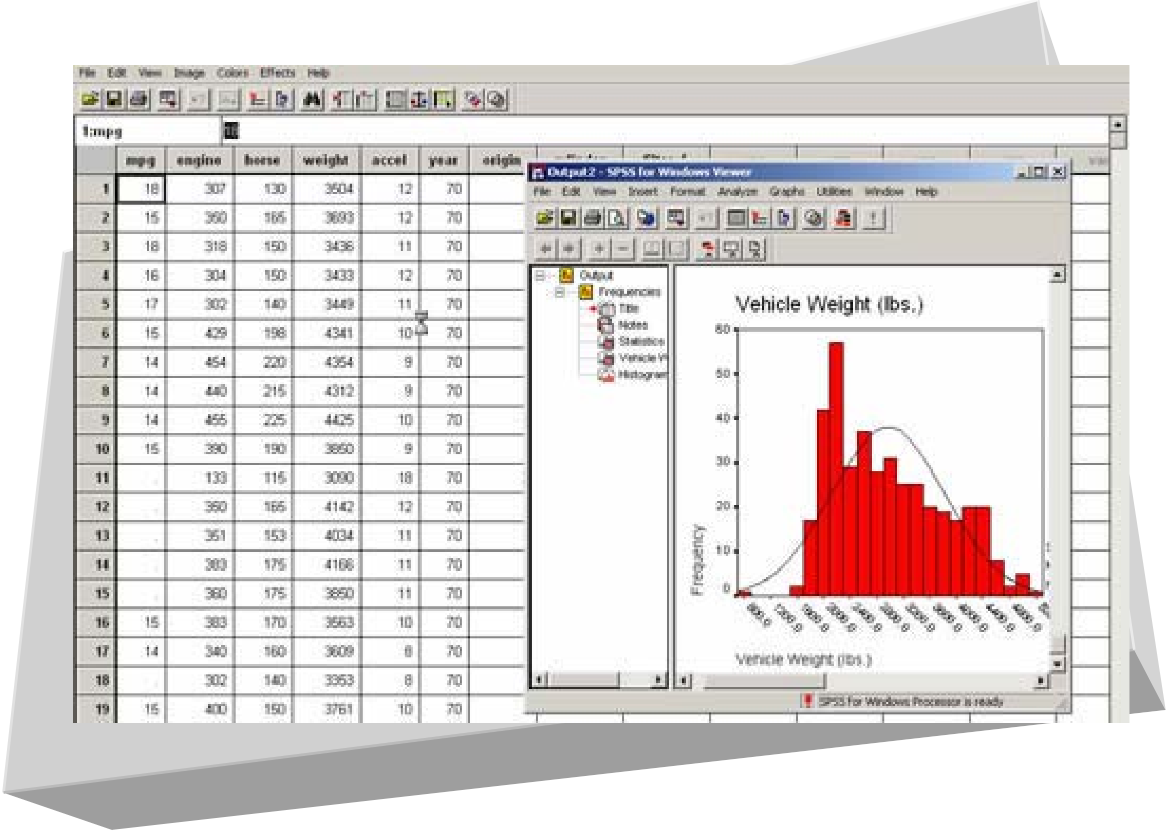
Center *for* Marketing Technology



	mpg	engine	horse	weight	accel	year	origin
1	18	307	130	3504	12	70	
2	15	350	165	3693	12	70	
3	18	318	150	3436	11	70	
4	16	304	150	3433	12	70	
5	17	302	140	3449	11	70	
6	15	429	198	4341	10	70	
7	14	454	220	4354	9	70	
8	14	440	215	4312	9	70	
9	14	455	225	4425	10	70	
10	15	390	190	3850	9	70	
11	.	133	115	3090	18	70	
12	.	350	165	4142	12	70	
13	.	351	153	4034	11	70	
14	.	383	175	4166	11	70	
15	.	360	175	3850	11	70	
16	15	383	170	3563	10	70	
17	14	340	160	3609	8	70	
18	.	302	140	3353	8	70	
19	15	400	150	3761	10	70	

SPSS
Data Mining/Data Analysis

Center for Marketing Technology, MOR 220, cmt@bentley.edu

Table of Contents

1. SPSS Overview

SPSS (Statistical Package for the Social Sciences) is a data mining tool that covers a broad range of statistical procedures that allow you to summarize data (compute means and standard deviations), determine whether there are significant differences between groups (t-tests, analysis of variance), examine relationships among variables (correlation, multiple regression), and graph results (e.g., bar charts, line graphs). In practice, it offers companies information about the most and least profitable customers, their purchase patterns, buying behaviors and demographic profiles, which are key to developing a successful marketing strategy.

