Australian National University

Digital literacy training

SPSS

Introduction

2020

Table of Contents

To start SPSS **1**
Importing Excel files **1**
The Data View window **1**
The Variable View window **2**
Create new variables **3**
The Output Viewer Window **3**
Create graphs **4**
Edit histograms 4
Edit scatterplots 6
Edit bar charts 7
Descriptive statistics **7**
Use a Syntax file **10**
Export graphs and tables to Word **10**
Shortcut to Recently Used Menu Items **10**
Data manipulation **11**
Other resources **12**

Presenter: Nyree Mason nyree.mason@anu.edu.au

The IBM Statistical Package for the Social Sciences (SPSS) was developed as a data management and analysis tool. This course will show you how to:

- Open and import data.
- Set-up data files using the *Variable View* window.
- Manage data using the *Data View* window.
- Create graphs and tables and export them to different file formats.
- Obtain descriptive statistics (i.e., percentages, means and standard deviations).
- Use the *Syntax Editor* to keep records of data manipulation & analyses and to run routine analyses on multiple data files.
- Compute new variables and recode existing ones.
- Sort variables.
- Select cases and split files to analyse subgroups separately.

The data files

ql.anu.edu.au/training

Save these 3 files:

Employee_Data.sav

Anxiety_2.sav *

Car_sales.sav *

* These files will be used for the Advanced Significance Testing SPSS course.

To start SPSS

1. Log on to the computer using university ID (e.g., 'u1234567') and password.
2. Click on the Windows Icon and search for **IBM SPSS Statistics 25**.
3. SPSS **File → Open → Data** find and open the **Employee_data.sav** data file and click **OK** (all data files have the ***.sav** extension in SPSS). Alternatively, you can double-click on the file and it will open SPSS and the file automatically.

Note: You can open a file any time within SPSS using **File → Open → Data**. SPSS allows you to have multiple data files open at once, and has reminders in the Output file to let you know which data file was used for each analysis you perform. You can also open any file from another **spreadsheet style** application (e.g., Excel) following the SPSS prompts.

Importing Excel files

Note: In Excel it's best if you have your variable names in the first row of the spreadsheet so that SPSS recognises them as labels. Do not have any annotations etc. underneath your data in the Excel spreadsheet, as this will confuse SPSS about your variable types during the import.

1. Go to **File → Open → Data** and in the drop-down menu for **Files of Type** choose **Excel** and then find your file.
2. Select **read variable names from the first row of data** if necessary, and if there are multiple sheets in the Excel file select the relevant one. Then click **OK**.
3. You will have to set up the data file afterwards in the **variable view** (see details below).

The Data View window

Click on the **Data View** tab at the bottom left corner of the screen. This is where you can view and enter your data in a similar way to any other spreadsheet-style program.

- There is one column per variable (e.g., date of birth, gender, salary).
- There should be only one person/case per row.
- Columns can be rearranged: highlight one column by clicking on the variable name, and then click and drag the name to another position (this can also be done in the **variable view**).
- Extra columns can be inserted into the data set by clicking on the **name** of the variable you want the new column to go next to, and then going through the menu: **Data → Insert Variable**.
- You can also delete whole columns (or rows) by clicking on the name of the variable (or row) you want to delete (thus highlighting it) and then pressing the **delete** key.

Note: The **Undo** option in the Edit menu may come in handy if you make a mistake.

The Variable View window

Click on the **Variable View** tab at the bottom left corner of the screen. This is where you tell SPSS about your data (i.e., names, labels, type, levels of measurement). The information you provide in the *Variable View* can help you avoid making common mistakes in data analysis: SPSS will limit the types of analyses you are allowed to perform to those which are appropriate to the statistical properties of your data.

