

Univariate Statistics

Univariate analysis, looking at *single* variables, is typically the first procedure one does when examining data being used for the first time. There are a number of reasons why it is the first procedure, and most of the reasons we will cover at the end of this chapter, but for now let us just say we are interested in the "basic" results. In other words, if we are examining a survey, we are interested in how many people said, "Yes" or "No", or how many people "Agreed" or "Disagreed" with a statement. We aren't really testing a traditional hypothesis with an independent and dependent variable; we are just looking at the distribution of responses.

The SPSS tools for looking at single variables include the following procedures: "Frequencies", "Descriptives" and "Explore" all located under the "Analyze" menu.

This chapter will use the 2004GSS.SAV file used in earlier chapters, so start PASW and bring the file into the Data Editor. (Forgot how? See the podcast in an earlier Announcement and lecture 3 on how to start PASW). To begin the process start PASW, then open the data file. Under the "Analyze" menu, choose "Descriptive Statistics" and the procedure desired: "Frequencies", "Descriptives", or "Explore."

Frequencies

Generally a frequency is used for looking at detailed information on nominal (category) data and describing the results. Categorical data is for variables such as gender i.e. males are coded as "1" and females are coded as "2." Frequencies options include a table showing counts and percentages, statistics including percentile values, central tendency, dispersion and distribution, and charts including bar charts and histograms. The steps for using the frequencies procedure is to click the "Analyze" menu, "Descriptive Statistics" then from the sub-menu choose "Frequencies" and select your variables for analysis. You can then choose statistics options, choose chart options, choose format options, and have PASW calculate your request.

For this example we are going to check out attitudes on the abortion issue. The 2004 General Social Survey, 2004GSS.SAV, has the variable "ABANY" with the label "ABORTION FOR ANY REASON." We will look at this variable for our initial investigation.

Choosing Frequencies Procedure:

From the "Analyze" menu, highlight "Descriptive Statistics", Figure 4-1, then move to the sub menu and click on "Frequencies."

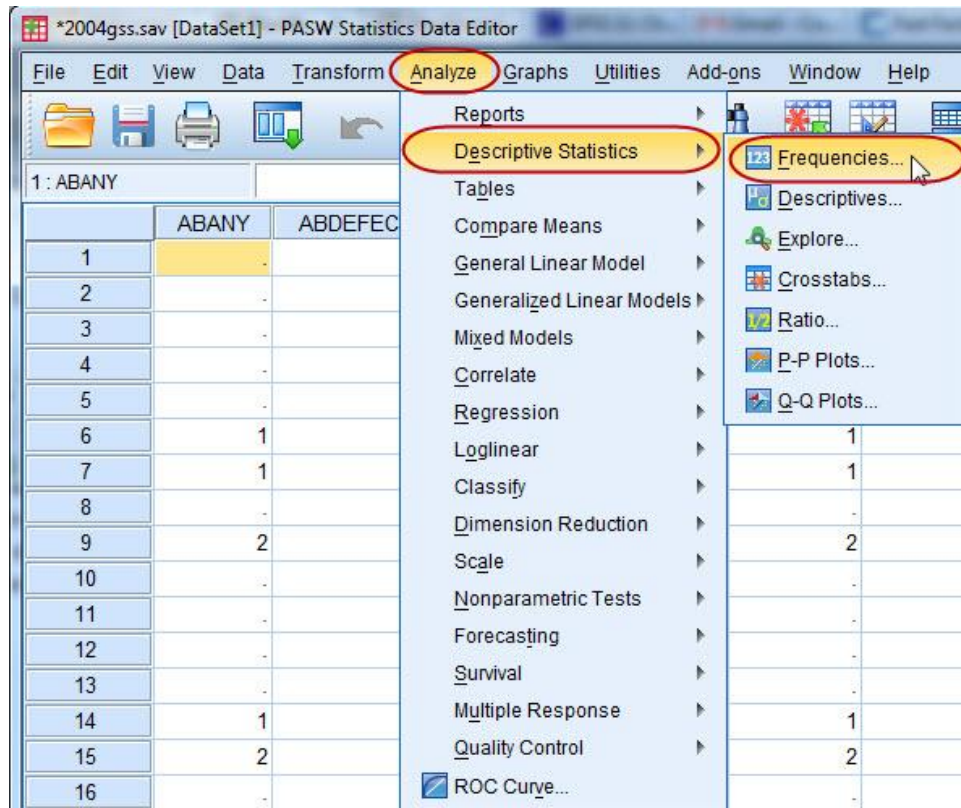


Figure 4- 1

A dialog box, Figure 4-2, will appear providing a scrollable list of the variables on the left, a "Variable(s)" choice box, and buttons for "Statistics", "Charts" and "Format" options.

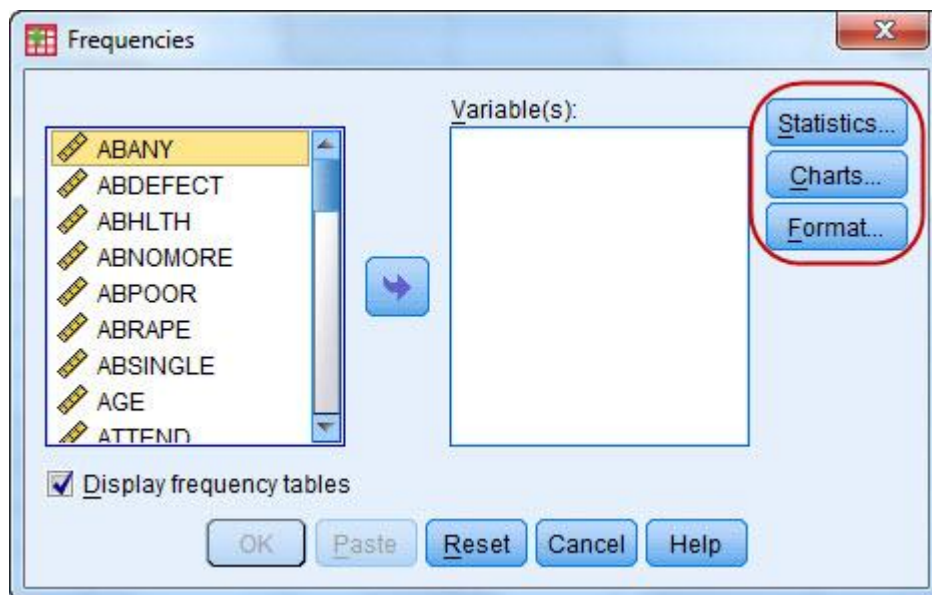


Figure 4- 2

Selecting Variables for Analysis:

First select your variable from the main frequencies dialog box, Fig 4-2, by clicking the variable name once. (Use the scroll bar if you do not see the variable you want.) In this case "ABANY" is the first variable and will be selected (i.e., highlighted). Thus, you need not click on it.

Click the arrow to the left of the "Variable(s):" box, Figure 4-2, to move "ABANY" into the box. All variables selected for this box will be included in any procedures you decide to run. We could click OK to obtain a frequency and percentage distribution of the variables, but in most cases we would continue and choose one or more statistics.

Choosing Statistics for Variables:

Click the "Statistics" button, bottom of Figure 4-2, and a dialog box of statistical choices will appear, Figure 4-3.

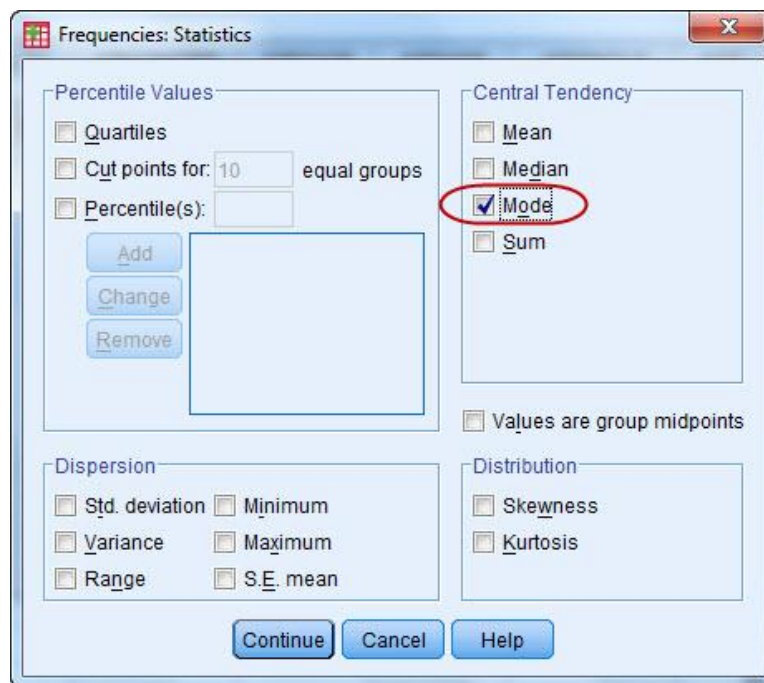


Figure 4- 3

This variable is a nominal (category) variable so click only the "Mode" box within the central tendency choices, as shown in Figure 4-3.

