DEBREMARKOS UNIVERSITY

College of Social Sciences and Humanities

Department of Geography and Environmental Studies

SPSS manual for version 20

July 2016

Contents

1 Introduction SPSS windows

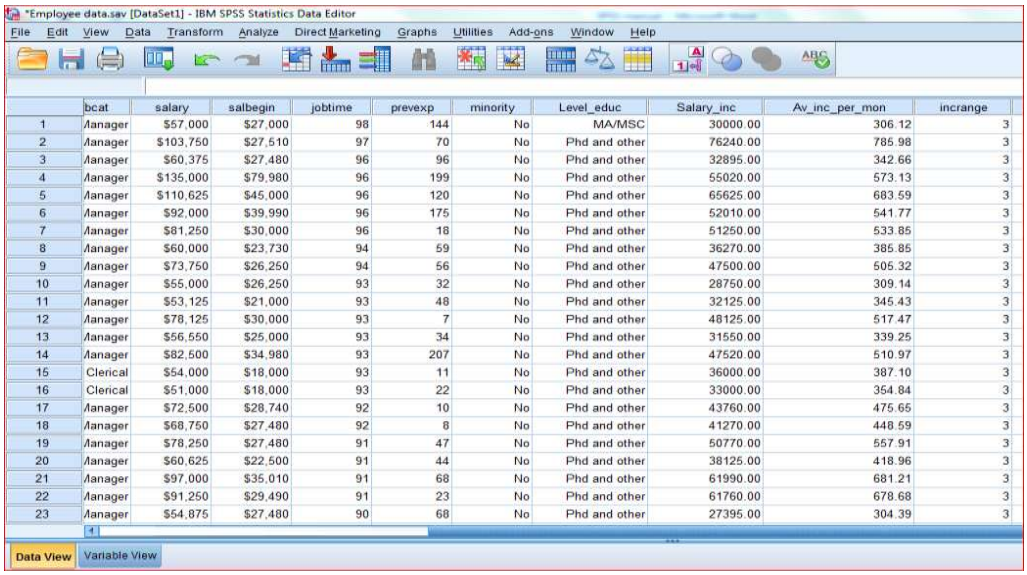
The "Statistical Package for the Social Sciences" (SPSS) is a package of programs for manipulating, analyzing, and presenting data; the package is widely used in the social and behavioral sciences.

SPSS has three basic windows***: the data editor, the syntax window and the output window***. The particular view can be changed by going to the Window menu. What you typically see first is the data window.

1. Data editor

The data editor window is where the data is either inputted or imported. The data editor window has two views the data view and the variable view. These two windows can be exchanged by clicking the buttons on the lower left corner of the data window. Information can be edited or deleted in both views.

- **In the data view-** your data is presented in a spreadsheet style very similar to Excel. The data is organized in rows and columns. Each row represents an observation and each column represents a variable. This view displays the actual data values or value labels.
- **In the variable view-** the logic behind each variable is stored. Each variable has a name (in the name column), a type (numeric, percentage, date, string etc.), a label (usually the full wording of the question), and the values assigned to the level of the variable in the "values" column. For example the name column we may have a variable called gender. In the label column we may specify that the variable is the "gender of the participants". In the values box, we may assign a"1" for males and a"2" for females.



	bcat	salary	salbegin	jobtime	prevexp	minority	Level_educ	Salary_inc	Av_inc_per_mon	incrange
1	Manager	$57,000	$27,000	98	144	No	MA/MSC	30000.00	306.12	3
2	Manager	$103,750	$27,510	97	70	No	Phd and other	76240.00	785.98	3
3	Manager	$60,375	$27,480	96	96	No	Phd and other	32895.00	342.66	3
4	Manager	$135,000	$79,980	96	199	No	Phd and other	55020.00	573.13	3
5	Manager	$110,625	$45,000	96	120	No	Phd and other	65625.00	683.59	3
6	Manager	$92,000	$39,990	96	175	No	Phd and other	52010.00	541.77	3
7	Manager	$81,250	$30,000	96	18	No	Phd and other	51250.00	533.85	3
8	Manager	$60,000	$23,730	94	59	No	Phd and other	36270.00	385.85	3
9	Manager	$73,750	$26,250	94	56	No	Phd and other	47500.00	505.32	3
10	Manager	$55,000	$26,250	93	32	No	Phd and other	28750.00	309.14	3
11	Manager	$53,125	$21,000	93	48	No	Phd and other	32125.00	345.43	3
12	Manager	$78,125	$30,000	93	7	No	Phd and other	48125.00	517.47	3
13	Manager	$56,550	$25,000	93	34	No	Phd and other	31550.00	339.25	3
14	Manager	$82,500	$34,980	93	207	No	Phd and other	47520.00	510.97	3
15	Clerical	$54,000	$18,000	93	11	No	Phd and other	36000.00	387.10	3
16	Clerical	$51,000	$18,000	93	22	No	Phd and other	33000.00	354.84	3
17	Manager	$72,500	$28,740	92	10	No	Phd and other	43760.00	475.65	3
18	Manager	$68,750	$27,480	92	8	No	Phd and other	41270.00	448.59	3
19	Manager	$78,250	$27,480	91	47	No	Phd and other	50770.00	557.91	3
20	Manager	$60,625	$22,500	91	44	No	Phd and other	38125.00	418.96	3
21	Manager	$97,000	$35,010	91	68	No	Phd and other	61990.00	681.21	3
22	Manager	$91,250	$29,490	91	23	No	Phd and other	61760.00	678.68	3
23	Manager	$54,875	$27,480	90	68	No	Phd and other	27395.00	304.39	3

There are 10 characteristics to be specified under the columns of the Variable View:

1. Name — the chosen variable name. This can be up to eight alphanumeric characters but must begin with a letter. While the underscore (_) is allowed, hyphens (-), ampersands (&), and spaces cannot be used. Variable names are not case sensitive.

2. Type — the type of data. SPSS provides a default variable type once variable values have been entered in a column of the Data View. The type can be changed by highlighting the respective entry in the second column of the Variable View and clicking the three-period symbol (…) appearing on the right hand side of the cell. This results in the Variable Type box being opened, which offers a number of types of data including various formats for numerical data, dates, or currencies. (Note that a common mistake made by first-time users is to enter categorical variables as type "string" by typing text into the Data View. To enable later analyses, categories should be given artificial number codes and defined to be of type "numeric.")

3. Width — the width of the actual data entries. The default width of numerical variable entries is eight. The width can be increased or decreased by highlighting the respective cell in the third column and employing the upward or downward arrows appearing on the right-hand side of the cell or by simply typing a new number in the cell.

4. Decimals — the number of digits to the right of the decimal place to be displayed for data entries. This is not relevant for string data and for such variables the entry under the fourth column is given as a grayed-out zero. The value can be altered in the same way as the value of Width.

5. Label — a label attached to the variable name. It is generally a good idea to assign variable labels. They are helpful for reminding users of the meaning of variables (placing the cursor over the variable name in the Data View will make the variable label appear) and can be displayed in the output from statistical analyses.

6. Values — labels attached to category codes. For categorical variables, an integer code should be assigned to each category and the variable defined to be of type "numeric." When this has been done, clicking on the respective cell under the sixth column of the Variable View makes the three-period symbol appear, and clicking this opens the Value Labels dialogue box, which in turn allows assignment of labels to category codes. For example, our data set included a categorical variable sex indicating the gender of the subject. Clicking the three-period symbol opens the then numerical code "0" may be assigned to represent females and code "1" males.

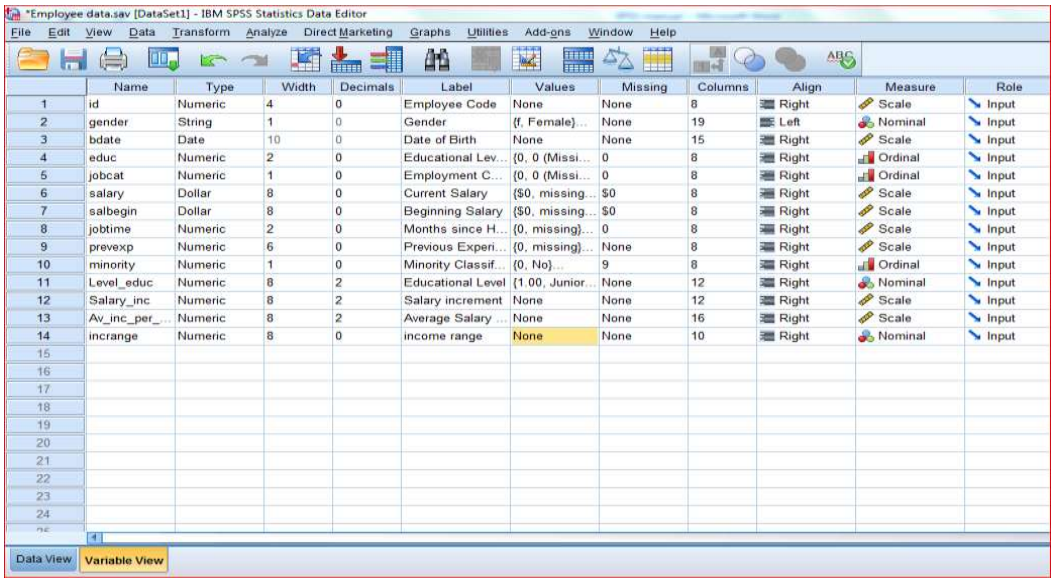
7. Missing — missing value codes. SPSS recognizes the period symbol as indicating a missing value. If other codes have been used (e.g., 99, 999) these have to be declared to represent missing values by highlighting the respective cell in the seventh column, clicking the three-periods symbol and filling in the resulting Missing Values dialogue box accordingly.

8. Columns — width of the variable column in the Data View. The default cell width for numerical variables is eight. Note that when the Width value is larger than the Columns value, only part of the data entry might be seen in the Data View. The cell width can be changed in the same way as the width of the data entries or simply by dragging the relevant column boundary. (Place cursor on right-hand boundary of the title of the column to be resized. When the cursor changes into a vertical line with a right and left arrow, drag the cursor to the right or left to increase or decrease the column width.)

9. Align — alignment of variable entries. The SPSS default is to align numerical variables to the right-hand side of a cell and string variables to the left. It is generally helpful to adhere to this default; but if necessary, alignment can be changed by highlighting the relevant cell in the ninth column and choosing an option from the drop-down list.

10. Measure — measurement scale of the variable. The default chosen by SPSS depends on the data type. For example, for variables of type "numeric," the default measurement scale is a continuous or interval scale (referred to by SPSS as "scale"). For variables of type "string," the default is a nominal scale. The third option, "ordinal," is for categorical variables with ordered categories but is not used by default. It is good practice to assign each variable the highest

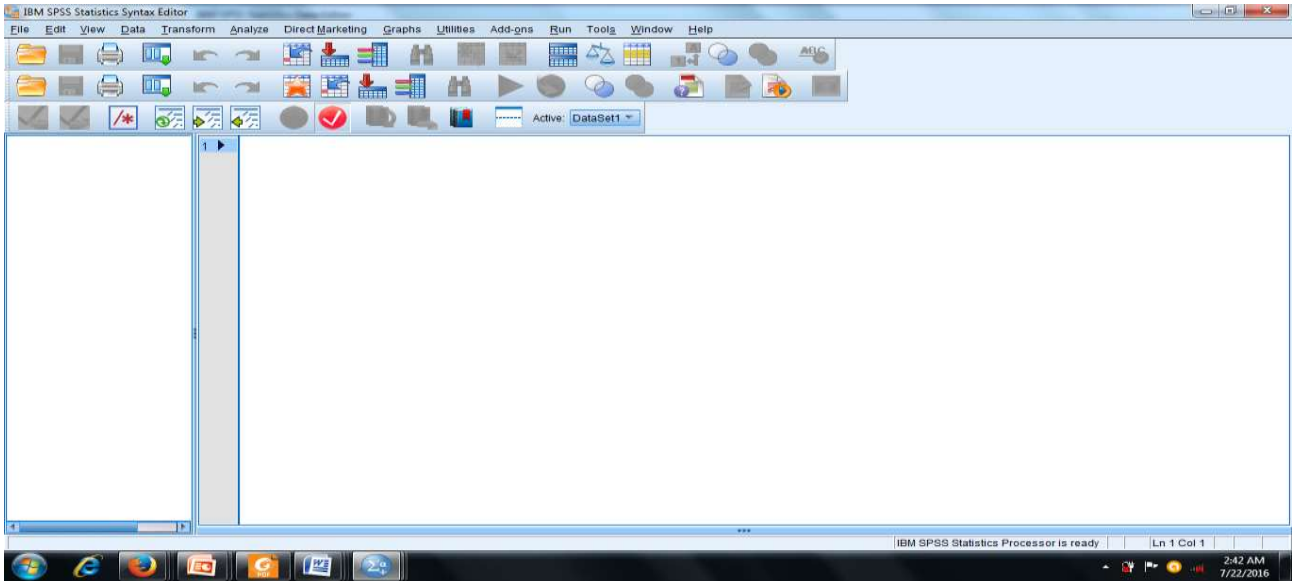
appropriate measurement scale ("scale" > "ordinal" > "nominal") since this has implications for the statistical methods that are applicable. The default setting can be changed by highlighting the respective cell in the tenth column and choosing an appropriate option from the drop-down list.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	4	0	Employee Code	None	None	8	Right	Scale	Input
2	gender	String	1	0	Gender	{f, Female}...	None	19	Left	Nominal	Input
3	bdate	Date	10	0	Date of Birth	None	None	15	Right	Scale	Input
4	educ	Numeric	2	0	Educational Lev...	{0, 0 (Missi...	0	8	Right	Ordinal	Input
5	jobcat	Numeric	1	0	Employment C...	{0, 0 (Missi...	0	8	Right	Ordinal	Input
6	salary	Dollar	8	0	Current Salary	{$0, missing...	$0	8	Right	Scale	Input
7	salbegin	Dollar	8	0	Beginning Salary	{$0, missing...	$0	8	Right	Scale	Input
8	jobtime	Numeric	2	0	Months since H...	{0, missing}...	0	8	Right	Scale	Input
9	prevexp	Numeric	6	0	Previous Experi...	{0, missing}...	None	8	Right	Scale	Input
10	minority	Numeric	1	0	Minority Classif...	{0, No}...	9	8	Right	Ordinal	Input
11	Level_educ	Numeric	8	2	Educational Level	{1.00, Junior...	None	12	Right	Nominal	Input
12	Salary_inc	Numeric	8	2	Salary increment	None	None	12	Right	Scale	Input
13	Av_inc_per_...	Numeric	8	2	Average Salary ...	None	None	16	Right	Scale	Input
14	incrange	Numeric	8	0	income range	None	None	10	Right	Nominal	Input
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											

