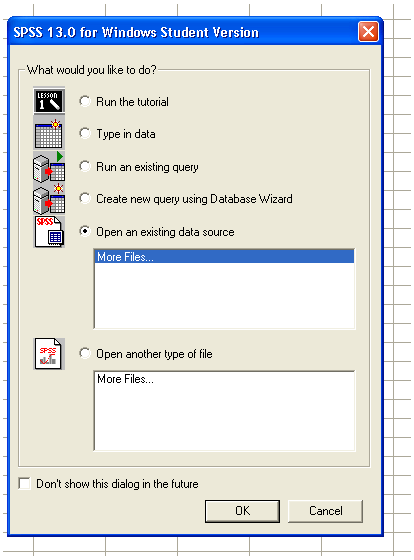
Table of Contents

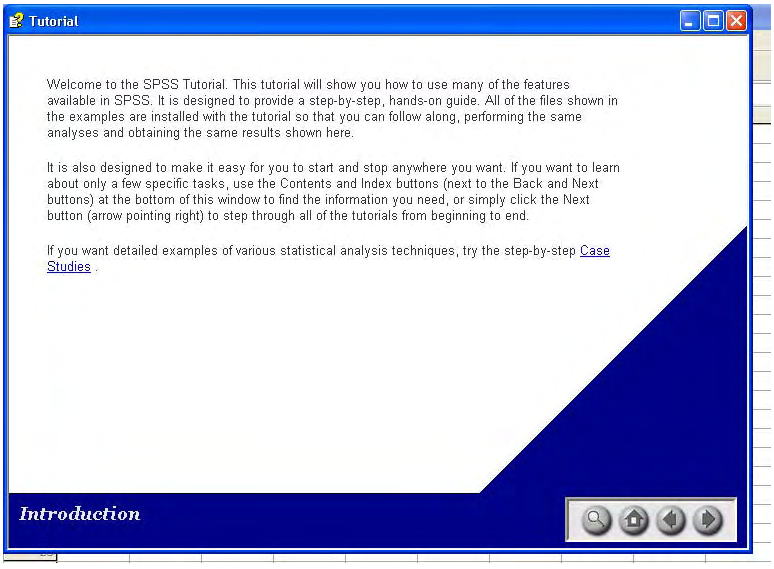
Getting Started with SPSS

►Opening SPSS

When you first open SPSS, the first screen you should see is the "What would you like to do?" window. This is asking for how you would like to enter the data.



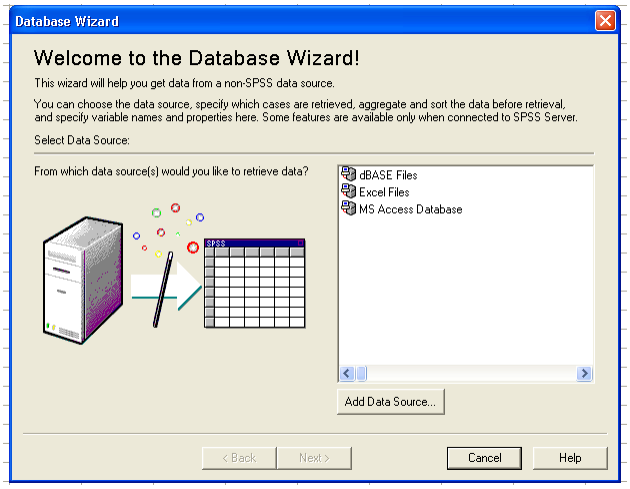
The first option, **Run the tutorial** will take you to a SPSS tutorial. This is a nice little guide on the basics of SPSS. It will walk you through entering data, doing basic analysis, and working with SPSS output. It also gives an introduction to using the SPSS help function.



The next option is to **Type in Data**. If you have data from a book, or your own research, or some other place that is not currently on the computer, this is the option you would want to choose.

Run an Existing Query will help you bring data in from another database, which you have used before. When you run a query into an Excel sheet, or another type of database, you can save the commands so that you don't have to start from scratch each time you want to read in the information. This is very useful when you are doing the same type of analysis over and over again on data that gets updated regularly.

Create a new query using Database Wizard walks you step by step in reading in data that exists on the computer in a non-SPSS format.



These formats include dBase Files, Excel Files, MS Access Databases, and any other database that you have a database (ODBC) driver for. These can be added through the Add Data Source button. This is an easy way to change from Excel format into SPSS.

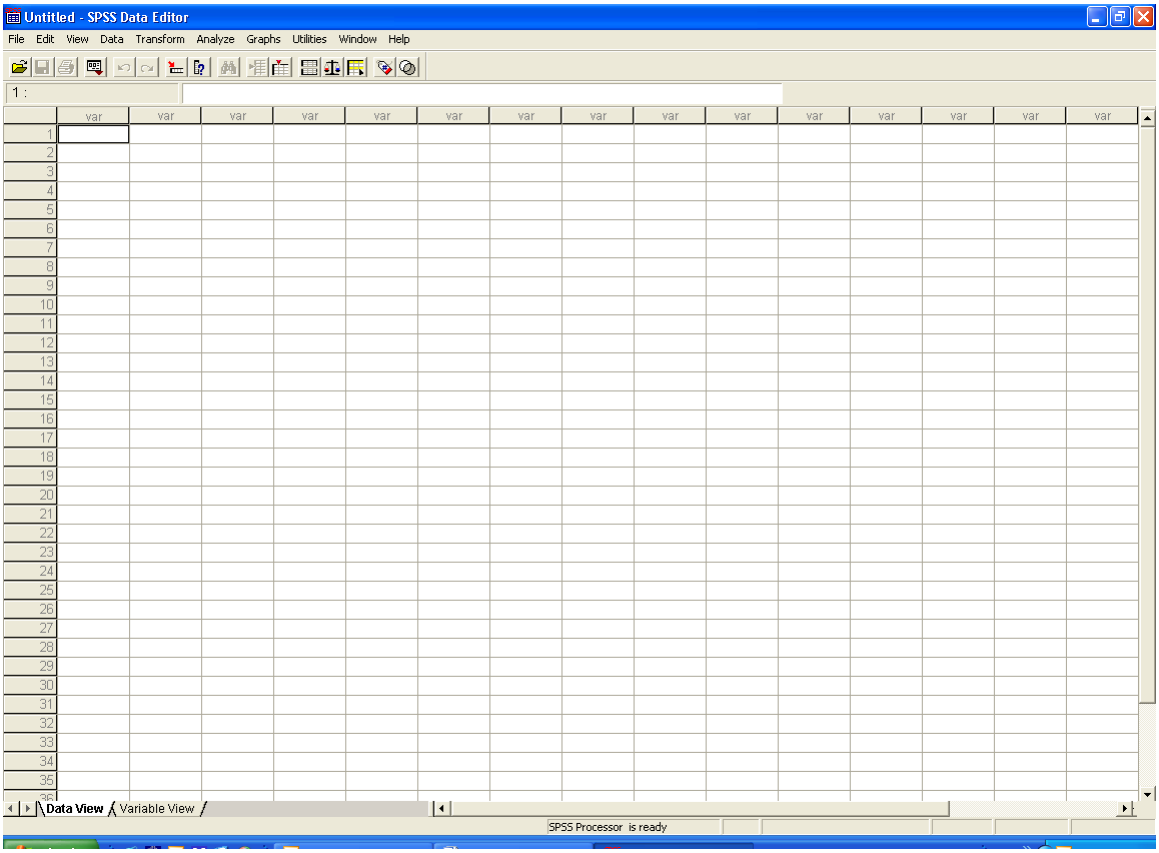
Open an existing data source, which is an SPSS data set. The recently used data sets will appear in the box for easy access to the data sets you were last utilizing.

Open another type of file will allow you to open SPSS scripts, which can be useful for randomly created data sets, or for specific programs you have saved.

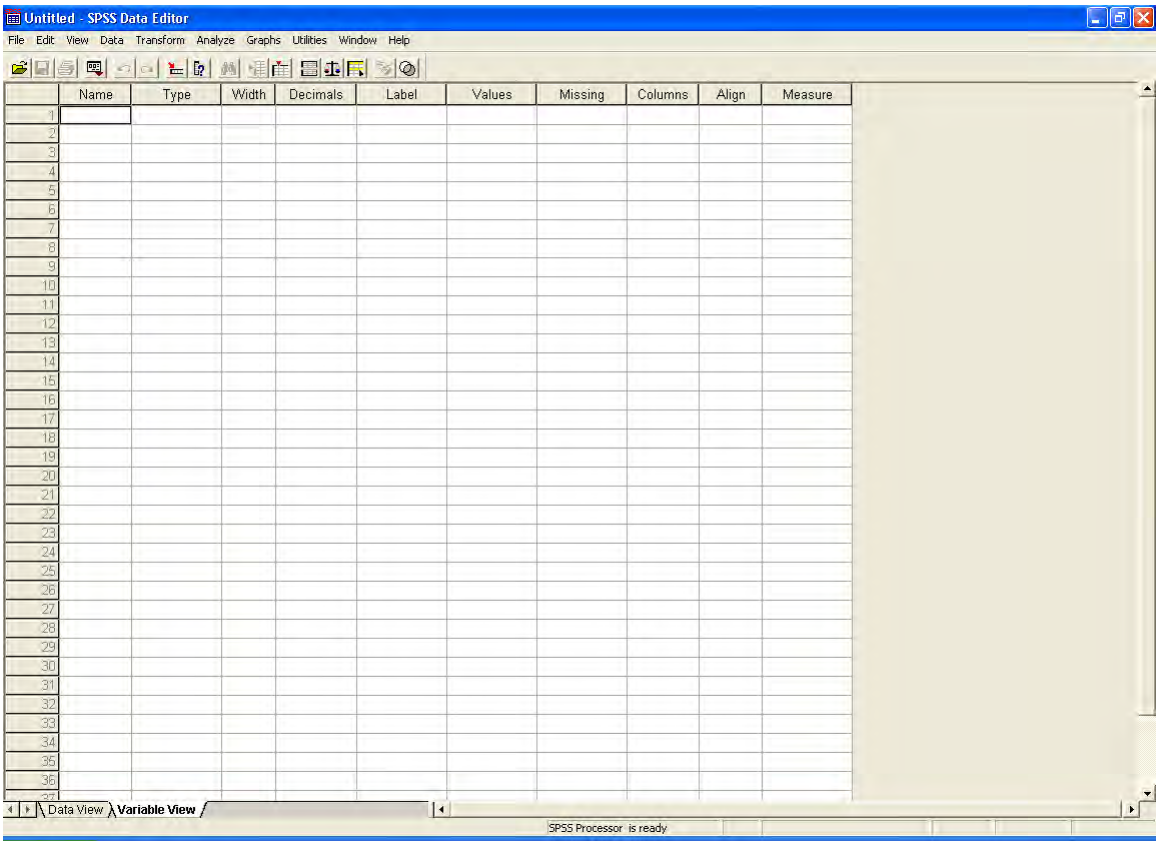
All of these options are available in SPSS through the file menu, and you can ask to not see this menu by clicking the Do not show this dialog in the future box. This would take you directly into SPSS on future executions.

►Using SPSS Files

SPSS begins with two views, the Data View, and the Variable View. Data View is the default start window, and looks like a spreadsheet. Each column is a variable, and each row is an observation. SPSS is set up to accept survey data, and so it expects data for every row for each column with data.



The Variable View window is accessed by clicking on the Variable View tab at the bottom of the Data View window. This is where you can add or change variable names, types, how they are viewed, etc.



Name is the name of the variable. It cannot be more than 64 characters long, must start with a letter, and cannot use any other symbols than a period, @, #, _, and $. No spaces are allowed, and the name cannot end in a period. While this may seem constricting, this is just a variable name. **Label**s can be used to give better detail on what is contained in the variable. Spaces are allowed in labels, but labels cannot be longer than 256 characters long.

The **Type** of variable can be selected from Numeric (example: 10000.001), Comma (example: 10,000.001), Dot (European convention, example: 10.000,001), Scientific Notation (example: 1.00E+004), Date (example: 01-June-2005), Dollar (example: $10,000), Custom Currency, and string (text, example: Male) .

Width is the width of the variable (how many characters). This is most useful when changing String, but can be used for long numbers as well.

The **Decimals** column gives how many decimal places to show. This does not change the values, just how they are seen in the Data View sheet.

Values allow you to give labels to specific values, such as 0-male 1-female. To view the label instead of the number, go to **View**, and check the **Value Labels**.

Missing allows you to specify what characters, numbers, etc. denote a missing value.

Columns dictate how wide the column in the Data View sheet will be.

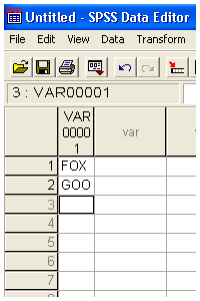
Align allows you to adjust the alignment of the values in the Data View Sheet (Right, Left, Center).

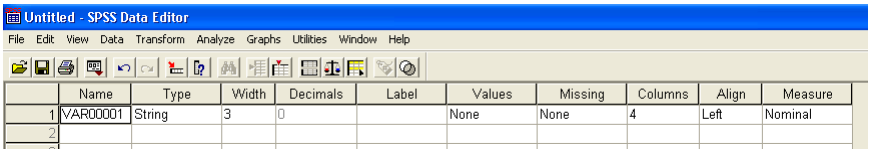
Measure is what type of variable you have. The choices are: Nominal (used for string, but can be used for numbers where the numbers just represent names), Ordinal (where order counts, but distance either varies or cannot be measured), and Scale (the normal numerical line). The type of analysis possible depends on which you chose.

►Entering Data into the Data View

Although it is not necessary to define the variables before typing in data, it is suggested to do so. Define the variables in the Variable View window, providing at least a name, and what Type of variable you wish to use. If you define a variable to be numerical, only numbers will be allowed in that column. If you define a variable to be a string, then you can type any text into the cells, provided they do not go past the length defined.

If you begin by typing in a cell in the Data View window instead, an automatic name will be created (usually VAR0001 if it is the first variable) and what you type will determine the type of variable. Putting in any numerical value will define the variable as numerical, and no value in that column will be accepted which is non-numerical. If you start by typing in text, the variable will be defined as a string, and the length will be set to be the length of the text you type in. (Example: Typing FOX into the first cell will define the variable to be a string, with a width of 3 characters. Typing in GOOSE into the next cell will give you GOO, as only 3 characters will be allowed). These definitions can still be changed in the Variable View window.



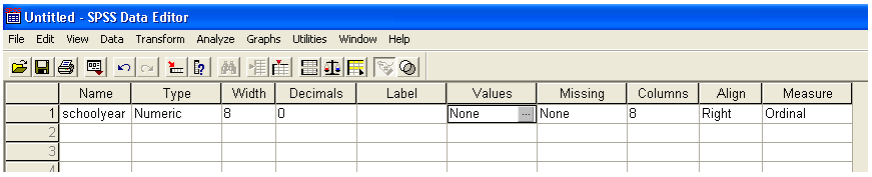


►Working with data

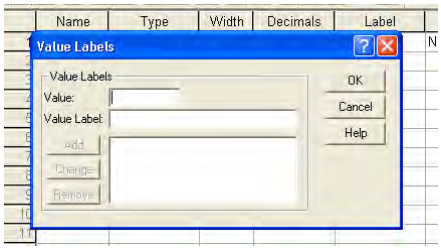
Before entering in data values, it is best to consider what type of variables you are looking at, and what type of analysis you want to run on the data. There are three major types of data to worry about.

-Nominal data: nominal data have no ordering issues, and can easily be entered as seen. For regression, or higher analysis, dummy variables should be used, but these can be created later.

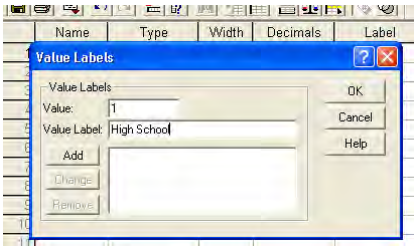
-Ordinal data: ordinal data have a natural ordering to it, and so, the ordering needs to be preserved. This can be accomplished by using numbers to represent the data, in order, and then attaching labels for the numbers. For example, a survey asks "What year of school are you in?" with responses ranging from high school to graduate school. In SPSS, we create a new variable, schoolyear, which will be numeric, no decimals, and a Measure of ordinal. This is all done in the Variable View page.



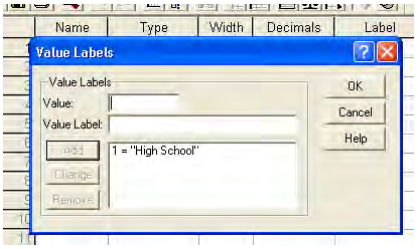
Clicking on the Values box will show a … button, and clicking on the … button will open the Value Labels dialog window.



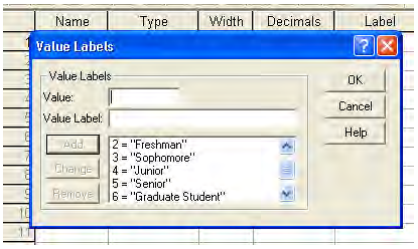
In the Value Labels dialog window, the first box is the Value of the variable, as you will be typing it into SPSS. We will use 1 for high school, so we type a 1 in the Value box, and High School in the Value Label box.



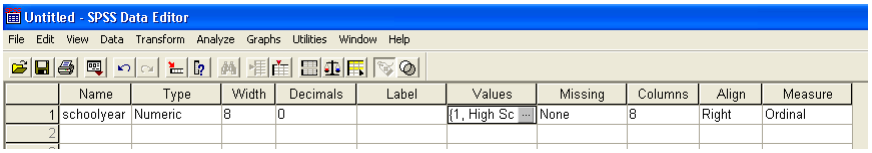
Once values are in both boxes, the Add button is available. Press Add, and any 1 in the column will be recognized by SPSS as High School.



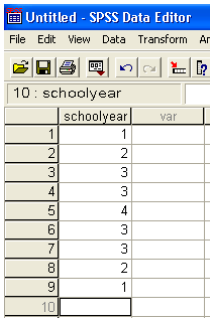
Once you add in all the possible values, push OK. If a value needs to be changed, you can highlight the value, which will put the information back into the Value and Value Label boxes, make the appropriate changes, and push the Change button. You can remove labels by highlighting them, and pressing the Remove button.



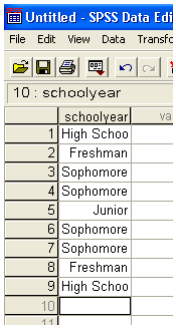
When labels are being used, you will see text in the Labels box, starting with the first value.



We can then switch to the Data View sheet, and enter in the information.



You will see the values as the numbers that you enter, but you can change that, by going to the View menu, and checking the Value Labels option. This will let you see the labels instead of the actual values. You will still type in a 1, 2, etc. but you will see the full names. This is also useful for any nominal values that you don't want to have to type multiple times. It is much easier to type in small numbers than long strings.

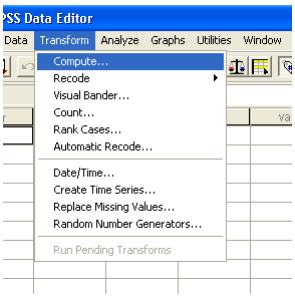


Although the analysis uses the numbers, thus preserving order, the output will print the labels.

-Numerical data: Numbers can be typed in directly, but be aware that you may not see the full number. You can change the number of characters shown in the Variable View sheet, as well as the number of decimal places shown. Reducing the number of decimal places does not change the value of the number. SPSS does not round the actual value, only the value that you see.

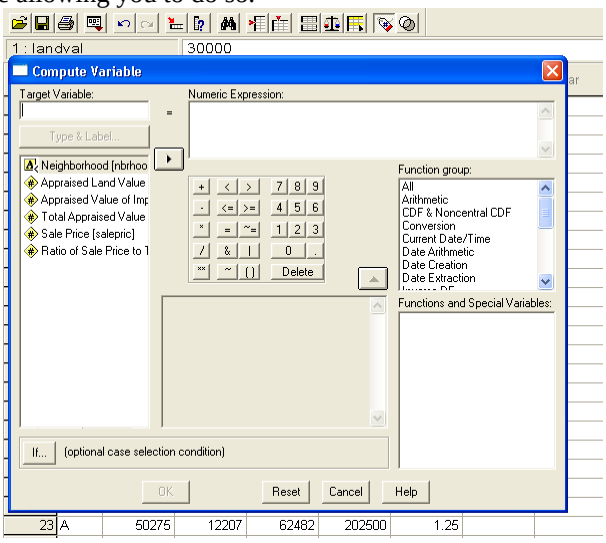
►Computations with data

There are times when you will want to perform some type of calculation using your data, such as squaring values, or finding probabilities, or creating a new variable with random values. SPSS has a calculator for such transformations. When you go to **Transform → Compute** you can make many types of changes to your data.

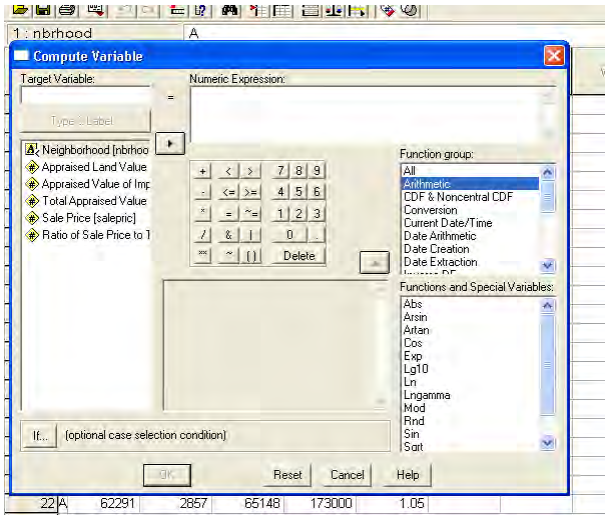


This will open up the Compute Variable window. You must already have data before this will open. The Compute Variable window contains five main boxes. The first is the Target Variable. This is a name for

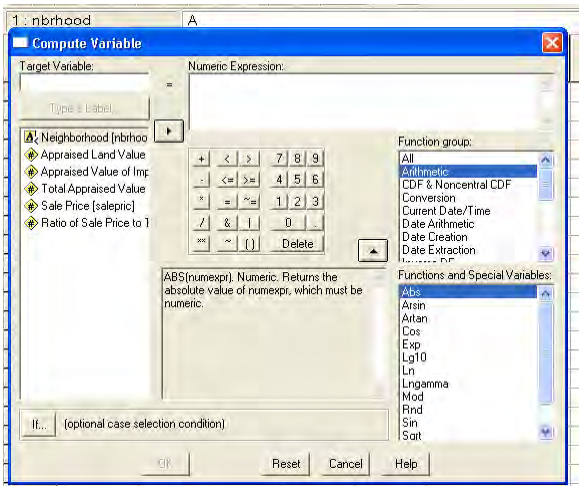
the new variable which you are creating. You can overwrite an existing variable if you so desire, but SPSS will give a warning before allowing you to do so.



The window below the Target Variable is a list of the current variables. All of these can be used as part of the function, and the list lets you come up with a name that will be different from what already exists. The Numeric Expression box is where you will put in the function you want to use. This could be as simple as 1+1, which would create a new variable, with the same number of rows as the existing data, all with the value of 2. The next box is the Function group box. This is the general category for mathematical or statistical functions that are available to use. For example, highlighting Arithmetic brings up a list of arithmetic operators in the Functions and Special Variables box.

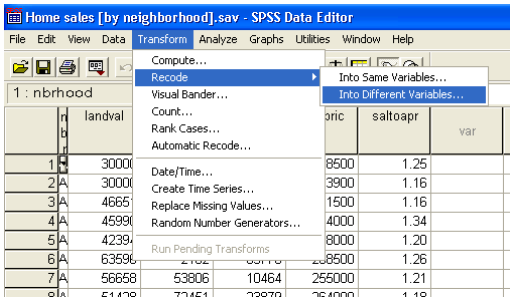


So, to calculate absolute values, we will highlight Abs in the list of Functions and Special Values, and an explanation of the function will appear in the box next to the function list.

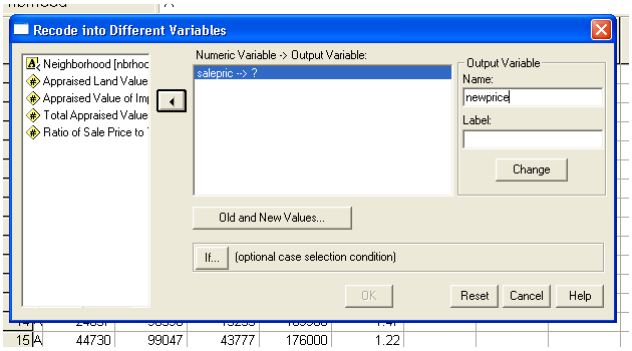


To use the Abs function, we will need a numerical expression to put into the parentheses. This value could be any of the variables or a function of the variables. So, we could use ABS(totval-landval) to get a new variable which is the absolute difference between total appraised value and appraised land value.

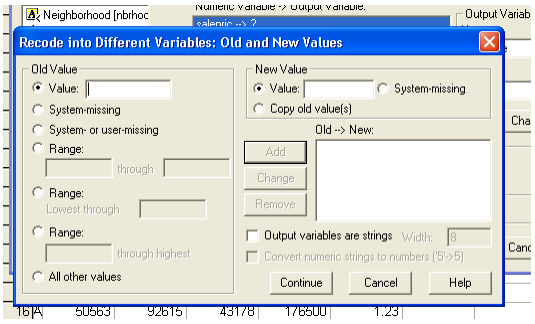
Another way to change the data is to recode the information. To recode, go to **Transform → Recode**, and select **Into Same Variables**… which overwrites an existing variable, or **Into Different Variables**… which creates a new variable. It is generally preferred to create a new variable in case of mistakes unless memory becomes a problem.



Selecting the different variables option will open up a Recode into Different Variables window. You select which variable you want to recode, and a name for the new variable. To create the variable, you must push the change button, which takes the new name to the right of the → in the Numeric Variable → Output Variable box.



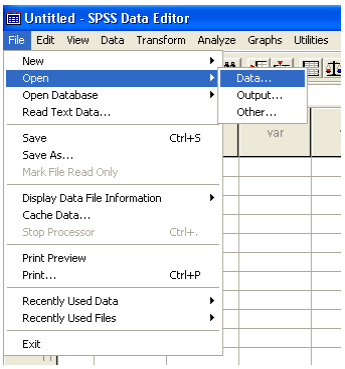
Once you have a name for the variable, click on Old and New Values. This will open the Recode into Different Variables: Old and New Values window. This is where you tell SPSS what changes you would like to make to the data.



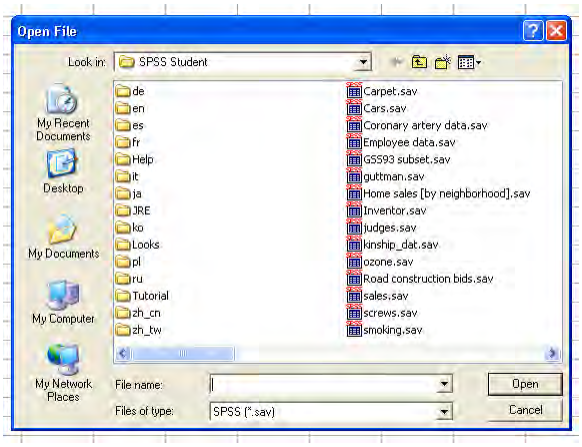
In the Old Value list, you let SPSS know what the old value was. If you have a short list, you can use the Value box. Type in the old value, put the new value in the New value box, and click Add. This is useful when you want to switch order, for example, you receive data that has a 5 point scale, where 1 is high and 5 is low, and you want 5 to be high and 1 to be low. You can then put 1 for the old value, and 5 for the new value, etc. If you have a list of numbers, you can use the Range values. This can be used to categorize numerical data, for example heights from height in inches to short, medium, tall. Once you have completed the full list, push the Continue button, and then you can push OK. You will now have a new, recoded variable. The Old and New Values works similarly for recoding into the same variable, but you don't have a window in which to name the variable.

►Opening Saved Data Sheets

To open a saved SPSS data sheet, click on **File → Open → Data …**



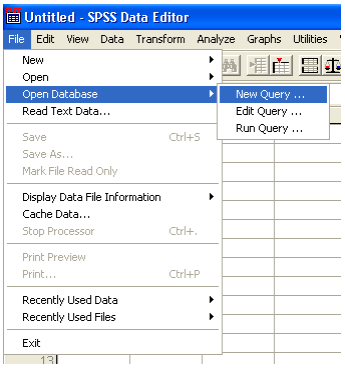
The following screen will appear.



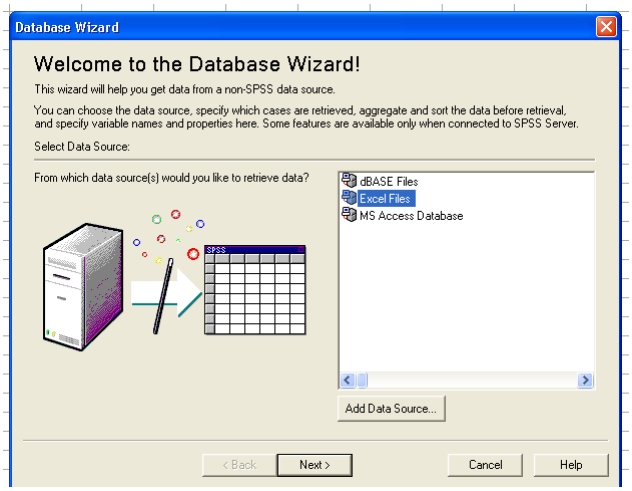
First, you must find where your data files are stored. The default location for SPSS to begin looking is in the data sets that came with the program. You can change the disk location by using the Look in: box. It is best to keep all of your data sets in a location that is easy to find. SPSS data sets will appear as little blue data boxes with SPSS in red above them. They also have the extension .sav following the name. No other data type can be read in as an SPSS data set.

►Importing Data from Excel

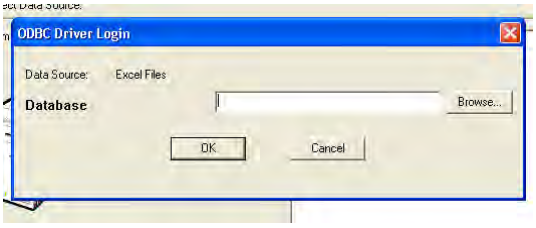
As most of the data come in three formats on the disk supplied with ASCII files, Excel files, and MINITAB files, but not SPSS files, it will be advantageous to import from one of these formats rather than recreating the data sets. You will need to save the files from the disk onto your hard drive, as they are not accessible by browsing the CD. Excel files are easier to import from than ASCII, as they follow specific conventions. To import from an excel file, go to **File → Open Database → New Query …**



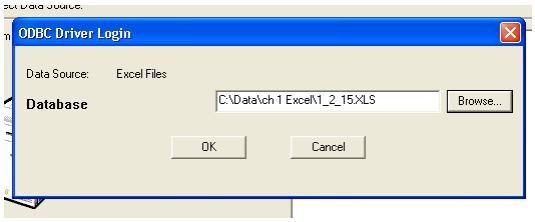
This will open up a Database Wizard window.



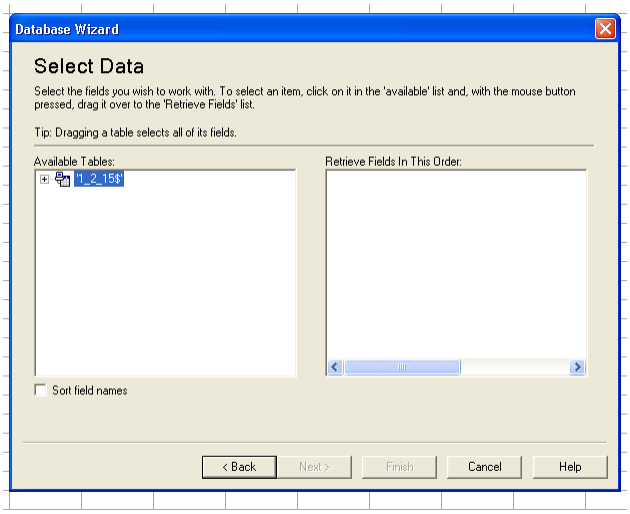
Highlight Excel Files, and push the **Next>** button. This will give you a dialog window.



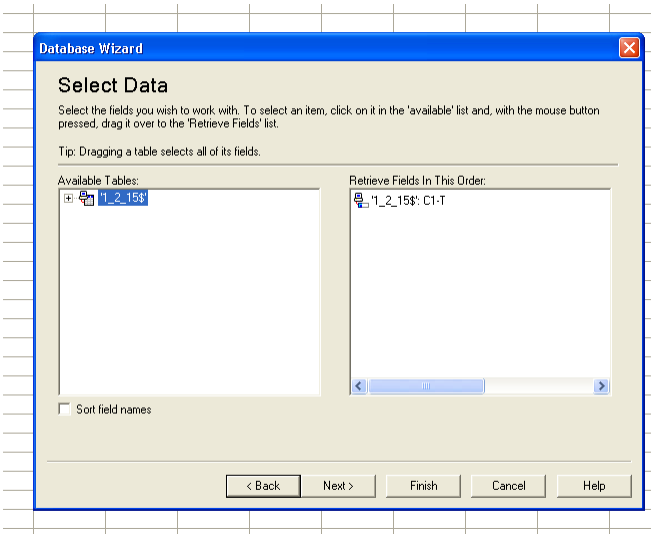
Use the Browse… button to find the file that you wish to open. The Excel file cannot be open in Excel when importing the data, as a sharing violation will occur.



The next screen is the Select Data screen. You may select all of a data set, or portions of the file. If Excel has more than one sheet available, this is also where you would specify which sheet contains the data that you want to use.

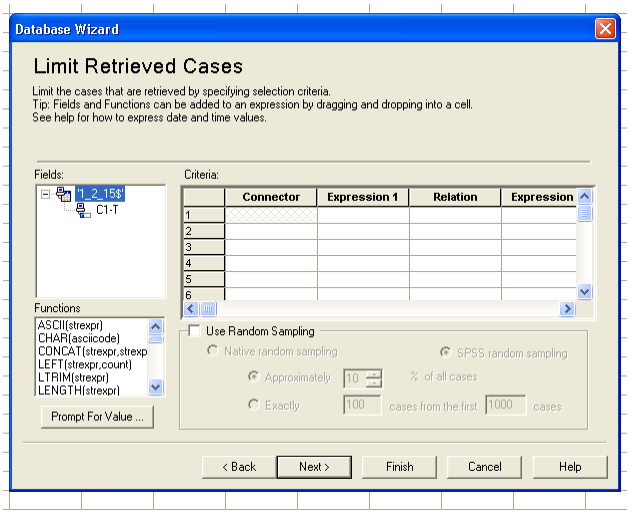


Drag the sheet over from the Available Tables box to the Retrieve Fields in This Order box. When you click on the available table, a hand will appear. Keep the mouse button pressed until the hand is over the Retrieve box. This will give a list of the variables in the data set.

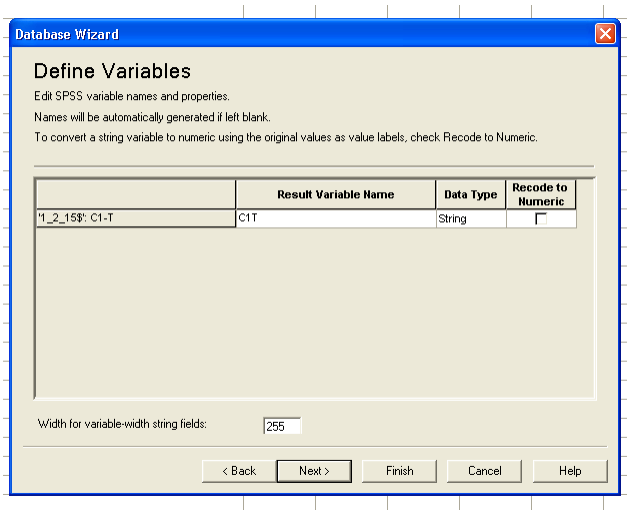


You may change the order of the variables, or remove variables at this point. When you are ready, push Next> or Finish.

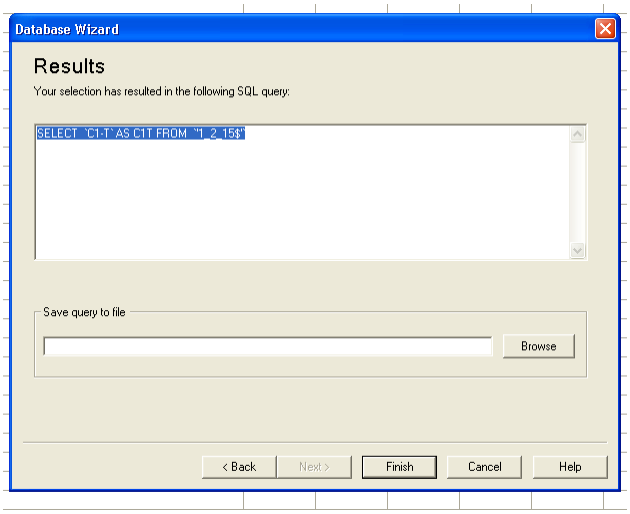
Pushing Next> will take you to a window where you can limit the number of observations to bring in, including the possibility of bringing in a random sample of the observations.



Pushing Next> again will take you to the Define Variables window. Here you can change the variable type, or length.



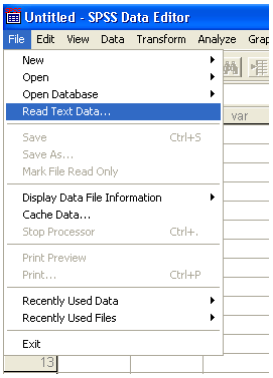
Pushing Next> again will take you to the Results window. Here you have the option of saving the query that you just ran. This is useful if you plan on opening the same Excel file many times. It is easier to save the data as an SPSS file than to re-import the data over and over again, but in business practice, there are times when you will want to open the same file numerous times, after updating the information in the file. When there are a large number of observations available in the data set, it is easier to save the query and just re-run it periodically than to re-make the query each time.