After clicking the "Mode" box click the "Continue" button and we return to the main "Frequencies" dialog box, Figure 4 2.

We could now click "OK" and SPSS would calculate and present the frequency and percent distribution (click "OK" if you want) but, in the more typical manner, we will continue and include choices for charts and check out the "Options" possibilities. If you clicked "OK", just press the "Analysis" menu then choose "Descriptive Statistics" and then "Frequencies" from the sub menu and you will be back to this point with your variable and statistics chosen.

Choosing Charts for Variables:

On the main frequencies window, click the "Charts" button, Figure 4 2, and a dialog box of chart choices, Figure 4-4, will appear.

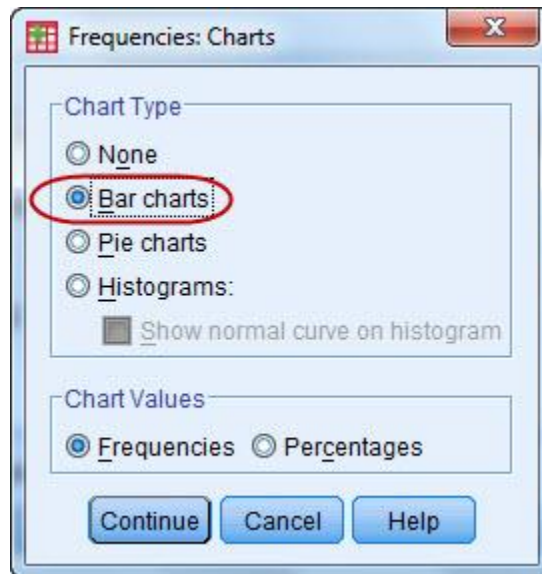


Figure 4- 4

Click "Bar Chart", as I have done, since this is a categorical variable, then click "Continue" to return to the main Frequencies window box. If this were a continuous variable I would choose "Histograms" and the "With Normal Curve" option would be available. I would choose the "With Normal Curve" option to have a normal curve drawn over my distribution so that I could visually see how close the distribution is to normal. Note: "Frequencies" is automatically chosen for chart values but if desired you could change that to "Percentages".

Now click "OK" on the main frequencies dialog box and SPSS will calculate and present a frequency and percent distribution with our chosen format, statistics, and chart. (Note: We could look to see if additional choices should be made by clicking the "Format" button. In this case we don't need to do this because all the "Format" defaults are appropriate since we are looking at only one variable.)

Looking at Output from Frequencies:

We will now take a brief look at our output from the SPSS frequencies procedure. (Processing time for SPSS to perform the analysis in the steps above will depend on the size of the data set, the amount of work you are asking SPSS to do and the CPU speed of your computer). The "SPSS Output Navigator", left side, and the output, right side, will appear when SPSS has completed its computations. Either scroll down to the chart in the right window, or click the "Bar Chart" icon in the outline pane to the left of the output as we did in Figure 4-5.

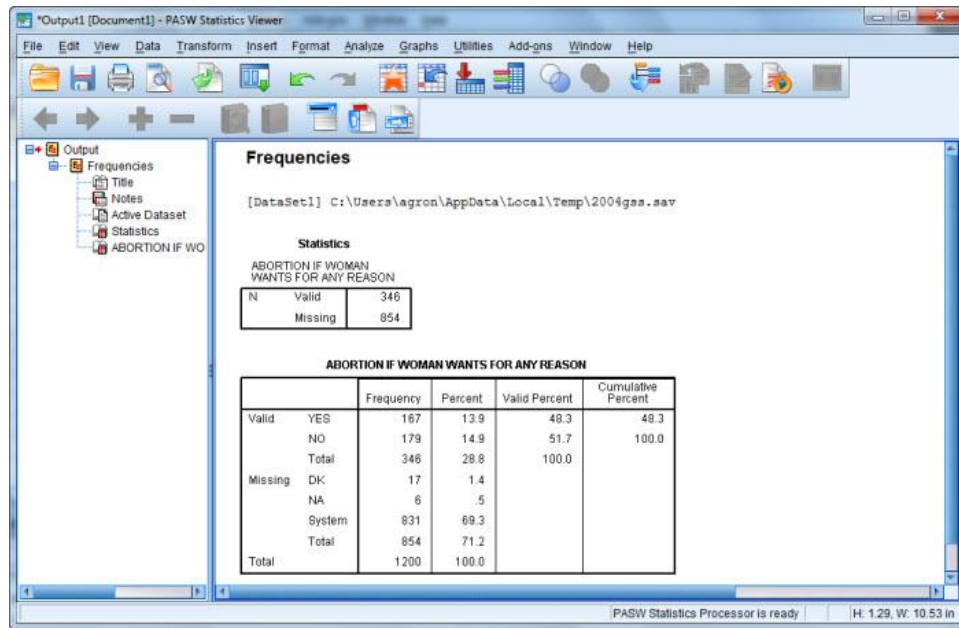


Figure 4- 5

Interpreting the Chart:

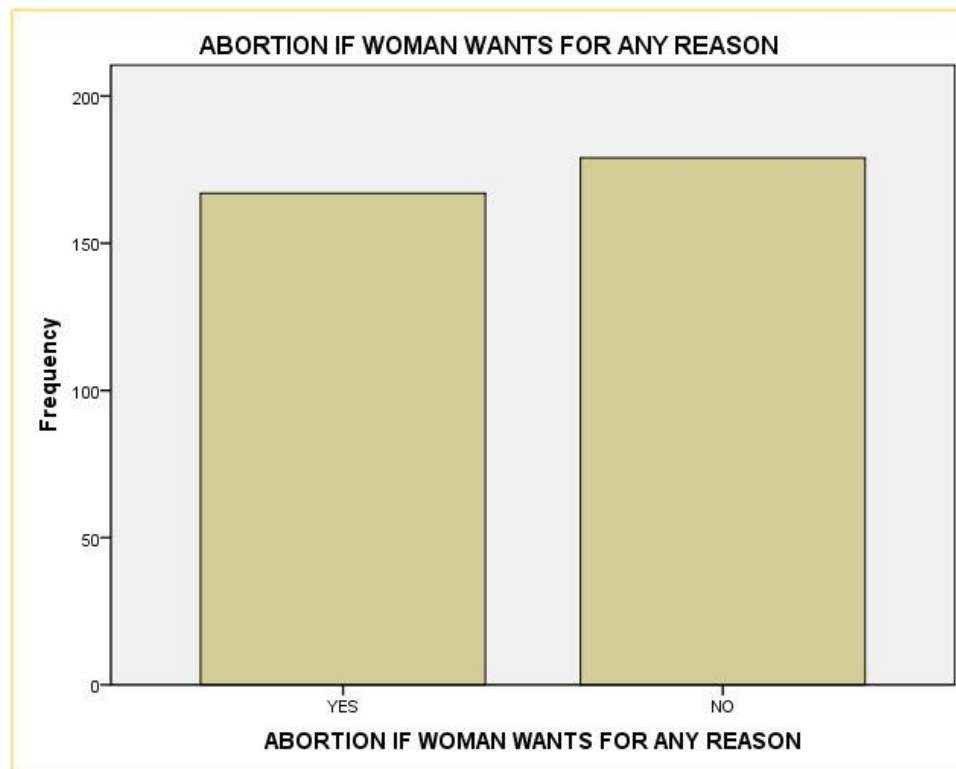
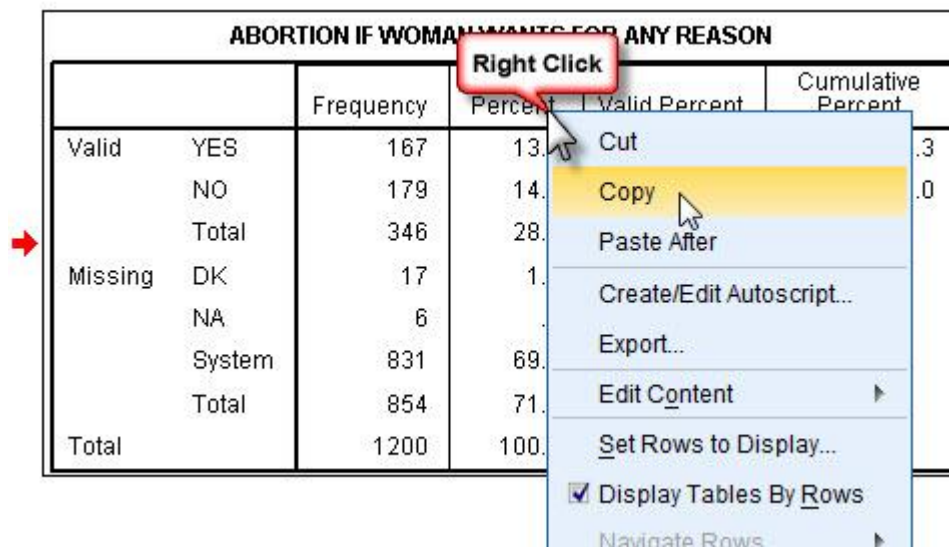