2. Loading SPSS on your computer

In order to access SPSS from your computer, you must first load the SPSS shell onto your computer. Once the shell is loaded, you can access SPSS **as long as you are connected to the network**. To get the shell loaded onto your computer do the following:

1. Click on the "Bentley College Virtual Lab Icon"
2. Once in the Virtual Lab, click on the "Tools" icon
3. Once in the Tools window, click on the "SPSS Version11 Base Setup." Follow the instructions given in the setup

The above procedure should give you an SPSS icon in your start menu. However, if for some reason it does not work try the following:

Click on the “Start” menu, click "Run" then type in:

\\electra\spss$\spss11\Setup\setup.exe

3. Using SPSS

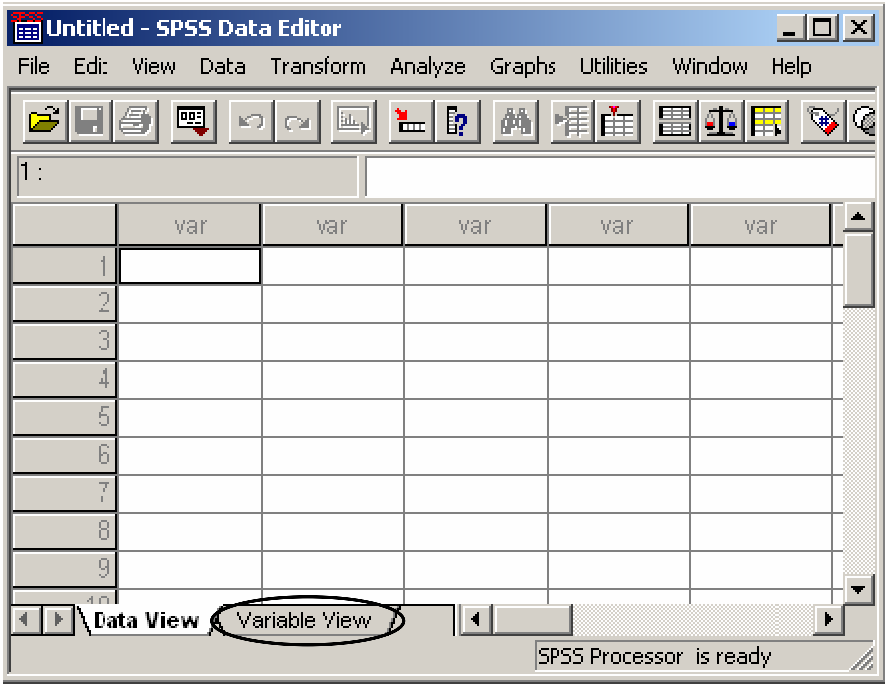
3.1. Startup Choices

Start SPSS by clicking on Start ⇨ Programs ⇨ SPSS 11.0 for Windows. Once the program launches, you will be presented with a few choices:

1. Run the tutorial
2. Type in data -- use this option to create and enter new data file
3. Run an existing query -- use this option to run syntax
4. Create new query using Database Wizard -- use this option to import data from other formats i.e. excel or access
5. Open an existing data source -- use this option to open a data file you have previously created in SPSS
6. Open another type of file – use this option to open a non-SPSS data file such as Excel

3.2. Entering Data on SPSS

1. In the Data Editor Window, each row is a case (or observation) and each column is a variable.



2. The cells in each row hold the value of a particular case (or observation) for a particular variable. At the top of the every column is the variable name – which can be no more than 8 characters long
3. Before filling in the observation values, you should define the variables you will be working with – see section below for thorough explanation
4. Once variables have been defined and you are ready to enter values, use the "Enter" key to move to the cell on the row below or the "Tab" key to move to the cell to the right on the same row