- **NAME** determines what the columns will be called in the **Data View**.
- **TYPE** determines what form the data in each column takes (e.g., **jobcat** is **numeric** because numbers are entered, **gender** is **STRING** because letters are entered, **bdate** is **DATE** because dates are entered).
- **WIDTH** determines how many digits/letters can go in each column (**Note:** NOT how many are displayed). **WARNING: Do not change the width after the data is entered** unless you are 100% sure you want to. Old versions of SPSS will delete any characters that are already entered that go beyond the new width limit **without warning!** (i.e., if the width is set to 4 a number such as 12,345 will be changed to 1,234).
- **DECIMALS** determines how many decimal places are displayed (i.e., 3 → 12.345).
- **LABEL** gives the variable a more descriptive label in the **Output Viewer**.
- **VALUES** labels the coding values entered in the column (i.e., m = male, 1 = Strongly Disagree), and will present these in the **Output**. **WARNING: SPSS prefers numerical codes, and some statistical analyses cannot be performed unless the data is coded with numbers.**
- **MISSING** determines which coding values will be considered missing and as such will be excluded from statistical analysis (e.g., 0 = "Other" Response, 77 = Invalid Answer, 88 = Not Applicable).
- **COLUMNS** determines how wide the columns will be *displayed* in the **Data View**. You can change this manually by clicking and dragging the columns wider in the data view also.
- **ALIGN** changes the alignment of the data characters (Left, Centre, and Right).
- **MEASURE** determines the level of measurement of each column, either: *nominal* (e.g., types of running shoes), *ordinal* (e.g., university grades: Pass, Credit etc.), or what SPSS terms *scale*, which includes both *interval* and *ratio* levels of measurement (e.g., real numbers: temperature, weight, time, frequency etc.).
- **ROLE** is designed to help you choose the right variables for analysis (**Input =** Independent variables, **Target =** Dependent Variables etc.). It is not fully functional in SPSS as yet.

Handy Hint: In the **Data View** you can see the labels associated with the raw data by going to **View → Data Labels**. When entering data you can also use this function: click on an empty cell, use the drop-down menu to select the appropriate label for your data and SPSS will fill in the number for you automatically (behind the scenes).

Create new variables

Try creating a new variable for job satisfaction called jobsat, which is represents levels of agreement to the statement: "I am satisfied with my job". It is measured on a 5-point scale with these value labels:

1. Strongly disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree
99. Not applicable

Any "Not applicable" values will be coded as a missing value of 99.

Handy Hint: You can copy one value and paste it to multiple cells at once in the Variable View: copy one missing value, select the cells you want to paste it to and then paste.

The Output Viewer Window

The charts and analyses you run appears in an **Output Viewer Window**, which can be saved separately with the extension ***.spv**. The material in the output viewer is **completely independent of the data**. If you change the data in any way, the relevant analyses will have to be run again.

Please note that output files from older versions of SPSS may not always be viewed properly in newer versions. Exporting as another file type (described later) will avoid any problems.

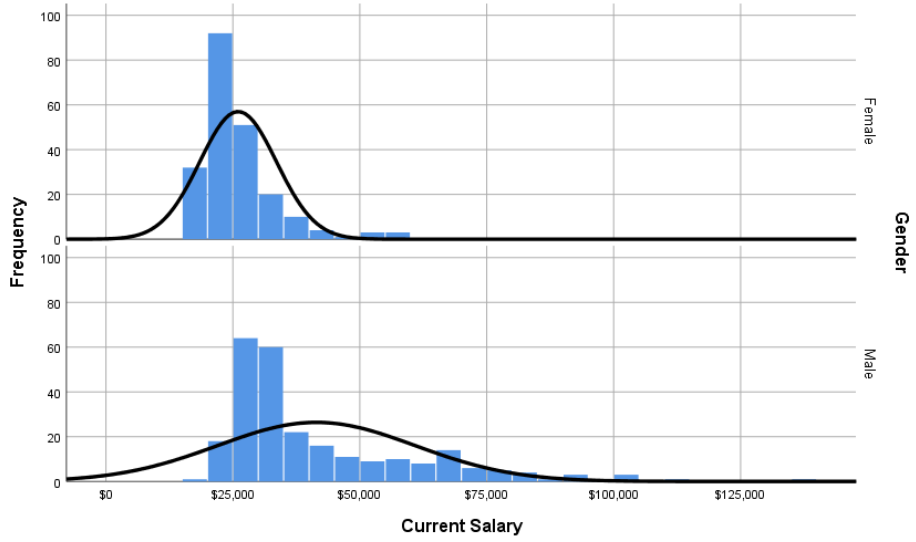
There is a navigation pane on the left hand side and the output appears in the main window. You can change the text anywhere within the Output Viewer by clicking twice on the text, and editing. A text formatting toolbar will also appear, which you can use to alter the text type.