Figure 4- 6

We now see the chart, Figure 4-6. The graphic is a bar chart with the categories at the bottom, the X axis, and the frequency scale at the left, the Y axis. To display the chart, scroll down the bar on the right of your table. The variable label "ABORTION FOR ANY REASON" is displayed at

the top of the chart. We see from the frequency distribution that there are more "no", 51.7%, answers than "yes", 48.3% answers (see Figure 4-7), when respondents were asked if a woman should be able to get an abortion for any reason-Remember! ALWAYS LOOK AT THE VALID PERCENT COLUMN. A much smaller number, which does not appear on this chart, 1.4% (see Figure 4 7), chose "don't know", "DK." If a chart were the only data presented for this variable in a report, you should look at the frequency output and report the total responses and/or percentages of "yes", "no" and "DK" answers as I did in the description of this chart. You could (should?) also label the chart with frequencies and/or percentages. There are lots of possibilities for enhancing this chart within PASW.

If we choose to copy our chart to a word processor program for a report, first select the chart by clicking the mouse on the bar chart. A box with handles will appear around the chart. Select "Copy" from the "Edit" menu. Start your word processing document, click the mouse where you want the chart to appear then choose "Paste Special" from the "Edit" menu. Choose "Picture" in the paste special dialog box that appears and click "OK" to paste the chart into your document, as shown below:



ABORTION IF WOMAN WANTS FOR ANY REASON

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	167	13.9	48.3	
	NO	179	14.9	51.7	
	Total	346	28.8	100.0	
Missing	DK	17	1.4		
	NA	6	.5		
	System	831	69.3		
	Total	854	71.2		
Total		1200	100.0		

Paste it in Word and edit it accordingly, as shown in the table below:

ABORTION IF WOMAN WANTS FOR ANY REASON

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	167	13.9	48.3	48.3
	NO	179	14.9	51.7	100.0
	Total	346	28.8	100.0	
Missing	DK	17	1.4		
	NA	6	.5		
	System	831	69.3		
	Total	854	71.2		
Total		1200	100.0		

Figure 4- 7

Interpreting Frequency Output:

To display the frequency distribution, move the scroll bar on the right of our output window until the frequency is displayed or click the "Frequencies" icon in the outline box to the left of the output window. To view a large table you may want to click on the Maximize Arrow in the upper right corner of the "SPSS Output Navigator" window to enlarge the output window. Use the scroll bars to display different parts of a large table. The most relevant part of the frequency distribution for ABANY is in Figure 4-7.

We can now see some of the specifics of the SPSS frequencies output for the variable "ABANY." At the top is the variable label "ABANY ABORTION IF WOMAN WANTS FOR ANY REASON." The major part of the display shows the value labels ("YES", "NO", "Total"), and the missing categories "NAP" [Not Appropriate], "DK" [Don't Know], "NA" [Not Answered] and "Total"), and the "Frequency", "Percent", "Valid Percent", "Cumulative Percent" (the cumulative % for values as they increase in size), for each classification of the variable. The "Total" frequency and percent is listed at the bottom of the table. When asked if a woman should be able to obtain an abortion for any reason, 13.9 %, of our sample answered "yes" while 14.9 % responded "no." "DK", don't know was chosen by 1.4 % and .5% were "NA" [Not Answered]. The 69.3 % "NAP" [Not Appropriate] or MISSING, was that portion of the sample that were not asked this question. In a paper report the "Valid Percent" excludes the "missing" answers and should be reported.

Variable Names, Variable Labels, Values, Value Labels, what is going on?!

Options in Displaying Variables and Values

It is important to use these concepts correctly so a review at this point is appropriate. Variable names are the short "handle" you gave to each variable, or question in a survey. The table below is designed to help you keep these separate.

Variable Name	Variable Label	Value	Value Label
SEX	Respondent's gender	1 or 2	Male, Female
AGE	Respondent's age at last birthday	18, 19, 20, 21... 89, 98, 99	None needed
AGED	Should aged live with their children	1, 2, 3, 0, 8, 9	A good idea, Depends, A bad idea NAP [Not Appropriate], DK [Don't Know], NA [Not Answered]

Understanding these concepts allows you to intelligently customize PASW for Windows so that it is easier for you to use. You can set PASW so that you can see the Variable Names when you scroll through a listing of variables, or so that you can see the Variable Labels as you scroll through the listing. You can set PASW so that you get only the Values, only the Labels, or both in the output. Below are two examples of a frequencies dialog box.

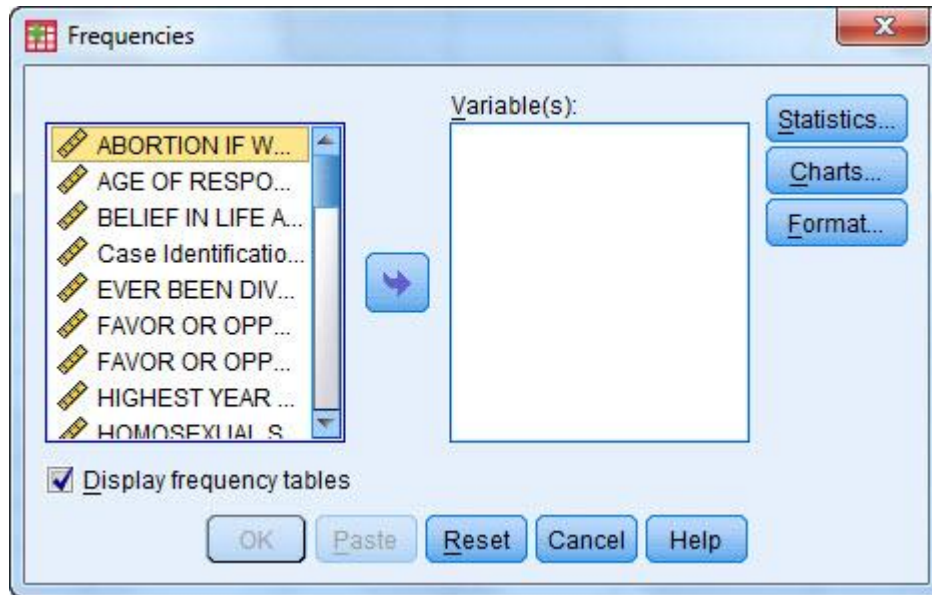


Figure 4- 8

Figure 4-8 shows the listing as the Variable Labels. This is the default setting when PASW for Windows is installed. You can change the listing however, so that you see only variable names as in Figure 4-9. Changing this is a matter of personal taste, but I suggest that variable names are easier to find, as long as you have spent some time in familiarizing yourself with these names. This Tutorial uses variable names, figure 4-9.