3.3. Defining Variables

1. Double-click on the first column heading (where it says "var") or click on the "Variables View" tab at the bottom of the page
2. To define a variable, you must complete the following fields:

   a) Variable Name - type the desired variable name, which can be *no more than 8 characters in length*. The first character must be alphabetic; the remaining characters can be alphabetic and/or numeric, and no spaces can appear in the name
   b) Type - allows you to define the type of variable, which can be:
   (Click on the ... icon to view the following choices)

i) Numeric - observations are numbers
ii) Comma - observations are displayed with commas delimiting every three places, and with the period as a decimal delimiter
iii) Dot - observations are displayed with periods delimiting every three places, and with the comma as a decimal delimiter
iv) Scientific notation - observations are displayed with an imbedded E and a signed power-of-ten exponent (for example, 123, 1.23E2)
v) Date - observations are displayed in one of several date formats (select one from the list)
vi) Dollar - observations are displayed in one of several formats (select one from the list)
vii) Custom currency - observations are displayed in one of the custom currency formats that you define by going to the top menu under Edit | Options | Currency tab
viii) String - observations are non-numeric, they can contain any characters up to the defined length

c) Width and Decimals – use these fields to limit the number of characters and define how numbers should be displayed

d) Label - assign descriptive variable and value labels

i) Variable Label - allows you to list a more comprehensive label for your variable that is not limited to 8 characters and may have spaces
ii) Value Label - allows you to provide labels for the various levels of a variable (click on the ... icon to proceed)
iii) For non-numeric variables, such as gender, it is very useful to set values such as 1=male, 2=female so that they are considered numeric after all and included in the calculations/data analysis
iv) Do so if appropriate, by placing the value (e.g., "1") in the Value field and the value label (e.g., "male") in the Value Label field and clicking "Add." You can do this process as many times as necessary for each string variable

e) Missing Values - enables you to designate certain observation scores as missing
f) Columns and Align - these fields allow you to change the maximum number of characters permitted in a column as well as determine the alignment of the text
g) Measurement - allows you to determine the kind of measure for that particular variable. It can be:

i) Scale - numeric data on an interval or ratio scale
ii) Ordinal – string and numerical variables with defined value labels
iii) Nominal - mostly string variables

A nice shortcut to label several variables at once is to use a global template.

1. First, in the Data window, highlight the columns that you want to have the same labels. Unfortunately, SPSS only allows you to highlight columns that are next to each other
2. Click on the "Data" menu. Click on "Templates"
3. In the "Apply" area, click on "Value Labels"

4. Click on "Define" In the window that appears, and then follow the same instructions given above

3.4. Computing a New Variable

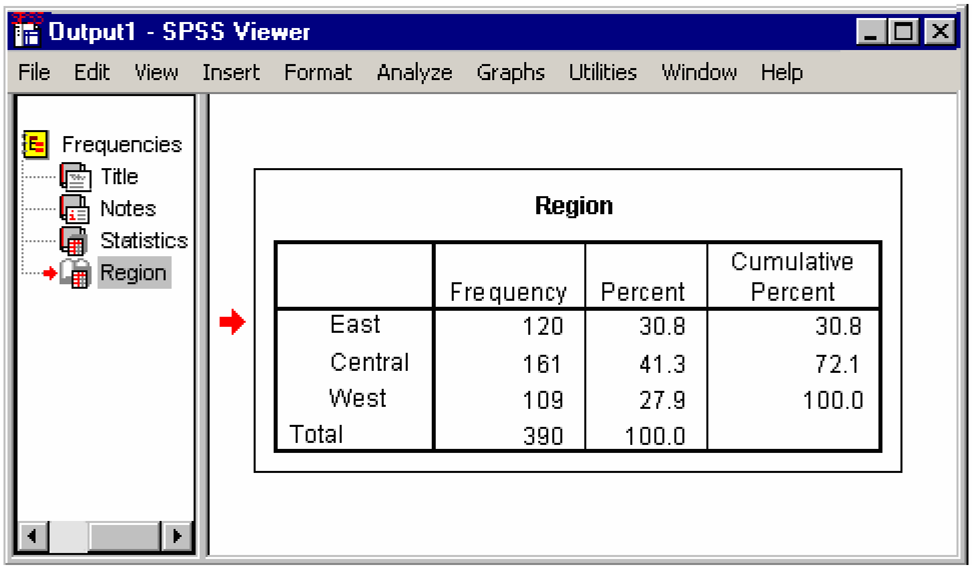
New variables can be created by manipulating the existing data set. For example, you might have weekly data but would like to analyze this same data in monthly time periods.