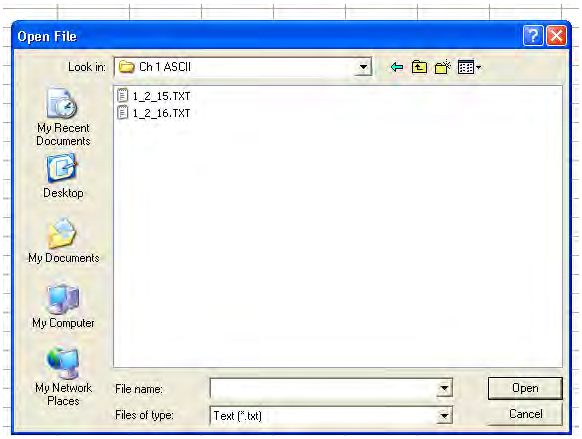
Pushing the Finish button at any time will bring the data into SPSS. The name of the column will be obtained from the first row in Excel, but will be changed to accommodate the SPSS naming conventions.

►Importing Data from ASCII

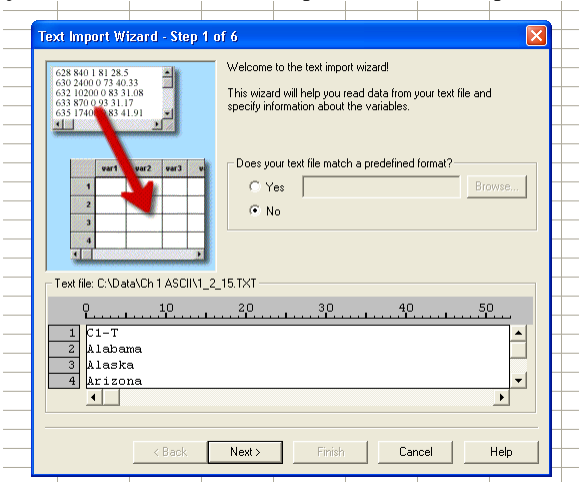
There are times when importing from an ASCII (text) file is necessary, especially when retrieving data from web sites. To read in a text file, go to **File → Read Text Data…**



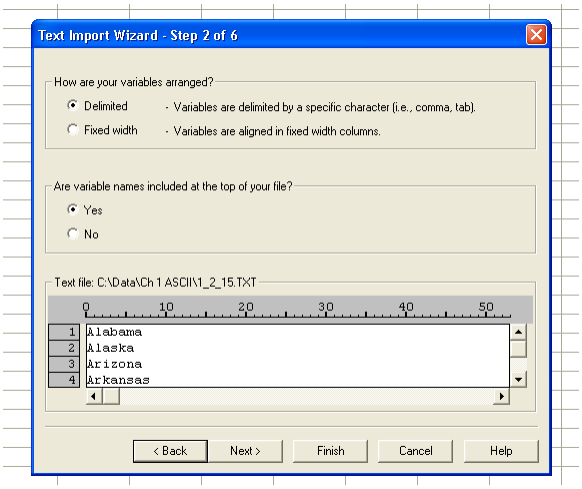
This will take you to an Open File window, where you will browse for the file that you want to import the data from.



Once you select which file you want to use, the text import wizard will open.

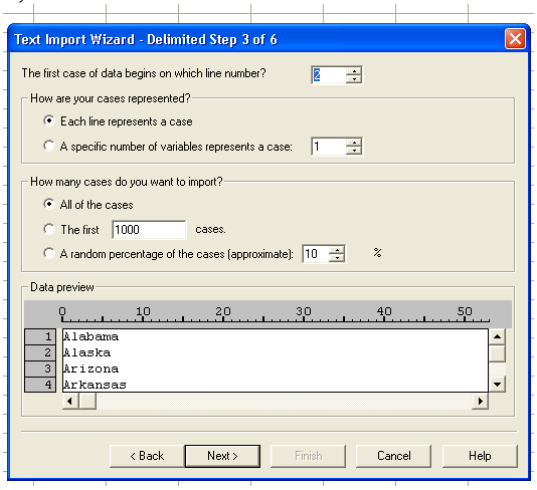


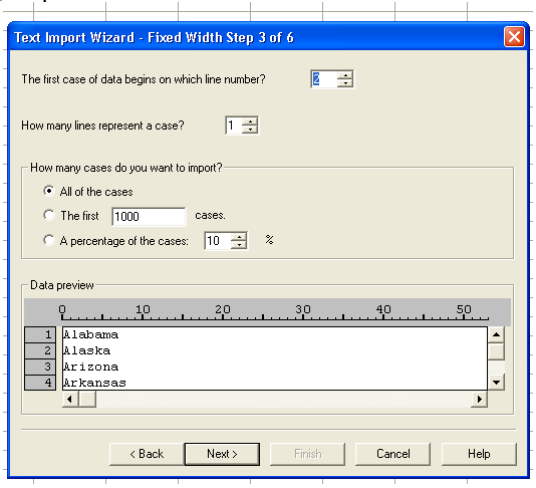
The first step is to see if the data is the correct one, you will see the first couple lines of the file in the bottom box. If you are going to read in the same document over and over, you can save the format, and use that in the predefined format box.



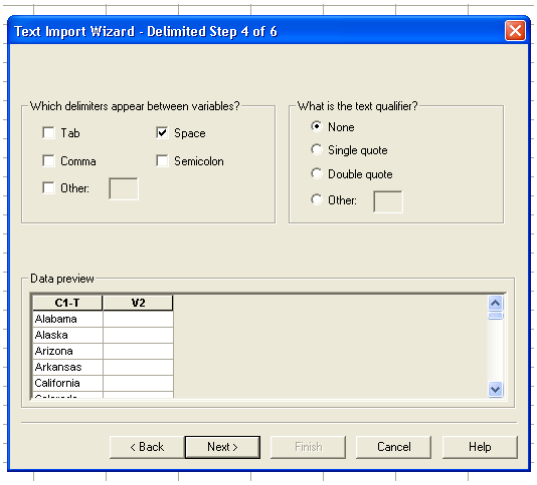
Step 2 is to tell SPSS how the data looks. If you are using a file which is Delimited (using the same character to distinguish between variables) then use the Delimited button. If the variables are EXACTLY

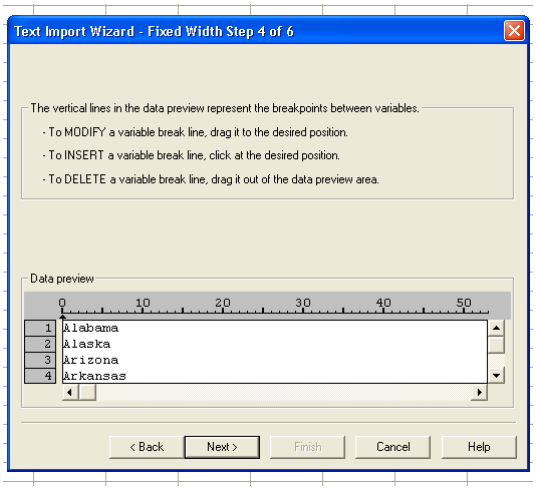
in line, you can use the Fixed width button. Files can be delimited with a space, but this means that spaces cannot be inside strings. If the data is not exactly lined up, the fixed width will not work very well. Also, define whether the first row is the variable name, or part of the data.



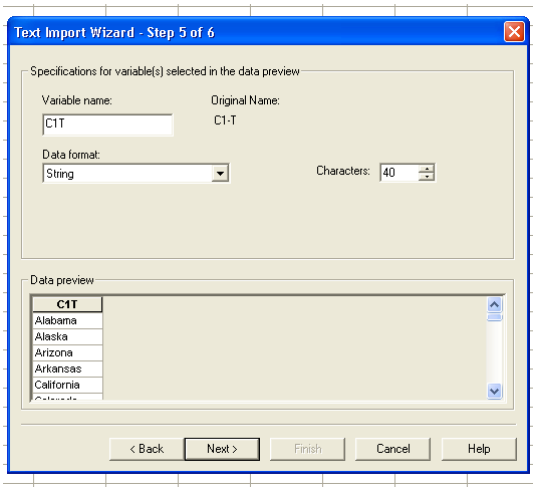


Step 3 is either to tell SPSS where the data starts. If the data starts on any row other than the second, you can change that information here as well. If you don't want to read in the whole data set, you can limit the number of cases.

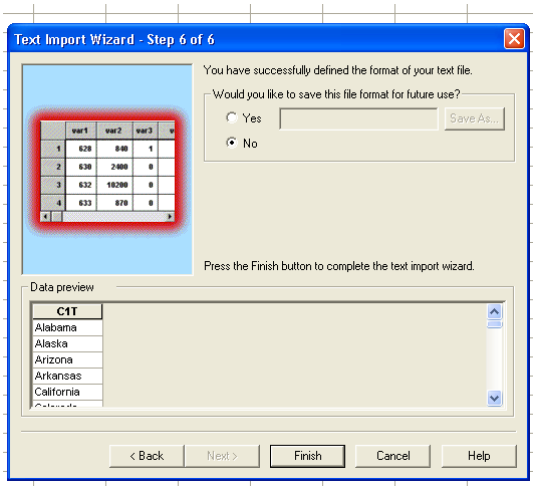




Step 4 is to define the delimiter or to define the widths. SPSS tries to define a delimiter based on general conventions, but these can be changed. The common delimiters are listed (Tab, Space, etc.) but a box is available for different delimiters. With the fixed width, you will put in a line delineating the variables.



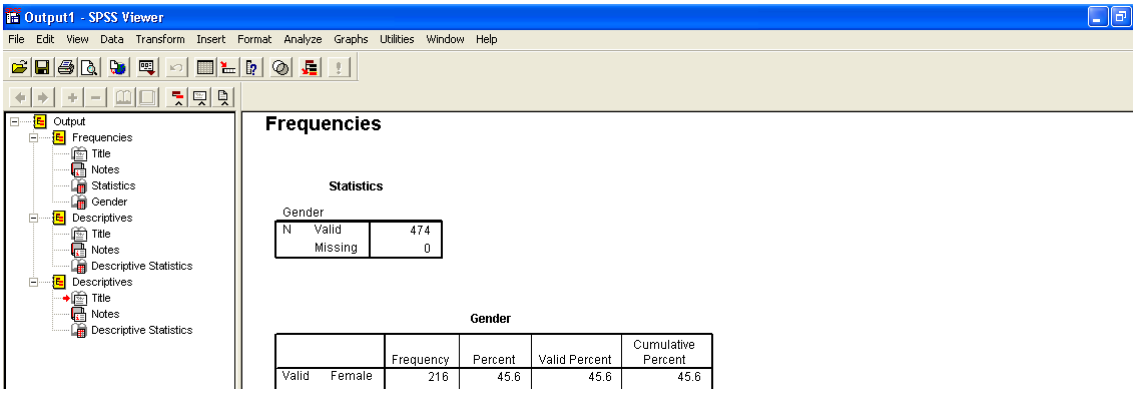
Step 5 is to specify a name. You will get an error message if the name in the first line (if you specified that the name was at the top of the file) does not satisfy the naming restrictions. SPSS will take out all illegal characters, but you have the possibility of renaming the variable altogether as well. Also, you can specify what type of data the variable should be. The default is numeric, so if there are any characters, such as commas or dollar signs, this will need to be changed.



Step 6 allows you to save the format for future use.
Pressing Finish at any time will read in the data.

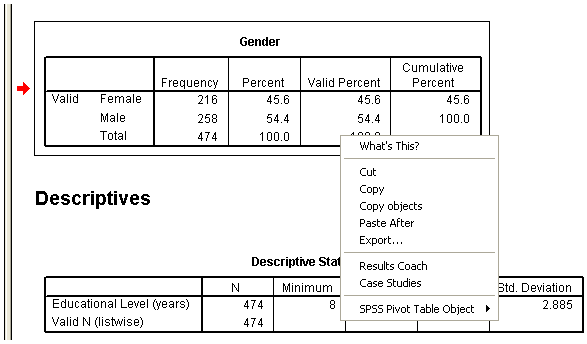
►Using the Output Viewer

After running an analysis in SPSS, you will usually see the results in a new window, called the Output Viewer window.



On the left hand side of the window is a list of all the analyses that have been run, and the pieces of the output that are available. Any of these can be removed, or, if the list is getting too long, you can push the – sign to the left of the analysis and it will hide, but not delete, the output. To show the analysis again, push the + button to the left of the closed output. You can continue to run analyses from the Output window. The same options appear in the output window as in the data view window, so you do not need to return to the data to keep going.

While in the output window, if you click on any piece of output with the right mouse button, the option will appear to use the Results Coach. This will take you to the tutorial for that type of analysis, and will walk you step by step on how to interpret the results of that analysis.



You can also copy the table from SPSS into Excel, or into Word for better formatting and printing options, but this should be done one table at a time, as problems arise trying to copy a whole page. You will want to look at the print preview when printing directly from SPSS, as SPSS may put page breaks at interesting times, and some output will run off of the page. Watching the print preview can save paper.

Chapter 1. Data Collection

SPSS is a data analysis tool, not a data creation tool. Because of this, it is not as powerful as other programs in generating data. It is easier to generate the data in Excel, or another tool, and import it into SPSS.

Section 1.1 Introduction to the Practice of Statistics

SPSS is particular about the type of data that you collect. Qualitative variables are broken up into two cases, Nominal, and Ordinal (see problem 57). Nominal comes from name, and represent qualitative data which have no specific order. Ordinal data on the other hand have an implied or specified order to them. Example:

a) Gender -Nominal, there is no order implied in being male or female.

b) Zip Code -Nominal, although the first digit can be used to specify a region in the US, 0 being east and 9 being west, the other digits relate to a city, but have no real order.

c) Class Standing -Ordinal, Freshmen have less credit hours than Sophomores, who have less credit hours than Juniors, who have less credit hours than Seniors, so an order is implied. You move from being a Freshman to being a Sophomore to being a Junior to being a Senior.

d) Leikert Scale -a scale from 1 to 5, or 1 to 7 etc. measuring agree to disagree, or like to dislike, or other opinions. Although there is an order specified, the distance between the values changes from view to view. In other words, a 5 doesn't dislike something twice as much as a 4, who dislikes the thing twice as much as a 3, etc.

SPSS does not break Quantitative variables into discrete and continuous. Both are given the title Scale. SPSS will identify text values as Qualitative variables, but must be told whether numbers are Nominal, Ordinal, or Quantitative. This is done in the Variable View window.

Section 1.2 Observational Studies, Experiments, and Simple Random Sampling

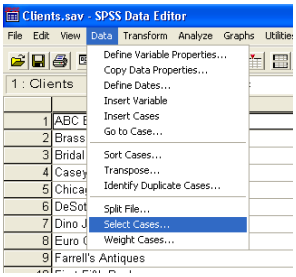
SPSS is a data analysis tool. It assumes you already have some data to analyze. Unlike some other software packages, it is difficult to create data from scratch in SPSS. SPSS will, however, take random samples.

►Obtaining a Simple Random Sample from a list

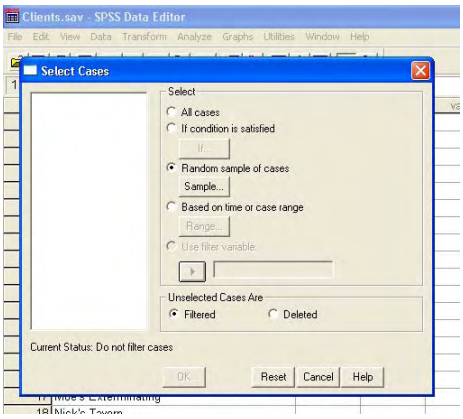
Example 2, page 17.

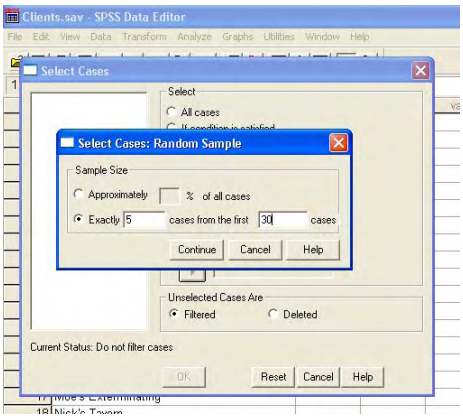
To select five clients out of the list of thirty, first we type all thirty clients into a blank SPSS sheet.

We then go to **Data → Select Cases** to take a sample.

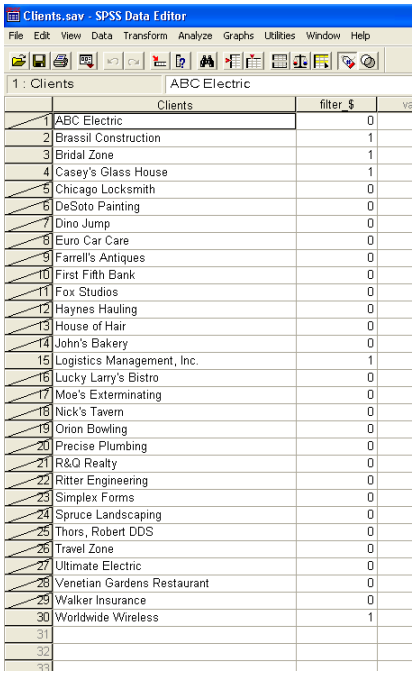


There are different types of selections we can take from SPSS. We can select specific values, or in this case, we can take a random sample of cases. In the Select Cases window, **select Random sample of cases**, and push the **Sample…** button. (The Sample… button will not be available until the Random sample of cases has been selected.)





Since the exact sample size is known, choose the Exactly row, and type in the sample size wanted (5) and the population size (30), and push continue. This will return you to the Select Cases window, where you can now push OK.

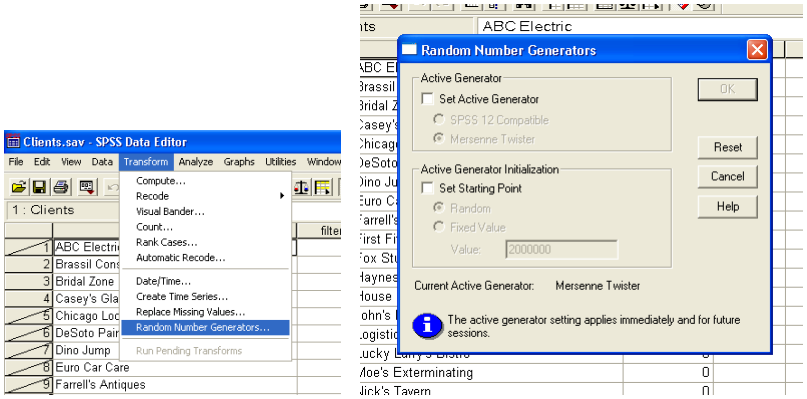


	Clients	filter_$
1	ABC Electric	0
2	Brassil Construction	1
3	Bridal Zone	1
4	Casey's Glass House	1
5	Chicago Locksmith	0
6	DeSoto Painting	0
7	Dino Jump	0
8	Euro Car Care	0
9	Farrell's Antiques	0
10	First Fifth Bank	0
11	Fox Studios	0
12	Haynes Hauling	0
13	House of Hair	0
14	John's Bakery	0
15	Logistics Management, Inc.	1
16	Lucky Larry's Bistro	0
17	Moe's Exterminating	0
18	Nick's Tavern	0
19	Orion Bowling	0
20	Precise Plumbing	0
21	R&Q Realty	0
22	Ritter Engineering	0
23	Simplex Forms	0
24	Spruce Landscaping	0
25	Thors, Robert DDS	0
26	Travel Zone	0
27	Ultimate Electric	0
28	Venetian Gardens Restaurant	0
29	Walker Insurance	0
30	Worldwide Wireless	1
31		
32		
33		

A new column has been added to the data, called filter_$. This has a 1 if the row is in the sample, and a 0 if it is not in the sample. The rows that have not been selected also have a line through the observation number. If the Deleted option was chosen in the Select Cases window, these values would no longer be in the data set. Although all values can be seen, only those with a filter_$ value of 1 will be used for any analysis. In this example, the clients to be surveyed are Brassil Construction, Bridal Zone, Casey's Glass House, Logistics Management, Inc., and Worldwide Wireless.

If another sample is desired, going back through the steps will select another sample, without having to clear the information from the first.

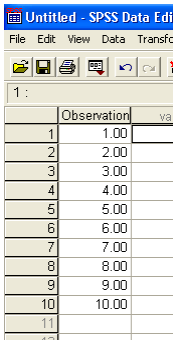
To change the random seed (starting location in the random number table), go to **Transform → Random Number Generators…**



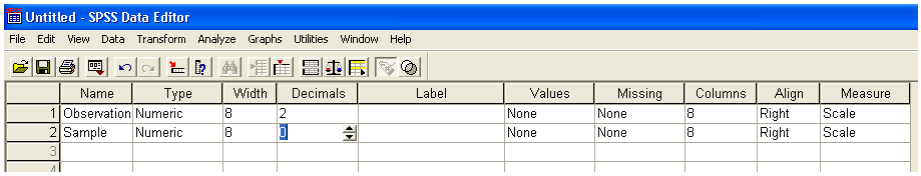
The Random Number Generators window has two options you can change. The first is the Active Generator. If you want to reproduce results as done in SPSS version 12, the SPSS 12 Compatible should be selected. Otherwise, the Mersenne Twister should be selected, as it is a more reliable randomization tool. Under Active Generator Initialization, you can select the starting point. The default is Random, which is generally based on the computer clock, down to a fraction of a second, so every person who takes a sample will get a different starting point. If you want more replicable results, a fixed value can be used, which will ensure the same starting location, and thus the same sample, each time.

►Obtaining a Simple Random Sample without a list in SPSS

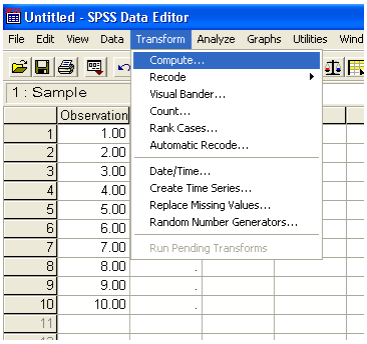
There are times that you do not want to enter a list into SPSS, due to length or other constraints. SPSS can still help chose the sample, although you must have a numbered list to choose from. SPSS has a random number generating tool, but data must exist before using it. So, the first step is to decide on the sample size. Similar to Excel, or Minitab, we will have to create a few extra observations in case of repeats. In a blank SPSS Data sheet, create a column of integers, from 1 to n. You will want to add a few more after this in case of repeated values.

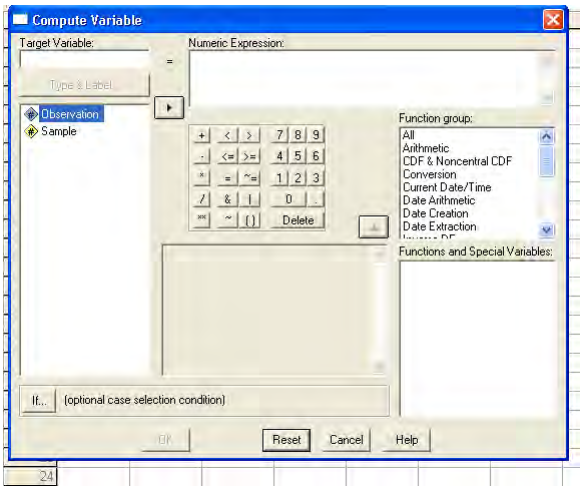


Here we want a sample of five from the thirty, but to be safe, we added five extra in case of repeats. In the Variable View window, create a new variable, which is numerical with no decimal places.

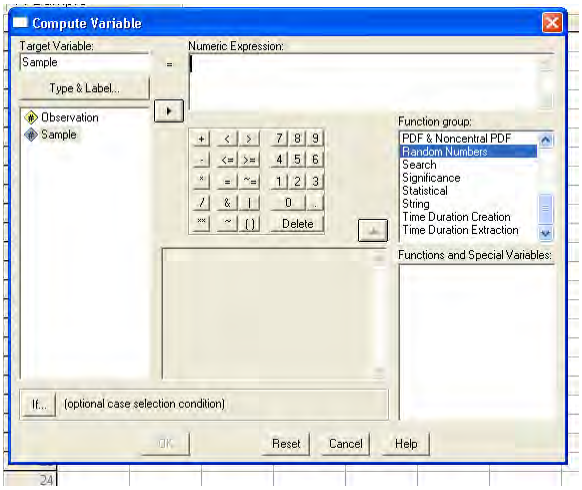


Go to **Transform → Compute…**

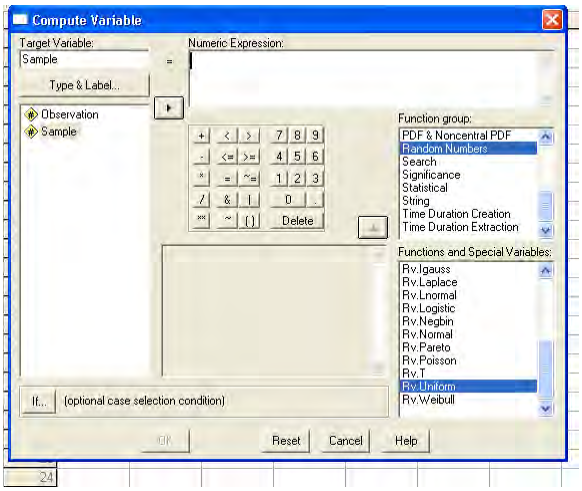




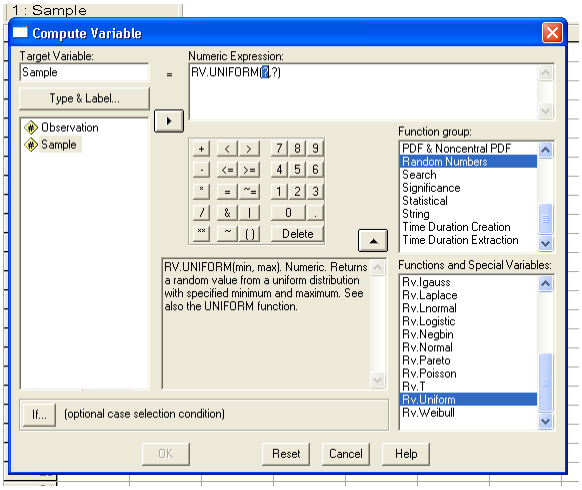
There are two windows that we must fill. The first is the Target Variable, for which we want to use the name of the blank column we just created.. In the Function Group box, find Random Numbers



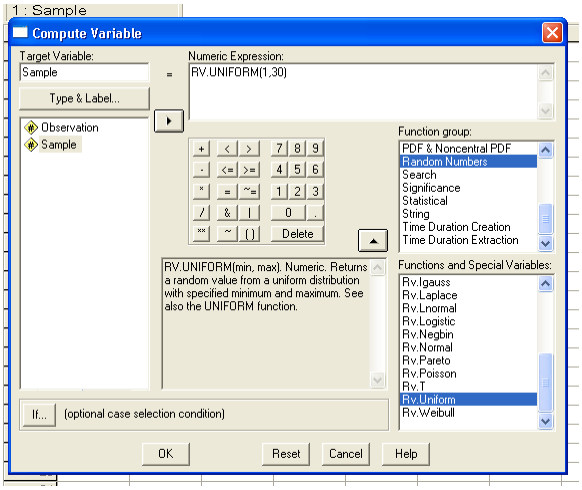
Clicking on Random Numbers will bring up a list of possible random variables. Find RV.Uniform. RV.Uniform will give every value an equal chance, which is what we want for a Simple Random Sample.



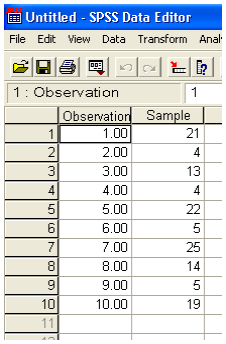
Click on the RV.Uniform until it appears in the Numeric Expression box.



The question marks in the Numeric Expression are values that you must provide. From the explanation in the box, we can see that we need a minimum value, and a maximum value. Since we are selecting values from 1 to 30, these are the values that we enter.



When the Numeric Expression looks correct, push OK. You will be asked if you want to change the existing variable. Since the existing variable Sample is currently empty, we can safely push OK. When you get this warning, you do want to make sure that you are not going to overwrite data that already exists. We will now see the values that have been selected in the new column.



Our sample in this example would be the 21st, 4th, 13th, 22nd, and 5th values in the list. Note that 4 was listed twice, so we ignore the second occurrence, and use the sixth value in the list. We can ignore the other values, as we only want the first five unique values.

Section 1.3 Other Effective Types of Sampling

SPSS only takes Simple Random Samples. Other types of sampling can be achieved through other methods, but are much harder to do. This is a result of SPSS being an analysis tool, where it is assumed that the data is already from a sample.

Obtaining a Stratified Sample:

To obtain a Stratified Sample, you can create a Simple Random Sample for each stratum. It must be remembered that SPSS will only create a value for a row that is already in use by another variable.

Obtaining a Systematic Random Sample:

SPSS can help create a Systematic Random Sample by helping you choose the first value from the first k observations. You will then choose every k^{th} individual after the randomly selected starting point.

Obtaining a Cluster Sample:

To obtain a Cluster Sample, you can use SPSS to help take a Simple Random Sample of the clusters.

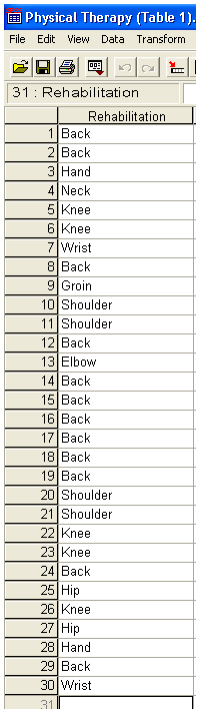
Chapter 2. Organizing and Summarizing Data

Section 2.1 Organizing Qualitative Data

There are two different ways to enter data into SPSS, in raw format, or summarized information. Raw format is the actual data, in a full list, each row representing an appearance of the observation. Summarized format is where one column is the value of the observation, and another column represents how many times that observation was observed.

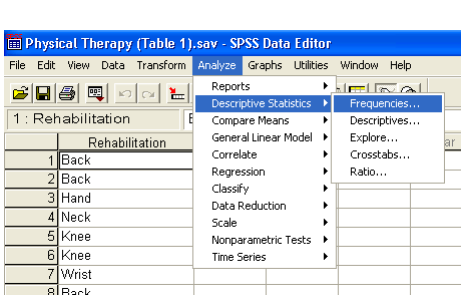
Example 1, page 61

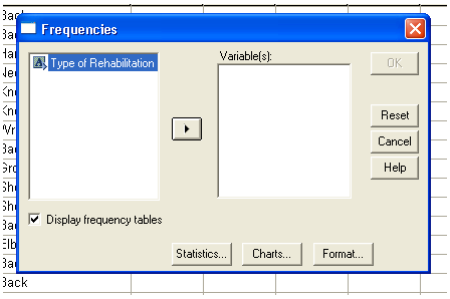
To construct a frequency distribution and/or a frequency bar graph in SPSS, you will first need to enter the data into a Data View sheet. For example 3, we will use the information as presented in Table 1 on page 61. This is in raw format. Each row represents an individual observation.



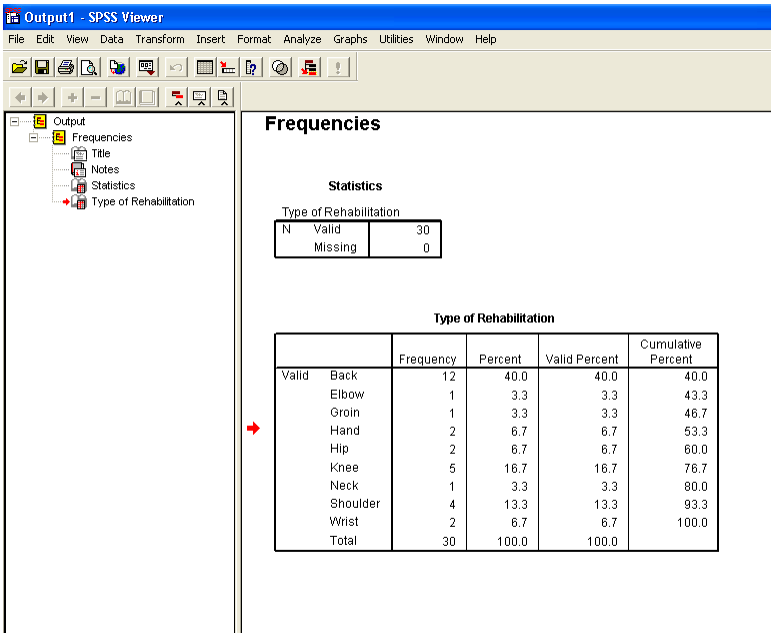
►Creating a Frequency Distribution

To create a frequency distribution, we select **Analyze → Descriptive Statistics → Frequencies…**

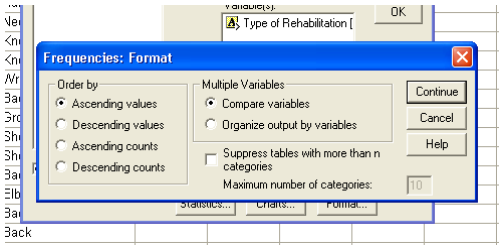




Highlight the variable that you want the frequency distribution of, in this case Type of Rehabilitation, and push the ► button. This will take the variable you chose into the Variable(s) list. Make sure that the Display frequency tables box is selected. Push OK. The output will appear a new window called the Output Viewer.



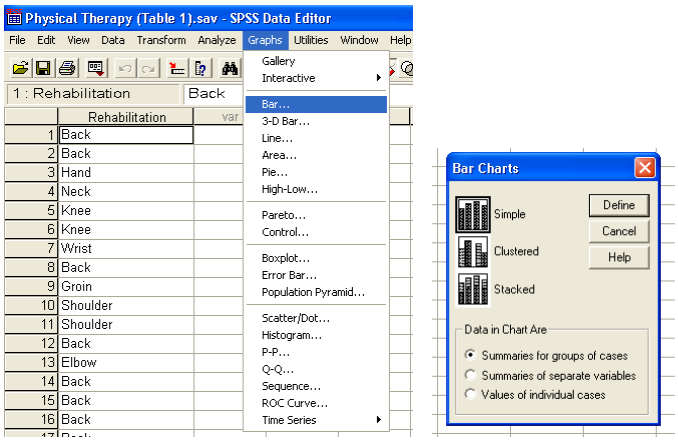
Included in the table are the Frequencies, Percent (Relative Frequency, but will take n to be the sample size including missing values), Valid Percent (which is the Relative Frequency without missing values) and Cumulative Percent (Cumulative Relative Frequency, based on the Valid Percent).
The order that SPSS uses by default is the alphabetical order. To change this, while in the Frequencies window (before pushing OK), select Format. This will open the Format window.



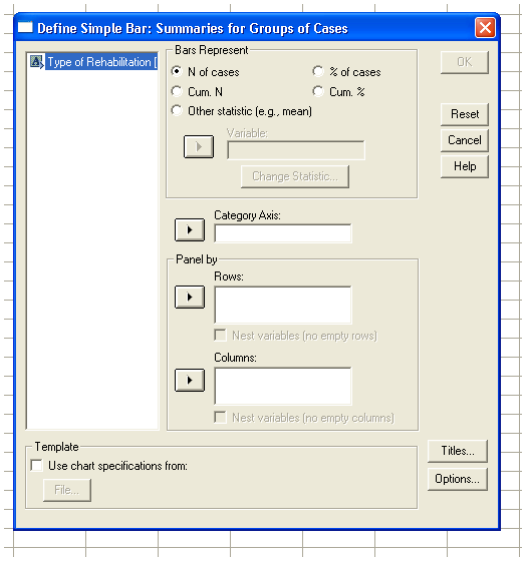
Here you can choose to order the variables by Ascending values (alphabetical), Descending values, Ascending counts, and Descending counts. You can also choose to not show values that appear infrequently.

►Creating a Frequency and Relative Frequency Bar Graph

Using the same data as for a frequency table, select **Graphs → Bar**

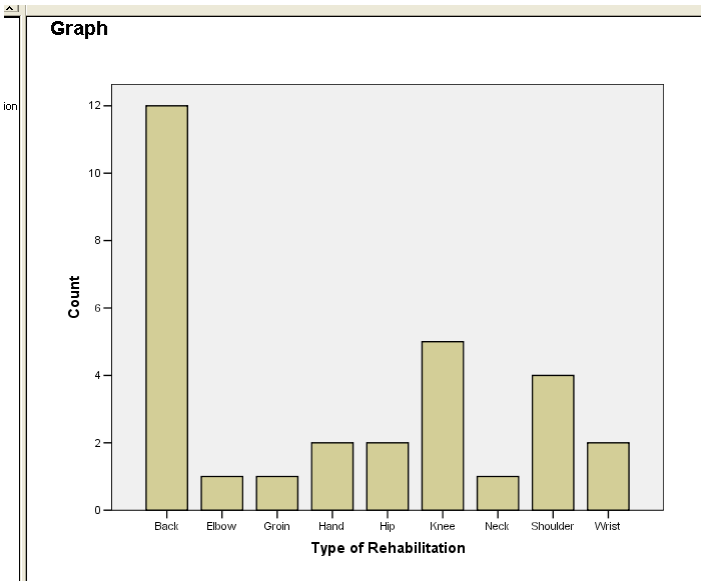


In the box, select **Simple**, and **Summaries for groups of cases**

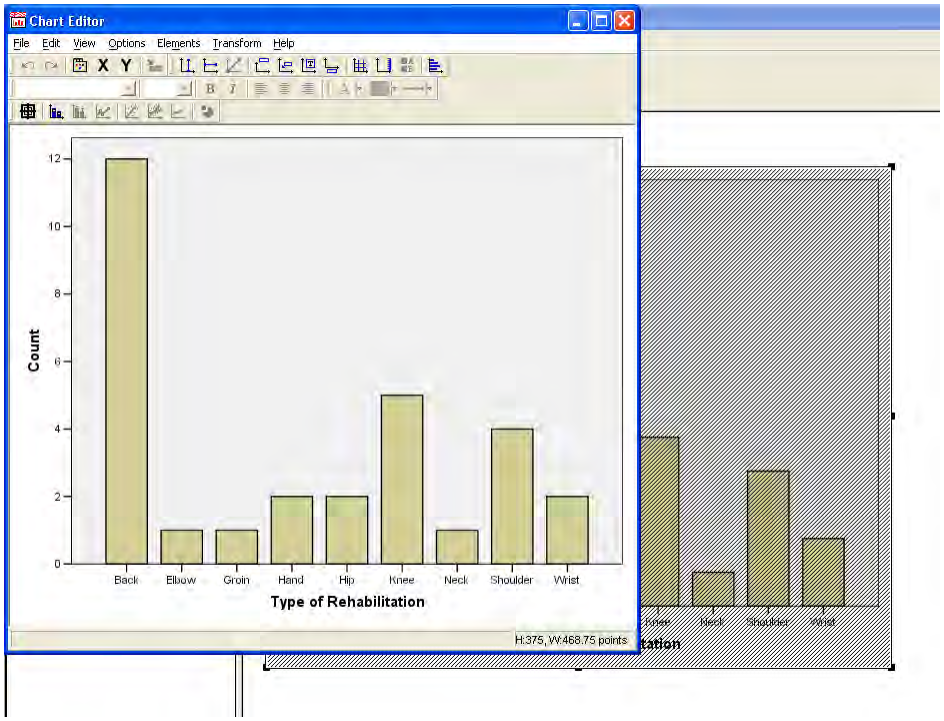


Highlight Type of Rehabilitation, and push the ► key by Category Axis. This selects the category for the bottom of the graph. You can panel, or split the bar plot by placing other variables in the Rows, and Columns boxes. This can be used to compare frequencies among different groups.