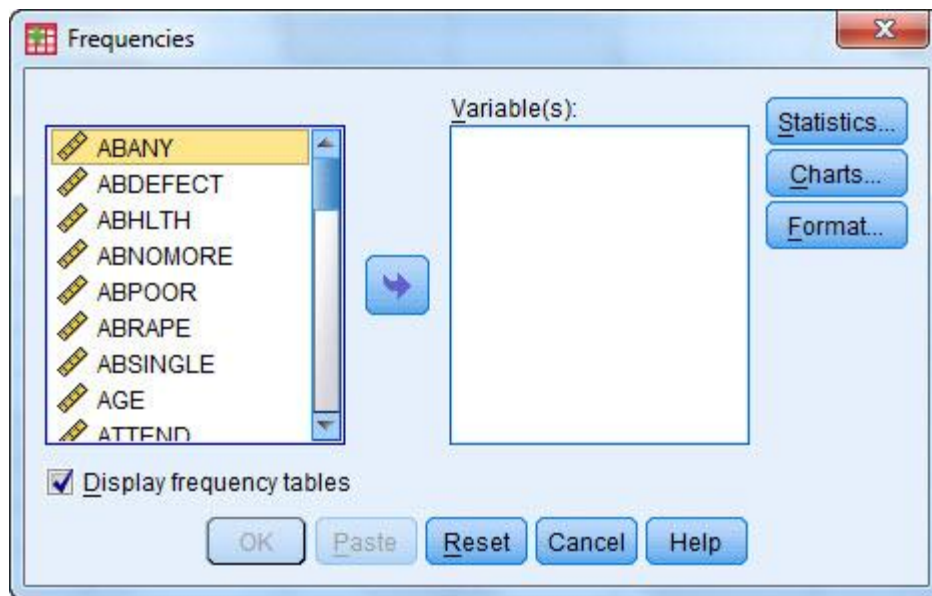


Figure 4- 9

Changing the display option for the variable selection dialog box must be done before the data file is opened. If you have PASW open with a data file click "File", "New" and "Data" and the data editor will be cleared.

To set the display option click "Edit" then chose "Options".

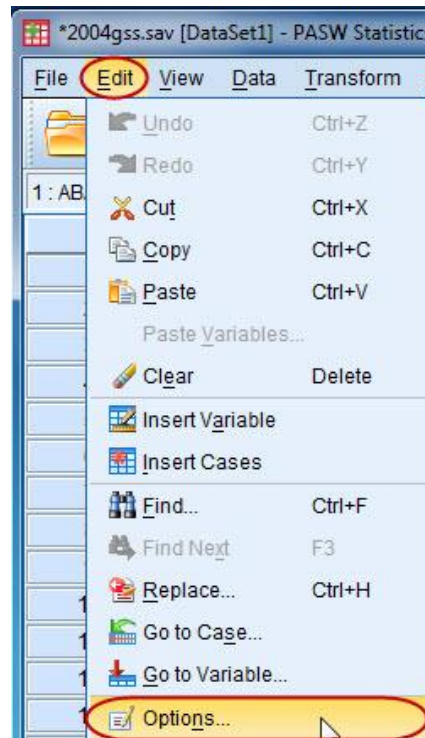
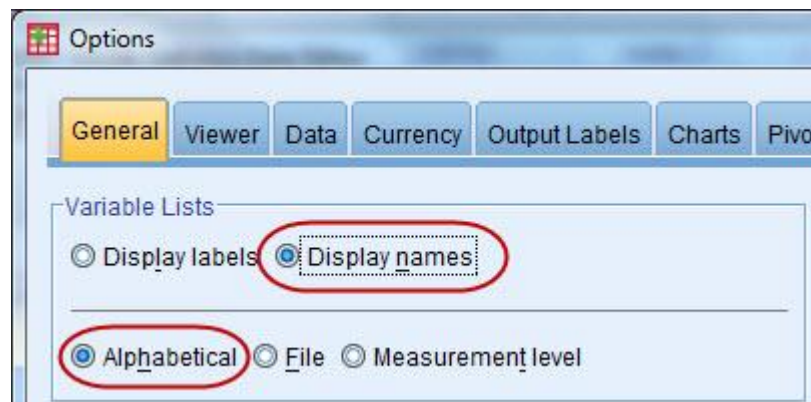


Figure 4- 10.1

The "General" tab on the options dialog box will appear, Figure 4-10.2. Under "Variable Lists" section, top left quadrant, click your choices then "OK". For this tutorial we choose "Display Names" and "Alphabetical" so that variable names will be displayed alphabetically, as in Figure 4-9 above.



Displaying Values, Value Labels or Both in Your Output

One other option you might want to make is in the table format for your SPSS output. You can choose to have displayed variable labels, values (e.g. 1, 2, 3, etc), Value Labels (YES, No, DK, etc.) or both values and labels (1 YES, 2 NO, 3 DK). To make these choices click the "Edit" menu and choose "Options", then click the "Output Labels", click the options dialog box and make your choices. You can also have the output display variable names and labels, as seen in Figure 4-11. However, we'll leave the default option (Labels) on.

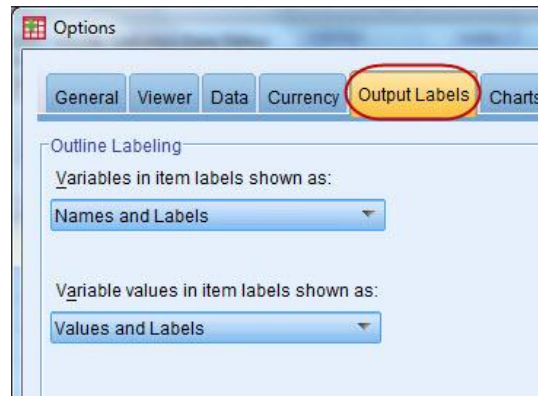


Figure 4- 11

Descriptives

"Descriptives" ("Analysis", "Descriptive Statistics", "Descriptives", Figure 4-12) is used to obtain summary information about the distribution, variability, and central tendency of *continuous* variables. Possibilities for "Descriptives" include mean, sum, standard deviation, variance, range, minimum, maximum, S.E. mean, kurtosis and skewness. For this example we are going to look at the distribution of age and education for the General Social Survey sample. Since both these variables were measured at interval/ratio level, different statistics from our previous example will be used.

Choosing Descriptive Procedure:

First click the "Analyze" menu and select "Descriptive Statistics", then move across to the sub menu and select "Descriptives" (see Figure 4-12).

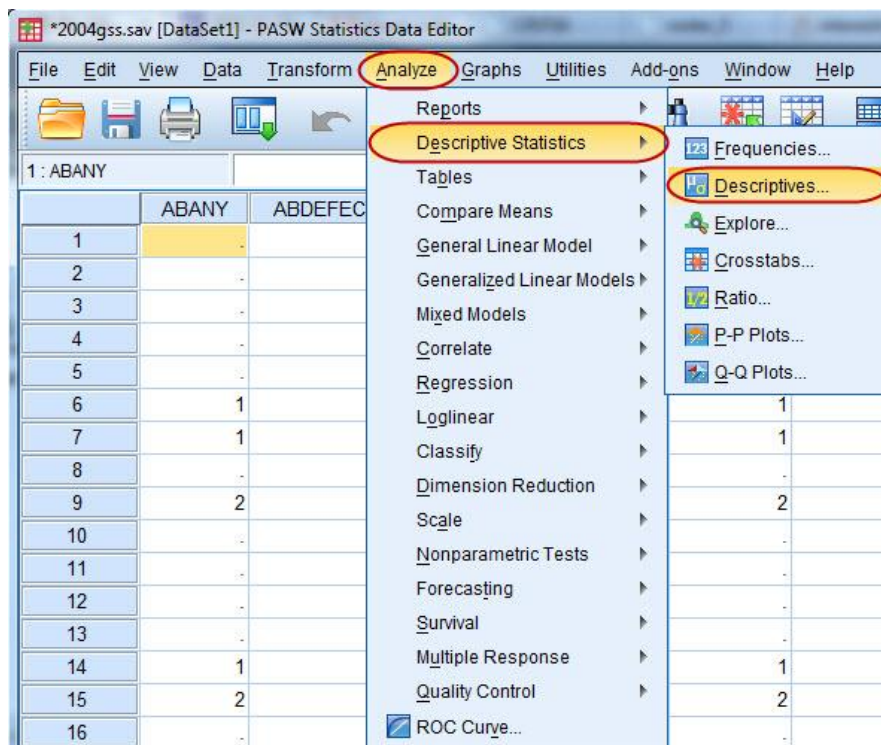


Figure 4- 12

Selecting Variables for Analysis:

First click on "AGE", the variable name for AGE OF RESPONDENT. Click the select arrow in the middle and PASW will place "AGE" in the "Variable(s)" box. Follow the same steps to choose "EDUC," the variable name for "HIGHEST YEAR OF SCHOOL COMPLETED." The dialog box should look like Figure 4-13.