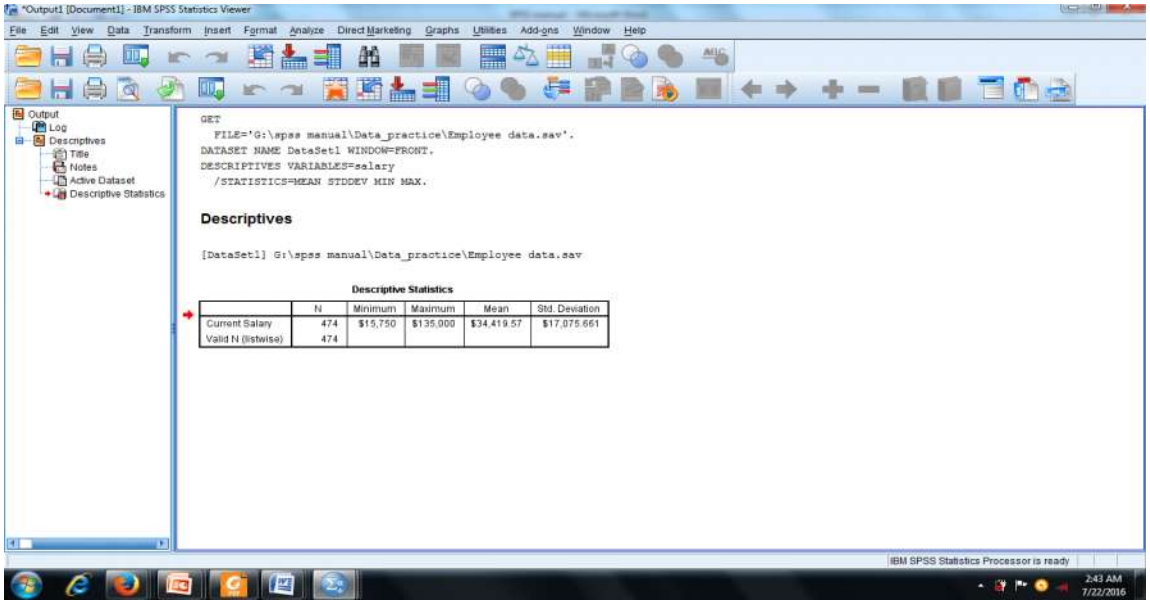
2. The Syntax Window

In the syntax editor window you can program SPSS. Although it is not necessary to program syntax for virtually all analyses, using the syntax editor is quite useful for two purposes: one to save your analysis steps. Second to run repetitive tasks. Firstly, you can document your analysis steps and save them in a syntax file, so others may re[run your tests and you can re[run them as well. To do this you simply hit the PASTE button you find in most dialog boxes. Secondly, if you have to repeat a lot of steps in your analysis, for example, calculating variables or re [coding, it is most often easier to specify these things in syntax, which saves you the time and hassle of scrolling and clicking through endless lists! of variables.



3. The Output Window

The output window is where SPSS presents the results of the analyses you conducted. Besides the usual status messages, you'll find all of the results, tables, and graphs in here. In the output window you can also manipulate tables and graphs and reformat them.



2 Data Entry

When SPSS of any version for Windows is first opened; a default dialogue box appears that gives the user a number of options. Most likely users will want to enter data or open an existing data file. This dialogue box can be prevented from opening in the future by checking this option

at the bottom of the box. When Type in data is selected, the SPSS Data Editor appears as an empty spreadsheet. At the top of the screen is a menu bar and at the bottom a status bar. The status bar informs the user about facilities currently active; at the beginning of a session it simply reads, "SPSS Processor is ready."

3 Creating primary reference lists

There is one set of outputs you'll create that is more important than anything else, and that is the set of primary references. Primary references describe your overall data set. In other words, how many in all? How many in each category? What are the maximums and minimums? Means? Standard deviations? Here's our rule: list out the summaries and put them somewhere where you refer to them quickly. You may not always want to print out all the details of your data set. For example, printing out every single income for a data set of one million people, would not be useful, economical, or nice to either your printer or the trees. So here are the basic rules: print frequencies for categorical variables and descriptive (also called univariate) statistics for continuous variables. In this exercise, we'll use the sample Employee dat.sav file.

3.1 Frequencies

1. If it's not already open, open the Employee dat.sav file by selecting File > Open and navigating to D\SPSS Tutorial Data\Employee data.sav.

2. From the menu, select File > New > Draft Output.

3. From the menu, select Analyze > Descriptive Statistics > Frequencies.

4. Double-click Gender, Employment Category, and Minority Classification to move them to the Variables list.

5. Click the check box labeled **Display frequency tables.**

6. Click **Statistics**.

7. Make sure all the check boxes are cleared (not checked).

8. Click **Continue.**

9. Click **Charts.**

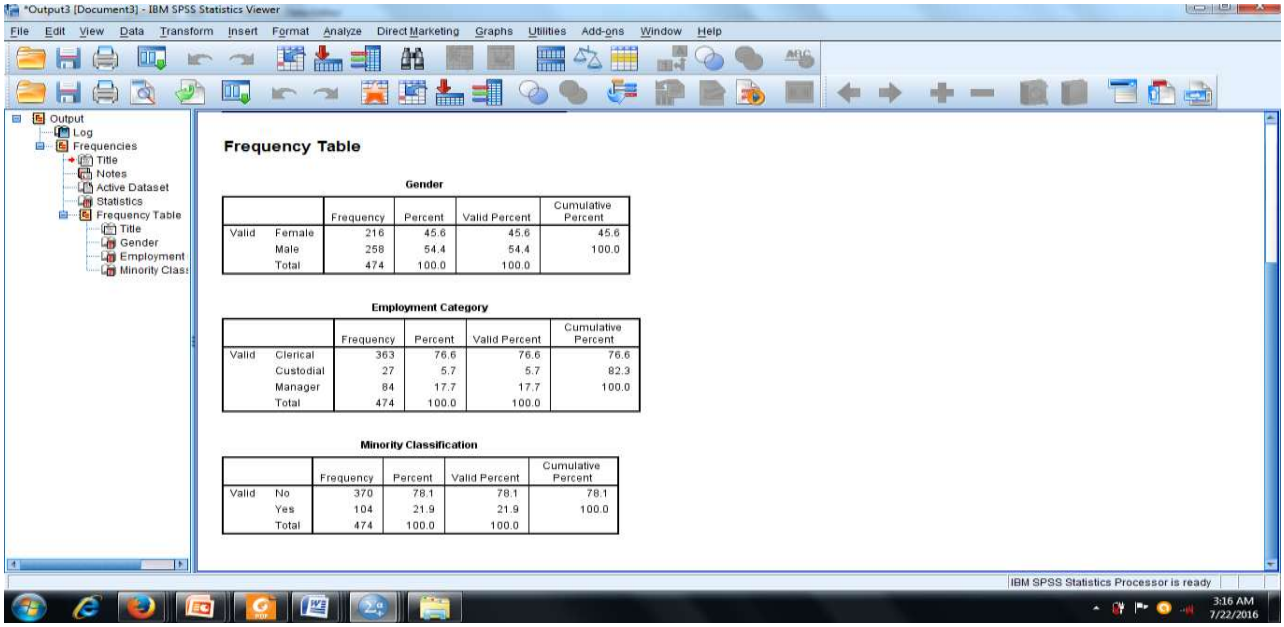
10. If it is not already selected, select **None** by clicking it.

11. Click **Continue.**

12. Click **OK.**

13. From the menu, select File > Save As.

14. Navigate to D:\SPSSTutorialData\ and save the file as All Freqs.



3.2 Descriptive statistics: descriptive (univariate)

The next step is to print the descriptive or univariate statistics for the continuous variables.

1. From the menu, select File > New > Draft Output.

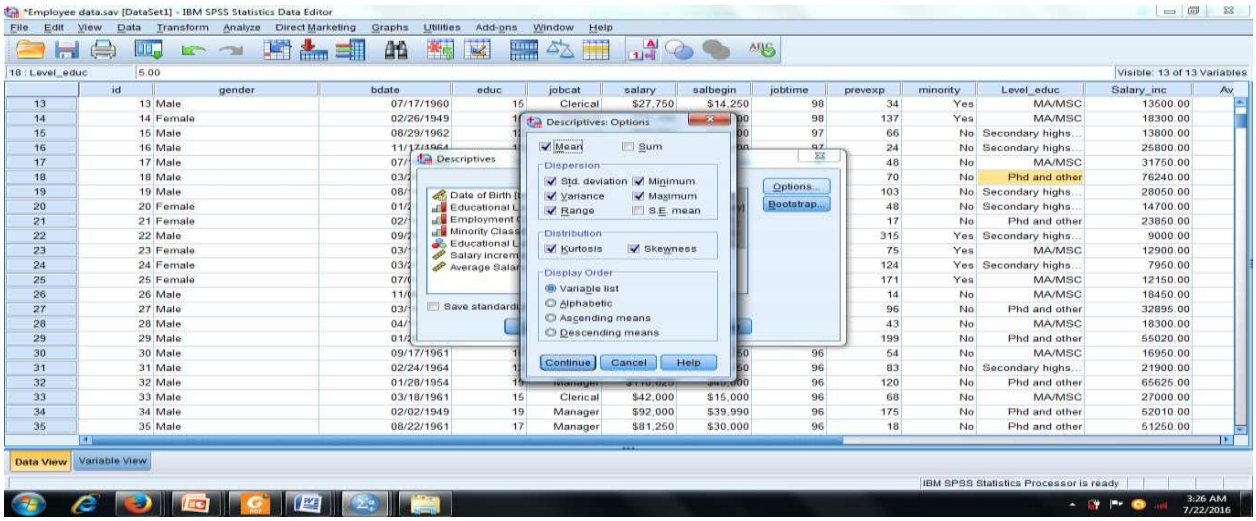
2. From the menu, select Analyze > Descriptive Statistics > Descriptives.

3. Click **Reset** to clear any previous selections.

4. Double-click Current Salary, Beginning Salary, Months since Hire, and Previous Experience to move them to the Variables list.

5. Click **Options.**

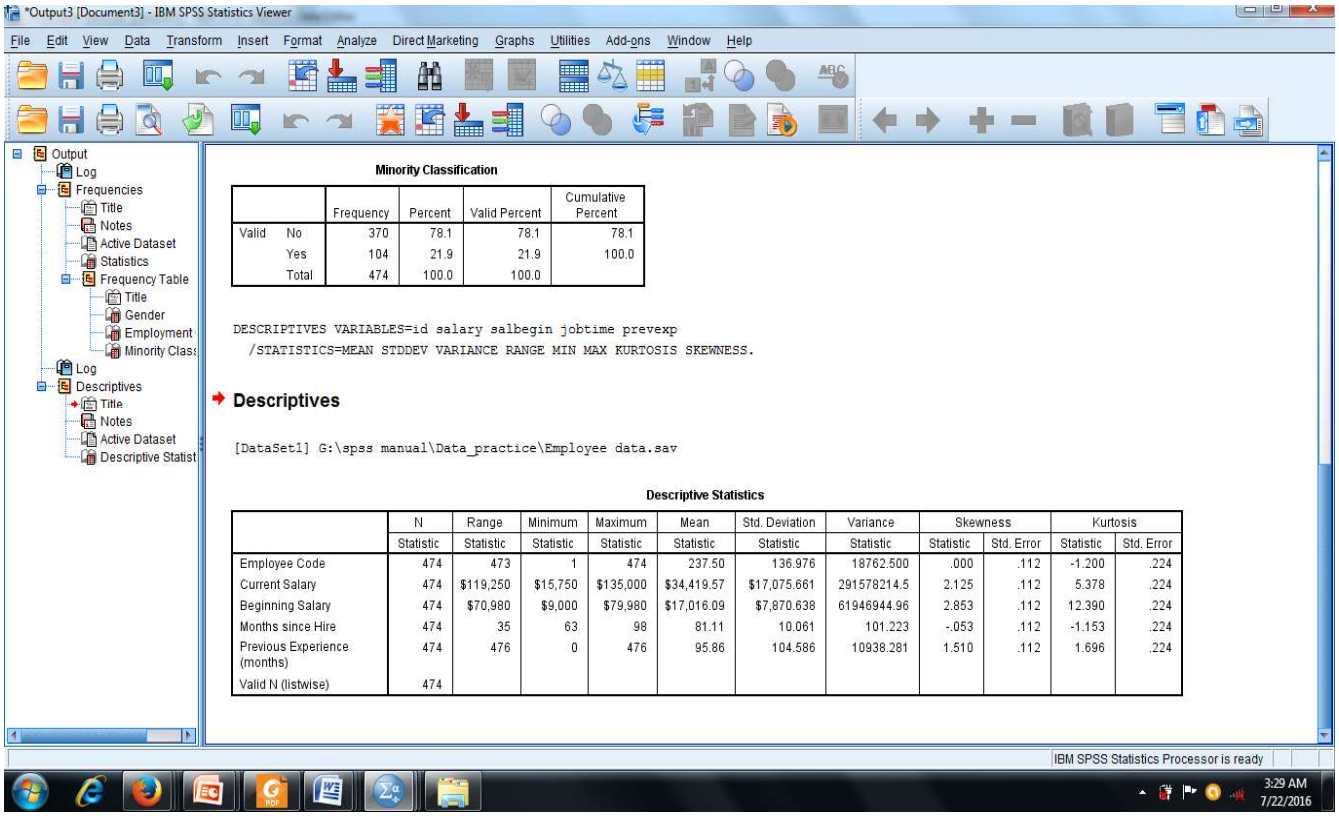
6. In the Descriptives: Options window, click Mean, Std. deviation, Variance, Range, Minimum, Maximum, Kurtosis, and Skewness



7. Click **Continue.**

8. Click **OK.**

9. When the resulting table is displayed, notice that the variables you selected are listed as rows, while the statistics are listed in columns.



