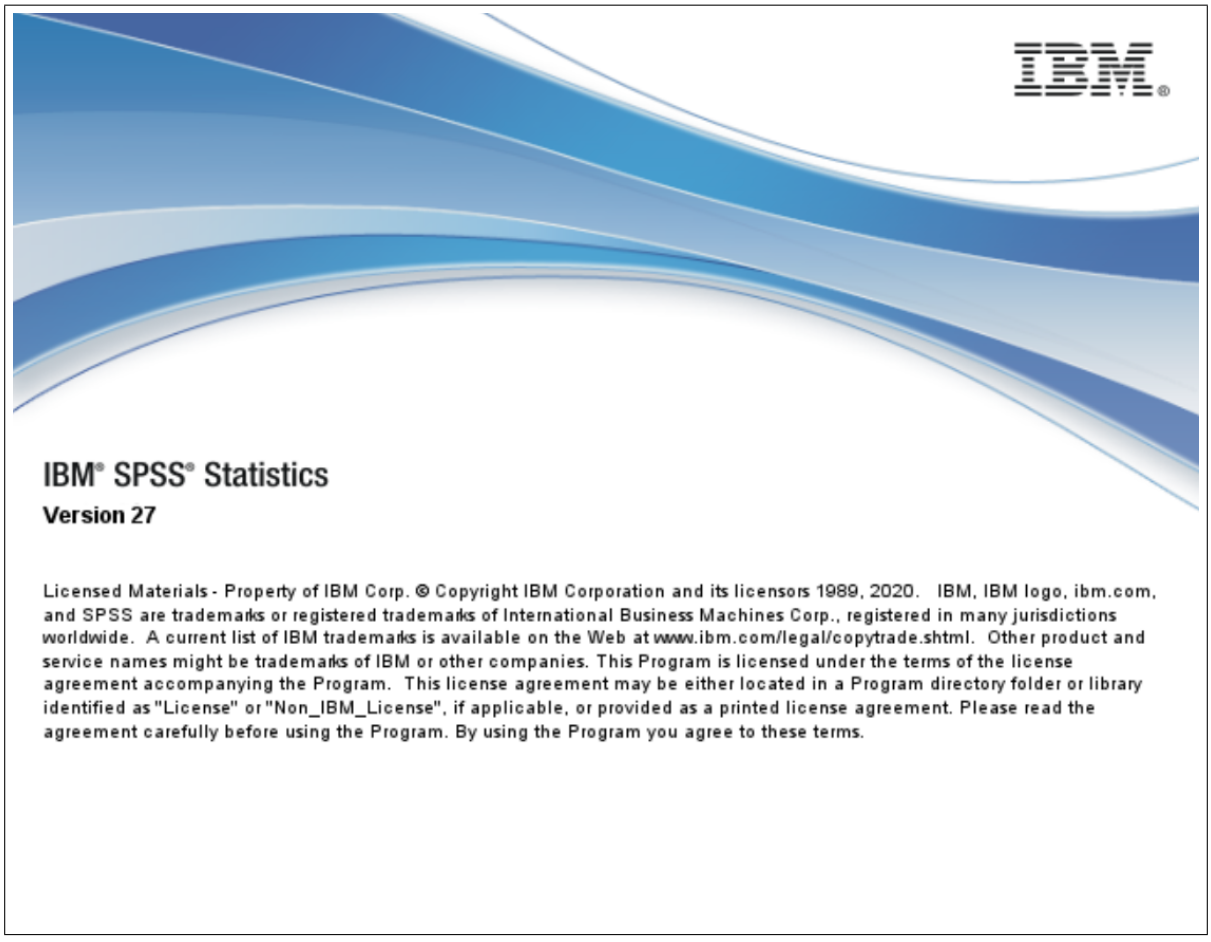
An Applied Introduction to SPSS v.27

Instructor: Dr. Matthew J. Sigal

Department of Psychology

Simon Fraser University

October 2022



Contents

Preface to the Course

0.1 Instructor Contact Information

Matthew J. Sigal, PhD
E-mail: msigal@sfu.ca
Homepage: www.matthewsigal.com
Department of Psychology, Simon Fraser University

0.2 Course Information

The goal of this workshop is to present the basics of SPSS. Some files for this course will be hosted online: http://www.matthewsigal.com/spss/. Using the PDF version of this text, all links should be live (including urls and those found in the index) and the text is fully searchable for easy navigation.

Part 1 of the workshop will introduce the computing concepts of SPSS, a few of the different facilities for reading data into an SPSS spreadsheet, and saving SPSS data files for future use. At the end of this section, participants should be able to run simple programs, including some basic statistical procedures, and feel comfortable with the SPSS environment.

Part 2 will cover basic and advanced data modifications, transformations, and other functions. These tools will allow you to open a variety of file types, and merge data frames together if necessary. By the end of this section, participants should feel comfortable defining the properties of their variables, and modifying them as necessary.

The focus of Part 3 is on the analysis of data. It will be assumed by this point that you have a basic grasp of the SPSS language and want to now utilize this skill. An emphasis will be made on the use of graphical methods for examining univariate and bivariate relationships. Some more advanced data analysis techniques will be covered, as time permits. Part 4 has some supplementary techniques that you might find useful as you delve deeper into using SPSS.

This course is designed as an applied introduction to a statistical program. As such, familiarity with statistical procedures is assumed. Practice exercises are also provided and students are encouraged to play with them outside of the workshop.

0.3 Some Useful Links for Learning SPSS

- SPSS Homepage: hhttps://www.ibm.com/spss
- SPSS Support Portal: https://www.ibm.com/products/spss-statistics/support
- SPSS Tutorials at UCLA: https://stats.oarc.ucla.edu/spss/

1 A Brief Introduction to SPSS

1.1 SPSS: An IBM Company

1.1.1 What is SPSS and why should I use it?

SPSS previously stood for the "Statistical Package for the Social Sciences", and was briefly referred to as PASW for "Predictive Analytics SoftWare", but is now an orphan initialism. It is a user friendly but powerful statistical analysis and data management system that was first developed in 1968. It largely operates through a graphical user interface, with users utilizing buttons and menus to run various statistical procedures. It is perhaps the most widely used statistical analysis program in the fields of psychology and political science, and is typically available and recommended at most educational institutions.

SPSS is a fantastic tool for learning how to conduct statistical analyses, with functions for providing rich graphical output. Other noteworthy features are the ability to program commands via a syntax interface, the easiness of data entry, the ability to assess boot-strapped estimates, and an interactive matrix calculator. Output is relatively thorough, and customizable. Finally, SPSS has some fantastic built-in help facilities, with sections pertaining to case study examples, general tips for working with SPSS, an interactive statistics coach, and syntax references.

Since IBM's purchase of SPSS, new versions of the application are typically released annually. Please note that the screenshots for this workshop were taken with version 27.0.1. It should be mentioned that, while this version may differ from that presently installed in the lab, the core functionality remains the same. The major differences between the last few versions have been entirely high-level additions of new statistical procedures, and backend tweaks. If discrepancies do arise between the lecture notes and the demonstration, please bring them to the attention of the professor.

1.1.2 How can I get it?

If you are affiliated with Simon Fraser University, there are multiple ways to obtain SPSS access beyond purchasing a very expensive license from IBM. You could:

- Log-on to an on-campus computer (SPSS is pre-installed on most of computers on campus; e.g., in AQ3148, WMC 2502, the lab in Bennett library, or the MicroLab in RCB).
- Access the program remotely using your student account and personal computer through Simon Fraser University's remote services at `https://www.sfu.ca/information-systems/services/computer-labs/remote/undergraduate.html`.
- Download to your personal computer using the educational license provided by Simon Fraser University from `https://www.sfu.ca/information-systems/services/software/spss.html` for free!

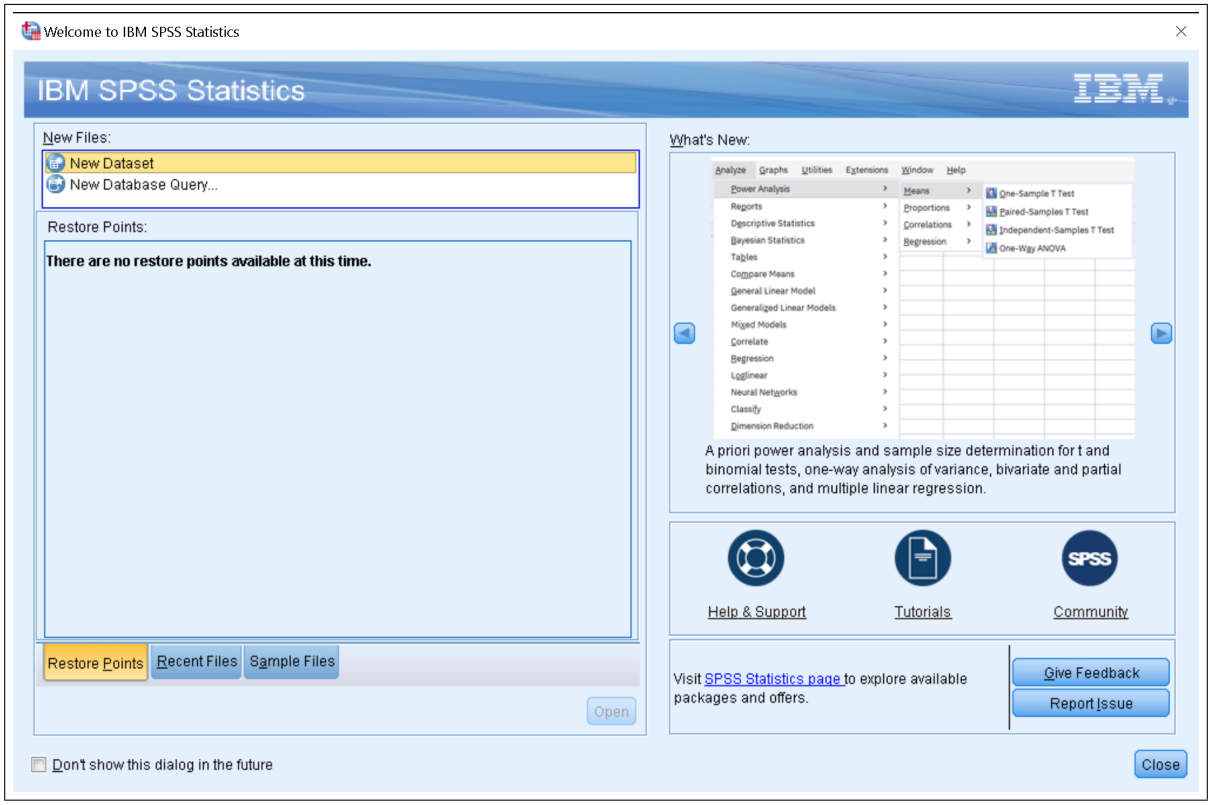
If you are affiliated with an educational institution besides SFU, you should contact your IT department and ask them if they offer SPSS licenses.

1.2 Launching and Exiting SPSS

SPSS is already installed on many of the lab computers throughout Simon Fraser University. To launch SPSS on a Windows computer, simply click on the application short-cut. This is often either located on the desktop or in the Start menu, under *IBM SPSS Statistics*. Likewise, on Mac OS X, SPSS can be launched either from a shortcut on the dock or from the *Applications* folder.

To quit SPSS, simply select **File** → **Exit** from the Windows menu, or **Quit SPSS Statistics** from the **SPSS Statistics** menu on OS X.

By default, standard installations of SPSS launch a "Welcome to IBSM SPSS Statistics" dialog box upon start-up. This behaviour can be disabled (by checking "Don't show this dialog in the future")[1], but it does allow for easy access to tutorials and other core functions.

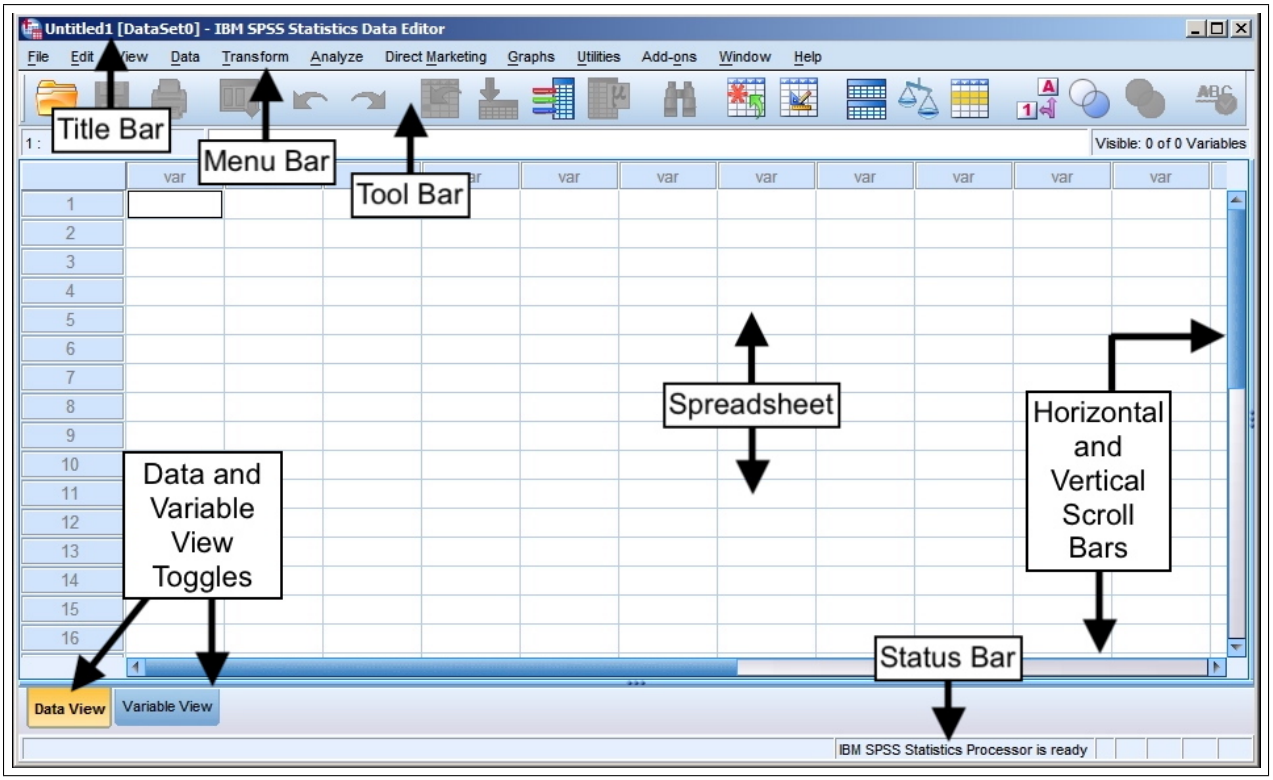


For now, launch SPSS, select 'New Dataset' and then click 'Close'.

[1] Caveat emptor: While it is easy (and often desirable) to disable this dialog box, once you have done so it is not so straight forward to get it back if you later change your mind!

1.3 The SPSS Environment

You should see this on your screen:



SPSS is spreadsheet based: most data entry work will be done using either the spreadsheet layout in Data View (used for inputting cases), or with Variable view (used for defining factors). Note that, by default, SPSS creates a new dataset upon launch entitled "Untitled1", which appears in the title bar. When we save a dataset with a new name, title bar will update.

Below the title bar is the menu bar with a series of headings (*File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window,* and *Help*). Below the menu bar is the tool bar, which contains one-click shortcuts to a variety of convenient features. Finally, note that, on the bottom of the screen (in the "status bar") the application reports that the "IBM SPSS Statistics Processor is ready".

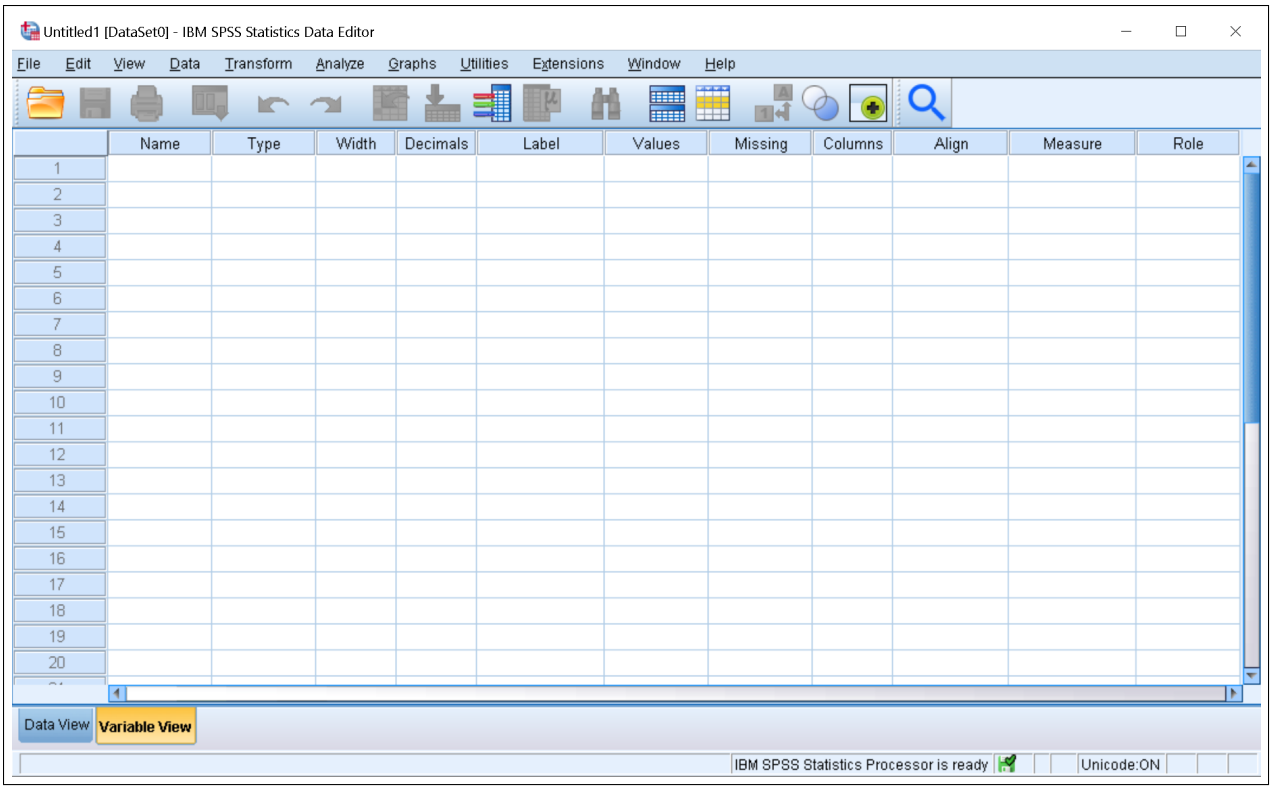
1.3.1 Data View

As pictured above, Data View has the appearance of a spreadsheet, with columns designated by the variable names of the active data set and rows numbered sequentially. Each column of the grid represents a single variable, and each row is used for a single case[2]. With large datasets, you may have to use the horizontal and vertical scroll bars to navigate through all of the data.

1.3.2 Variable View

If you are starting a project from scratch, you will typically want to begin by defining your variables. This is done in "Variable View", which can be selected via the button in the lower lefthand corner.

[2]This is called 'wide' format, as information pertaining to a particular participant is entered horizontally across columns. Some advanced analyses (such as multilevel modeling) require the data in 'long' format, in which the participant data typically spans across multiple time points, and are entered over multiple rows. This is detailed in Section 2.3.5.



Each variable can be defined in terms of the following properties:

- **Name:** Variable name (appears as the column heading in Data View; must begin with a letter and be unique)
- **Type:** Defines the type of variable (commonly choose either numeric or string)
- **Width:** The number of digits SPSS should store for a variable (8 digits is the default)
- **Decimals:** Specifies the number of decimals to show in Data View for a new variable (e.g. use 0 for integer data)
- **Label:** A long variable name that is used to designate the variable in output
- **Values:** Apply labels for values (useful w/ordinal data, e.g. 7 indicates "strongly agree")
- **Missing:** Defines how you specify missing data (e.g. use 999 to indicate missing)
- **Columns:** Defines the width of the column for the variable in Data View
- **Align:** Specifies how to align values in column ("left", "centre", "right")
- **Measure:** Specifies the level of measurement ("scale", "ordinal", "nominal")
- **Role:** Tthis property allows for the automatic inclusion/coding of a variable in certain dialog boxes (not generally recommended)

At **minimum**, you should always define a variable name and measure type!

Some of these options are aesthetic (e.g. width, number of decimals, columns, and align), and can be used to change how your variables appear in Data View (note: altering the number of decimals will not affect the precision of your numerical results, just how they are displayed).

Typically, you will want to choose: a **name** that is useful for differentiating the item from your other items, an appropriate **type** and **measure**, a **width** that is as wide as the widest variable

entry, the number of **decimals** as large as the precision in your data (e.g., if you are entering categorical variables, it is OK to use decimals = 0), and give it a meaningful **label**. If you have missing data, it is *extremely* important that you define how you will enter it (e.g. 999).

Copying Variable Attributes: Once entered, variable attributes (such as labels) can be copied from one variable and pasted onto another. This can be done from the keyboard, or via the **Edit** menu and is very handy for applying Values labels to multiple questions.

Rearranging Variables: Variables can be reordered in this view by left click and dragging on the numbered cells that appear to the left of the variable names.

1.3.3 The Menubar

The following is a brief reference guide to each of the menus and, for illustrative purposes, a selection of the options that they contain:

- **File:** Saving or loading data, graphs, or output; print functions; data export.
- **Edit:** Copy/Paste/Search values in your dataset and gives access to the Options menu.
- **View:** A few system specifications such as whether to show grid lines and whether to show variables names or labels.
- **Data:** Allows batch manipulation of the entries in Data View (most frequently used features include adding variables and/or cases; split file, which is used to split the dataset by a grouping variable; and select cases, which is used to run analyses on only a subset of the cases).
- **Transform:** This menu is useful if you want to manipulate a variable, e.g. to conduct a transformation of some sort. Manipulations can be done using "Compute Variable".
- **Analyze:** This menu serves as the backbone of SPSS and grants access to all of the statistical procedures included in the software. Below is a brief guide to some of these:
  - *Descriptive Statistics:* Measures of central tendency, frequencies, general data exploration. Also, measures of expected frequency (e.g. chi-square tests) can be conducted using "Crosstabs".
  - *Compare Means:* This is where you can find t-tests (independent and repeated measures) and one-way independent measure ANOVAs.
  - *General Linear Model:* This is the menu for complex ANOVA designs such as two-way (unrelated, related, or mixed), one-way with repeated measures, and multivariate analyses of variance.
  - *Mixed Models:* This menu can be used for running multilevel linear models.
  - *Correlate:* This menu contains methods to calculate Pearson's R, Spearman's ρ, and Kendall's τ, as well as partial correlations.
  - *Regression:* A variety of regression models, ranging from simple linear regression, to multiple linear regression, to more advanced techniques like logistic regression.
  - *Data Reduction:* For conducting factor analyses.
  - *Scale:* For conducting reliability analyses.
  - *Nonparametric Tests:* There are a variety of non-parametric statistics available, such as the Mann-Whitney test, the Kruskal-Wallis test, Wilcoxon's test, and so on, for use if you have violated the assumptions of parametric tests.
  - *Power Analyses:* Provides interfaces for calculating sample sizes for a variety of univariate designs.
- **Graphs:** This menu is used to construct visual representations of your data. This can be conducted through the more dynamic "Chart Builder" or the older Legacy Dialogs.
- **Utilities:** This menu has a few interesting options (e.g. allowing for the inclusion of data file comments, coding a "scoring wizard", running a spellcheck, or defining a variable set),

although most are not typically used.
- **Extensions:** SPSS sells several additional packages that can be accessed through this menu.
- **Window:** This menu is primarily used to switch between windows (e.g. to go from data view to the output window or to another open datafile).
- **Help:** Primarily used to access the "Topics" link, which is an online resource with pages devoted to general "Reference" topics, "Tutorials", "Case Studies", and a "Statistics Coach".

1.3.4 The Toolbar

The toolbar allows for quick access to a variety of useful commands. Under **View** → **Toolbars** → **Customize** you can customize the icons that appear as well their size (large or small). Also note that, by default, SPSS enables a feature called "ToolTips", which will tell you what a button does if you hover your mouse over it for a few seconds.

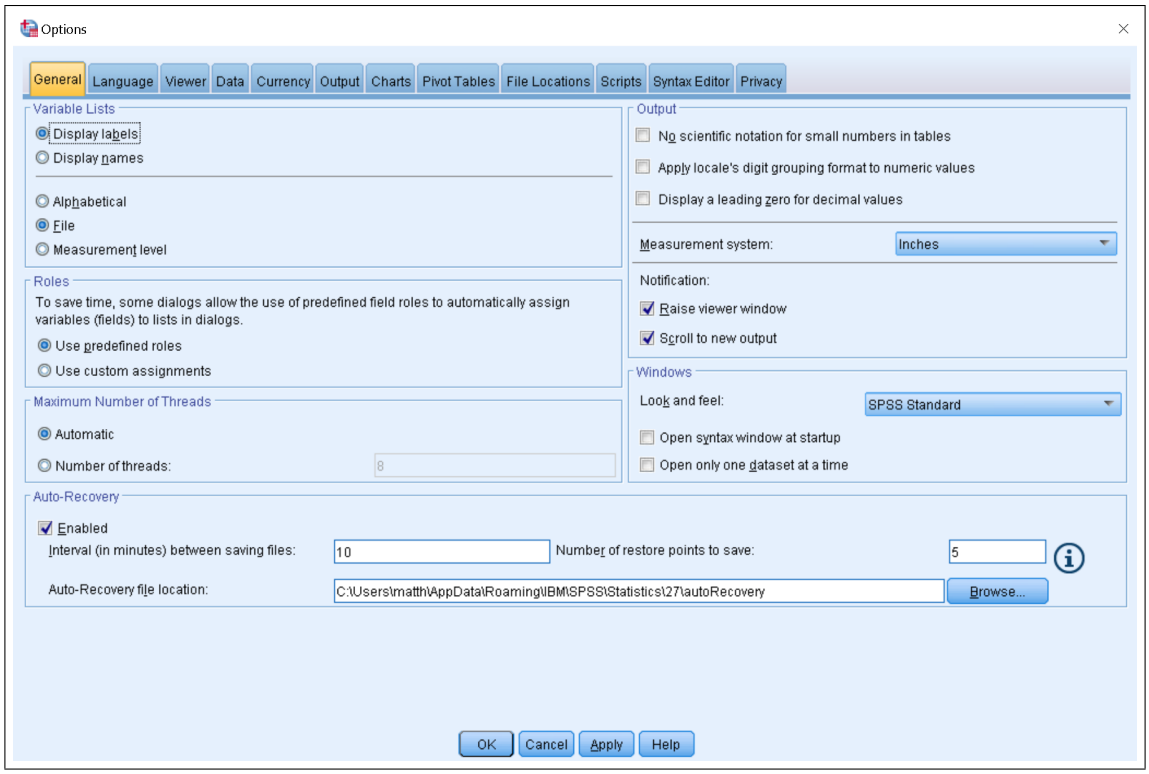
By default, the toolbar includes:

- **Open:** Open a previously saved file.
- **Save:** Save the current dataset or output.
- **Print:** Print whatever you are working on (dataset or output).
- **Recently Used Commands:** Displays your 12 most recently used commands.
- **Undo:** Undo your last action.
- **Redo:** Redo your last action.
- **Go To Case:** Go to a particular case number (row).
- **Go To Variable:** Go to a particular variable (column).
- **Variables Summary:** Show a summary of the variables included in a dataset.
- **Run Descriptive Statistics:** Calculate some relevant descriptive statistics on selected data.
- **Find:** Find and possibly replace for a given query.
- **Split File:** Quickly split output by a grouping variable (e.g. "Young" and "Old").
- **Select Cases:** This is useful for only running analyses on a particular subset of your data file (e.g. run the analysis only on cases where age category = 1 to look at that particular age band)
- **Value Labels:** Toggles between showing either the numeric values or the value labels specified in the Variable View when looking at your data in Data View.
- **Use Variable Sets:** This allows for the designation of multiple "sets" or clusters of variables to toggle between (useful with a dataset that has many variables, akin to Excel's "hide columns").
- **Search:** Find help topics, dialogs, and case studies for a given keyword.

Note: Some of these are only available under certain instances (e.g. you can't click **save** until you have entered data, and you can't switch between **value labels** and variable names in Variable View).

1.3.5 Options in SPSS

Under **Edit** → **Options** is a series of preferences you can set within the SPSS environment. By and large, most of the default values are completely acceptable although you may be inclined to tweak them here and there.



For instance, under the "General" table, you can change whether to display variables by names or by their labels in output and select a theme under "Look and feel". Under "Viewer", you can specify font options for the output file. Under "File Locations", you can change the default directory that SPSS will look in for finding saved datasets and output. You can select a default "chart template" (e.g. APA style) under "Charts". In their tab, "Pivot tables" can be customized in a similar fashion.

Some recommended settings:

- In the **General** tab, variables can be displayed using their labels (default) or their names, and sorted either alphabetically or as they are within the file (default). It is recommended that you use "Display names" instead of "Display labels".
- In the **Viewer** tab, it is recommended that you leave checked "Display commands in the log" option, which will produce the syntax SPSS generated by your selections. Further, you can manipulate the fonts used in the output.
- Under **Output**, it is recommended to change all four boxes to include both "Labels" and either "Names" or "Values", as available. These are often easier to interpret in your output than just one or the other.

1.4 Inputting and Saving a New Dataset

When you want to input data manually, this is called inputting raw data. When doing this, it is useful to remember a few things:

1. Data from different things (e.g. participants, or units of observation) should go in different rows in Data View, whereas data from the same things (e.g. a particular participant's age, education institution, and IQ) should go in different columns of the same row.
2. Nominal and ordinal variables should be inputted numerically (e.g. code relationship status as 0 for single, 1 for in a relationship, and so on), with the levels labeled in Variable View!

3. Variable NAMES must begin with a letter, cannot contain spaces, and are limited to 64 bytes (approximately 64 characters, but shorter is best!). Variable LABELS don't have such strict restrictions, and should be used to provide more details.

> Exercise: Try making a new dataset using the following information.

Participant Name	Age	Professional Sector	IQ	Income	Neuroticism
Tiki	25	Academia	110	20,000	10
Lulu	28	Academia	122	40,000	17
Mya	26	Government	98	15,000	14
Hogarth	31	Private	118	35,000	13
Mouchie	40	Government	105	22,000	21
Oscar	22	Private	130	10,000	13

Steps:

1. Create a new dataset and go to Variable View.
2. Define variables (taking care to set appropriate types, labels, values, and scales of measurement as necessary).
3. Input data in Data View.
4. Toggle between value and numeric labels to ensure that data was coded properly (see 1.3.4).

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	NAME	String	8	0	Name of Participant	None	None	8	Left	Nominal	Input
2	AGE	Numeric	8	0	Age of Participant	None	None	8	Right	Scale	Input
3	SECTOR	Numeric	8	0	Work Sector	{0, Academia}...	None	8	Right	Nominal	Input
4	IQ	Numeric	8	0	IQ	None	None	8	Right	Scale	Input
5	INCOME	Numeric	8	0	Income	None	None	8	Right	Scale	Input
6	NEUROTICISM	Numeric	8	0	Neuroticism	None	None	8	Right	Ordinal	Input

Figure 1.1: Variable View for the exercise data. Notice how the nominal sector values are defined (0 indicates Academia, 1 indicates Government, 2 indicates Private Sector), and how Neuroticism is labeled as ordinal rather than scale.

NAME	AGE	SECTOR	IQ	INCOME	NEUROTICISM
Tiki	25	0	110	20000	10
Lulu	28	0	122	40000	17
Mya	26	1	98	15000	14
Hogarth	31	2	118	35000	13
Mouchie	40	1	105	22000	21
Oscar	22	2	130	10000	13

NAME	AGE	SECTOR	IQ	INCOME	NEUROTICISM
Tiki	25	Academia	110	20000	10
Lulu	28	Academia	122	40000	17
Mya	26	Government	98	15000	14
Hogarth	31	Private Sector	118	35000	13
Mouchie	40	Government	105	22000	21
Oscar	22	Private Sector	130	10000	13

Figure 1.2: Data View. The left display shows the data as entered, with sector inputted using a numeric key. On the right, is the same data with value labels toggled on (via the toolbar).

> Exercise: Try saving this dataset with any filename you want. What is the default file type?

1.5 Importing a Pre-Existing Dataset

1.5.1 Loading an SPSS save file

If you have an SPSS data file (*.sav, the default extension), all you need to do is double click on it, and SPSS will launch and load it for you. Similarly, you can use **File** → **Open** or **File** → **Recently Used Data** to quickly access any previously saved file on your computer.

If you wish to import a dataset from another program (such as Microsoft Excel), there are typically four approaches:

1. Open the file in its native application and attempt to save it as an SPSS file (*.sav).
2. Open the file in SPSS using **File** → **Open** and selecting the appropriate filetype.
3. Open the file in its native application and simply copy and paste the data into SPSS.

Unfortunately, the first approach is not presently available using the latest version of Microsoft Excel, but is from other applications, such as SAS.