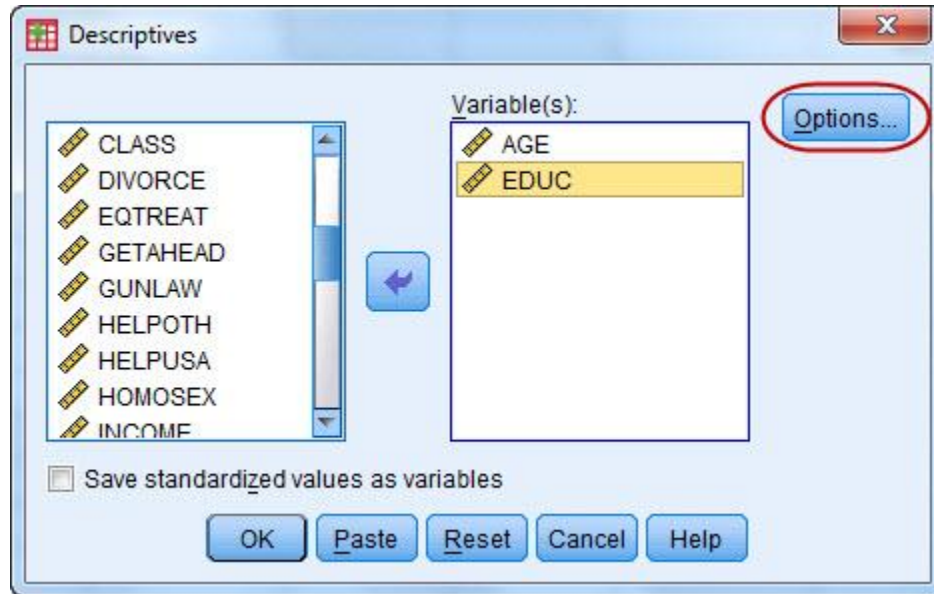


Figure 4- 13

We could click "OK" and obtain a frequency and percentage distribution, but we will click the "Options" button and decide on statistics for our output. Click "Options" and the "Descriptives: Options" dialog box, Figure 4-14, will open.

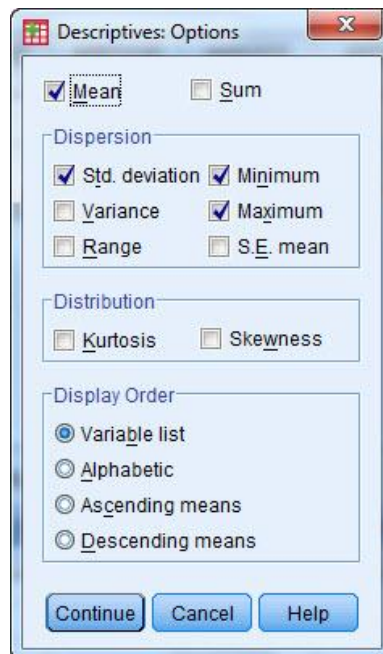


Figure 4- 14

Since these variables are interval/ratio measures, choose: "Mean," "Std. deviation," "Minimum" and "Maximum." We will leave the defaults for the "Distribution" and "Display Order."

Next, click the "Continue" button to return to the main "Descriptives" dialog box, (Figure 4-13). Click "OK" in the main "Descriptives" dialog box and SPSS will calculate and display the output seen in Figure 4-15.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
AGE OF RESPONDENT	1195	18	89	44.79	16.248
HIGHEST YEAR OF SCHOOL COMPLETED	1199	0	20	14.09	2.772
Valid N (listwise)	1194				

Figure 4- 15

Interpretation of the Descriptives Output

In the Interpretation of Figure 4-15, "AGE OF RESPONDENT" has a mean of 44.79 and a standard deviation of 16.248. The youngest respondent was 18 and the oldest was 89. "HIGHEST YEAR OF SCHOOL COMPLETED", has a mean of 14.09 (little more than 2 years and a standard deviation of 2.8. Some respondents indicated no "0" years of school completed. The most education reported was 20 years.

Explore

Explore is primarily used to visually examine the central tendency and distributional characteristics of continuous variables. "Explore" statistics include M estimators, outliers, and percentiles. Grouped frequency tables and displays, as well as "Stem and leaf" and box plots, are available. "Explore" will aid in checking assumptions with Normality plots and Spread vs. Level with the Levene test.

Choosing the Explore Procedure:

From the "Analyze" menu choose "Descriptive Statistics", drag to the sub menu and select "Explore."

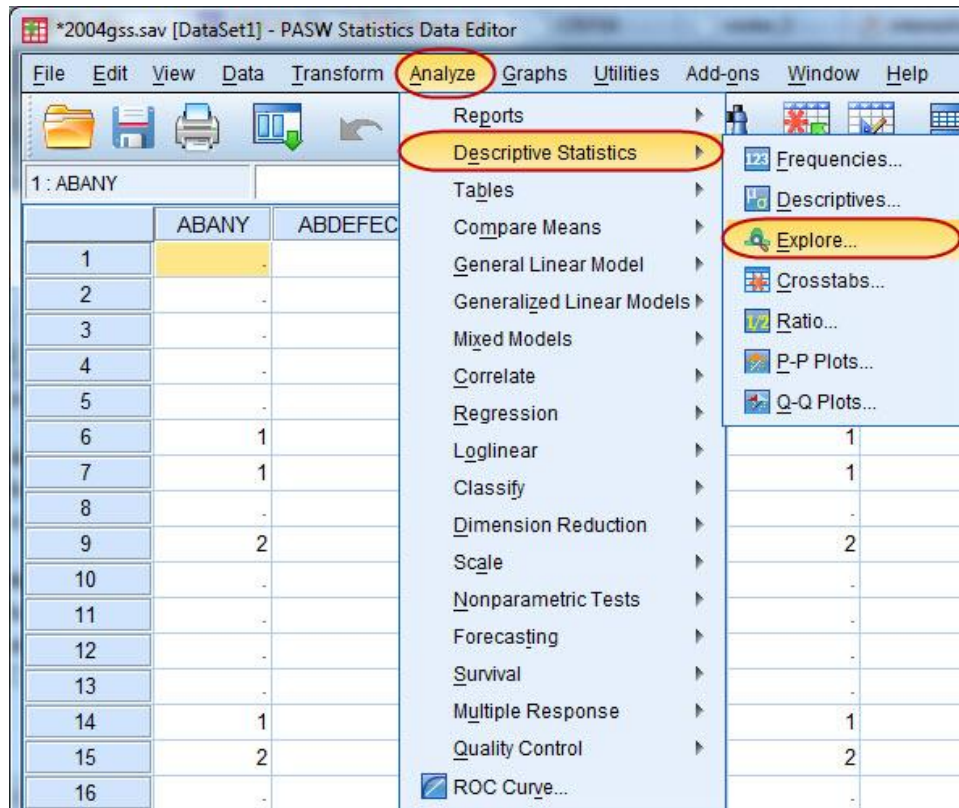


Figure 4- 16

Selecting Variables:

As in the other procedures, find and click the variable you want to explore, then click the select arrow to include your variable in the "Dependent List" box. Choose the variable "EDUC." The dialog box should look like Figure 4-17.