For a Frequency Distribution, make sure that **N of cases** is selected. Select OK. For a Relative Frequency Distribution, select % of Cases. Cum. N will provide a cumulative frequency plot, and Cum. % will provide a cumulative relative frequency plot. You can use the Titles button to add titles to the plot.



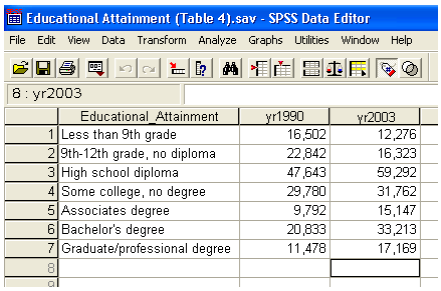
Note that SPSS puts values into Alphabetical Order, so although values are the same, the graph may not look exactly like you would expect. This is ok, as order doesn't matter in a bar graph.



If you click the right mouse button on the graph, you can get to SPSS Chart Object, which opens the graph editor. Here you can change the order of the bars in the **Edit → Properties** menu, as well as changing bar width, changing color, etc.

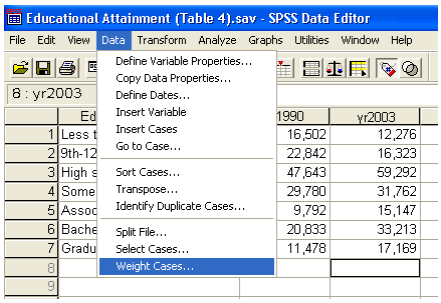
Example 5, page 65

Sometimes data is already summarized. When this occurs, SPSS must know where the summarized information lies. This form of data entry is useful for large amounts of information, and can save on many hours of typing, as each value only needs to be entered once, with another column describing the actual frequency of observations.
To construct a frequency distribution: Type in the data (do not include totals)

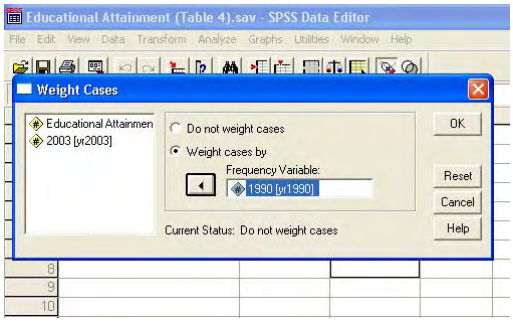


	Educational_Attainment	yr1990	yr2003
1	Less than 9th grade	16,502	12,276
2	9th-12th grade, no diploma	22,842	16,323
3	High school diploma	47,643	59,292
4	Some college, no degree	29,780	31,762
5	Associates degree	9,792	15,147
6	Bachelor's degree	20,833	33,213
7	Graduate/professional degree	11,478	17,169
8			

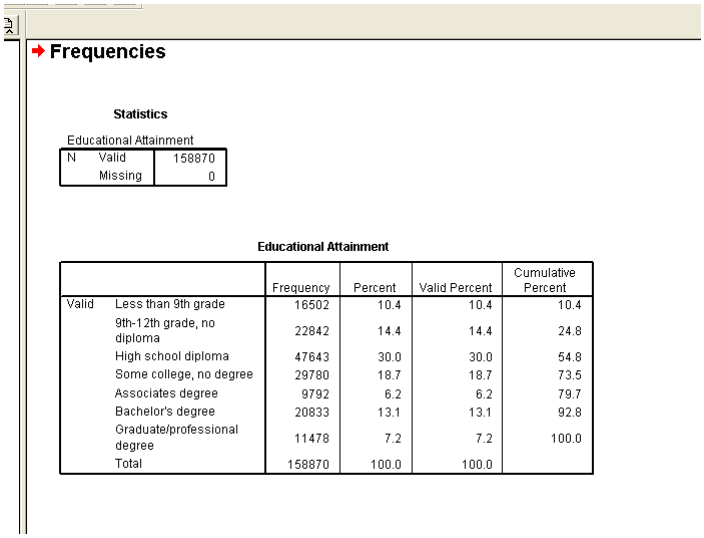
Select Data → Weight Cases



Adding weights allows SPSS to "see" the data a different number of times than what appear in data sheet. Although the actual value was only typed one time, the computer will "see" it as many times as the value actually occurred in the sample. We must select which variable to weight cases by, but only one variable may be selected at a time. You cannot have SPSS see two different sets of data at once.



Select OK. The methods above can now be used to find a Frequency distribution.



Frequencies

Statistics

Educational Attainment

N	Valid	158870
	Missing	0

Educational Attainment

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than 9th grade	16502	10.4	10.4	10.4
	9th-12th grade, no diploma	22842	14.4	14.4	24.8
	High school diploma	47643	30.0	30.0	54.8
	Some college, no degree	29780	18.7	18.7	73.5
	Associates degree	9792	6.2	6.2	79.7
	Bachelor's degree	20833	13.1	13.1	92.8
	Graduate/professional degree	11478	7.2	7.2	100.0
	Total	158870	100.0	100.0	

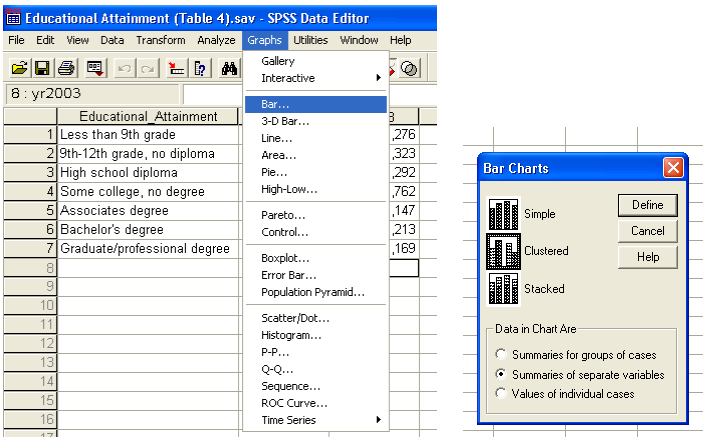
The frequencies should match the original data, and relative frequencies (percent), valid percent (percent when there are missing values), and cumulative relative frequencies (cumulative percent) are given. Note again that the values will be in alphabetical order. If the data is recorded as nominal, SPSS will put these values into either alphabetical order, or ordered by counts, so when typing in the values, it is usually best to include an ordering value, or to use labels. Using numbers for the data and attaching a value to the number will preserve the order without having to have the numbers out in front of the text.

If values are too large to be seen in the window, SPSS will use scientific notation for them. You can click on the table, and enlarge the width of the columns so that the values can be seen well.

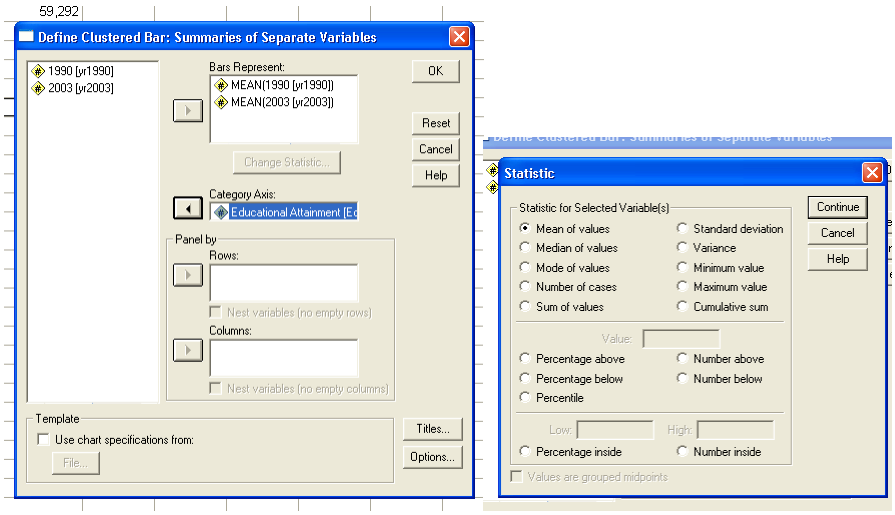
►Creating a Side-by-Side Frequency Bar Graph

We can use the same steps as above to create a regular Bar Graph for an individual set of data. However, this is not the case for a Side-by-Side Bar Graph. We cannot use the weighting method, as we can only weigh data by one set of values at a time. So, to get a side-by-side plot, we must first take off any weights we have used. This is accomplished by going to the weight cases window and clicking on Do not weigh cases. As above, to make sure the order of the bar graph is in the order that we want, we must use variable labels. You can view the data values still by using **View** and checking **Value Labels**.

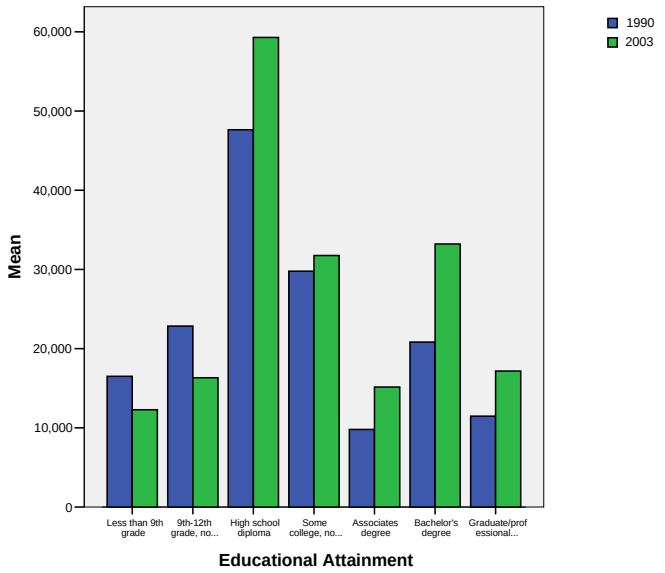
As in the regular bar graph, select **Graphs → Bar…** but, unlike regular bar graphs, select Clustered, for side-by-side, and Summaries of separate variables. This utilizes the summarized information instead of looking for raw data.



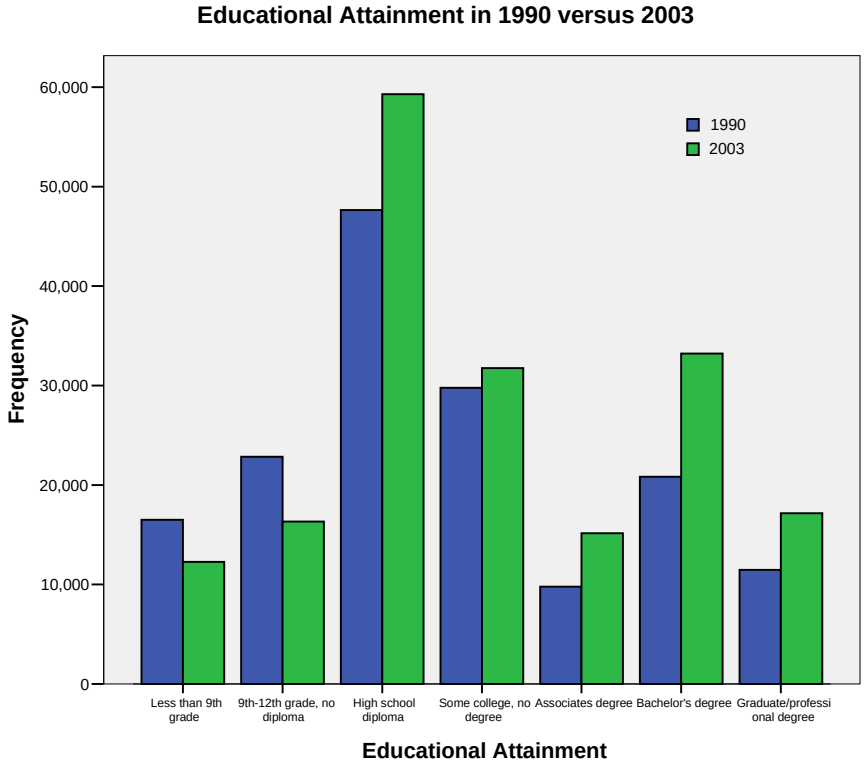
Select Define, and select the columns with the counts for the Bars Represent box, and the values as the Category Axis. The bars must represent a number, and there is a list of functions that can be used. Mean works for the type of data we have, since the mean of a single value is the value itself. Other functions, such as the median, mode, number of cases, sum, standard deviation, variance, minimum, maximum, and cumulative sum can be used. Number of cases sounds like the one that should be used, but would require raw data.



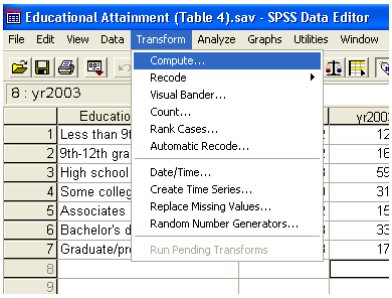
Push OK and the side-by-side frequency graph will be in the output window.



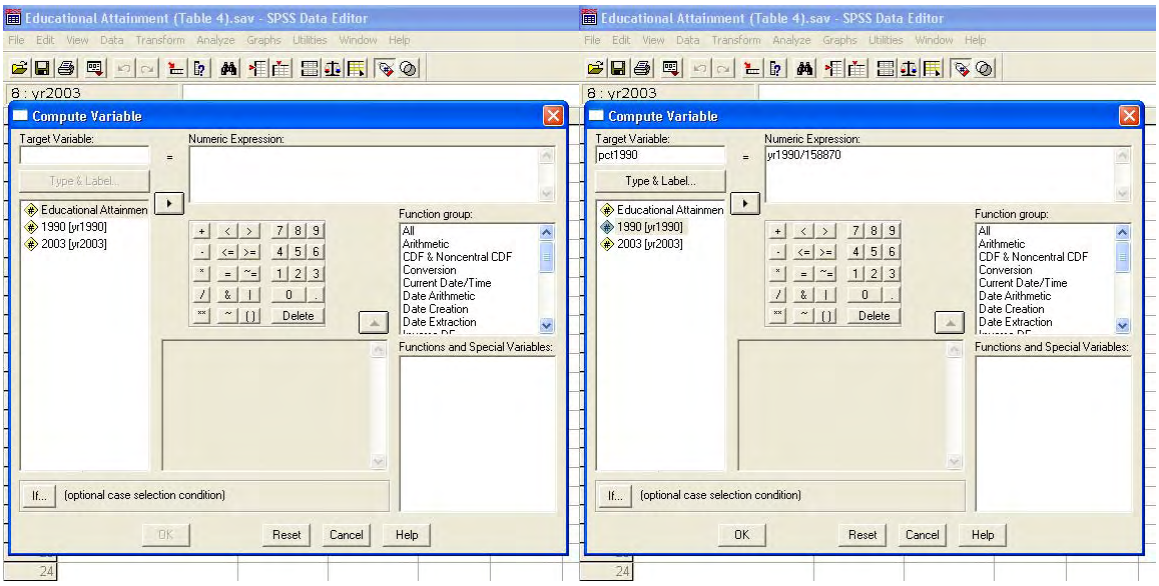
We can use the chart editor (right click on the graph to get to chart editor) to change the label on the y-axis to Frequency instead of Mean. You can also add a title, move the legend, change colors, and make other alterations that may seem necessary.



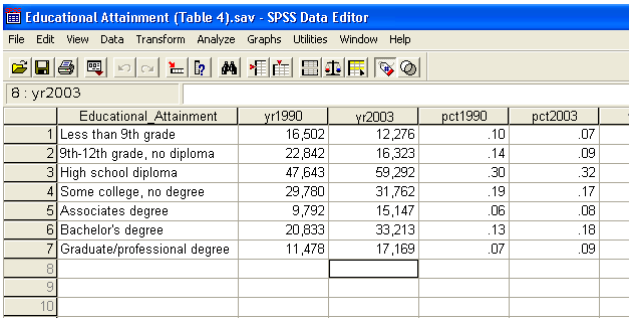
Of course, with side by side bar plots, it is better to use relative frequency rather than actual frequency, so better comparisons may be done. To do this, we will use the SPSS calculator to compute the relative frequencies. Go to **Transform → Compute…**



In the calculator, there are four boxes. The Target Variable is the name for the new column you are going to create, such as pct1990. This name must follow SPSS naming conventions. The second box is the Numeric Expression box. This is where the equation will go. Under the Target Variable box is the current variable box. This provides a list of all the variables currently being used, so you can use them in an equation, and also make sure you don't copy over an existing variable. The fourth box is the function box. This is a list of functions available in SPSS. Also provided is a calculator keypad, although the keyboard can provide all necessary mathematical functions.

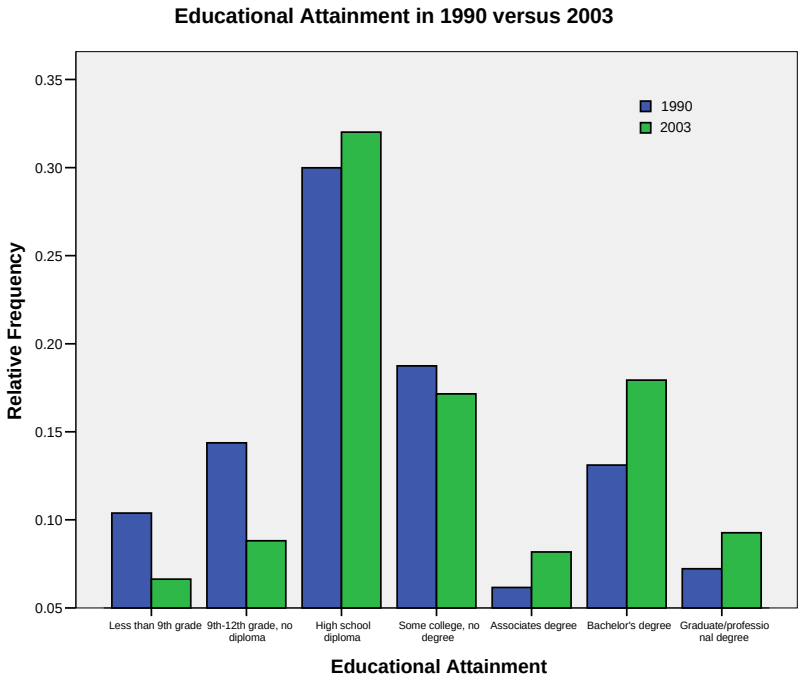


Type the function into the Numeric Expression box, you can either type in the variable name that you want to use, or click on the variable name in the list and press the ► key. This will perform the same calculation for each cell in the variable selected. Doing this for each year will provide us with two new columns in the Data View sheet, the relative frequencies for each year.



	Educational_Attainment	yr1990	yr2003	pct1990	pct2003
1	Less than 9th grade	16,502	12,276	.10	.07
2	9th-12th grade, no diploma	22,842	16,323	.14	.09
3	High school diploma	47,643	59,292	.30	.32
4	Some college, no degree	29,780	31,762	.19	.17
5	Associates degree	9,792	15,147	.06	.08
6	Bachelor's degree	20,833	33,213	.13	.18
7	Graduate/professional degree	11,478	17,169	.07	.09
8					
9					
10					

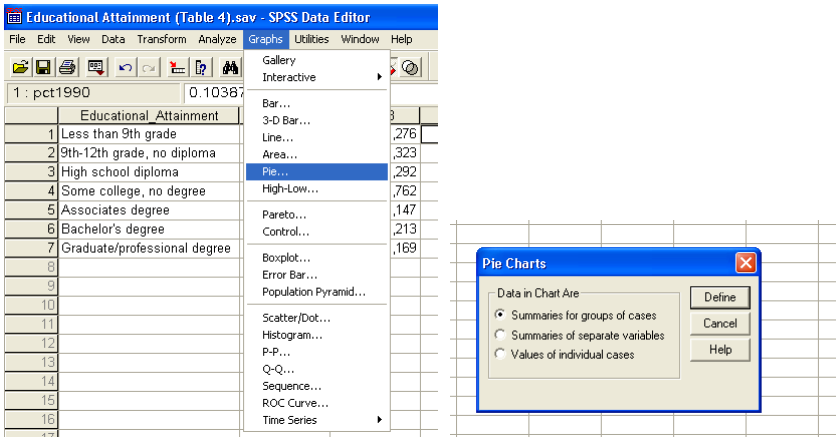
We can then redo the bar plot, using the new variables instead of the counts.



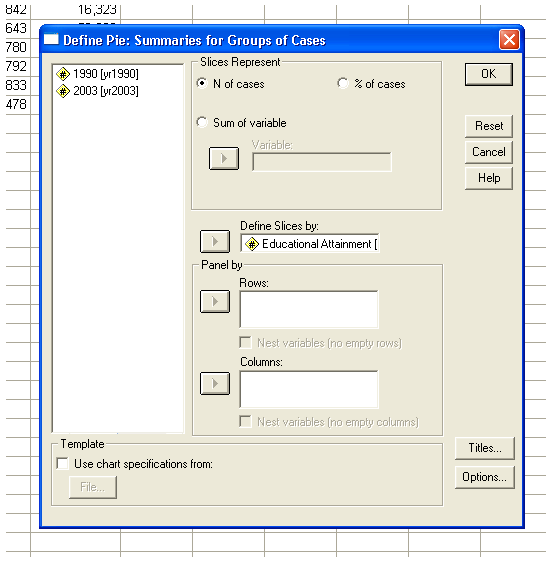
►Creating a Pie Chart

Example 6, page 66

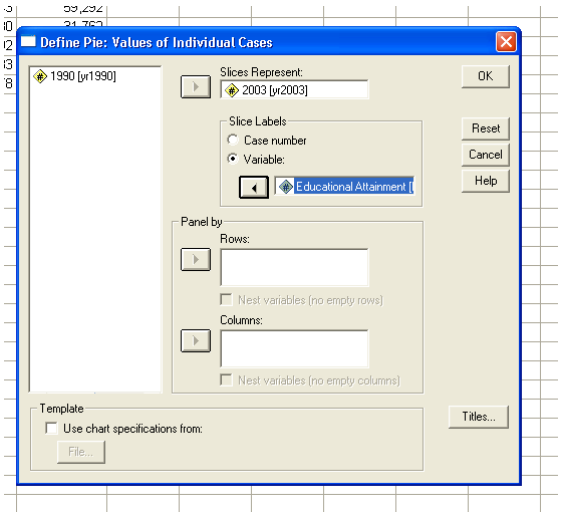
Using the same data as above, the methods for producing a pie chart are similar to producing a frequency table. We follow all the steps of producing a frequency table until we get to the table itself. Instead of creating a table, we go to **Graph → Pie…**,



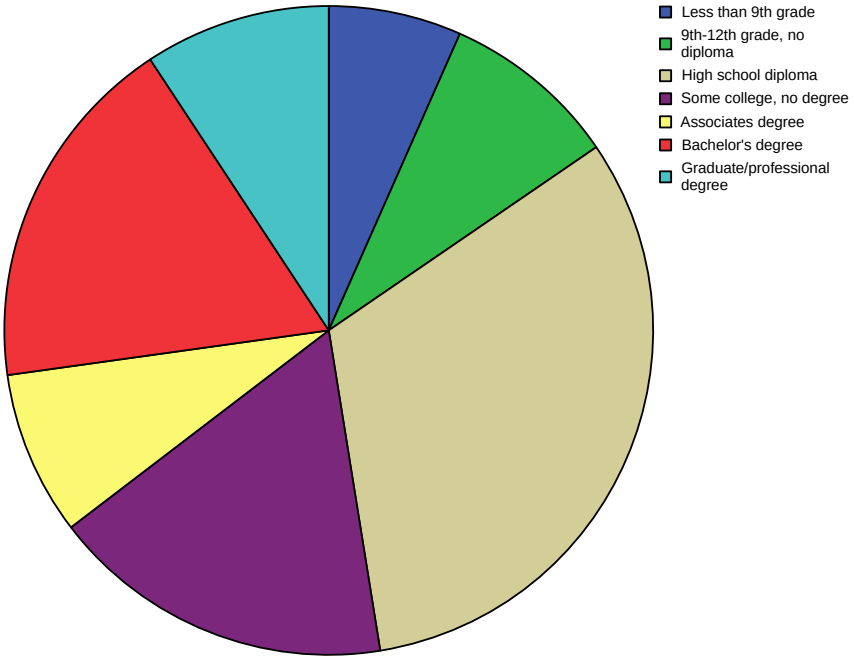
For the raw data, or weighted summarized data, use the Summaries for groups of cases, and push Define. This will open the Summaries for Groups of Cases window. We then select the variable as the Define Slices by, and if we want to create a pie chart based on N(umber) of cases, % of cases, or the sum of another variable.



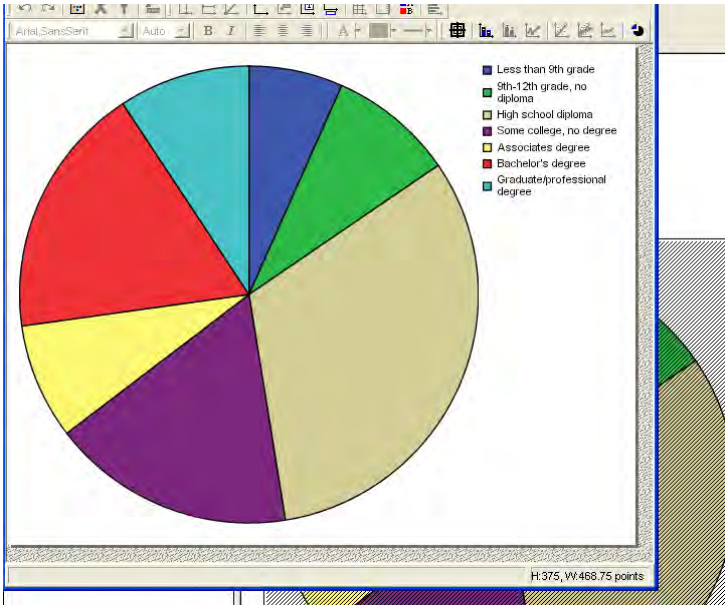
For un-weighted summarized data, use the Values of individual cases.

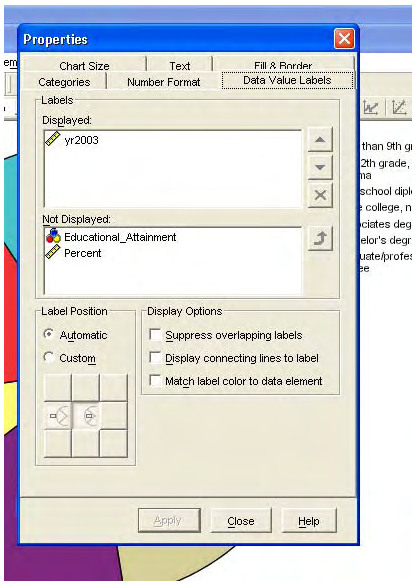


Select the counts as the Slices Represent, and the variable categories as the Variable in the Slice Labels, and push OK.

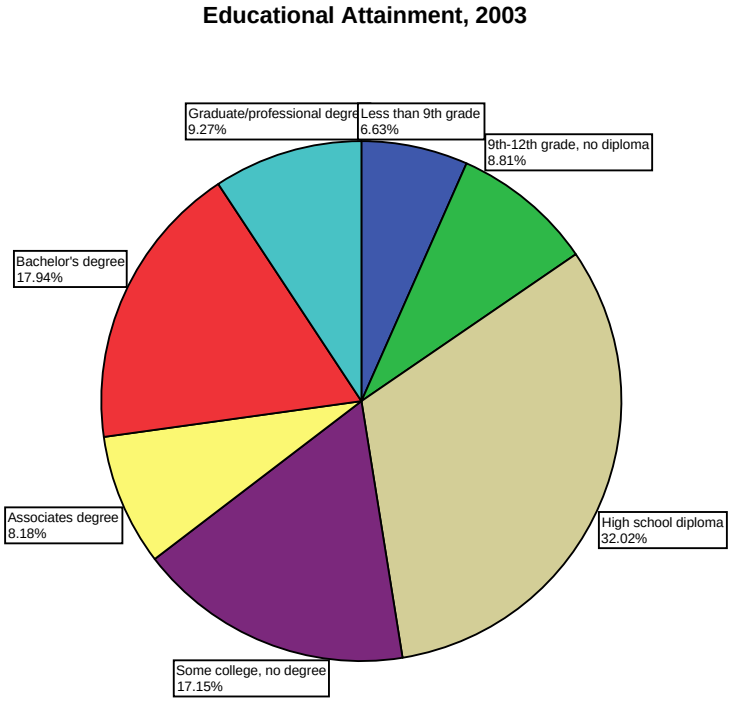


The colors and order can be changed by going to the SPSS chart object (right mouse button), and the percent labels can be added there as well. To add the labels, go to the SPSS chart object, and go to **Elements →Show Data Labels**





In the Properties window, select Percent and hit the green arrow button. The values in the Displayed box will be displayed on the graph. The ruler by the variable represents a numerical value, while the three colored balls represent a string. The Label Position will let you control where on the plot the value will appear (inside the wedge or outside the wedge). Other options may be played with as well.

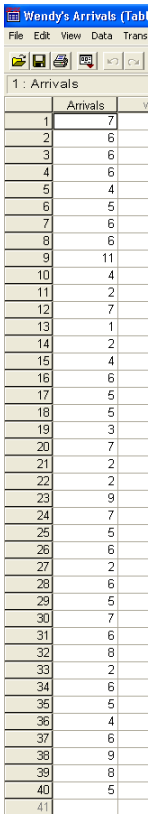


Section 2.2 Organizing Quantitative Data: The Popular Displays

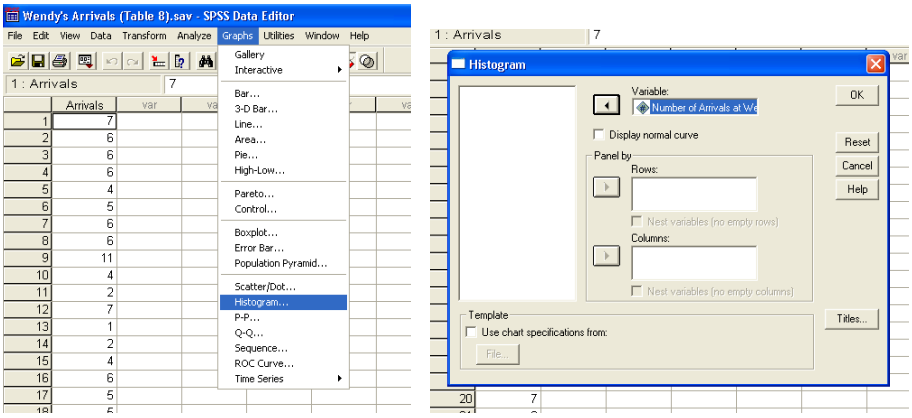
►Creating a Histogram

Example 2, Page 78

To create a histogram, first, we enter the data values.

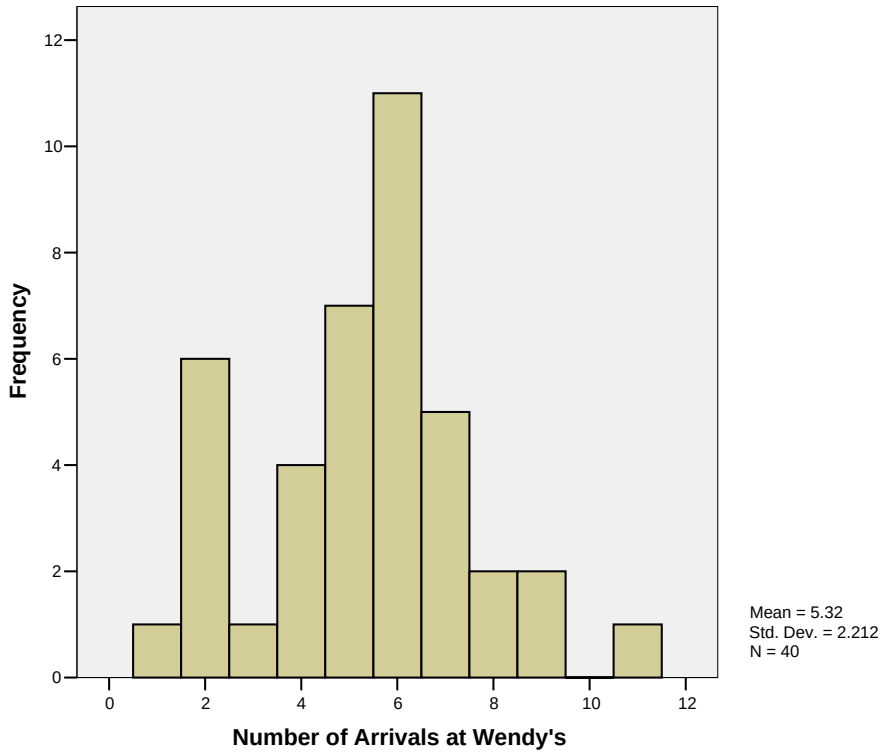


Once all 40 values have been entered, we are ready to create the histogram. Go to **Graphs → Histogram…**

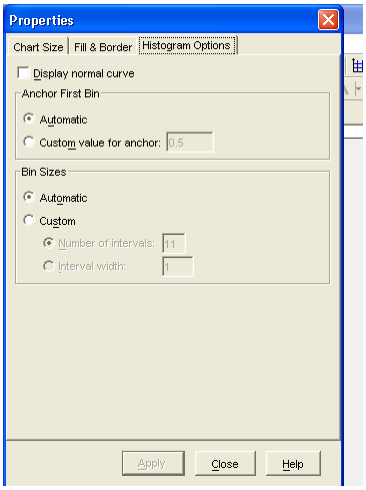


This will open up the Histogram window. Highlight the variable you are using to create the histogram and push the ► button next to the Variable box. You can add titles by clicking on the Titles… button. This

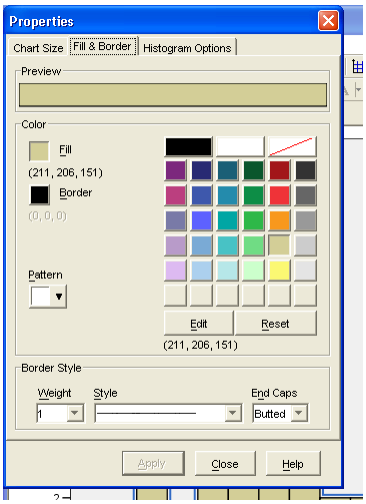
allows for a two line title, a subtitle, and two lines of footnotes. Once you are ready, push OK. The histogram will appear in the Output window.



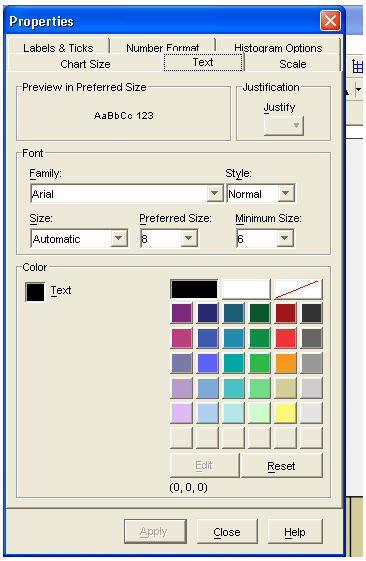
You can change the colors, and how many values are displayed on the horizontal axis by opening the SPSS Chart Object. Using the right mouse button, click on the histogram to open the chart object. In the chart editor window, click on the histogram, and select **Edit → Properties**. This will open the Histogram Options box.



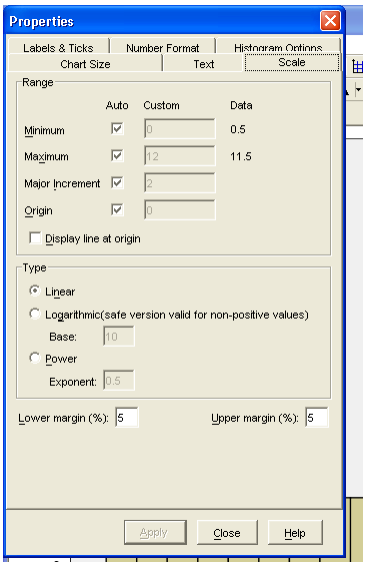
The histogram Options will allow you to change where the lowest point on the graph will be (but you cannot set it at a higher point than already is set by SPSS, and once you change it to a lower value, it cannot be increased again). You can change the number of classes, as well as the class width as well. To change colors, click on the Fill & Border tab.