The most recommended way of importing data is to attempt to open them directly within SPSS. This simply involves selecting **File** → **Open** → **Data** and then changing the file type from *.sav to what you are trying to import (e.g. an Excel file, *.xls, *.xlsx, and *.xlsm). If the spreadsheet you are importing contains variable names in the first row check the "Read variable names" option, otherwise leave it unchecked.

The third method of importing data has no real benefits over using the built-in Open command, although it can be quicker - especially if you are only copying portions of a dataset. It is possible to simply highly a column (or matrix) of data in Excel, select **Edit** → **Copy** and paste it into an SPSS data frame. The main downside of this procedure is that variable names are not retained.

> For each of the applied examples during the remainder of the course, a different data file type will be used. For future reference, use the index 'Data' entries for guidance on importing particular datatypes.

1.5.2 Loading a sample dataset from Excel

This example uses the same data that we manually entered in Section 1.4. Feel free to open the datafile, `example.xlsx`, with Excel to see how it is organized.

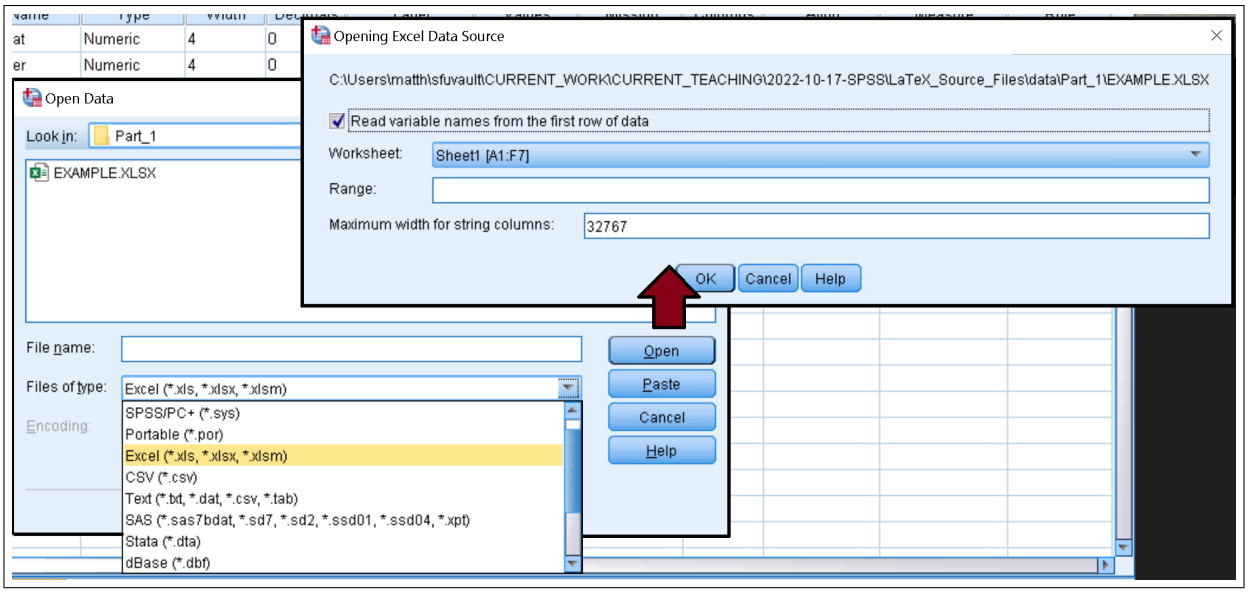
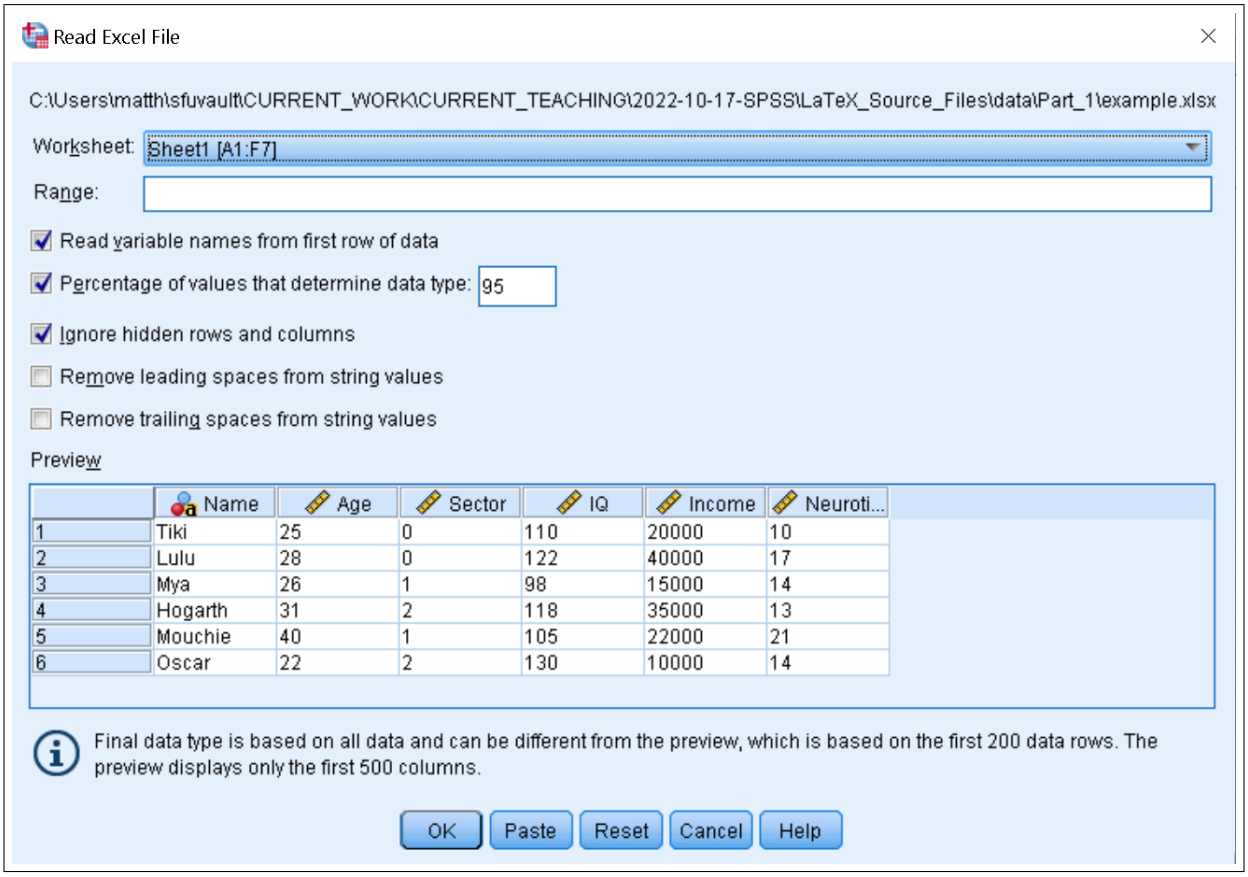


Figure 1.3: Loading an Excel dataset. Make sure to change the "Files of type" option in the first window, and to select "Read variable names from the first row of data" in the second.



	Name	Age	Sector	IQ	Income	Neuroti...
1	Tiki	25	0	110	20000	10
2	Lulu	28	0	122	40000	17
3	Mya	26	1	98	15000	14
4	Hogarth	31	2	118	35000	13
5	Mouchie	40	1	105	22000	21
6	Oscar	22	2	130	10000	14

Figure 1.4: Importing an Excel spreadsheet summary view.

Exercise:

1. Try importing `example.xlsx`, with "Read variable names" checked.
2. Note the differences in labeling between the manually entered data and this one.
3. Ensure each variable has the proper type and level of measurement specified.
4. Add variable labels to the six variables.
5. Add value labels to the Sector variable.
6. Save this dataset as "fakedata.sav"

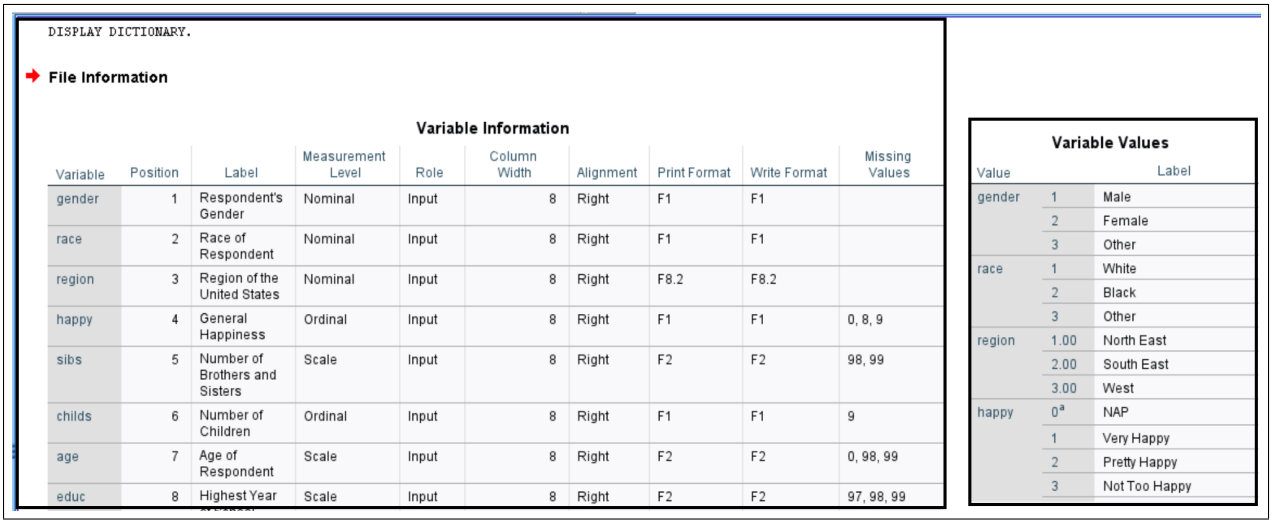
1.6 Working with Data

Now that we have gone over the basics of SPSS and basic file import, lets look at a more interesting dataset and what can be done with it. Find and open the `1991 US Social Survey.sav` datafile.

1.6.1 Data Overview

When opening a datafile, it is a good idea to:

1. Switch to Variable View and look at how each variable was entered. Ask yourself:
   - Do the codings make sense?
   - Are the variables appropriately labeled?
   - How are missing values dealt with?
2. Flip back to Data View and use the Variable Labels toggle to see if any values 'stick out'.
3. Look at the SPSS Codebook, which summarizes the dataset by selecting **File** → **Display Data File Information** → **Working File**. Does anything strike you as interesting or notable?



DISPLAY DICTIONARY.

File Information

Variable Information

Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format	Missing Values
gender	1	Respondent's Gender	Nominal	Input	8	Right	F1	F1	
race	2	Race of Respondent	Nominal	Input	8	Right	F1	F1	
region	3	Region of the United States	Nominal	Input	8	Right	F8.2	F8.2	
happy	4	General Happiness	Ordinal	Input	8	Right	F1	F1	0, 8, 9
sibs	5	Number of Brothers and Sisters	Scale	Input	8	Right	F2	F2	98, 99
childs	6	Number of Children	Ordinal	Input	8	Right	F1	F1	9
age	7	Age of Respondent	Scale	Input	8	Right	F2	F2	0, 98, 99
educ	8	Highest Year	Scale	Input	8	Right	F2	F2	97, 98, 99

Variable Values

Value		Label
gender	1	Male
	2	Female
	3	Other
race	1	White
	2	Black
	3	Other
region	1.00	North East
	2.00	South East
	3.00	West
happy	0[a]	NAP
	1	Very Happy
	2	Pretty Happy
	3	Not Too Happy

Figure 1.5: Selections from our "Codebook".

Note: If this window does not automaticaly "pop-up" when selected, click on the SPSS icon (or use the "Window" menubar) to change the view to the "Output" window.

1.6.2 Descriptive Statistics, Introduction

SPSS has many functions for describing a dataset. For instance, we might want to know about the mean age of our participants. Or we might be interested in the median level of education. To

generate such output, we typically use two functions under **Analyze** → **Descriptive Statistics**, **Frequencies** and **Descriptives**.[3] These, and two other procedures for generating descriptive reports, will be covered in more detail in Section 3.1.

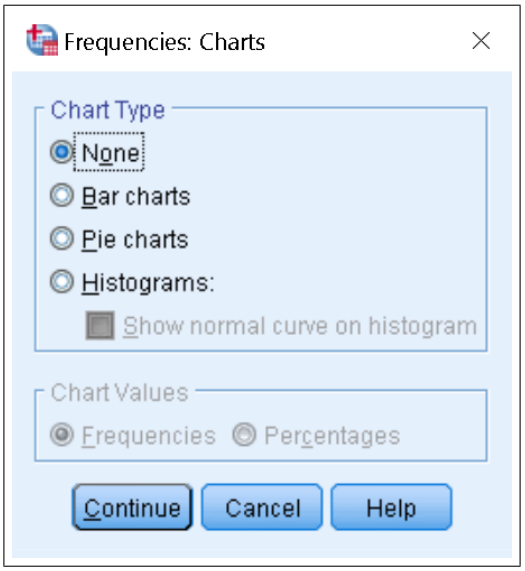


Figure 1.6: The Frequencies dialog box, with Statistics and Charts expanded.

To request SPSS to analyze a variable, simply bring it from the left box (the available variables) to the right box, either by double clicking on it or by selecting it and clicking on the blue arrow. Once some variables are selected, we can ask to look at their quartiles, their measures of central tendency (mean, median, mode), dispersion (standard deviation, variance), and distributional properties (skewness and kurtosis), via the "Statistics" button. We can also obtain histograms with normal curves overlaid through the Charts dialog box, and request or ignore frequency tables (which are handy for categorical variables, but not very useful for continuous ones).

1.6.3 Missing Data and SPSS

In world of real data analysis, missing data is an almost always present problem. Participants will fail to return for a repeated measures study, or they will simply skip an important question (of course, hopefully not on purpose!). SPSS has two default behaviours in terms of dealing with missing data — neither of which are actually optimal. Throughout the program, typically in the 'Options' tab of a statistical analysis panel, you will be asked to choose between **Pairwise deletion of missing cases** and **Listwise deletion of missing cases**.

Listwise deletion entails dropping a case (observation) from the analysis because it had a missing value in at least one of the specified variables in the model. The analysis is only run on the observations that have a *complete* set of data. In contrast, **pairwise deletion** means that the case with the missing data is only ignored when the procedure is estimating the parameters for the particular variable that the case is missing on. The remaining information is used, if possible. This means that more of your data is being used than with listwise deletion, but that each computed statistic may be based upon a different subset of observations!

[3] Both **Descriptives** and **Frequencies** produce similar output, but, unless you want to save standardized scores, "Frequencies" is more flexible.

In both cases, the default behaviour of SPSS is to throw away your hard-earned data!

Entire courses can be taught on approaches to missing data. If missingness is a problem for your research, it is highly recommended that you look into alternative procedures to deal with it, such as **multiple imputation**.

1.6.4 Working with Syntax

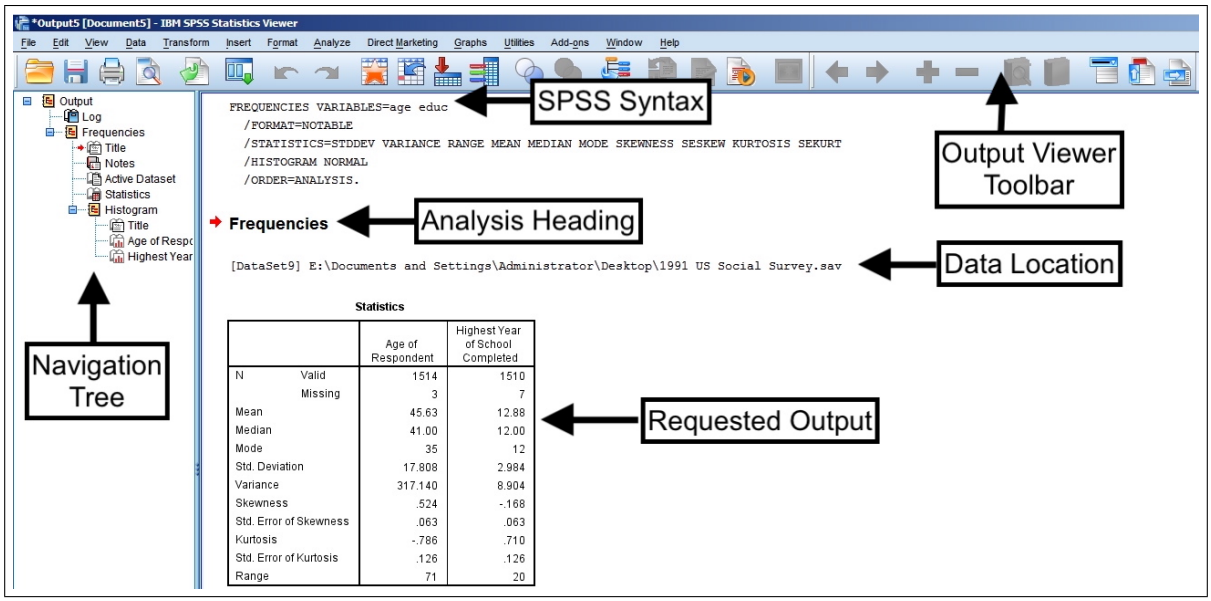
A quick note should be made here about the syntax editor. Syntax is a feature often invoked by power users. It allows for any operation that you can run in SPSS via a menu selection to be done through textual commands. Benefits of using syntax include being able to conduct analyses (or variations of analyses) quickly and being able to specify options that might not be available in the conventional dialog box.

To get started with syntax, it is highly recommended to every-so-often use the "paste" button when conducting an analysis (seen in Figure 1.4 to the right of the "OK" button). This will open a new syntax window and show you what commands SPSS is running for that particular test, based upon your selections. This way you can see how the code should look and how you might build upon it. One minor (but essential) thing to note is that syntax commands require a full stop (a period character) at the end of the requested program. If it is forgotten, SPSS may output a warning that it was unable to complete the requested commands. Also, to run submitted code, you should include an "EXECUTE." command on the final line.

1.6.5 The SPSS Viewer

Before proceeding, choose a few variables, and request some descriptive statistics on them.

Once you have made your selections for some descriptive output, SPSS will automatically open a new window. This window is called the **SPSS Viewer** and has the following basic properties: a tree diagram of the current output in the left-hand column, and the actual output of the analysis on the right-hand side. The following is output produced from the example data set with "Frequencies":



		Age of Respondent	Highest Year of School Completed
N	Valid	1514	1510
	Missing	3	7
Mean		45.63	12.88
Median		41.00	12.00
Mode		35	12
Std. Deviation		17.808	2.984
Variance		317.140	8.904
Skewness		.524	-.168
Std. Error of Skewness		.063	.063
Kurtosis		-.786	.710
Std. Error of Kurtosis		.126	.126
Range		71	20

Figure 1.7: Sample output window produced in the SPSS Viewer.

1.6.6 Navigating and working with the SPSS Viewer

- To edit an item (a title, a graph, or pretty much anything else that appears in the output viewer), double click on it and the appropriate window will open.
- To move quickly through output, use the navigation tree. You can safely open or close (or minimize) branches with the plus and minus sign icons, respectively. This comes in handy when a series of analyses are run on the same dataset.
- Click on an item in the navigation tree to jump directory to it.
- To delete an object, select it on the tree diagram and press the delete key on the keyboard.
- The menu bar items in the output viewer are similar to those in the Data and Variable View. A few new additions include buttons to manage the output tree, such as allowing the user to promote or demote an item in the hierarchy or to show and hide components.[4]
- Like the data itself, SPSS output can be saved to a separate file via **File** → **Save**. The result is an .spv file that can be opened again with SPSS. It is also possible to export your results to other file types (e.g. HTML or PDF) via the **File** → **Export** command.
- Output can also be directly printed using **File** → **Print**.

[4]Remember, simply hover the mouse over an icon to activate ToolTips and learn more about it!

Exercise:

1. Find the mean, median, mode, and range for the age of respondent and the number of years of education completed.
2. Think about why "Number of Children" is classified as an ordinal measure rather than scale.
3. Find the standard deviation and variance of the paeduc and maeduc variables.
4. Find the 75th percentile for happiness.
5. Look at a frequency table to see the breakdown for the race of the participants.
6. Compare the histograms for the levels of education for the participant's mother and father with normal curves overlaid. Which appears more normal? Which parent, on average, seemed to have a higher level of education?

At this point, you have the tools needed to work through Hands On Exercise I. Good luck!

2 Data Manipulation

2.1 Basic Data Manipulation

This section will go over some essential of data editing, filtering, transformations, and other useful features found in the SPSS environment.

2.1.1 Importing Plain Text Data

> Start by using **File** → **Open** → **Data** and the filetype "Text" to access `CLASS.DAT`

This filetype, *.DAT, is a generic data file. It is a plain text file, as you might observe if you were to open it with NotePad or a similar plain text editor. You might be surprised upon opening the file that data doesn't just appear in the Data View. Instead, the **Text Import Wizard** will appear to guide you through the process. This wizard is designed to help read in data stored in two basic ways: in **'Fixed Width'** format or in **'Delimited'** format.

The major difference between these two formats is how space was allocated when the data was inputted. If data is entered as **fixed width**, it means that each variable is recorded in the same location on the same line for each case in the data file - each value lines up along a vertical line. In contrast, **delimited** data use a special character (often a comma, as in *.csv files, which stands for 'comma-separated values') to indicate separation between values.

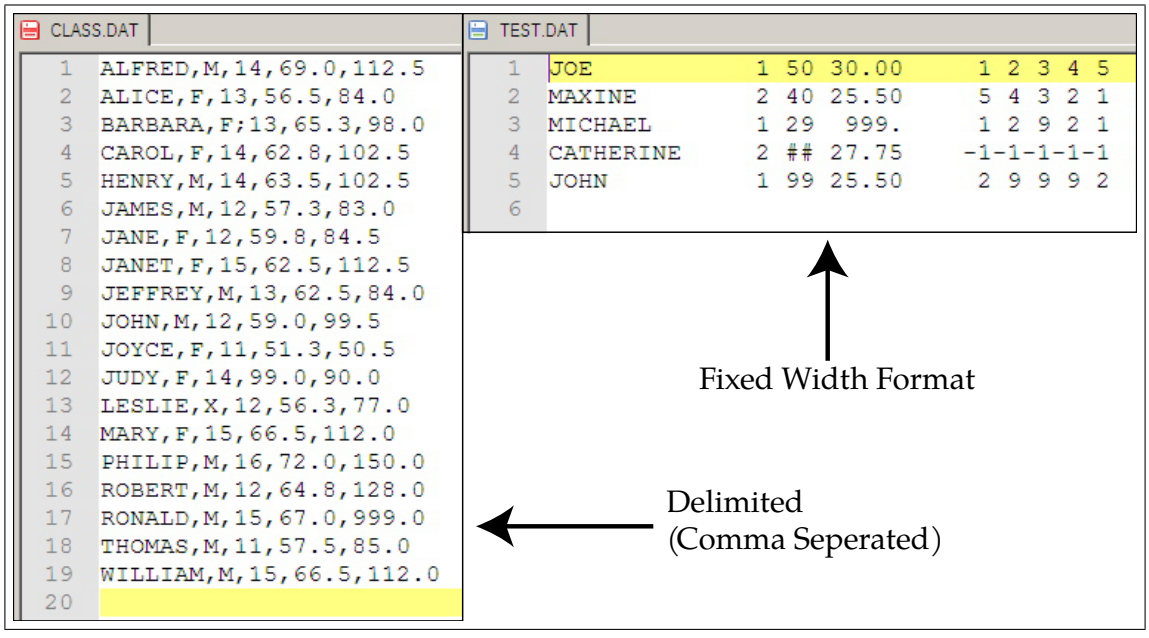


Figure 2.1: Visual difference between a Delimited and a Fixed Width raw text file. Note how, in both cases, each line/row represents a particular case.

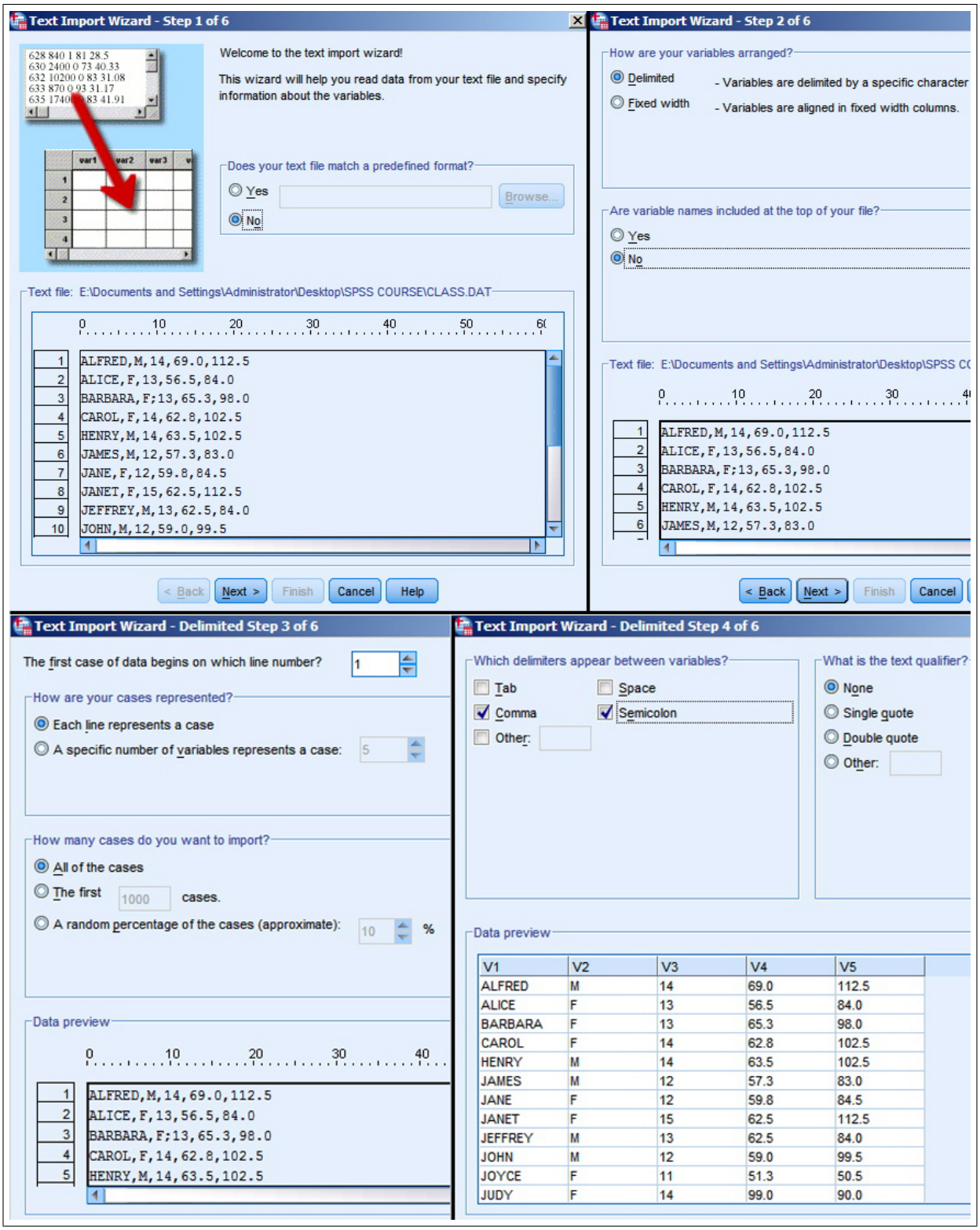


Figure 2.2: The Text Import Wizard, Steps 1 through 4.

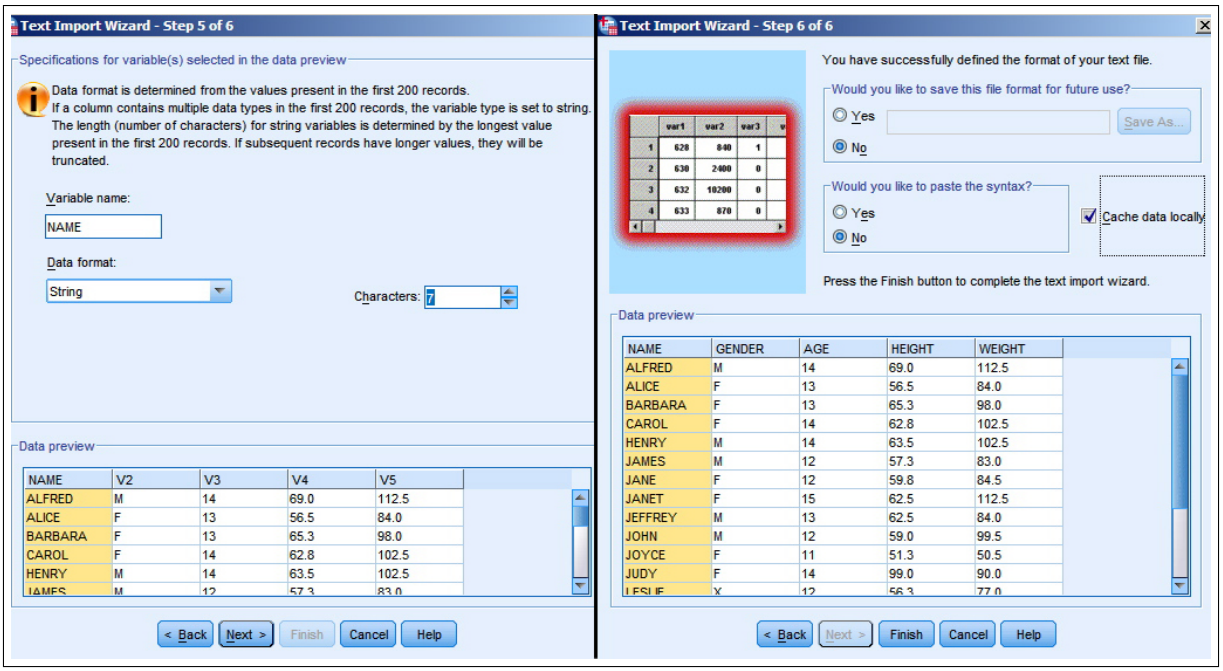
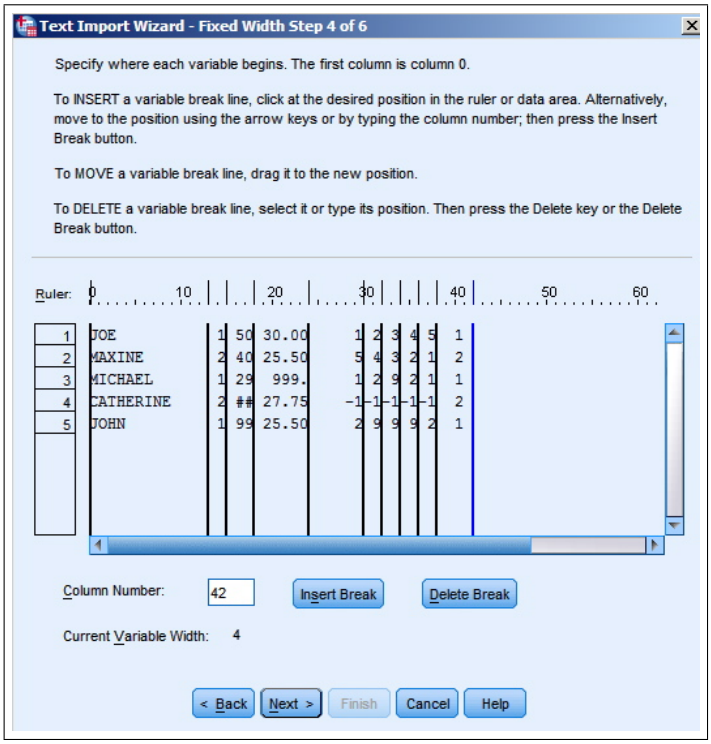


Figure 2.3: The Text Import Wizard, Steps 5 and 6.

Exercise: Finish the import of `CLASS.DAT`, remembering to define the variable names as in Figure 2.3.

2.1.2 Importing Data in Fixed Width Format

The only difference between importing data from a delimited file and a fixed width file is in Step 4 of the Text Import Wizard. Instead of indicating the delimiter used (or having SPSS autodetect it for you), you will instead be prompted with a ruler and a series of bars. It is up to you to indicate where each of the 'variable break lines' occur.



This is accomplished using the **Insert Break** and **Delete Break** commands, and by nudging each break appropriately. The space between each line indicates the data for a particular variable.

2.1.3 Inserting and Removing Data

The following steps use your imported `CLASS.DAT` file.

Adding Variables: To add a new variable to an open dataset, use **Edit** → **Insert Variable**. This will add a new column *to the left* of the currently selected variable.