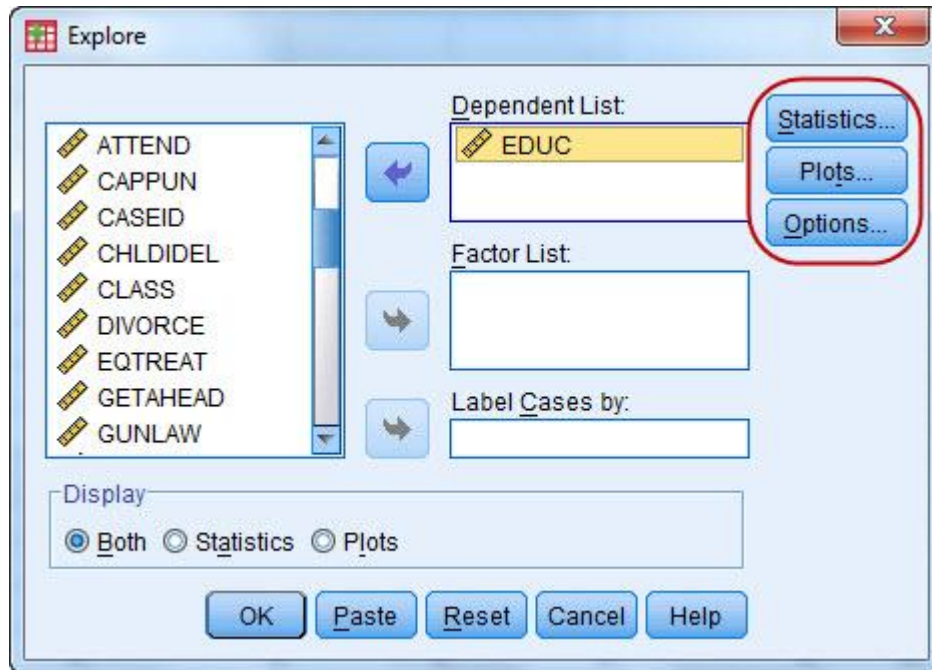


Figure 4- 17

Selecting Displays:

In the "Display" box on the bottom left, you may choose either "Both", "Statistics", or "Plots." We left the default selection, "Both" to display statistics and plots.

Selecting Statistics:

Click the "Statistics" button, upper right of Figure 4-17, and the "Explore: Statistics" dialog box will open, Figure 4-18.

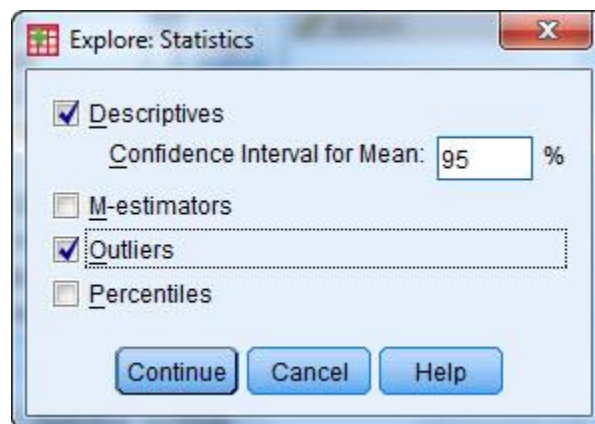


Figure 4- 18

Leave checked the default box for "Confidence Interval for the Mean 95%, " and click the "Outliers" box so we can look at the extreme observations for our variable. Click "Continue" to return to the main explore dialog window.

Selecting Plots:

Click the "Plots" button on the main Explore Dialog Box, Figure 4-17, and the "Explore: Plots" dialog box, Figure 4-19, will open.

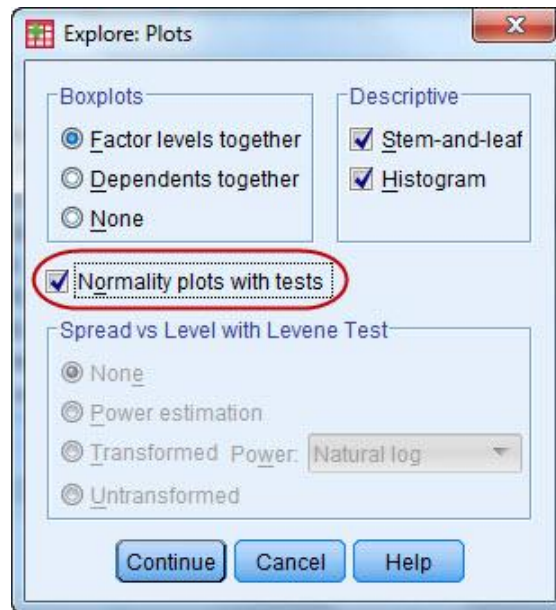


Figure 4- 19

Leave the default choices in the "Boxplots" box and then click "Stem and leaf" and "Histogram" in the "Descriptive" box. Click on "Normality Plots with Tests" so we can see how close the distribution of this variable is to normal. Leave the default for "Spread vs Level with Levene Test". Click "Continue" to return to the main explore dialog box.

Selecting Options:

Click the "Options" button in the main explore dialog box, Figure 4-17, and the "Explore: Options" dialog box, Figure 4-20, will be displayed.

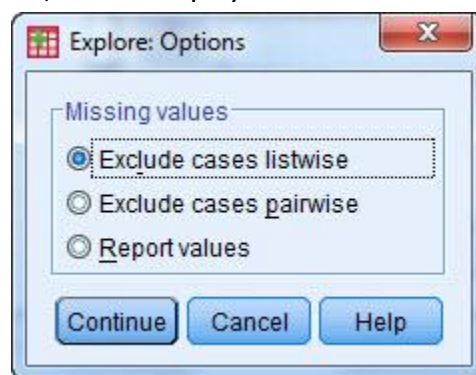


Figure 4- 20

No changes are needed here since the default of "Exclude cases listwise" is appropriate. Now click "Continue" to return to the main "Explore" dialog box, Figure 4-17. Click "OK" in the main "Explore" dialog box and PASW will perform the chosen tasks and display the data in the "PASW Statistics Viewer".

Interpretation of Explore Output:

Use the scroll bar to view any part of the output. The first part of the output is the "Case Processing Summary", Figure 4-21.

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
HIGHEST YEAR OF SCHOOL COMPLETED	1199	99.9%	1	.1%	1200	100.0%

Figure 4- 21

We can see that 1199 (99.9%) of our respondents answered this question. The other 1, .1% of the sample, was "Missing", not answering the question in this case. The GSS in recent years has had a split sample where not all respondents in the sample are asked the same questions. This is a question where all respondents were asked the question, so the total sample size was 1200 (100%).

"The "Descriptives" statistics output should look like Figure 4-22.

Descriptives					Statistic	Std. Error
HIGHEST YEAR OF SCHOOL COMPLETED	Mean				14.09	.080
	95% Confidence Interval for Mean	Lower Bound			13.94	
		Upper Bound			14.25	
	5% Trimmed Mean				14.15	
	Median				14.00	
	Variance				7.681	
	Std. Deviation				2.772	
	Minimum				0	
	Maximum				20	
	Range				20	
	Interquartile Range				4	
	Skewness				-.423	.071
	Kurtosis				1.587	.141

Figure 4- 22

We can see all the typical descriptive statistics on this output: mean (14.09), lower bound (13.94) and upper bound (14.25) for a 95% confidence of the mean (in polling terminology this says that we are 95% confident that the mean for the population is between 13.94 and 14.25), median (14.00), variance (7.68), standard deviation (2.77), minimum (0), maximum (20), range (20), inter quartile range (4.00), skewness (-.423 [negatively skewed]), kurtosis (1.587). A narrative explaining the education for the US population in 2004 would be somewhat like the following:

Our sample from the General Social Survey of 2004, indicates that the average education for those over 18 in the US in 2004 was a little over 14 years with a 95% confidence that the real average would fall between 13.94 and 14.25 years. The least years of education reported was found to be 0 and the most was 20. The exact middle point of the population with 50% falling below and 50% above, the median, was 14.00.

Extreme Values				
			Case Number	Value
HIGHEST YEAR OF SCHOOL COMPLETED	Highest	1	37	20
		2	63	20
		3	65	20
		4	101	20
		5	117	20 ^a
	Lowest	1	901	0
		2	528	0
		3	762	2
		4	380	2
		5	465	3

a. Only a partial list of cases with the value 20 are shown in the table of upper extremes.