10. From the menu, select File > Save As.

11. Navigate to D:\SPSSTutorialData\ and save the file as All Descriptives.

12. Notice that the statistic **Range** displays the distance between the minimum and maximum.

4 Recodes and Transformations

Sooner or later, no matter how carefully you planned your data design, you'll probably want to work with some variables in different forms. If you collected income or age data, for example, you might want to group the continuous variables into categories. Or you might want to create a variable that combines various conditions, say, all minority managers by gender. This type of data manipulation is called transforming or recoding. In this exercise, you'll create several new variables, some that indicate multiple conditions and some that recode continuous variables into categorical variables.

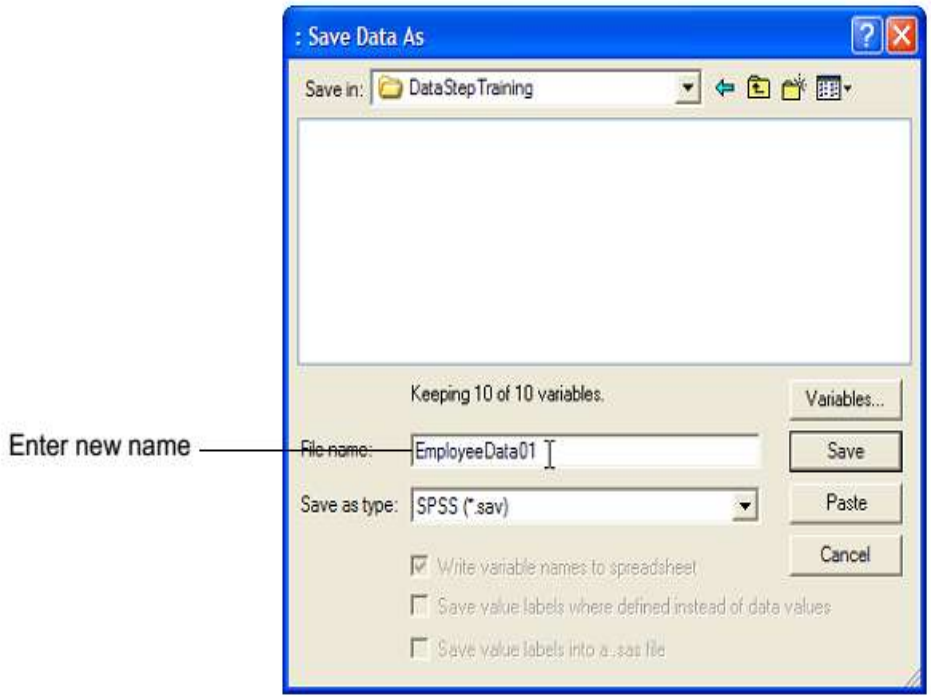
Backup the original file

The first step before making ***any*** changes to your data file is: BACK UP YOUR DATA. And the easiest way to back up your data is to save it under another name.

1. If you don't have the data view open, select it from the menu by selecting Window > Employee data.sav - SPSS Data Editor.

2. From the menu, select File > Save As, and then navigate to D:\SPSSTutorialData\.

3. Name the new file EmployeeData01



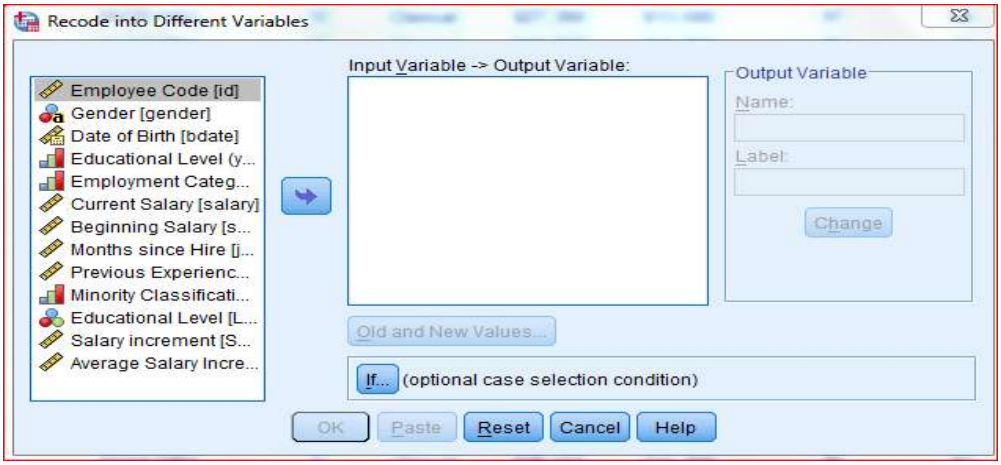
4. Click **Save.** Notice that the title bar of your window now identifies the file as EmployeeData01.sav. Now you can begin your transformations.

4.1 Recoding existing variables

Recoding refers to assigning codes (or different codes) to an existing variable. **NEVER** ever, ever, EVER recode your variables into the same variable name (with one exception). That way lays madness. And chaos. For one thing it deletes your existing data. And for another it destroys the history of the data. Always create a new variable to contain the new codes.

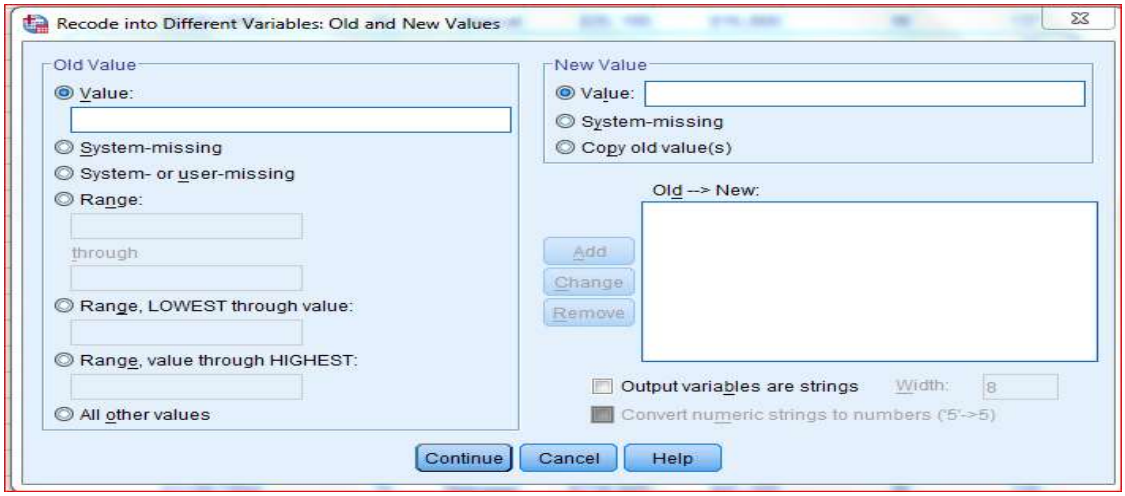
Recode income data

1. From the menu, select Transform > Recode > Into Different Variables to open the Recode window

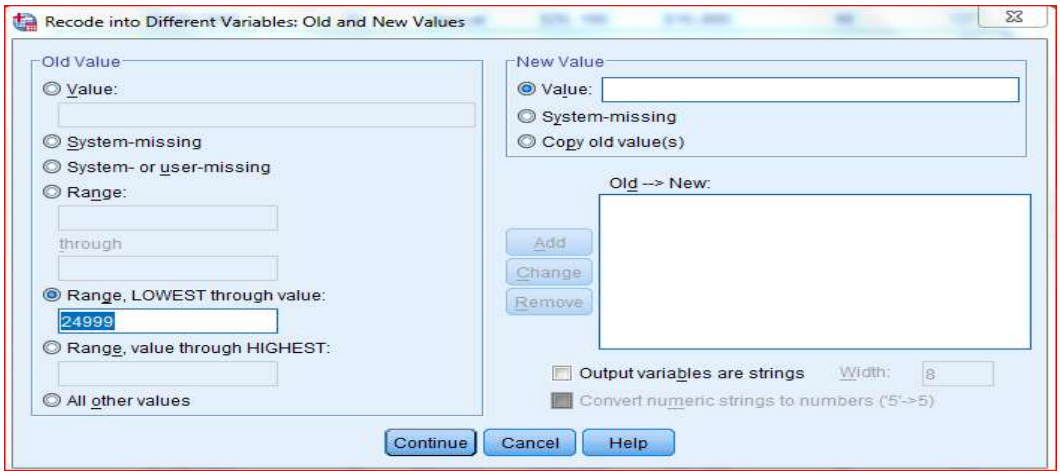


2. Double-click Current Salary to move it to the Input Variable --> Output Variables pane.

3. Click **Old and New Values** to open window where you'll create the new codes



4. Click the second **Range** radio button (lowest through) to activate its field

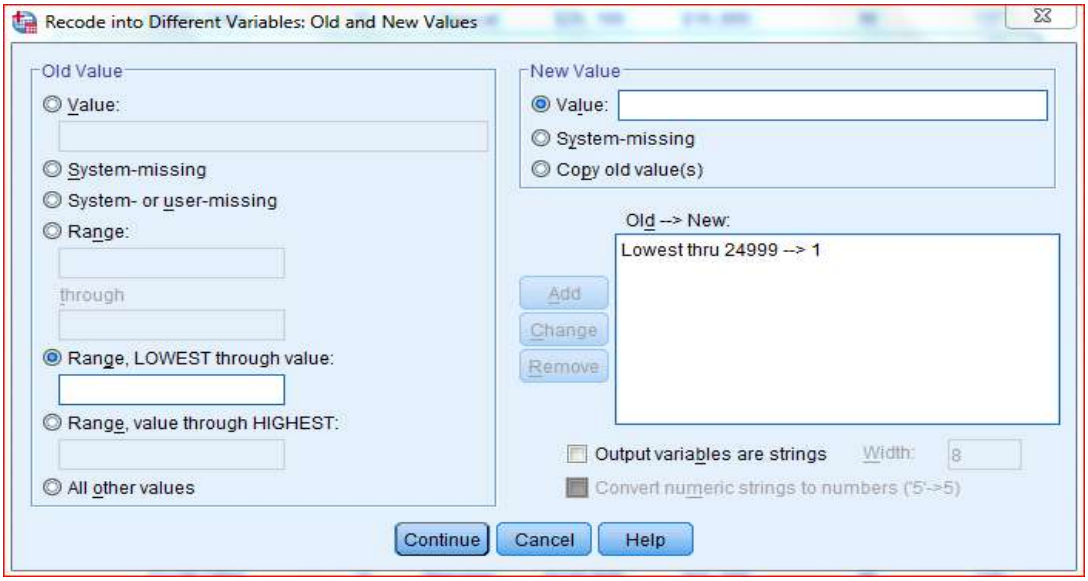


5. Click in the Lowest through range and type: **24999**

6. Click in the Value field under New Value and type: **1**

In the new field, any incomes less than 25000 will have a value of 1

7. Click **Add.** Notice that the complete definition of old and new values appears in the Old --> New pane (Figure 16). In the next steps you'll define three more income ranges.



8. Under Old Value, click the first Range radio button to active the minimum and maximum values

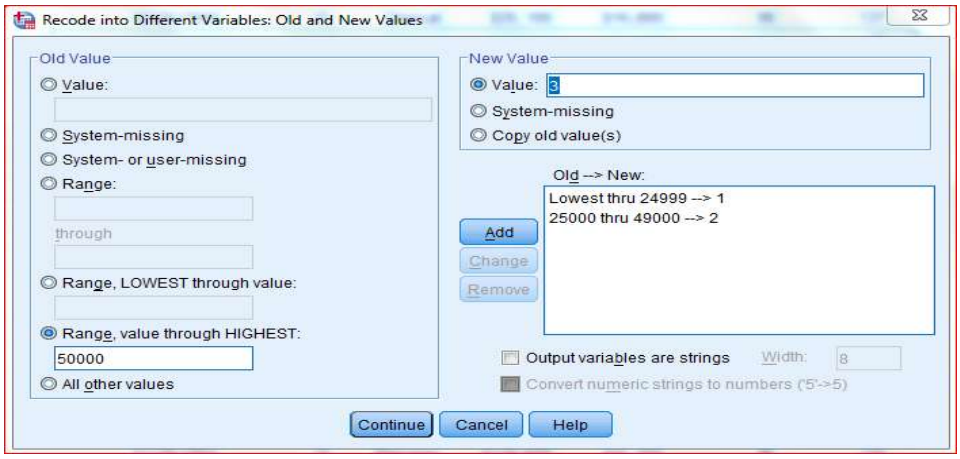
9. In the first range field, type: **25000**

10. in the second range field, type: **49999**

11. Under New Value, type: **2**

12. Click **Add.** Notice that the new definition is added to the Old --> New pane.

13. Under Old Value, click the **third** range radio button to activate the Range ... through highest field



14. In the field to the left of *through Highest* type: **50000**

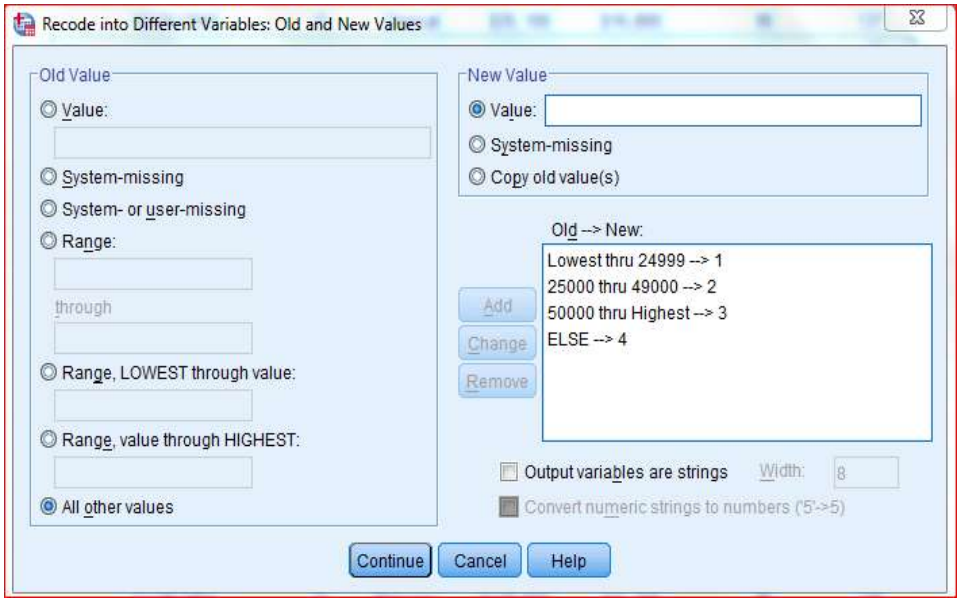
15. Under New Value, type: **3**

16. Click **Add.** Again, the new definition is added to the Old --> New pane. In the next steps, you'll create a value to accommodate any odd values that might have been entered into the file. Even if you're sure there aren't any, check anyhow.