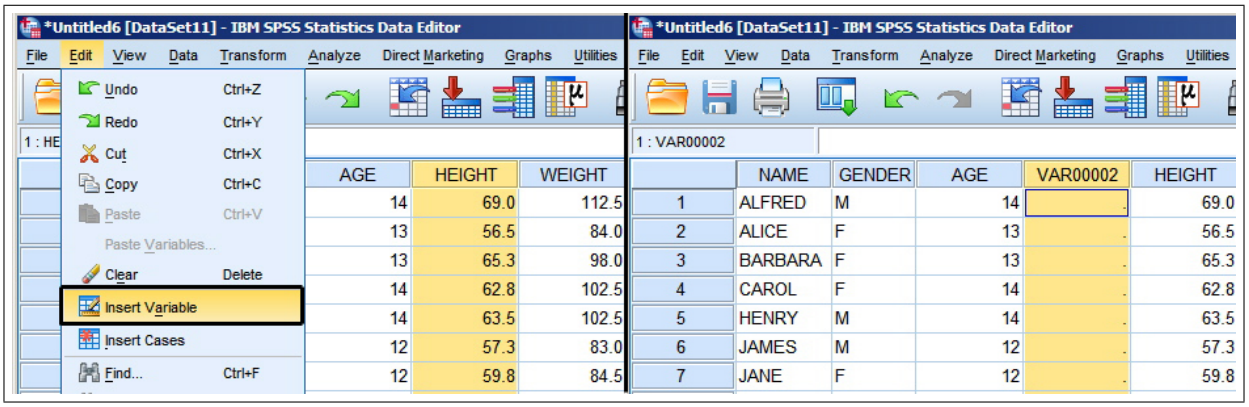


Figure 2.4: Inserting a new variable.

Removing Variables: Likewise, to delete a variable, click on the name of the column to select it, and then go **Edit** → **Clear**.

Adding Cases: To add a case (or new observation) to an open dataset, you can scroll to the bottom and simply type in the new data. If you want the observation to appear at a particular point in the dataset, scroll to where you want the data to appear and select the *row below* where you want the enter the data. Then select **Edit** → **Insert Case**.

Removing Cases: To remove a case, select the row by clicking on the desired row number, and then go **Edit** → **Clear**.

> **Exercise:** Using `CLASS.DAT`, try adding a new variable between GENDER and AGE. Call it 'ATTITUDE' and give each individual a score between 0 and 100. Also, add a new entry for 'Huxley', between Henry and James. Save this file as `MYCLASS.SAV`.

2.1.4 Merging Files

What would happen if, now that we have our dataset, we find out that we actually have *more* data, but it was hiding in another file? Or, we had a team of undergraduates tasked with inputting our data and now we want to collect all the files into one dataset? Luckily, SPSS has facilities to accommodate this dilemma.

Adding Cases: Probably the most common issue would be adding more cases to a pre-existing file. This assumes that both files have the **same variables** but **different data**. For instance, lets add `NEWSTUDENTS.SAV` to `MYCLASS.SAV`.[1]

[1] Warning: Make sure you completed the previous exercise and added the 'Attitude' variable before attempting this merge!

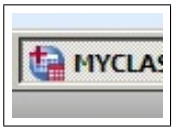
MYCLASS.SAV

	NAME	GENDER	ATTITUDE	AGE	HEIGHT	WEIGHT
1	ALFRED	M	80.00	14	69.0	112.5
2	ALICE	F	85.00	13	56.5	84.0
3	BARBARA	F	50.00	13	65.3	98.0
4	CAROL	F	30.00	14	62.8	102.5
5	HENRY	M	40.00	14	63.5	102.5
6	HUXLEY	M	90.00	10	72.0	125.0
7	JAMES	M	34.00	12	57.3	83.0
8	JANE	F	81.00	12	59.8	84.5
9	JANET	F	66.00	15	62.5	112.5
10	JEFFREY	M	14.00	13	62.5	84.0
11	JOHN	M	2.00	12	59.0	99.5
12	JOYCE	F	50.00	11	51.3	50.5
13	JUDY	F	55.00	14	99.0	90.0
14	LESLIE	X	61.00	12	56.3	77.0
15	MARY	F	77.00	15	66.5	112.0
16	PHILIP	M	70.00	16	72.0	150.0
17	ROBERT	M	33.00	12	64.8	128.0
18	RONALD	M	50.00	15	67.0	999.0
19	THOMAS	M	60.00	11	57.5	85.0
20	WILLIAM	M	40.00	15	66.5	112.0

NEWSTUDENTS.SAV

	NAME	GENDER	ATTITUDE	AGE	HEIGHT	WEIGHT
1	HOGARTH	M	39.00	12	62.5	115.0
2	LULU	F	86.00	13	58.5	100.0
3	MALOU	F	80.00	9	60.0	90.0
4	MOUCHIE	M	20.00	12	63.0	105.0
5	MYA	F	68.00	11	59.0	77.0
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						

Figure 2.5: The original datafile and the new students datafile that we wish to add.



To merge these two files, have both the primary datafile (`MYCLASS.SAV`) and `NEWSTUDENTS.SAV` open. Make sure `MYCLASS.SAV` is in the foreground by looking for a plus sign on the taskbar. Then select **Data** → **Merge Files** → **Add Cases**. All open datasets[2] will appear in the Add Cases dialog box. We want to add our new students, so select `NEWSTUDENTS.SAV` and click continue.

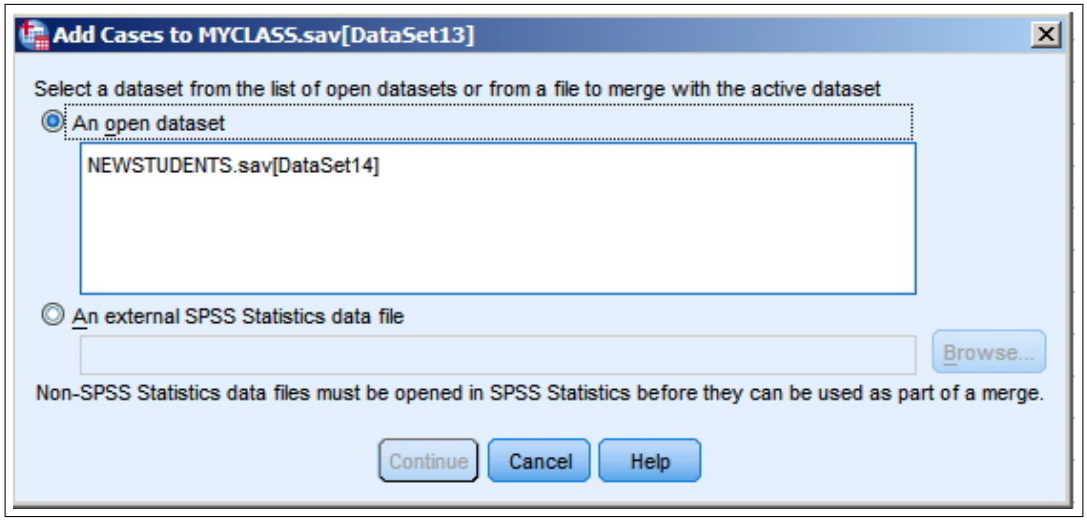


Figure 2.6: The Add Cases dialog box.

In the next dialog box, you have to select which variables are paired between the two datasets. This feature allows you to selective match between the two files (for instance, if the second file has an additional variable that you don't want merged, you can leave it in the 'Unpaired Variables' pane on the left).

[2]It is recommended to only have the datasets you are actively working with open at a time.

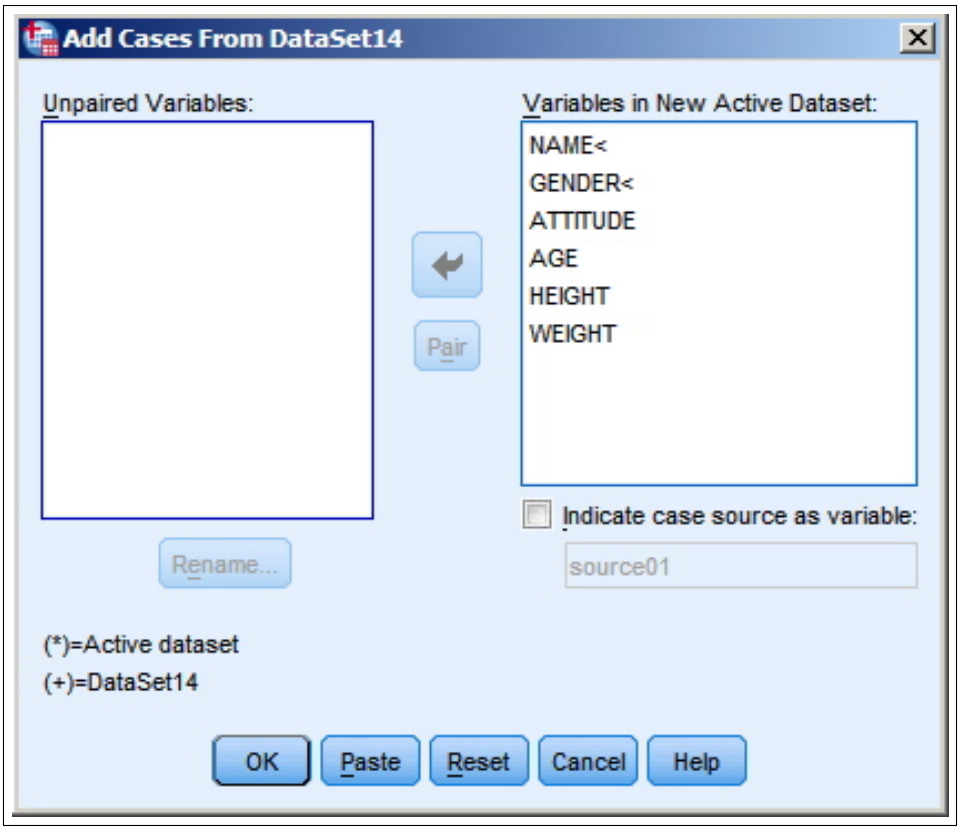


Figure 2.7: Selecting the variables to add.

- If there are variables that don't match between the two datasets, ones in the active dataset will be marked with an asterisk (*), while variables in the dataset being imported will be marked with a plus sign (+).
- If the two datasets share a variable you want merged but they have *different names*, select it in the 'Unpaired Variables' pane and click **Rename**. Enter the name that the variable has in the primary dataset and click OK.
- If you want to keep track of the file each case came from, check **Indicate case source as variable**. This is especially useful if you are merging more than two datasets together.

Click OK and SPSS will import the new cases into `MYCLASS.SAV`. Save the new dataset as `MYCLASS2.SAV` using **File** → **Save As**.

Adding Variables: Sometimes, the data pertaining to our participants exists in multiple files. For instance, we have **MYCLASS2.SAV**, but to our dismay, we learn that all of participant's IQ scores were stored in **IQ.SAV**! Luckily, the participant's names were entered in both files, which we can use as a **key variable**.

This process is similar to adding cases. First, open both datasets, making sure that `MYCLASS2.SAV` is the active window. Second, run **Data** → **Merge Files** → **Add Variables**. This will open a window similar to Figure 2.6, in which we choose the dataset we wish to import the variables from (`IQ.SAV`).

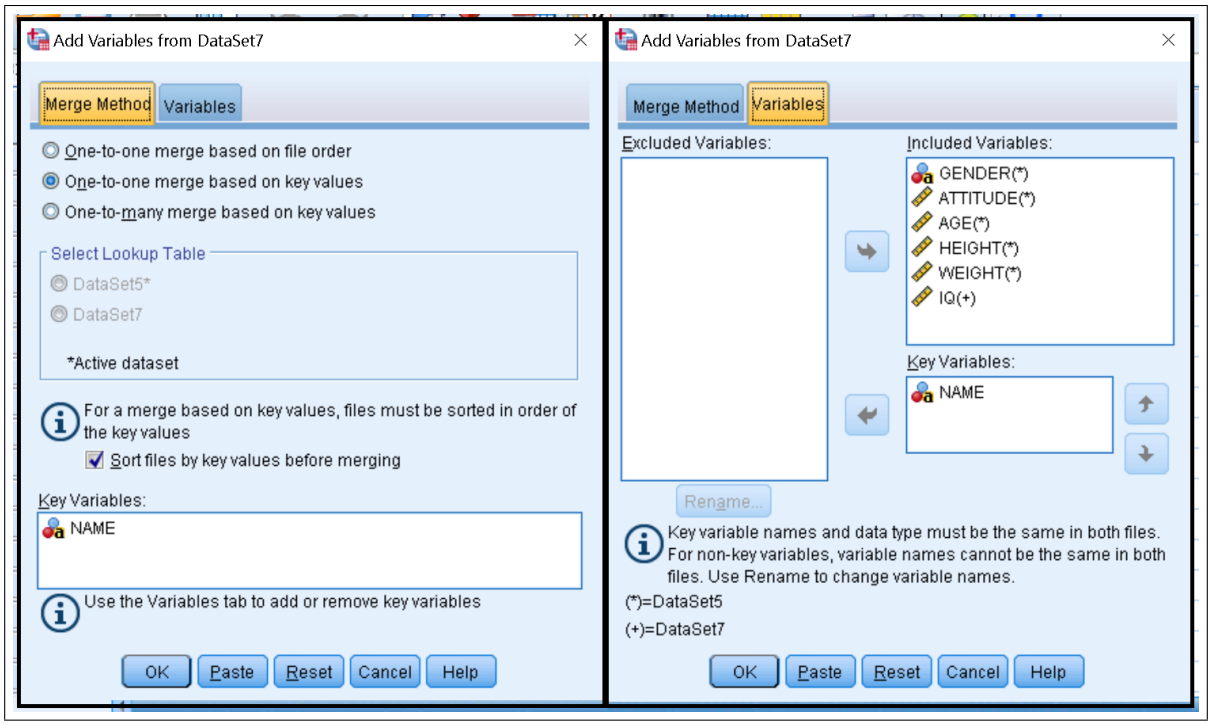


Figure 2.8: The Add Variables dialog box.

- Variables in the two datasets that have the same name will automatically appear in the 'Included Variables' pane.
- If a variable is a **Key Variable**, meaning it can be used to identify particular participants (e.g. by name, or by participant ID) and can be selected in the "Variables" pane.
- If you do use a key variable, make sure to specify how the data is sorted depending on how the data is organized. The safest approach is to have the cases sorted in order of the key variables in both datasets[3], and to have both files provide the case keys. If the data is not sorted, SPSS will try to do it for you during this process.

Exercise: Import the IQ variable into `MYCLASS2.SAV`. Make sure NAME is indicated as a key variable (cases are sorted; both files provide cases). Save the new untitled dataset as `MYCLASS3.SAV`. The resulting datafile should have **25 cases** with **7 variables**.

2.1.5 Sorting Data

With `MYCLASS3.SAV` open and in Data View, it is easy to notice that the data is now in alphabetical order in regard to the participant names. If we would like the data sorted by another variable, this ban be done using **Sort Cases**, which is found under **Data**.

[3]See Section 2.1.5 for more information on sorting a datafile.

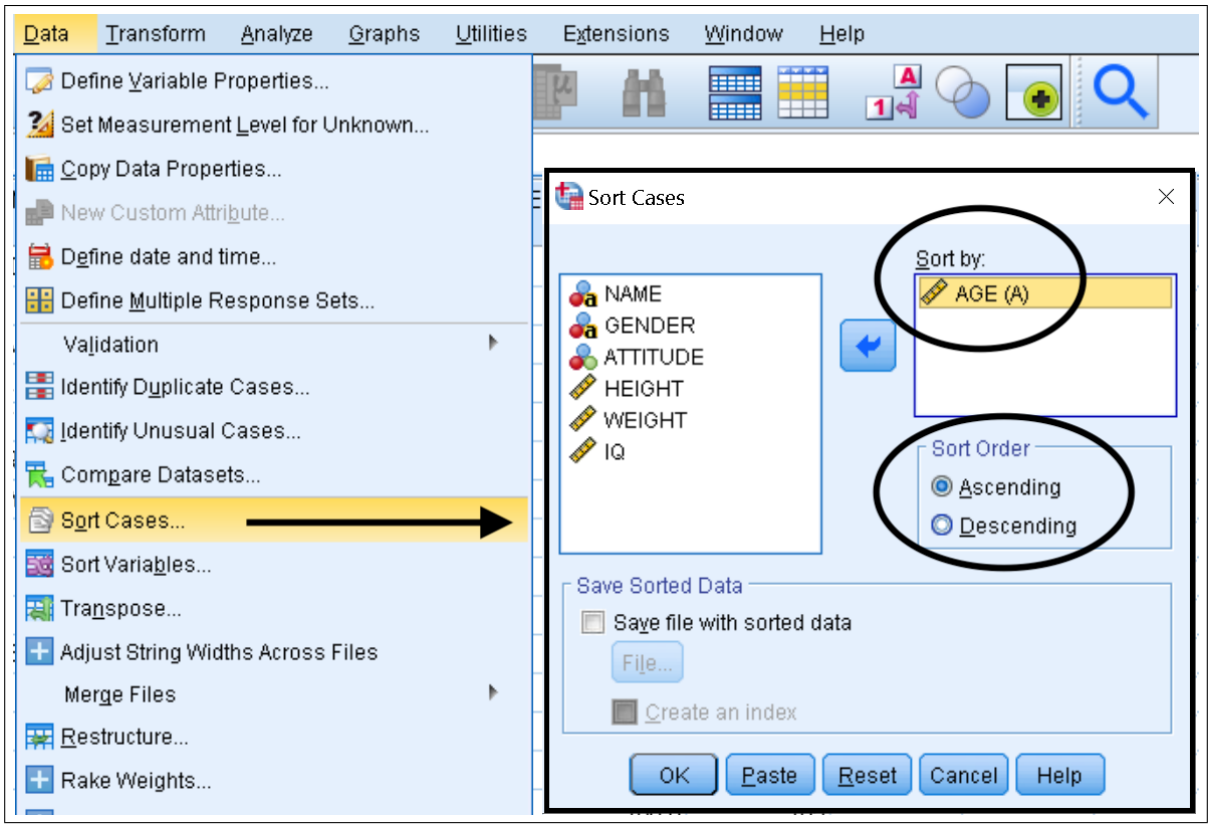


Figure 2.9: The Sorting Cases dialog box.

To Sort:

1. Select the variable to sort by (e.g. 'Name').
2. Click the arrow to bring it to the 'Sort By' field.
3. Select the 'Sort Order' (e.g. 'Ascending' to sort the file alphabetically, from A to Z).
4. Click 'OK'.

Before proceeding, ensure your datafile is sorted by name (from Alfred to William). Save this dataset as `MYCLASS4.SAV`.

2.1.6 Recoding Data

With `MYCLASS4.SAV` open and in Data View, it is easy to see some anomalies in the data. In particular, note how the GENDER variable is coded. What could be improved in this variable?

In this instance, gender was been entered using letters, not numbers! This will cause problems for certain features of SPSS, so it would be beneficial to recode the variable.

This involves selecting **Transform → Recode into Different Variables**.

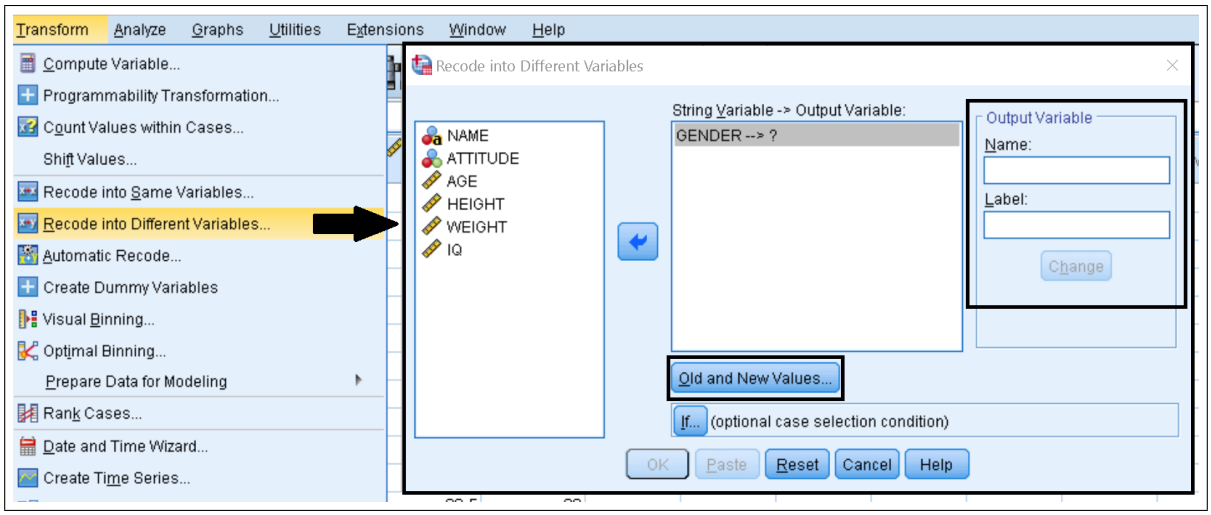


Figure 2.10: Recoding a Variable into a Different Variable.

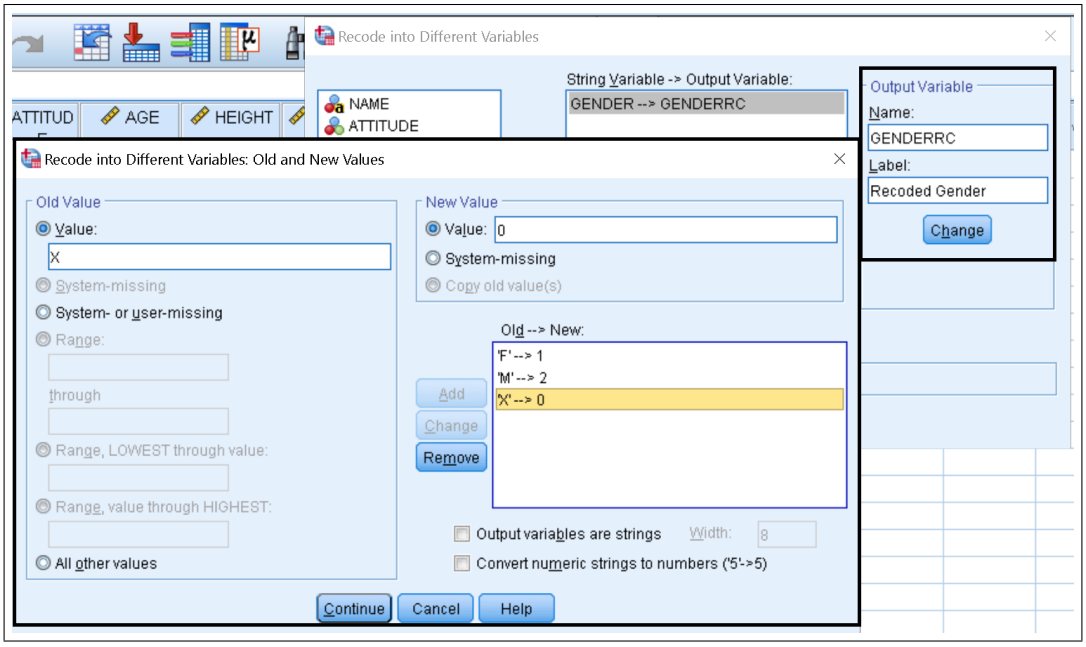


Figure 2.11: Recoding a Variable into a Different Variable, Step 5.

Notes on Recoding:

- It is good practice to recode a variable into a *different variable* (e.g. recoding GENDER into the new variable GENDERRC). SPSS can recode a variable into the *same variable*, however if you make a mistake in the recoding the original data might be over-written for good!
- More than one variable can be recoded simultaneously! This is extremely useful for recoding a series of reverse coded Likert scale items.
- The new recoded variable can be defined in a variety of ways. For instance, the new values could be based upon particular values (as in our example), a range of values (e.g. converting IQs lower than 90 into 'low'), properties of the old values (e.g. by their missingness, or otherness).

Exercise: Make sure you have properly recoded GENDER into 1s (for female) and 2s (for male), and 0s (for other). Define the variable properties in Variable View (remember to include value labels!). Save this dataset as `MYCLASS5.SAV`.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	NAME	String	7	0		None	None	7	Left	Nominal	Input
2	GENDER	String	1	0		None	None	6	Left	Nominal	Input
3	ATTITUDE	Numeric	8	2		None	None	8	Right	Nominal	Input
4	AGE	Numeric	2	0		None	None	8	Right	Scale	Input
5	HEIGHT	Numeric	4	1		None	None	8	Right	Scale	Input
6	WEIGHT	Numeric	5	1		None	None	8	Right	Scale	Input
7	IQ	Numeric	8	0		None	None	8	Right	Scale	Input
8	GENDERRC	Numeric	8	2	Recoded Gender	{.00, Other}...	None	10	Right	Nominal	Input

Figure 2.12: Recoded Gender in Variable View.

2.2 Data Filtering and Weighting

As will be used in some of the later exercises, it is often of interest as an analyst to look at particular subsets of data. For instance, we might be interested in gender differences, or about the intelligence of participants who have attitudes over a particular threshold, and we might want to see the results of our analyses looking only at a subset of individuals. SPSS has a few ways to approach such analyses.

2.2.1 Selecting Cases

Perhaps the most well-used approach for filtering a dataset is to simply use **Select Cases**, accessed via **Data** → **Select Cases**. The method for selecting cases can be customized, as shown in the image on the next page.

The default selection in the the **Select Cases** dialog box is for **All Cases** to be displayed. This means that SPSS will use all available data in each subsequent analysis.

This is useful to remember in case you run **Select Cases** and later wish to return to analyzing the entire dataset!

The most common approach if you wish to select cases is to use an IF statement based upon a categorical variable in the dataset. For instance, we could select all the females in the dataset by clicking on **If condition is satified** and specifying `GENDERRC = 1`. Likewise, if we wanted to analyze only the data from male participants, we would enter instead `GENDERRC = 2`.

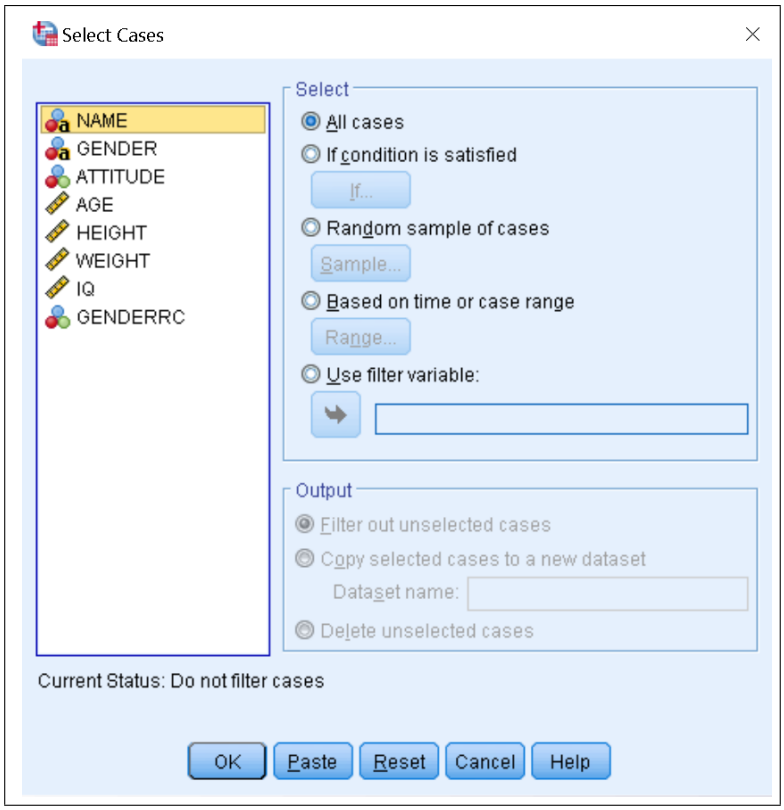


Figure 2.13: Using Select Cases to Filter a Dataset.

Similarly, through this dialog box we could request a random sample of cases (e.g. 90%), or a specific range of cases (e.g. only use Case 1 through 10).

Finally, we could use a **Filter Variable**. This is a variable that is a series of 0s (indicating *not* selected) and 1s (selected). If you have previously run **Select Cases** using any of the above selection, a filter variable will be automatically generated for you with the variable name `filter_$` (which you can then rename to something more meaningful).

After defining a subset of cases, in the 'Output' portion of the window you can choose whether the unselected cases should be filtered (typical behaviour), or deleted entirely.

Back in Data View, if you asked for the unselected cases to be filtered, you should see a diagonal strikethrough overlaid on unselected case row numbers. You should also note that in the Status Bar, a **Filter On** message has been enabled, and a new variable, `filter_$` has been created. For selected cases, `filter_$` = 1, for unselected cases the `filter_$` = 0. This new variable can be edited in Variable View (e.g., it is often useful to change its name and label), and saved for future use (applied via **Data** → **Select Cases**, and then 'Use filter variable').

To de-activate the filter, return to the **Data** → **Select Cases** dialog box and change the filter selection back to **'All cases'**. The 'Filter On' message in the Status Bar should disappear.

Exercise: Using `MYCLASS5.SAV`, try selecting only individuals with an ATTITUDE score greater than or equal to 50 then find their mean IQ.

2.2.2 Using Split File

When particular subsets are of primary interest, it is often desirable to set a global option pertaining to data filtering. In this regard, the **Split File** feature can be convenient and time saving as it literally will split our output into a series of groups, that can each be analyzed separately or compared against each other. For instance, this could be useful if you have a large dataset with a variable that pertains to a range of ethnic backgrounds. Using split file, our output could automatically be delimitated by ethnicity.

This feature is accessible via **Data** → **Split File**, which produces the following dialog box:

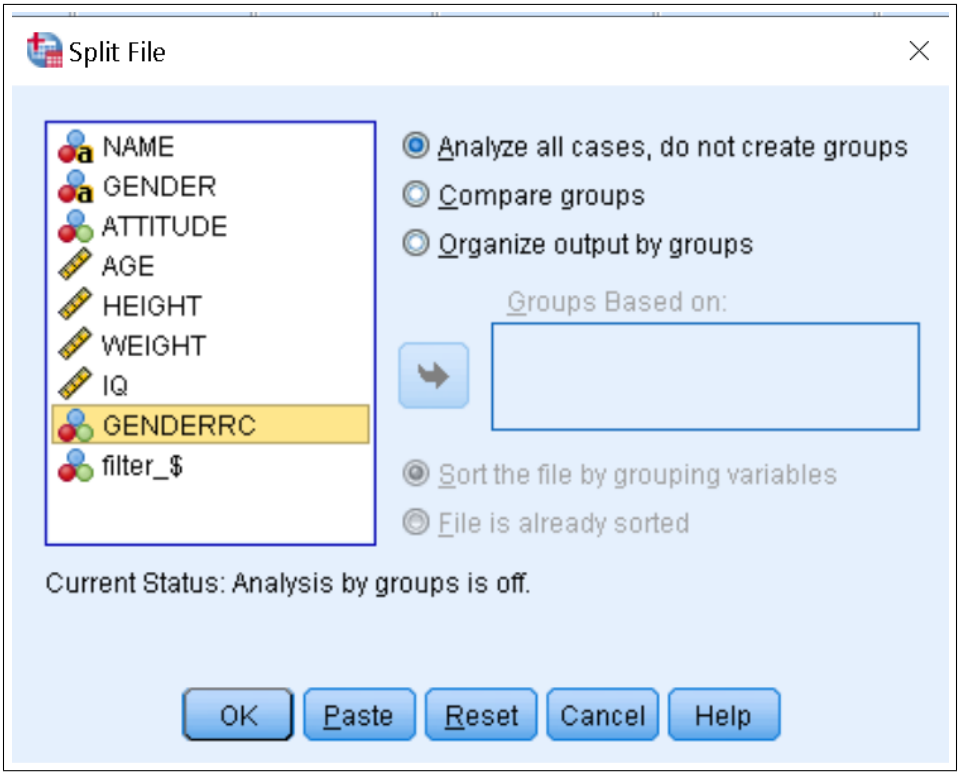


Figure 2.14: Splitting a Dataset into Separate Files.

Split File Options:

- *Analyze all cases, do not create groups* – Default behaviour, nothing is split.
- *Compare groups* – Split file groups are presented together for comparison purposes.
- *Organize output by groups* – All results from each procedure are displayed separately for each split-file group.

Exercise: Using the full dataset `MYCLASS5.SAV` (make sure the ATTITUDE filter has been disabled!), use **Split File** to compare output by groups, with GENDER as the basis. Use **Analyze** → **Descriptive Statistics** → **Frequencies** to request the mean ATTITUDE scores. Rerun the same analysis, but having using **Split File** to organize output by group, and note the differences in output. Finally, turn off **Split File**.

2.2.3 Weighting Observations

By default, each case in an analysis has a weight of 1. This is often appropriate, especially if our data pertains to individual observations. However, sometimes observations should be weighted. For example, if our dataset pertained to different cities, we might want to weight each city by its respective population in our analysis. Similarly, in sample surveys, observations are thought to selected through a random process, but different observations may have different probabilities of selection[4]. In this case, weights are equal (or proportional) to the inverse probability of being sampled. In either case, the goal of weighting is to ensure that our observations are as comparable or unbiased as possible, which might not be the case if the weightings were omitted.

Assuming we have a weighting variable in the dataset, we can instruct SPSS to 'weight' each case by the value of this variable in all available SPSS procedures.

As a quick (but entirely unrealistic) exercise to demonstrate the function of the **Weight Cases** procedure, create a new variable in `MYCLASS5.SAV` creatively called WEIGHTS. Pseudo-randomly assign each individual a weight value between 1 and 3. The use **Data** → **Weight Cases** to weight the cases by this new variable.