Figure 4- 23

The "Extreme Values" can be seen in Figure 4-23. This Figure shows the five highest and the five lowest values for our variable. More than five respondents listed their years of education as 20-as you can read in footnote "a". On the low end, there were two with 0 education, two with 2 years of education and one with 3 years in our sample. The "Test of Normality" is shown next (see Figure 4-24). This shows that this distribution is significantly different from the expected normal distribution. "Sig." which is our p-value is .000, smaller than 0.05-therefore we reject our null hypothesis of the distribution being normal. This is a pretty stringent test, most researchers would not require the distribution to be this close to normality.

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
HIGHEST YEAR OF SCHOOL COMPLETED	.118	1199	.000	.954	1199	.000

a. Lilliefors Significance Correction

Figure 4- 24

The histogram, Figure 4-25, shows a rough bell shaped distribution. SPSS divided our distribution into nine groups with a width of 2.5 years of education for each group.

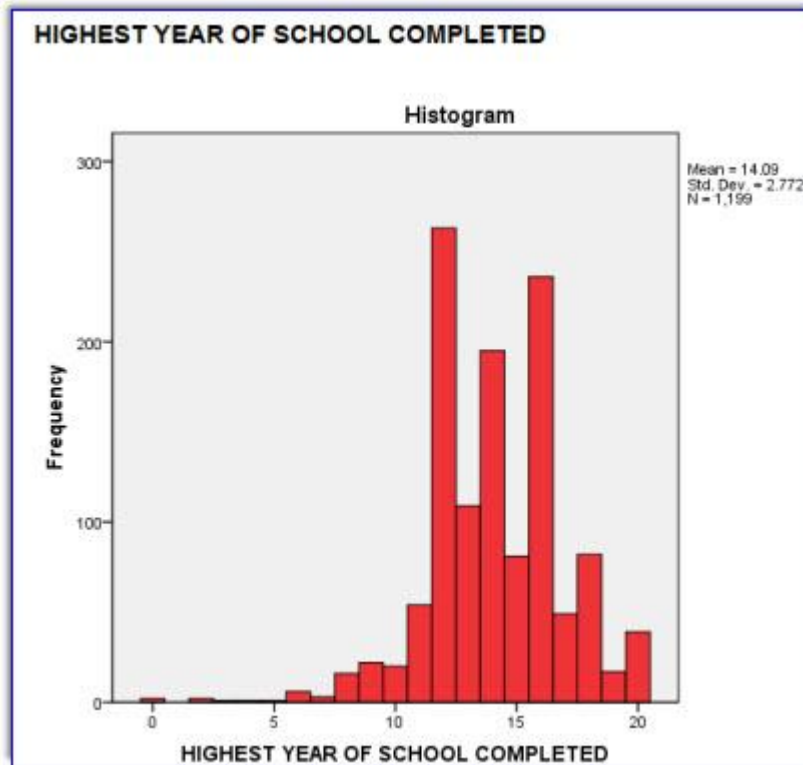


Figure 4- 25

The largest group has a little more than 250 cases, a visual estimate. The smallest group has very few cases (we know there are a number of respondents who reported 3 years of education from our Extreme Values and the 2.5 bar). The statistics on the histogram tell us that the standard deviation is 2.77 with a mean of 14.09 for a total N of 1199.

The Stem and Leaf is next. Figure 4-26, once again, shows a close but not quite normal distribution with significant outliers on the end of the distribution and a high number of observations above the mode. The stem and leaf is sort of a bar chart but instead of being vertical is a horizontal representation of values of the sample.

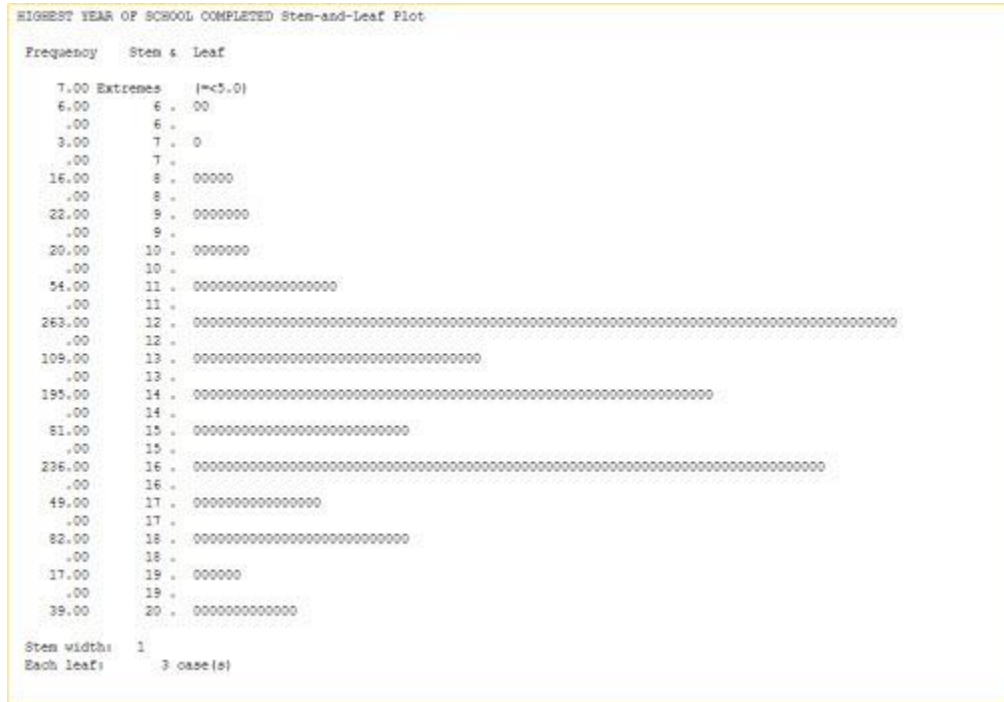


Figure 4- 26

Interpretation of the Q Q Plot of Age:

Continue scrolling down the "PASW Statistics Viewer" to the "Normal Q Q Plot of HIGHEST YEAR OF SCHOOL COMPLETED" (see Figure 4-27).

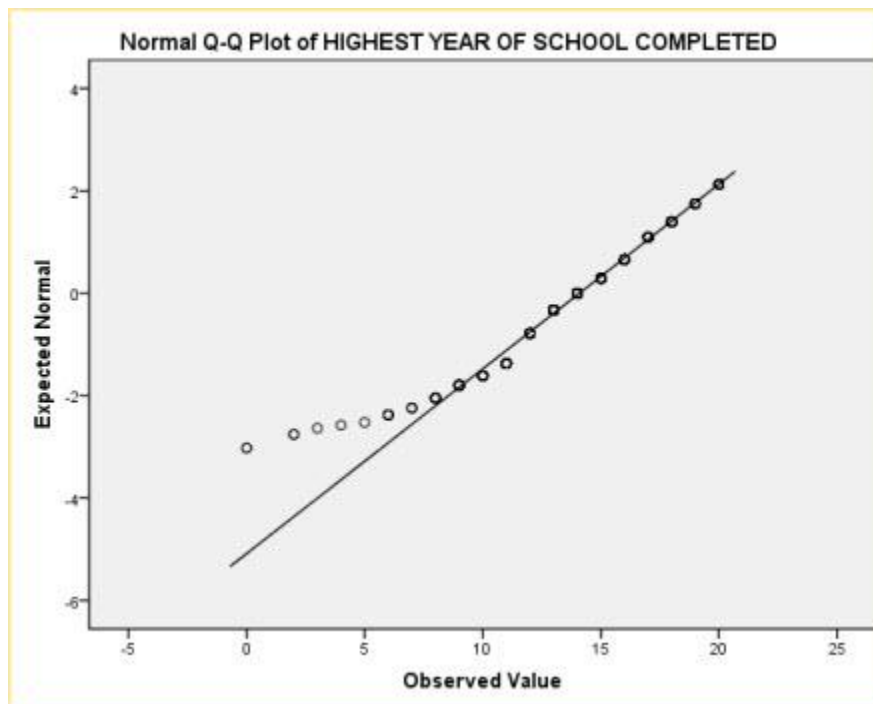


Figure 4- 27

A Q-Q plot charts 'observed values' against a known distribution, in this case a normal distribution. If our distribution is normal, the plot would have observations distributed closely around the straight line. In Figure 4-27, the expected normal distribution is the straight line and the line of little boxes is the observed values from our data. Our plot shows the distribution deviates somewhat from normality at the low end. The high end of the distribution is pretty much normal.