17. Under Old Value, click the radio button for **All other values.** Notice that there is no range to enter.

18. Under New Value, type: **4**

19. Click **Add.** Now all possible definitions are entered under the Old --> New pane



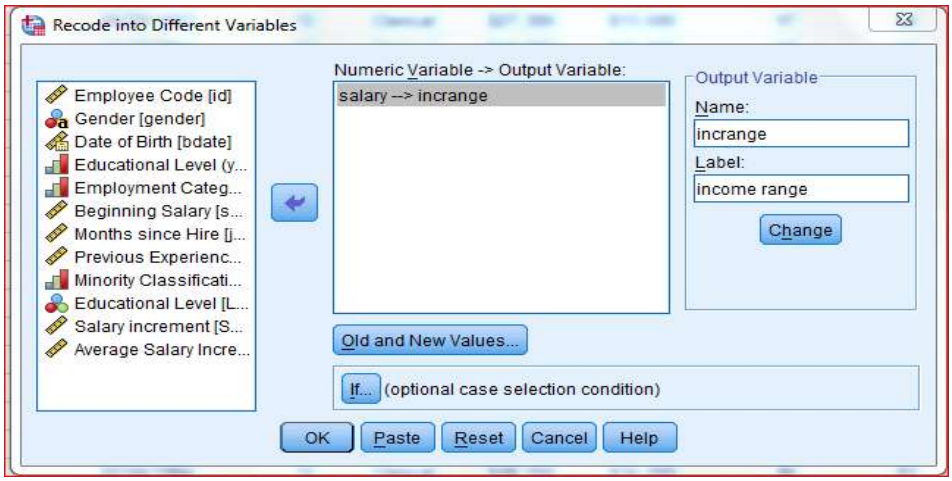
20. Click **Continue.** The Old and New Values window closes and the original Recode into Different Variables window is displayed.

21. In the field for Output Variable Name type: **incrange**

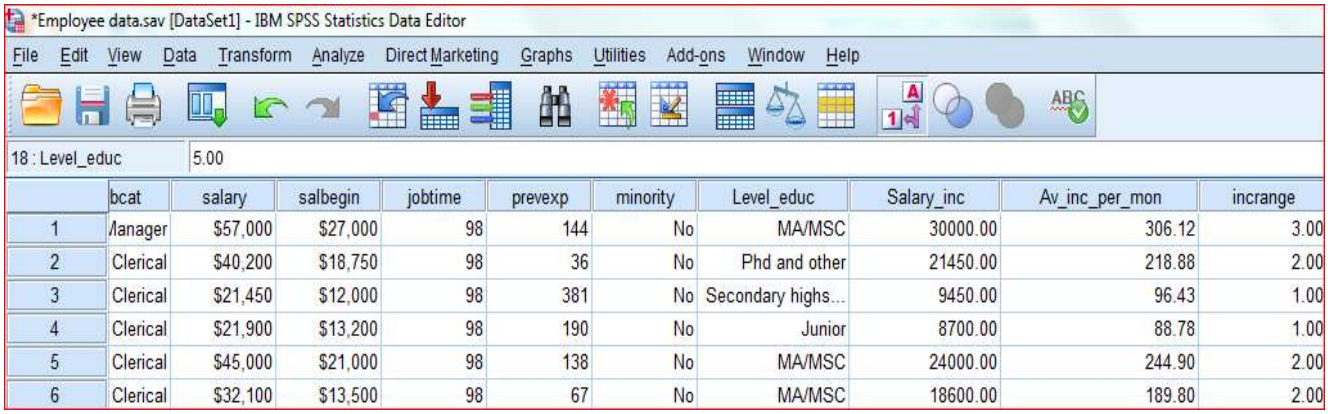
22. In the field for Output Variable Label type: **Income Range**

The label is the name that will be displayed on all output, so you'll want to make sure it's informative and correctly formatted.

23. Click **Change.** The new name is now listed in the Numeric Variable --> Output Variable pane



24. Click **OK.** The Recode window closes and the data view is displayed. Notice that there is now a new column on the right containing the new range codes and the following will be displayed.

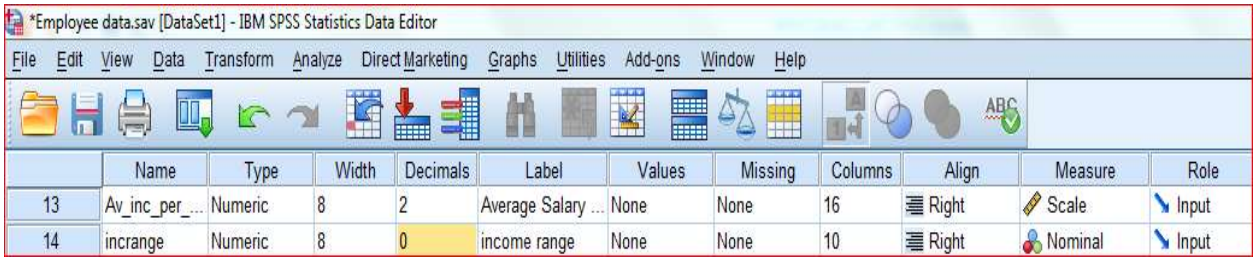


	bcat	salary	salbegin	jobtime	prevexp	minority	Level_educ	Salary_inc	Av_inc_per_mon	incrange
1	Manager	$57,000	$27,000	98	144	No	MA/MSC	30000.00	306.12	3.00
2	Clerical	$40,200	$18,750	98	36	No	Phd and other	21450.00	218.88	2.00
3	Clerical	$21,450	$12,000	98	381	No	Secondary highs...	9450.00	96.43	1.00
4	Clerical	$21,900	$13,200	98	190	No	Junior	8700.00	88.78	1.00
5	Clerical	$45,000	$21,000	98	138	No	MA/MSC	24000.00	244.90	2.00
6	Clerical	$32,100	$13,500	98	67	No	MA/MSC	18600.00	189.80	2.00

25. Noticed that the codes are displayed with two decimal places. These should be simple integer codes, so in the next step you'll change the format of the variable.

26. Double-click the name **incrange** to open the Variable View with that variable selected

27. Double-click the **2** in the Decimals field and type: **0**



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
13	Av_inc_per_...	Numeric	8	2	Average Salary ...	None	None	16	Right	Scale	Input
14	incrange	Numeric	8	0	income range	None	None	10	Right	Nominal	Input

28. Press **tab** or **Enter** to leave the field.

29. Click the Data View tab to see the corrected data. In the next step, you'll sort the cases in ascending and descending order of **incrange** to see how the values were applied and to see if there are any values that were not included.

30. Click anywhere in the incrange column.

31. From the menu select Date > Sort Cases to open the sort window

32. Scroll down to Income Range and double-click it to move it to the Sort by pane.

33. Click **OK.** The cases are now sorted according to the lowest income range value.

34. From the menu select Data > Sort cases. Notice that your previous choices are still selected.

35. Click once on Income Range in the Sort By pane.

36. Click **Descending** under Sort Order.

Now let's put the new variable to work and display the distribution of cases by income range.

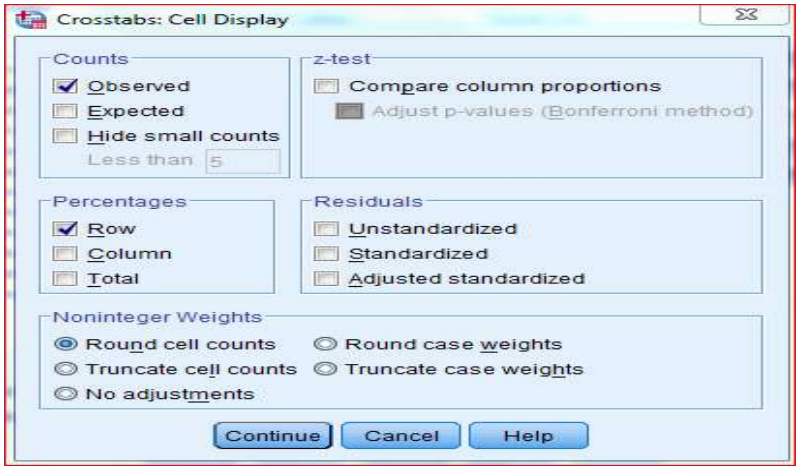
1. From the menu, select New > Draft Output.

2. From the menu, select Analyze > Descriptive Statistics > Crosstabs.

3. In the Crosstabs window, click once on Gender, then click the right arrow to move it into the Rows pane.

4. Click once on Income Range, then click the right arrow to move it into the Columns pane.

5. Click **Cells** to open the Crosstabs: Cell Display window



6. In the Percentages pane, click the check box for **Row**. In this case, row percentages will display the percent within gender in each income range. For counts, make sure Observe is checked.

7. Click **Continue.**

8. Click **OK.** The Crosstabs window will close and the new crosstabs will be displayed in the draft output window as follow.

Crosstabs

[DataSet1] G:\spss manual\Data_practice\Employee data.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * income range	474	100.0%	0	0.0%	474	100.0%

Gender * income range Crosstabulation

			income range			
			1	2	3	Total
Gender	Female	Count	124	86	6	216
		% within Gender	57.4%	39.8%	2.8%	100.0%
	Male	Count	19	173	66	258
		% within Gender	7.4%	67.1%	25.6%	100.0%
Total		Count	143	259	72	474
		% within Gender	30.2%	54.6%	15.2%	100.0%

Notice that the income ranges are listed as 1, 2, and 3. Not very informative. So let's go back and assign value labels for the new variable.

1.. Close the draft output window without saving.

2. In the Data View, double-click the column heading for **incrange** to open the Variable View for that variable.

3. In the Values column, click the build button to open the Value Labels dialog box.

4. In the Value field, type: **1**

5. In the Value Label field, type: **< $25,000**

6. Click **Add.**

7. In the Value field, type: **2**

8. In the Value Label field, type: **$25,000 - $49,999**

9. Click **Add.**

10. In the Value field, type: **3**

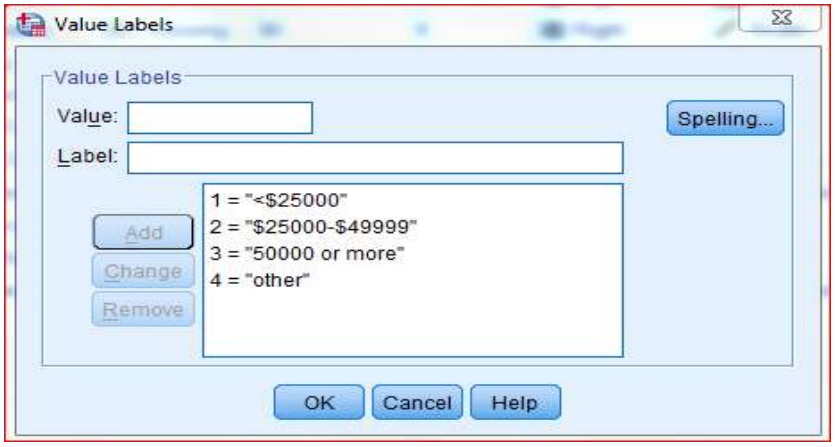
11. In the Value Label field, type: **$50,000 or more**

12. Click **Add.**

13. In the Value field, type: **4**

14. In the Value Label field, type: **Other**

15. Click **Add.**

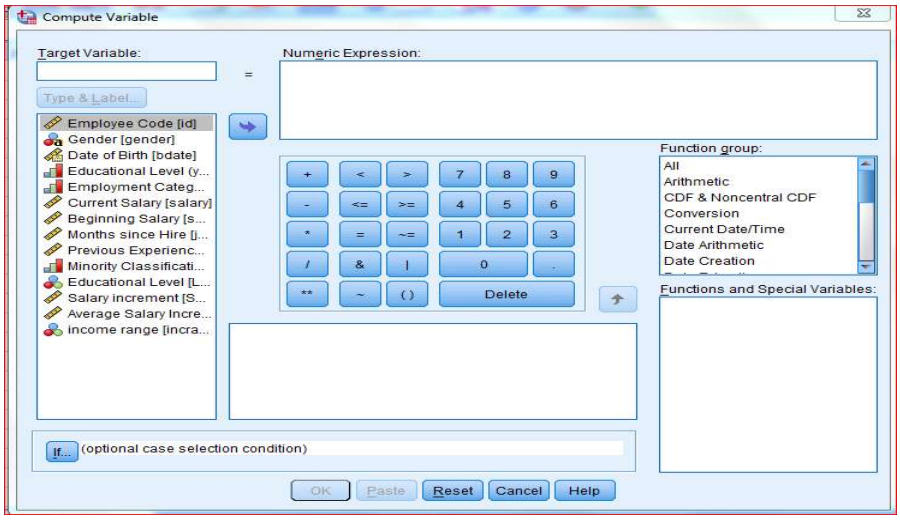


5 Computing new variables

5.1 Performing calculations with a variable and a function

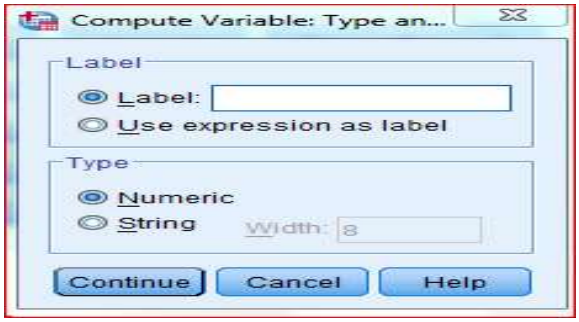
In some cases, you might want to calculate new variables based on values in existing variables and some arithmetic function like multiplying or dividing. For example, if you have a variable that contains an annual salary, you might want to calculate a monthly salary. To create the new variable, you use the **Compute** function.

