the SQL Lab tool and perform some simple queries

This is a mandatory tutorial. In order to pass the coursework, you must achieve a score of 50% or higher on the associated Blackboard quiz (Lab Quiz 1-3). You have three attempts (3) for every quiz.

***Note: We strongly suggest you do not start the Lab Quiz for this tutorial before you have all your answers ready.**

Preamble

Your colleague has passed you a data file containing data from a survey that was recently carried out to explore the factors that impact on people's level of optimism (as measured by a six-item scale). This survey was administered to members of the general public in Melbourne, Australia. The final sample size was 439, consisting of 42% males, 58% females, with ages ranging from 18 to 82 (mean = 37.4).

Unfortunately, there are several problems with the data file that you need to address:

1. Your colleague doesn't use SPSS so the file is in Excel format. You will therefore need to import the data into SPSS.
2. The file contains only raw data, no metadata (e.g. variable names, types etc) is included in the file. Fortunately, your colleague has prepared a codebook. You will need to use the information in this codebook to define the variables.
3. The data was entered in a hurry, so there may be some errors present. You will need to detect these errors and correct them.

Exercise 1: Import some raw data from Excel

The file is called “SurveyDataPrep.xls” and can be found within the Laboratory → Resources sub-folder. Download this file and open it in Excel.

Note the number of rows and columns and whether the variables (columns) are named or not. As the sample size is 439, you would expect to see the same number of data rows. Check this is the case before proceeding.

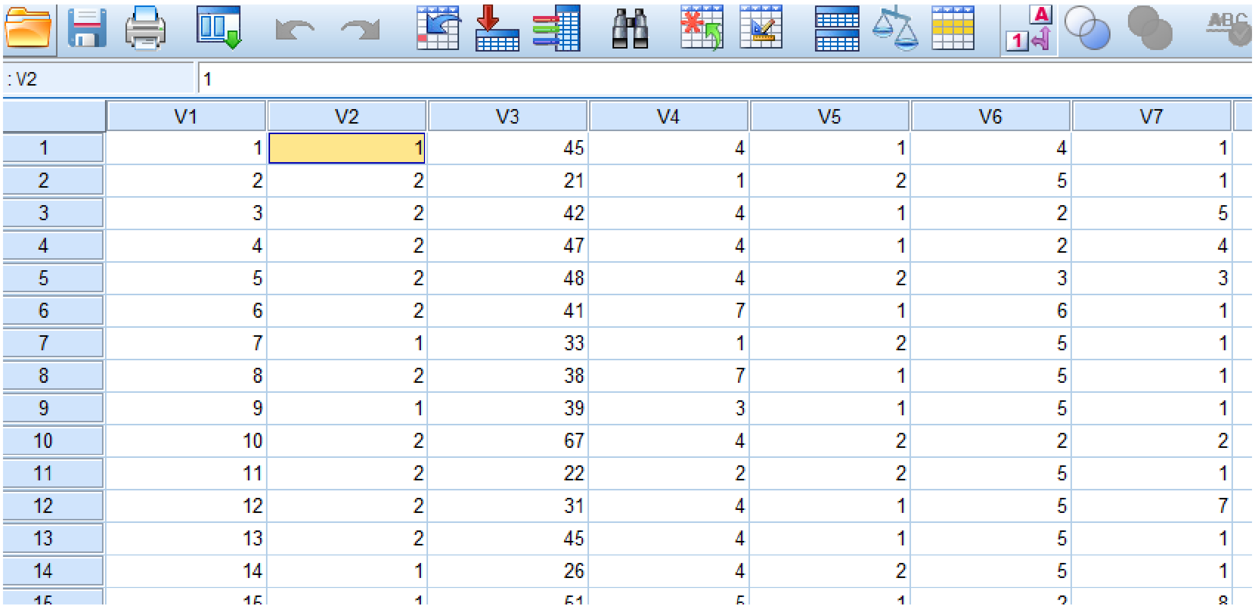
Start SPSS and go directly into the main data table view (Type in Data). Then begin the importing process by following the menu path: “**File → Open → Data**”. From the dialogue box select “Excel” from the “Files of type” menu. Select the file called “SurveyDataPrep.xls” and click “Open”.

You will now see a dialogue box called “Opening Excel Data Source”. Note there are various options presented here. Do not press “**OK**”.

Q1: There is check box for “Read variable names from the first row of data”. Do you leave this box checked (True) or not (False)?

Immediately below this option is a menu for selecting the worksheet to load. If there were more than one sheet in the file, then you would use this menu to select the correct one. As there is only one sheet in this Excel file, you can ignore this menu for now. We can also ignore the boxes for selecting the “Range” (as we want all cases) and the “Maximum width”.

Click “OK”. The first few rows and columns of the resulting data table should look something like this:



: V2 1

	V1	V2	V3	V4	V5	V6	V7
1	1	1	45	4	1	4	1
2	2	2	21	1	2	5	1
3	3	2	42	4	1	2	5
4	4	2	47	4	1	2	4
5	5	2	48	4	2	3	3
6	6	2	41	7	1	6	1
7	7	1	33	1	2	5	1
8	8	2	38	7	1	5	1
9	9	1	39	3	1	5	1
10	10	2	67	4	2	2	2
11	11	2	22	2	2	5	1
12	12	2	31	4	1	5	7
13	13	2	45	4	1	5	1
14	14	1	26	4	2	5	1
15	15	1	51	5	1	2	8

Look carefully at the first row of the table view shown above. If the values in your view of the data differ, then go back to Q1 and reconsider your answer.