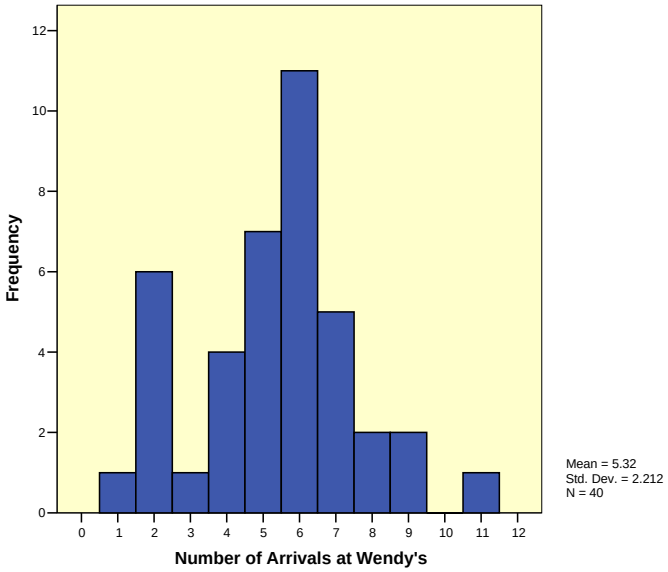
Since you clicked on the histogram before choosing the preferences, the color that appears is the color for the bars of the histogram. To change the color of the background, click on the background and select preferences. To change the numbers on the horizontal axis, click on the numbers on the current axis. Make sure that only the numbers are highlighted. Then go to preferences.



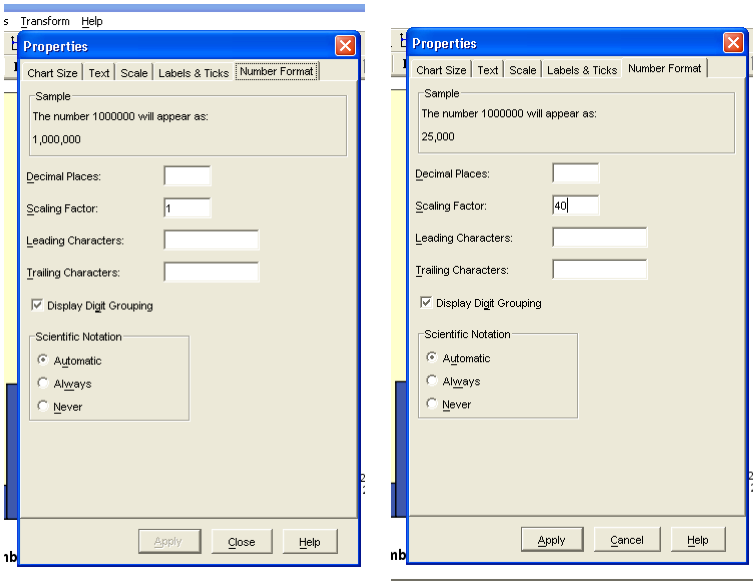
By default, you should enter the Text properties. Here you can change the font, style, and size for the numbers. Clicking on the Scale tab will take you to where you can change the numbers themselves.



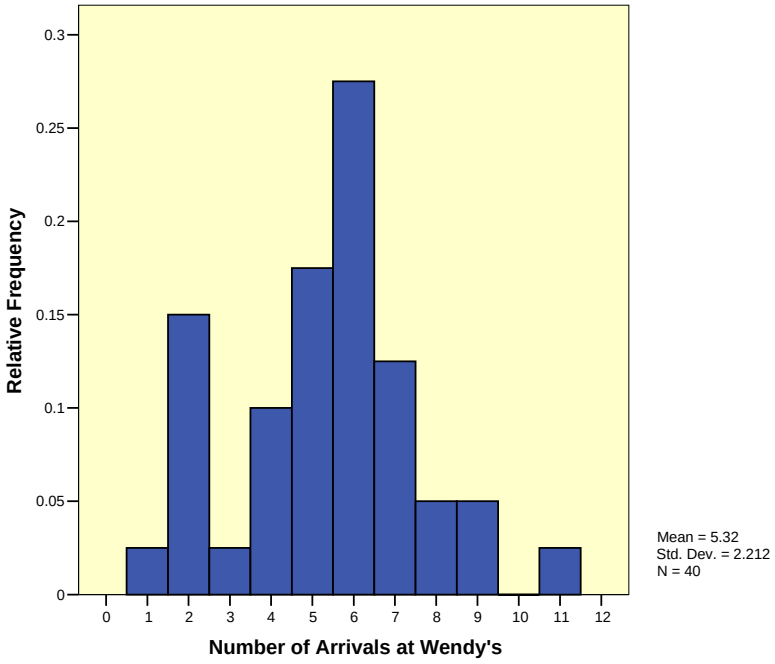
By changing the Major Increment, you change how many values will be shown. For example, to show all values in the Wendy's histogram, change the Major Increment to 1 instead of 2.



So, by changing the color, and the numbers on the horizontal axis, we can create a histogram that looks like the one on page 79. To change this into a relative frequency histogram, again go to the SPSS Chart Object. Click on the vertical axis (the numbers on the vertical (Frequency) axis should be the only things highlighted) and go to properties. Click on the Number Format Tab.



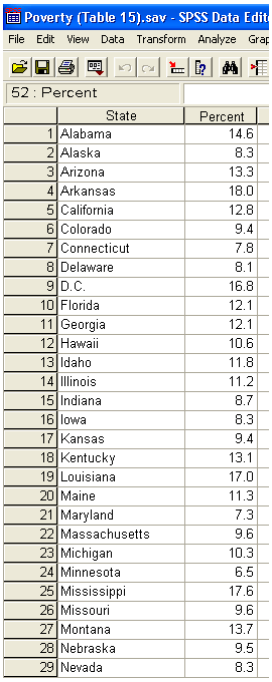
Where you see the scaling factor of 1, change this to the sample size. If you are unsure of the sample size, it appears to the right of the histogram, in this example, 40. This will change the scale to be divided by 40, which is the relative frequency. You will also need to change the label. After pushing Apply and close, highlight the Frequency label until it opens into a text box (it will change to horizontal to make typing easier), and type in what you want the label to be.



►Creating a Stem-and-Leaf Plot

Example 6, Page 82

To create a stem-and-leaf plot, first, we enter the data values.

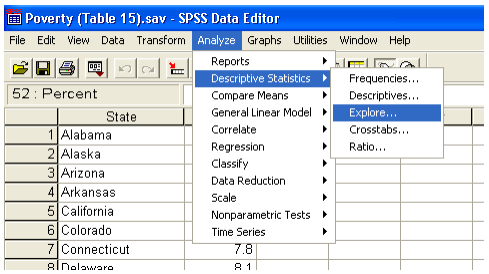

Poverty (Table 15).sav - SPSS Data Edit

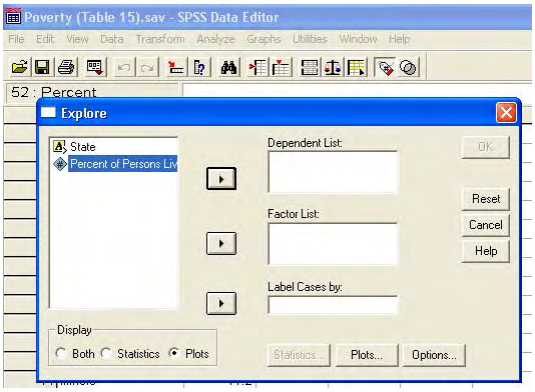
File Edit View Data Transform Analyze Grap

52 : Percent

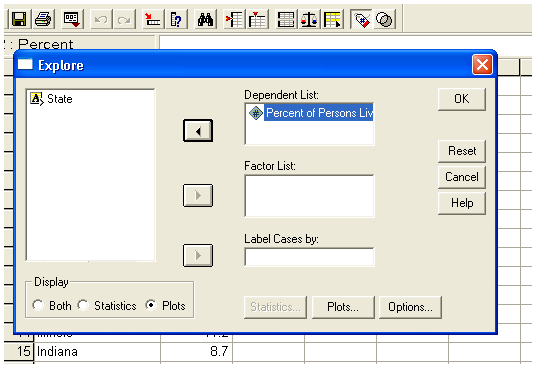
	State	Percent
1	Alabama	14.6
2	Alaska	8.3
3	Arizona	13.3
4	Arkansas	18.0
5	California	12.8
6	Colorado	9.4
7	Connecticut	7.8
8	Delaware	8.1
9	D.C.	16.8
10	Florida	12.1
11	Georgia	12.1
12	Hawaii	10.6
13	Idaho	11.8
14	Illinois	11.2
15	Indiana	8.7
16	Iowa	8.3
17	Kansas	9.4
18	Kentucky	13.1
19	Louisiana	17.0
20	Maine	11.3
21	Maryland	7.3
22	Massachusetts	9.6
23	Michigan	10.3
24	Minnesota	6.5
25	Mississippi	17.6
26	Missouri	9.6
27	Montana	13.7
28	Nebraska	9.5
29	Nevada	8.3

Once the data have been entered, go to **Analyze → Descriptive Statistics → Explore…**.





In the Explore window, highlight the value you wish to use, which must be numerical, and click on the ► button by the Dependent List box. Clicking on the Plots button, you will see that stem-and-leaf is a default plot for the data exploration. A box-plot is also default, but you can remove the box-plot by checking the **none** radio button under box-plots. To only get the stem-and-leaf plot, click on the Plots button. Statistics provides some summary descriptive statistics, and the both will provide both the statistics and the plots.



Once you are ready, push OK. The stem-and-leaf plot will appear in the output window.

Percent of Persons Living in Poverty, 2002 Stem-and-Leaf Plot

Frequency	Stem & Leaf
1.00	0 . 5
4.00	0 . 6777
16.00	0 . 8888888999999999
11.00	1 . 00000011111
8.00	1 . 22233333
5.00	1 . 44445
5.00	1 . 66777
1.00	1 . 8

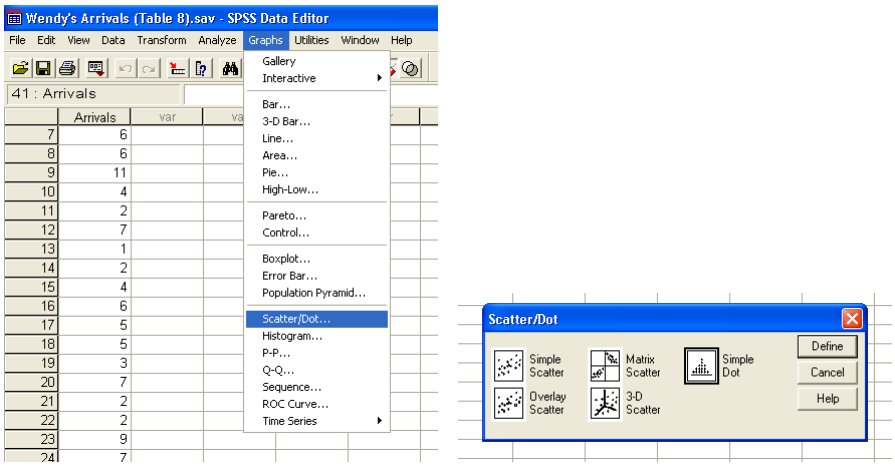
Stem width: 10.0
Each leaf: 1 case(s)

Unfortunately, SPSS does not have a method to change the stem-and-leaf plot. What the program decides to use is what you get. In the SPSS output, the first column is the frequency of leaves for each stem. This is to help show where the median would be. It is not cumulative. The next column is the Stem. SPSS has chosen to use a stem width of 10, so each stem represents the tens place. To the right of the stem is the leaf. Since the stem is in units of 10, the leaves are in units of 1. In this example, SPSS chose a split stem, which means each stem value appears more than once, to make the plot easier to read. So, the first data point in this example has 0 tens, and one 5, so it would be 5. The sixes and sevens appear in the next stem, the eights and nines in the next, etc.

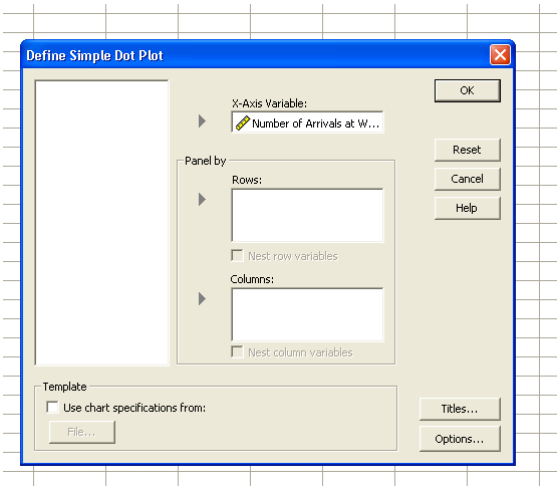
►Creating a Dot Plot

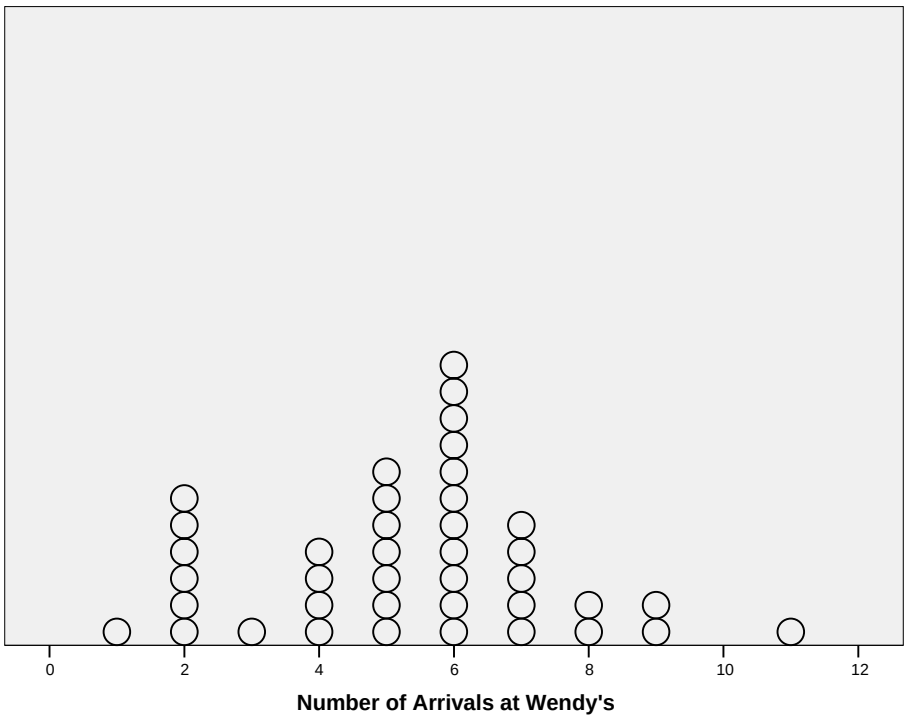
Example 9, page 86

To create a dot plot, we first enter the data, which in this example is the same as in example 1. Once we have the data, go to **Graphs → Scatter/Dot…**. From the five choices for type of graph, select Simple Dot.



Select the variable that you want to make the dot plot of and click on the ► button to place it in the X-Axis Variable. Then, push OK.



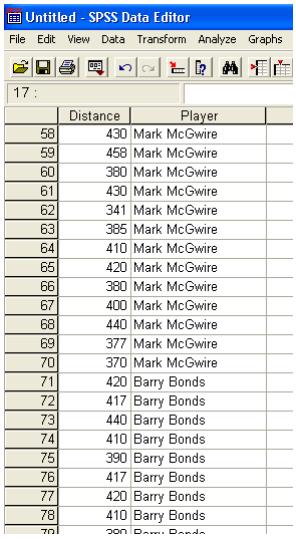


You will have a dot plot of the data. Using the chart editor, you can change the size and shape of the dots, as well as color, labels, titles, etc.

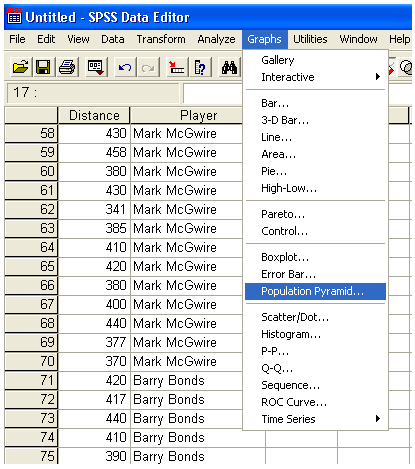
►Creating a Back-to-Back Histogram

Problem 44, page 96

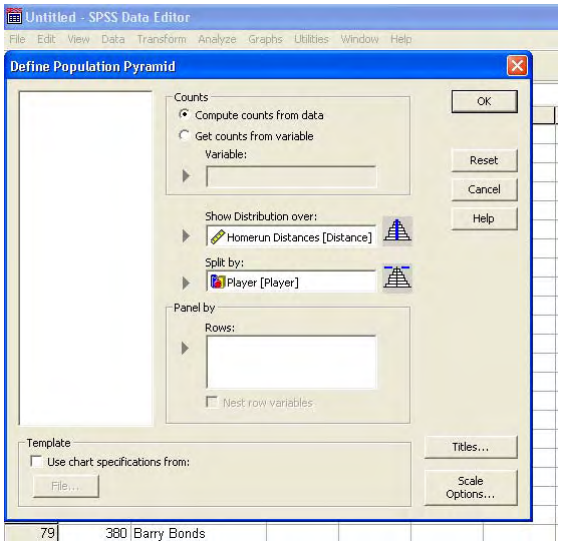
Creating a back-to-back histogram is new to SPSS version 13. If you are using a previous version of SPSS, you cannot use these methods. As will often be the case with SPSS, the data must follow a specific format. This is due to the survey analysis nature of SPSS. In this case, we need to create two columns. One column will be the value of interest, in this case homerun distance. The second column will be the player's name.



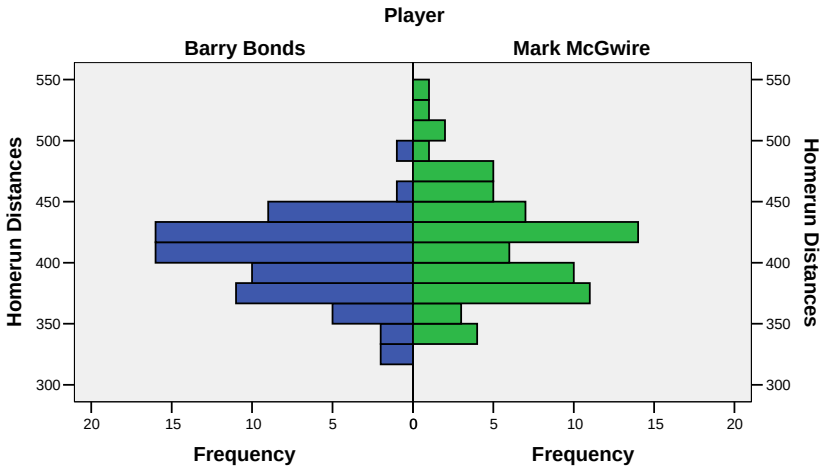
Once the data have been entered, go to **Graphs → Population Pyramid…**



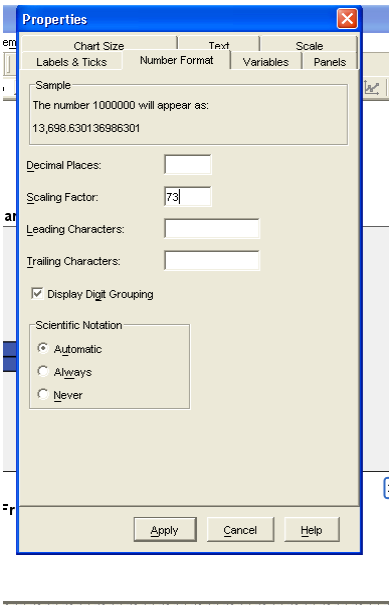
The population pyramid is the back-to-back histogram. Once you get to the Define Population Pyramid window, highlight the category (player) and the ► button by the **Show Distribution Over:** box. Highlight the grouping variable (homerun distance) and the ► button by the **Split by**: box. Once the variables are in the correct places, push OK.



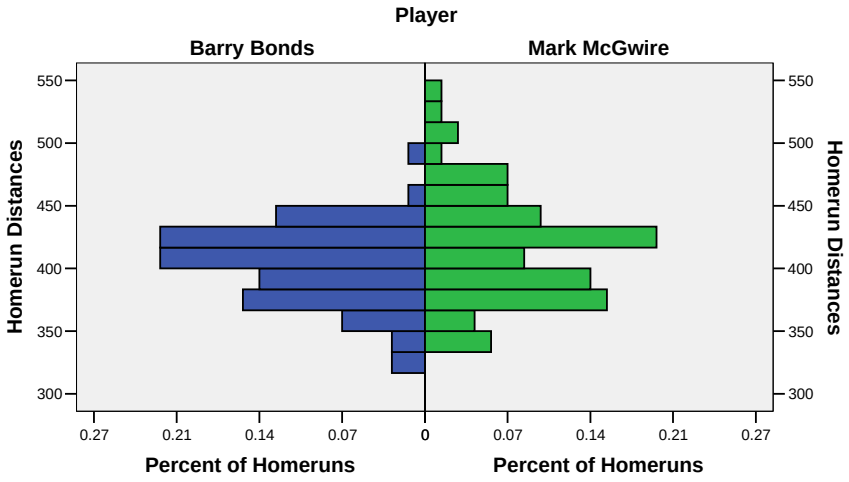
The output window will now contain the back-to-back frequency histogram.



To change this to a relative frequency histogram, go to the Chart Editor, highlight the numbers on the horizontal axis (the frequency), and go to properties. Click on the Number Format tab, and change the Scaling Factor to the larger of the two sample sizes, in this case you can use 73.



This will change the frequency histogram into a relative frequency histogram. You will also want to change the label from sum to percent. Changing the label for one group will change the label for both.

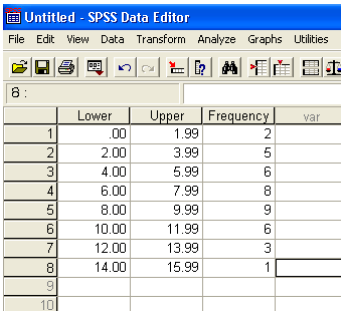


Section 2.3 Additional Displays of Quantitative Data

►Creating a Frequency Polygon

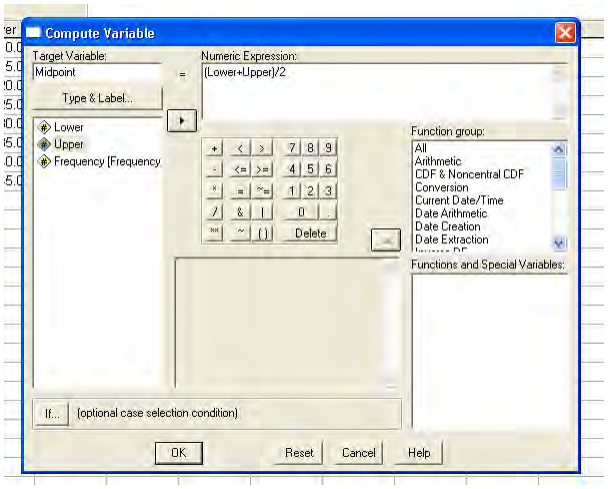
Example 1, Page 97

To create a frequency polygon, first we enter the data. We will create two columns for the classes, and one column for the frequency. Creating two columns for the classes will allow SPSS to do the work for us.

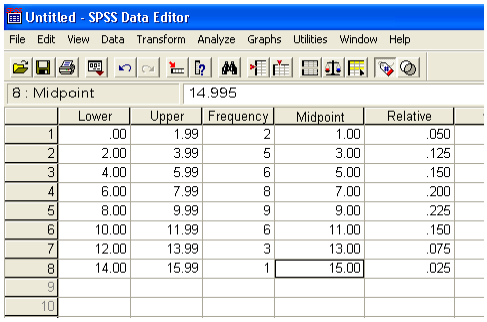


	Lower	Upper	Frequency	var
1	.00	1.99	2	
2	2.00	3.99	5	
3	4.00	5.99	6	
4	6.00	7.99	8	
5	8.00	9.99	9	
6	10.00	11.99	6	
7	12.00	13.99	3	
8	14.00	15.99	1	
9				
10				

Once the data have been entered, we go to **Transform → Compute** to calculate the midpoints. We type the name for the new variable, midpoint, in the Target Variable box. In the Numeric Expression box, we type in the formula. We just need to remember to use parentheses for the addition, so that the numbers are added before dividing by 2.

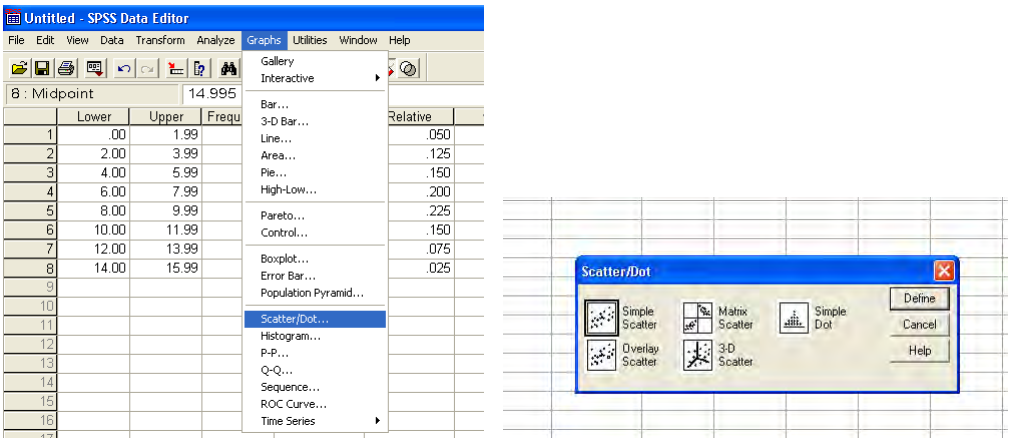


This creates the column of midpoints for us. Similarly, we can use the **Transform → Compute** to calculate the relative frequency.

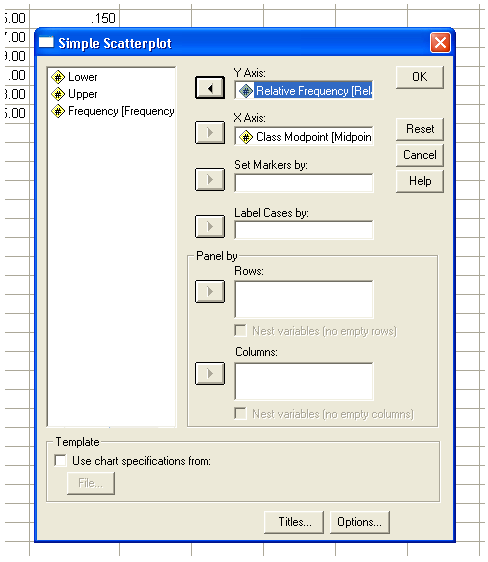


	Lower	Upper	Frequency	Midpoint	Relative
1	.00	1.99	2	1.00	.050
2	2.00	3.99	5	3.00	.125
3	4.00	5.99	6	5.00	.150
4	6.00	7.99	8	7.00	.200
5	8.00	9.99	9	9.00	.225
6	10.00	11.99	6	11.00	.150
7	12.00	13.99	3	13.00	.075
8	14.00	15.99	1	15.00	.025
9					
10					

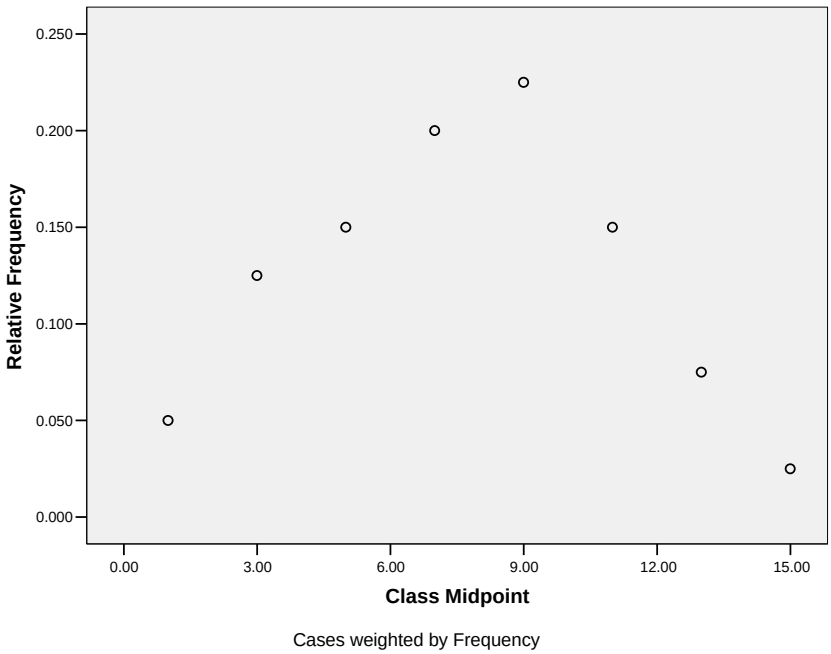
Once the data have been entered, we go to **Graphs → Scatter…**



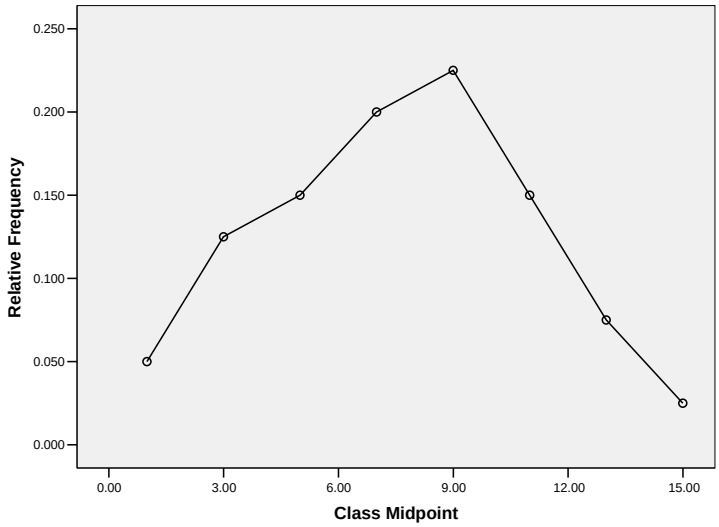
Select Simple Scatter and push the Define button. In the Simple Scatterplot window, select the frequency, or relative frequency for the Y-Axis, depending on whether you want a frequency polygon or a relative frequency polygon. Select the Midpoints as the X—Axis.



Once the variables have been selected, press OK.



This provides the "dots" for the polygon. Go to the Chart Editor. Once in the Chart Editor, click on any of the dots, and go to **Elements → Interpolation Line**. The line is automatically added, so push close, and close the Chart Editor. This connects the dots, and provides your (relative) frequency polygon.

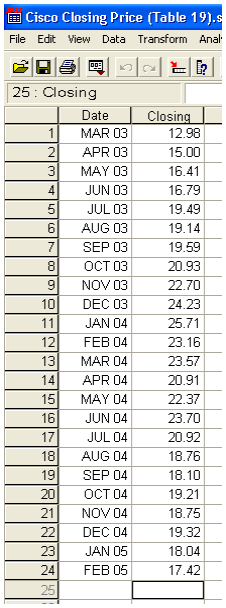


Ogives can be created in the same way, using cumulative frequencies instead of frequencies.

►Creating a Time Series Plot

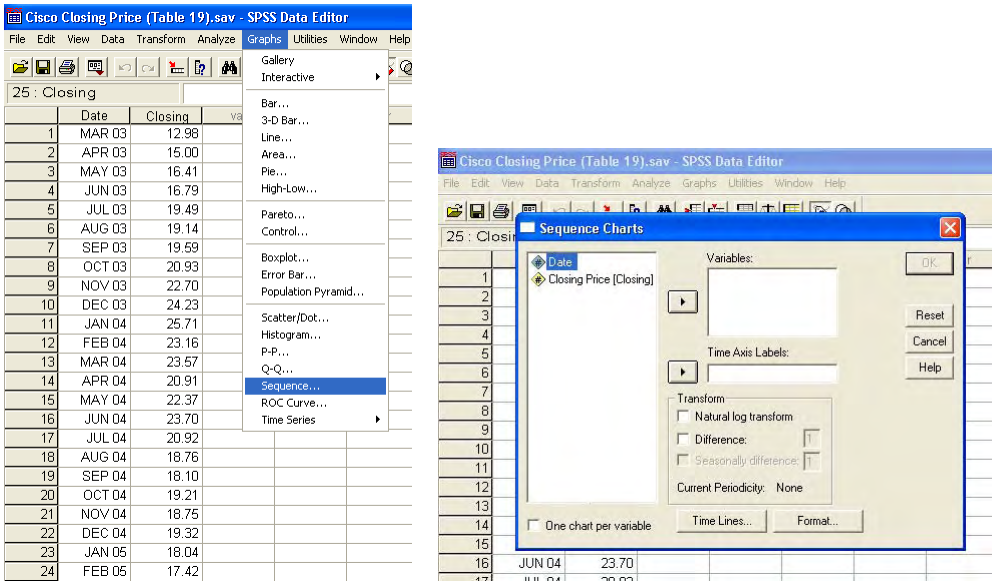
Example 1, Page 100

To create a time series plot, enter the data into SPSS, one column for the dates, and one column for the values.

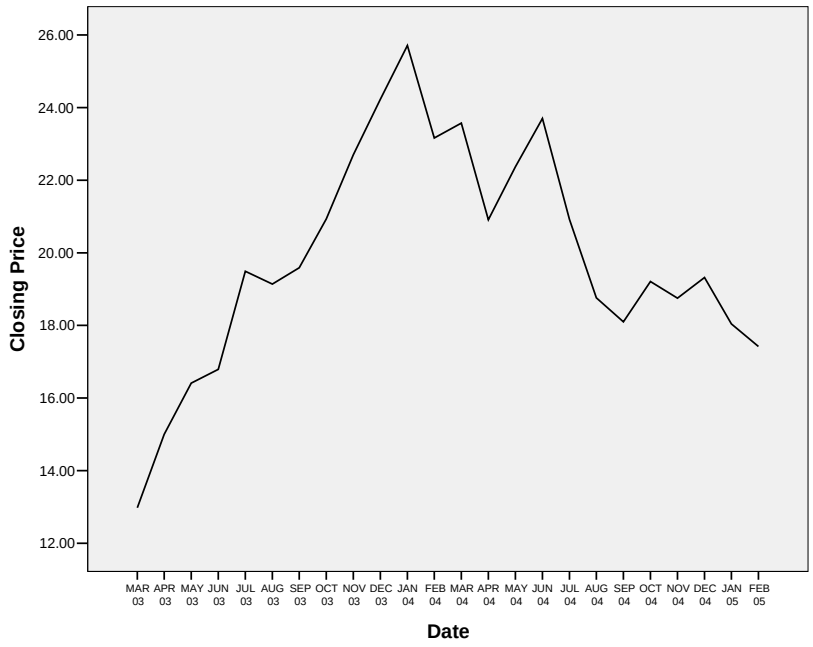


	Date	Closing
1	MAR 03	12.98
2	APR 03	15.00
3	MAY 03	16.41
4	JUN 03	16.79
5	JUL 03	19.49
6	AUG 03	19.14
7	SEP 03	19.59
8	OCT 03	20.93
9	NOV 03	22.70
10	DEC 03	24.23
11	JAN 04	25.71
12	FEB 04	23.16
13	MAR 04	23.57
14	APR 04	20.91
15	MAY 04	22.37
16	JUN 04	23.70
17	JUL 04	20.92
18	AUG 04	18.76
19	SEP 04	18.10
20	OCT 04	19.21
21	NOV 04	18.75
22	DEC 04	19.32
23	JAN 05	18.04
24	FEB 05	17.42
25		

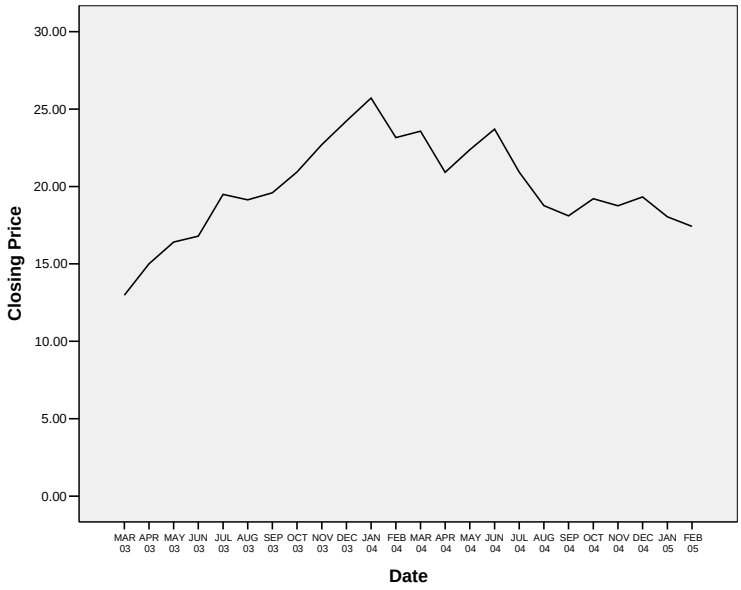
Once the data have been entered, go to **Graphs → Sequence…**. Highlight the date variable and click the ► button next to the Time Axis Labels: box. Highlight the values variable, and click the ► button next to the Variable: box.



Once the variables have been entered, press OK. The time series plot will appear in the Output window.



Using the Chart Editor, you can change the scale, add a title, etc.



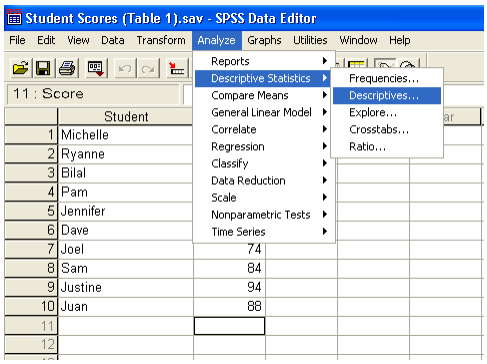
Chapter 3. Numerically Summarizing Data

Finding descriptive statistics is very easy to do in SPSS, but there are many possible ways to do so. SPSS is an analysis tool, not a scholastic tool, and as such, populations are not available in SPSS. SPSS automatically assumes that the information comes from a sample, as most real data does. If the data does represent a population, some adjustments to the SPSS values must be made.