1. Click "Transform" on the top menu line then click "Compute" from the drop down menu.
2. In the “Target Variable” dialogue box, enter the name of the new variable (e.g., "Monthly").
3. Select the variable (“Weekly” in this case) you want to compute for by highlighting it and clicking on the arrow to place it into the “Numeric Expression box.”
4. Select the appropriate formula for the computation by using the keypad or pre-defined functions. Continuing with our example, click "*" and then "4."
5. Click "OK" and SPSS will compute a new variable called “Monthly” by multiplying each obervation under the variable “Weekly” by 4

3.5. SPSS Windows

When using SPSS, you will have two types of windows open, the Data Editor window (as illustrated previously), where the data set is displayed and the Output Viewer window.

The Output Viewer window is where you will always find your output (i.e. test results) and any messages or commands from SPSS.



Region

	Frequency	Percent	Cumulative Percent
East	120	30.8	30.8
Central	161	41.3	72.1
West	109	27.9	100.0
Total	390	100.0	

To get back to your Data window, you can click on the menu option titled Window, and then select your data file. Or, you can click on the little button that looks like a data matrix. Whenever you conduct an analysis procedure (e.g. get a frequency distribution), SPSS will automatically bring up the Output Viewer window. You do not need to go back to the Data Editor window to run a different analysis on the same data set.

3.6. Saving

You can save any of the windows just like you would on any other program by clicking on the disk icon or going to the file menu. To save a window, make that particular window active by clicking on it. SPSS automatically adds a three-letter suffix to the end of the file name (".sav" for data editor files, ".spo" for output files). Thus, with a particular data set, it is recommended to use a single name for both files.

Make sure you save it to a disk and a directory that you will be able to access whenever needed. Save on a regular basis.

3.7. Printing

Be careful printing in SPSS. Unless you tell it otherwise, SPSS will print everything you have done up to that point. It is better to always highlight the output you wish to print and then tell SPSS to only print the selected areas. To select the areas you wish to print, SPSS offers several options. One is to choose Edit from the menu, then Select, then Last Output. This selects your most recent analysis. Another way to select output for printing is to use the branching diagrarn that appears on the left part of the Output window. By clicking on a part of a branch in the diagram, this part of the actual output becomes selected and then when you print, only this part will be printed.

4. SPSS and DATA ANALYSIS

4.1. Purpose of Statistical Analysis

To calculate statistics or create graphs, you must first select the appropriate procedure from the "Analyze" menu on the top toolbar. The drop down menu will give you many options (some of which are outlined below). **Please note:** SPSS does a lot more than what's described here. Use the SPSS help menus at any time to find out more.

Statistical Analysis	Procedures Available	Description
Descriptive Statistics	Frequencies	Gives you frequency counts. Useful for summarizing the number of occurrences and percentages of observations under each variable.
	Descriptives	Gives you various descriptive statistics such as means, standard deviations, minimum and maximum values.
	Crosstabs	Calculates frequency tables combining two variables. A Chi-square test may be added to identify whether or not the difference in frequency counts are statistically significant or not.
Compare Means	Means	Get means (and other statistics) for the entire data set or for subsets of the data (e.g., an average GPA for males and for females, as opposed to the overall average).
	One-Sample T-Test	Test if the mean of some variable is significantly different from some hypothesized number.

	Independent-Samples T-Test	Test if the means of two groups are significantly different.
	One-way ANOVA	Determine if categorical independent variables have an effect on a dependent variable.
Correlate	Bivariate	Determine if two variables are positively or negatively correlated such that they "move together" or "in opposite directions"
Regression	Linear	Calculate a "line" that will use one or more independent variables to predict the value of a dependent variable.

The choice of which procedure to run really depends on the type of data you are analyzing and what exactly you are trying to find out.