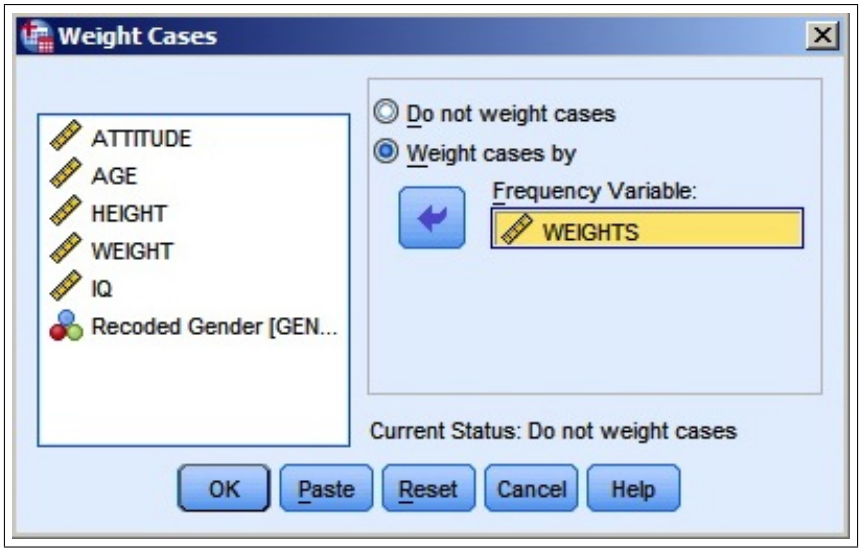


Figure 2.15: Weighting Variables using Weight Cases.

After you make this selection, note in the **Status Bar** that a 'Weight On' message now appears (see Figure 3.22). From now on, the weighting of cases will be applied in all statistical data analyses. As an example, look at output from the **Frequency** procedure on the next page. It was asked to produce a frequency table of the NAME variable. Notice how the frequency (or weight) of the individual observations now range between 1 and 3, and are worth between 2.2 and 6.7 percent, instead of all being weighted equally.

Remember: The larger the weight, the more influence that observation has on the analysis!

[4]For instance, maybe statisticians are more likely to respond to a cold call survey!

NAME

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ALFRED	1	2.2	2.2	2.2
	ALICE	2	4.4	4.4	6.7
	BARBARA	3	6.7	6.7	13.3
	CAROL	2	4.4	4.4	17.8
	HENRY	2	4.4	4.4	22.2
	HOGARTH	2	4.4	4.4	26.7
	HUXLEY	1	2.2	2.2	28.9

Figure 2.16: An Excerpt from the NAMES Frequencies Table, Weighted.

2.3 Advanced Data Manipulation

Before beginning this section, ensure that all weights, filters, and split file commands have been disabled.[5]

2.3.1 Using COMPUTE Statements

The COMPUTE feature, found under **Transform** → **Compute Variable**, is one of the fundamental SPSS workhorses for data manipulation. In general, you will specify a 'Target Variable' and some expression which defines it.

The expression can be based upon pre-existing variables (e.g. converting weight in pounds to kilograms, or summing scale items into an overall score), arithmetic (operations and functions), or chosen from a multitude of pre-defined functions (e.g. generating random numbers, cumulative distribution probabilities, dates, etc.).

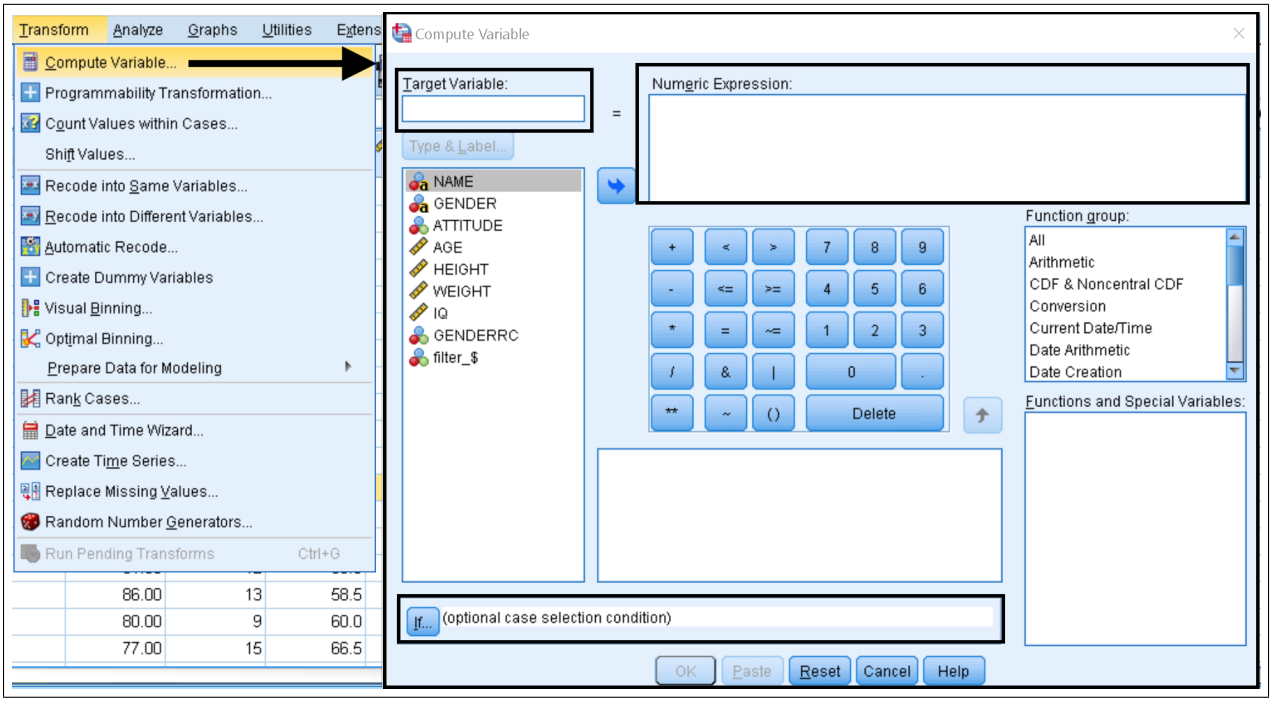


Figure 2.17: The Compute Variable dialog box.

[5]Remember, you can look at the Status Bar to quickly observe if any of these are active!

Steps for Computing a Variable:

1. In the **Compute Variable** dialog box, supply the name of the new variable you wish to create as the **Target Variable**.
2. Define the data type and labels for the new variable via the **Type & Label** dialog box.
3. Build up the **Numeric Expression**. This can be done by pasting, typing, or using the available functions (accessible via the lists on the right).
4. Click 'OK' to compute the new variable, or 'Paste' to view the command's syntax.

Commonly Used Operations and Functions:

Arithmetic Operations	**Arithmetic Functions**
+ → Addition	ABS(arg) – Absolute Value
− → Subtraction	RND(arg) – Round to nearest integer ($-4.7 \rightarrow -5$)
∗ → Multiplication	TRUNC(arg) – Truncate to an integer ($-4.7 \rightarrow -4$)
/ → Division	MOD(arg1,arg2) – Modulo (remainder of arg1/arg2)
∗∗ → Exponentiation	SQRT(arg) – Square Root
	EXP(arg) – Raise e to the power of the argument
Logical Operators	LG10(arg) – Base 10 logarithm
< and > → Less/Greater	LN(arg) – Natural log (base e)
<= and >= → L/G or Equal to	ARSIN(arg) – Arc Sine transformation (in radians)
= and ˜= → Equal/Not Equal to	ARTAN(arg) – Arc Tangent transformation (in radians)
& → And	SIN(arg) – Sine transformation (arg. must be in radians)
| → Or	COS(arg) – Cosine transformation (arg. must be in radians)

One of the primary uses of the COMPUTE variables dialog box is to transform variables. Many statistical analyses assume that your data are relatively normally distributed.

If you have variables where this is not the case (e.g. a salary variable might be highly positively skewed, with many respondents of aveage income but a few with extremely high incomes), it is often a good idea to transform the variable to make it more normal. For example, taking the square root or log of a positively skewed variable will help normalize it.

Example COMPUTE Statements:

- `TOTALSCORE = ITEM1 + ITEM2 + ITEM 3 + ITEM4 + ITEM5`
- `AVERAGE1 = MEAN(ITEM1, ITEM2, ...)` – found in the Statistical function group.
- `AVERAGE2 = MEAN(ITEM1 TO ITEM5)` – only works if items are together in data list.
- `ABSVAL = ABS(ITEM)` – found in the Arithmetic function group.

Using Statistical Functions:

- SPSS also has a set of statistical functions built-in to this dialog box (summarized on the next page).
- Each argument to a statistical function (expression, variable name, or constant) must be separated by a **comma** (e.g. MEAN(A,B,C,D).
- The optional **[.n]** suffix can be used with all statistical functions to specify the number of valid arguments required for the calculation.
  - For example, MEAN.2(A,B,C,D) returns the means for the valid values for variables A, B, C, and D but *only if at least two* of the variables have valid values (non-missing).
- The keyword **TO** can be used to refer to a set of variables in the argument list.

Commonly Used Statistical Functions:

SUM[.n](arg list) – Sum of non-missing values across argument list.
MEAN[.n](arg list) – Mean of non-missing values across argument list.
SD[.n](arg list) – Standard deviation of non-missing values across argument list.
VARIANCE[.n](arg list) – Variance of non-missing values across argument list.
CFVAR[.n](arg list) – Coefficient of variation of non-missing values (SD/MEAN).

MIN[.n](arg list) – Minimum non-missing value across the argument list.
MAX[.n](arg list) – Maximum non-missing value across the argument list.
NMISS(arg list) – Count of missing values across the argument list.
NVALID(arg list) – Count of valid values across the argument list.

Exercise: Using `MYCLASS5.SAV`, compute a new variable WEIGHTKG, which is based upon WEIGHT multiplied by 0.4536.

2.3.2 Arithmetic Operations vs. Statistical Functions

When using COMPUTE, there are often multiple ways of accomplishing a particular task. For example, if we were interested in calculating a sum score for a participant on a 5 item survey, we could do this in two ways – either via Arithmetic Operations or with a Statistical Function.

The arithmetic approach would be to enter:

```
SUM = ITEM1 + ITEM2 + ITEM3 + ITEM4 + ITEM5
```

While using a **function**, we could enter:

```
SUM = SUM(ITEM1,ITEM2,ITEM3,ITEM4,ITEM5)
```

Or, more concisely (if the items are in order in the datafile),

```
SUM(ITEM1 TO ITEM5)
```

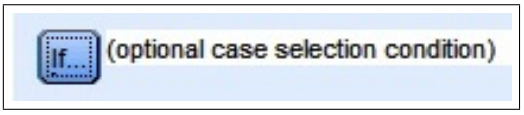
You might think that these two approaches would yield the same result. Often they do, but **not necessarily so**! The main difference is in how they deal with missing values. If the **arithmetic approach** comes across a missing value for a participant, it will **stop** – and place a missing value for the sum, even if only one value out of 50 is missing!

Contrarily, the **statistical functions** are designed to handle missing data. Depending on the procedure, they will simply ignore the missing cell (e.g. the SUM function will add up the remaining cases), or they will modify the procedure to account for the missingness (e.g. the MEAN function will calculate the average for the participant out of 4 instead of 5). As mentioned before, you can use the [.n] suffix to tell SPSS the minimum number of valid arguments required.

Exercise: Import `TEST.DAT`, with the following variable names: ID, GENDER, AGE, INCOME, and SCORE1 through SCORE6. Account for the missing values in the dataset (for the SCORE variables, 9 and -1 indicate missingness). Calculate a sum score with both the arithmetic approach and using SUM. Save the data as `MYTEST.SAV`.

2.3.3 The Logic of IF Statements

Another way of manipulating data through the COMPUTE dialog box is with **IF** statements, via the **IF (optional case selection condition)** dialog box. These use the logical operations listed on the previous page. In general, they have the form:




```
IF(logical expression) then TARGET VARIABLE = EXPRESSION
```

To reiterate, the following relations and logical operations are available:

EQ or = [equal to]	LE or <= [less than or equal to]	& [AND]
LT or < [less than]	NE or ~= [not equal to]	| [OR]
GT or > [greater than]	GE or >= [greater than or equal to]	! [NOT]

Example COMPUTE IF Statements:

- `IF (X LT1) Y1=1` → for all X values equal or greater than 1, Y1 is not defined (system-missing). Otherwise, Y1 = 1.
- `IF (X=1 OR X=2) Y3=3` → Checks X for 1s and 2s. If found, sets Y3 to 3.

Using COMPUTE IF To Recode a Variable:
An applied example of when this might come in handy is if you wanted to categorize a continuous variable. For example, we code recode an AGE variable into a new variable, AGECAT, with 'low', 'medium', and 'high' categories.[6] For instance, maybe we want 'low' to pertain to ages 30 or less, 'middle' from 31 to 40, and 'high' to be greater than 40.[7]

Exercise: Open `MYTEST.SAV`

To accomplish this we need to go **Transform** → **Compute Variable**. As before, type in the new target variable name (AGECAT). For the **numeric expression**, this will either be 0, 1, or 2 (pertaining to 'low', 'medium', and 'high').

For the first run through, type 0 into the **numeric expression** box. Then open the **IF** dialog box. Select 'Include if case satisfies condition' and type or paste **AGE <= 30**. Click 'Continue'.

When you select 'OK' in the **Compute Variable** window, the Data Editor will show a partially filled new column with your new variable name.

[6] But, in real life, please don't! In almost all situations, working with a continuous variable is substantially better than chopping it into a categorical one.

[7] Of course, this is even easier to do with the RECODE function, but this illustrates how IF/ELSE style programming works.

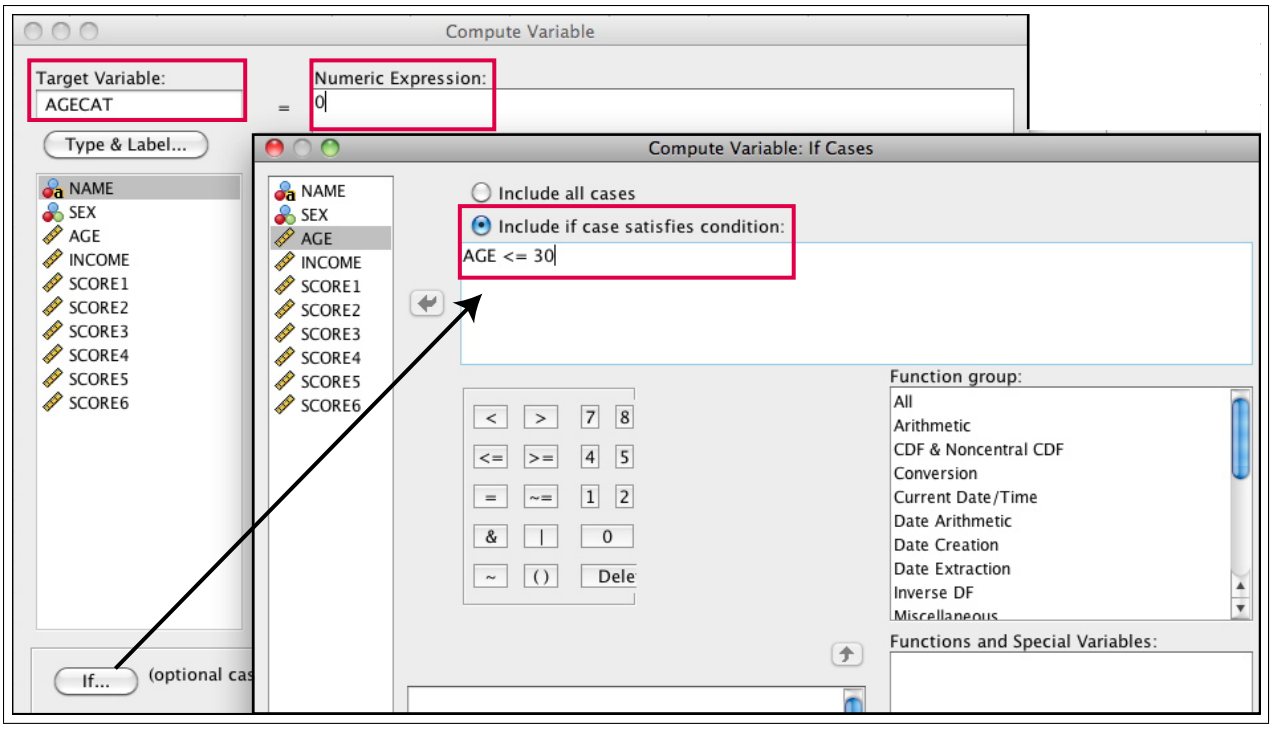


Figure 2.18: Computing the 'Low' Category.

Now, we have to repeat the procedure for the 'medium' and 'high' categories. For medium, you will want to include the condition **AGE >= 31 & AGE <= 40**, which should yield a 1 in the **numeric expression** box.

Note: when you click 'OK' on this, and any subsequent, run throughs SPSS will give you a warning about how you are changing an existing variable. Click 'OK' to get rid of the warning, as our goal is to fill in some of the blanks of the new variable.

Exercise: Figure out how to apply the 'high' label to the dataset. Make sure your new variable is properly defined – with value labels. Save the data as `MYTEST2.SAV`.

2.3.4 The COUNT Command

The COUNT command does what it sounds like - it will create a new variable, which will contain the counts of a particular attribute of the other variables in the dataset. It is accessible via **Transform → Count Values within Cases**.

Countable attributes (selectable through the **Define Values** dialog box) include:

- A particular VALUE (e.g. the number of responses that were 2s)
- The number of values within a range (e.g., between two values, or from a specified value through the LOWEST or HIGHEST values)
- The number of SYSMIS (system missing) values in the dataset.
- The number of MISSING (user defined missing, as well as system missing) values in the dataset.

Exercise: Using `MYTEST2.SAV`, use the **COUNT** procedure to determine the number of missing values in the SCORE variables.

2.3.5 Restructuring Data from Wide Format to Long

Most datasets come in 'wide format', with each row pertaining to a different unit of observation (e.g., participant). However, some analyses[8] require the data to be set-up in a different manner. This orientation still uses rows for observations, but now information from a participant can span multiple rows – typically, one for each time point that they participate in during the study. This is called 'long format'.

Exercise: Open the dataset `WIDE.SAV` and note how the variables are organized.

Long format for this dataset means that we want three rows for each participant (one for each time point), rather than just one. To do this in SPSS is relatively easy using the **Restructure Data Wizard**, which is found via **Data** → **Restructure**.

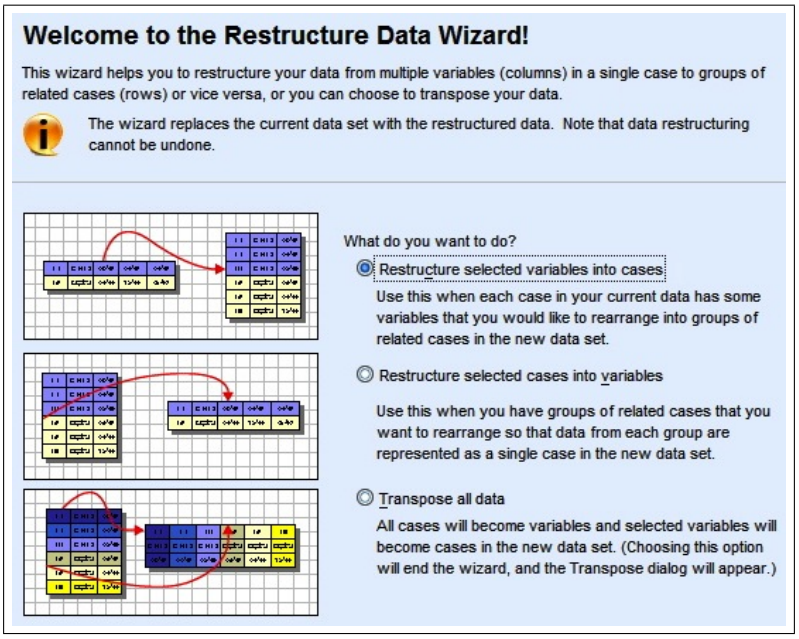


Figure 2.19: The Restructure Data Wizard.

Using the Wizard is fairly straight forward. Choose how you want to do the restructuring (look at the pictures for a reference on what each option does; in this example we want the first option – 'restructure selected variables into cases'), and click 'Next'.

[8] Most importantly in multilevel modeling with data that features repeated observations.

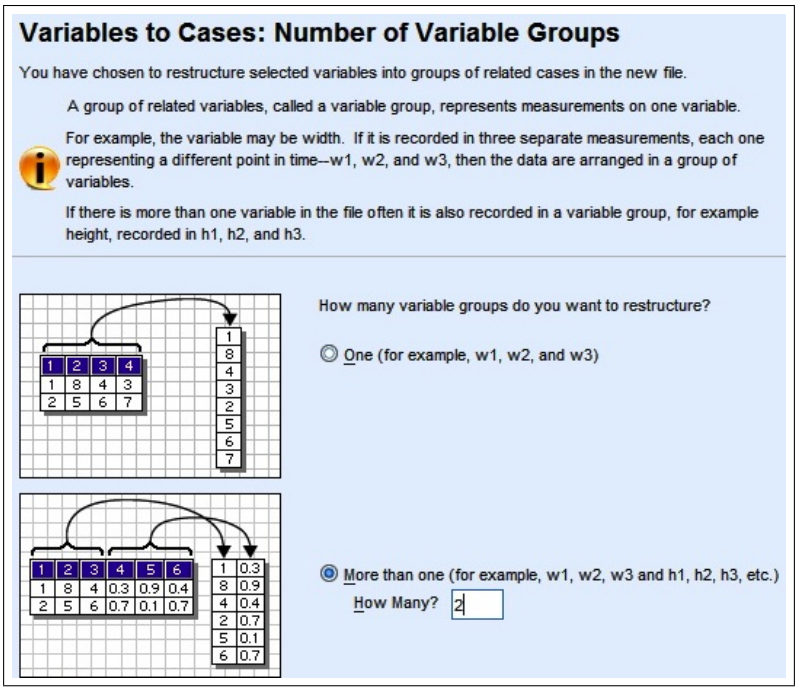


Figure 2.20: The Restructure Data Wizard: Variables to Cases.

In this pane, you need to tell SPSS how many time varying variables you have (there are 2 in `WIDE.SAV`, one for the 'study' variable, and one for 'score').

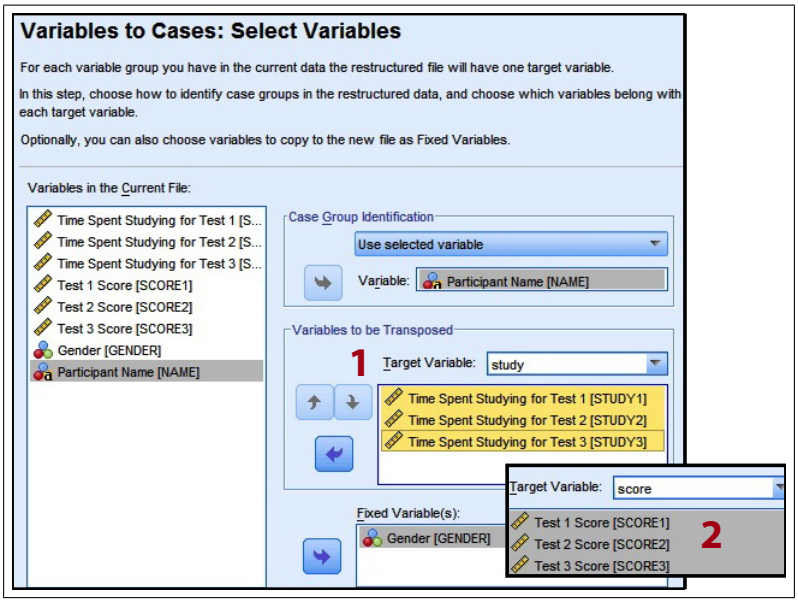


Figure 2.21: The Restructure Data Wizard: Step 3.

In this pane you need to define some form of case identification (usually a participant ID variable, or use the case number); the variables that require transposing (note: you need to define *both* STUDY and SCORE here!); and, the variables that are fixed across time points (e.g., GENDER).

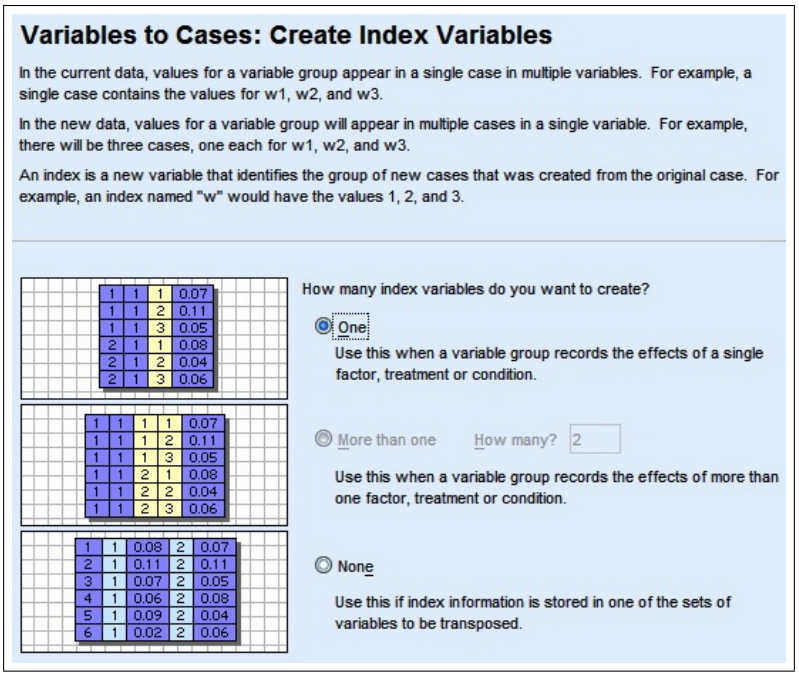


Figure 2.22: The Restructure Data Wizard: Step 3.

Since all of the variables being transposed are either from Time 1, 2, or 3, we only need one index variable here – which will take on the values 1, 2, or 3. A few more options follow this one, but the defaults are fine (and what they do is relatively straightforward). When you finish with the Wizard, you should return to **Data View**. Your data should now look like this:

NAME	GENDER	Index1	study	score
Lulu	.00	1	200.00	3500.00
Lulu	.00	2	190.00	3300.00
Lulu	.00	3	195.00	3100.00
Hoga	1.00	1	160.00	3000.00
Hoga	1.00	2	150.00	2900.00
Hoga	1.00	3	150.00	2800.00
Mya	.00	1	180.00	3050.00
Mya	.00	2	175.00	3100.00
Mya	.00	3	170.00	2900.00

Figure 2.23: Long Form Data.

Exercise: Now that we have converted our data into long format, try to reverse the process. Use the **Restructure Data Wizard** to return the cases to wide format!

By this point, you now should be able to work through both Hands On Exercises I and II. Good luck!

3 Statistical Analyses

3.1 Introduction to Data Analysis in SPSS

After inputting, manipulating, and cleaning data, the next step in a typical data analysis pipeline is to figure out what to *do* with it. This often involves: summarizing the dataset to make sure everything looks as it should, producing some representative graphics, and then running statistical analyses. The remainder of the course will be devoted to opening data, and learning about how to run, configure, and interpret some basic to intermediate level statistical procedures.

> Unless otherwise specified, the following techniques and exercises will be demonstrated using the `1991 US Social Survey.SAV` data file.

3.1.1 Getting Help!

Before getting into the thick of data analysis, it should be quickly reiterated that SPSS has some fantastic facilities for aiding you. Beyond the manuals and tutorials found in the menubar under **Help**, each dialog box has a 'Help' button. Clicking this will launch your default web browser and bring you to a page that relates to the dialog box you were working with. Further, in the output from a statistical analysis, you can double click on a table to select it, and then right click on a particular element in the table. If you select 'What's This?', SPSS will give you a brief definition of the element in question.

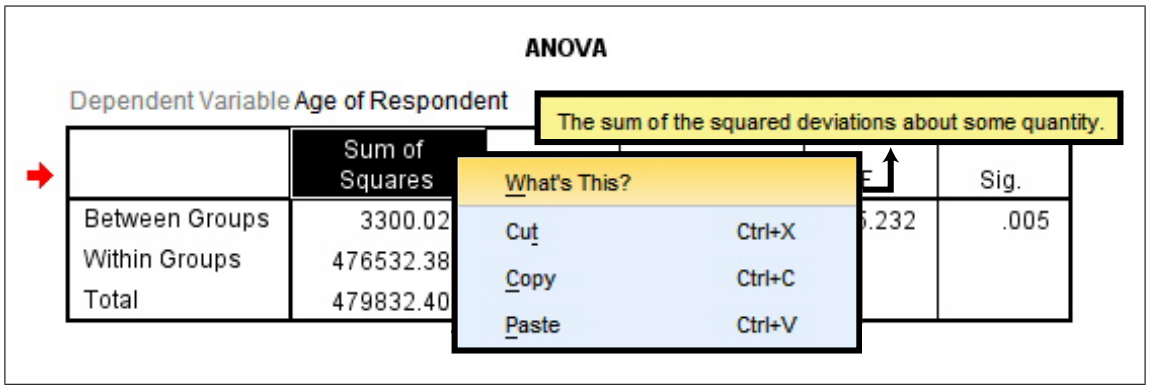


Figure 3.1: 'What's This?' Style Help.

3.1.2 Bootstrapped Estimates

As you work through the following sections, you will notice a button that will reappear over and over again on the statistical analysis dialog boxes. This button is titled 'Bootstrap' and allows the user to request what are known as 'bootstrapped estimates' for coefficients from their statistical analysis. This is thought to be a method

for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation and regression coefficients, and for constructing hypothesis tests.

The idea is to treat your dataset (or sample) as a population of scores, instead of a sample from a population. The procedure then randomly samples from your dataset, *with* replacement, a predetermined number of times (or for a set amount of time). For instance, we might request 1,000 samples from our dataset, each being an independent draw. The bootstrapped estimate for the mean, for example, is then the average of the means of the 1,000 estimates.

This is thought to be most useful as an alternative to parametric estimates when the assumptions of those methods are in doubt (e.g. when using regression models that have heteroscedastic residuals fit to small samples). This is a useful addition to SPSS and while it won't be used during the applied examples in this course, it is something worth being aware of.

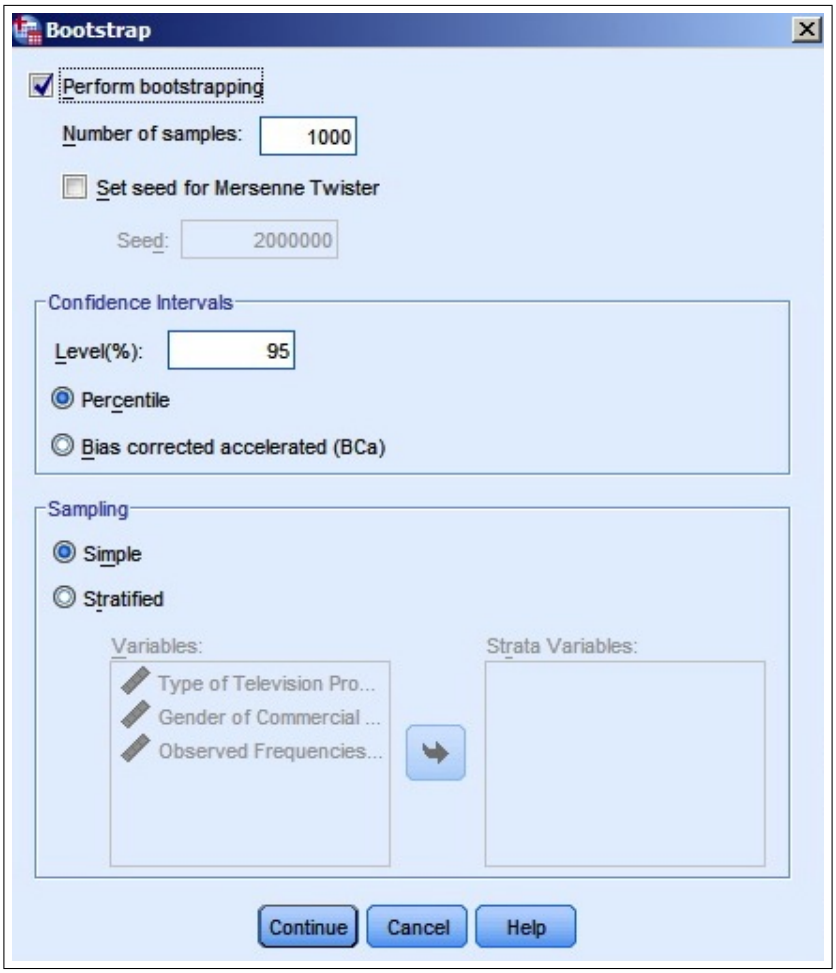


Figure 3.2: Requesting Bootstrapped Estimates for an Independent Samples t-test.

3.2 Summarizing a Dataset

There are multiple approaches to summarizing a new dataset. I will present four in fairly rapid succession, so you can see their similarities and differences. The choice between them is largely one of personal preference. The approaches are to use the **Codebook** function, to run **Case Summaries**, to use the **Explore** procedure, or to simply begin with **Descriptive Statistics**.