If you wish to add additional text (e.g., Annotations), click on the object in the Output **under which** you wish the new text box to appear. Then go to the **Insert** menu and select **New Text**.

By default, SPSS will add the syntax used to create the output as text in the Output Viewer for future reference. It will also record the file path of the data file used to create the output.

Create graphs

HISTOGRAMS describe the distributions of **scale** data best but are also used for **ordinal** data. It can display what the normal curve would look like given the mean and standard deviation of your data. **Graphs → Legacy → Histogram** select a continuous variable (e.g., current salary) and tick **display normal curve** then **OK** or **Paste**. These can be split into groups (e.g., by gender) using **Panel by Rows** (graphs one above the other) or **Panel by Columns** (graphs side by side). This has the advantage of maintaining identical X and Y axis value ranges. Then click **OK**.



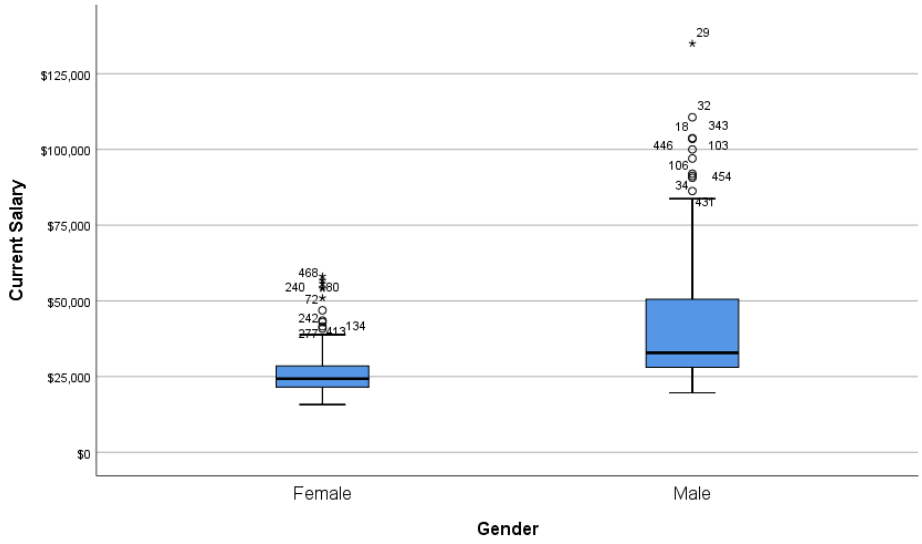
Comparing the histograms to the normal curve suggests the distribution of salary is positively skewed for males and closer to normal for females.

Edit histograms

- Double-click on the graph to open the **Chart Editor** and **Properties** windows. If the **Properties** window doesn't open automatically, just click on the graph in the **Chart Editor** and it will appear.
- Whatever you want to change, the easiest option is to click on it and the **Properties** window will give you all the appropriate formatting options for that feature in separate tabs.
- For example, to change the range and increments used on either axis, click on one of the values, select the **Scale** tab in the **Properties** window and change the values you wish.
- To change the colour of bars for example, click on the bars and **Properties** window will give you a tab for **Fill & Border** options. This allows you to change the look of the bars – even add a pattern. You can also change the binning here (width of the bars).