There are two main types of data: categorical and continuous. Categorical data represents types of data that can be divided into groups. Examples include race, sex, age group, and educational level. Continuous data represents types of data that are associated with some sort of measurement. Examples include dollar amounts, time and units. Briefly, frequency tables and crosstabs are indicated for categorical data and descriptives and correlations are indicated for continuous data.

4.2. How to run a statistical analysis

The following analyses are based on customer surveys conducted by "PC's Unlimited" Catalog Company. This dataset is available at the CMT for practice purposes. To open this file, click on the "My Computer" icon on the desktop, then click on ***C Drive (C:) / Data/ mk 721/ PC'sUnlimited.sav***

1. Frequencies

The "Frequencies" procedure is useful for summarizing the number and percent of observations that have each value of a categorical variable. For example, the PCsUnlimited's dataset includes a variable called REGION indicating which REGION each customer lives in.

To determine out how many customers are from each REGION:

1. From the pull-down menus, select ***Analyze / Descriptive Statistics / Frequencies***
2. In the dialog box that appears, select REGION for the variable list (click on REGION, then click on the arrow to move it to the Variable List)
3. Make sure that the "Display Frequency Tables" box contains a check
4. Click on the ***Charts*** button. Select "None"
5. Click "Continue"
6. The ***Statistics*** and ***Format*** buttons allow you to change the output produced. For this example, no changes from the default values are required
7. Click "OK". The Output window will display both the number and percentage of customers with different values for the REGION variable

REGION

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	NorthEast	227	22.7	22.7	22.7
	MidAtlantic	167	16.7	16.7	39.4
	SouthEast	81	8.1	8.1	47.5
	MidWest	263	26.3	26.3	73.8
	SouthCentral	98	9.8	9.8	83.6
	West	164	16.4	16.4	100.0
	Total	1000	100.0	100.0	

The Table above is an output of the Frequency run on the Region variable of PCs Unlimited, it clearly shows the customer distribution of the company. The highest being in Midwest region, followed by Northeast and so on.

2. Descriptives

The "Descriptive" procedure produces summary statistics such as mean, standard deviation, min and max values, sum and range. These statistics are useful for variables measured on a continuous scale – such as dollars, years, units, pounds, or meters.

Suppose you want to know the total dollars spent by all customers and the average dollars spent per customer as well as the largest and smallest amounts spent by any customer. You also want the same summary statistics for number of purchases

1. From the pull-down menus, select ***Analyze / Descriptive Statistics / Descriptive***
2. In the dialog box that will appear, select TOTDOL and NUMPURCH for the variable list
3. Click on the ***Options*** button. Check the "Mean," "Sum," "Minimum," and "Maximum" boxes, and remove the checks from the remaining boxes
4. Click "Continue"
5. Click "OK". The Output window will display following result:

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean
Total Dollars Spent	1000	50	8763	335229	335.23
Total Number of Purchases	1000	1	15	1822	1.82
Valid N (listwise)	1000				

The Table above is the result after a Descriptive Statistics is run; it shows the Number of Observations (N) as 1000. Since we are measuring the $ spent, we have the result showing that the Minimum $ spent by a customer is $50 and Maximum is $8763, the Average (Mean) spending by a customer is $335.23. Similar interpretation is applicable for Number of Purchases.

3. Crosstabs

By forming two-way (or multi-way) tables, the "Crosstabs" procedure count records for two (or more) categorical variables. In addition, statistical tests (including the commonly used chi-square test) can be used to test for significant associations between categorical variables.

For example, if you want to know what percentage of females vs. males responded to the latest offering from PCsUnlimited, we need to calculate four percentages: (1) % of females who responded; (2) % of females who did not respond; (3) % of males who responded; and (4) % of males who did not respond.