1. In the Data window, select from the menu Transform > Compute variable and you will see the following window



2. In the Target Variable field, type: **salmonth**

3. Click **Type & Label.**

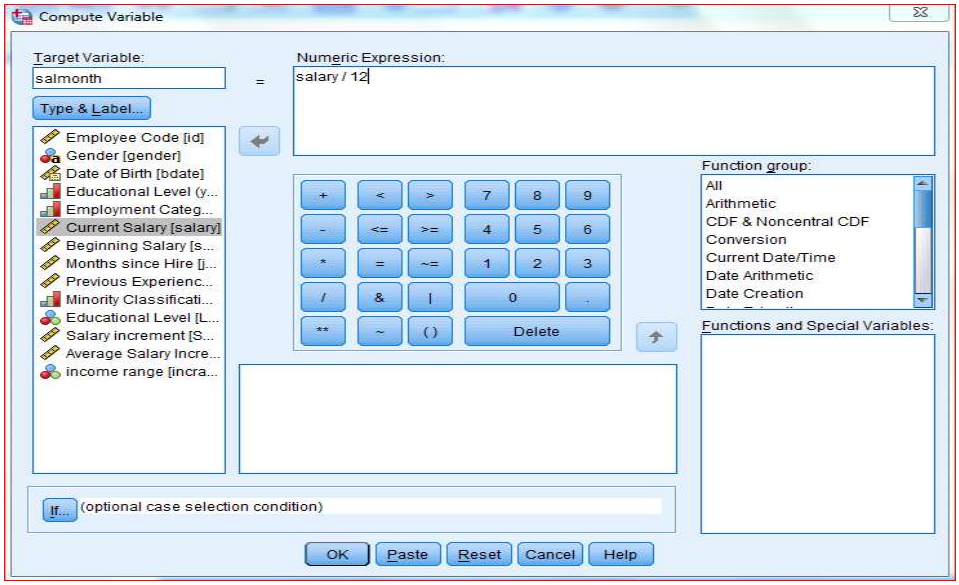


4. In the Label field, type: **Average monthly salary**

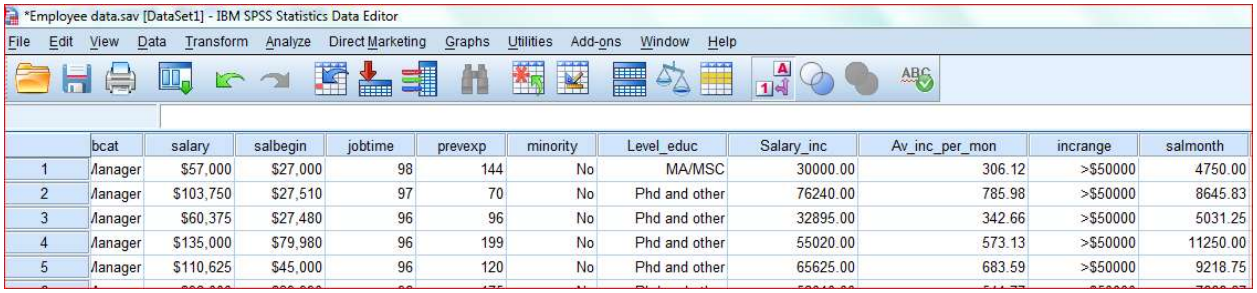
5. Click **Continue.**

6. In the Compute Variable window, select Current Salary and move it to the Numeric Expression pane by clicking the right arrow.

7. In the Numeric Expression pane, click the cursor after **salary** and type: **12**



8. Click **OK.** The Compute Variable window closes and the new variable is displayed in the Data window as the following window. You can now use the new variable in procedures such as crosstabs or in further calculations. For example, you could create a new variable for monthly withholding that calculates withholding as a percentage of monthly salary. You could then subtract the new withholding variable from the monthly salary to create still another variable for monthly net.



	bcat	salary	salbegin	jobtime	prevexp	minority	Level_educ	Salary_inc	Av_inc_per_mon	incrange	salmonth
1	Manager	$57,000	$27,000	98	144	No	MA/MSC	30000.00	306.12	>$50000	4750.00
2	Manager	$103,750	$27,510	97	70	No	Phd and other	76240.00	785.98	>$50000	8645.83
3	Manager	$60,375	$27,480	96	96	No	Phd and other	32895.00	342.66	>$50000	5031.25
4	Manager	$135,000	$79,980	96	199	No	Phd and other	55020.00	573.13	>$50000	11250.00
5	Manager	$110,625	$45,000	96	120	No	Phd and other	65625.00	683.59	>$50000	9218.75

5.2 Creating expressions with more than one variable

Let's use the previous example of calculating withholding and net to compute variables based on more than one variable. First you'll compute the withholding variable, then you'll compute the net variable.

1. From the menu, select Transform > Compute.

2. In the Target Variable field, type: **withhold**

3. Click **Type & Label.**

4. In the Label field, type: **Monthly withholding**

5. Click **Continue.**

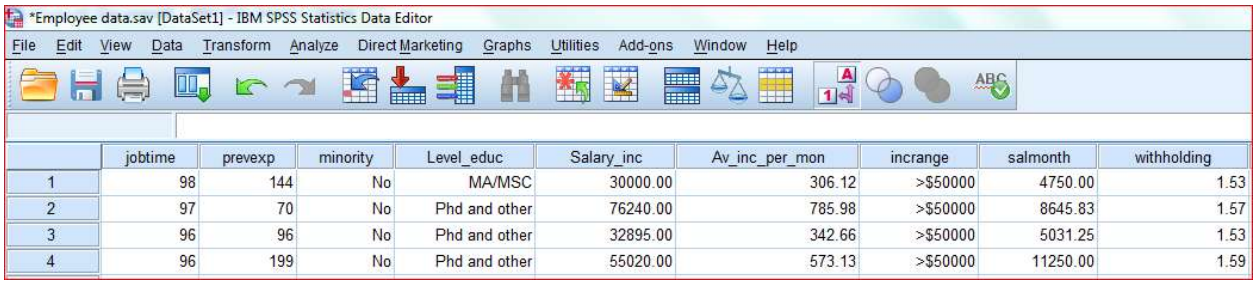
6. Select all the text in the Numeric Expression field and delete it.

7. Move the new variable, Average Monthly Salary, to the Numeric Expression field.

8. Click after **salmonth** and type: *** .05**

Note: If you haven't worked with computer programs before to make calculations, the asterisk denotes multiplication. A double asterisk (**) denotes exponentiation. In SPSS, a vertical bar (|) denotes "OR", and the ampersand (&) denotes "AND".

9. Click **OK.** The new variable appears in the data view as in the following table. In the next step, you'll use two variables to calculate a third.



	jobtime	prevexp	minority	Level_educ	Salary_inc	Av_inc_per_mon	incrange	salmonth	withholding
1	98	144	No	MA/MSC	30000.00	306.12	>$50000	4750.00	1.53
2	97	70	No	Phd and other	76240.00	785.98	>$50000	8645.83	1.57
3	96	96	No	Phd and other	32895.00	342.66	>$50000	5031.25	1.53
4	96	199	No	Phd and other	55020.00	573.13	>$50000	11250.00	1.59

10. From the menu, select Transform > Compute.

11. In the Target Variable field, type: **netmonth**

12. Click **Type & Label.**

13. In the Label field, type: **Monthly net**

14. Click **Continue.**

15. Select all the text in the Numeric Expression field and delete it or reset it.

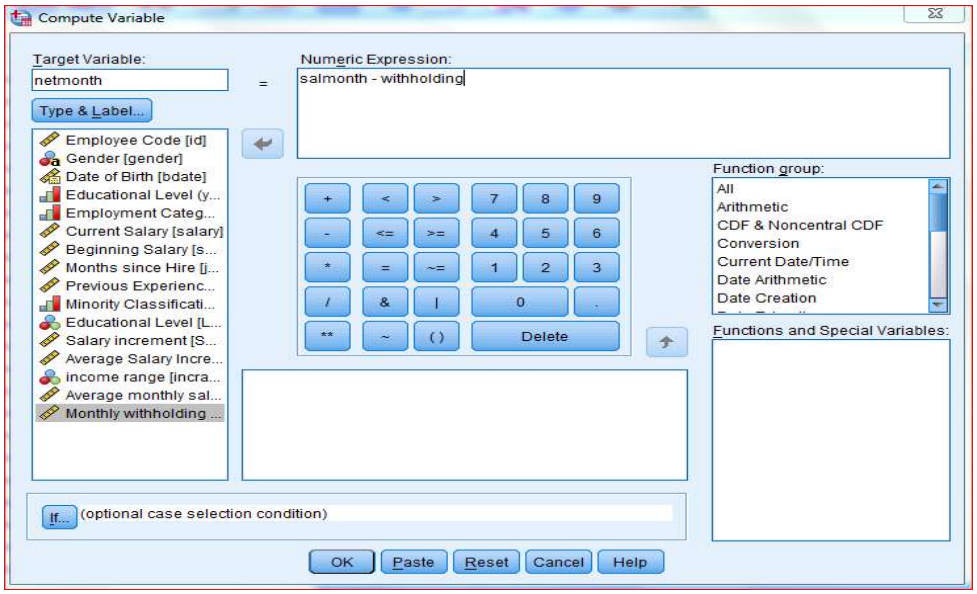
16. From the list of variables, select Average Monthly Salary and move it to the Numeric Expression field.

17. Click after salmonth in the Numeric Expression field.

18. Using the keypad in the Compute Variable window, click **"-".**

19. From the list of variables, select the new variable Monthly Withholding.

20. Click the right arrow to move it to the Numeric Expression pane. Your Compute Variable window should now look like as in the following window.



21. Click **OK.** The new variable appears in the data view see it

5.3 Conditional expressions

In some cases, you might want to look at only a specific subset of your data. Say you want to send a monthly newsletter to only female clerical staff. To identify these staff, you'll calculate a new binary variable (one that has only two values) using the IF statement to set the condition.

1. From the menu, select Transform > Compute.

2. In the Target Variable field type: **femclerk**

3. Click **Type & Label.**

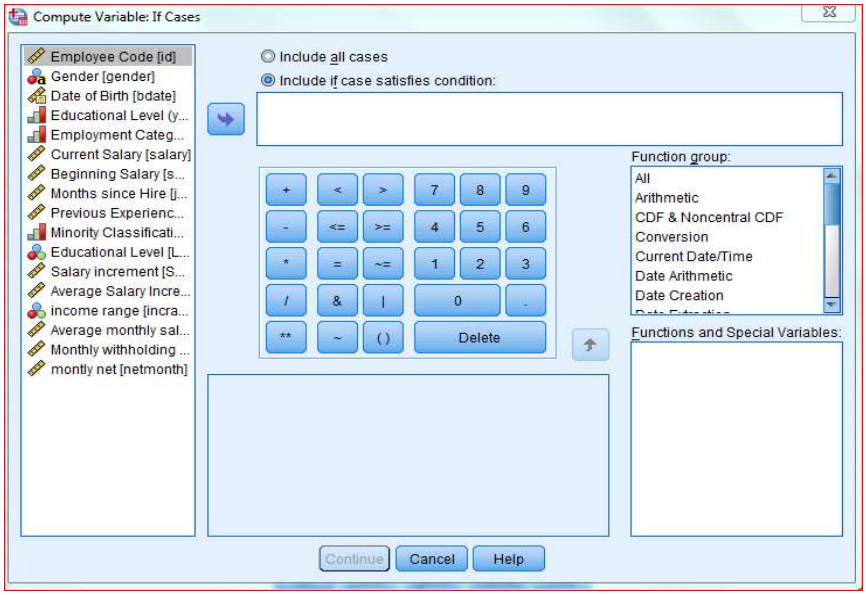
4. In the Label field type: **Female Clerical**

5. Click **Continue.**

6. Select all the text in the Numeric Expression field and delete it.

7. In the Numeric Expression field type: **1**

8. Click **If** to open the Compute Variable: If Cases window and you will see the following window



9. Select **Include if case satisfies condition.**

10. Double-click **Gender** to move it to conditions field.

11. Click after Gender in the conditions field and type: **= "f"**

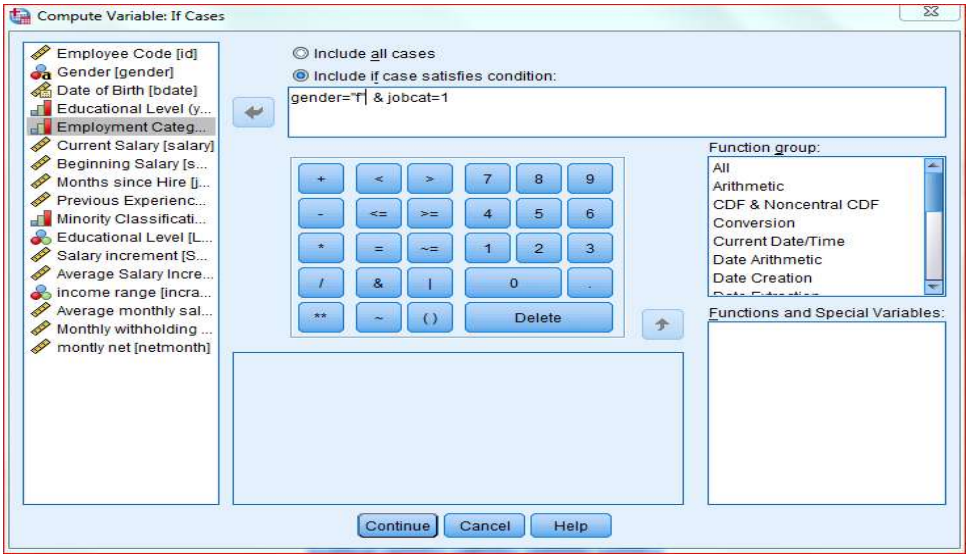
Note: Whenever you create a condition, you must use the actual values in the variable, not their labels. Thus, setting a condition to **gender = "Female"** would not select any cases.