3.2.1 Using the Codebook

The **Codebook** is a basic tool for surveying a new dataset. It produces a report that summarizes all of the variables in the dataset, one at a time. One nice feature about the **Codebook** is its ability to include meta-data, or information about the file itself (e.g. the datasets location on the hard drive, when it was last modified, et cetera). This function is found under **Analyze** → **Reports** → **Codebook**. The initial screen is used for selecting the variables you want analyzed and what would you like to see in the output:

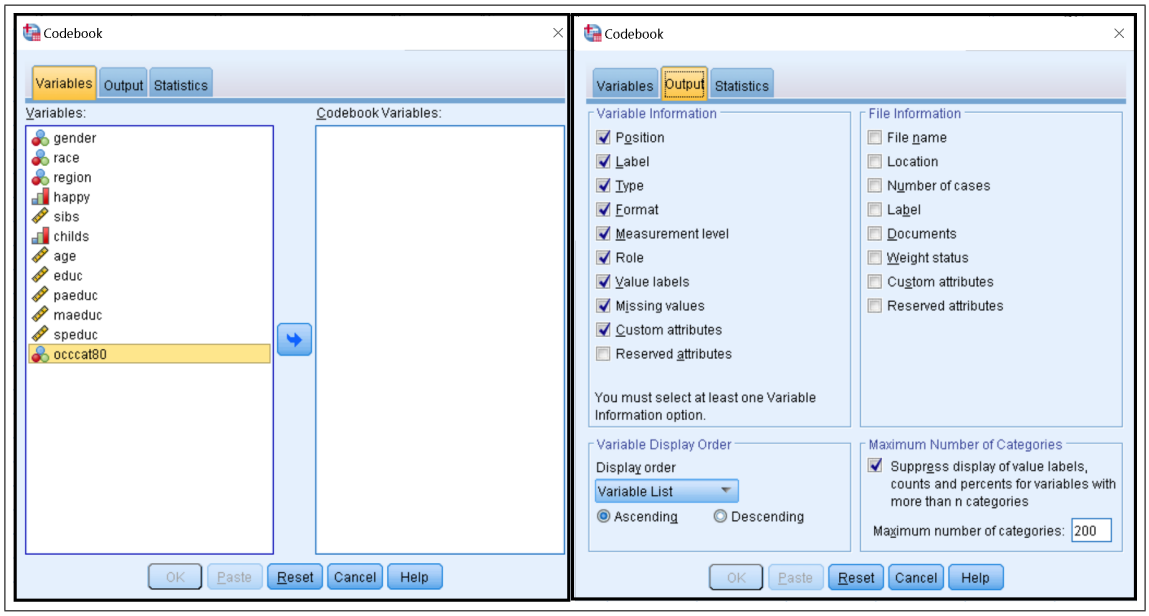


Figure 3.3: The Codebook interface: Selecting variables and Output.

The final **Statistics** tab allows you to toggle counts and some descriptive statistic calculations.

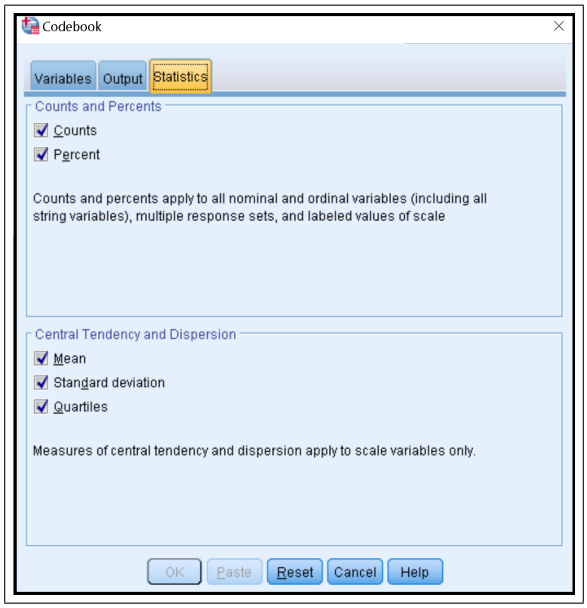


Figure 3.4: The Codebook interface: Statistics.

Note the 'File Information' section, which allows you to store information about the datafile itself. Also note the somewhat lacklustre selection of statistical options. For now, the default options for both of these tabs is fine. When you press 'OK', the procedure will produce something like the following in your SPSS Output Viewer window:

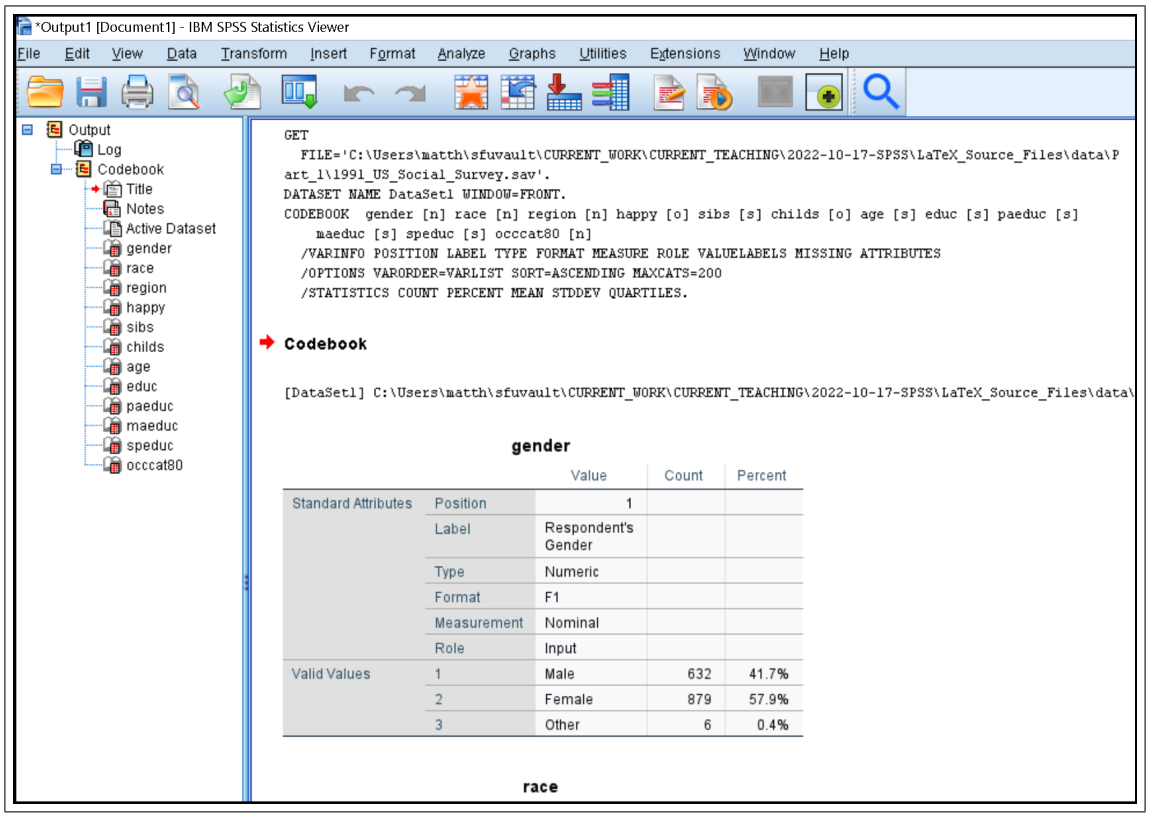


Figure 3.5: Codebook Output.

Exercise: Produce a basic **Codebook** for the `1991 US Social Survey.SAV` datafile .

Notice how each variable in the **Codebook** output has its own heading in the navigation tree, and that the options we chose in the **Output** tab appear for each. Scroll through the output, and note how it differs for continuous variables.

3.2.2 Preparing Case Summaries

In contrast, **Case Summaries** lumps variables together, allowing for easier direct comparisons between them. This function is also found under **Analyze** → **Reports**. Notable features of **Case Summaries**:

- The ability to limit the summary to a maximum number of cases. This procedure will literally output a table with each participant's response for the chosen variables, so you may want to simply uncheck 'Display cases'.
- The ability to specify a grouping variable. This allows the summaries to be partitioned by some nominal variable (e.g. show all of the college student responses in a separate summary than the high school student responses).
- A *wide* variety of statistical procedures are available from the 'Statistics' tab!

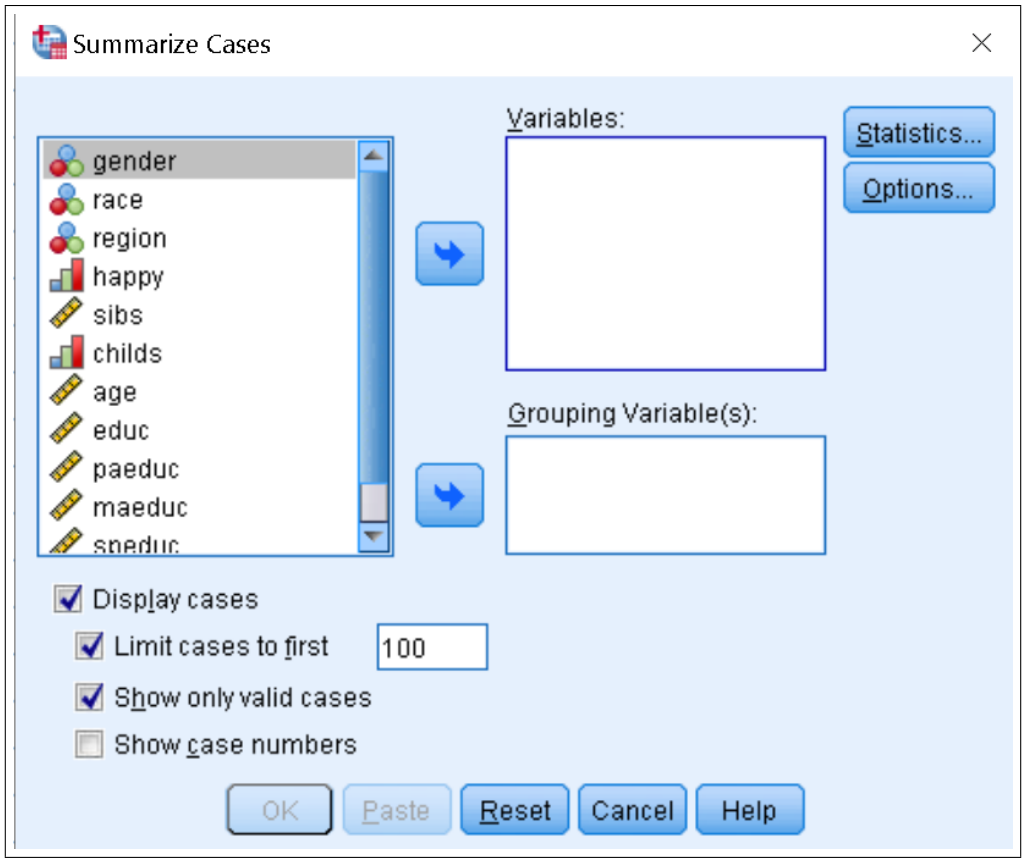


Figure 3.6: The Case Summary interface.

The output from **Case Summaries** is divided into two major components: a 'Case Processing Summary', which totals the number of valid and missing cases out of the total asked for (e.g. if you limited the summary to 100 cases, these totals will only pertain to the first 100 cases!); and, a 'Case Summaries' section. The later begins with the case list of responses (if it was asked for), and has a numerical summary for each variable at the bottom with the requested statistics. This has a benefit over the Codebook, since it allows a side by side comparison of the variables, using statistics not available from the other procedure.

Exercise: Use **Case Summaries** to display the first 75 cases of the Respondent's GENDER, RACE, and HAPPINESS, grouped by where the live (REGION). Make sure to obtain the mean, standard deviation, and skewness. Are all of the regions of the United States in the dataset represented by this summary?

3.2.3 Using Explore

Many textbooks rely on the **Explore** feature to introduce datasets. It doesn't have wonky defaults like **Case Summaries** (if you don't notice 'Limit cases to first 100' you might get some surprising results!), and has more flexibility in terms of available procedures than the **Codebook**. Other benefits include the ability to screen the data, visually examine the distributions of values for the total population or for various groups, and request tests for normality and homogeneity of variance. It is found under **Analyze** → **Descriptive Statistics** → **Explore**.

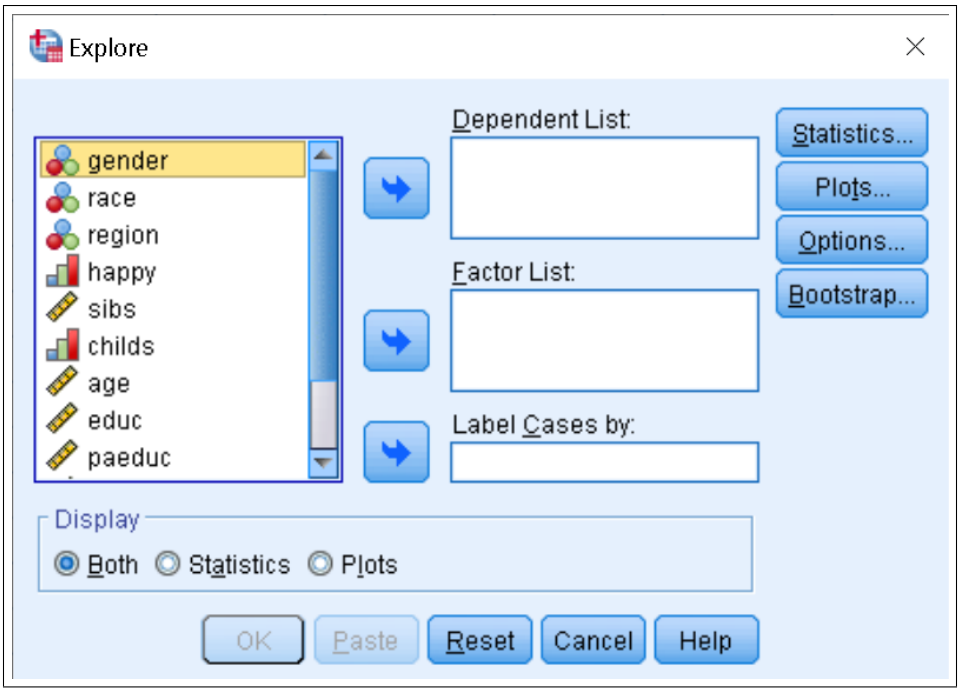


Figure 3.7: The Explore Procedure.

If you simply want univariate displays of particular variables, add them to the 'Dependent List' box. If you have a grouping variable (e.g. you might have a control and a treatment group), you can add it to the 'Factor List' to have the output separated by it. The 'Label cases by' feature allows you to specify an identification variable to be used in the plots (e.g. by their name).

The 'Statistics' option box has four selections that can be toggled:

- **Descriptive Statistics** - This gives a standard set of output (some measures of central tendency, dispersion, and shape).
- ***M*-estimators** - Display a selection of 'robust alternatives' to the sample mean and median for estimating the location.
- **Outliers** - Display the five largest and smallest values with case labels.
- **Percentiles** - Display the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentiles.

In addition to a useful array of statistics, **Explore** is also capable of producing a variety of useful plots. By default, the procedure produces a boxplot (with factor levels together), and a series of stem-and-leaf plots for each variable.

You can ask for the boxplot to be plotted with the 'Dependents together', which simply means that each variable will appear in the same plot as opposed to separate ones. You can also request histograms, and normality plots with tests, as well as spread vs. level plots that allow you to assess the impact of various common transformations on your data.

Exercise: Use **Explore** to look at the number of children the survey respondents had, and their age. Factor the output by the respondant's gender. Make sure to request histograms as well as normality plots with tests, and try to interpret the output.

3.2.4 Descriptive Statistics

As mentioned in Section 1, there are two primary methods for running particular descriptive statistics, **Frequencies** and **Descriptives**. Both are found under **Analyze** → **Descriptive Statistics**. While can they produce similar output, a few more details about their differences are given below.

Frequencies

Generally, a 'frequency' is used for looking at detailed information on nominal (or categorical) data and describing the results. This entails producing tables showing counts and percentages, statistics including percentile values, measures of central tendency, dispersion and distribution, and charts, including bar charts and histograms. The steps for using the frequencies procedure is to select your variables for analysis, and then choose the desired statistics, chart, and format options.

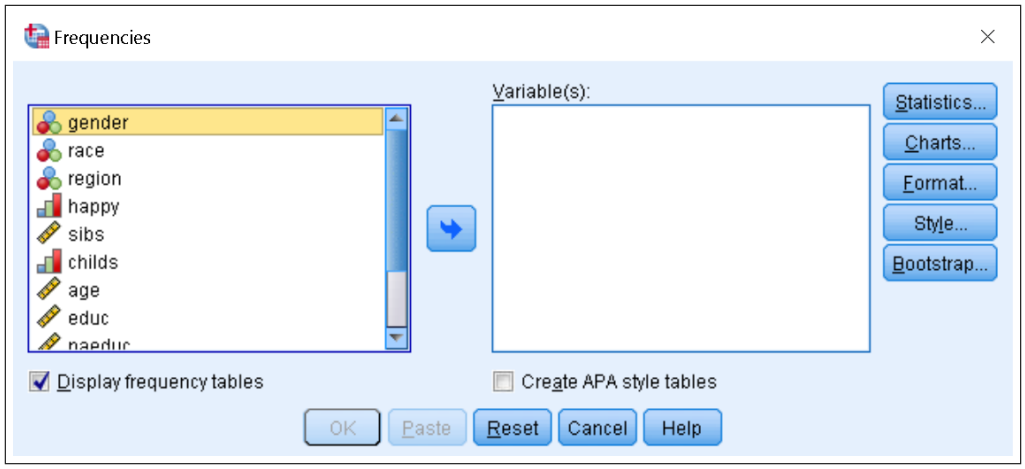


Figure 3.8: The Frequencies Dialog Box.

Some features of the **Frequencies** procedure are:

- The ability to request 'Frequency Tables', which summarize the number of responses in each category (*very* useful for categorical responses; *very* lengthly if used on a continuous variable with a wide range!).
- The ability to choose which descriptive statistics to conduct (although not as customizable as the options found using **Case Summary**).
- The ability to request bar charts, pie charts, or histograms (with or without normal curves overlaid), and in terms of either raw frequencies or percentages.

happy General Happiness

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very Happy	467	30.8	31.1	31.1
	2 Pretty Happy	872	57.5	58.0	89.0
	3 Not Too Happy	165	10.9	11.0	100.0
	Total	1504	99.1	100.0	
Missing	9 NA	13	.9		
Total		1517	100.0		

Figure 3.9: Sample Frequencies Table for Happiness Variable.

Descriptives

On the other hand, 'Descriptives' are used to obtain summary information about the distribution, variability, and central tendency of continuous variables. Possibilities for this procedure include finding the mean, sum, standard deviation, variance, range, minimum, maximum, standard error of the mean, kurtosis and skewness of each variable.

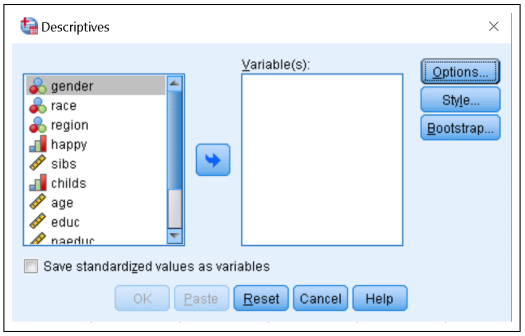


Figure 3.10: The Descriptives Dialog Box.

The only other benefit of the **Descriptives** procedure is the ability to 'Save standardized values as variables'. A standardized score has the following form:

$$X' = \frac{(X - \overline{X})}{s}$$

Where X is the raw score on a particular variable, $\overline{X}$ is that variable's mean, and s is the variable's standard deviation. More simply, the standardized score is then a score subtracted by the mean of the variable it belongs to, and then divided by the standard deviation of that variable.[1] If selected, the standardized scores will be entered as new columns back in your dataset's Data View.

[1] While sometimes regarded as unnecessary, standardizing variables has the benefit of putting different things on a similar scale – one with a mean of 0, and a standard deviation of 1 – that allows them to be directly compared and some analyses are designed to work best with standardized variables!

Exercise: Use **Frequencies** to look at the frequency tables for RACE and REGION. Present your findings graphically. Similarly, use **Descriptives** to produce summary statistics for the highest year of education variables, and the age of the respondent. Save the standardized values as variables, and re-run the **Descriptives** procedure on these new variables. Note the differences in the measures of central tendancy and variability!

3.3 Graphical Techniques

An essential element of statistical inquiry is to be able to visually inspect, interpret, and present the results of a study. This section will introduce the primary facility in SPSS for building graphics, the **Chart Builder**, as well as give an applied example on how to conduct some light post-processing/refinements using the **Chart Editor**, and demonstrate how to export your graphic in a format that can be utilized by other applications (such as LaTeX or Microsoft Word).

Remarks on further graphical techniques will arise during the notes on the statistical analyses that utilize them. For instance, please refer to the section on one-way ANOVAs for notes on producing a means plot. By the end of this section, you should be comfortable building your own plots using the templates provided by SPSS, and customizing them via the **Chart Editor's** plethora of options, such as: adding identification labels, changing the axis scaling, colours, and shading, adding titles, subtitles, footnotes, and changing the fonts, and incorporating fit lines and error bars.

3.3.1 Chart Builder

The **Chart Builder** is accessible from the menu bar under **Graphs** → **Chart Builder**. The first time you run the procedure you will most likely have a warning box appear:

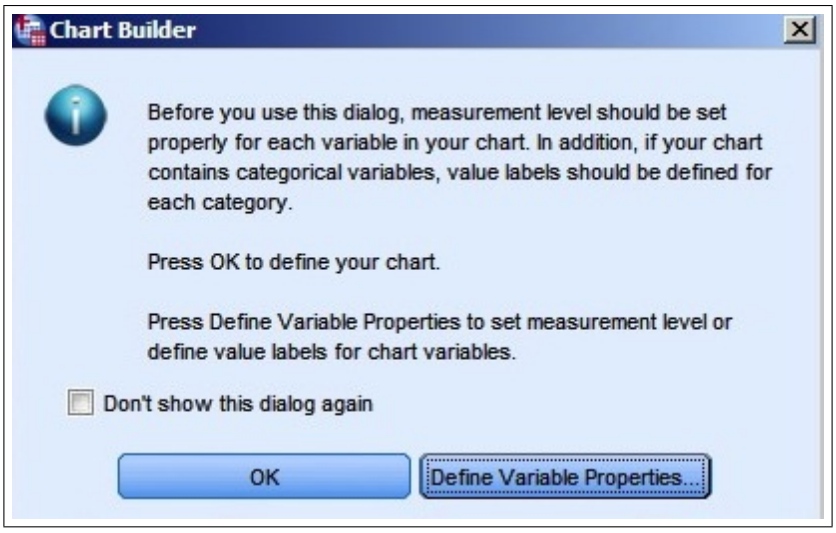


Figure 3.11: Chart Builder Warning.

The only purpose of this warning is to reiterate the importance of defining the levels of measurement for your variables (since categorical variables will be interpreted differently for graphical procedures than continuous ones). Since we are experts in defining variable properties, you can feel comfortable checking 'Don't show this dialog again' and 'OK'.

Once you have done so, the main window for the **Chart Builder** (see next page) will appear, with most of the screen devoted to a blank canvas on the right, and your variables listed on the left

(with category levels listed below the list, if applicable). This procedure uses a interactive interface, where you can browse the available types of graphics (found in the 'Gallery' tab along the bottom of the window), or you can start from scratch using tools found in the the 'Basic Elements' tab. Most often, it is easiest to select the type of graph you wish to create from the 'Gallery' by first selecting a category, and then simply left clicking and dragging the chosen chart style to the blank canvas. To find out the name of a particular type of graph, hold your cursor over the image and a ToolTip will appear letting you know what it is.

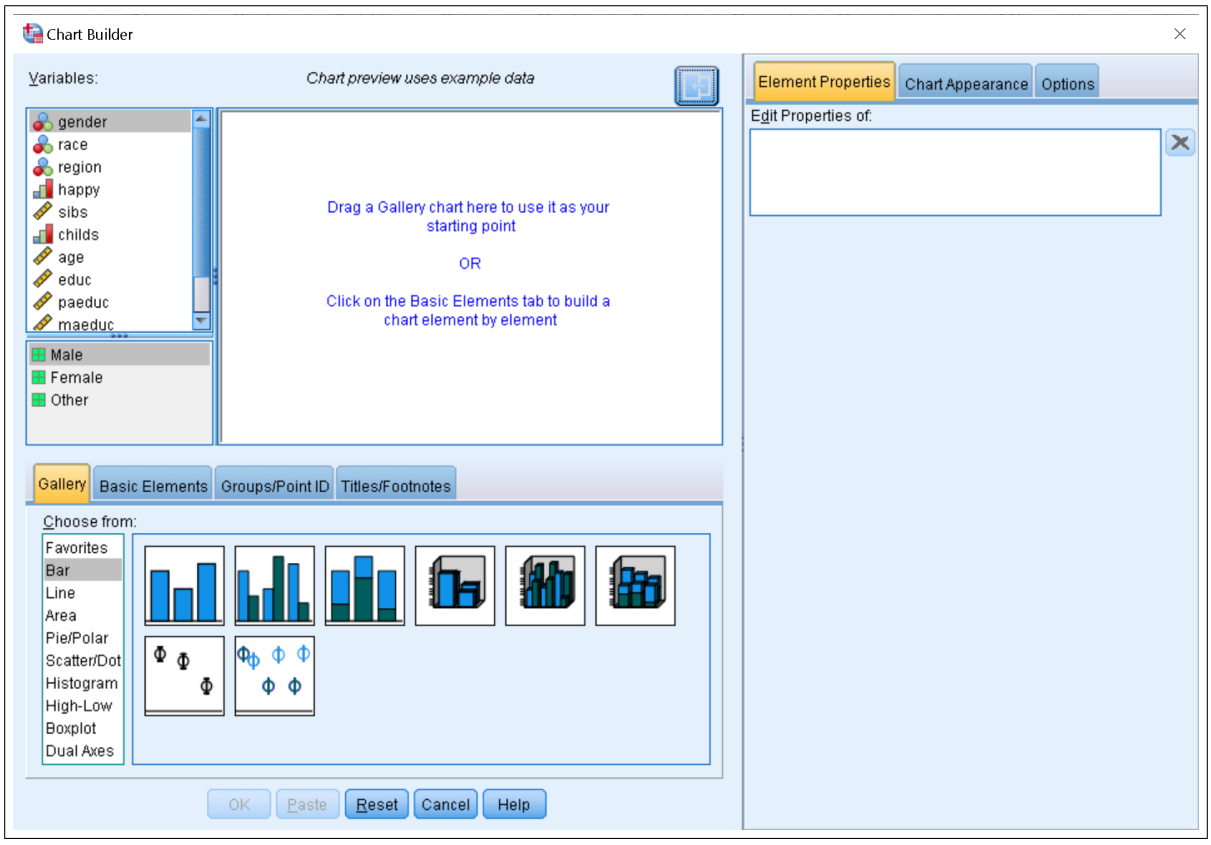


Figure 3.12: The Chart Builder Interactive Tool.

To begin, select a chart type and drag it into the 'starting point' window. The most frequently used plots are histograms, boxplots, scatterplots, and line graphs – although the appropriate plot is highly dependent on the data you are trying to interpret. A rough guide is that with one or two categorical variables, you should use a bar (or clustered bar) graph or pie chart.[2] For tests of mean differences, bar charts, line graphs, or boxplots are often useful. To show the relationship between two continuous variables, use a scatterplot. With more than two continuous variables, a scatterplot matrix is often desirable for a quick overview of each pairwise relationship.

Exercise: How might we generate a scatterplot for our dataset with PAEDUC (father's education) on the X-axis and MAEEDUC (mother's education) on the Y-axis, with the points shaded by OCCAT80 (occupational category)?

[2]But, really, you should almost never use pie charts.

This plot can be made by choosing the 'Scatter/Dot' heading from the Gallery, and then dragging the first icon (corresponding to 'Scatter Plot') up to the workspace.

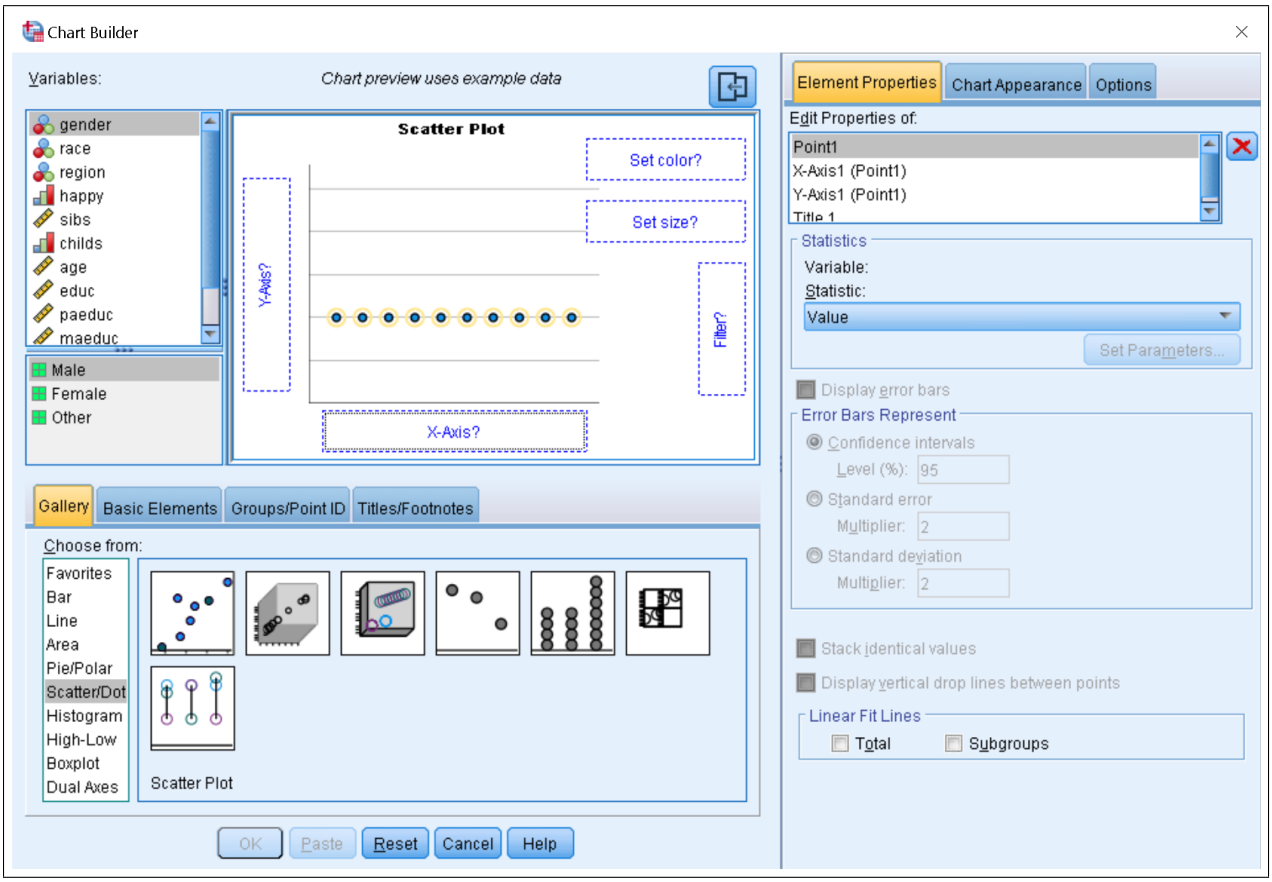
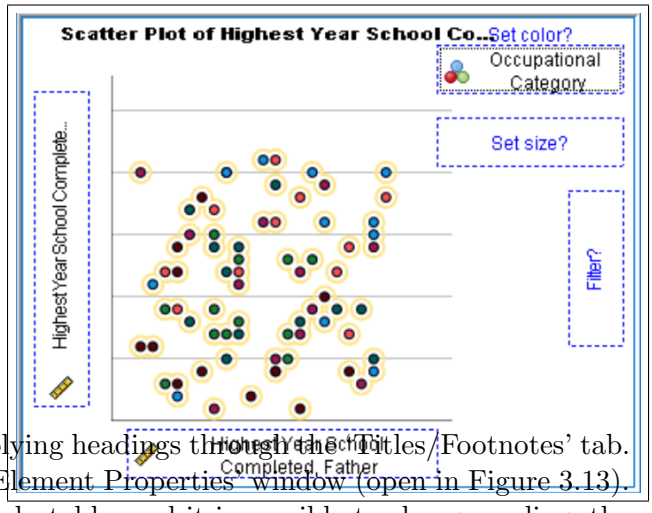


Figure 3.13: The Chart Builder: Scatterplot Template.

With our template in the builder window, you should notice three dashed outline boxes – one for the X-axis, one for the Y-axis, and one for 'Set colour'. We can simply find the variables of interest in the list and drag them onto the respective box.



Note: The chart preview does *not* live update with your data! It only uses a set of built-in 'example data' to give you an idea of what the chart will look like.

We could click 'OK' to generate the plot onto the Output Window, or we could further customize it from within the **Chart Builder** by applying headings through the 'Titles/Footnotes' tab. Even more tweaks can be made by opening the 'Element Properties' window (open in Figure 3.13). In this window, every element of the graphic is selectable, and it is possible to change scaling, the order of variables, metrics, and axis attributes.