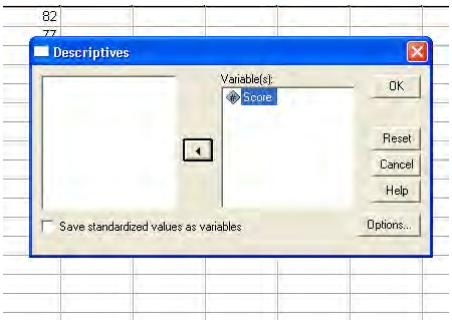
►Finding the basics: mean, standard deviation, percentiles, etc.

Example 1, Page 122

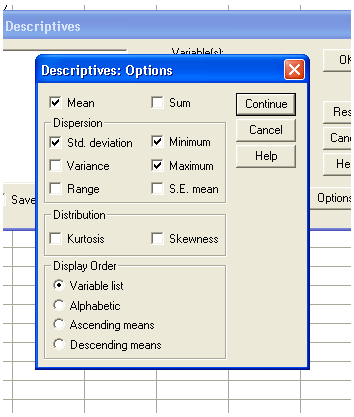
The basic descriptive statistics are the mean and standard deviation. These are very easy to obtain in SPSS. Once the data have been entered, go to **Analyze → Descriptive Statistics → Descriptives…**.



Once in the Descriptives window, highlight the variable(s) you want the mean and standard deviation of, and push the ► button by the Variable(s) window.



Pushing the Options button will allow you to select which descriptive statistics you want to use.

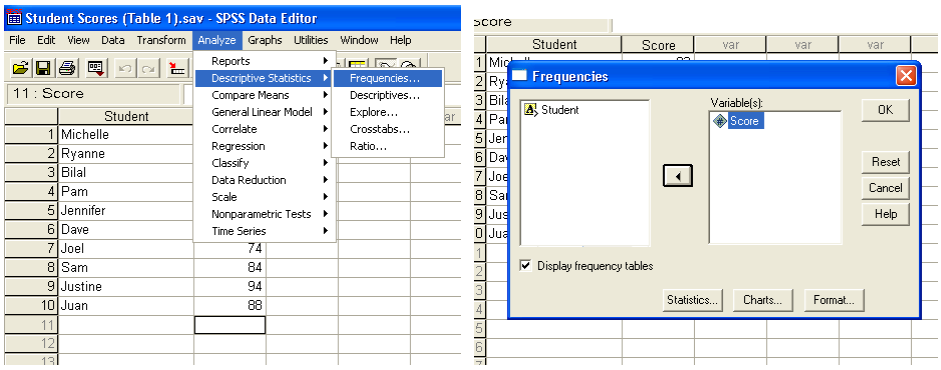


The mean, standard deviation, minimum, and maximum are default values. You can also select the Sum, or total of all the values, the Variance, Range, and standard error of the mean (S.E. mean), which is the standard deviation divided by the square root of the sample size. Also available are the Kurtosis, which measures peakedness, or the amount to which the data values cluster around the center, and the skewness, which measures how asymmetric the values are.
Once you have selected your variable(s), push OK. The results will appear in the output window.

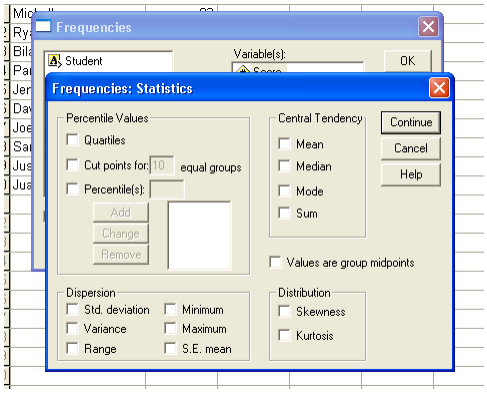
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Score	10	62	94	79.00	10.349
Valid N (listwise)	10				

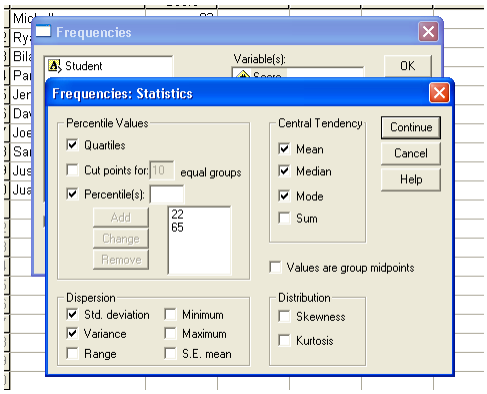
If the list of possible statistics in descriptives is not satisfying, there are more options available by using **Analyze → Descriptive Statistics → Frequencies…**. Once in the Frequencies window, highlight the variable(s) you want the mean and standard deviation of, and push the ► button by the Variable(s) window.



You do not usually want to see a frequency table of numerical data, as this is not a grouped frequency, but actual frequency. Instead you want to select some statistics. If you uncheck the Display frequency tables before selecting some statistics, you will get a warning, but it is only to remind you that without selecting something, there will be no output. Click on the Statistics button.



Once in the Frequencies: Statistics window, you can select which statistics you want. These are grouped into Central Tendency (Section 1) on the top right, Dispersion (Section 2) on the bottom left, Percentile Values (Section 3) on the top left, and Distribution (not in book) on the bottom right.
For measures of Central Tendency, you can select the mean, median, mode (if one exists), and the sum. From dispersion, you can select standard deviation, variance, range, maximum, minimum, and standard error of the mean (standard deviation divided by the square root of n). Again, it must be remembered that these are the sample values, not population values.
From the Percentiles, you can select the quartiles, values to cut the data into equal groups (using both quartiles and cut points for 4 equal groups will be redundant, and SPSS only prints each value once), or specific percentile values. For percentiles, check the box to the left of Percentile(s), and type the value in the box to the right of Percentile(s), and push the Add button. You can ask for as many percentiles as you wish. Again, asking for the quartiles, along with the 25^{th}, 50^{th}, and 75^{th} percentiles is redundant, and will only be printed once.



Once you have selected which statistics you want, push Continue, make sure the Display Frequency Tables is unchecked, and push OK. The results will appear in the output window.

Statistics

Score

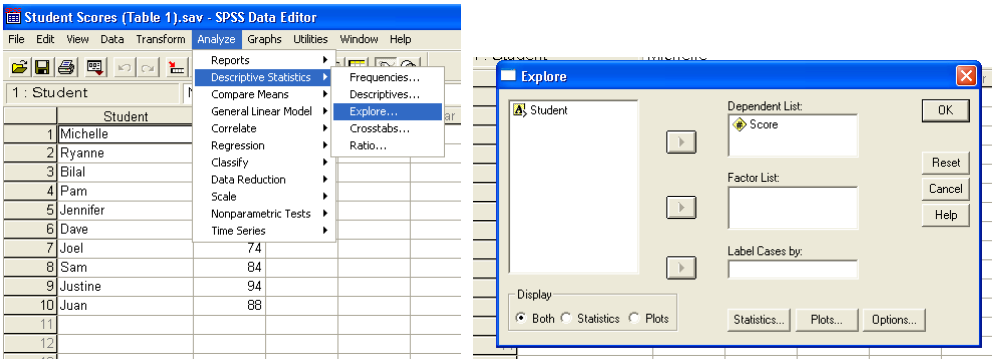
N	Valid	10
	Missing	0
Mean		79.00
Median		79.50
Mode		62[a]
Std. Deviation		10.349
Variance		107.111
Percentiles	22	69.26
	25	70.25
	50	79.50
	65	84.60
	75	88.50

a. Multiple modes exist. The smallest value is shown

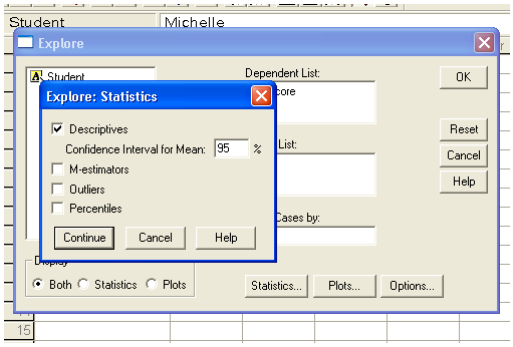
Note that SPSS will attempt to find a mode, but that does not mean that the mode actually exists.

Other calculations, such as the population variance, population standard deviation, inter-quartile range, and z-scores, must be found by hand. SPSS does not calculate these directly.

Another method to obtain summary statistics is **Analyze → Descriptive Statistics → Explore**.

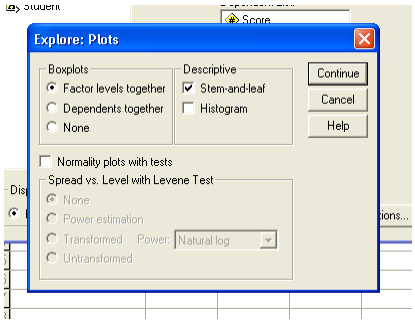


Choose the variable of interest and put it in the Dependent List. To select which statistics you want displayed, click on the **Statistics** button.



Descriptive will provide the mean, median, mode, trimmed mean (which is the mean after dropping out the points which make the outer 5% of the data), the standard error (which is the standard deviation divided by the square root of the sample size, and will be used in later chapters), variance, standard deviation, minimum, maximum, range, IQR, skewness (a measure of how skewed the distribution is, which is not used in this book), skewness standard error, kurtosis (which is a measure of how much of the data is in the center, also not used in the book), and kurtosis standard error. You will also be provided with a Confidence interval for the Mean, which will be addressed in chapter 8. M-estimators are better estimators of the center when there are extreme values, but are not addressed in this book. Outliers will provide the 5 smallest and 5 largest observations so you can check if they are outliers. Percentiles provides the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentiles.

Clicking on the Plots option will provide the options for creating a Boxplot, Stem-and-Leaf plot, and Histogram. You can choose to have only statistics, only plots, or both in the output.



The output will look similar to the below, using the default options.

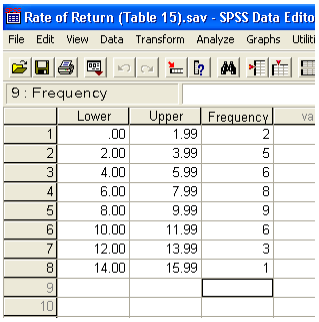
Descriptives

			Statistic	Std. Error
Score	Mean		79.00	3.273
	95% Confidence Interval for Mean	Lower Bound	71.60	
		Upper Bound	86.40	
	5% Trimmed Mean		79.11	
	Median		79.50	
	Variance		107.111	
	Std. Deviation		10.349	
	Minimum		62	
	Maximum		94	
	Range		32	
	Interquartile Range		18	
	Skewness		-.163	.687
	Kurtosis		-1.002	1.334

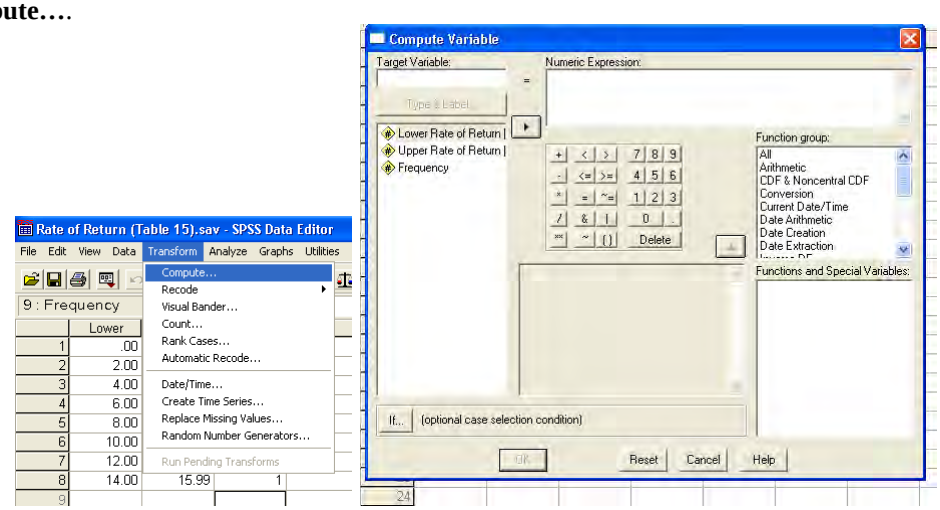
►Finding measures based on Grouped data

Example 1, Page 158

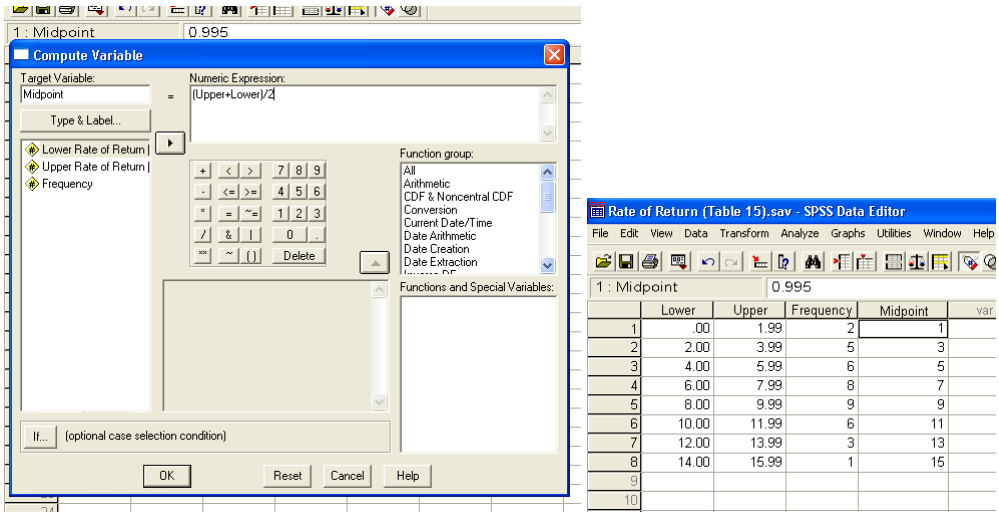
When entering the data, it makes things a little easier to enter the lower class limit and upper class limit as separate variables. This will allow you to make calculations in SPSS rather than by hand.



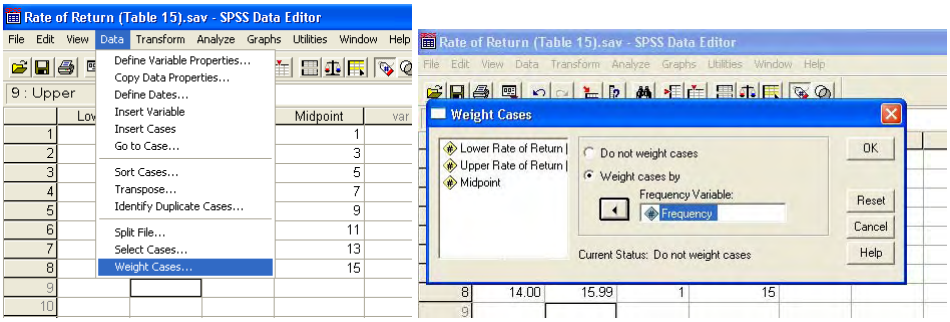
Once the data have been entered, we must find the class midpoints. We can do that by going to **Transform → Compute…**.



In the Target Variable box, create a name for the midpoint. In the Numeric Expression box, we will create the equation (Upper + Lower)/2, which is why it is easier to have these as separate variables. This will not work if you create one text variable to define the classes.



When the equation looks correct, push OK. You will now have a variable of frequencies, and a variable of midpoints. **Go to Data → Weight Cases…**. We want to weigh the cases by the frequency. This tells SPSS to see the midpoints as many times as we have a frequency instead of once.



Now, we can use the same procedures as for ungrouped data, but we must remember that these are only approximate values.

Descriptives

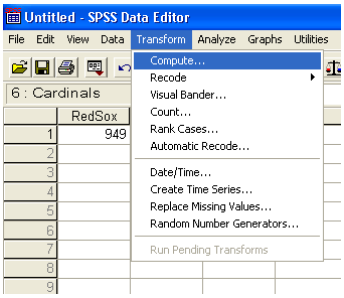
			Statistic	Std. Error
Midpoint	Mean		7.5950	.54632
	95% Confidence Interval for Mean	Lower Bound	6.4900	
		Upper Bound	8.7000	
	5% Trimmed Mean		7.6061	
	Median		6.9950	
	Variance		11.938	
	Std. Deviation		3.45521	
	Minimum		1.00	
	Maximum		15.00	
	Range		14.00	
	Interquartile Range		5.50	
	Skewness		-.020	.374
	Kurtosis		-.581	.733

This is now the weighted mean, median, etc. Percentiles should refer to the group rather than a specific number, as they are referring to midpoints, not actual values.

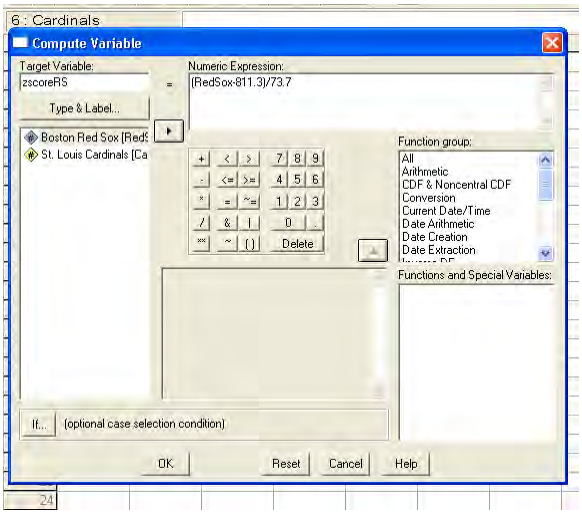
►Finding a z-score

Example 1, Page 166

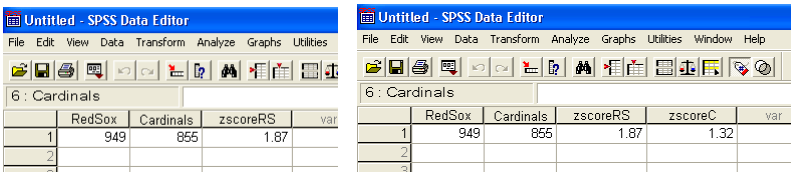
Calculating a single z-score in SPSS is the same as using a calculator. The only real advantage of using SPSS is when you have a list of values to calculate z-scores for, all with the same mean and standard deviation. As SPSS does not calculate population means and variances, these must be done by hand. To calculate the z-score, type in the x value into a cell, then go to **Transform → Compute**, which is the SPSS calculator. You must have at least one value in a cell before being able to use the SPSS calculator.



In the calculator, you create a new variable, so you need to name the variable. This can be any name under the normal SPSS naming conventions (so z-score will not work, as it has a -, but zscore will be fine). This name goes in the Target Variable box. You will type in the z-score equation in the Numeric Expression box. You can use the value(s) of a current variable by typing the name of the variable, or selecting the variable from the current variable list and clicking the ► button. You must watch the order of operations. Use parentheses to ensure that the order in which SPSS will do the math is what you expect.



Once the formula has been correctly entered, push OK. This will create a new column in the Data View window, with the value of the z-score.

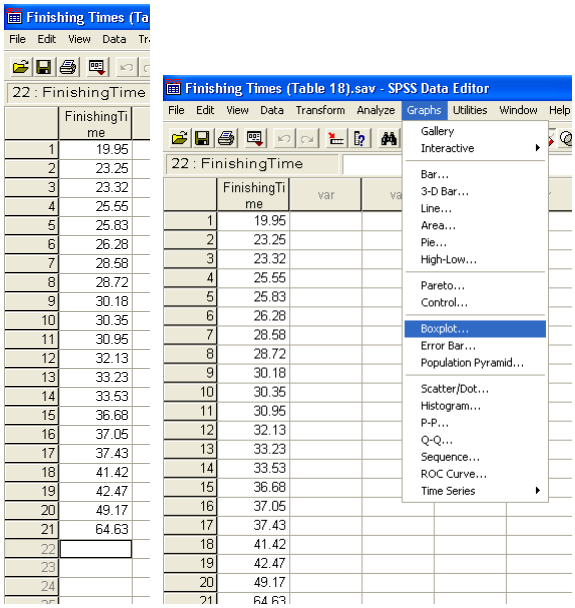


We can then compare the z-values. The Boston Red Sox had a z-score of 1.87, or a score 1.87 standard deviations above the mean, while the St. Louis Cardinals had a z-score of 1.32, or 1.32 standard deviations above the mean, so the Red Sox have a relatively higher score than the Cardinals. The advantage of using SPSS will come when calculating the z-scores for a list of values, such as calculating the z-values for all the teams in the American League, or National League.

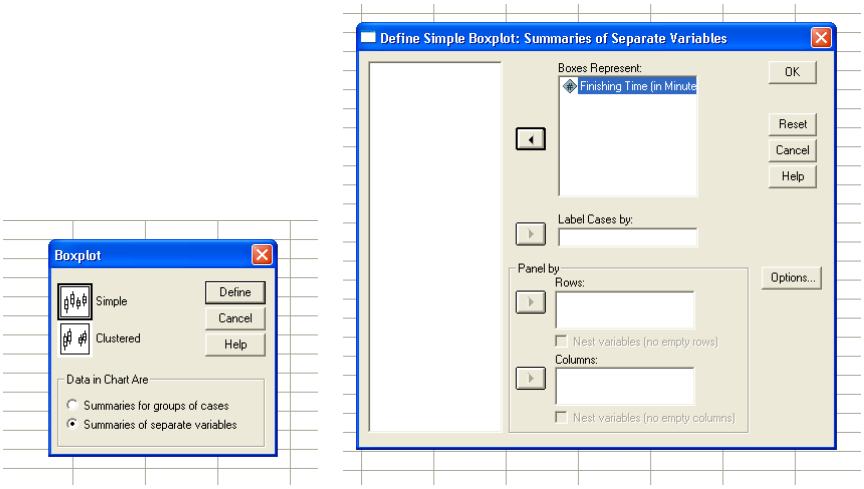
►Creating a box-plot

Example 1, Page 176

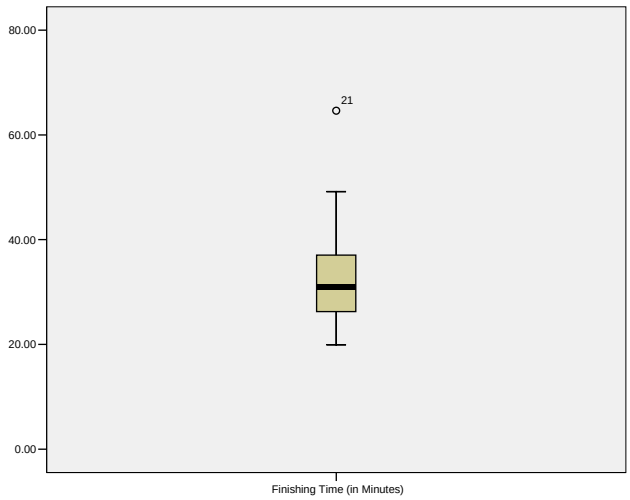
Once the data have been entered, go to **Graphs → Boxplot…**.



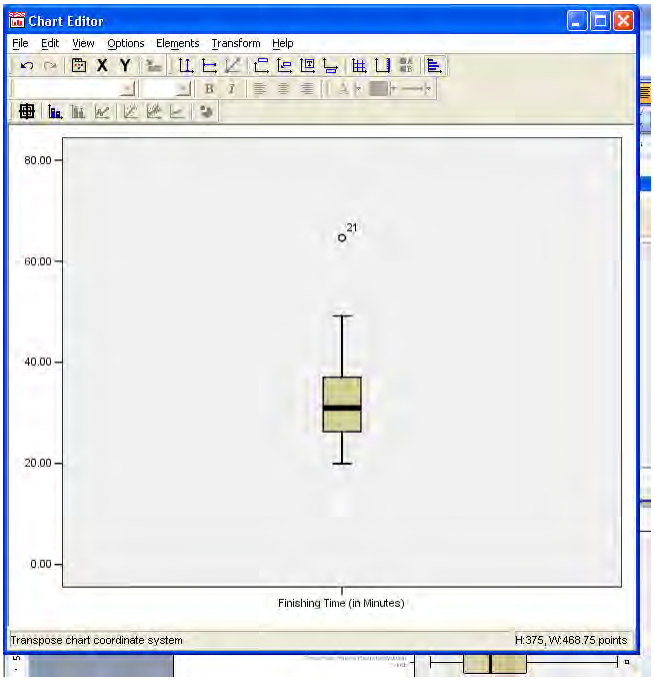
In the Boxplot window, you will see two choices for type of box-plot. The first is the simple box-plot. This is the one that you will use most often. The simple box-plot is for a single, or for grouped data. The clustered box-plot is used when looking at groups within a cluster. An example could be box-plots for males and females within each class (freshman, sophomore, junior, senior). After selecting Simple, the data can be presented in one of two ways. **Summaries for groups of cases** is used when you have all the values for all groups in one column, and which group they belong to in a separate column. **Summaries of separate variables** is used when the data are in separate columns. In this example, since we don't have a group, we will use summaries of separate variables.

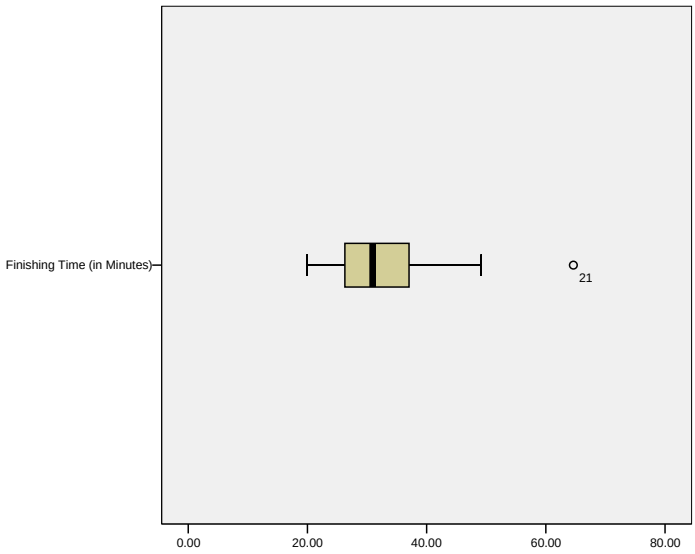


Once in the Define Simple Boxplot: Summaries of Separate Variables window, select the variable of interest, and push the ► button next to the Boxes Represent box. As we don't have a grouping variable, which would go in the Label Cases by: box we just push OK.



The box-plot will appear in the output window. SPSS creates the box-plot by identifying the 1st, 2nd, and 3rd Quartiles, and then the maximum and minimum values inside the lower and upper fences. The median is marked by the solid black line inside the box. Any values that lie outside the fences are marked as outliers. SPSS uses two sets of fences. If a value falls outside 1.5*IQR but inside 3*IQR then an O is used to label the observation as an outlier. If a value falls outside 3*IQR the outlier is marked as an extreme outlier, and an asterisk is used. The observation number is printed by the outlier label so you can check to make sure that observation was entered correctly. Most people don't like to read box-plots from bottom-to-top, as is presented by SPSS. This can be changed by going to the SPSS Chart Object editor. Once in the editor, there is a button on the far right that looks like a histogram on its side. Clicking on this button will transpose the x and y axes, so that the box-plot will be easier to read. The same option appears under **Options → Transpose Chart**.





To conserve space, the output box can also be resized, although it may take some practice to get it to come out how you want it.

Chapter 4. Describing the Relation between Two Variables

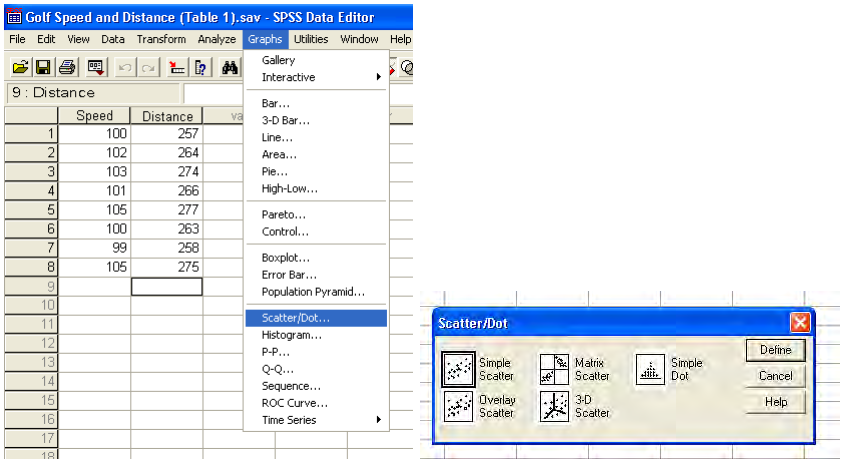
Similar to descriptive statistics, there are various ways to calculate the relationship between variables. Only the simplest methods will be shown here.

Section 4.1 Scatter Diagrams and Correlation

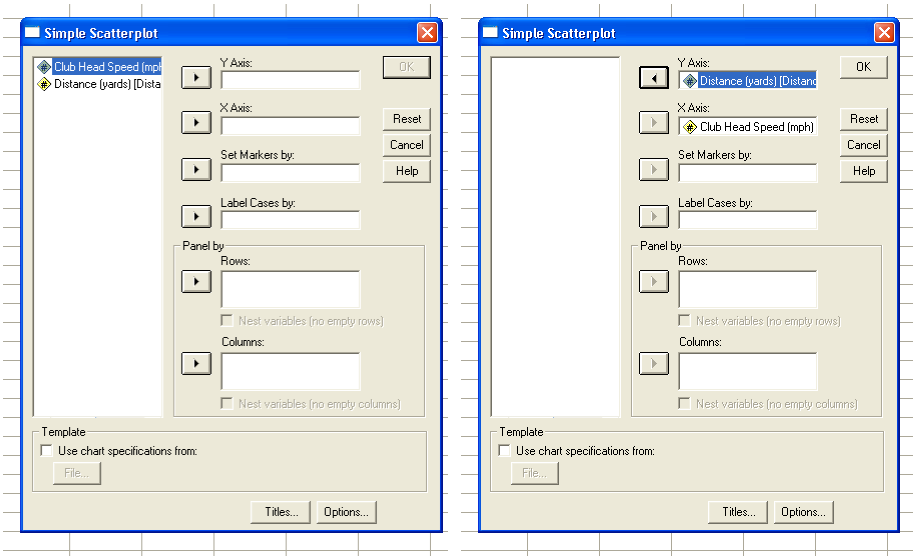
►Creating a Scatter diagram

Example 1, Page 195

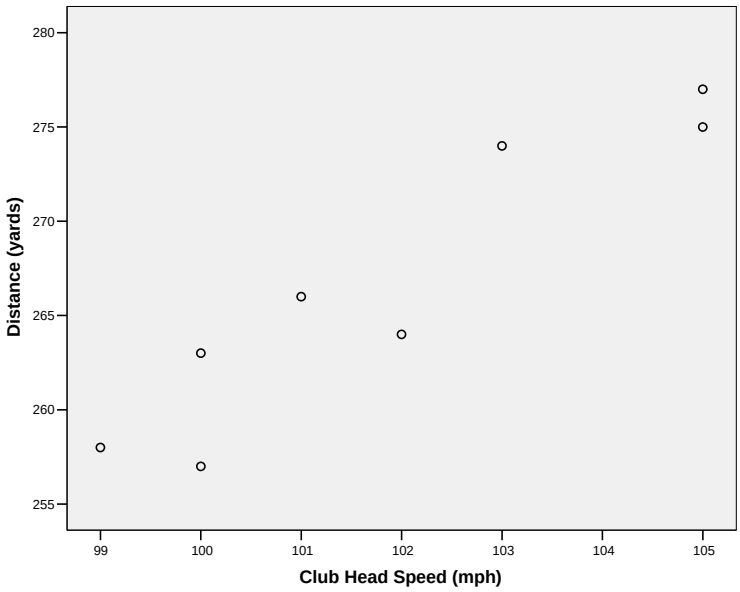
After typing the data into SPSS, go to **Graphs → Scatter/Dot…** and select Simple Scatter.



Selecting simple scatter will open the Simple Scatterplot window. Here you are asked to define the Y-axis, or the response variable, and the X-axis, or the predictor variable. You can also Set **Markers by:** which allows you to set different markers (color or shape) for a group (such as males and females). The **Label Cases by** can be used to identify specific points, if a categorical variable exists.



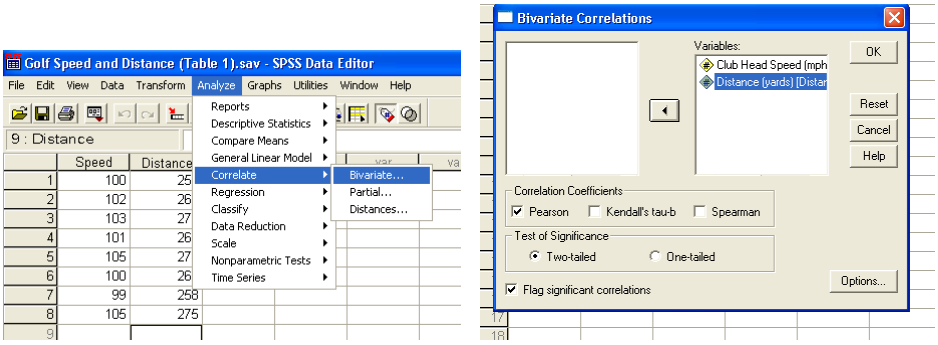
Once you have selected the x-axis and y-axis values, you can create the graph. To add labels, you must select a variable to label cases by, and go to options, and select **Display Chart with case labels**. Once you are ready, push OK. The scatter plot will appear in the output window.



►Finding Correlation

Example 2, Page 200

After entering the data into SPSS, go to **Analyze → Correlate → Bivariate...** (two variables)



Highlight the variables you wish to find the correlation between and push the ► button.

Make sure that Pearson is selected. Spearman may be selected for chapter 13 (Non-Parametric Correlation). Other options (in the Options button) allow you to get the means, and standard deviations, as well as the covariance and cross product deviations.

When you have the options you want, push OK

Correlations

		Club Head Speed (mph)	Distance (yards)
Club Head Speed (mph)	Pearson Correlation	1	.939**
	Sig. (2-tailed)		.001
	N	8	8
Distance (yards)	Pearson Correlation	.939**	1
	Sig. (2-tailed)	.001	
	N	8	8

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

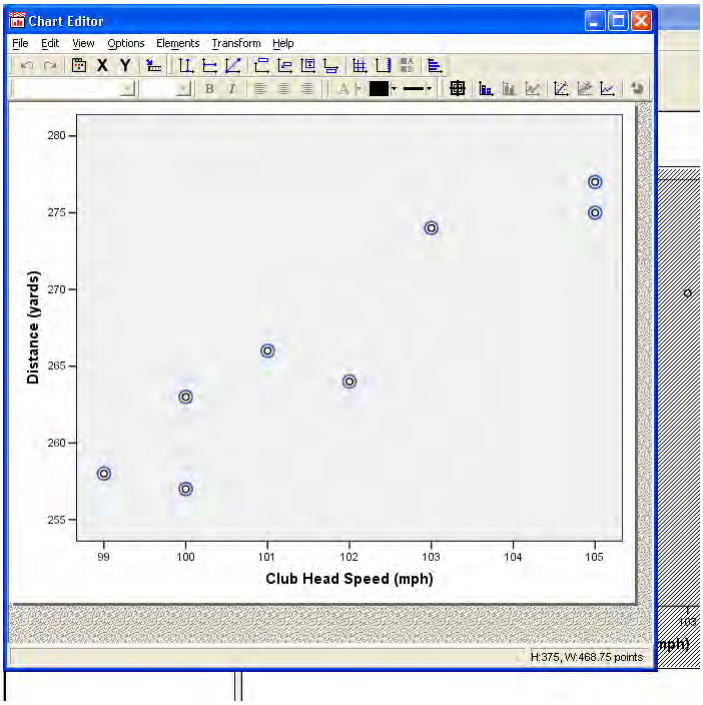
The correlation will appear in the Output Viewer Screen, in this case 0.939. The correlation is provided for all combinations of variables (you can have more than two at a time). The correlation of something with itself is always 1, which appear in the boxes relating the variables to themselves. The other values, Sig. (2-tailed) and the asterisks will not be used until after chapter 9, when hypothesis testing is discussed.