12. Click after "f" and type a space.

13. Using the keypad in the Compute Variable window, click **&.** You use the ampersand to add a second condition.

14. From the field list, double-click **Employment category** to move it to the calculation pane.

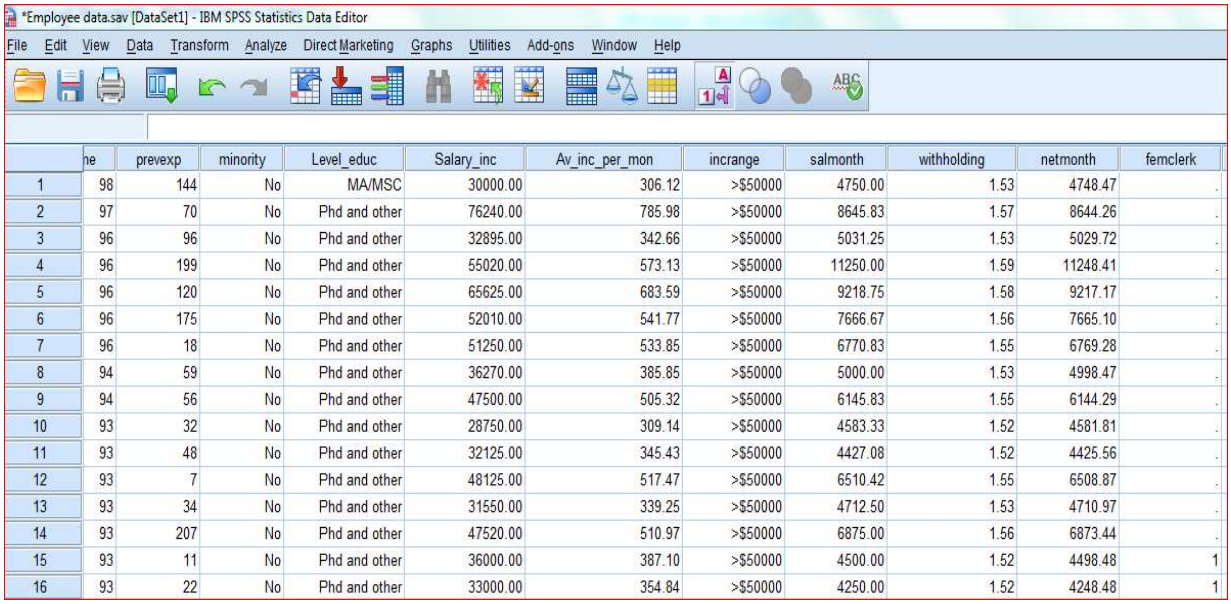
15. In the calculation pane, type: **= 1**



Note that you don't use quotation marks this time because is a numeric variable.

16. Click **Continue.**

17. Click **OK.** The new variable appears in the Data window. Scroll through the records to see how the values in the new variable. Notice that cases where gender is not female and job category is not manager have only a period, indicating a missing value. Only those cases where gender is female **and** jobcat is manager contain a 1 in the new variable like you see in the following window.

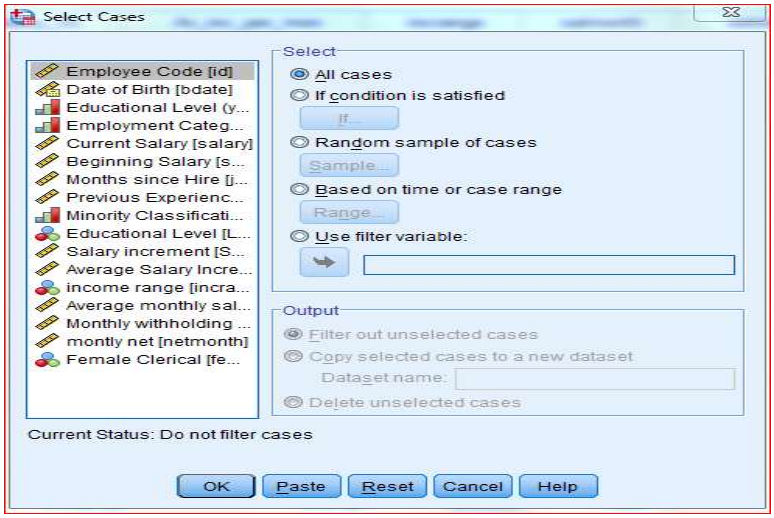


	he	prevexp	minority	Level_educ	Salary_inc	Av_inc_per_mon	incrange	salmonth	withholding	netmonth	femclerk
1	98	144	No	MA/MSC	30000.00	306.12	>$50000	4750.00	1.53	4748.47	.
2	97	70	No	Phd and other	76240.00	785.98	>$50000	8645.83	1.57	8644.26	.
3	96	96	No	Phd and other	32895.00	342.66	>$50000	5031.25	1.53	5029.72	.
4	96	199	No	Phd and other	55020.00	573.13	>$50000	11250.00	1.59	11248.41	.
5	96	120	No	Phd and other	65625.00	683.59	>$50000	9218.75	1.58	9217.17	.
6	96	175	No	Phd and other	52010.00	541.77	>$50000	7666.67	1.56	7665.10	.
7	96	18	No	Phd and other	51250.00	533.85	>$50000	6770.83	1.55	6769.28	.
8	94	59	No	Phd and other	36270.00	385.85	>$50000	5000.00	1.53	4998.47	.
9	94	56	No	Phd and other	47500.00	505.32	>$50000	6145.83	1.55	6144.29	.
10	93	32	No	Phd and other	28750.00	309.14	>$50000	4583.33	1.52	4581.81	.
11	93	48	No	Phd and other	32125.00	345.43	>$50000	4427.08	1.52	4425.56	.
12	93	7	No	Phd and other	48125.00	517.47	>$50000	6510.42	1.55	6508.87	.
13	93	34	No	Phd and other	31550.00	339.25	>$50000	4712.50	1.53	4710.97	.
14	93	207	No	Phd and other	47520.00	510.97	>$50000	6875.00	1.56	6873.44	.
15	93	11	No	Phd and other	36000.00	387.10	>$50000	4500.00	1.52	4498.48	1
16	93	22	No	Phd and other	33000.00	354.84	>$50000	4250.00	1.52	4248.48	1

5.4 Creating subsets

In some instances, you might want to use only part of the file in an analysis. For example, you might want to look at changes in income among single working mothers. Or you might want to consider only staff born before a specific date. To select a *subset* of the cases in your file,

1. From the menu, select Data > Select Cases

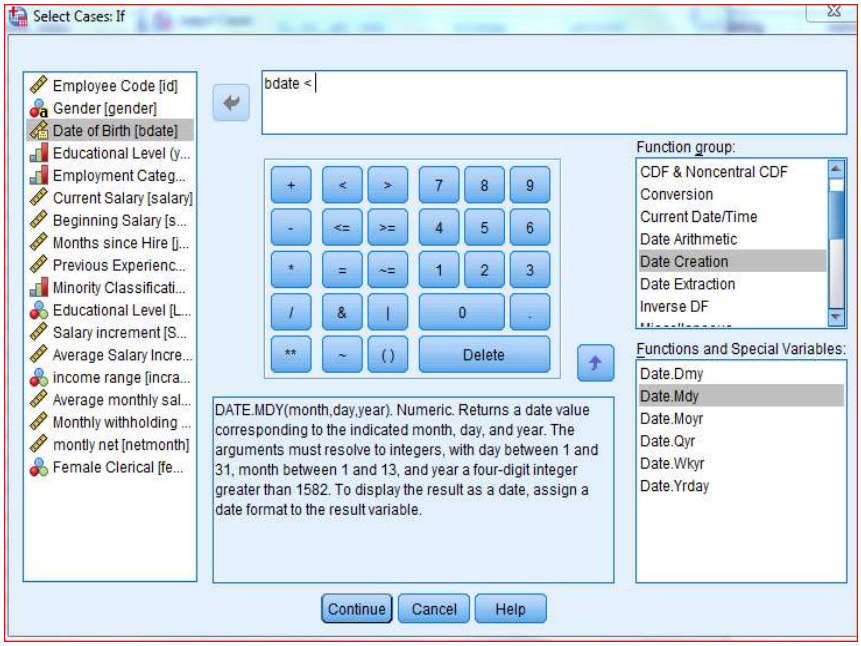


2. Select **If condition is satisfied** by clicking its radio button.

3. Click **If....** Notice that the **Select Cases: If** window looks exactly like the If window you used in the earlier compute procedures.

4. From the variable list, double-click Date of Birth.

5. Click the cursor anywhere after **bdate** in the calculation pane.



6. Type (or select from the keypad): **<**

7. Scroll through the Function group and highlight date creation and double-click DATE.MDY (month,day,year).

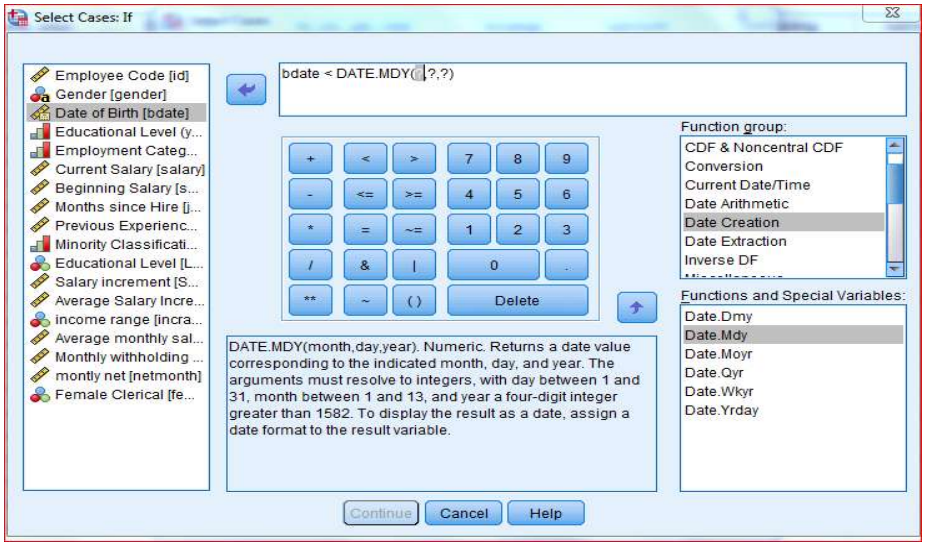
In the next step, you'll set the date criterion. SPSS adds the function to the calculation pane, substituting question marks to indicate that you need to specify

the values.