The Detrended Normal Q-Q plot, shows the differences between the observed and expected values of a normal distribution. If the distribution is normal, the points should cluster in a horizontal band around zero with no pattern. Figure 4-28, of HIGHEST YEAR OF SCHOOL COMPLETED, indicates some deviation from normal especially at the lower end. Our overall conclusion is that this distribution is not normal. Many researchers would see this as close enough to treat as a normal distribution.

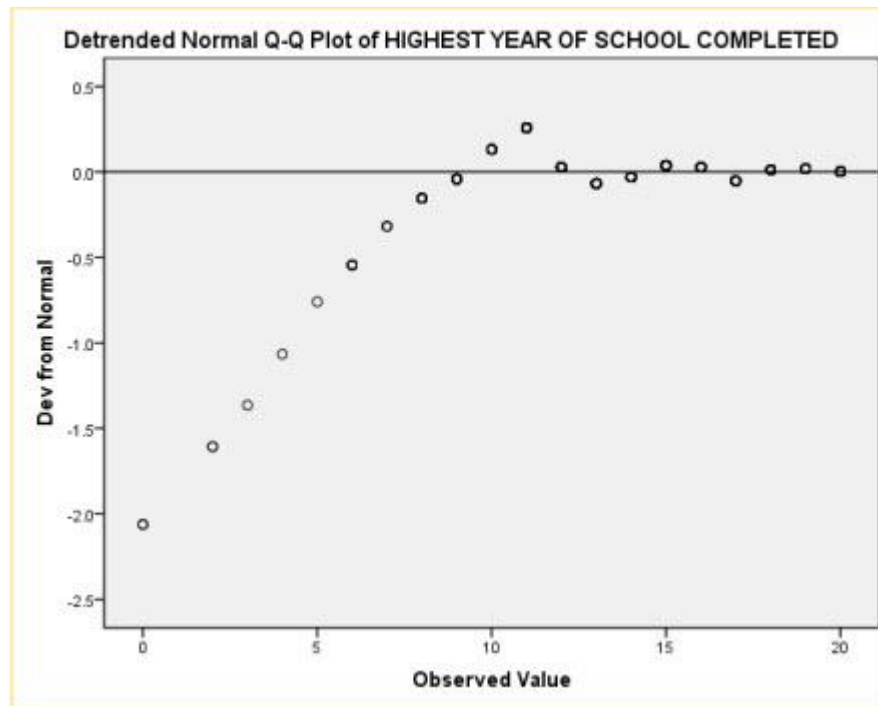


Figure 4- 28

Interpretation of the Boxplot:

In the PASW Statistics Viewer, scroll to the boxplot of HIGHEST YEAR OF SCHOOL COMPLETED. The boxplot should look like Figure 4-29.

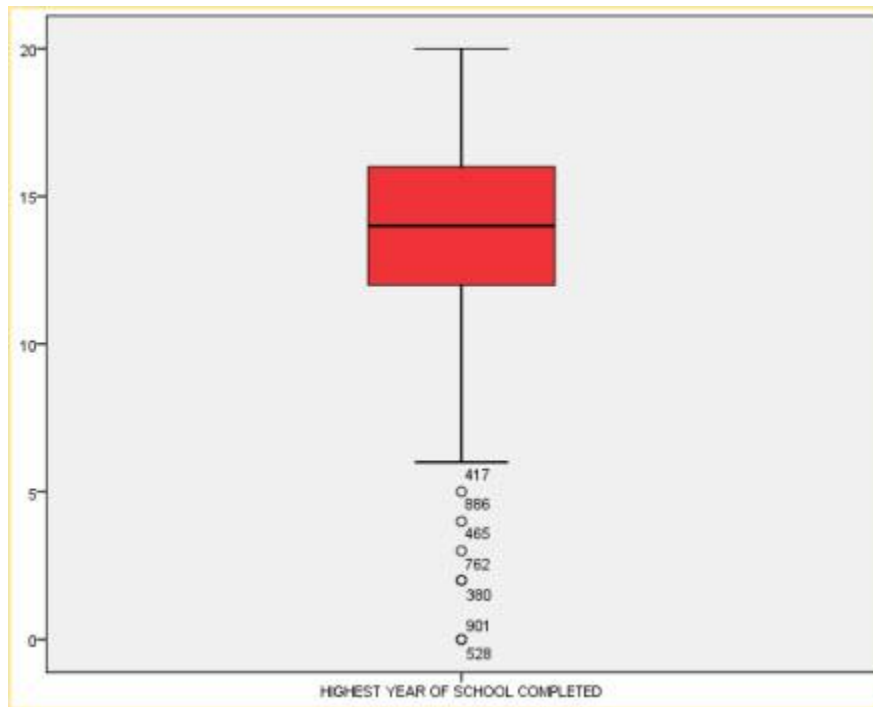


Figure 4- 29

Once again the major part of our distribution is not normal and there are significant outliers, the cases beyond the lower line of our boxplot. Our outliers are at the lowest end of the distribution, people with little or no education. There are also more observations above than below the mode.

Conclusion

In performing univariate analysis the level of measurement and the resulting distribution determine appropriate analysis as well as further multivariate analysis with the variables studied. The specific output from SPSS one uses in a report is chosen to clearly display the distribution and central tendencies of the variables analyzed. Sometimes you report a particular output to enable comparison with other studies. In any case, choose the minimal output that best accomplishes this goal. Don't report every SPSS output you obtained.

Univariate Analysis as Your First Step in Analysis

Why do univariate analysis as your first step in data analysis? There are five reasons:

1. As discussed at the beginning of this chapter, the frequency distribution may actually be all you are interested in. You may be doing research for people with little statistical background and/or they are really only interested in the percentage or count of people that said "Yes" or "No" to some question.
2. You can check for "dirty" data. Dirty data is incorrectly entered data. "Data cleaning" is correcting these errors. Remember, each case must have an ID number. One primary reason for the ID number is to help us clean our data in case there are data entry errors. One way to do this is by determining when there are codes in the data outside the range of the question asked and determining which cases, the ID number, is in error. You can then check all the way back to the original questionnaire and correct the entry or if that's not possible change the erroneous code to the "Missing values" code.

An example might be if you had a question in a questionnaire where responses were coded in the following way:

- 1 is the code used for "Strongly Agree"
- 2 is the code used for "Agree"
- 3 is the code used for "Neutral"
- 4 is the code used for "Disagree"
- 5 is the code used for "Strongly Disagree"

But suppose you run a frequency distribution and find that two respondents have a code of "6." That wasn't one of the codes! What happened? Your data entry person, who may have been you, hit the 6 on the keyboard instead of some other number. We can correct this error. In fact, when we locate this error, we may find others because often errors occur in streaks. The data entry person gets something out of order, or they get their fingers on the wrong keys. These problems can happen to any of us. The trick is to correct the errors as best possible.

You can have SPSS for Windows select only those cases that have the code of "6" for that variable, and then tell it to do a Frequencies on the variable ID. This will tell you the case number(s) that has the error and you can correct it. Be sure to double check the codes, before and after, to make sure they are correct.

3. A third reason for running Frequencies on your variables as your first step in analysis is that you can tell if you need to combine categories, and if so, what codes should be combined. You would know if there were too few respondents giving "Strongly Agree" or "Strongly Disagree" and for analysis they should be folded into either "Agree" or "Disagree". Another common combination of categories is for age groups. For example you would do this if you wanted to compare age groups born before and after a significant event (i.e. those born before Vietnam compared to those born after Vietnam).
4. Related to number 3 is that you can find if everything that should be defined as "Missing" is actually defined as missing. For example, if you find that 8 "Don't Know" is a response that has been left in your calculations, your analysis will include all of the eights. Even your mean statistics will have these "extra" eight's included in the calculation. You need to go into the definition of the variable and make these codes "Missing values" or recode these so they are not included, say as a "System Missing" value.
5. Finally, you may want to examine the distributions for your variables. This should help you determine characteristics of your sample, make some conclusions and decide further steps in your analysis. You might find that in a 1-5 agree/disagree question, discussed in step 2 above, almost everyone disagreed. You may discover you do not have a normal distribution and may decide that you want to "fix" the distribution using various transformation techniques to convert the data into a normal distribution. These and related techniques are often referred to as "exploratory data analysis" and are beyond the scope of this text.