For now, click 'OK' to view our scatterplot:

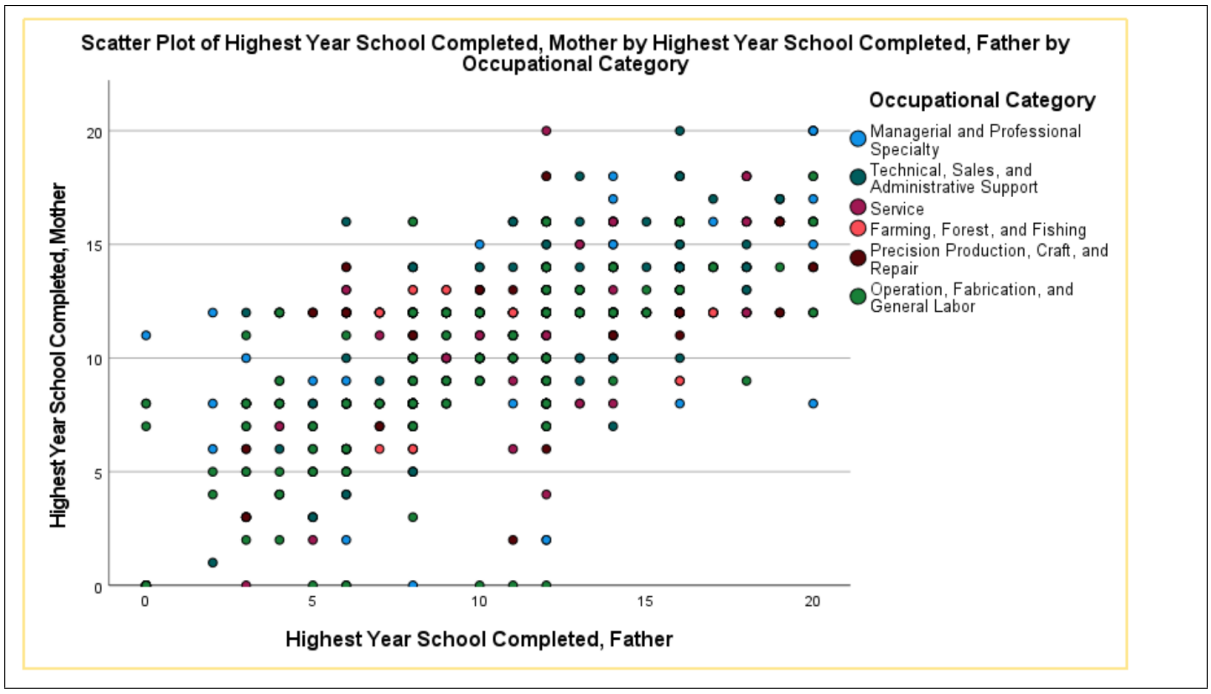


Figure 3.14: A Grouped Scatterplot.

Some interesting things become apparent with this visualization. First, there is a high variability in the number of years of education that each parent completed - both apparently from 0 to 20.[3] Secondly, it appears that most of the participants whose parents failed to complete any schooling were from the Operation, Fabrication, and General Labor occupational category, while individuals with both parents having had 20 years of schooling were typically found in the Managerial and Professional Specialty sector.

If this dataset was the focus of our research, spending some time with this graphic could reveal even more interesting observations. However, as it stands, perhaps we don't like the colours used to mark the occupational categories, or the fact that the points that indicate mothers with zero years of education lie along and are partially obscured by the X-axis/table border. How might we fix these annoyances?

3.3.2 Customizing Graphics with the Chart Editor

Any graphic that appears in the Output Window of SPSS can be further customized by simply double clicking on it. This opens the **Chart Editor** window.

[3]This could either be a realistic maximum, or a consequence of the scale used in the survey, where it is possible 20 was the maximum available selection.

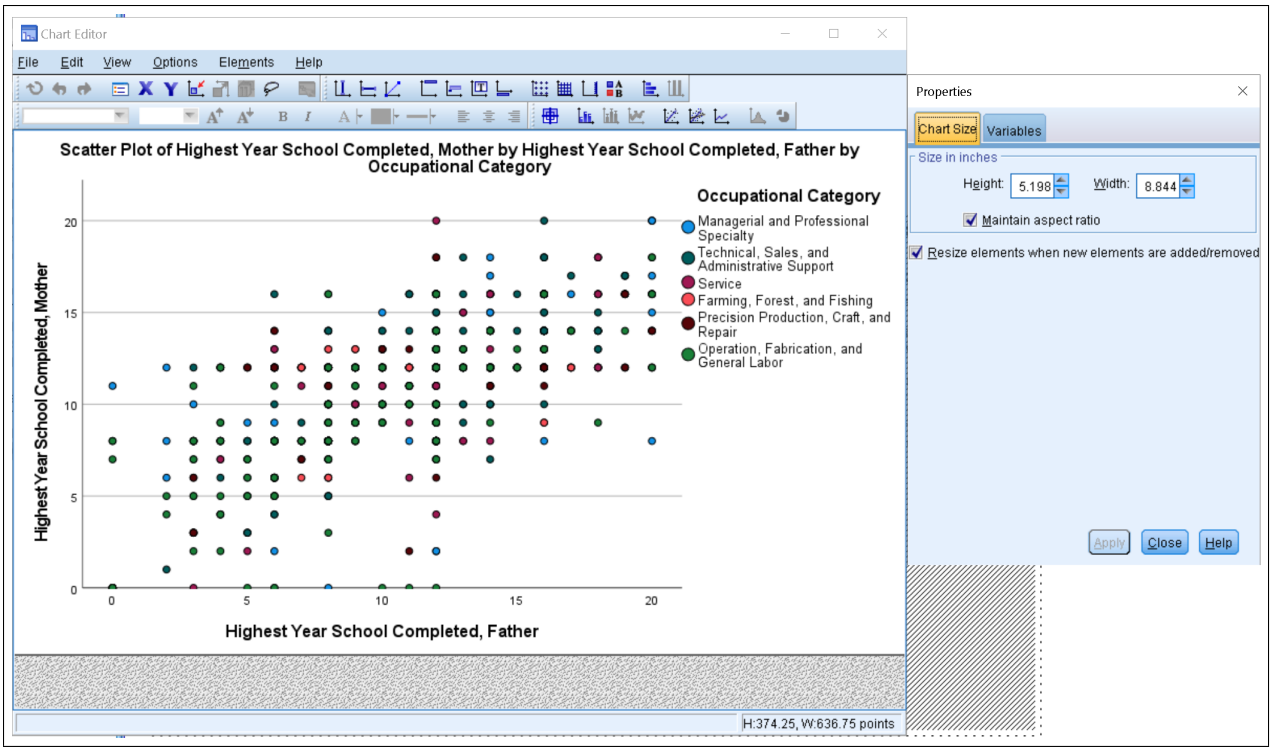


Figure 3.15: The Chart Editor Window.

This editor has numerous features, with its default toolbar presenting three full lines of buttons and toggles, as well as a context sensitive 'Properties' window![4] Again, you can hover your cursor over any particular button of interest to get a ToolTip about what it does.

To get started, try changing the scaling of the Y-axis. To do so, click the button that looks like a **Y**. This selects the Y-axis, and you should notice that the properties window has updated with a new series of tabs. All of these pertain to customizing the Y-axis:

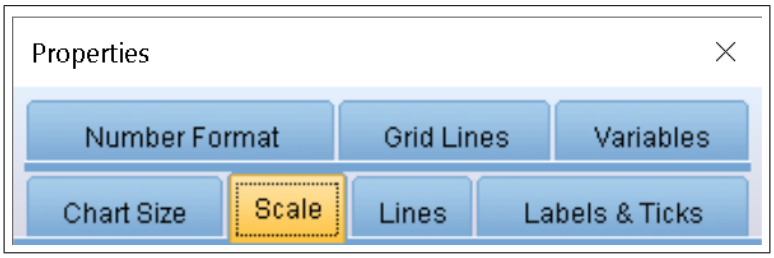


Figure 3.16: The 'Properties' Window: Customizing the Y-axis.

To change the numbering on the axis, we want to look for the 'Scale' tab.

There are a few ways of changing the axis to make the values that lie along the X-axis more visible.

[4]If this window does not appear or if you close it by mistake, it can be reopened using **Edit** → **Properties Window** from the **Chart Editor** menubar.

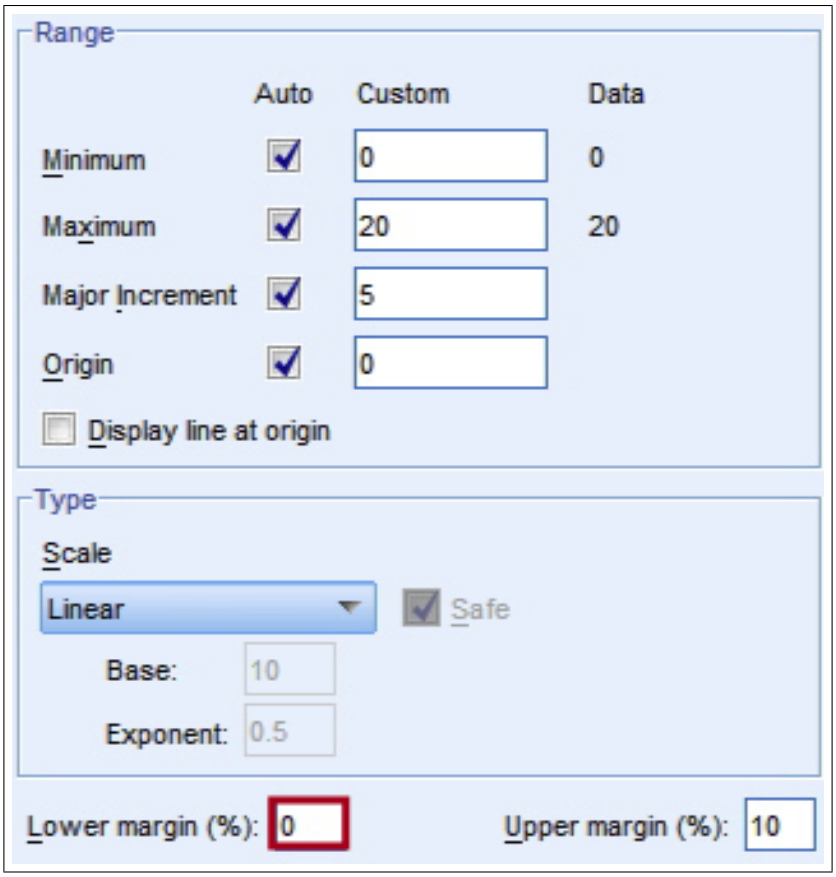


Figure 3.17: Customizing the Y-axis Scale.

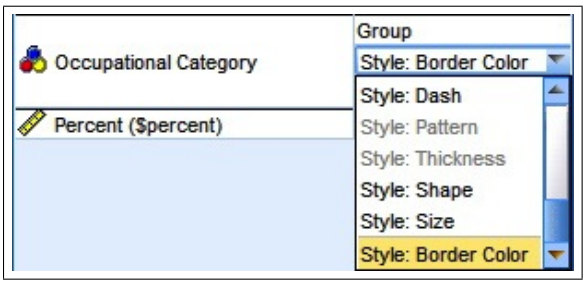
- We could change the 'minimum' range value to a negative integer (e.g. replace the 0 with a -2). However, this causes some problems as the X-axis will now begin at -2 instead of 0 – a number that is not logically possible for this variable.
- A better solution is to notice the two input boxes at the bottom of the 'Properties Window'. These pertain to the 'Lower' and 'Upper Margin'. To get the points off the X-axis, try changing the 'Lower Margin' to 10%.
- Note: If our minimum value was above the lowest *possible* minimum (for example, if we were looking at income and our poorest participant only made $10,000 a year when, obviously, someone could conceivably make less), and this was the value SPSS chose to use for the X-axis, then the first approach (changing the range of the axis to a smaller, but still realistic, value) would be entirely appropriate.

Next, perhaps we want to edit the points themselves in our graphic. To do this, we can double click on any point[5], and you should notice that all of the points *from that occupational sector* will become highlighted. The 'Properties' window will now update to refer to the marker/general chart properties. In the 'Marker' tab, we can change the colour for that category as we see fit.

But, perhaps, we are submitting this graphic to a journal that only prints in black and white!

Is everything ruined forever?!

[5]It is good practice to select points from the LEGEND as opposed to the data plot itself. This is because the occupational categories overlap on many of the points, so selecting a particular dot may select more than one category.

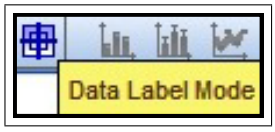


Nope! In this case, we would want to go to the 'Variables' tab in the general 'Properties' window. This pane lists all of the available variables in the dataset, and can be used to add/remove/change the variables displayed in our plot.

We want to find our Occupational Category variable, and change its style from its current setting ('Style: Border Colour').

Feel free to experiment with choosing from the other selections (e.g. 'Shape' and 'Size'), and clicking 'Apply' to see how they affect the plot.

While there are many other ways we can tweak this plot, one last thing I would like to mention here is the labelling of points. For instance, maybe we want to know which cases appear at the minimum (0,0) and maximum (20,20) parental education coordinates.



To do this, we can activate the 'Data Label mode' on the menubar. Using this useful tool, we can click on a particular case on the plot and a window will appear telling us the cases that exist at the location. If there is only one data point at that location, it will be labeled instantly. If more than one data point exists, another window will appear with a list of all the points at that location. You can then click 'Select All' and 'OK' to have SPSS label that point with all of the cases that belong to it.

Here is an example with the 'Shape' style selected for Occupational Category, the maximum and minimum points labeled, as well as Case 854 (the only point with that particular combination of parental education values).

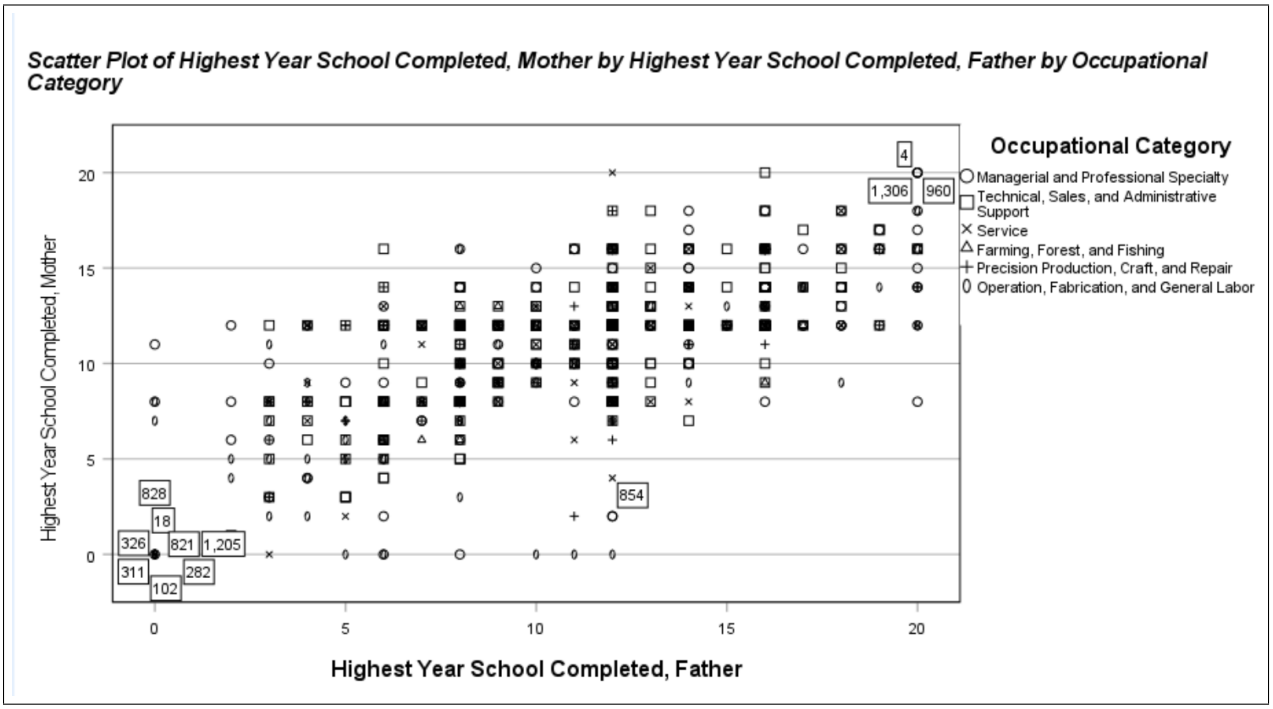


Figure 3.18: Data Labeling in the Chart Editor.

3.3.3 Exporting Graphics

Now that we have a plot that we might want to keep or distribute, it is important to know how to save it. There are a few options:

1. You can simply save **the entire SPSS output file**. This will keep all of the output produced during your SPSS session, and put it in a 'Viewer' file with the extension *.spv. Since the entire file is saved, every object within remains editable from within SPSS.
2. You can save an **image file**. If you no longer want to edit your graphic, you can use **File** → **Export** on the menu bar to save the graphics in your output as image files. In the Export window, scroll to the bottom of the 'Document' → 'Type' list to find **None (Graphics Only)**. Then select the desired filetype from the 'Graphics' drop down menu (e.g. as *.jpgs).
3. Similarly, if you have another application open, you can often select a chart and right click on it to copy it. Then you can simply paste it into your other document but this is not a recommended approach!

Exercise: Try to save your plot as `MYCHART.JPG`. Make sure to note where SPSS is going to place the saved file!

3.4 Statistical Procedures for Categorical Data

The remainder of the course will focus on importing various data types, and running appropriate statistical analyses based upon the nature of the available data. As stated in the course objectives, this is not meant to be an introductory statistics course, so we will not be delving into the mathematics of these analyses, but rather focusing on how to use them in SPSS.

The first set of data we are going to use is `CAT1.DTA`. The *.DTA filetype is the default extension used by STATA, another statistical software package. Try opening this file with SPSS and save it as `MYCAT1.SAV`.

This dataset has two variables: PATIENCE and GENDER. Perhaps we are interested in the research question: is there a relationship between patience, as measured by some validated metric, and the gender of the participant?

3.4.1 Using Frequencies for Categorical Predictors

The first approach to investigating this hypothesis would be to produce some visualizations. Since both of these variables are categorical (PATIENCE only has three levels; GENDER only has two), we can use the **Frequencies** procedure to investigate this relationship.

1. Bring both PATIENCE and GENDER into the 'Variables' box.
2. Open 'Charts' and choose either 'Bar charts' or 'Pie charts' to visualize categorical variables.[6]

3.4.2 Using Crosstabs

The graphics produced above are essentially univariate displays. They will show the proportion of males and females in the dataset, and the proportion of people with low, medium, and high levels of patience, but not how the two variables interact. To investigate the relationship between gender

[6] There are more advanced techniques for visualizing categorical data. Dr. Michael Friendly has written extensively on the topic and his books are highly recommended!

and patience, we will use **Analyze** → **Descriptive Statistics** → **Crosstabs** to run a chi-square analysis (χ^2), which looks cell-wise at the observed number of cases in each as compared against the expected number of observations if the variables were entirely independent.

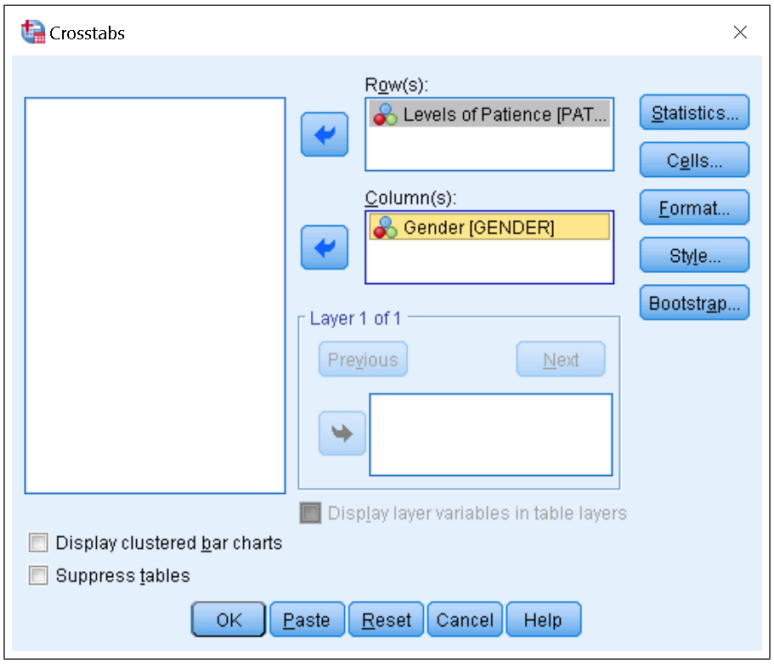


Figure 3.19: Categorical Variable Analysis via Crosstabs.

- It doesn't matter which variable you select for 'Row(s)' and which you use for 'Column(s)'. However, an aesthetic rule of thumb is to put the variable with more levels as 'Row(s)', since a printed page has more vertical space available than horizontal.
- 'Display clustered bar charts' produces a useful graphic with the 'Row(s)' variable along the X-axis, while the count of the other variable defines the bar height.

Some useful options: In the 'Statistics' dialog box, request **Chi-Square** to test the independence between the variables and **Phi** (ϕ) and **Cramer's V** as effect sizes. Under 'Cells', check **Expected Values** and **Standardized Residuals** to help understand the relationship between the two categorical variables.

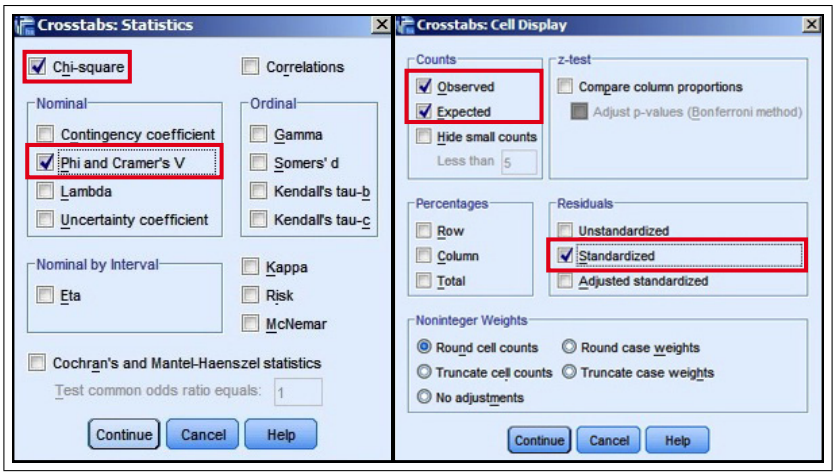


Figure 3.20: Crosstabs Options for a Chi-Square Analysis.

When you click 'OK', the SPSS output window should appear with the requested analyses. Does gender seem to have a relationship with patience?

3.4.3 Alternative Approach: Using Weights to Analyze Aggregated Count Data

Sometimes categorical data is given in an aggregated format (e.g. when the original cases or observations are unavailable). In this scenario, you can use a "weight" variable to account for the number of observations for each factor combination. For instance, open `TVAGG.SAV`, which pertains to (fictitious) data on how many individuals watched an entire commercial (narrated either by a male, female, or no narrator) during a particular type of broadcast (a science fiction show or a sports event). The dataset has three variables: PROGRAM, GENDER (of narrator), and COUNT.

	PROGRAM	GENDER	COUNT
1	1.00	1.00	2.00
2	1.00	2.00	10.00
3	1.00	3.00	3.00
4	2.00	1.00	5.00
5	2.00	2.00	7.00
6	2.00	3.00	3.00

To test to see if there is a relationship between PROGRAM and GENDER, we should use a Chi-Square (χ^2) test of independence. This is found in SPSS via **Analyze** $\rightarrow$ **Descriptive Statistics** $\rightarrow$ **Crosstabs**. Use PROGRAMS and GENDER as the rows and columns, and check Chi-Square off in the 'Statistics' options window. See Figure 3.21 for an overview of these steps.

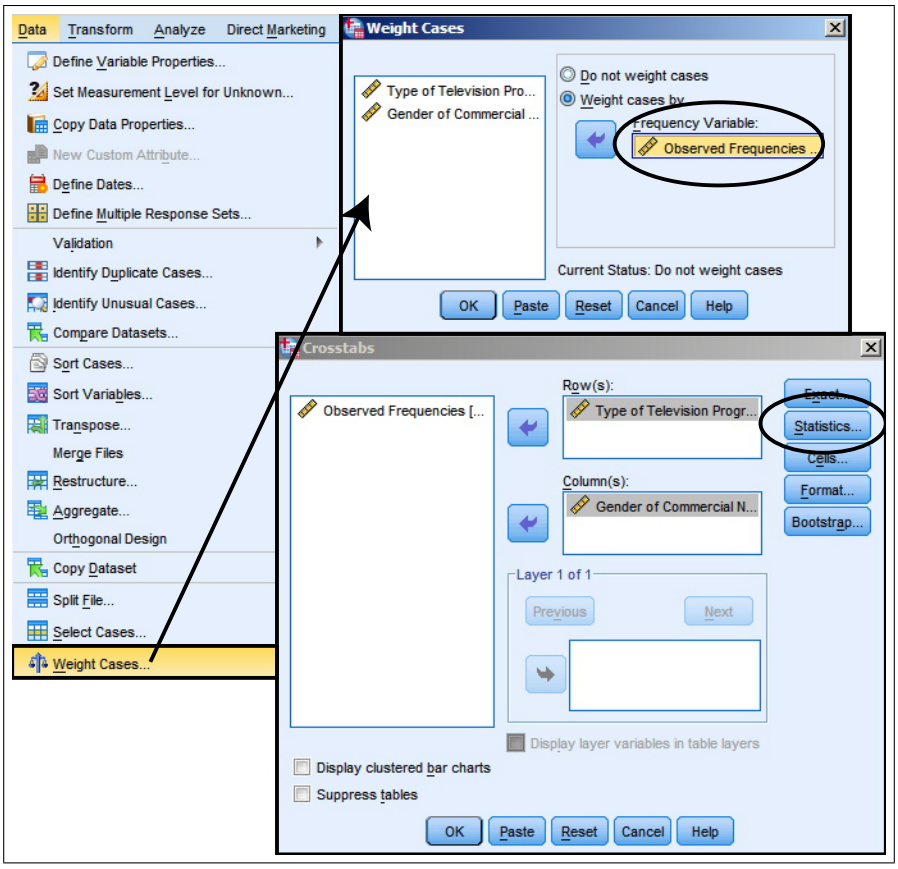


Figure 3.21: Using Aggregated Data to Conduct a χ^2 Statistic.

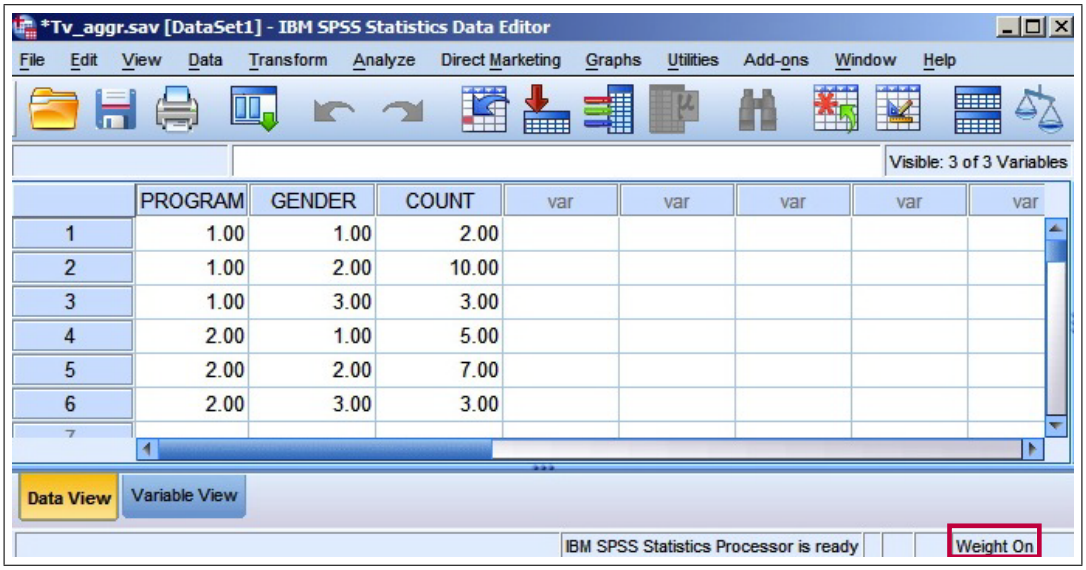


Figure 3.22: Weights Enabled.

Exercise: Try replicating this analysis, and interpret the results.

3.5 Categorical by Continuous Data

Most datasets are not purely categorical. We often have a measure that is on a continuous (or pseudo-continuous) scale. For example, intelligence, and age have a wide range of possible values. The choice of statistical procedure for these studies depends on the variables available, and the nature of their relationship.

> Open `CATCON1.SAV`. This file has two variables ANIMAL (categorical, with 2 levels: person identifies more as a "dog person" or as a "cat person"), and a measure of RESILIENCE, which is continuous.

3.5.1 Investigating two-group data using Explore

To get a sense of the variables we are working with, again, it is a good idea to build up from numerical and visual summaries. The **Analyze** → **Descriptive Statistics** → **Explore** procedure allows us to do just that.

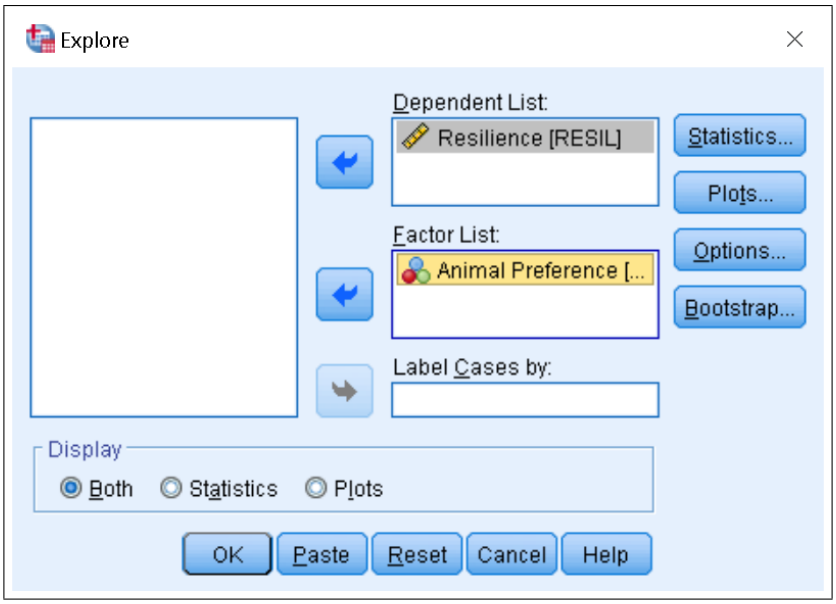


Figure 3.23: The Explore Procedure for Grouped Data.

Now that we have a grouping factor (ANIMAL), we can use it to investigate the descriptives of a continuous variable at the levels of the categorical one. We do this by including ANIMAL in the 'Factor List' box, while putting RESIL in the 'Dependent List'.

Some useful options: In the 'Statistics' dialog box, request **Descriptives**. If you wish to test for the presence of **outliers**, also check that box. Under 'Plots', choose **Histogram** and **Normality plots with Tests**. You can also safely uncheck the **Stem-and-Leaf** diagrams option.

The Output: When you click 'OK' to obtain the output from this procedure, you will see four tables: a case processing summary (with information on missing values, if any); descriptives (e.g. measures of central tendency and dispersion, separated by your categorical variable); extreme values (if you checked the 'Outliers' option in the 'Statistics' window); and, a 'Tests of Normality' table, also with entries for each level of the categorical variable. If either of these tests are significant,

some authors recommend you look at your distributions and transform them as necessary.

After these tables, you will have a series of plots: a histogram for each gender; a series of Q-Q plots (which are supposed to help you diagnose non-normality, but are sometimes hard to interpret); and, finally, a boxplot with the categorical variable on the X-axis. Boxplots are fantastic tools for visualizing grouped data, as they show outliers and extreme values (with asterisks), the smallest and largest observations that aren't outliers, the 25^{th}, 50^{th} (the median), and 75^{th} percentiles, and allow comparisons to be quickly made between groups.