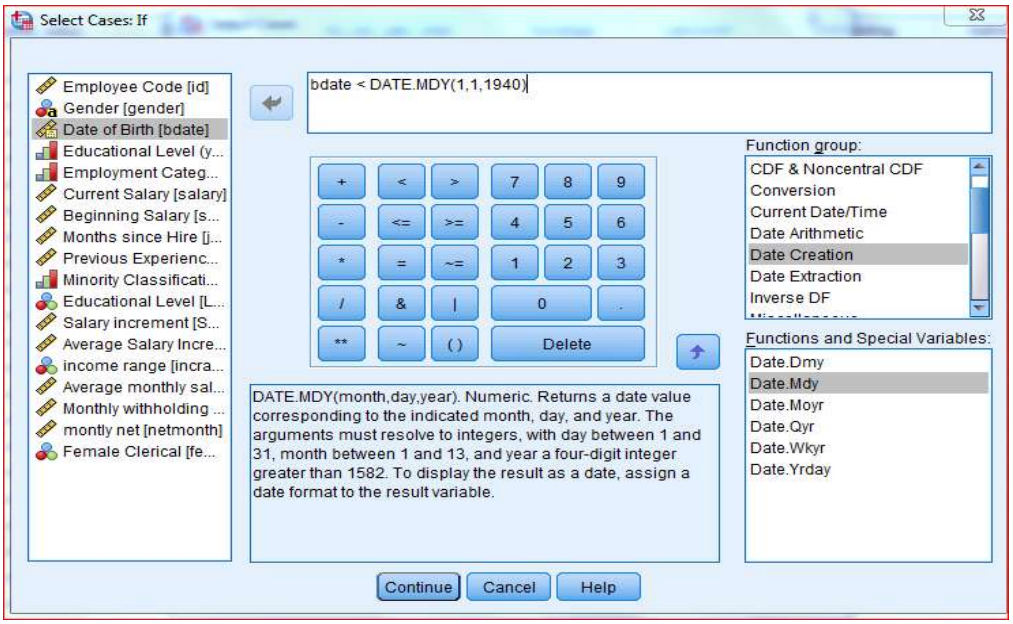
8. Select the first question mark and type: **1**

9. Select the second question mark and type: **1**

10. Select the third question mark and type: **1940**

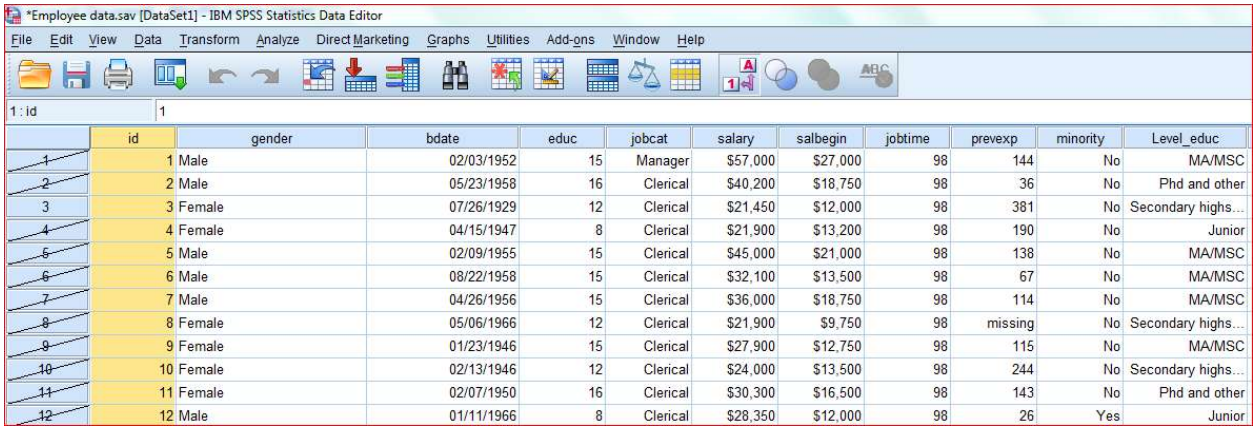


Your completed window should look like



11. Click **Continue.**

12. Click **OK.** Notice that many of the records are marked with a diagonal line through the record number as shown below. These cases are excluded from any further calculations until you specifically include them again.



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	Level_educ
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	144	No	MA/MSC
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	36	No	Phd and other
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	381	No	Secondary highs...
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	190	No	Junior
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98	138	No	MA/MSC
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98	67	No	MA/MSC
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98	114	No	MA/MSC
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98	missing	No	Secondary highs...
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98	115	No	MA/MSC
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98	244	No	Secondary highs...
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98	143	No	Phd and other
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98	26	Yes	Junior

13. To see the effect of the subset selection, right click the heading for **bdate.**

14. From the pop-up menu, select Sort Ascending. Notice that all employees born before 1940 are selected, except for the person with the missing date of birth. In the next step, you'll instruct SPSS to include all cases until otherwise instructed.

15. From the menu select Data > Select Cases.

16. Select All Cases by clicking its radio button

6 Measuring differences

Typically, differences in one or more continuous dependent variables based on differences in one of more categorical variables are evaluated using a t-test or an analysis of variance. If we have only one continuous dependent variable and only one categorical independent variable with no more than two values, the t-test can be used to look for differences. If we have more than one dependent continuous variable or more than two values across the categorical independent variables or we have both categorical and continuous independent variables, we need to use an analysis of variance (ANOVA).

6.1 T-Tests

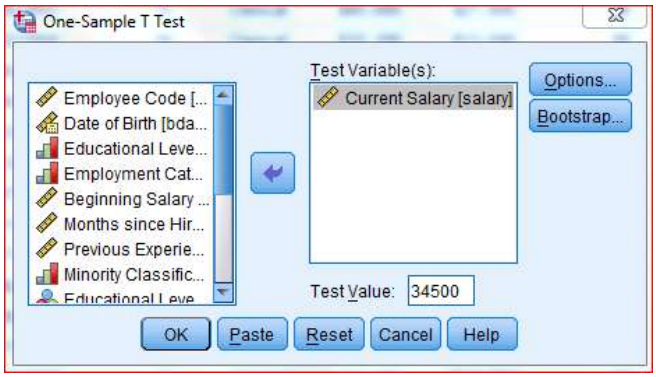
There are three types of t-tests: (A One Sample t-test, A Paired Samples t-test and Independent Samples t-test)

6.1.1 A One Sample t

A One Sample t-test is used to evaluate whether the mean of a continuous dependent variable is different from zero. To test if the mean is different than some other value, subtract that value from each observation and the test to see if the mean of the new values is zero.

Steps to compute one sample t-test on SPSS

1. Click Analyze menu-compare means –one sample t-test
2. Move current salary to test variable(s)
3. Type 34500 at test value as you see in the following window



4. Click option button and set confidence interval percentage
5. Then click **OK** and see the result in the output window as show below

T-Test

[DataSet1] G:\spss manual\Data_practice\Employee data.sav

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	67	$24,717.16	$7,752.314	$947.096

One-Sample Test

	Test Value = 34500					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Current Salary	-10.329	66	.000	-$9,782.836	-$11,673.77	-$7,891.90

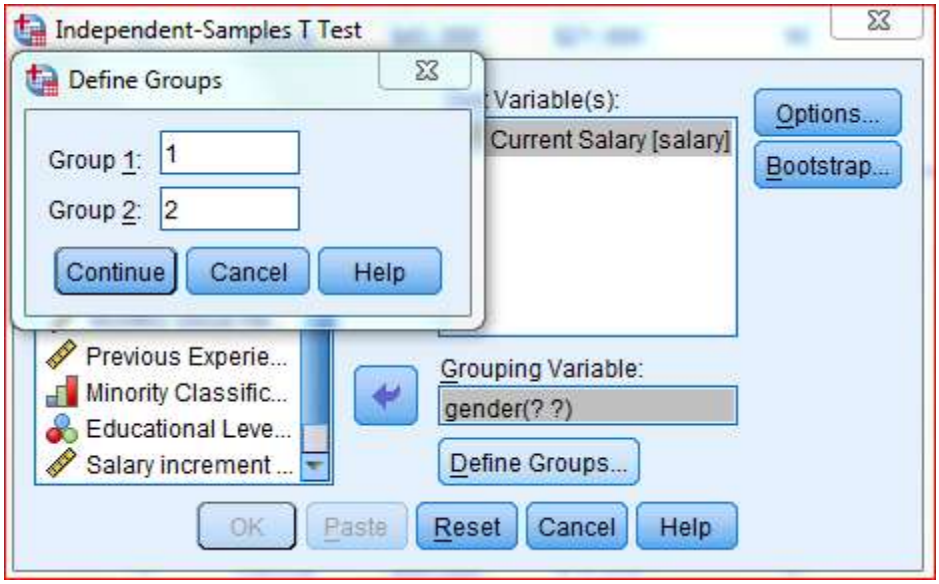
Result: The estimated average salary (34500) is statistically significant. Meaning the average current salary is equal or nearly equal to $34500

6.1.2 Independent Samples t

An Independent Samples t-test is used when cases are randomly assigned to one of two groups. After a differential treatment has been applied to the two groups, a measurement is taken which is related to the effect of the treatment. The t-test is calculated to determine if any difference between the two groups is statistically significant.

Steps to compute independent samples t-test on SPSS

1. Click analyze-compare means - independent samples t-test
2. Move current salary to test variable(s) and Gender to grouping variables to test if there is difference between the two genders in terms of their current salary.
3. Highlight gender under grouping variables and click on define groups



4. Then decide which will be subtracted from which. For example is its salary of females from males or vice versa. So if 1 is for group one and 2 for group 2.It means that the salary of males will be subtracted from females'. Let we decided here to subtract salary of males from females
5. Click **continue and then Ok**
6. Then you will see the following result in the output window

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Current Salary	Equal variances assumed	119.669	.000	-10.945	472	.000	-$15,409.862	$1,407.906	-$18,176.401	-$12,643.322
	Equal variances not assumed			-11.688	344.262	.000	-$15,409.862	$1,318.400	-$18,002.996	-$12,816.728

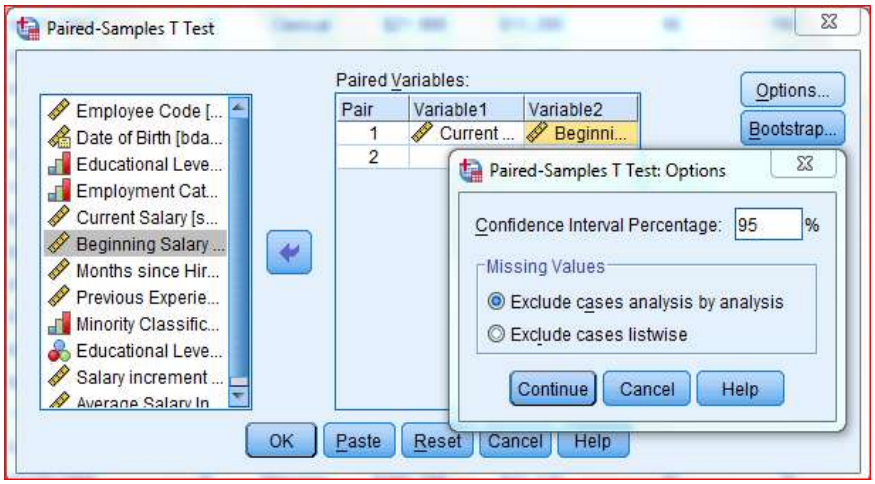
Result: Since p<0.05 the difference between females' and males' average current salary is statistically significant. If males' were labeled as group 1 and females as group 2 the mean difference (-$15,409.862) would have been positive you can test it by re running independent samples t-test again. As a result the average salary of females averagely less by 15,409.862 than males' average salary.

6.1.3 Paired Samples t-test

A Paired Samples t-test can be used to evaluate differences between two groups who have been matched on one or more characteristics or evaluate differences in before/after measures on same person. If you want to use pre/post measures, make sure the post-test is the same as the pre-test.

Steps to compute paired samples t-test on SPSS

1. Analyze-compare means –paired samples T-test
2. Move current salary first and then beginning salary to paired variables as variable 1 and variable 2 respectively
3. Click on option button and set the confidence interval percentage as shown in the following window.



4. Click continue and then Ok
5. Then you see the result from output window

T-Test

[DataSet1] G:\spss manual\Data_practice\Employee data.sav

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Current Salary	$34,419.57	474	$17,075.661	$784.311
	Beginning Salary	$17,016.09	474	$7,870.638	$361.510

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Current Salary & Beginning Salary	474	.880	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Current Salary - Beginning Salary	$17,403.481	$10,814.620	$496.732	$16,427.407	$18,379.555	35.036	473	.000

Result: Since p<0.05 and t-statistics is out of the lower and upper limit the difference between the current and beginning salary is significant. Hence the current employees' salary is averagely higher than the beggnning by $496.732.

6.2 ANOVA

If we have more than one dependent continuous variable or more than two values across the categorical independent variables or we have both categorical and continuous independent variables, we need to use an analysis of variance to measure differences. There are a number of particularly useful special cases for the general analysis of variance model.

If we have only one dependent continuous variable and one independent categorical variable we can use a One Way Analysis of Variance or One-Way ANOVA. If we have only one dependent continuous variable, but more than one independent categorical variable we can use a general Univariate Analysis of Variance or Univariate ANOVA. If we have one or more independent continuous variables, we can use the Univariate Analysis of Covariance or Univariate ANCOVA More advanced models are available for more than one dependent continuous variable. If the model has one or more independent categorical variables, we would use a Multivariate Analysis of Variance or MANOVA. If the model also included one or more continuous independent variables, we would use a Multivariate Analysis of Covariance or MANCOVA.

6.2.1 One-Way ANOVA

One-way analysis of variance is an extension of the t-test in that, in a t-test you have two groups, one that received a treatment and one that did not, while in a one-way analysis of variance you have more than two groups where the groups received different variation of the same treatment.

Steps to compute One way ANOVA on SPSS

1. Analyze-compare means-One way ANOVA
2. Move current salary to dependant list and job category to factor
3. Click on post hoc and activate one of LSD or Bonferroni or Tukey. Let we activate LSD
4. Click continue
5. Click option button and activate the statistics you want. Let we activate homogeneity of variance test and click continue

6. Then click Ok and see the result from the output window

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	89438483926	2	44719241963	434.481	.000
Within Groups	48478011510	471	102925714.5		
Total	1.379E+11	473			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Current Salary
LSD

(I) Employment Category	(J) Employment Category	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Clerical	Custodial	-$3,100.349	$2,023.760	.126	-$7,077.06	$876.37
	Manager	-$36,139.258*	$1,228.352	.000	-$38,552.99	-$33,725.53
Custodial	Clerical	$3,100.349	$2,023.760	.126	-$876.37	$7,077.06
	Manager	-$33,038.909*	$2,244.409	.000	-$37,449.20	-$28,628.62
Manager	Clerical	$36,139.258*	$1,228.352	.000	$33,725.53	$38,552.99
	Custodial	$33,038.909*	$2,244.409	.000	$28,628.62	$37,449.20

*. The mean difference is significant at the 0.05 level.

Result: Since p<0.05(see ANOVA table) the overall current salary difference among different employment categories is significant. But the difference between clerical and custodial is not significant. The salary of clerics averagely less than managers by $36139.258.The custodial's averagely less by $33038.909 than managers.

7 Measuring association

Typically the association between two variables is evaluated by using a bivariate correlation procedure. If the two variables are continuous and you want to predict one variable using the value of the other, a simple linear regression or some method of curve estimation can be used. If there are more than two variables and they are continuous, use a partial or a multiple

correlation procedure. If you want to predict one of the variables using the values of the other variables, a multiple regression can be used.

If you have frequency distributions based upon one or more categorical variables, you should consider cross tabulation or Chi-square.