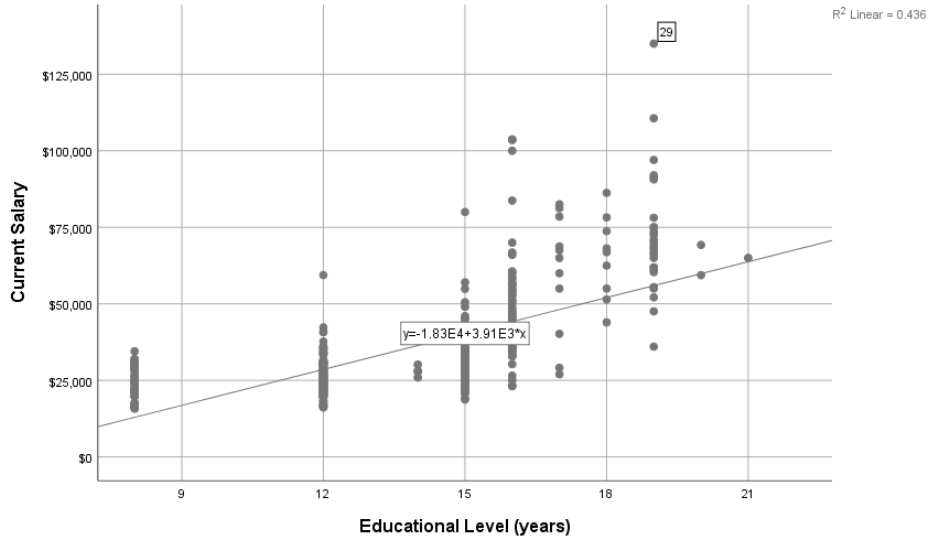
Note: When you are finished editing the chart, you only need to close the **Chart Editor**, and the **Properties** window will close also. If you accidentally close the **Properties** window, the easiest way to get it back is to double-click on the chart in the Editor. Alternatively, you can go to the **Edit** menu in the Editor and select **Properties**.

BOXPLOTS are also useful for describing the distributions of **ordinal** data but are also good way to describe **scale** data. To produce the same comparison of current salary distribtions for males and females separately click on **Graphs → Legacy → Barcharts** and choose **Simple**. Leave Data in Chart Are **Summaries for groups of cases** then click **Define**. Select an ordinal or scale variable (e.g., current salary) and move it into the **Variable** box. Choose a grouping variable (e.g., **Gender**) and move it into the **Category Axis** box, then click **OK**.



The thick line represents the **median** and the size of the box is the **interquartile range (IQR)** (between the **25th** and **75th percentiles**). The lower whisker is the last data point *within* 1.5 times the IQR below the 25th percentile, and the upper is the first 1.5 times above the 75th. Any point outside the whiskers is flagged as a potential outlier, and its line number in the **Data View** is shown. The boxplots should look symmetrical if the data is normally distributed. Again, the one for females is acceptably normal, but not that for males.

SCATTERPLOTS describe the relationship between two scale or ordinal variables. **Graphs → Legacy → Scatter → Simple** select one scale variable for the x-axis (usually the IV, e.g., education level) and another for the y-axis (usually the dependent variable, e.g., beginning salary), then **OK**. These can also be panelled if you wish.

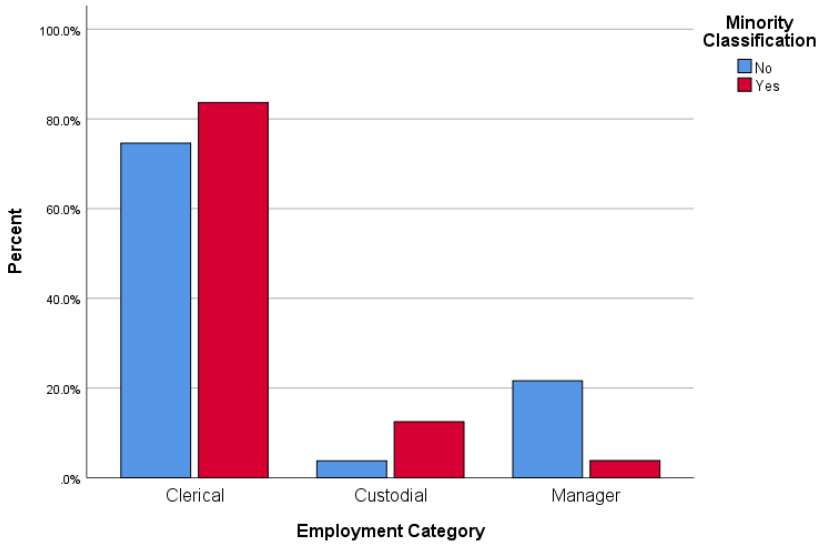


The **linear regression line** shows the strength and direction of the relationship between the two variables. This is a positive relationship (as one increases, so does the other). The R^2 value for the relationship indicates the amount of variance in salary being explained which is 43.6%.