In this example, the boxplot looks like:

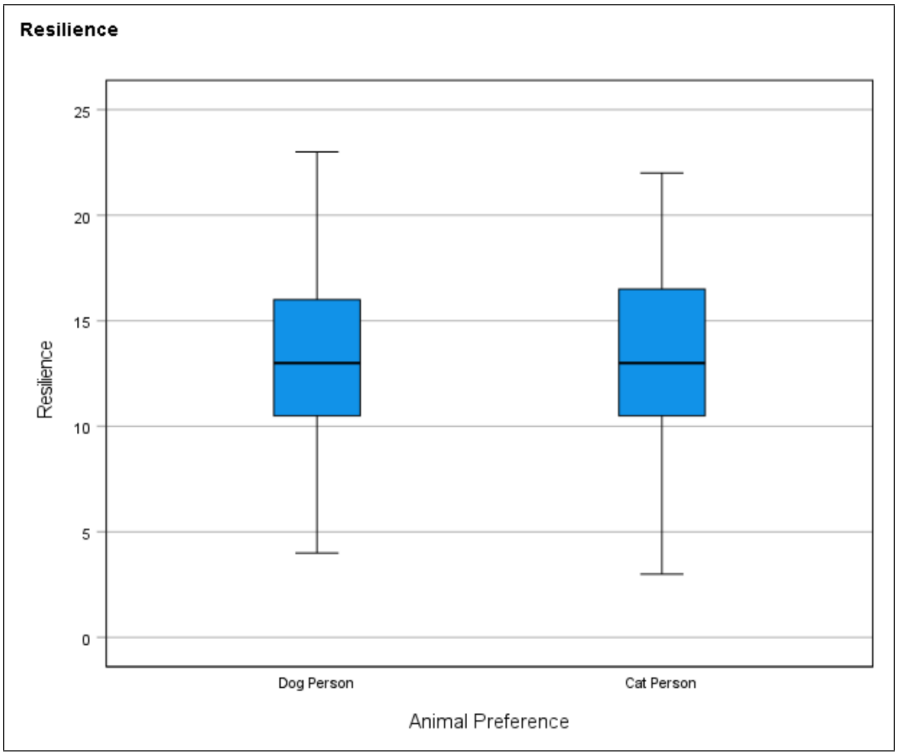


Figure 3.24: A Boxplot for a Study with a 2-Group Design.

We can quickly see that the two groups appear relatively equal in terms of resilience - the medians are pretty much identical, they have similar amounts of variability (ranging from around 4 to 24), and neither group has any substantial outliers.

3.5.2 The Independent Samples *t*-test via Compare Means

However, we spent all of this time collecting our data – and so we would still like to statistically test for a difference between the means of resilience between these so-called cat people and dog people. To do so, we will use the **Analyze** → **Compare Means** → **Independent Samples *t*-test** procedure.

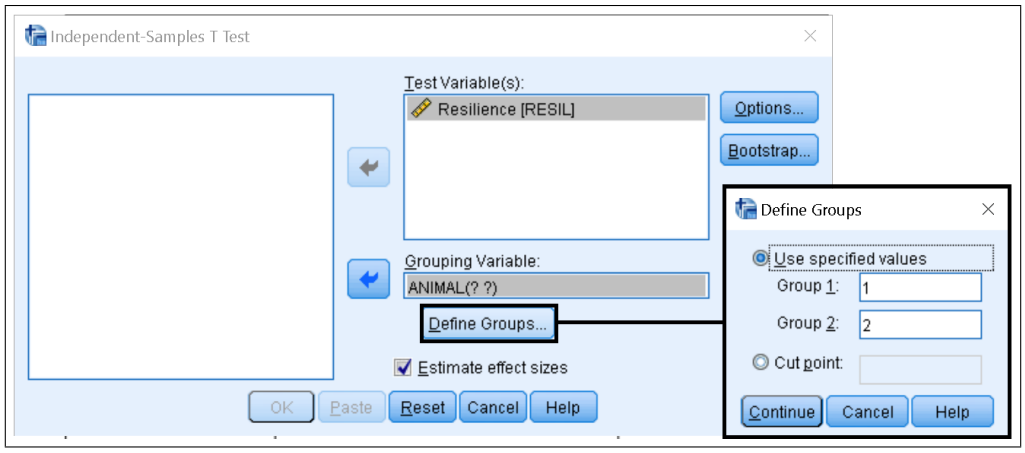


Figure 3.25: The Independent Samples t-test.

The t-test is only used for two-group factor variables. However, you can select a variable with more than two groups in the 'Grouping Variable' dialog box. That is why it is important to always click on 'Define Groups...' and make sure to enter the PROPER values for the two groups you wish to compare. In our example, this would be 0 (for dog people), and 1 (for cat people).

The Output: Compared to the other procedures we have run so far, the output from an Independent Samples t-Test is downright minimal. It yields two tables: one on 'Group Statistics' (which gives the sample size, mean, standard deviation, and standard error of the mean for both groups), and the 'Independent Samples t-Test' table, which reports the Levene's Test for Equality of Variances (which is used to check the homogeneity of variance assumption), as well as the t-Test test statistics, their significance, and their confidence intervals.

> In this example, does resilience significantly differ due to animal preference?

3.5.3 The Dependent or Repeated Measures t-test

Another form of the two-group design is the Repeated Measures t-test. In this set-up, each participant is given the test stimuli twice. This is often thought in terms of a pre-/post-design, where a participant comes into the laboratory, takes a pre-test, then is exposed to some sort of intervention, and takes a post-test afterward. The goal is to see the influence of the intervention.

> Open `PREPOST.SAV`. This file has two variables ANXPRE and ANXPOST, both of which pertain to anxiety scores – one before our intervention, the other afterward.

Note: In a repeated measures t-test datafile, there is no grouping variable! This is because every individual participates in both conditions.

The Repeated Measures t-test works by looking at the difference between the pre- and post-scores for each individual (e.g. do the anxiety scores typically get larger, indicating that our intervention was a catastrophic failure? Or do they diminish after our intervention?). These differences are averaged, and divided by a measure of variability that expresses how much we would expect these differences to vary from zero if the null hypothesis (that our intervention had no effect) was true.

To run this variant of the t-Test, go **Analyze** → **Compare Means** → **Paired-Samples *t*-test**.

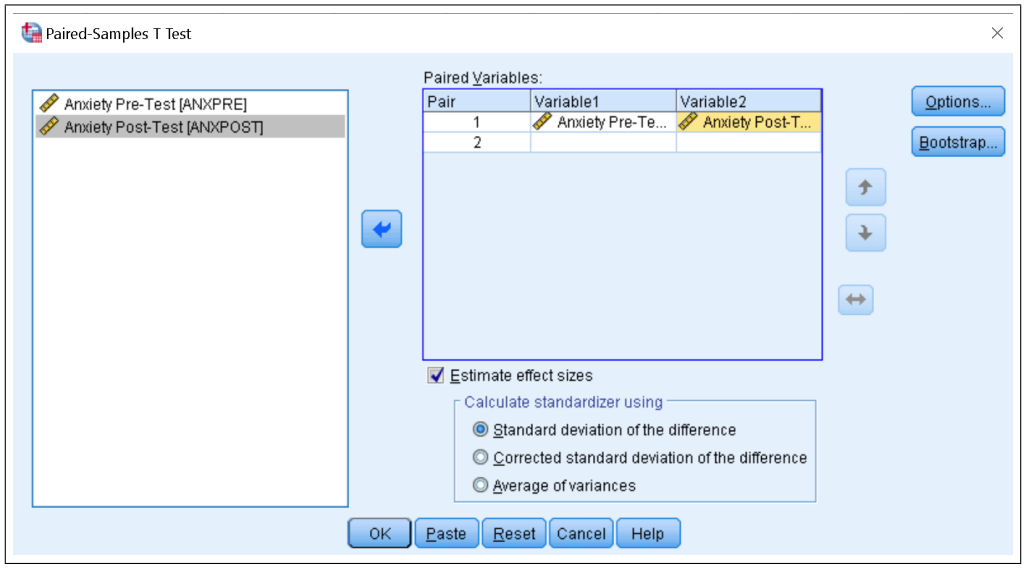


Figure 3.26: The Repeated Measures t-Test Dialog Box.

This dialog box may look slightly different, due to the fact that you are asked to provide a pair of variables instead of an arbitrary number. However, since this procedure is about looking at the difference between a pair of variables, it makes sense that they would lay it out this way. Simply drag or double-click the Pre-Test score into the Variable1 box, and the Post-Test score into the Variable2 box (or vice versa) to set-up the procedure.

The Output: Similar to the output from the Independent Samples t-Test, the Repeated Measures t-test is also fairly straightforward. It yields three tables: one of the paired samples statistics (the mean, sample size, standard deviation, and standard error of the mean for the pre-test scores and for the post-test scores); one of the correlation between the pre-test and post-test scores;[7] and, one for the paired samples test statistic.

> Was there a significant difference between our pre-test and our post-test scores? In which direction was the effect?

3.5.4 One-Way ANOVA

The t-test is appropriate if we only have two groups that we are comparing on some measure. If we have more than two groups, we instead turn to a one-way analysis of variance (ANOVA). This procedure compares two proportions of variability – that found between the group means (often referred to as the $MS_{between}$, and represents differences between groups) to that found, on average, within groups (MS_{within}, or inter-group noise). If the variance between groups is substantially larger than the variance within groups, we reject the null hypothesis that there were no differences between them.

[7] If the correlation was equal to zero, then the repeated measures t-test would be equivalent to the independent measures t-test.

Open `ONEWAY.SAV`. Like the Independent Samples t-test data, this file contains only two variables: HAPPINESS, the dependent variable that is a continuous measure that ranges from 3 to 24, and a categorical variable, FACULTY, which has three levels.

Note: The *only* difference between a 3 level one-way ANOVA and one with 4 or more levels is the number of categories the categorical variable can take on.

To run a one-way ANOVA, run **Analyze** → **Compare Means** → **One-Way ANOVA**.

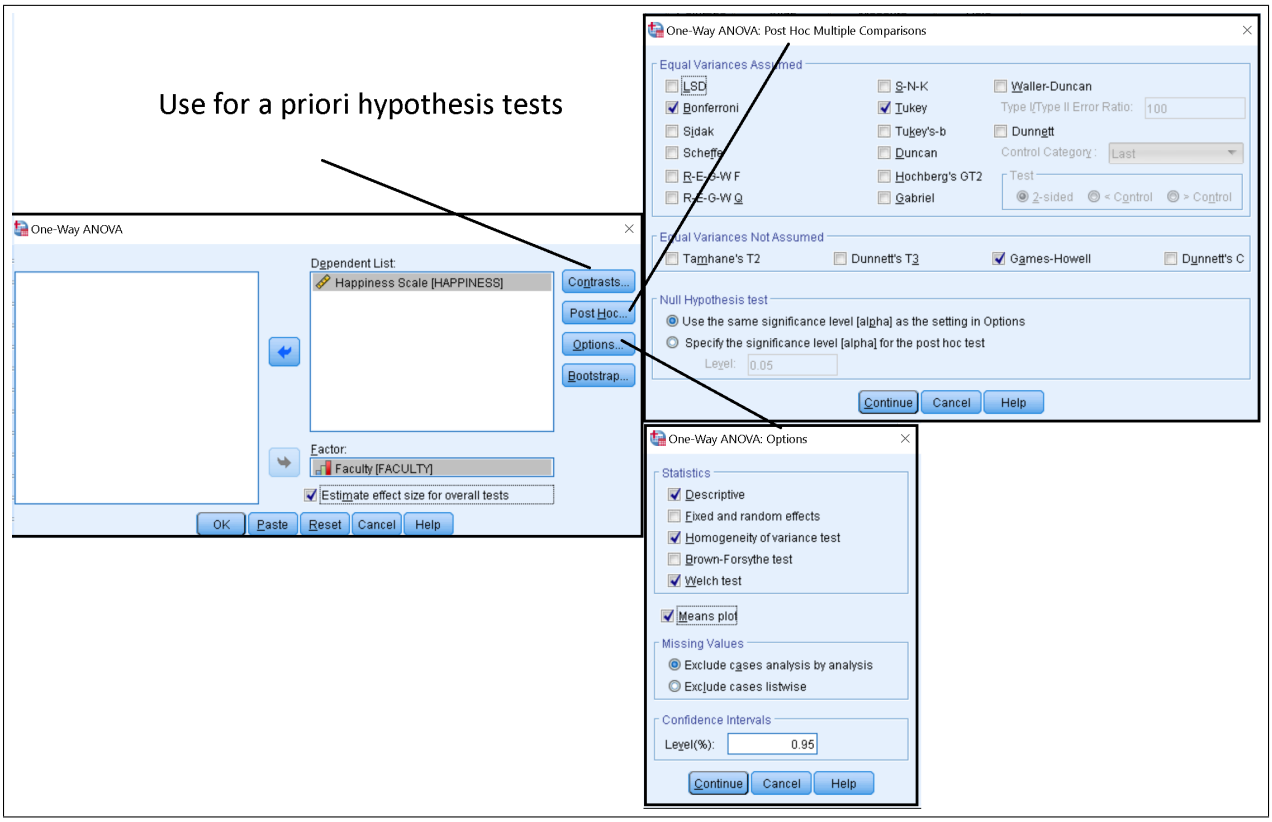


Figure 3.27: One-Way ANOVA: Post-Hocs and Options.

Some commonly selected options are illustrated in Figure 3.27. Note that, unlike for the t-test, descriptives and homogeneity of variances tests are not produced by default. You have to request them under the 'Options' menu. You can also find the Welch procedure (for conducting a 'robust' version of the one-way ANOVA), and the means plot function here.

The ANOVA is an omnibus test, meaning that if you obtain a significant result, you know that there is a significant difference between at least two of the groups, but you do not know where that difference might be (this is especially tricky with a large number of groups). Post-hoc, or multiple comparison, procedures are somewhat equivalent to running multiple t-tests among the different group combinations looking for the significant difference, while attempting to control the family-wise Type I error rate. The selection of a specific post-hoc test depends on many factors, ranging from the discipline you are publishing in, to personal preference, but a few of the more typical ones are highlighted here. Note that most of the procedures assume relatively equal variances within each group, while the four tests along the bottom do not.

The Output: The output for the one-way ANOVA, especially with the recommendations above, is lengthier than any other procedure so far. Like in the previous analyses, a 'Descriptives' table summarizes the basic characteristics of the dependent variable for each level of the categorical variable. The Levene's Statistic tells you about the homogeneity of variance of the scores, with insignificant values indicating a lack of violation.

Next is the ANOVA summary table, which yields the overall omnibus results. Remember, a significant result here only means there is a difference somewhere among the groups. It is followed by the 'Robust Test of Equality of Means', since we asked for the Welch correction.

The final two tables are used for following up a significant omnibus ANOVA. All of the request post-hoc tests are put into a 'Multiple Comparisons' table for comparison, which is followed by a homogeneous subsets summary - which is another method for showing which groups are different from the others.

Finally, SPSS provides a means plot, which is a quick and somewhat inelegant graphic that shows how the means of each of the categorical variable groups differ.

WARNING: As you many notice, the omnibus ANOVA in this example was *not* significant. And yet, if you look at the means plot, it looks like there were substantial differences between the physical sciences and the business students! This is illusory and is entirely caused by SPSS choosing an inappropriate minimum and maximum value for the Y-axis scale. Try changing the axis scaling to begin at 0 instead of 13.0!

3.5.5 Univariate General Linear Models

The ANOVA procedure extends into many other models. If we begin incorporating other factors – e.g. FACULTY, as before, as well as GENDER – this is called a **Factorial ANOVA**. If it has two categorical variables, it is a two-way ANOVA. If it has three, it is a three-way, and so forth. Factorial ANOVAs are typically more interesting that one-way ANOVAs because you can not only have a main effect for either variable on the outcome, but also study the interaction between the two variables. This allows us to ask questions like: is the relationship between FACULTY and HAPPINESS different for males and females?

Another extension is ANCOVA, or the Analysis of Covariance. This procedure is used when we have a typical ANOVA framework, but want to control for a continuous 'covariate'. For example, perhaps we want to remove the influence of age. We can include age in our model as a covariate, and its effects will be averaged out of the analysis.

A third approach is called a 'Random Effects' model ANOVA. This model is used if the treatment levels used in the categorical variable cannot be considered fixed. This occurs when the various factor levels have been sampled from a larger population, and we want to infer back to the population at large – not just at the levels we happened to sample. For example, we might randomly select three drug dosage levels to administer in our study (5 mL, 15 mL, and 30 mL). We are less interested in those exact levels, as we are about the overall effect of the drug.

These and other more complicated methods are all available in the **Analyze** → **General Linear Model** → **Univariate** area.

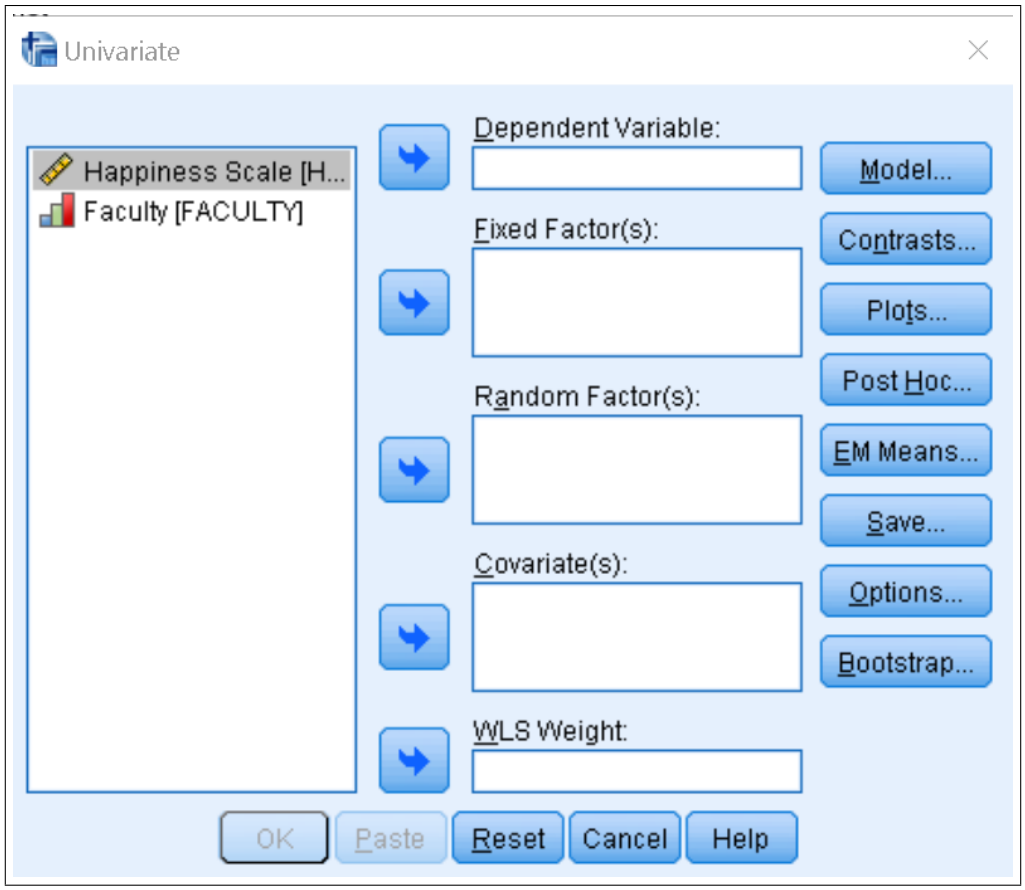


Figure 3.28: The General Linear Model Dialog Box.

As you can see, this dialog box is substantially more configurable than the previous ones. In addition to the less complex ANOVA procedures, in the Univariate GLM options you can specify the entire model from scratch (using 'Model'), incorporate multiple fixed and random factors, covariates, and weights into the model, ask for the exact contrasts you want tests ('Contrasts'), request customized plots ('Plots'), and save predicted values, residuals, and case level diagnostics information. All in all, while more daunting to get started with, this procedure is substantially more useful and customizable than the One-Way ANOVA procedure.

3.6 Continuous Predictors

The statistical approaches taken in Section 3 so far focused on looking for differences between groups. Did our treatment group outperform our control? Did the new teaching method work better than the old? And so on. However, there are times when we are not looking for how things *differ*, but how they are *related*.

This especially common when using non-experimental data. Is the number of years of education some has related with happiness? Are anxiety scores stable across time? Does the temperature predict ice cream sales? Is temperature a better predictor of ice cream sales than geographic location? Questions like these are addressed using correlation and regression-based techniques.

Load `IQSURV.SAV`. This dataset has five variables: ID, which is a simple numeric identification for participants, and five continuous variables: ESTEEM (self-esteem), PIQ (performance IQ), VIQ (verbal IQ), and TIQ (total IQ).

In this study, we are primarily interested in total IQ. We want to know which of the other predictors – ESTEEM, PIQ, and/or VIQ – are related to or predict TIQ.

3.6.1 Using Frequencies with Continuous Predictors

Since each of the variables in this dataset are continuous, we can use **Analyze** → **Descriptive Statistics** → **Frequencies** to obtain summary statistics on each of the variables in a condensed format (a single table). This is always a good first step, since it allows us to diagnose data entry errors before we begin modeling in earnest.

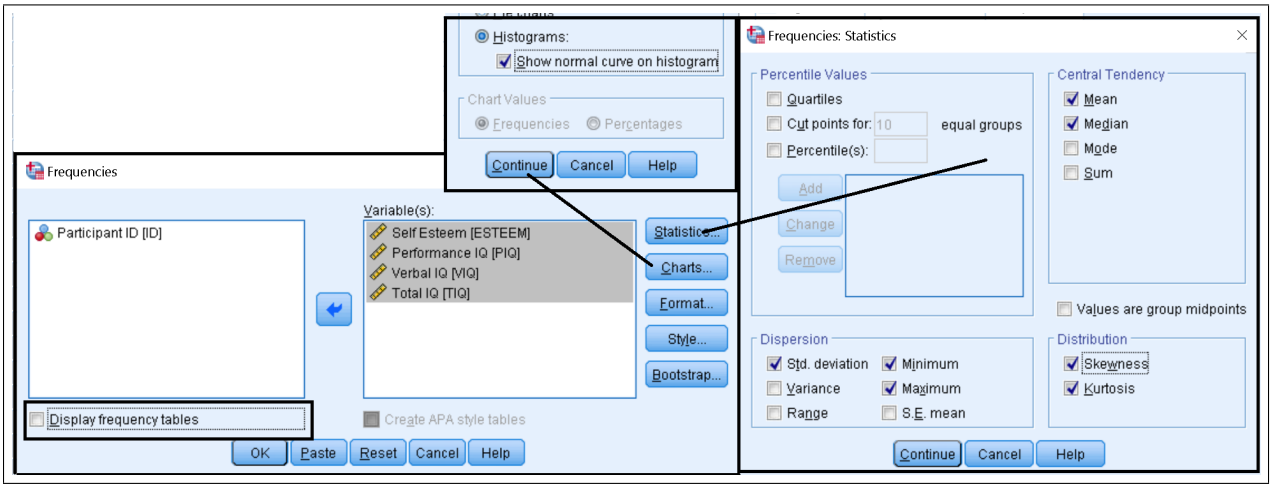


Figure 3.29: Using Frequencies for Summary Statistics.

It is often a good idea to turn off 'Display Frequency Tables', since it produces a *lot* of output that isn't really useful.

Exercise: Look at the summary statistics produced by **Frequencies**. Are there any values that look out of place? Are each of the distributions relatively normal?

Note: If any of the distributions were substantially skewed, it might be a good idea to transform the variable (see Section 2.3.1), and rerun **Frequencies** to see if the transformation helped.

3.6.2 Correlation Coefficients

Correlation is a statistical technique used to estimate and describe the relationship between two variables. It is expressed in terms of direction (positive/negative), and in strength (between 0 and 1). For example, it be expected that sugar or caffeine intake and activity level might be moderately positively correlated, or that coffee sales and temperature to be negatively related. The primary idea here is of covariance, or how two variables vary *together*. As the coefficient approaches ± 1, the relationship between the two variables becomes stronger and stronger. It is also possible that the variables are not related at all, which produces a coefficient around 0.

Bivariate Correlation Coefficients

The most common method for investigating the relationship between two variables is to look at their *correlation coefficient*. There are two common metrics for this measure: *Pearon's* r, which is the standard method for correlating continuous variables; and *Spearman's* r_S, which runs the correlation on a ranked version of the data. Spearman's r_S is typically used with small n, non-normal data, or when there are outliers present that will bias the typical correlation coefficient.

Both of these procedures[8] are available under **Analyze** → **Correlate** → **Bivariate**. Note that, while these procedures are all bivariate (two variable exclusive), you *can* select more than two variables in the **Bivariate Correlation** dialog box. This simply means that you will get a *correlation matrix*, with all pairwise coefficients.

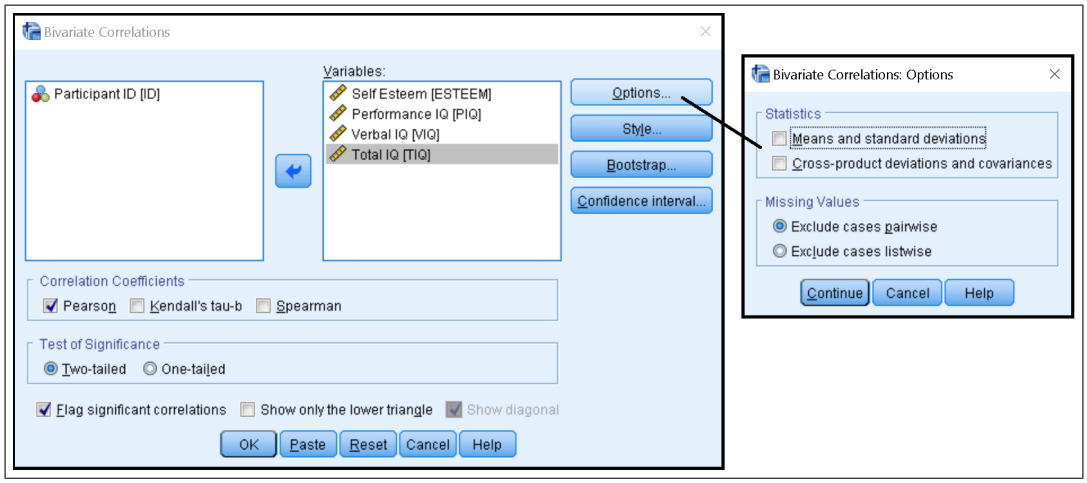


Figure 3.30: Bivariate Correlations Dialog Box.

Under 'Options' you can also request the means and standard deviations of the variables, and their cross-product deviations and covariances.

The Output: When you click 'OK' with the options listed as above, you will only obtain one table: the correlation matrix. This table is necessarily symmetric (since, for instance, the correlation of ESTEEM with PIQ is the same as PIQ with ESTEEM).

> Which variables are most and least strongly related? Are any of the coefficients significant?

Partial Correlation Coefficients

A partial correlation coefficient is the test for correlation between two variables while controlling for others. While this is only available for Pearson's r (and so the two variables of interest should be continuous), the 'Controlling for' variables can be either continuous or categorical. This procedure is found under **Analyze** → **Correlate** → **Partial**. You can request 'Zero-order Correlations' be displayed as well for comparison under 'Options'.

[8] As well as Kendall's τ, which is a non-parametric correlation coefficient.

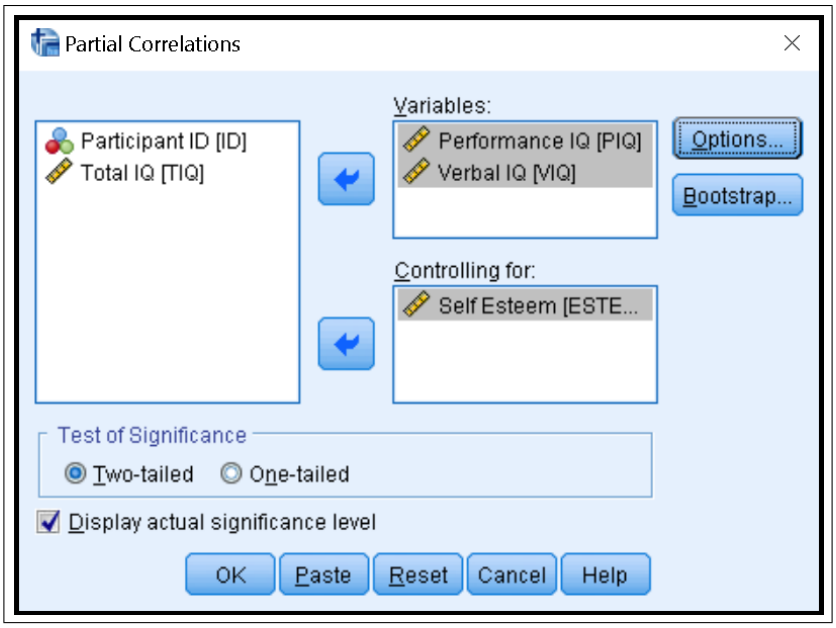


Figure 3.31: Partial Correlation Example: PIQ and VIQ, controlling for ESTEEM.

Exercise: Does controlling for ESTEEM substantially change the correlation between PIQ and VIQ? Why or why not?

3.6.3 Simple and Multiple Linear Regression

While correlation is about *describing* the relationship between a set of variables, regression is about *predicting* it. For instance, we might want to see if PIQ, VIQ, or ESTEEM predict TIQ. If we do this with one predictor, it is simple regression. With two or more predictor variables, it is multiple regression. In either case, the goal is to find the equation for a line of best fit that will allow us to predict Y (our dependent variable), given specific levels of X (our predictor variables). Both simple and multiple regression are accessible in SPSS from the same dialog box: **Analyze** → **Regression** → **Linear**, depending on the number of 'Independent' variables selected.

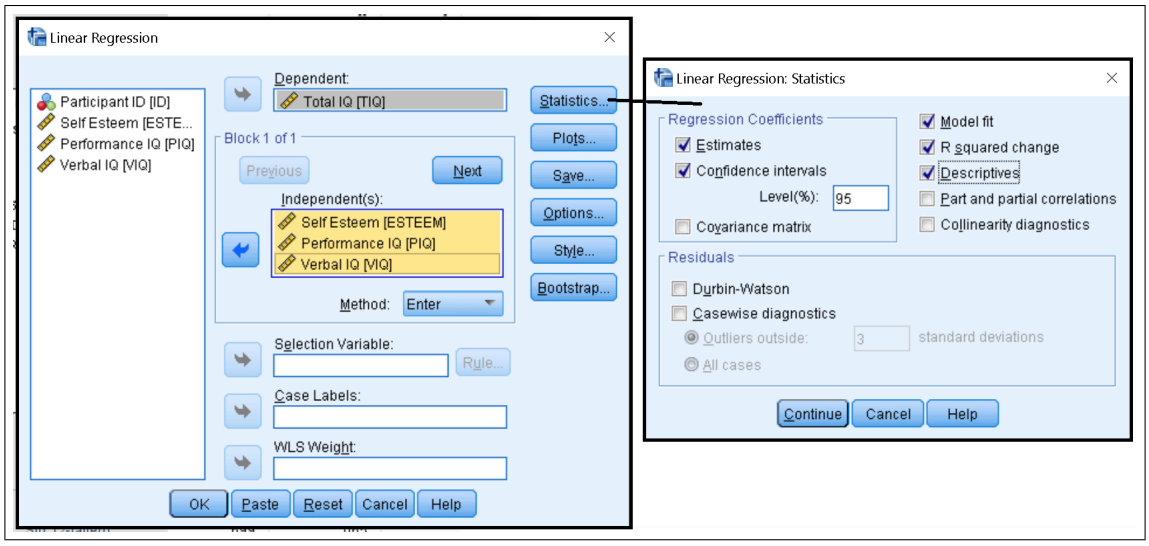


Figure 3.32: Linear Regression Dialog Box.

The Output: The output from a basic multiple regression consists of six tables:

1. **Descriptive Statistics**: Summarizes the entered variables (both predictors and dependent) in terms of their mean, standard deviation, and sample size;
2. **Correlations**: Provides their correlation matrix;
3. **Variables Entered**: Relates how the model changes as variables are entered/removed in the multiple regression[9];
4. **Model Summary**: Summarizes how the model changes during this process, if it was used;
5. **ANOVA**: This table tests the strength of the model (do the coefficients, taken together, predict a significant proportion of the variation in the dependent variable?); and,
6. **Coefficients**: The final table breaks the regression down by predictor.

> **Exercise:** Is the overall regression of PIQ, VIQ, and ESTEEM on TIQ significant? Which predictors seem to be most important, controlling for the others? What is the regression equation produced by this model?

3.6.4 Visualizing Simple Regression

Simple regression is relatively straight-forward to visualize. This is because we are using one variable (e.g., PIQ) to predict another (e.g., TIQ).[10] To visualize the relationship between these two variables, we can place PIQ along the X-axis of a scatterplot, with TIQ on the Y-axis. Then we can request our regression line be overlaid on the plot. The basics of this plot can be selected via **Graphs** → **Chart Builder**, by choosing the basic scatterplot template and dragging PIQ to the X-axis and TIQ to the Y-axis.

[9] The regression we conducted is called a 'simultaneous' regression, since all the variables are entered/evaluated at once. There are other so-called 'model building' techniques for regression that computationally decide if predictor variables are important or not. These 'step-wise' techniques typically have multiple steps, summarized in this section of the output.

[10] *Multiple* regression, on the other hand, is notoriously tricky to visualize in this manner since we need an extra dimension for *each* predictor variable in the model!