7.1 Bivariate correlations

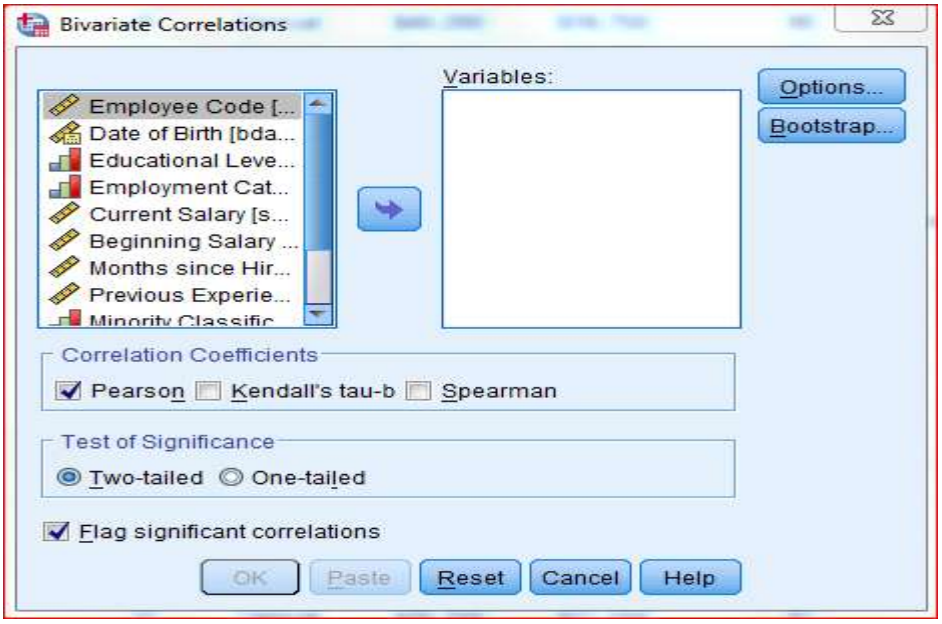
Bivariate correlations measure the degree of association between two variables. If the two variables are continuous, the Pearson product moment correlation is an appropriate measure. If they are not continuous (that is, if they are discrete or categorical), it would be more appropriate to use Spearman's rho or Kendall's tau-*b*. The correlation coefficient, which ranges from 1 to 1, is both a measure of the strength of the relationship and the direction of the relationship. A correlation coefficient of 1 describes a perfect relationship in which every change of +1 in one variable is associated with a change of +1 in the other variable. A correlation of -1 describes a perfect relationship in which every change of +1 in one variable is associated with a change of -1 in the other variable. A correlation of 0 describes a situation in which a change in one variable is not associated with any particular change in the other variable. In other words, knowing the value of one of the variables gives you no information about the value of the other. The correlation squared is another measure of the strength of the relationship. In fact, *the correlation squared is the percent of the variance in the dependent variable that is accounted for or predicted by the independent variable.*

You can also determine the statistical significance of the correlation coefficient. If the direction of the association is hypothesized in advance, you can use a one-tailed test to determine whether the correlation is statistically significantly different from zero, otherwise use a two-tailed test.

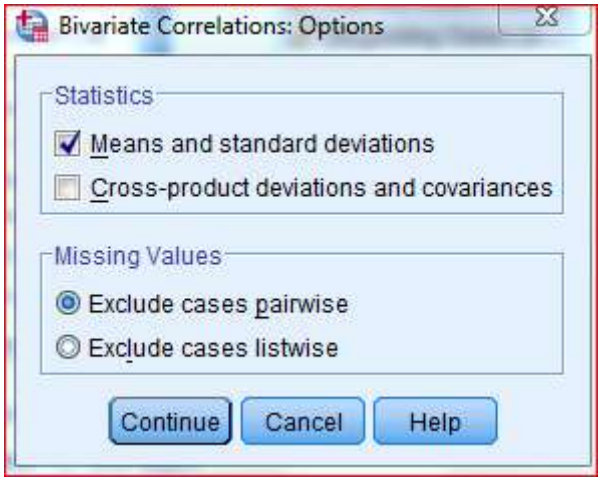
Correlation is not causation. If we looked at the correlation of the time the paper boy delivers the morning paper and the time of the sunrise, we would find a very strong positive correlation. And yet, we would be reluctant to claim that the newspaper boy causes the sun to rise.

Steps

1. In the data window, open the file named Employee data.sav in the folder named SPSS Tutorial Data
2. In the data window, from the menu select Analyze > Correlate > Bivariate



3. Double-click **Current Salary** to move it to the Variables list.
4. Double-click **Beginning Salary** to move it to the Variables list.
5. Click **Options.**
6. Activate Pearson correlation coefficient
7. In the Statistics pane, select **Means and standard deviations** by clicking its check box.



8. Click **Continue**
9. Click **OK.** The output is displayed in the Output window as follow

Correlations

		Current Salary	Beginning Salary
Current Salary	Pearson Correlation	1	.880**
	Sig. (2-tailed)		.000
	N	474	474
Beginning Salary	Pearson Correlation	.880**	1
	Sig. (2-tailed)	.000	
	N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

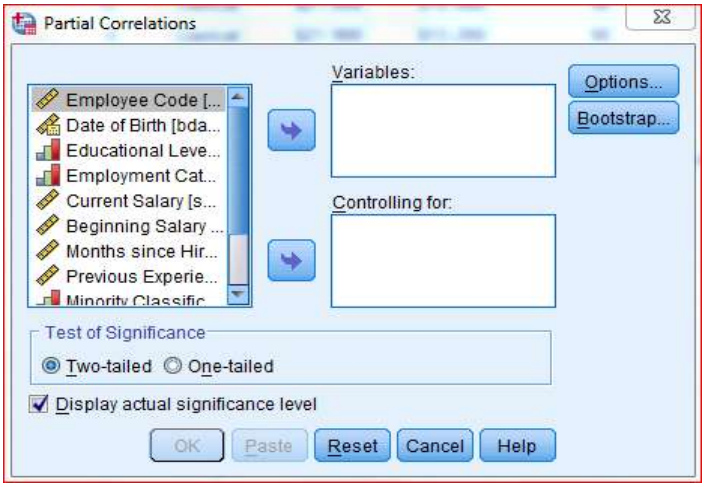
10. Notice that the correlation is particularly high (.880). The footnote to the table indicates that correlation is significant at the .01 level. (And, no, we can't find any documentation on why SPSS has highlighted the particularly high correlation. Just another one of those moments of cryptic helpfulness.)

7.2 Partial correlation

Partial correlation is used to measure the association of two continuous variables after controlling for the association of other variables. Conceptually, what is being done is to first calculate the variance in the dependent variable that can be explained or accounted for by all of the *control* variables. The variance accounted for by the control variables is then removed from the dependent variable. Finally, the degree of association is measured between the variance remaining in the dependent variable and the non-controlled variable.

Steps

1. From the menu, select Analyze > Correlate > Partial.



2. Notice that this time, in addition to selecting the variables to be compared, you can also select **Controlling for.**
3. Select **Current Salary** and **Beginning Salary** and move them to the **Variables** pane.
4. Select **Previous Experience** and move it to the **Controlling For** pane.
5. Click **Options** to select the statistics you want displayed.
6. Select **Means and standard deviations** by clicking its check box.
7. Click **Continue.**
8. Click **OK.** The new output is displayed in the output window as follow

Correlations

Control Variables			Current Salary	Beginning Salary
Months since Hire	Current Salary	Correlation	1.000	.885
		Significance (2-tailed)	.	.000
		df	0	471
	Beginning Salary	Correlation	.885	1.000
		Significance (2-tailed)	.000	.
		df	471	0

7.3 Multiple correlations (multiple regressions)

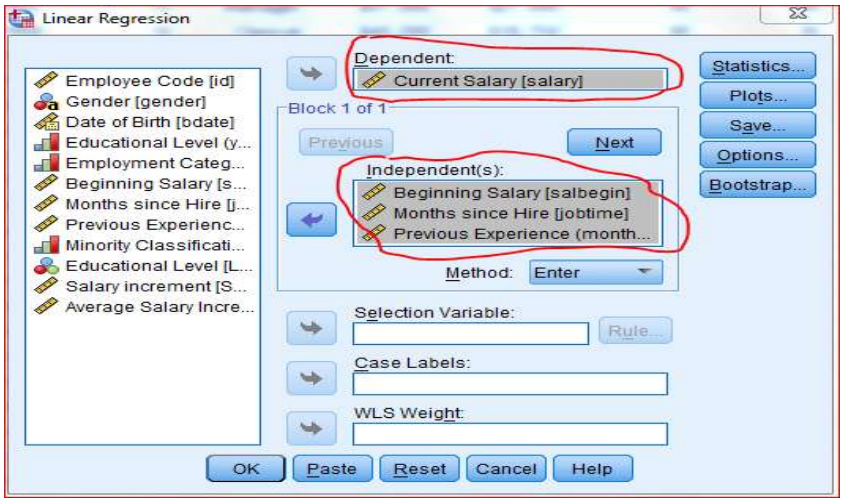
Multiple correlation looks at the association between one continuous variable (often called the dependent variable) with a group of two or more continuous variables (usually called predictors). One use for a multiple correlation is to find out if there is a relationship between an independent variable and a dependent variable *after controlling* for a subset of all other variables. In this sense the multiple correlations or multiple regressions is used as a more sophisticated method of exploring partial correlations.

When you run a step-wise multiple regression, SPSS will find the one variable in the group of predictors which has the highest correlation with the dependent variable. It will then statistically remove that variance from the dependent variable that the predictor variable accounts for. The procedure will then go to the list of remaining predictors and select the variable which has the highest correlation with the remaining variance in the dependent variable, remove *that* variance, then select the next predictor and so on until some criterion is met. Typical criteria that you can specify are the amount of additional variance accounted, the level of statistical significance for the change in variance accounted for, and the maximum number of predictors that can be

selected. Example: In a study of the effectiveness of entitlement programs, you want to find out which set of variables can best predict client's income once they are no longer receiving benefits. All entitlement data are quantitative, including time receiving benefits, individual benefit values, length of job training, and family size. A single categorical variable — minority/non-minority — is included in the calculations as a binary variable. In this example, post-eligibility income is the dependent variable,

Steps

1. Analyze>regression>linear
2. Move current salary to dependant pane and beginning salary, month since hire and previous experience to independent(s) pane to examine the effect of beginning salary, month since hire and previous experience on current salary.



3. Click statistics and activate regression coefficients and residuals you want
4. Click Ok and then you will see model summary and coefficient tables in the out put as shown below

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.897[a]	.804	.803	$7,586.187	.804	642.151	3	470	.000

a. Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-10266.629	2959.838		-3.469	.001	-16082.782	-4450.475
	Beginning Salary	1.927	.044	.888	43.435	.000	1.840	2.015
	Months since Hire	173.203	34.677	.102	4.995	.000	105.062	241.344
	Previous Experience (months)	-22.509	3.339	-.138	-6.742	.000	-29.070	-15.949

a. Dependent Variable: Current Salary

Result: since $p<0.05$ and t-statistics are not within the region of accepting the null hypothesis it rejected. Hence all- beginning salary, month since hire and previous experience have statistically significant effect on the current salary. While beginning salary and month since hire has positive association with current salary, previous experience has negative association.

The model is the following

$$\hat{y} = A + B_1X_1 + B_2X_2 + B_3X_3$$

$$\hat{y} = -10266.629 + 1.927Beggnning\ salary + 173.203month\ since\ hire - 22.509previous\ experience$$

7.4 Crosstabs

"A cross tabulation is a joint frequency distribution of cases according to two or more classificatory variables. The display of the distribution of cases by their position on two or more variables is the chief component of contingency table analysis *and is indeed the most commonly used analytic method in the social sciences.*"

The Chi-square test can be used to determine whether the frequency distributions of one or more categorical variables are statistically independent. The crosstab can be used to provide measures of the associations of categorical variables. Some of the measures of association are the contingency coefficient, phi, tau, gamma, etc..

These measures describe the degree to which the values of one variable predict or vary with those of another.

Data requirements: Crosstabs require categorical data or continuous data recoded into categories, such as income or age ranges. The frequencies for each variable in the population should be approximately normal.

Steps

1. From the menu, select File > Open > Data.
2. Navigate to the file named employee. Data. sav and open it.
3. From the menu, select Analyze > Descriptive Statistics > Crosstabs.
4. Select employment category and move it to the Columns pane.
5. Select gender and move it to the Rows pane.
6. Click **Statistics.**
7. Select Chi-square by clicking its check box.
8. Click **Continue.**
9. Click **Cells.**
10. Select **Observed** and **Expected.**
11. Click **Continue.**
12. Click **OK.** The Chi-square results appear in the Output window. Notice that all the significance levels are less than .001. Something is definitely going on here.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Employment Category	474	100.0%	0	0.0%	474	100.0%

Gender * Employment Category Crosstabulation

			Employment Category			
			Clerical	Custodial	Manager	Total
Gender	Female	Count	206	0	10	216
		Expected Count	165.4	12.3	38.3	216.0
	Male	Count	157	27	74	258
		Expected Count	197.6	14.7	45.7	258.0
Total		Count	363	27	84	474
		Expected Count	363.0	27.0	84.0	474.0

Chi-Square Tests

				Monte Carlo Sig. (2-sided)		
					99% Confidence Interval	
	Value	df	Asymp. Sig. (2-sided)	Sig.	Lower Bound	Upper Bound
Pearson Chi-Square	79.277[a]	2	.000	.000[b]	.000	.000
Likelihood Ratio	95.463	2	.000	.000[b]	.000	.000
Fisher's Exact Test	90.869			.000[b]	.000	.000
N of Valid Cases	474					

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.30.

b. Based on 10000 sampled tables with starting seed 2000000.

Result: Since p<0.01 the association between gender and job category is significant. The number of female clerics is higher than their counterpart males and the number of male custodial and managers is higher than females.