Edit scatterplots

- Double-click on the graph to open the **Chart Editor**
- To get the linear regression line, go to **Chart → Elements → Fit Line at Total**. The R squared value will be displayed in the top right-hand corner to indicate the amount of variance explained by the line. With the line highlighted you can change the type you want in the **Properties Window** in the **Fit Line** tab (in this example a Quadratic equation was used as it provided a better fit to the data). The Regression Equation for the line is displayed in the centre of the chart – this can be removed by unticking the box that says **Attach Label to Line** in the **Fit Line** tab.
- Other elements can be changed as per the Histogram example above. To find the case number of potential outliers, go to **Chart → Elements →** and tick **Data Label mode**. This changes your cursor to a target, and click on the data point in question to see which case number (row number in the data view) it is.
- Close the **Chart Editor** window to return to the output.

BAR CHARTS are useful for describing the frequencies/percentages of nominal and ordinal data. They look like histograms but have fixed bars which can be ordered in any way you wish. To produce a Clustered Bar Chart go to **Graphs → Legacy → Bar → Clustered** with a Nominal/Ordinal variable as the **Category Axis** (e.g.,job category) and a Nominal/Ordinal variable (e.g., minority) in the **Define Clusters by** box. Check the **% of cases** radio button to display percentages on the Y-axis rather than counts. Click **Continue** and then **OK**.



When choosing to show percentages rather than frequencies, the bars represents the **% of data in each clustered group** (minority) **within each axis group** (job category); e.g., here just over 80% of people who are in a Minority group are employed in Clerical positions.

Edit bar charts

- Double-click on the graph to open the **Chart Editor**
- You can change the order of the categories on the X-axis by clicking on a category label, then in the **Properties Window** go to the **Categories** tab and rearrange the order by clicking on a category and using the up and down arrows.
- You can do the same for the order of the bars by clicking on a bar, selecting the **Categories** tab and rearranging the order the same way.
- Other elements can be changed as per the Histogram example above.
- Close the **Chart Editor** window to return to the output.

Descriptive statistics

Note on Terminology: **Dependent variables** (a.k.a. Response) are those with which you expect to measure some effect. **Independent variables** (sometimes called **Factors** in SPSS) are those that you expect to have an effect on the Dependent Variables. For example, if you think males will be taller on average than females, sex will be the Independent Variable and height will be the Dependent Variable.

DESCRIPTIVES are used to primarily to produce means and standard deviations for scale data. For example, to find the mean and standard deviation for the scale levels of measurement click **Analyze → Descriptive Statistics → Descriptives**. Choose all the **Scale** levels of measurement (e.g., **education level, current salary** etc.) and move them into the **Variable(s)** box with the arrow button. In the **Options** area you can also choose to display sum, variance, standard errors, range, skewness and kurtosis. Then click **OK** to run the analysis.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Educational Level (years)	474	8	21	13.49	2.885
Current Salary	474	$15,750	$135,000	$34,419.57	$17,075.661
Beginning Salary	474	$9,000	$79,980	$17,016.09	$7,870.638
Months since Hire	474	63	98	81.11	10.061
Previous Experience (months)	474	0	476	95.86	104.586
Valid N (listwise)	474				

FREQUENCIES Used to describe nominal/ordinal data. To find the frequencies and percentages for the nominal levels of measurement click **Analyze → Descriptive Statistics → Frequencies**. Choose the nominal variables (e.g., **jobcat, gender** etc.) and move them into to **Variable(s)** box. Then click **OK** to run the analysis.

Employment Category

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Clerical	363	76.6	76.6	76.6
	Custodial	27	5.7	5.7	82.3
	Manager	84	17.7	17.7	100.0
	Total	474	100.0	100.0	

The **Percent** column calculates the percentage out of the total number of cases *including* missing cases. The **Valid Percent** column calculates the percentage out of the number of valid cases which *excludes* missing cases.