(1) % Of females who responded
(2) % Of females who did not respond
(3) % Of males who responded and
(4) % Of males who did not respond

1. From the pull-down menus, select ***Analyze / Descriptive Statistics / Crosstabs***
2. In the dialog box that appears, select GENDER for the Row variable and REPOND for the column variable
3. Click "Cells." Click on **"Row percentages"**. This will compute the percentage of males and females who responded
4. Click "Continue" and then click "OK" to produce the crosstabs listing:

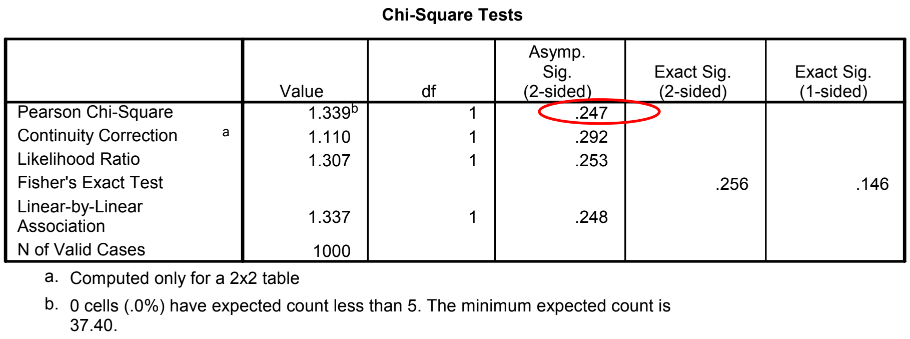
GENDER * RESPOND Crosstabulation

			RESPOND		
			no	yes	Total
GENDER	female	Count	232	43	275
		% within GENDER	84.4%	15.6%	100.0%
	male	Count	632	93	725
		% within GENDER	87.2%	12.8%	100.0%
Total		Count	864	136	1000
		% within GENDER	86.4%	13.6%	100.0%

The Table above gives us the % of responders and non-responders by **Rows** that is why we selected the **Row %** option. The table can be interpreted as out of 275 females only 43 (15.6%) responded and 232 (84.4%) did not; where as out of 725 males 93 (12.8%) responded and the rest 632 (87.2%) did not.

Now, suppose we want to know whether the higher percentage of females who responded to an offer (15.6%) than males (12.8%) is statistically significant. We can repeat the analysis adding the following:

1. Click on the "Statistics" box at the bottom of the Cross-tabs dialog box
2. In the dialog box that appears, click on Chi-square to produce the following:



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.339[b]	1	.247		
Continuity Correction[a]	1.110	1	.292		
Likelihood Ratio	1.307	1	.253		
Fisher's Exact Test				.256	.146
Linear-by-Linear Association	1.337	1	.248		
N of Valid Cases	1000				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 37.40.

The p-value for the **Pearson Chi-square** [1] **is .247**, which is well above the **typical .05** cutoff, indicating that the difference in response rate between males and females is **not statistically significant.**

Finally, suppose that instead of finding the % of females vs. males who responded, we want to know the % of responders and non-responders who were male or female. In other words, we want to know:
(1) % Of responders who are female
(2) % Of responders who are male
(3) % Of non-responders who are female and
(4) % Of non-responders who are male

To calculate these values, use the above procedure, with one small change: Instead of clicking on "Row percentages," click on "Column percentages". You can also create a crosstabs listing with **both** row and column percentages, though some find this confusing!

GENDER * RESPOND Crosstabulation

			RESPOND		Total
			no	yes	
GENDER	female	Count	232	43	275
		% within RESPOND	26.9%	31.6%	27.5%
	male	Count	632	93	725
		% within RESPOND	73.1%	68.4%	72.5%
Total		Count	864	136	1000
		% within RESPOND	100.0%	100.0%	100.0%

The Table above shows that there are 275 (27.5%) females in the database. Out of 136 responses 43 were female (31.6%) and 93 (68.4%) were males and similar interpretation for the 'no'.

It doesn't matter which variable is the row variable and which is the column variable – although you will need to think through whether you want to request row or column percentages depending on which is your row and which is your column variable.

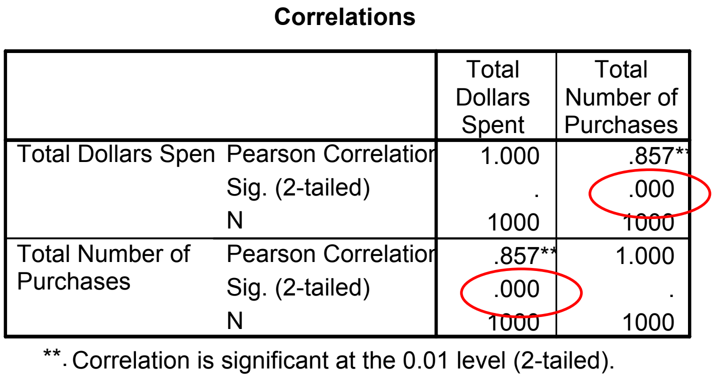
[1] The Chi-Squared is a "Goodness of Fit" Test. The Chi-Square Test allows the comparison of two variables in a sample of data to determine if there is any relationship between them. Lower the Chi-Square level more significant the relationship and vice-versa. A standard range accepted for Chi-Square level is between .000 - .05.