Section 4.2 Least Squares Regression

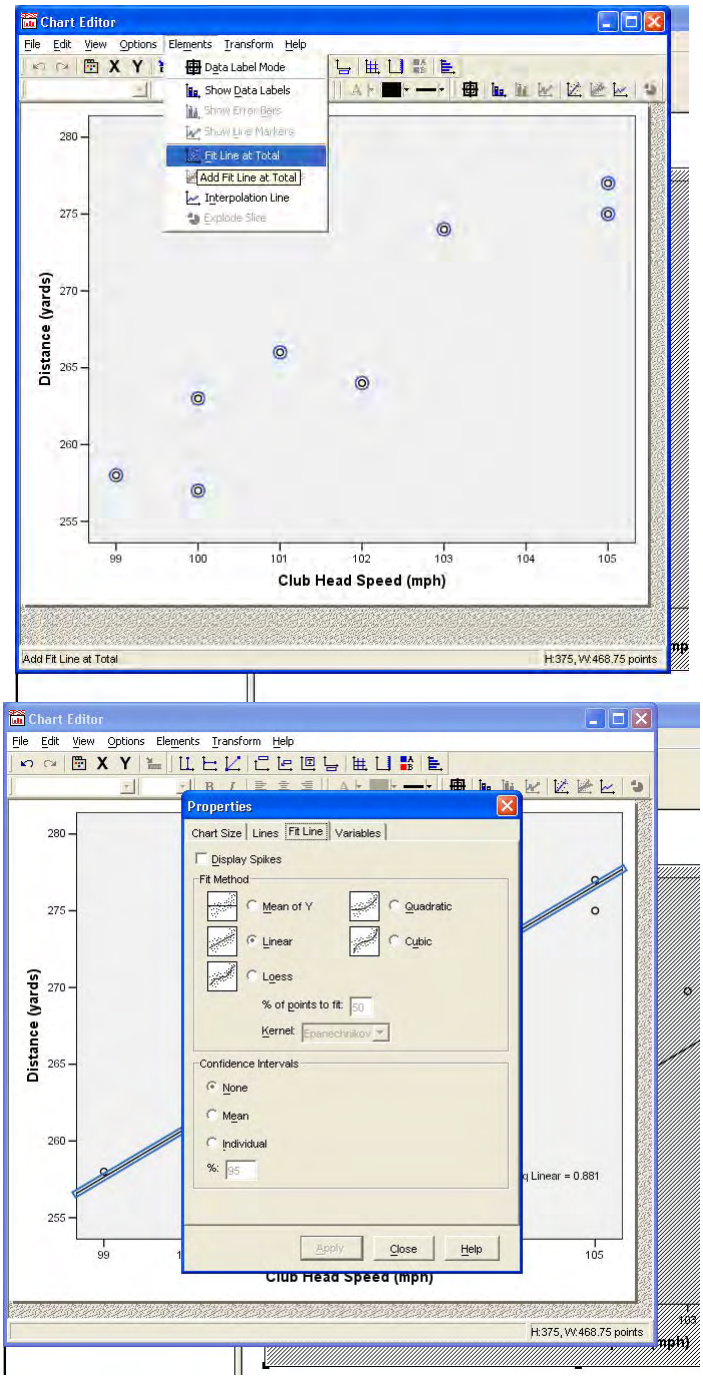
►Fitting a line in a Scatter diagram

Example 1, Page 213

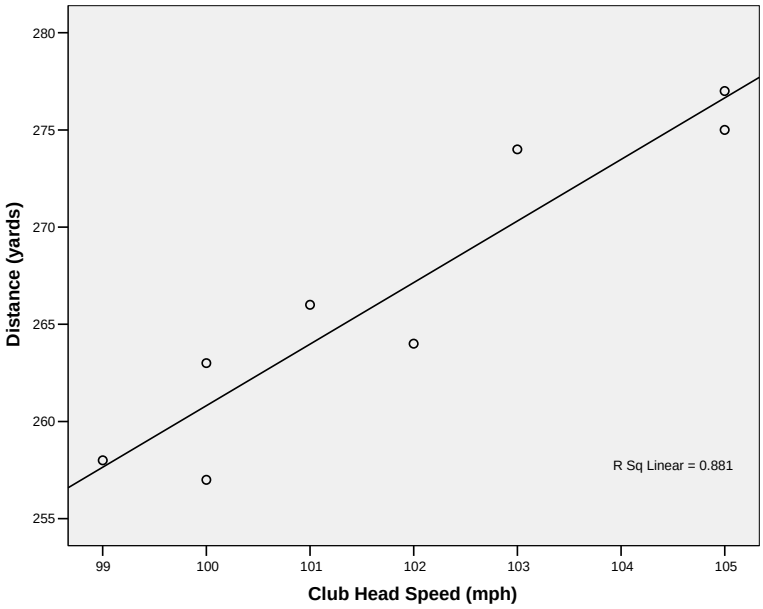
After creating the scatter plot (see Creating a Scatter diagram), go to the chart editor and click on any point in the graph, so that the points are highlighted.



Once the points are highlighted, select Elements → Fit Line at Total



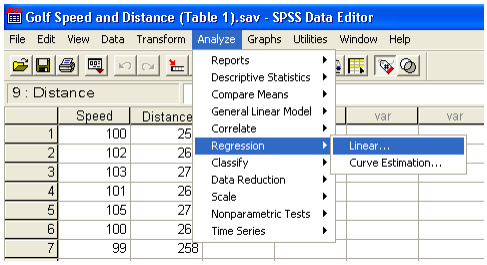
Select Linear (which is the default). This will automatically add in the Least Squares Regression line to the plot. Mean of Y will create a horizontal line at the mean of the response variable. Loess will create a line that will draw a fit line using iterative weighted least squares. At least 13 data points are needed. This method fits a specified percentage of the data points, with the default being 50%. In addition to changing the percentage, you can select a specific kernel function. The default kernel (probability function) works well for most data. Quadratic will fit a quadratic curve to the data, and Cubic will fit a cube function to the data. These are beyond what you should need for this class. The Linear option is all you really need at this point. When you have selected the option you want, push Close.



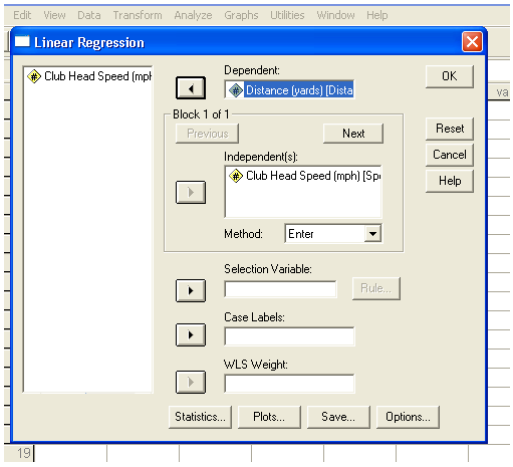
►Finding the Least-Squares Regression Equation

Example 2, Page 216

Once the data have been entered into SPSS, go to **Analyze → Regression → Linear…**



This will open the linear regression window. Select the response variable, and click the ► button by the Dependent box. Select the predictor variable and click the ► button by the Independent(s) box.



Although there are many other things we can, and will, do while in this window, for now we will just proceed by clicking OK. This will provide the following output.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Club Head Speed (mph)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Distance (yards)

This first box is only for other types of regression, and we will ignore it.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.939[a]	.881	.861	2.883

a. Predictors: (Constant), Club Head Speed (mph)

This table provides the correlation coefficient, similar to what was provided by the **Analyze → Correlate → Bivariate...** command used previously. If you plan on obtaining both the correlation and the regression equation, only the **Analyze → Regression → Linear…** command needs to be used. The other values will be discussed later.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	369.642	1	369.642	44.484	.001[a]
	Residual	49.858	6	8.310		
	Total	419.500	7			

a. Predictors: (Constant), Club Head Speed (mph)

b. Dependent Variable: Distance (yards)

The third table provides information that will not be discussed until chapter 12, but provides some important information for later.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-55.797	48.371		-1.154	.293
	Club Head Speed (mph)	3.166	.475	.939	6.670	.001

a. Dependent Variable: Distance (yards)

The last table is the one of most interest for this section. The first column, under the Unstandardized Coefficients, is the B values. The first of these is b_0, or the intercept, in this example -55.797. The second,

next to the predictor variable, is the slope, in this example 3.166. So our regression equation would be Distance (yards) = -55.797 + 3.166 * Club Head Speed (mph). The other values in the table will be used in later chapters.

Section 4.3 Diagnostics on the Least-Squares Regression Line

►Coefficient of Determination

Example 1, Page 228

Using the information in the regression, the R^2 value is given in the second table, right next to the correlation coefficient.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.939[a]	.881	.861	2.883

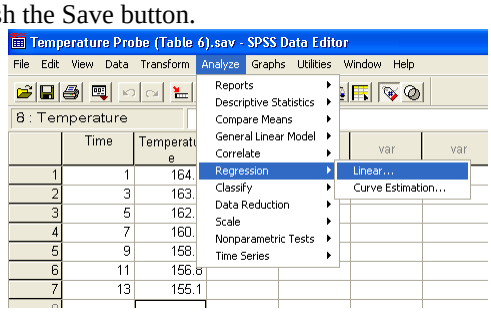
a. Predictors: (Constant), Club Head Speed (mph)

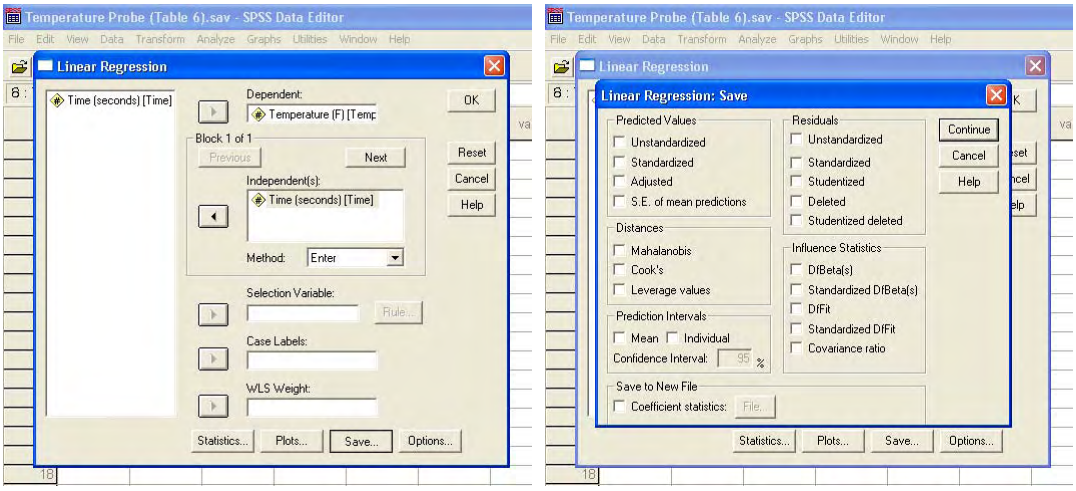
So, in this example, R^2 is 0.881, or 88.1%. The adjusted R^2 value is used in more complicated models.

►Producing a Residual Plot

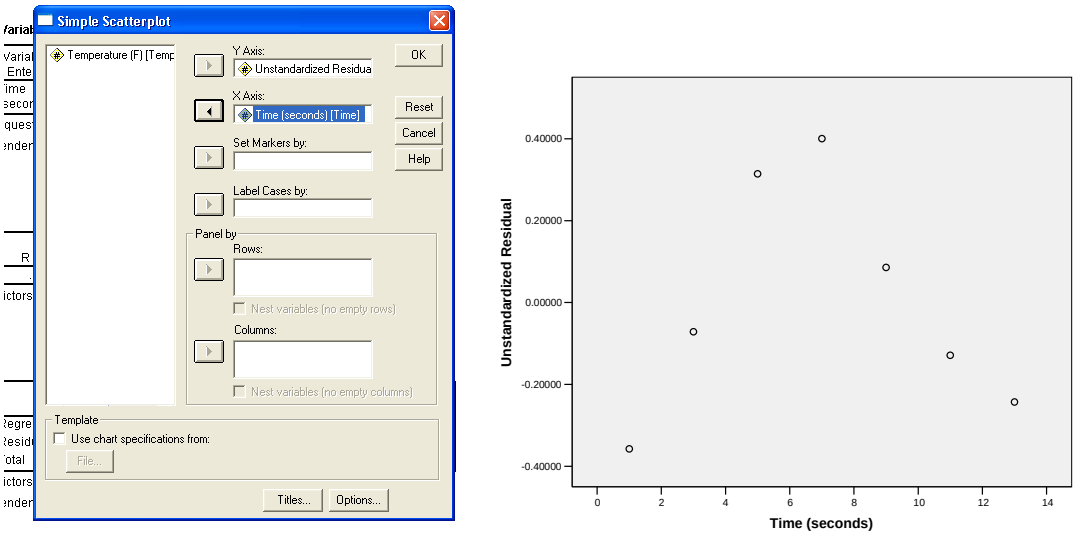
Example 2, Page 229

After entering the data into SPSS, go to **Analyze → Regression → Linear…**, select the dependent and independent variables, and push the Save button.

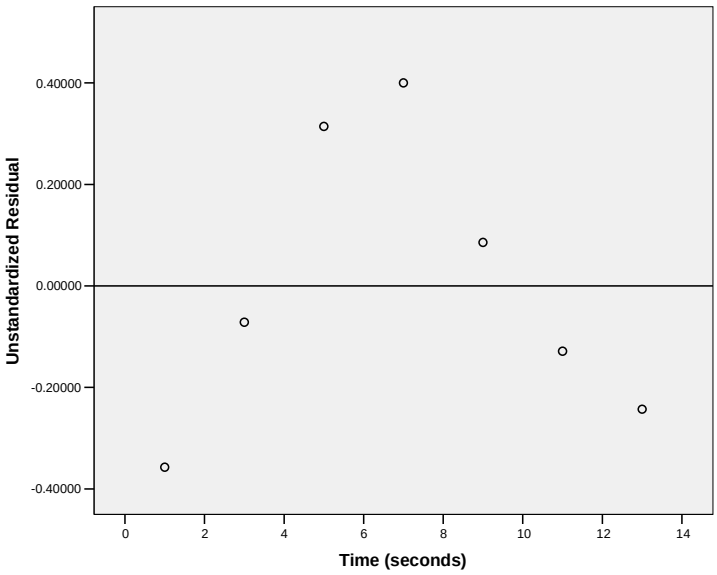




In the Linear Regression: Save menu, you will see Residuals in the top Right section of the window. Select Unstandardized. This will not put the residuals into the output window, but will create a new variable, RES_1, which are the residuals at each point. After selecting Unstandardized, press continue, and OK. You will then go to Graph → Scatter/Dot, (you do not have to switch back to the data window to do this, all menus are available in the output window as well) and select simple scatter. Select the new variable, RES_1 as the Y-Axis, and the predictor variable as the X-Axis. This will create the Residual plot.



You can add the horizontal line at 0 by going to the chart editor, clicking on the points, as if you were going to add the least squares line, Go to **Elements → Fit Line at Total** and select Mean of Y. The mean of the residuals is 0, so it will create a horizontal line at 0.



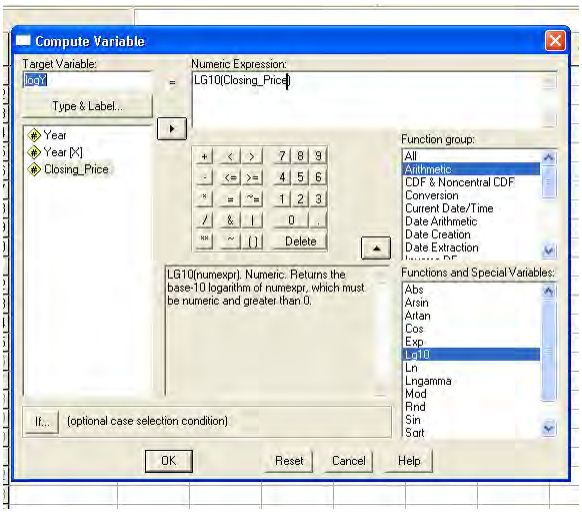
SPSS does have its own plots that it produces, under the plots button in the regression window, which are similar in function, but are different than the ones provided in the book.

Section 4.4 Nonlinear Regression: Transformations (on CD)

►Non-linear transformations

Example 4 (4-4)

There are two ways to find regression equations for non-linear transformations in SPSS. The first is to create a new variable and using the simple linear regression methods already produced. To do this, enter the data into SPSS. Then go to **Transform → Compute…** to create the new variable. Type a new variable name in the Target Variable box, and the transformation in the Numeric Expression box.



In this case, the transformation is the LG10, or $\log_{10}$ transformation. Other transformations are possible, such as using the natural log, ln. These are listed under Arithmetic functions. Once the variable has been

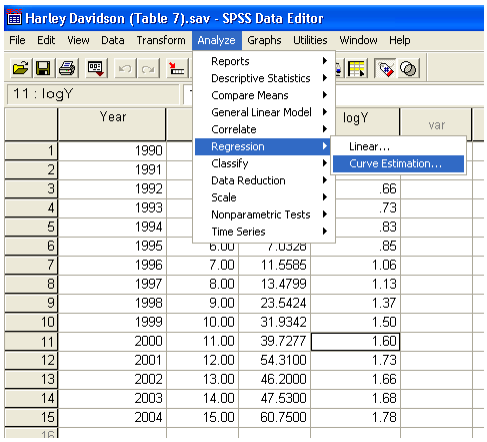
created, a regular simple linear regression can be used, using the new variable as the response, and the regular x as the predictor.

Coefficients[a]

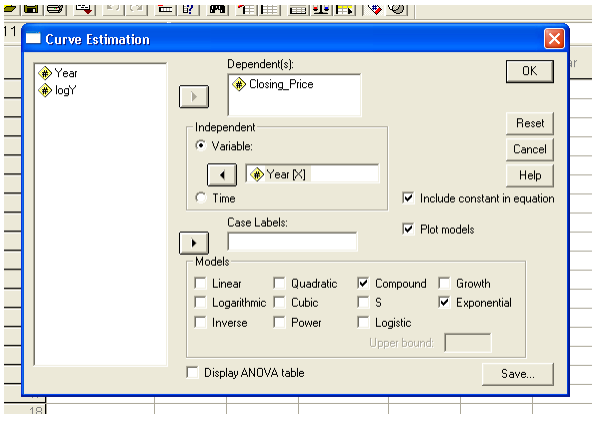
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.211	.068		3.084	.009
	Year	.116	.008	.974	15.445	.000

a. Dependent Variable: logY

SPSS knows that sometimes transformations must be used. They have some transformations built in. Go to **Analyze → Regression → Curve Estimation**.



This will open the Curve Estimation window. Here you can select from a variety of possible transformations. The advantage of this is, when you are unsure of which transformation to use, you can select more than one at a time, and compare the results to see which is best.



Here, the method that was used was the Compound, which uses $\log_{10}$. Also possible is the exponential, which uses the natural log. Here we can compare the two directly to each other. Using the right mouse button on any of the models will provide you with information on what the model looks like. For example, using the right mouse button on the Compound will provide the following information:

Model whose equation is Y=b0*(b1**t) or ln(Y)=ln(b0)+(ln(b1)*t),

which tells us that the equation will be $Y = b_0 b_1^x$ which is the equation we used in the example. The output provided is

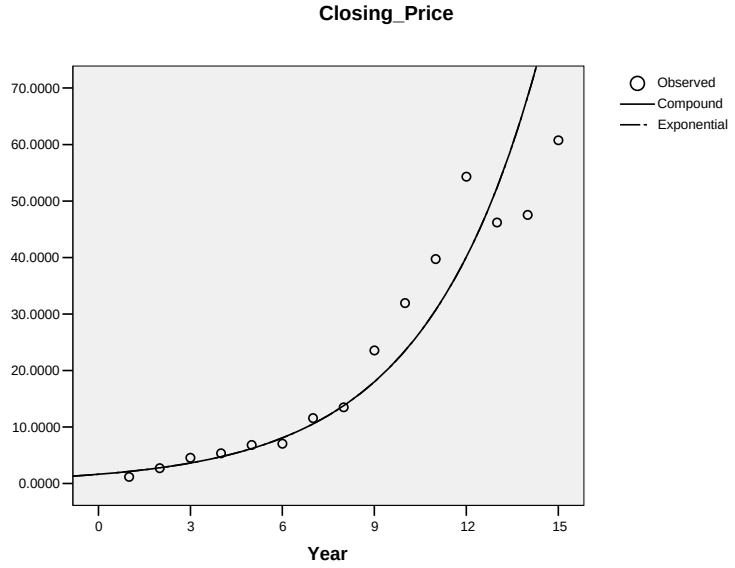
Model Summary and Parameter Estimates

Dependent Variable: Closing_Price

	Model Summary					Parameter Estimates	
Equation	R Square	F	df1	df2	Sig.	Constant	b1
Compound	.948	238.536	1	13	.000	1.624	1.306
Exponential	.948	238.536	1	13	.000	1.624	.267

The independent variable is Year.

Here we can compare, using R^2 values, the different equations. In this case, both models are the same, although the models look slightly different. (In the Compound model, we are using $1.624 * (1.306^x)$ and in the Exponential model we are using $1.624 * e^{.0.267*x}$) Using the curve estimation also provides a graph, in which we can see that both lines are the same.



Similarly, we could fit power models, or other models using this same method, without having to change the original data.

Chapter 5. Probability

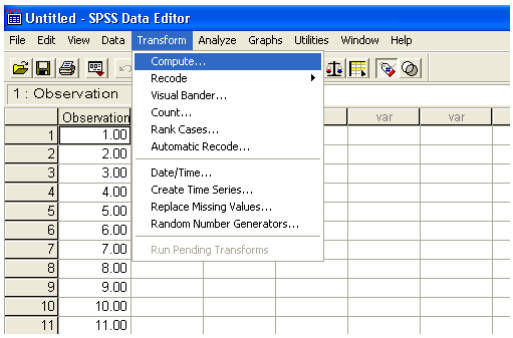
SPSS is not a data creation tool, and there are some difficulties in creating data sets in SPSS. SPSS is a data analysis tool, and should be used as such. Despite this, once you have a data set, you can create new variables that can be used with this section.

Section 5.1 Probability Rules

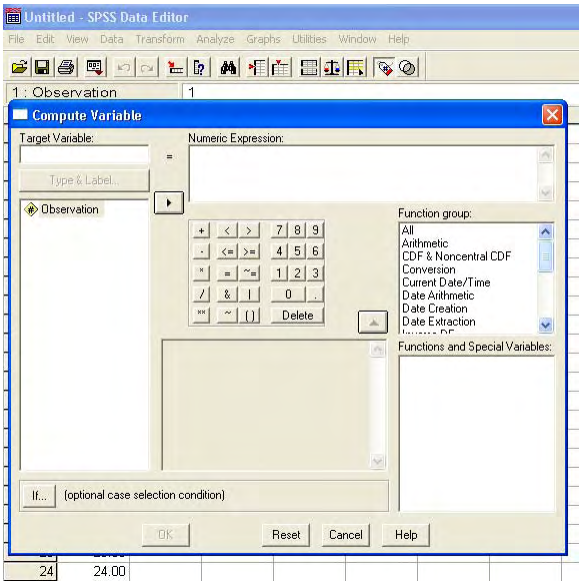
►Simulating Probabilities

Example 8, Page 259

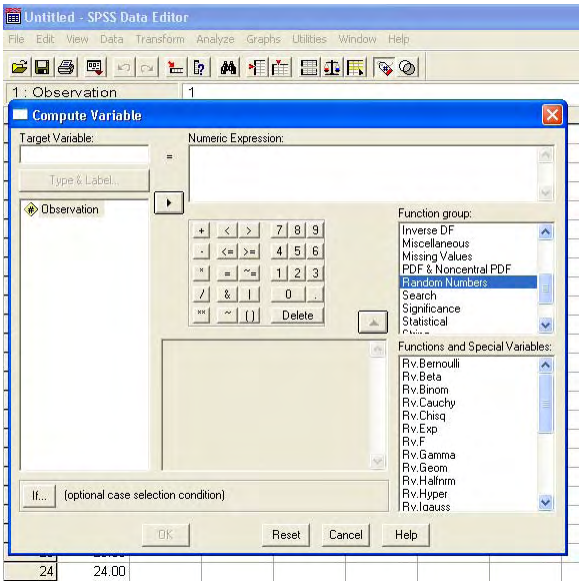
To create 100 random numbers, you must first create a data set with 100 observations. SPSS will create new variables, with the same sample size as already exists, but will not create more observations than are already in the data set. To create the original sample of 100, you can type in the values 1 to 100 by hand, or use another program, such as Excel, where data generation is easier. After creating the data set with 100 observations, go to **Transform → Compute…**



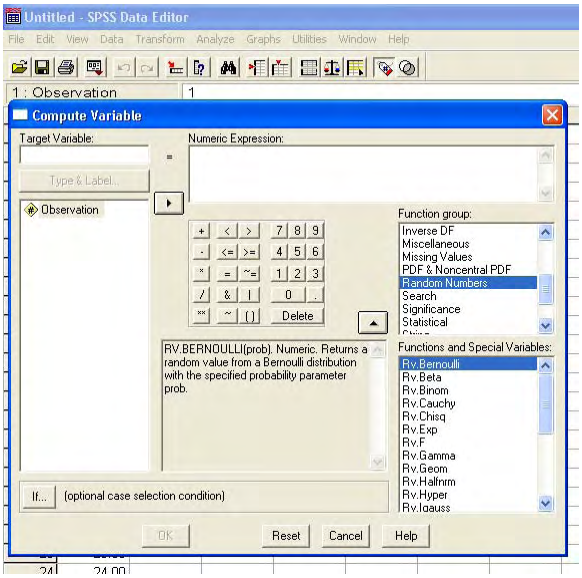
Once in the Compute Variable window, you can choose a name for the new variable you wish to create. This name must follow SPSS naming conventions (no strange characters or spaces).



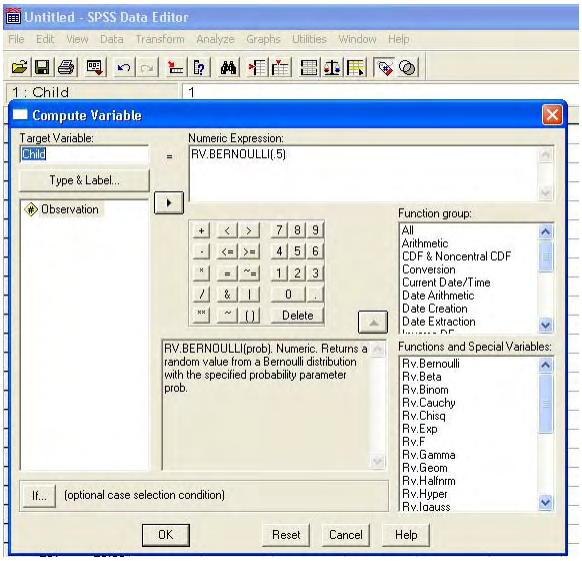
The name of the variable goes in the Target Variable box. In the Function group box, find and select Random Numbers.



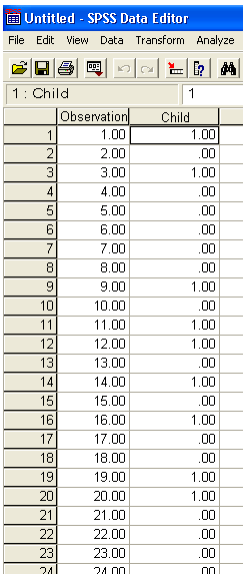
This provides a list of all of the random number distributions possible. We will discuss more about what distributions are in the next chapter. To choose a value of 0 or 1, select Rv.Bernouli. In the box to the left of the Functions and Special Variables box will be a description of what the function does.



In this case, Bernoulli will provide a value of either 0 or 1 (a Bernoulli trial) randomly.



The random variable requires an expected probability (prob), which for this example we will select to be .5 (half girls and half boys). Any value between 0 and 1 can be selected as the prob value. Once you have selected a name for the variable, and the function you wish to use, push OK. This will add a new column to the data window.



	Observation	Child
1	1.00	1.00
2	2.00	.00
3	3.00	1.00
4	4.00	.00
5	5.00	.00
6	6.00	.00
7	7.00	.00
8	8.00	.00
9	9.00	1.00
10	10.00	.00
11	11.00	1.00
12	12.00	1.00
13	13.00	.00
14	14.00	1.00
15	15.00	.00
16	16.00	1.00
17	17.00	.00
18	18.00	.00
19	19.00	1.00
20	20.00	1.00
21	21.00	.00
22	22.00	.00
23	23.00	.00
24	24.00	.00

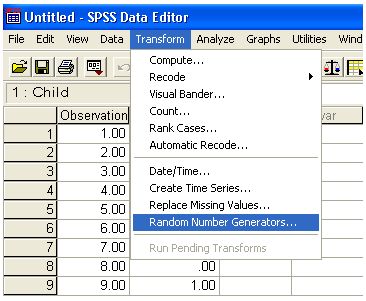
You may go to the variable view to define 0 to be boys and 1 to be girls, as in the example. To get the count, we use the frequency methods discussed in chapter 2.

Child

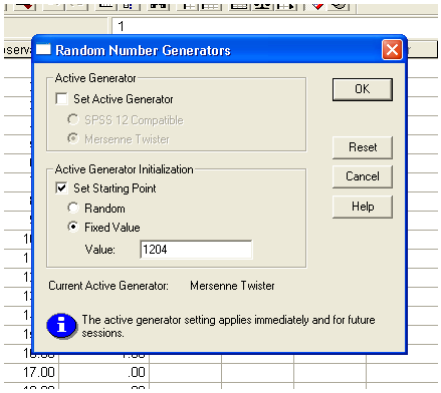
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	54	54.0	54.0	54.0
	1.00	46	46.0	46.0	100.0
	Total	100	100.0	100.0	

So, in this example, we had 54 boys and 46 girls out of the 100 children.

To set a seed value, or starting place in the random number table, go to **Transform → Random Number Generators…**

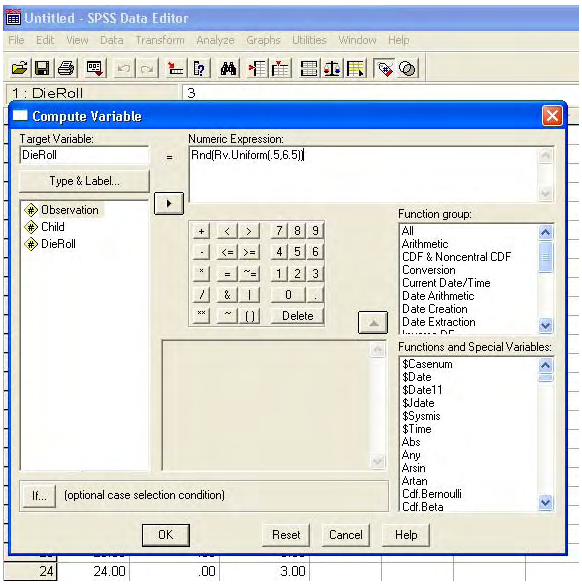


In the Random Number Generators window, you can select Set Starting Point, and Fixed Value, and type in the seed value. This will make it so every time you open SPSS, the starting point in the random number table will be the same. Using Random will make it so the starting point changes every time you open SPSS. Using a set value is useful if you want everybody to get the exact same results, while using the Random option is better for actual simulation.



Technology Step-by-step Page 265

To create a simulated die roll, follow the steps above, except instead of choosing Bernoulli, we will use a uniform distribution, which means all values are equal. Again, the sample size will be the same as how many samples are currently in the data set, so we will create 100 tosses, as we already have 100 observations.



One of the problems with using the uniform distribution is that it is a continuous instead of discrete uniform distribution. Because of this, instead of selecting numbers between 1 and 6, as we would think, we will select between .5 and 6.5. This gives 1 and 6 equal chances instead of giving them only ½ of a chance each. We then round the random numbers to the nearest whole number. (This is the Rnd function).

This provides us with a new column of values between 1 and 6. Again using the frequency procedures from chapter 2, we can count the number of times we roll each number.

DieRoll

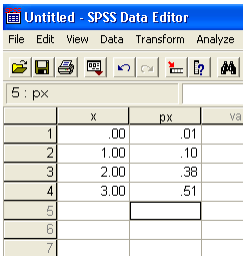
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	14	14.0	14.0	14.0
	2.00	18	18.0	18.0	32.0
	3.00	17	17.0	17.0	49.0
	4.00	18	18.0	18.0	67.0
	5.00	14	14.0	14.0	81.0
	6.00	19	19.0	19.0	100.0
	Total	100	100.0	100.0	

Chapter 6. Discrete Probability Distributions

Many of the methods required for this chapter are found in previous chapters, including frequency distributions, and histograms.

Section 6.1 Expected Values

You can find the expected value of a distribution in SPSS, but you cannot find the variance. This is because SPSS is still assuming that the data comes from a sample, not a population. To find the expected value, enter the values of x and the probabilities. We then follow the methods of finding a weighted mean (from chapter 3, section 4). We weight the data by the probability, and then find the descriptive statistics on x.



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
x	1	.00	3.00	2.3900	.
Valid N (listwise)	1				

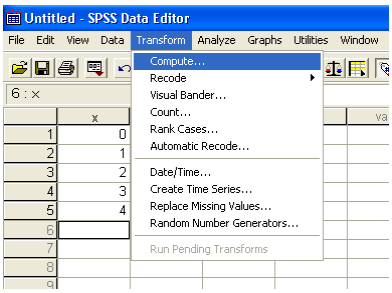
The standard deviation is missing, because n-1 is 0, so cannot be calculated. Even if you changed the weights to be whole numbers (say by multiplying by 100), the standard deviation would be calculated using n-1 instead of n, and would provide the wrong value. This value may be adjusted for, however. To get the population variance from the sample variance, use $\frac{N-1}{N}s^2$.

Section 6.2 The Binomial Probability Distribution

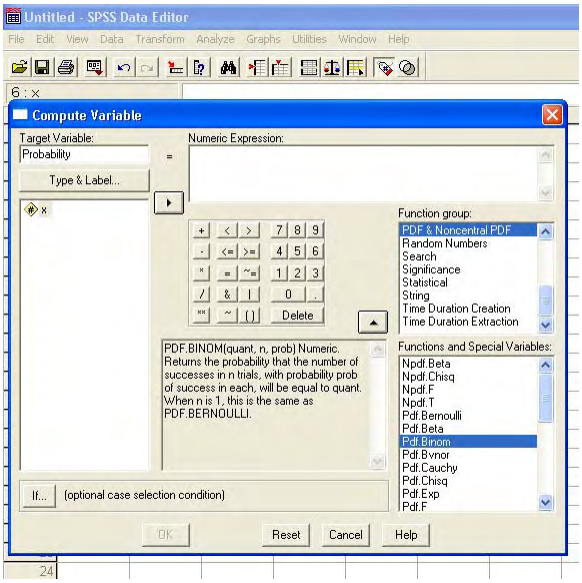
►Binomial Distribution

Example 2, Page 330

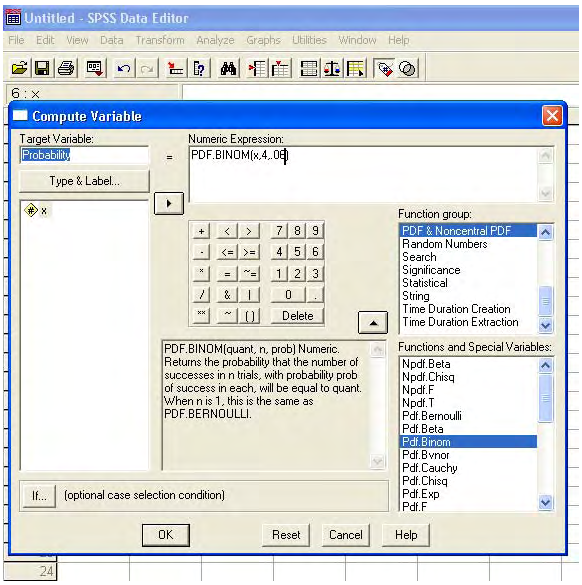
To create a Binomial Probability Distribution, create a variable with the possible values of x, in this example, 0, 1, 2, 3, and 4. Once the variable has been defined, go to **Transform → Compute…**



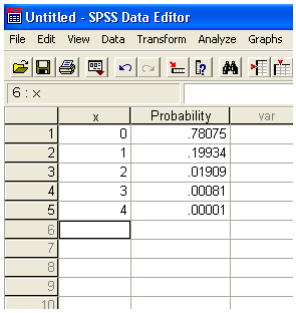
While in the compute variable window, create a name for the probabilities in the Target Variable box. In the Function Group box, select PDF and Noncentral PDF. PDF is the Probability Distribution Function. In the Functions and Special Variables, select Pdf.Binom.



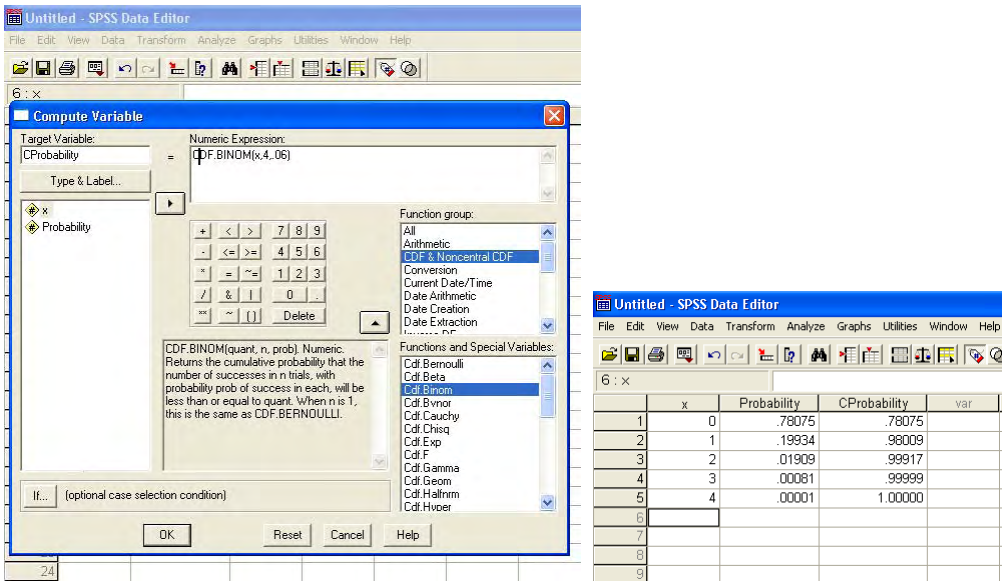
For this function to work, you need three values. The first is x, which is the variable we created earlier. The second is the number of trials, n, and the third is the probability of a success, or p. So, we will select Pdf.Binom, and replace the question marks that appear with x, 4, and .06. When selecting the name of the variable we wish to use, we can either type the name (if it is short), or highlight the name, and push the ► button next to the Numeric Expression box.



Once the name and expression have been completed, push the OK button. This will create a new column in the data view window with the probabilities of observing each x value. (You may need to change the number of decimal places shown in the variable view window to see some of the values).



The probabilities provided are the probabilities of being equal to x for the pdf. We can also find the Cumulative Distribution Function by changing the P in PDF to C in the Numeric Expression box. This is the probability of getting a value less than or equal to x.



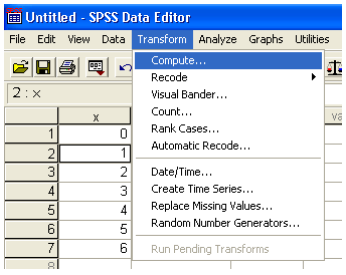
To create a Binomial Probability Histogram, follow the methods used in chapter 2.