You can also produce **multiple bar charts**, histograms or pie charts using a Frequencies Analysis. Click on the **Charts** button and select **Bar** and tick **Percentages**. For **scale** or **ordinal** data, you can produce a **histogram** with a **normal curve** but be careful to untick ''display tables' for scale data before you run the analysis, or you will probably get very large and unwanted tables as well.

Frequencies can be used to produce the same "Descriptives" for **scale** data also but provides extra information particularly useful for **ordinal** data: quartiles, percentiles, median and mode. Select a scale or ordinal variable, untick "display tables", click on the **Statistics** button and select the statistics you require.

EXPLORE is useful when you want **descriptive statistics** for scale data broken into groups (e.g., means for males vs females). To find the difference in current salary between males and females: **Analyze → Descriptive Statistics → Explore**. Put the dependent variable (which must be scale) **beginning salary** into **dependent list** and the independent variable(s) (which must be nominal/ordinal) **gender** into **factor list** box. In **Statistics** choose **Descriptives** and you'll note that **95% confidence interval** is the default. You can also select **outliers** if you want SPSS to show you the cases that have values greater than 3 standard deviations away from the mean. In **plots** *unselect* **stem-and-leaf** and select **box plots** (so you can see the outliers and assess normality). Click **Continue** and then **OK**.

Descriptives

	Gender			Statistic	Std. Error
Current Salary	Female	Mean		$26,031.92	$514.258
		95% Confidence Interval for Mean	Lower Bound	$25,018.29	
			Upper Bound	$27,045.55	
		5% Trimmed Mean		$25,248.30	
		Median		$24,300.00	
		Variance		57123688.27	
		Std. Deviation		$7,558.021	
		Minimum		$15,750	
		Maximum		$58,125	
		Range		$42,375	
		Interquartile Range		$7,013	
		Skewness		1.863	.166
		Kurtosis		4.641	.330
	Male	Mean		$41,441.78	$1,213.968
		95% Confidence Interval for Mean	Lower Bound	$39,051.19	
			Upper Bound	$43,832.37	
		5% Trimmed Mean		$39,445.87	
		Median		$32,850.00	
		Variance		380219336.3	
		Std. Deviation		$19,499.214	
		Minimum		$19,650	
		Maximum		$135,000	
		Range		$115,350	
		Interquartile Range		$22,675	
		Skewness		1.639	.152
		Kurtosis		2.780	.302

Editing Tables

- Double-click on the table so that it gets a dotted line around it. Move your mouse onto the right line of the **statistic** box. When you get the double-headed arrow, click and hold out the mouse while dragging it towards the right. This should expand the box so that more figures are displayed. You can minimise and maximise the width of all columns this way.
- To change the text, double-click on a label type the new text. You also have a **Formatting Toolbar** which you can use to change the font etc..
- To delete a value from a table click once on a value and press the delete button. If you delete the corresponding statistic for all categories, the whole line will disappear. You can delete all unnecessary/unwanted information in the table this way.
- To change the number of decimals displayed, highlight the appropriate values and right click on the selection, and select **Cell Properties** from the menu. In the **Format Value** tab there is a decimals option.

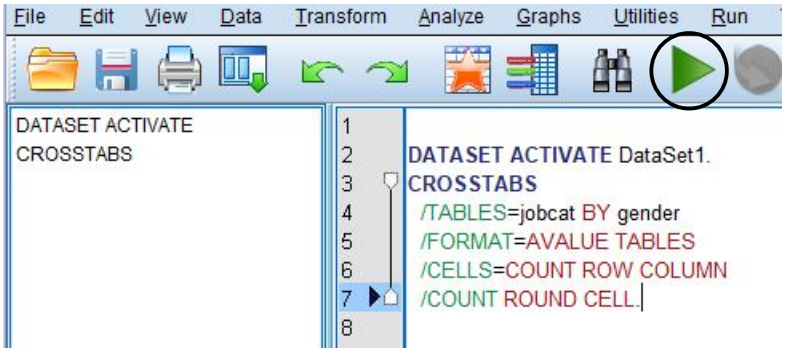
CROSSTABS can be used for nominal/ordinal data only. To find the relationship between employment category and minority classification: **Analyze → Descriptive Statistics → Crosstabs**. In the **Rows** box enter a Nominal/Ordinal variable (e.g., jobcat) that you want represented in the rows of the contingency table. In the **Columns** box enter the Nominal/Ordinal variable you want in the columns (e.g., minority). Click the **Cells** button and note than **Observed** is the default in order to display the actual count in each cell. You can also choose to display **Row and/or Column Percentages** here. **Format** again, is not necessary. Click **Continue** and then **OK**.