Hint: if you are cross-tabbing two variables and one has many categories – it is best to make that one the row variable. For example, consider a cross-tabs of gender by profession (and you have 25 professions). That will either produce a table with 2 columns and 25 rows or one with 25 columns and 2 rows – and the first (2 columns and 25 rows) will fit on a sheet of paper much better!

4. Bivariate correlation

One measure of the relationship between continuous variables is their correlation. To compute the correlation between total dollar purchases and number of purchases:

1. From the pull-down menus, select ***Analyze / Correlate / Bivariate***
2. Select NUMPURCH and TOTDOL for the Variables list
3. Under "Correlation Coefficients," select "Pearson"
4. If you want statistically significant correlations marked with an asterisk, select "Flag significant correlations
5. Click "OK" to produce the output:



Correlations

		Total Dollars Spent	Total Number of Purchases
Total Dollars Spen	Pearson Correlatior	1.000	.857**
	Sig. (2-tailed)	.	.000
	N	1000	1000
Total Number of Purchases	Pearson Correlatior	.857**	1.000
	Sig. (2-tailed)	.000	.
	N	1000	1000

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

The Table is showing the correlation between total dollar and total purchases. The only thing of importance here is that **Sig. (2-tailed)** is .000, which means that there is a significant correlation between the two variables. The Sig. (2-tailed) should not be more that .01

4.3. Graphs

1. Histogram

The "Descriptive" procedure gives us information about the distribution of the values of a variable. As we saw, we can find the minimum and maximum values and the mean; we can also display the variance and standard deviation, to give us some idea of how tightly values are clustered around the mean.

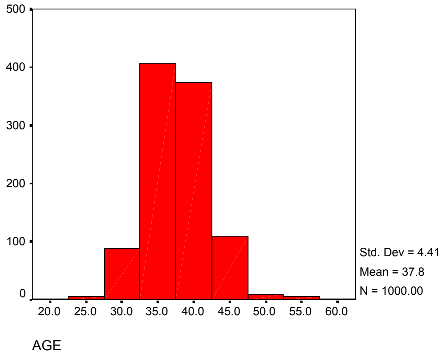
It is often more useful, though, to see a histogram showing the distribution of the values. There are two ways to create histograms in SPSS:

(1) Using the histogram option under the Charts menu, or
(2) With the "Frequencies" procedure under the Analyze menu

The following examples show how to create a histogram showing the distribution of customer ages.

(1) Using histogram option:

1. Open the dataset in SPSS
2. From the pull-down menus, select ***Graphs / Histogram***
3. In the dialog box that will appear, select AGE for the variable
4. Click on the ***Titles*** button if you want to add a title
5. Click "OK". A histogram showing the distribution of AGE will appear in the Output window. The mean and standard deviation for AGE is also printed:



(2) Using the frequencies procedure (this option allows you to get frequency tables as well as a histogram):

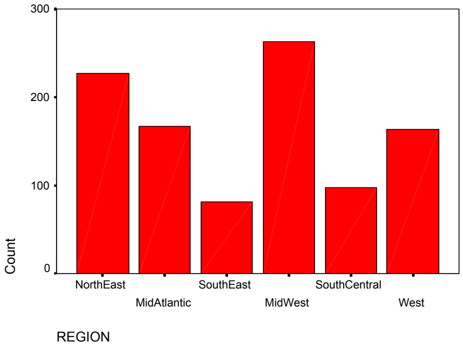
1. From the pull-down menus, select ***Analyze / Descriptive Statistics / Frequencies***
2. In the dialog box that appears, select AGE for the variable list
3. Click on the ***Charts*** button. Select "Histograms"
4. Click "Continue"
5. The "Display Frequency Tables" box is probably checked. Click this box to "un-select" this option (if you forget to do this you will get a long – and probably useless – frequency table)
6. The ***Statistics*** and ***Format*** buttons allow you to change the output produced.
7. Click "OK". A frequency histogram showing the distribution of AGE will appear in the Output window. As before, the mean and standard deviation for AGE will also be printed