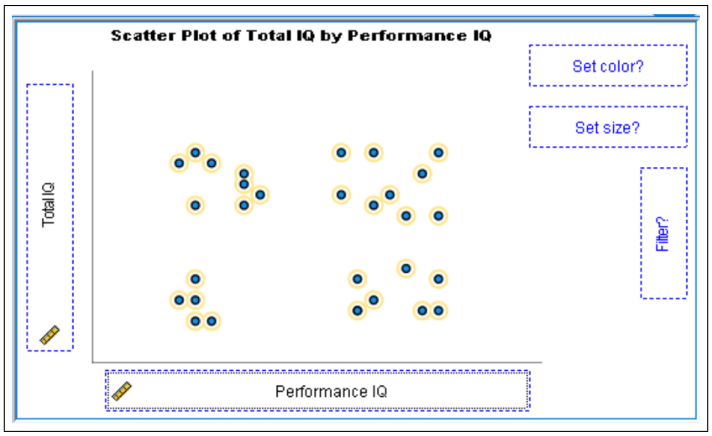


Figure 3.33: Visualizing Regression 1: Chart Builder Template

Click 'OK' to have your scatterplot appear in the **SPSS Output Viewer**, and double click on it to open the **Chart Editor**.

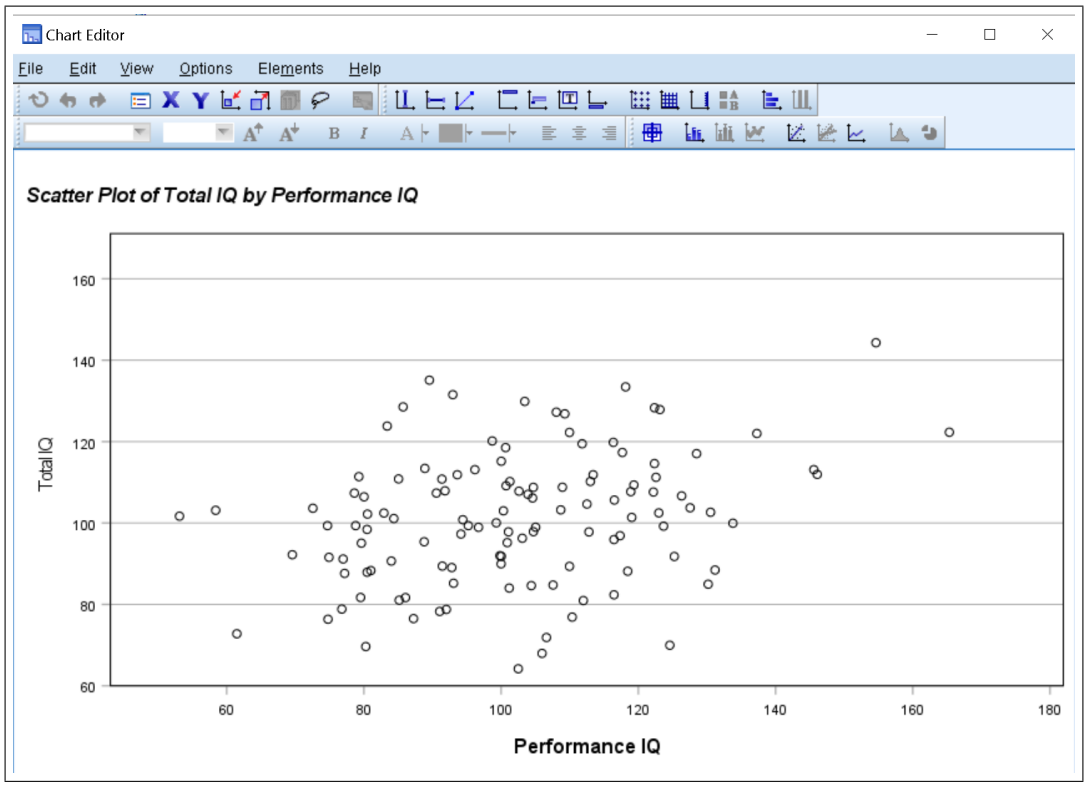


Figure 3.34: Visualizing Regression 2: Chart Editor.

Visually, we can see that there seems to be a moderate positive trend happening between PIQ and TIQ (as PIQ scores increase, so do TIQ). The steepness of our regression line will demonstrate the extent of that relationship. The next step is to click on the 'Add Fit Line at Total' button on the ToolBar. This will automatically add the regression line, and bring up a panel with more options. For example, you could add different types of

lines (e.g. LOESS curves, quadratic or cubic functions, or simple means), or confidence bands to the plot.

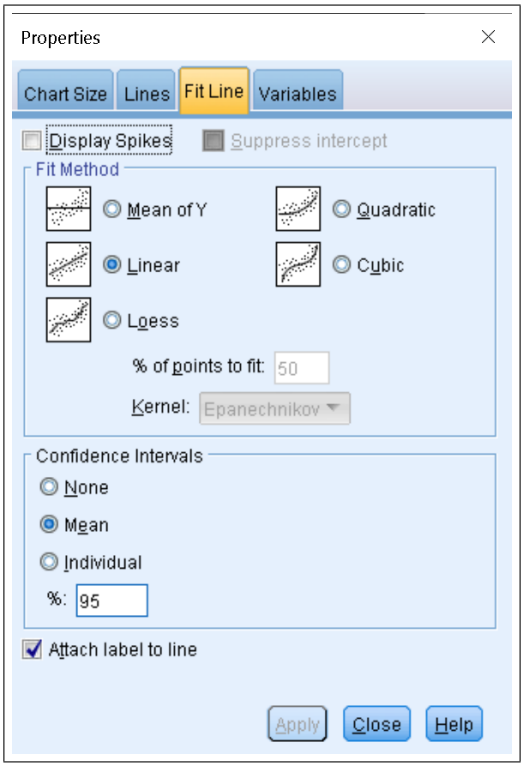


Figure 3.35: Fit Line Properties

I have chosen here to add 95% confidence bands around the mean, as well as have moved the labelling of the line to make it less distracting.[11] As soon as you add a regression line, SPSS also gives an index of its goodness in terms of R^2. This is basically an effect size measure for regression, with a value closer to 1.0 meaning a near perfect relationship between X and Y. Its interpretation is in terms of the proportion of variance of Y that is account for by X.

[11] This was done simply by click and dragging it to the new location. The only caveat is that it is 'stuck' to the regression line – you can't put it anywhere you would like on the plot.

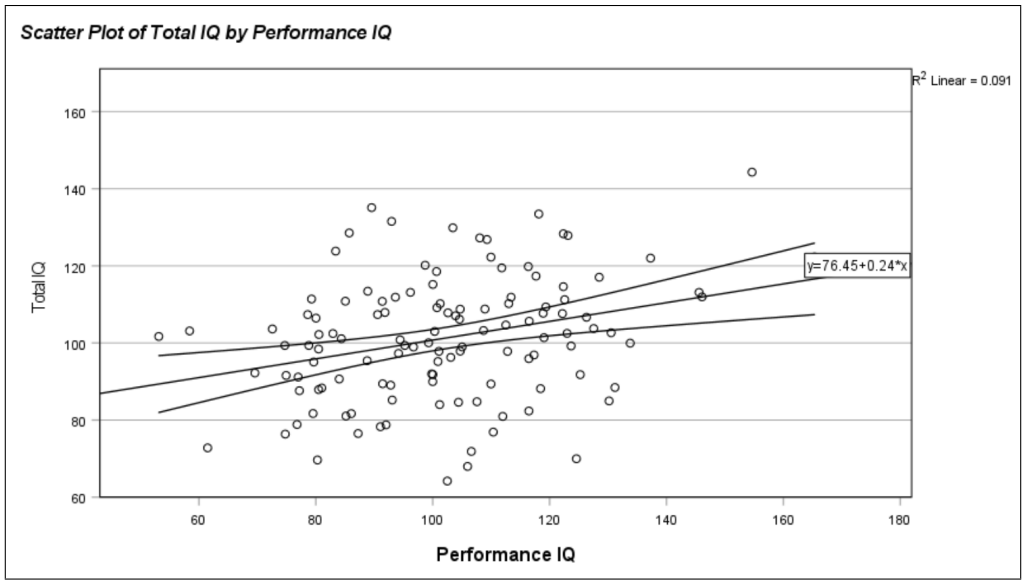


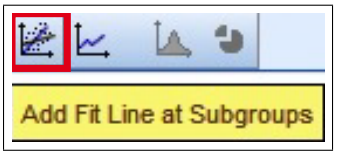
Figure 3.36: Visualizing Regression 3: The Plot.

3.6.5 Visualizing Multiple Group Simple Regression

But what if our data actually pertained to different occupations? Could we incorporate more information into this plot?

> Load `OCCUP.SAV`, which has an ID variable (same as in `IQSURV.SAV`), but also an occupation variable. Import OCCUP into your dataset and save it as MYIQSURV.SAV.

What we want to do know is to reproduce the previous graphic, but also tell SPSS that we know have a categorical grouping factor (OCCUP) that should be displayed. To do this, rerun the **Chart Builder**, and choose 'Grouped Scatter' instead of 'Simple Scatter'. Drag OCCUP to the 'Set Colour' box in the upper right corner and click 'OK'. The symbols on the new plot will now be in 5 different colours – one for each occupational category.



To add fit lines for each occupation separately, click on the previously greyed out but now selectable 'Add Fit Line at Subgroups' button on the ToolBar. This will automatically add five regression lines to the plot (one for each occupation), as well as report the strength of the relationship between PIQ and TIQ for each.

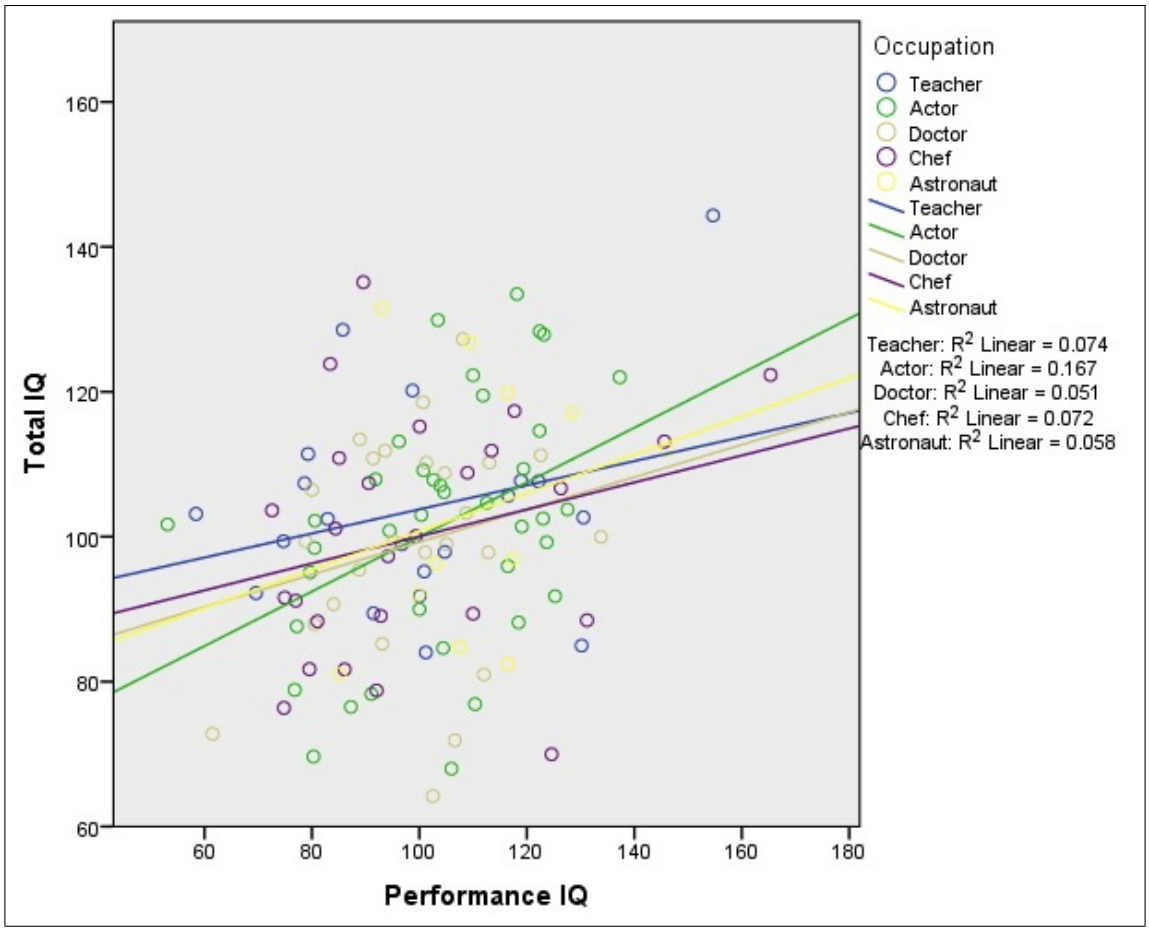


Figure 3.37: Multiple Group Regression with Fit Lines.

This plot is nice because it breaks down the relationship between PIQ and TIQ for us. We can see that the relationship is fairly minimal for the participants from most of the occupational categories – except for actors, who have a substantially steeper regression line. Note that, in order to maintain legibility, the confidence intervals and the equation labels have been removed from the plot.

3.6.6 Regression Diagnostics

While Sections 4.1.3-4.1.5 gave an example of how to conduct and visualize a regression analysis in SPSS, it skipped an imperative step: that of closely inspecting our data to ensure it meets the assumptions of the statistical test we are interested in. This process is generally referred to as 'diagnostics', and it should be a focus during any analysis.

Diagnostics for regression is an advanced topic, and it is recommended that you learn more about this methodology if it is unfamiliar to you.[12] However, since it is such an important topic, a few recommendations will be discussed here.

Lets return to our original regression of ESTEEM, PIQ, and VIQ on TIQ. With **Analyze** → **Regression** → **Linear** open, click on 'Statistics'. Things we are now interested in are:

- Collinearity Diagnostics: Tests our independent variables to see if any of them are a linear function of the others.

[12]Good resources include John Fox's *Applied Regression Analysis and Generalized Linear Models* (2008) and *Regression Diagnostics* (1991), and Barbara Tabachnick and Linda Fidell's *Using Multivariate Statistics* (2007).

- Durbin-Watson Residuals: Test for the serial correlation of residuals.
- Casewise Diagnostics, flagging outliers outside 2 standard deviations: Checks for cases that are unusual in the model.

Before running the regression, open the 'Save' tab, and choose the following options:

- Unstandardized Predicted Values: To be plotted to help identify the influence of outliers.
- Studentized Residuals: To test if the residuals are normally distributed.
- Cook's D(istance): Also, to be plotted to help identify the influence of outliers.

Exercise:

1. In the **Model Summary** table, look at the Durbin-Watson statistic. Use the 'What's This?' feature to find out how to interpret this result.

2. Look at the two **Collinearity Statistics** columns in the **Coefficients** table. Values close to 0 on Tolerance, or very large on VIF (as a heuristic, VIFs greater than 5 are typically considered 'large'), indicate possible problems with multicollinearity. Are there any issues of that sort with this model?

3. Look at the **Residual Statistics** table. Use the 'What's This?' feature to learn about some of the diagnostic measures.

4. Return to the datafile. Note the three new variables: **PRE_1** (unstandardized predicted values), **SRE_1** (studentized residuals), and **COO_1** (Cook's Distances).

5. Test if SRE_1 is normally distributed using the **Explore** procedure.

6. Make a scatterplot with PRE_1 on the X-axis and SRE_1 on the Y-axis. This is a traditional residuals diagnostic plot for multiple regression. You want it to look like a band with no discernible pattern.

7. Finally, plot COO_1 on the X-axis of a scatterplot and SRE_1 on the Y-axis. Points the have a large SRE_1 (± 2) and a large COO_1 is said to be an *influential* outlier, and may have a dramatic impact on your regression results.

At this point, you have the tools to work through Hands On Exercises I through IV. Good luck!

4 Miscellaneous Topics

4.1 Reliability Analyses

It is pretty common to want to investigate the **reliability** of a psychological inventory. **Reliabilty** generally refers to the consistency of results and can be assessed within an instrument or across time points. For instance, you might have 30 questions pertaining to a psychological construct, like anxiety, and want to know if all 30 of your questions are really tapping the same idea. Alternatively, perhaps you have three raters or judges coding a behavioural study. Reliability analyses can help address the level of agreement (or disagreement!) between the raters.

To run a reliability analysis, select **Analyze** then **Scale** then **Reliability Analysis** to see the following dialogue box:

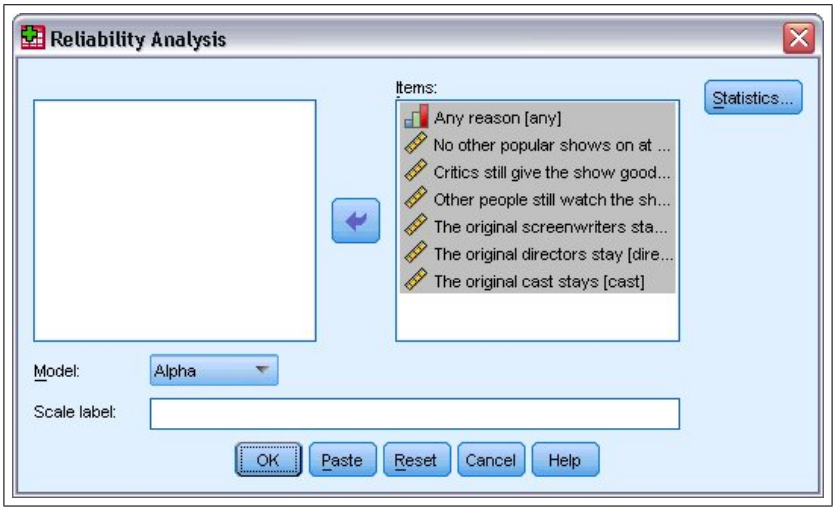


Figure 4.1: Reliability Analyses.

This is used to select the variables you want to include in your analysis. The `Statistics` screen allows you to configure the model in more detail:

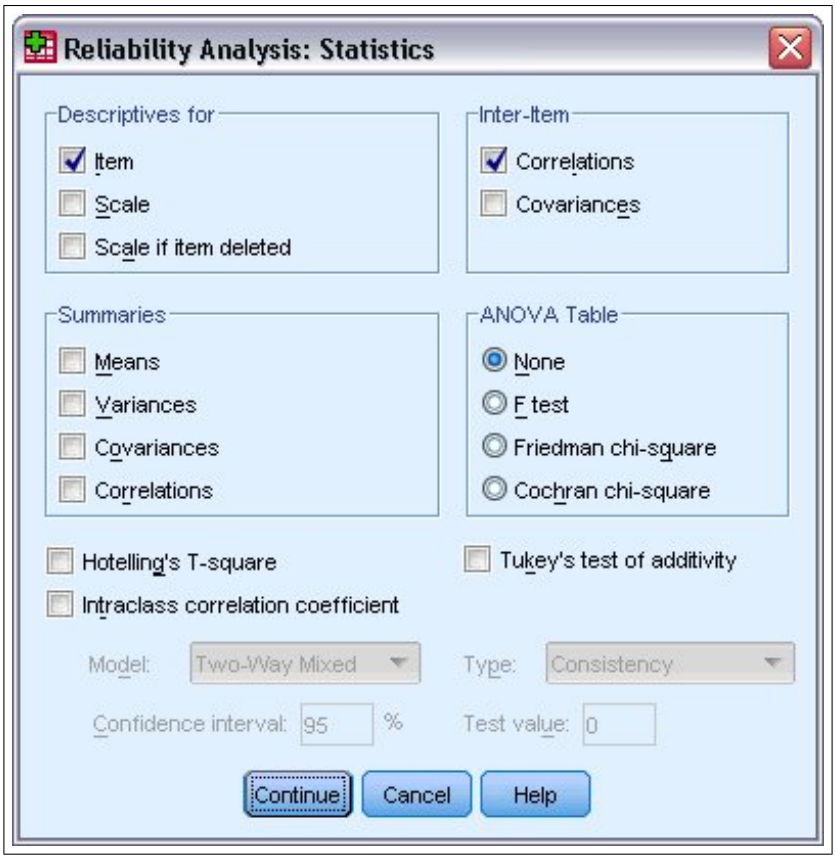


Figure 4.2: Reliability Analyses.

If you are looking for the unidimensionality of a scale, people will often report Chronbach's α. If you are interested in inter-rater agreement, then you would choose to see the Intraclass Correlation Coefficient.

4.2 Syntax Cribsheets

As mentioned in Section 1.6.4, SPSS features a powerful backend that allows users to access more features of the application than they can via the regular interface. The downside is that SPSS has its own programming language, with its own rules, quirks, and idiosyncrasies. That being said, learning how to use syntax can be extremely useful,[1] and if you plan on using SPSS frequently, it is highly recommended you become comfortable with it.

As previously mentioned, the 'Paste' command that appears in almost all of the analysis windows is extremely useful in this regard. It will take whatever settings you have chosen in the dialog box, and pastes the syntax that would produce those results into the **Syntax Editor**. From there you can save, edit, or run those commands. The following sections are not meant to be an extensive lesson on syntax, but to show you some of the things we can do with it, and how to approach command-based programming.

> Note: If you want to run these quick examples, make sure you have a dataset open.

[1] And could even be a stepping stone to learning other command based applications, like R!

4.2.1 Quickly Create an ID Variable

When collecting your data, it is useful to assign each participant a unique identifier. This allows you to go back and match your raw surveys back to the dataset, which is extremely handy if you begin your analyses and realize some values are amiss (e.g. on a scale of 1 to 5, you notice Lulu has a recorded score a 53!).

If you didn't assign an identifier in advance, you can easily use syntax to create a new variable, `ID`, which gives a consecutive, numeric identifier to each case in the dataset. To do this, go **File** → **New** → **Syntax**. In the syntax window, type:

```
COMPUTE ID=$CASENUM.
  FORMAT id (F8.0).
  EXECUTE.
```

Run the above code, and when you return to **Data View** you should notice a new column, with a unique identifier for each participant! Remember: you can always change the ordering of the variables in **Variable View** by dragging the variable row number to the desired location.

4.2.2 Some Basic Data Modications

A series of commands that can come in handy when working with a new datafile.

SORT CASES Reorders the sequence of cases in the datafile in either ascending (A, default) or descending (D) order.

General form: `SORT CASES BY` *varlist*(A or D).

Example: `SORT CASES BY SEX INCOME (D).` (Sorts by sex than income, in descending order)

SELECT IF Selects cases based upon a logical expression (further analyses will only be based upon the cases for which the expression is true).

General form: `SELECT IF` (*logical expression*).

Example: `SELECT IF (X GT 600 OR Y GT 600).)` (Only select cases where either X or Y is greater than 600)

SPLIT FILE Requests that output be split by a particular (categorical) variable.

General form: `SPLIT FILE BY` *varlist*.

Example: `SPLIT FILE BY SEX.` (Analyses will produce output for males and females separately)

SAMPLE Randomly sample a set proportion or number of observations from the datafile.

General form: `SAMPLE` *percentage*. OR `SAMPLE` *n* `FROM` *m*.

Example 1: `SAMPLE .25.` (Sample 25% of the cases from the datafile)
Example 2: `SAMPLE 80 FROM 200` (Sample 80 cases from the 200)

WEIGHT Weight cases by a particular variable.

General form: `WEIGHT BY` *varname*.

Example: `WEIGHT BY WTFACTOR.` (Analyses will be weighted by the variable WTFACTOR)

4.2.3 The IF Command

IF statements are a basic feature of computer programming languages. They are extremely useful for quickly manipulating data based upon some logical expression (e.g. that can be expressed as true or false). Their general form is:

```
IF (some logical expression) target variable = expression.
```

Some examples:

```
IF (X LT 1) Y1 = 1. (For values of X less than 1, set Y1 to 1)

COMPUTE Y2 = 0. (Quickly create a variable, Y2, and set all values to 0)
IF (X=1 OR X=2) Y2 = 3. (Set Y2 to 3 if variable X is 1 or 2)
```

4.2.4 Using SPSS as a Matrix Algebra Calculator

One of the other features of SPSS that isn't widely known is that it features a matrix algebra calculator that can be used via the syntax editor. This can be extremely helpful, especially if you take any graduate level statistics courses!

In order to use this feature, make sure your output default option is set so that commands are printed with the output results. To check this go to **Edit** → **Options** then to the 'Viewer' tab. In the 'Initial Output State' pane check 'Display commands in the log'. Click 'Apply', then 'OK'.

Open a new syntax window via **File** → **Open** → **Syntax**.

To start a SPSS Matrix session, your first command must be:

```
Matrix.
```

Your last command should be:

```
End Matrix.
```

To display a vector, matrix, or the result of some computation, enter:

```
PRINT [objectname].
```

Comments (remarks that SPSS won't try to process) can be made via:

```
/* comment */
```

A matrix is entered using COMPUTE statements. Commas indicate row elements, semi-colons result in new columns. The following would create a 3x3 square matrix object called NEWMAT:

```
COMPUTE NEWMAT = {1, 2, 3; 0, 4, 5; 1, 0, 6}.
PRINT NEWMAT.
```

A matrix object can be transposed:

```
COMPUTE NMPRIME = T(NEWMATRIX).
PRINT NMPRIME.
```

The rank of a matrix can be requested:

```
COMPUTE RANKX = RANK(NEWMAT).
PRINT RANKX.
```

The determinant of a matrix can be requested:

```
COMPUTE DETX = DET(NEWMAT).
PRINT DETX.
```

The inverse of a matrix can be requested:

```
COMPUTE INVX = INV(NEWMAT).
PRINT INVX.
```

Using a vector Y and a matrix X, we can solve a system of equations:

```
COMPUTE Y = {1;2;3}.
PRINT Y.
COMPUTE A = INVX*Y.
PRINT A.
```

. . . And so on. If this is a feature that interests you, a quick web search for "SPSS MATRIX tutorials" will yield many fruitful results.

4.3 Conclusion

SPSS is a fairly vast program that is constantly expanding. This workshop was intended to introduce you to some of the basic features, and get you running analyses that would be expected during a typical introductory statistics course. There are numerous other procedures for more advanced analyses (including multilevel or hierarchical modeling, factor analysis, survival analysis, even neural networking!) that await you if you require them but, unfortunately, are too complex to be summarized here. We hope that you found this material helpful!

5 Hands On Exercises

5.1 Exercise I

Data Entry

Create an SPSS data file containing the data below. Make it slightly more difficult by pretending to forget to enter age until all the other data is already entered, and then put the age column in the position it appears on the page.

- Code hair colour and eye colour as 0 (light) and 1 (dark)
- Code missing values (e.g. for Stephen, Molly, and Mildred) as 9

Name	Age	Hair Colour	Eye Colour
Joe	24	light	light
Mary	56	light	dark
Ted	42	light	light
Sally	52	light	light
Sue	12	light	light
Jack	18	light	light
Ed	65	light	dark
Mandy	28	light	light
Eric	46	light	dark
Janice	35	light	light
Diane	24	dark	dark
Kelly	15	dark	dark
Norman	62	dark	dark
Candice	78	dark	dark
David	32	dark	light
Michael	25	dark	dark
Terry	36	dark	dark
Colin	14	dark	dark
Reg	19	dark	dark
Rosemary	43	dark	dark
Stephen	25		
Molly	41		
Mildred	65		

Save the data as `EXERCISE1.SAV`.

Frequencies

With the data entered, run the **Frequencies** procedure on hair colour to make sure that you have 10 light haired, 10 dark haired, and 3 unknown haired participants.

Hair and Eye Colour Cross-Tabulation

1. First, create a **bar chart** that plots hair colour on the X-axis, with separate bars for eye colour. Label the bars with the percentage of the sample in that combination.
2. Do a statistical test of the relationship between hair and eye colour by running the **crosstabs** procedure, and seeing if the chi-square statistic is significant.

Mean Comparison

Investigate if the light haired members of our sample tend to be, on average, older or younger than the dark haired members.

1. First, look at the data by requesting a **boxplot** of the mean age for each group.
2. Test the difference in mean age by running an **independent groups *t*-test**.
3. Interpret the results.

5.2 Exercise II

Open the SPSS data file `DEPRESS.SAV`. This data comes from a study comparing depression levels in males and females. Each participant filled out a depression questionnaire which had 5 statements that participants had to rate according to how much they agreed to them (1 was coded as strongly disagree; 5 was strongly agree). Items 1, 3, and 5 were worded so that higher scores reflect depression (e.g. "I feel sad"), while items 2, and 4 were worded so that higher scores reflect happiness (e.g. "I feel happy").

1. Recode 'depress2' and 'depress4' so that higher scores will reflect depression.

2. Count how many of the questions on the survey participants left blank.

3. Compute a summary depression score for each participant by computing a sum of their scores on each question – but only if they answered *at least* 3 of the 5 questions.

4. Sort the data file by summary depression score, from highest to lowest.

5. Run a list cases procedure, showing participant id, number of questions missing, and summary depression for each case.

6. Find the average summary depression score for males and then for females by splitting the file by sex.

7. Attempt to replicate that result by using 'Select Cases' instead of 'Split File'.

5.3 Exercise III

The following data are reading test scores obtained from three grade 7 classes at the same school:

Grade 7A: 8, 10, 9, 8, 11, 7, 12, 13, 13, 9, 11, 9
Grade 7B: 11, 9, 6, 12, 15, 9, 8, 7, 14, 7, 10, 9
Grade 7C: 10, 14, 12, 7, 10, 15, 8, 8, 9, 6, 8, 15

1. After entering the data into an SPSS spreadsheet, code the scores from Class 7A as 0, Class 7B as 1, and Class 7C as 2.

2. Determine what the mean, median, mode, standard deviation, and variance is for each class using the split file function.

3. Ensure that there are 12 scores entered for each class. Using SPSS (e.g. not counting by hand), determine how many children obtained scores between 12 and 14.

4. Use the Explore function to construct a histogram, a QQ plot, and a boxplot for each class. What can these figures tell you about your data?

5. Are the scores normally distributed for all three classes? If not, address the non-normality by transforming the data (e.g. you could try a square root or Log10 transformation and determine which is more appropriate).

6. Plot the means for each class using a bar graph with standard error bars.

7. Analyze this data using a one-way ANOVA. Are there differences between the classes? If so, where are the differences?

8. Add a 'Previous Test Score' variable for the Grade 7A and 7B classes to your dataset:

   Grade 7A Previous Test Score: 15, 9, 15, 8, 6, 11, 12, 7, 13, 10, 13, 9
   Grade 7B Previous Test Score: 10, 7, 11, 13, 14, 14, 8, 8, 9, 11, 5, 15

9. Construct a scatterplot between this new 'Previous Test Score' variable and the test scores for the classes. Set the markers by class. Ensure that they symbol representing Class 7A is an X, while the symbol for 7B is a square. Add separate fit lines for both classes.

10. From looking at this scatter plot, do you think the previous and the new scores are related? Determine if they are related using a correlation analysis.

11. Run a regression analysis to asses the contribution of 'Previous Test Score' on the new test scores for the 7A and 7B classes.

5.4 Exercise IV

As a final exercise, please try importing the `GUERRY.TXT` dataset into SPSS. This dataset contains real, historical information collected by Andre-Michael Guerry in 1833. It pertains to the the literacy levels, suicide and crime rates, and other "moral variables" in the 86 departments of France. The data frame contains 85 observations (one for each department) on the following variables:

- DEPT: Department identification number.
- REGION: Region of France (North, South, East, West, Central)
- DEPARTMENT: Department name.
- CRIME_PERS: Population per crime against persons.
- CRIME_PROP: Population per crime against property.
- LITERACY: Percent of population that can read and write.
- DONATIONS: Amount of donations made to the poor.
- INFANTS: The number of infants born.
- SUICIDES: The number of suicides.
- MAINCITY: Ordinal variable pertaining to department city size from 1 (small) to 3 (large).
- WEALTH: The per capita tax on personal property (a ranked index based on taxes on personal and movable property per inhabitant).
- COMMERCE: The number of patents per population.
- CLERGY: Ranked index of the number of Catholic priests in service to population size.
- CRIME_PARENTS: Crimes against parents; rank of ratio of crimes vs. parents to all crimes.
- INFANTICIDE: Ranked ratio of infanticides to population per capita.
- DONATION_CLERGY: Amount of donations made to the clergy.
- LOTTERY: The per capita wage on the Royal Lottery.
- DESERTION: A ranked index of military desertion (the ratio of soldiers accused to the force of the military contingent, minus the deficit produced by insufficiency of available billets).
- INSTRUCTION: The inverse of literacy.
- PROSTITUTES: Number of prostitutes registered in Paris, classified by department of birth.
- DISTANCE: Distance from Paris in km.
- AREA: The area of the department in 1000 km^2.
- POP1931: The department population in 1831.

Some Ideas for Analyses:

1. Load in the TAB-DELIMITED datafile `GUERRY.TXT`.
2. Determine if SPSS properly interpreted the variables, based upon the variable names.
3. Label the variables according to their description and determine if there is any missing data.
4. Fix the coding of MAINCITY.
5. Find the mean, median, and mode of INFANTS, SUICIDES, and POP1831.
6. Find the standard deviation and variance of CRIME_PERS and CRIME_PROP.
7. Find the 75^{th} percentile for WEALTH.
8. Generate a scatterplot that has INFANTS on the X-axis and LOTTERY on the Y-axis. Recreate this graphic with the points labeled by department.
9. Generate a box-plot for DONATIONS by REGION. Which departments appear as outliers?
10. Generate a scatterplot matrix for CRIME_PERS, CRIME_PROP, and CRIME_PARENT, and add the appropriate lines of best fit.
11. Correlate DONATIONS and CLERGY.
12. Attempt a multiple regression predicting SUICIDE using any variables you think might be relevant. Interpret the output.

Index