Employment Category * Minority Classification Crosstabulation

			Minority Classification		
			No	Yes	Total
Employment Category	Clerical	Count	276	87	363
		% within Employment Category	76.0%	24.0%	100.0%
		% within Minority Classification	74.6%	83.7%	76.6%
	Custodial	Count	14	13	27
		% within Employment Category	51.9%	48.1%	100.0%
		% within Minority Classification	3.8%	12.5%	5.7%
	Manager	Count	80	4	84
		% within Employment Category	95.2%	4.8%	100.0%
		% within Minority Classification	21.6%	3.8%	17.7%
Total		Count	370	104	474
		% within Employment Category	78.1%	21.9%	100.0%
		% within Minority Classification	100.0%	100.0%	100.0%

Use a Syntax file

- Repeat the previous **Crosstabs** analysis but click **Paste** instead of **OK** this time.
- **PASTE** places the commands that SPSS uses to execute its tasks to a **Syntax File** (***.sps**).
- Useful for when you want to run the same analyses again on different variables or data sets. It is also a useful record of what you have done to the data file.
- In the **Crosstabs** syntax, try changing the variable name **gender** to **minority** and then highlight the **syntax** and then **Run → Selection**. You can also click the green "Play" button.



Export graphs and tables to Word

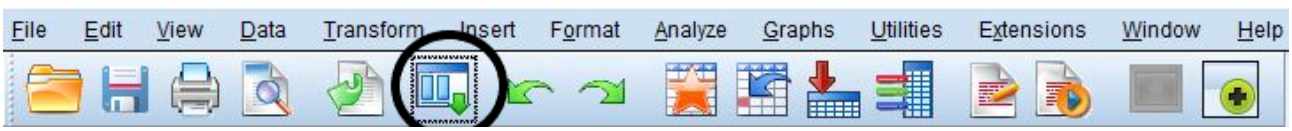
CUT-AND-PASTE METHOD Open a **Word** document (**Start → Programs → Word**). When you are happy with how the table/graph looks in SPSS, in the **Output Viewer** select the table, then **Copy**. In **Word** go to **Edit → Paste**

EXPORTING METHOD You can export the whole output document in one go, or select certain objects in the Outcput to export by using CTRL-Click method. In the **Output window**, go to **File → Export**, and you have the option to save the file as a Word, HTML, pdf, PowerPoint, Text (no graphs), Excel (no graphs) or "None" (separate picture files for graphs only) file types. It's usually best to **export visible objects only** otherwise you get a lot of non-visible SPSS syntax in the new file.

Please note that graphs will be exported as picture files and can no longer be edited (apart from size).

Shortcut to Recently Used Menu Items

There is a handy icon on the main toolbar for **Recently Used Dialogues** as a short cut to the menu items you have used the most.



Data manipulation

COMPUTING NEW VARIABLES is useful for creating new variables based on the data you have or transforming non-normal data, etc.. To transform skewed variable (e.g., salary) for further analysis, go to **Transform → Compute** then type **log_salary** in the **target variable** box. This will be the name of the new variable in your data file. In the **Function Group** box click on **Arithmetic**, and this will display all the arithmetical options in the **Functions and Special Variables** box. Here, select **Ln** (for Natural Log) and click on the up arrow to move it into the **Numeric Expression Box**. Replace the "?" symbol with the variable **salary** (type or select from the list and move it into the box). You should have the formula: **LN(salary)**. Then click **OK** or **Paste** and **Run**, and a new variable will be displayed at the end of your data set. **Save** the file again, so that this change will not be lost.

Please Note: To raise a number to a power, use the symbols ** (NOT ^ as you would in Excel).