2. <u>Bar Charts</u>

Bar charts are commonly used to provide a visual summary for categorical data. For example, to create a bar chart showing the number of customers in each region:

1. From the pull-down menus, select ***Graphs / Bar***
2. In the dialog box that appears, select SIMPLE
3. Click on the "Summaries for groups of cases" button and click on *Define*
4. Select REGION from the variable list and click on the arrow next to the *Category Axis* box
5. Click on the "N of cases" button under "Bars Represent"
6. (Optional) Click on the Titles box and enter a title

Click "OK" to create this chart:



5. Helpful Hints

5.1.Analyzing a Subset

In some cases, you may only want to analyze a subset of the total data collected. For example, you may only be interested in women's opinions on a certain matter. You can tell SPSS to only analyze that group.

1. Click "Data" on the top menu line, then click "Select Cases" from the drop down menu.
2. In the resulting dialogue box, click the "if condition is satisfied" check box.
3. Click the "If' button. In the resulting dialogue box, highlight the variable that you want to split into a subset and click the arrow.
    - Complete the If Statement. For example, if females are coded as "1 " type or click "= 1" Therefore you should have a statement that reads, gender (if that's how you named the variable) = 1.
    - Click "Continue."
    - Click "OK."
    - **VERY IMPORTANT -** this restriction will remain in place until you go through this process again and click the "All Cases" check box.

5.2. Recoding a variable into a new variable

Suppose you want to divide your customers into two groups: one-time purchases and repeat customers or those who have purchased more than once.

An easy way to do this in SPSS is to use the ***Transform / Recode/ Into Different Variables*** to create a new Yes/No variable:

1. Make sure you are in the Data Editor window.
2. From the pull-down menus, select ***Transform / Recode / Into Different Variables.***
3. Select NUMPURCH for "Input Variable → Output Variable."
4. Under 'Output variable,' type a name for the new variable (perhaps REPEAT) and a brief description. Click "Change."
5. The "Numeric Variable → Output Variable" box should show:
   NUMPURCH → REPEAT
6. Click the "Old and New Values" box.
7. Click the "Output Variables are Strings" box on the lower right-hand side of the dialog box.
8. Under "Old Value," click "Value" and type a 1 in the box.
9. In the "New Value" box, type No.
10. Click "Add." This specification will set REPEAT to 'No' whenever NUMPURCH is equal to one (i.e., the customer made only one purchase
11. Under "Old Value" click "Range: ___ through highest" and type a 2 in the box.
12. In the "New Value" box, type Yes.
13. Click "Add." This specification will set REPEAT to 'Yes' whenever NUMPURCH is 2 or greater (i.e., the customer has made 2 or more purchases).

14. The “Old→New” box should show two lines:
    1→’No’
    2 through Highest→’Yes’
15. Click “Continue.”
16. Click “OK.” The new variables will be added. Depending on the file size, this can take from a few seconds to several minutes.

You can now run a Crosstabs to see how many one-time versus repeat customers responded to the latest offer. If you want to keep your new variables for future use, remember to save the SPSS dataset!

5.3. Transferring Data from Excel into SPSS

To open a file in SPSS from Excel:

1. In Excel, make sure the first row includes the names of each variable (e.g., Name, Gender, etc..).Format the first row as text.
2. In Excel, format all the numbers as numbers and any text as text.
3. Save and close the file.
4. In SPSS, go to File⇨Open. Select the folder to “Look in” and choose File Type - Excel (*.xls).
5. Choose your file to open. Click Open.
6. A window will open. Click on "Read variable names"
7. Click OK and the file will open on SPSS

5.4. Conclusion

Understanding SPSS is generally a matter of just sitting down and using it for awhile. Yes, it may be a little painful at first, but after a short time, you will be able to use the program fairly well. *Remember*, the help menus in SPSS are extremely useful.