Section 6.3 The Poisson Probability Distribution

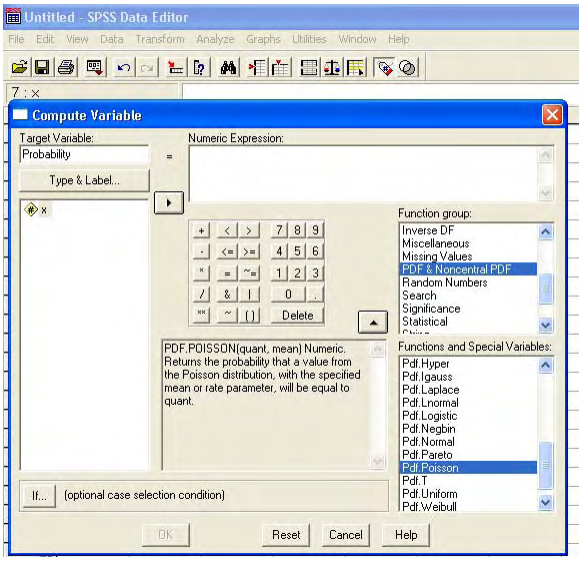
►Poisson Distribution

Example 2, Page 346

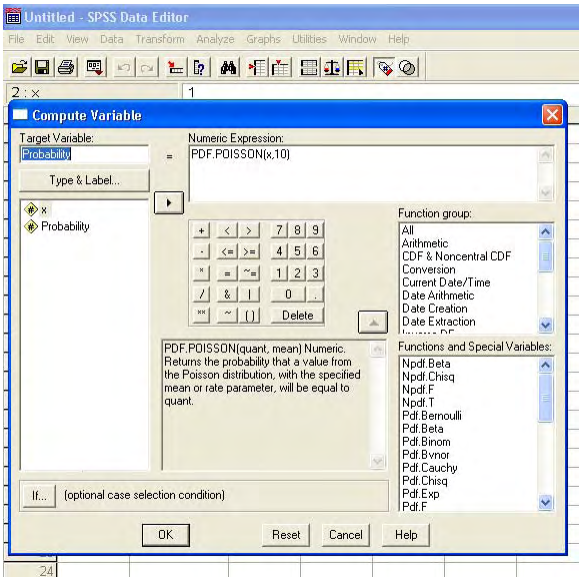
To create a Poisson Probability Distribution, create a variable with the possible values of x. This is a little more difficult than for the Binomial distribution, as the values of x can go on to infinity. Choose appropriate x values for the problem. For example, to find the probability that exactly 6 cars arrive between 12 noon and 12:05 P.M., one of the x values needs to be 6. To get the probability of fewer than 6, we need an x value of 5 (since the cumulative probability up to 5 is the probability of less than 6). To get the probability of at least 6 cars arriving, you need to have a value of 5, and use the complement rule. We will use the values 0 through 6 for this example, so that the values can be related to the ones in the book. Once the variable has been defined, go to **Transform → Compute…**



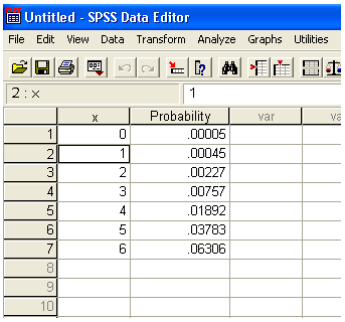
While in the compute variable window, create a name for the probabilities in the Target Variable box. In the Function Group box, select PDF and Noncentral PDF. PDF is the Probability Distribution Function. In the Functions and Special Variables, select Pdf.Poisson.



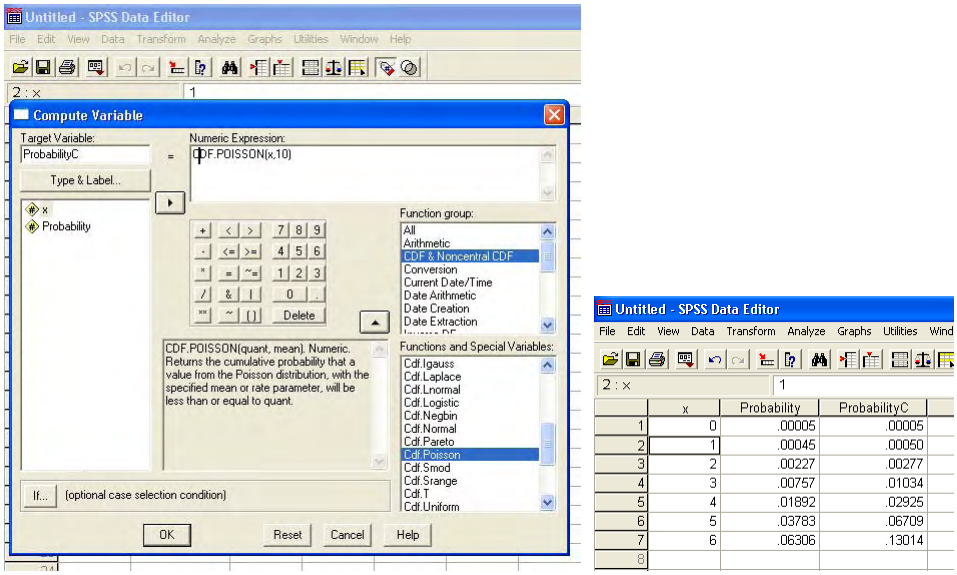
For this function to work, you need two values. The first is x, which is the variable we created earlier. The second is the mean, or the average rate, which is λt, or 10 for this example. So, we will select Pdf.Poisson, and replace the question marks that appear with x, and 2. When selecting the name of the variable we wish to use, we can either type the name (if it is short), or highlight the name, and push the ► button next to the Numeric Expression box.



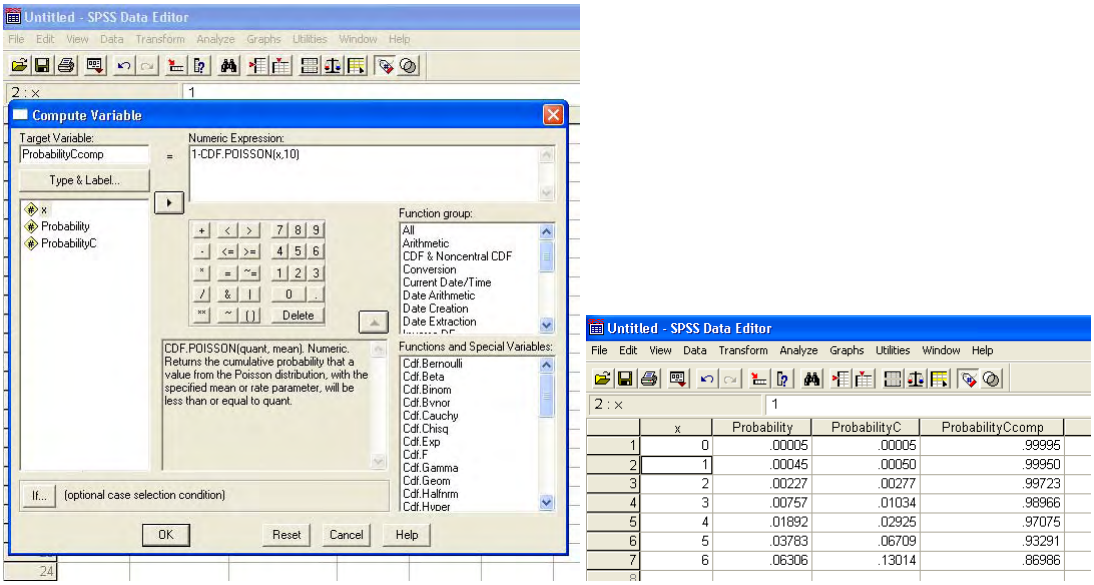
Once the name and expression have been completed, push the OK button. This will create a new column in the data view window with the probabilities of observing each x value. (You may need to change the number of decimal places shown in the variable view window to see some of the values).



The probabilities provided are the probabilities of being equal to x for the pdf. We can also find the Cumulative Distribution Function by changing the P in PDF to C in the Numeric Expression box. This is the probability of getting a value less than or equal to x.



To find the complement, we take the same function, and subtract it from 1.



So, using the tables we have just created, the probability of having exactly 6 cars arrive between 12 noon and 12:05 P.M. is .06306. The probability of less than 6 cars arriving between 12 noon and 12:05 P.M. is .06709, which is the cumulative probability where x=5. The probability that at least 6 cars arrive between 12:00 noon and 12:05 P.M. is .93291, which is the complement of the cumulative probability when x=5.

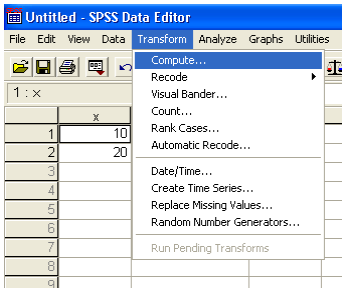
Chapter 7. The Normal Probability Distribution

Section 7.1 Properties of the Normal Distribution

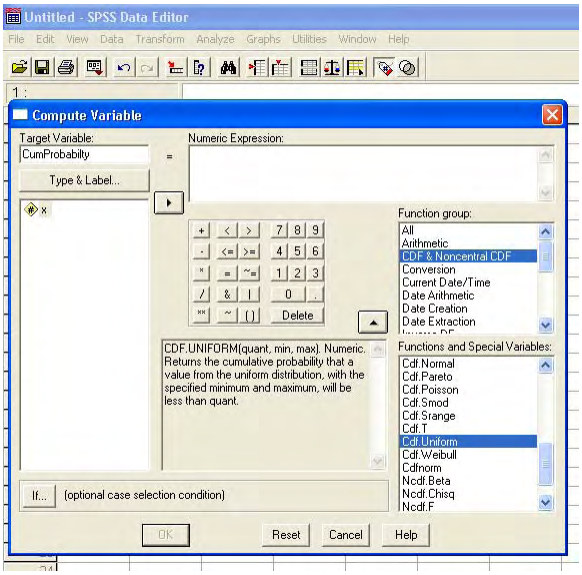
►Uniform Distribution

Example 2, Page 362

To find the probability between two values in a continuous distribution, we must use the cumulative distribution functions. In SPSS, type the two values that you want to find the area between as a new variable. Once the variable has been defined, go to **Transform → Compute…**

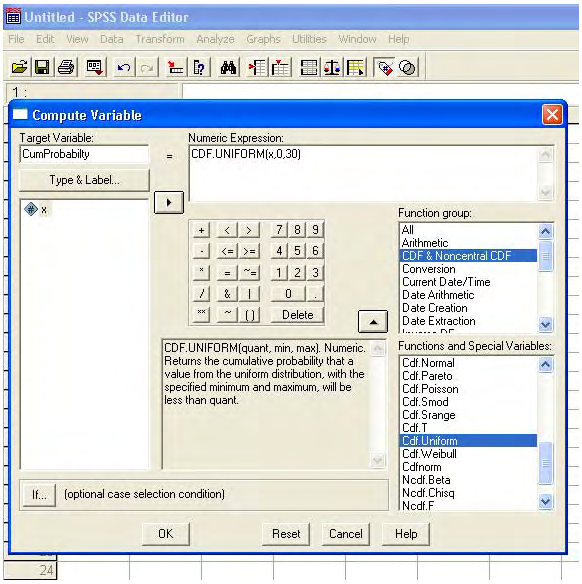


While in the compute variable window, create a name for the probabilities in the Target Variable box. In the Function Group box, select CDF and Noncentral CDF. CDF is the Cumulative Distribution Function. In the Functions and Special Variables, select Cdf.Uniform.

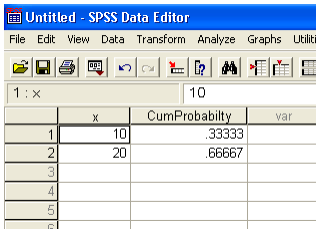


For this function to work, you need three values. The first is x, which is the variable we created earlier. The second is the minimum value for the distribution, and the third is the maximum value for the distribution. These are the maximum and minimum values over the whole distribution, not the values of x. So, we will select Cdf.Uniform, and replace the question marks that appear with x, 0, and 30. When

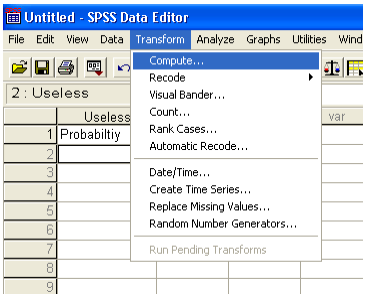
selecting the name of the variable we wish to use, we can either type the name (if it is short), or highlight the name, and push the ► button next to the Numeric Expression box.



Once the name and expression have been completed, push the OK button. This will create a new column in the data view window with the probabilities of observing less than or equal to each x value. (You may need to change the number of decimal places shown in the variable view window to see some of the values).

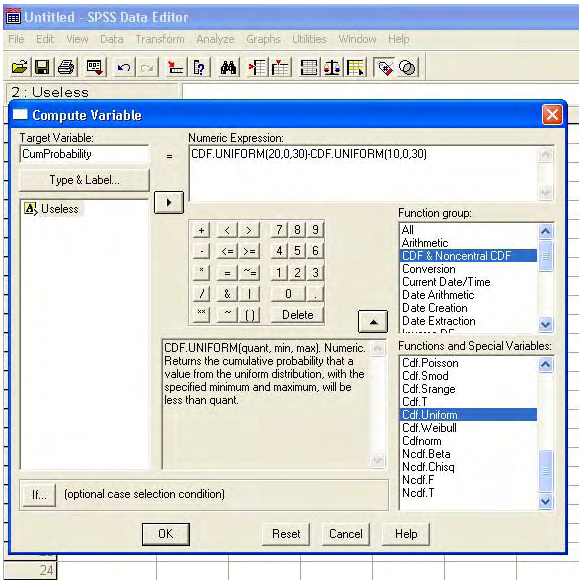


To get the area between the two, you will need to find the difference. .666667-.33333=.33333 or 1/3. Using SPSS is only an advantage in this case when there are many x values that you will use.
To find the probability between two points, you can start with typing in one value in the first column. This value is irrelevant, as it serves only to define the sample size. . Once the variable has been defined, go to **Transform → Compute…**

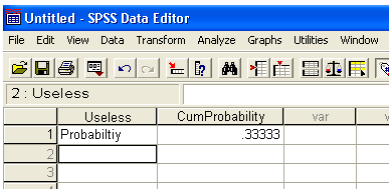


While in the compute variable window, create a name for the probabilities in the Target Variable box. In the Function Group box, select CDF and Noncentral CDF. CDF is the Cumulative Distribution Function. In the Functions and Special Variables, select Cdf.Uniform. Instead of using a variable, x, for the first

value, we type in the larger of the two x values. We then subtract the Cdf.Uniform of the smaller of the two x values from the first.



Once the name and expression have been completed, push the OK button. This will create a new column in the data view window with the probabilities of observing an x value between 10 and 20. (You may need to change the number of decimal places shown in the variable view window to see some of the values).

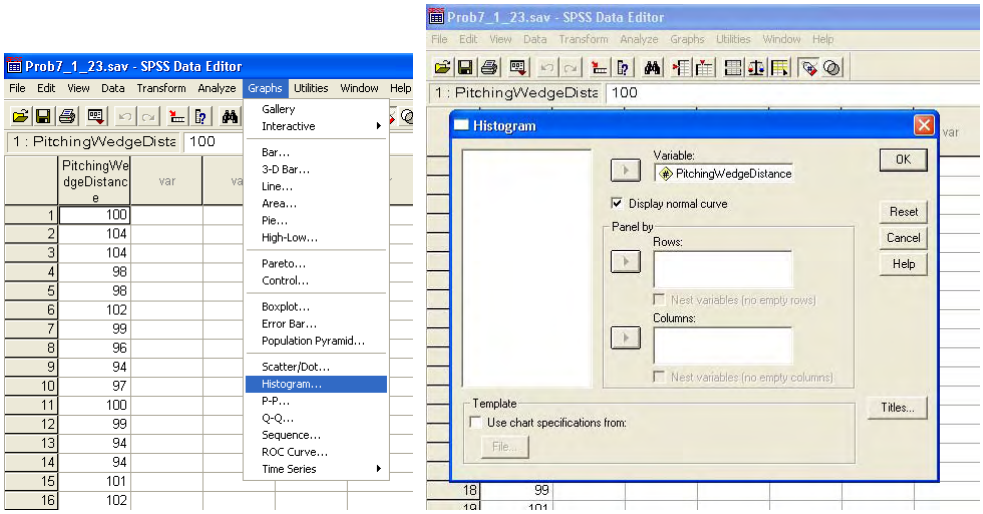


If you don't create the dummy variable first, you will get an error, as SPSS will not know how many answers to create. If you have more than one value, the answer will be repeated as many times as you have values in the data set.

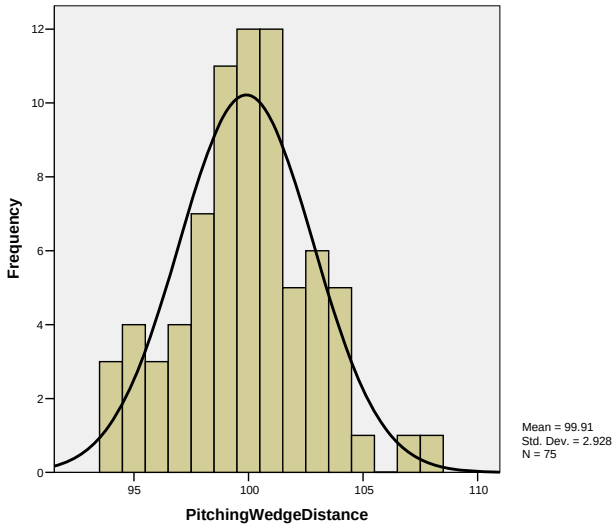
►Adding a Normal Curve to a Histogram

Problem 37, Page 371

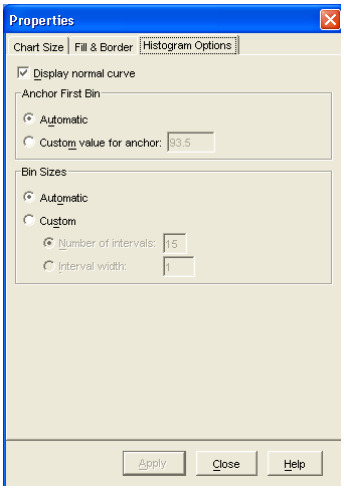
To create a histogram, see the information presented in chapter 2. Once all of the values have been entered, we are ready to create the histogram. Go to **Graphs → Histogram…**



To add the Normal Curve, check the Display normal curve box underneath the Variable box. Then push OK. The histogram will be provided in the Output window, with the normal curve added.



If you forget to click the Display normal curve when creating the histogram, the curve may be added to the histogram through the SPSS Chart Editor. While in the editor, click on the histogram and go to **Edit → Properties**. Here you will be given the option of adding the normal curve. Again, all you need to do is check the Display normal curve box.



Section 7.2 The Standard Normal Distribution

►Finding the Area under the Standard Normal Curve

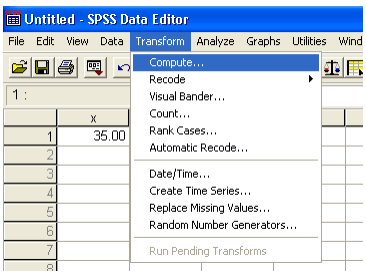
Rather than going into detail of how to find the area under the Standard Normal Curve, we will show how to find the area under any Normal curve. To find areas under the Standard Normal, use the same methods provided below with a mean of 0 and a standard deviation of 1.

Section 7.3 Applications of the Normal Distribution

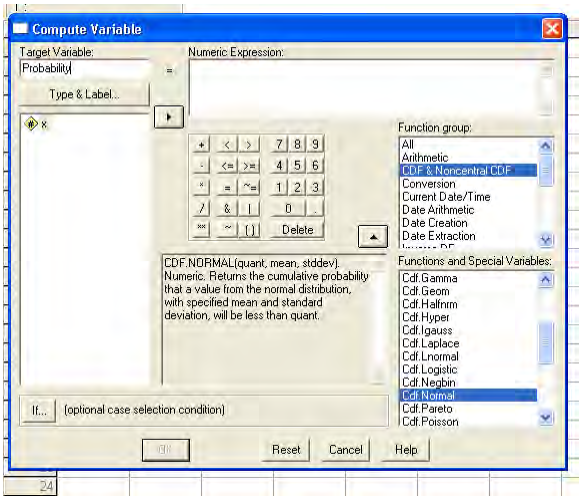
►Finding the Area under a Normal Curve

► *Area less than or equal to a value*:
Example 1, Page 385

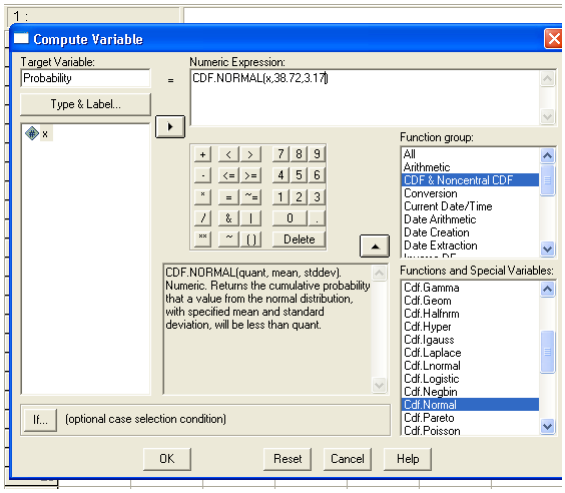
To find the proportion of three-year-old females who have a height less than 35 inches, we create a new variable with the value 35. Once the variable has been defined, go to **Transform → Compute…**



While in the compute variable window, create a name for the probabilities in the Target Variable box. In the Function Group box, select CDF and Noncentral CDF. CDF is the Cumulative Distribution Function. In the Functions and Special Variables, select CDF.NORMAL.

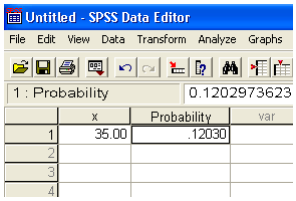


For this function to work, you need three values. The first is x, which is the variable we created earlier. The second is the mean of the Normal distribution, and the third is the standard deviation of the Normal distribution. So, we will select CDF.NORMAL, and replace the question marks that appear with x, 38.72, and 3.17. When selecting the name of the variable we wish to use, we can either type the name (if it is short), or highlight the name, and push the ► button next to the Numeric Expression box.



Once the name and expression have been completed, push the OK button. This will create a new column in the data view window with the probabilities of observing an x value less than or equal to 35 inches. (You may need to change the number of decimal places shown in the variable view window to see some of the values).

Note: SPSS answers may vary slightly from book answers due to the book rounding to 2 decimal places for the z-score. SPSS does not round in its calculations.



So, 12.03% of three-year-old females have a height less than 35 inches.

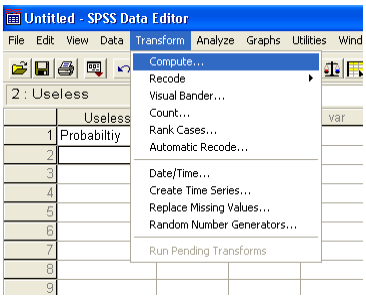
► *Area greater than or equal to a value*:

To find the area greater than or equal to a value, use 1-CDF.NORMAL as the equation. The CDF is always the area to the left, so we can use the complement rule to find the area to the right.

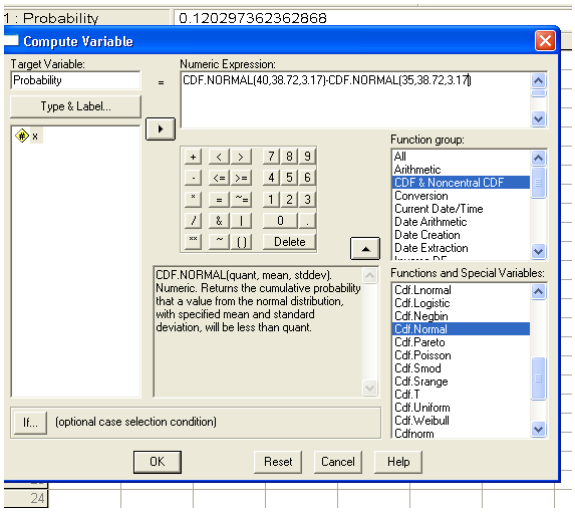
► *Area between two values*:

Example 3, Page 387

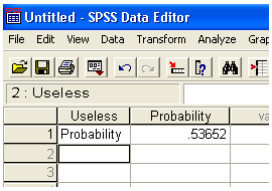
To find the area between two values, you can start with typing in one value in the first column. This value is irrelevant, as it serves only to define the sample size. . Once the variable has been defined, go to **Transform → Compute…**



While in the compute variable window, create a name for the probabilities in the Target Variable box. In the Function Group box, select CDF and Noncentral CDF. CDF is the Cumulative Distribution Function. In the Functions and Special Variables, select CDF.NORMAL. Instead of using a variable, x, for the first value, we type in the larger of the two x values. We then subtract the CDF.NORMAL of the smaller of the two x values from the first.



Once the name and expression have been completed, push the OK button. This will create a new column in the data view window with the probabilities of observing an x value between 35 and 40. (You may need to change the number of decimal places shown in the variable view window to see some of the values).



If you don't create the dummy variable first, you will get an error, as SPSS will not know how many answers to create. If you have more than one value, the answer will be repeated as many times as you have values in the data set.

The probability of finding a three-year-old female between 35 inches and 40 inches is .53652.

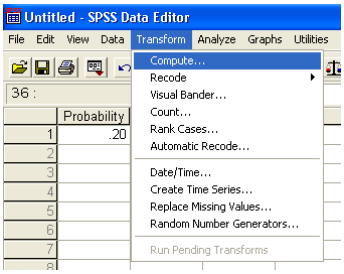
►Finding Values of Normal Random Variables

► *Area less than or equal to a value*:

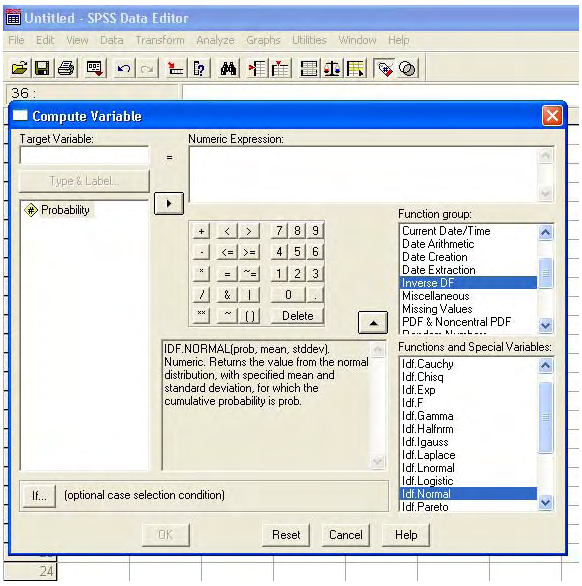
Example 4, Page 388

When looking for a specific value of x under the Normal Distribution, SPSS must look for values using a cumulative distribution. So, probabilities must be areas to the left of x. If the information does not come in this way, it must be changed to fit how the computer can use it.

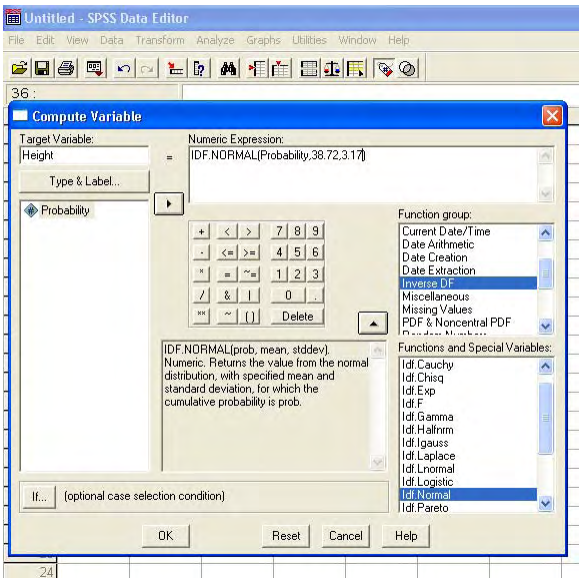
In this example, we are looking for the 20^{th} percentile, which is the point at which 20% of the information is less than or equal to x, or the cumulative probability is .20. To find the height of three-year-old females which marks the 20^{th} percentile, we create a new variable with the value .20. Once the variable has been defined, go to **Transform → Compute…**



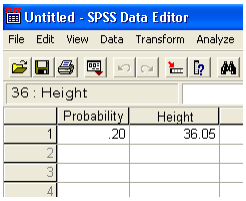
While in the compute variable window, create a name for the probabilities in the Target Variable box. In the Function Group box, select Inverse DF. Inverse DF is the Inverse Distribution Function. In the Functions and Special Variables, select IDF.NORMAL.



For this function to work, you need three values. The first is the probability to the left, which is the variable we created earlier. This only works with the cumulative probability. The second is the mean of the Normal distribution, and the third is the standard deviation of the Normal distribution. So, we will select IDF.NORMAL, and replace the question marks that appear with probability, 38.72, and 3.17. When selecting the name of the variable we wish to use, we can either type the name (if it is short), or highlight the name, and push the ► button next to the Numeric Expression box.



Once the name and expression have been completed, push the OK button. This will create a new column in the data view window with the x value. (You may need to change the number of decimal places shown in the variable view window to see some of the values).

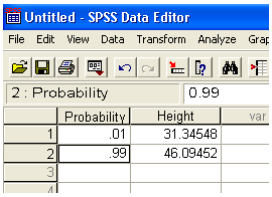


The height of a three-year-old female that separates the top 80% from the bottom 20% is 36.05.

► *Area more than or equal to a value*:

Example 6, Page 389

To get the x value which marks where 1% of the information is to the right of that value, we have to use the complement rule. We know that if 1% is to the right, then 99% is to the left, and that is the information we give to SPSS. We can then follow the steps for area to the left.



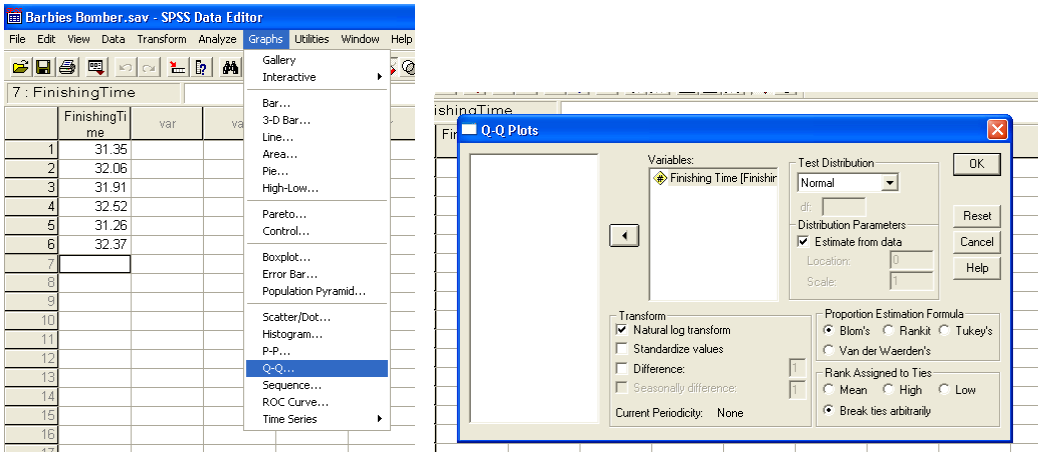
There are two advantages of using SPSS in these situations. The first is that we don't need to find the closest value in the table, as SPSS has the complete table to work from. In other words, we don't have to worry about rounding for the z-scores. The second is we can look up the values of more than one probability at a time, without adding work.

Section 7.4 Assessing Normality

►Creating a Normal Probability or Q-Q Plot

Example 1, Page 395

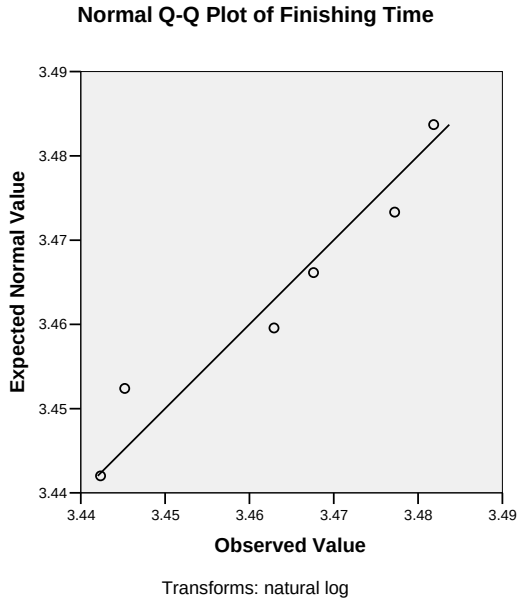
Once we have entered the data, we go to **Graphs → Q-Q…** (Q-Q stands for Quantile - Quantile).



Highlight the variable of interest, and push the ► button. Make sure that the Test Distribution is Normal (if you want to test if the distribution is other than normal, there are other choices possible).

The Blom's Proportion Estimation Formula is the one used in the book, $f_i = \frac{i - 0.375}{n + .25}$.

Rankit uses $f_i = \frac{i - 0.5}{n}$, Tukey's uses $f_i = \frac{i - 0.333}{n + .333}$, and Van der Waerden's uses $f_i = \frac{i}{n + 1}$. We will just use Blom's Proportion Estimation Formula, as on page 397. Also, to make the Q-Q plot look like the ones in the book, select the Break ties arbitrarily under the Rank Assigned to Ties options. The default is to give values that have the same rank the mean of the values, where breaking ties arbitrarily will provide two different values for the same observed value, based on the different indexes.
Push OK

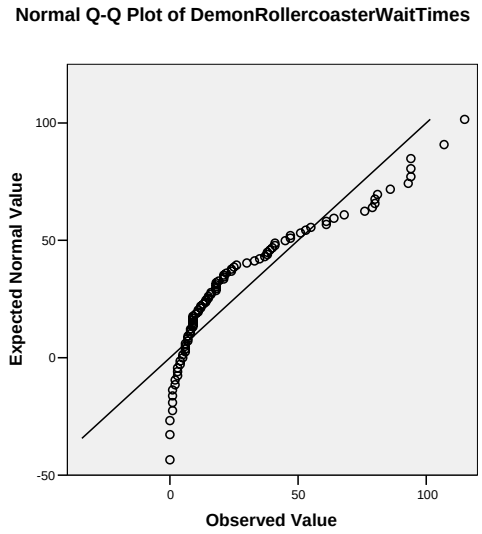


The plot will appear in the Output window, with a de-trended plot. We will work with the Q-Q plot, as in the book. SPSS does not put the interval curves, as Minitab does. Instead, we must use our best judgment as to how close the points are to the line. Some general guidelines are as follows:

Quantile - Quantile Plot Diagnostics

Description of Point Pattern	**Possible Interpretation**
All but a few points fall on a line	Outliers in the data
Left end of pattern is below the line; right end of pattern is above the line	Long tails at both ends of the data distribution
Left end of pattern is above the line; right end of pattern is below the line	Short tails at both ends of the distribution
Curved pattern	Data distribution is skewed
Staircase pattern (plateaus and gaps)	Data have been rounded or are discrete

Here is the graph using same method for example 3, showing a curved pattern, indicating a skewed distribution.



Section 7.5 The Normal Approximation to the Binomial Probability Distribution

►Normal Approximation to the Binomial

You can use SPSS to get the actual probability for the Binomial, so that the Normal Approximation is no longer necessary. It was more useful before computers. But, the approximation can still be done using the same methods as in section 7.3, but you must remember to make the continuity corrections.

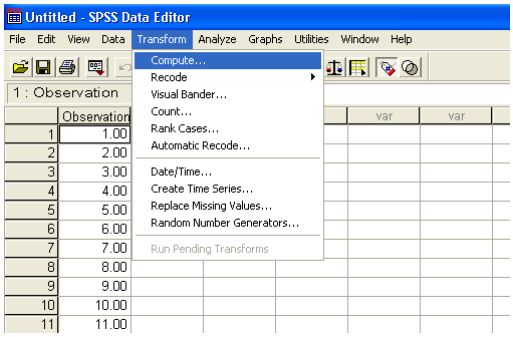
Chapter 8. Sampling Distributions

Section 8.1 Distribution of the Sample Mean

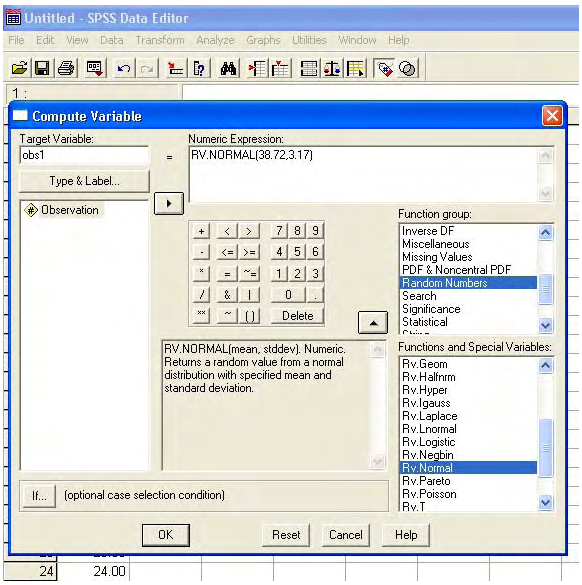
►Sampling Distribution Simulation - Mean

Example 2, Page 420

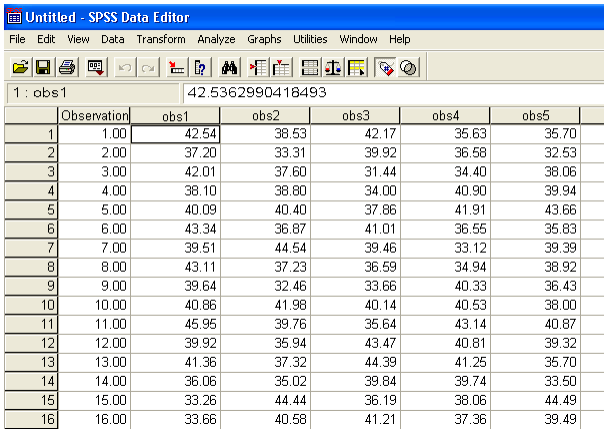
Similar to Chapter 5, section 1, to create 100 random numbers, you must first create a data set with 100 observations. SPSS will create new variables, with the same sample size as already exists, but will not create more observations than are already in the data set. To create the original sample of 100, you can type in the values 1 to 100 by hand, or use another program, such as Excel, where data generation is easier. After creating the data set with 100 observations, go to **Transform → Compute…**



Once in the Compute Variable window, you can choose a name for the new variable you wish to create. This name must follow SPSS naming conventions (no strange characters or spaces). You will be creating five variables in the same way, so it may be easiest to add a number to the end of the name, to keep the sample straight. In the Function Group box, select Random Numbers. In the Functions and Special Variables box, select Rv.Normal. This will create a random observation from a Normal distribution. There are two pieces of information that this requires, which are the mean and standard deviation of the distribution. We replace the two question marks with 38.72 and 3.17, for a mean of 38.72 and standard deviation of 3.17.