RECODING VARIABLES is useful for grouping or regrouping variables (e.g., if you want to turn **education** in years into **education levels <=**12, 13-15, and >=16). **Transform → Recode → Into Different Variables** (so you don't overwrite an existing one) and in **Name** type **edulevel** and click **Change**. In **Label**, type **Education Level** to give it a name in the output. Then go to **Old and New Values** and click the **Range** radio button. Select **Lowest through ?** radio button and type in12, then in **New value** type 1, then click **Add**. Select the **? through ?** radio button, and type in 13 **through** 15. Give this range the **New value** 2, then **Add**. Finally, select the **? through highest** radio button, type in 16 and give the range the **New value** of 3, then **Add**. Then Press **Continue**, and then **OK** (or **Paste** and **Run**). You should have a new variable called **edulevel** in your Data View. If you go into the **Variable View** by clicking the tab at the bottom left of the screen, you can label these values (when a **Values** cell is selected click on the blue square and choose the appropriate options (e.g., type in **value** "1", and label it "8-12" then **Add** and so on).

SELECTING CASES is useful when you want to perform statistics on only a subsample of your data set. For example, to make sure that any analysis performed on the data is only performed on those cases that have been in the job for more than or equal to 72 months (6 years), follow these steps: **Data → Select Cases** then choose the **If condition is satisfied** radio button. Then click **If...** and move Months since hire (jobtime) to the box. Then select the greater than or equal to symbol (**>=**) and type "**72**" (so you get this formula: jobtime >= 72), then **Continue**, then **OK**. Note that in order to perform statistics on the whole data set from now on requires going back into **Data > Select Cases** and checking the **All Cases** radio button, then **OK**. SPSS shows you that it has done what you asked by putting diagonal lines through the case labels being excluded. In the bottom right hand corner it should also say **Filter On**.

SPLITTING FILES is useful when you want to analyse two or more groups within the data set separately. Split the files into those groups by **Data → Split File** and select the **Compare groups** radio button. This option splits tables by the variable selected. You could also use **Organize output by groups** and this gives you the same information, just in separate tables in the output. Move the grouping variable into the **Groups based on** box (e.g., gender), make sure the **Sort file by grouping variables** is checked, then press **OK**.

Other resources

Training notes

To access training notes, visit the Research & learn webpage anulib.anu.edu.au/research-learn and select the skill area followed by the relevant course. You can register for a workshop and find other information.

Research & learn how-to guides

Explore and learn with the ANU Library's how to guides (anulib.anu.edu.au/howto). Topics covered are:

- Citations & abstracts
- E-books
- EndNote
- Evaluating Sources
- Finding books and more
- Finding journal articles and more
- Finding theses
- Increasing your research impact
- ORCID iD (Open Researcher and Contributor ID)
- Research Data Management
- Text and Data Mining
- Topic analysis

Subject guides

Find subject-specific guides (anulib.anu.edu.au/subjectguides) and resources on broad range of disciplines. Such as:

- Asia Pacific, Southeast Asia and East Asian studies
- Business, economics, art, music and military studies
- Criminal, human rights and taxation law
- History, indigenous studies, linguistics and philosophy
- Biological, environment, physical & mathematical sciences, engineering & computer science, health & medicine

Navigating the sea of scholarly communication

An open access course designed to build the capabilities researchers need to navigate the scholarly communications and publishing world. Topics covered include finding a best-fit publisher, predatory publishing, data citations, bibliometrics, open access, and online research identity. Five self-paced modules, delivered by international and local experts/librarians (anulib.anu.edu.au/publishing).

Online learning

Online learning is available through ANU Pulse, which can be accessed from both on and off campus by all ANU staff and students (ql.anu.edu.au/pulse).

Modules available in ANU Pulse

- Microsoft Office (Access, Excel, OneNote, Outlook, PowerPoint, Project, Visio, Word)
- Microsoft Office (Mac)
- Adobe suite (Illustrator, Photoshop)
- Type IT

Training

A range of workshops are offered to help with your academic research and studies (anulib.anu.edu.au/training-register).

Feedback

Please provide feedback about webinars on the online feedback form (ql.anu.edu.au/libwebinar).