Exercise 2: Define variables using a codebook

You can see that the data table is quite raw. To begin with, none of the variables have meaningful names, just V1, V2 etc. These are the default names assigned by SPSS in the absence of further information. Fortunately, we have a codebook which will help us to name the variables appropriately and to add other relevant metadata. You can find the codebook at the end of this document.

- Go to the "Variable View". You'll see a table where variables are depicted as rows. Variable properties are described by the columns.
- Enter the correct **names** for each variable using the ("SPSS variable name") in your codebook.
- Entering a more descriptive **label** (i.e. "full variable name") will make it easier to identify variables and their meaning

Remember what you learnt about in the previous lecture about levels of measurement. In some cases, the variable is represented by quantitative or **scale** level data. For instance, age (years), height (cm), weight (kg) would all be considered scale level measures because they are real numbers that can be added, subtracted, multiplied and divided. This data can be entered 'as is' into the data editor.

Some variables are **ordinal** level measures, where categories are ordered (e.g. age or BMI categories) but the differences/ratio between categories are inconsistent or difficult to quantify. We will see in later weeks that when choosing between two different classes of statistical test – parametric and non-parametric – is in part dependent upon whether a variable is measured at scale or ordinal level.

The final level of measurement is **nominal**, whereby values simply refer to unordered categories. For instance, if we had a variable for hair colour we might assign a value of 1 to brown, 2 to black and so on. These values are chosen arbitrarily; there is no reason to believe that black hair is more (or less) than brown hair, the numbers simply identify the different categories.

Hence:

- Decide, using the codebook information, the level of measurement used for each variable
- Set the correct level in the **Measure** column within the variable view

Q2: Which of the following are nominal variables?

- **sex**
- **age**
- **marital**
- **child**
- **op1**
- **educ**

Within the codebook, look at the column called “Coding Instructions”. This contains information that tells you how the data was numerically encoded for each variable. This is important when it comes to certain ordered (ordinal) and most unordered (nominal) category data where numerical codes have been assigned to qualitatively defined responses/observations. If a coding scheme is given (e.g. 1=Male, 2=Female), you need to assign the labels based on this scheme using the “Values” property. You’ve done this before back in Lab 1-1. If you have problems, refer back to these notes.

Exercise 3: Check for errors within the data table

Recall from the lecture that the **Frequencies** command is often a good way to detect data entry errors.

- Follow the menu path: **Analyze → Descriptive Statistics → Frequencies**
- Select all variables in the data table except ‘ID’.
- Click on the “Statistics” button. Tick the “Minimum” and “Maximum” options in the “Dispersion” section
- Click on “Continue” then “OK” to execute the Frequencies command

Now look at the output window. You’ll see that the “Statistics” table appears first and provides summary statistics for each selected variable, including number of valid (non-missing) and missing cases along with the minimum and maximum values as specified in the statistics options (any other statistics you selected in the statistics menu will also appear here). Use this summary information to identify any variables that have out values outside of their expected range.

Q3: Which variables contain out of range values?

Note: the legitimate ranges for categorical variables can be found in the codebook. The age range of the subject sample is described at the start of this tutorial.

The statistics also show the number of missing values for each variable. Two variables have a relatively large number of cases with no data.

Q4: Which variable has the highest number of missing value cases?

Q5: How many missing cases exist in your answer to Q4?

After the statistics table, you should see “Frequency Tables” for each selected variable. A frequency table shows a breakdown of frequency counts by each distinct occurring value. You’ll see that for most of the variables that contained out of range values, there is just one offending case. There is one exception, however:

Q6: Which categorical variable has the largest number of out of range values?

Q7: How many cases are affected all together in your answer to Q6?

Having now identified the offending variables and the nature of likely errors you will of course want to track down and (if possible) correct the offending cases. This could be achieved by manually scanning the data table; however, this is both laborious and error-prone. Two more efficient methods were described in the lecture: The **Sort** and **Find** commands. Which is the best method for finding missing and large numbers of out of range cases?

Q8: What is the subject ID number of the case with an out of range value for the variable "educ"?

Q9: Assuming the entered value is a typo, which of the following categories is most likely to be the correct one?

- **1 – primary**
- **2 – some secondary**
- **3 – completed high school**
- **4 – some additional training**
- **5 – completed undergraduate**
- **6 – completed postgraduate**

Summary

In this tutorial, you learned how to import raw data from an Excel spreadsheet, define the variables using information in the codebook, and detect and correct errors within the data table. You also learnt how to export your data table to the dBase database format. Before you move on to the next tutorial, make sure you have successfully answered all questions in this tutorial and submitted your answers using the Blackboard quiz. If you have any queries, please ask a tutor during your laboratory session. **Don't forget to save your data file** before you close SPSS**.**

Further Reading

Pallant, J. (2007) SPSS survival manual : a step by step guide to data analysis using SPSS for Windows (Version 15), Chapters 4 & 5.

Codebook for survey.sav

Full variable name	***SPSS variable name***	***Coding instructions***
Identification number	id	subject identification number
Sex	sex	1 = males; 2 = females
Age	age	in years
Marital	marital	1 = single; 2 = steady relationship; 3 = living with a partner; 4 = married for the first time; 5 = remarried; 6 = separated; 7 = divorced; 8 = widowed
Children	child	1 = yes; 2 = no

Highest level of education	educ	1 = primary; 2 = some secondary;
		3 = completed high school;
		4 = some additional training;
		5 = completed undergraduate;
		6 = completed postgraduate

Major source of stress	source	1 = work; 2 = spouse or partner; 3 = relationships; 4 = children; 5 = family; 6 = health / illness; 7 = life in general
Do you smoke?	smoke	1 = yes; 2 = no
Cigarettes smoked per week	smokenum	Number of cigarettes smoked per week
Optimism Scale	op1 to op6	1=strongly disagree , 5=strongly agree