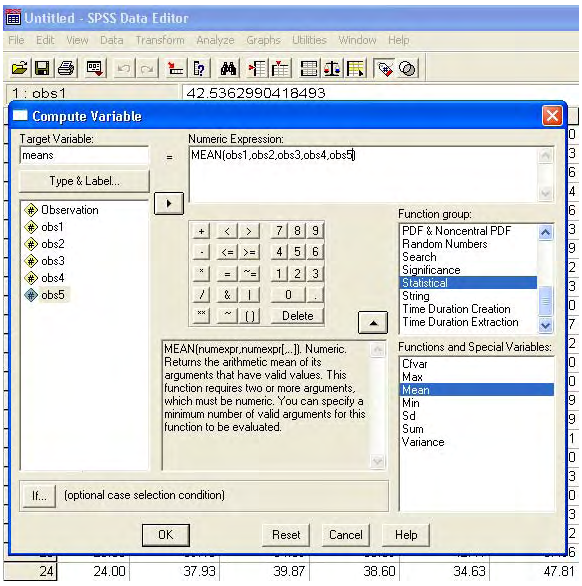


Once the equation and target variable name is correct, push OK. This will create a new variable of random Normal values. To create the other four observations, go back to **Transform → Compute…** and repeat the process. The equation will still be there, so all you have to do is change the variable name.

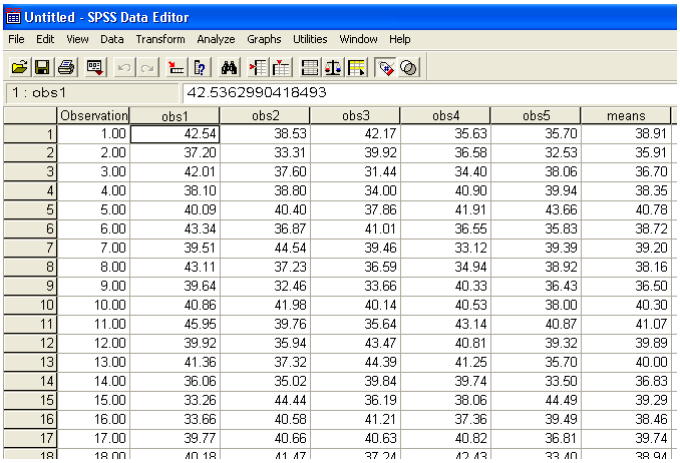


	Observation	obs1	obs2	obs3	obs4	obs5
1	1.00	42.54	38.53	42.17	35.63	35.70
2	2.00	37.20	33.31	39.92	36.58	32.53
3	3.00	42.01	37.60	31.44	34.40	38.06
4	4.00	38.10	38.80	34.00	40.90	39.94
5	5.00	40.09	40.40	37.86	41.91	43.66
6	6.00	43.34	36.87	41.01	36.55	35.83
7	7.00	39.51	44.54	39.46	33.12	39.39
8	8.00	43.11	37.23	36.59	34.94	38.92
9	9.00	39.64	32.46	33.66	40.33	36.43
10	10.00	40.86	41.98	40.14	40.53	38.00
11	11.00	45.95	39.76	35.64	43.14	40.87
12	12.00	39.92	35.94	43.47	40.81	39.32
13	13.00	41.36	37.32	44.39	41.25	35.70
14	14.00	36.06	35.02	39.84	39.74	33.50
15	15.00	33.26	44.44	36.19	38.06	44.49
16	16.00	33.66	40.58	41.21	37.36	39.49

Once you have all five observations, return to **Transform → Compute…**, but this time, we will be changing both the name and equation. Choose a name to represent the mean of the five observations. After clearing the equation in the Numeric Expression box, select Statistical in the Function Group box. From the Functions and Special Variables box, select Mean. The mean requires the data to find the mean of, so type in the name of all five variables that we just created, separating each name with a comma.

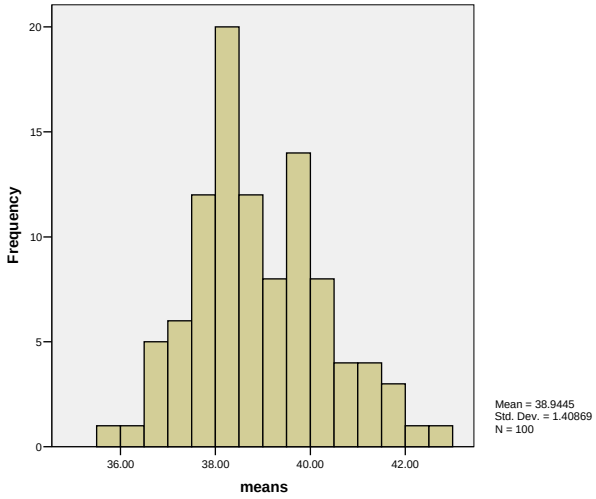


When the equation looks correct, and the name is as you want it, click OK. We now have a variable of sample means.



	Observation	obs1	obs2	obs3	obs4	obs5	means
1	1.00	42.54	38.53	42.17	35.63	35.70	38.91
2	2.00	37.20	33.31	39.92	36.58	32.53	35.91
3	3.00	42.01	37.60	31.44	34.40	38.06	36.70
4	4.00	38.10	38.80	34.00	40.90	39.94	38.35
5	5.00	40.09	40.40	37.86	41.91	43.66	40.78
6	6.00	43.34	36.87	41.01	36.55	35.83	38.72
7	7.00	39.51	44.54	39.46	33.12	39.39	39.20
8	8.00	43.11	37.23	36.59	34.94	38.92	38.16
9	9.00	39.64	32.46	33.66	40.33	36.43	36.50
10	10.00	40.86	41.98	40.14	40.53	38.00	40.30
11	11.00	45.95	39.76	35.64	43.14	40.87	41.07
12	12.00	39.92	35.94	43.47	40.81	39.32	39.89
13	13.00	41.36	37.32	44.39	41.25	35.70	40.00
14	14.00	36.06	35.02	39.84	39.74	33.50	36.83
15	15.00	33.26	44.44	36.19	38.06	44.49	39.29
16	16.00	33.66	40.58	41.21	37.36	39.49	38.46
17	17.00	39.77	40.66	40.63	40.82	36.81	39.74

Now, we can use the methods in chapter 2, section 2 to create a histogram of the means.

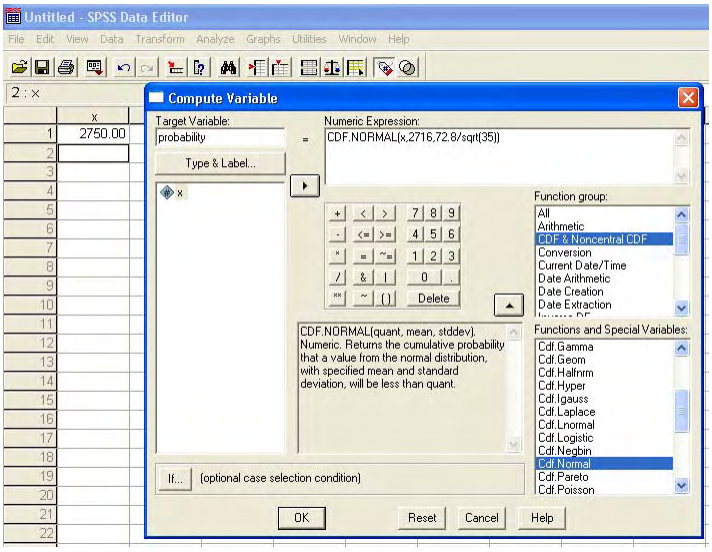


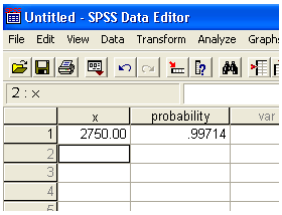
To sample from a non-Normal distribution, you can select from any of the other random number functions provided by SPSS, and repeat the same process.

►Applying the Central Limit Theorem

Example 6, Page 429

To find probabilities applying the Central Limit Theorem, we can use the same methods as in chapter 7, except now we must calculate the correct standard error first. So, where we selected Cdf.Normal, and replaced the question marks that appear with x, 2716, and 72.8 for a single observation, we will now use x, 2716, and 72.8/sqrt(35) , (SPSS uses sqrt(value) to calculate square roots). SPSS will accept an equation for the standard error, so we don't have to worry about rounding. SPSS does not care if the variable x contains single observations or sample means, as it assumes that you know which you wish to use, and that you will adjust the standard error appropriately. The same goes for the Idf.Normal. As long as you know if you are putting in the standard deviation or standard error, SPSS will give you the correct values.



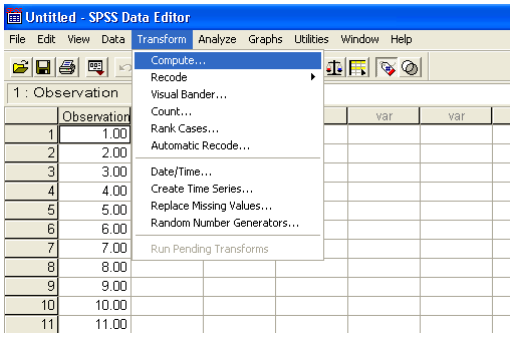


Section 8.2 Distribution of the Sample Proportion

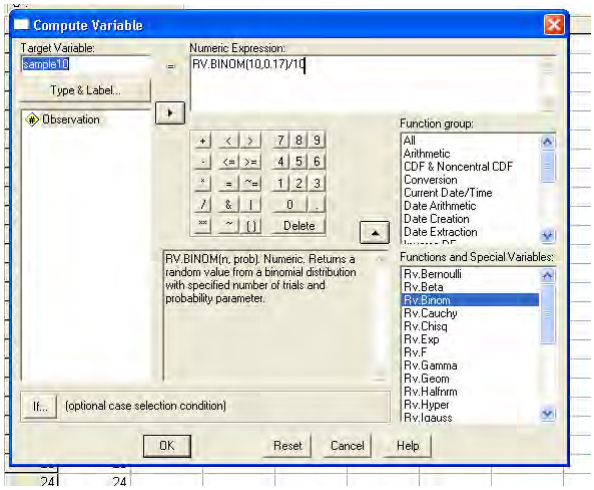
►Sampling Distribution Simulation - Proportion

Example 2, Page 435

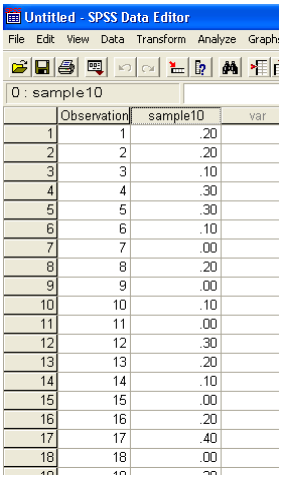
Similar to Chapter 5, and the sampling distribution of means, to create 100 random numbers, you must first create a data set with 100 observations. SPSS will create new variables, with the same sample size as already exists, but will not create more observations than are already in the data set. To create the original sample of 100, you can type in the values 1 to 100 by hand, or use another program, such as Excel, where data generation is easier. After creating the data set with 100 observations, go to **Transform → Compute…**



Once in the Compute Variable window, you can choose a name for the new variable you wish to create. This name must follow SPSS naming conventions (no strange characters or spaces). In the Function Group box, select Random Numbers. In the Functions and Special Variables box, select Rv.Binom. This will create a random observation from a Binomial distribution. There are two pieces of information that this requires, which are the sample size and the probability of a success. We replace the two question marks with 10 (40, 80) and 0.17. This will give a binomial value, which will be a number between 1 and 10. Because we are looking for proportions, we will divide the random variable by the sample size (10).

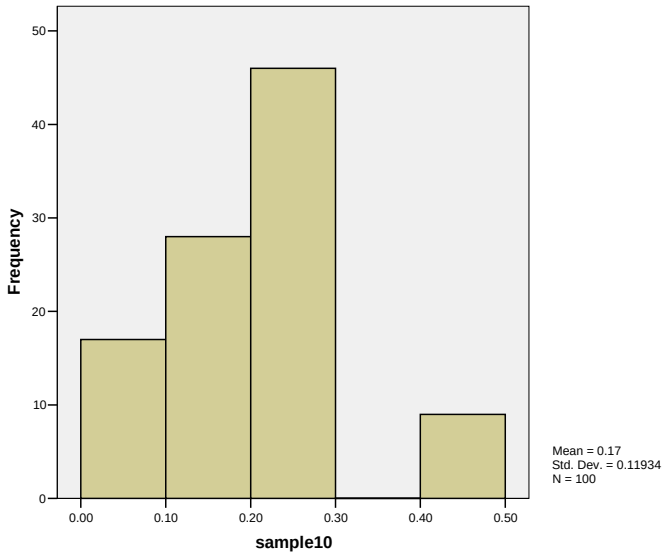


Once the equation and target variable name is correct, push OK. This will create a new variable of random proportions based on the Binomial distribution.

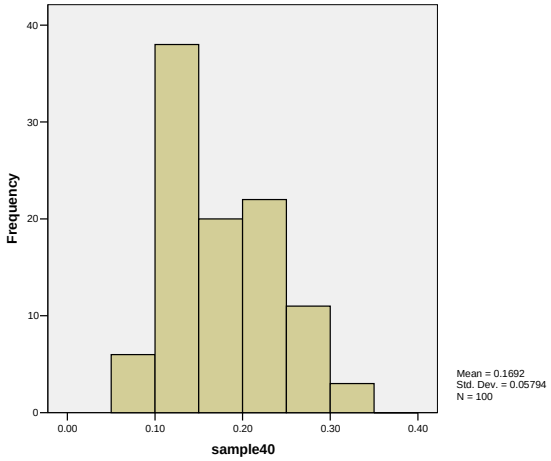


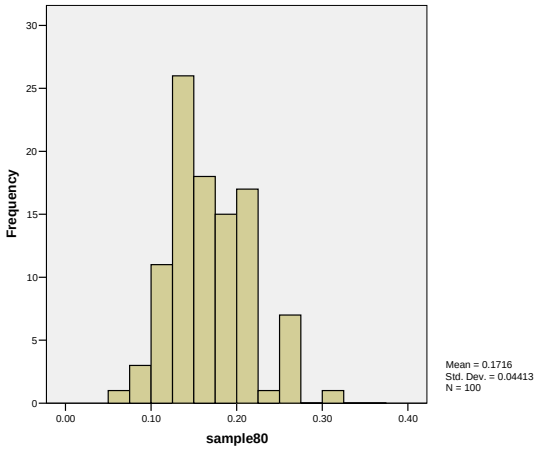
	Observation	sample10	var
1	1	.20	
2	2	.20	
3	3	.10	
4	4	.30	
5	5	.30	
6	6	.10	
7	7	.00	
8	8	.20	
9	9	.00	
10	10	.10	
11	11	.00	
12	12	.30	
13	13	.20	
14	14	.10	
15	15	.00	
16	16	.20	
17	17	.40	
18	18	.00	

Now, we can use the methods in chapter 2, section 2 to create a histogram of the means.



To obtain the histograms for sample of size 40 and 80, repeat the process, changing the sample size in the equation.





To sample from other non-Normal distribution, you can select from any of the other random number functions provided by SPSS, and repeat the same process.

Chapter 9. Estimating the Value of a Parameter

SPSS assumes that all data come from a sample. Because of this, all packaged methods assume σ to be unknown. Methods where σ is known must be done, at least in part, by hand.

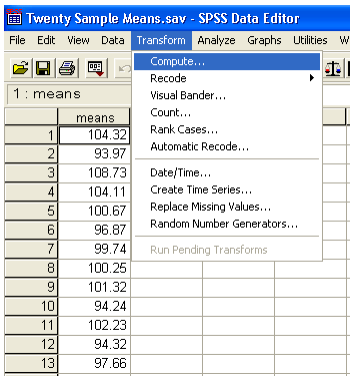
Section 9.1 The Logic in Constructing Confidence Intervals about a Population Mean

The point estimate, or sample mean, can be obtained through the methods covered in chapter 3.

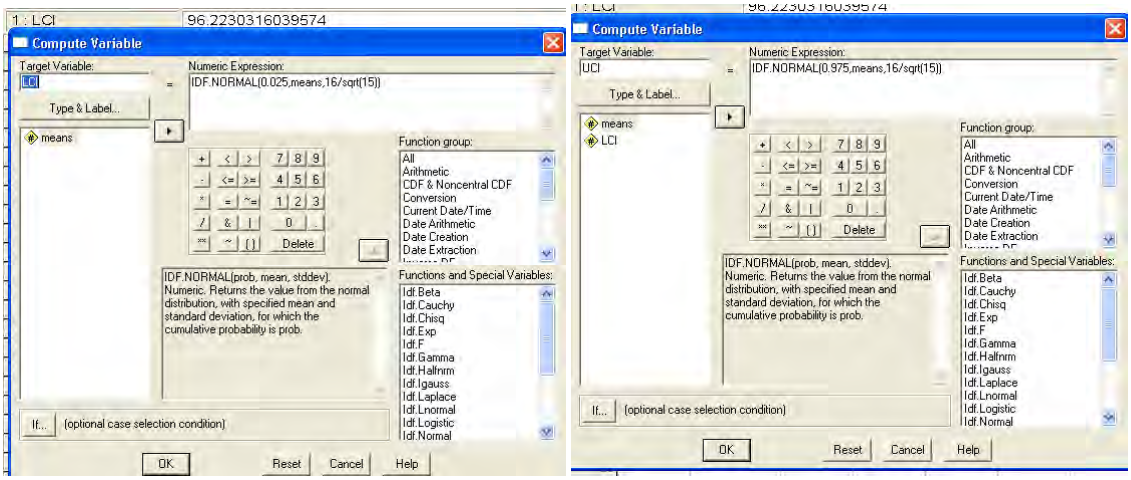
►Constructing 20 95% Confidence Intervals Based on 20 Samples

Example 2, Page 450

We can create samples using the method provided in chapter 8. Once we have the sample means in our data set, got to **Transform → Compute…**



We will have to create two values, one for the lower confidence limit, and one for the upper confidence limit. Create a name for the lower limit. There are two ways we can now create the confidence interval. We can type in the full equation after looking up the z-value on the table, which would be means – 1.96*16/sqrt(15) and means+1.96*16/sqrt(15), or, we can let SPSS find the z-value for us, using the IDF, or Inverse Distribution Function. IDF.NORMAL finds the value of z (when we use a mean of 0 and standard deviation of 1) where the area **to the left** of z is the probability we provide. In the case of a 95% confidence interval, the area in the tails is .05, so the area to the left for the lower limit is .025, or α/2. The area to the left for the upper limit will be 1-α/2, or 1-.025, which is .975. Here we have means + IDF.NORMAL(0.025,0,1)*16/sqrt(15), and means + IDF.NORMAL(0.975,0,1)*16/sqrt(15). Note that we have to use the area to the left, not the confidence level. We can also use the IDF for the full equation. Knowing that the IDF.NORMAL uses the probability to the left, a mean, and a standard error, we can use IDF.NORMAL(0.025, means, 16/sqrt(15)) as our equation, and receive the same values. Also, in both cases we are **adding** the IDF.NORMAL value, since the z-value at .025 will be negative. If we subtract the IDF.NORMAL(0.025,0,1) it will provide the same value as adding the IDF.NORMAL(0.975,0,1). An advantage of allowing SPSS to find the values is that SPSS won't round the values, as we have to when using the table.



When you have the equation as you want it, push OK. Do this again for the upper limit, changing the probability value. This will provide the lower and upper limits for the confidence interval.

Twenty Sample Means.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities

1 : LCI 96.2230316039574

	means	LCI	UCI
1	104.32	96.22	112.42
2	93.97	85.87	102.07
3	108.73	100.63	116.83
4	104.11	96.01	112.21
5	100.67	92.57	108.77
6	96.87	88.77	104.97
7	99.74	91.64	107.84
8	100.25	92.15	108.35
9	101.32	93.22	109.42
10	94.24	86.14	102.34
11	102.23	94.13	110.33
12	94.32	86.22	102.42
13	97.66	89.56	105.76
14	101.44	93.34	109.54
15	98.19	90.09	106.29
16	107.15	99.05	115.25
17	100.38	92.28	108.48
18	95.89	87.79	103.99
19	104.43	96.33	112.53
20	102.28	94.18	110.38
21			

SPSS does not work directly in computing confidence intervals using the standard normal table, as SPSS assumes that the population standard deviation is not known. SPSS can provide the point estimate from a data set, but the interval must be calculated by hand.

Example 4, Page 455

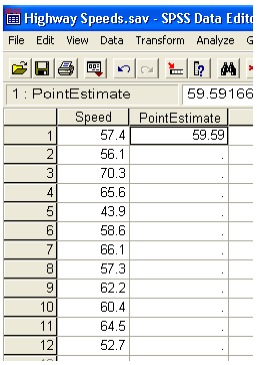
We follow the methods in chapter 3 to obtain the following information.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Speed	12	43.9	70.3	59.592	7.0171
Valid N (listwise)	12				

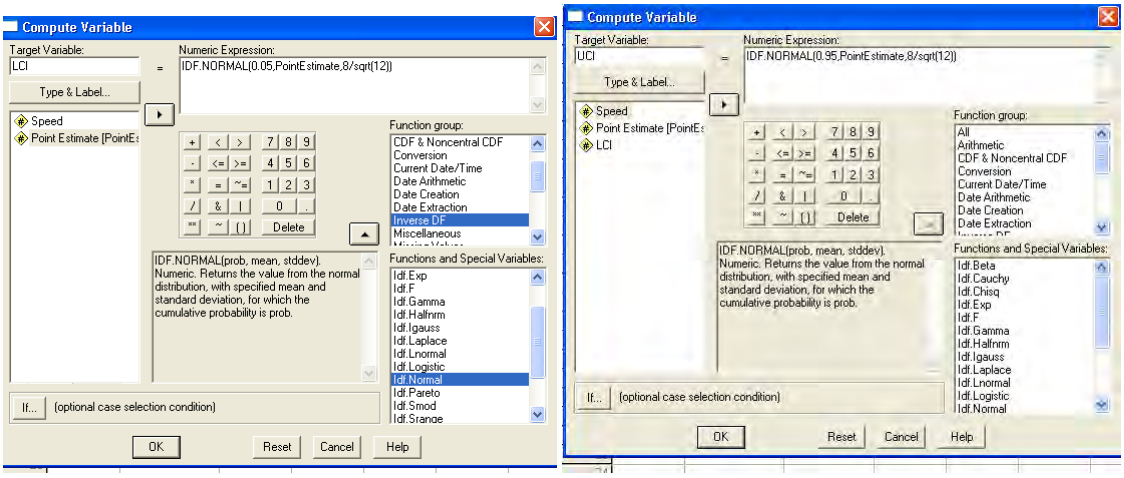
We can also use SPSS to create the box-plot, and Normal Probability Plot to check the data. We will note that the sample standard deviation is not the same as the population standard deviation. SPSS does not

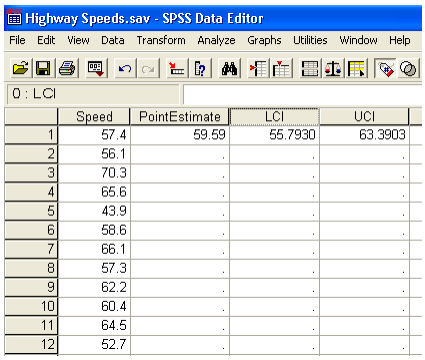
know the population standard deviation. To obtain the confidence interval we create a new column which contains the value of the point estimate.



	Speed	PointEstimate
1	57.4	59.59
2	56.1	.
3	70.3	.
4	65.6	.
5	43.9	.
6	58.6	.
7	66.1	.
8	57.3	.
9	62.2	.
10	60.4	.
11	64.5	.
12	52.7	.

We can then follow the same steps as example 2 to create the confidence interval.





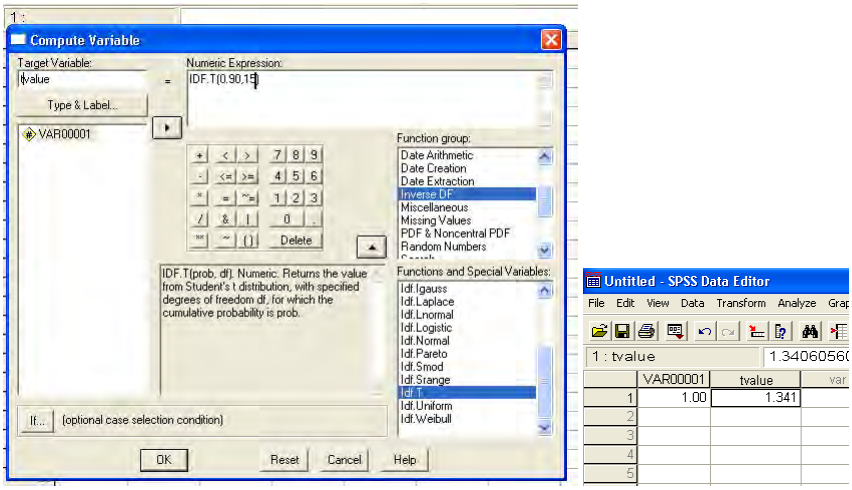
	Speed	PointEstimate	LCI	UCI
1	57.4	59.59	55.7930	63.3903
2	56.1	.	.	.
3	70.3	.	.	.
4	65.6	.	.	.
5	43.9	.	.	.
6	58.6	.	.	.
7	66.1	.	.	.
8	57.3	.	.	.
9	62.2	.	.	.
10	60.4	.	.	.
11	64.5	.	.	.
12	52.7	.	.	.

So, we are 90% confident that the mean speed of all cars traveling on the highway outside the subdivision is between 55.79 and 63.39 miles per hour.

Section 9.2 Confidence Intervals about a Population in Practice

►Finding t-values

Example 2, page 468
We can use the IDF.T to find t-values. We must remember that IDF is always looking for area to the left, so if we want to find the t-value where the area under the t-distribution to the right of the t-value is .10, assuming 15 degrees of freedom, we can use IDF.T(0.90,15). The IDF.T requires the area to the left, and the degrees of freedom. Unlike the IDF.NORMAL, you cannot get back to an x-value directly by using the IDF.T.

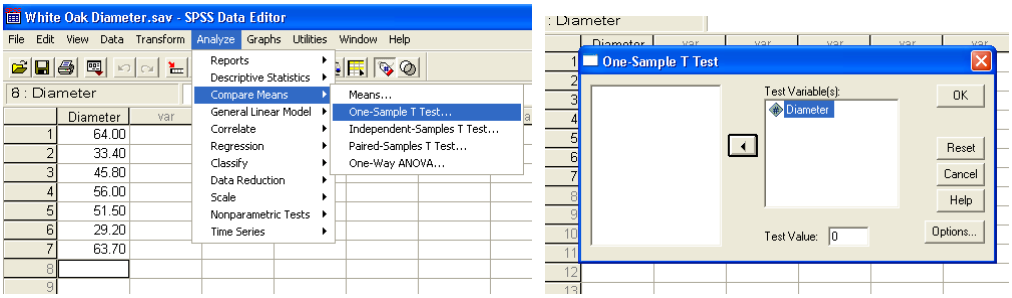


The value of $t_{0.10}$ with 15 degrees of freedom is 1.341.

►Constructing Confidence Intervals when σ is unknown

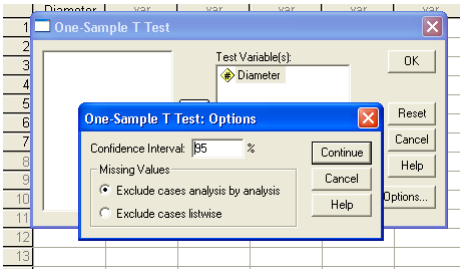
Example 3, page 470

First, we must enter the data. We can then do all the steps of a confidence interval.
Step 1. To create the box-plot and normal probability plot, see previous sections.
Step 2-Step 4. These are done in the computer. Go to **Analyze → Compare Means → One-Sample T Test…**



Highlight the variable of interest, and push the ► button. We will ignore the Test Value box for now. For the confidence interval, we want to leave this at 0.

To chose the level of confidence, click on the Options button.



Type the level of confidence in the box. Note-this is **not** in decimal form. When you have chosen the correct level of confidence, click Continue, and OK. The output window will have the following:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Diameter	7	49.0857	13.79570	5.21429

This provides the sample size, sample mean, sample standard deviation, and standard error, which is the sample standard deviation over the square root of the sample size. If you want to perform a confidence interval by hand, these are all the values you will need, except the table value.

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Diameter	9.414	6	.000	49.08571	36.3268	61.8446

The confidence interval is provided in the **last two boxes**. For now, we will ignore the first four boxes. NOTE- the t value given in the first box is not the t-value used for the confidence interval. The df is the correct degrees of freedom, or sample size -1.

Step 5. Interpretation: We are 95% confident that the mean diameter from the base of mature white oak trees is between 36.33 and 61.84 centimeters

Section 9.3 Confidence Intervals about a Population Proportion

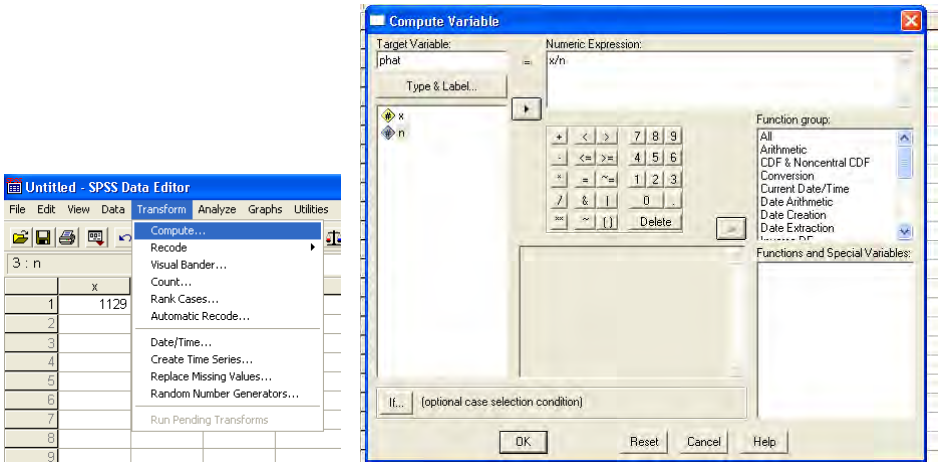
►Constructing Confidence Intervals about a Population Proportion

Oddly enough for a package originally designed for the Social Sciences, SPSS does not work directly with one or two proportions. There is very little advantage to using SPSS over a calculator unless you are working with multiple proportions at the same time, or you are worried about rounding.

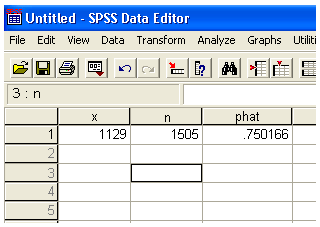
Example 2, Page 480

Step 1: Compute the value of $\hat{p}$

In SPSS, create one variable for the value x, and another for the value n. Go to **Transform → Compute…**, create a new variable, called phat, by putting the name phat in the Target Variable box, and the equation x/n in the Numeric Expression box.

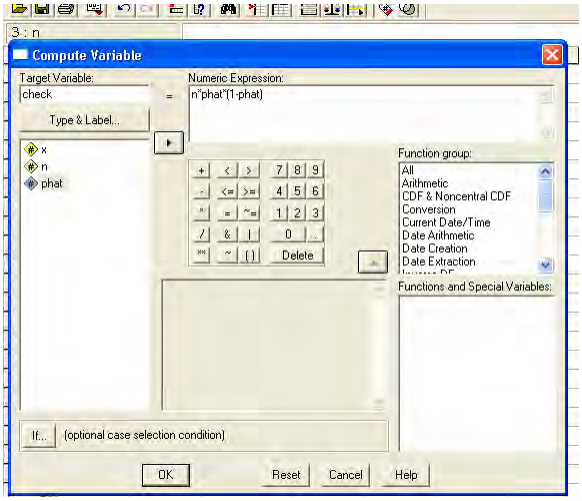


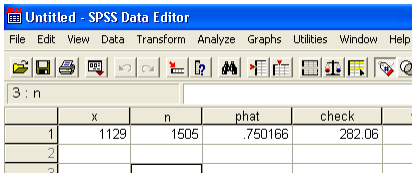
Remember that SPSS will not accept a name with odd characters, so p-hat is not accepted.



The point estimate will appear in the Data View sheet, although you may need to change the number of decimal places, using the Variable View sheet to see it better. In this example, $\hat{p}$ is 0.7502, or, rounding to 4 decimal places, 0.0750. SPSS does not round.

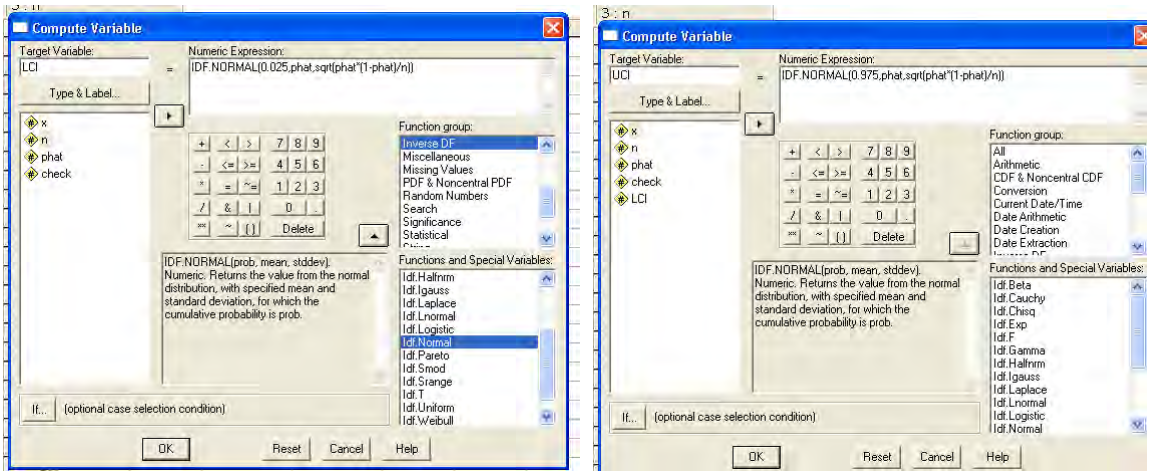
Step 2: We can check this using the **Transform → Compute…** as well. Create a new variable, check, by using the equation n*phat*(1-phat) in the numeric expression box.

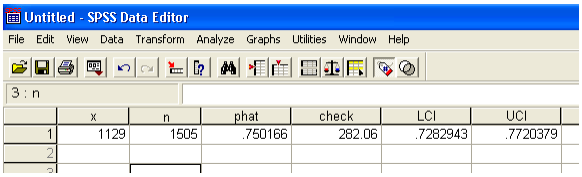




As the value is 282.06, which is greater than 10, we can proceed to construct the confidence interval. One advantage of using the computer is that you don't have to worry about rounding, but this may present slight differences from what you see in the book.

Step 3-4: This is similar to constructing a confidence interval with σ known. We will create two new variables, one for the lower bound, and one for the upper bound. Returning to **Transform → Compute…**, we will start with the lower bound, so we create a name, and in the Numeric Expression box, type the equation. We can again either use a table value, or use the SPSS tables. For a 95% confidence interval, the areas to the left are 0.025 and 0.975, leaving 0.95 in the middle. So, our equation for the lower bound will be phat + IDF.NORMAL(.025,0,1)*sqrt(phat*(1-phat)/n), and the equation for the upper bound will be phat + IDF.NORMAL (.975,0,1)*sqrt(phat*(1-phat)/n), or, similar to means, we can use IDF.NORMAL(probability to the left, point estimate, standard error).





So, in this example, we are 95% confident that the proportion of Americans who are in favor of tighter enforcement of government rules on TV content during hours when children are most likely to be watching is between .728 and .772.

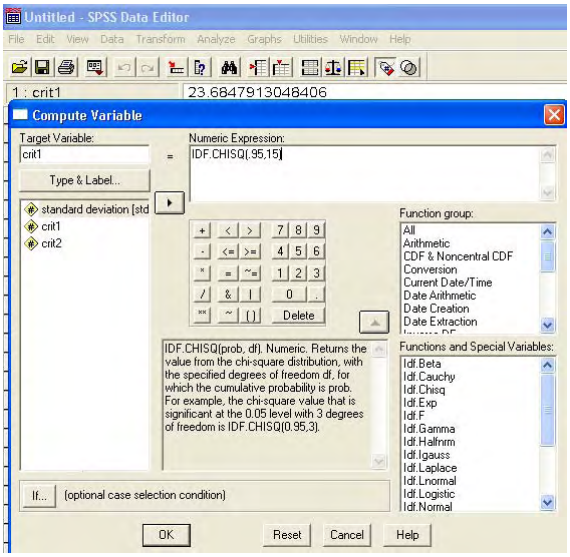
Section 9.4 Confidence Intervals about a Population Standard Deviation

SPSS does not have a method to directly create confidence intervals for standard deviations. You can use the IDF.CHISQ in **Transform → Compute…** to find the critical values for the Chi-Square distribution, but the rest must be done either by hand, or similar to the method for proportions.

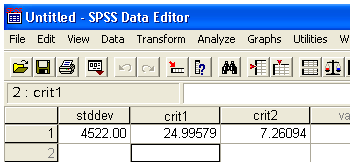
►Finding Critical values for the Chi-Square Distribution

Example 1, Page 488

Before calculating the critical values, you must have a variable already existing. To find the critical values, go to **Transform → Compute…**. Create a name for the new variable, such as critical1. In the Function Group, select Inverse DF, and from the Functions and Special Variables, select IDF.CHISQ.



The Inverse Chi-Square requires two pieces of information. The first is the probability to the left. This will be opposite of the tables in the book, which are areas to the right. So, for $\chi^2_{.95}$ in the book, we will look up the value with an area of .05 to the left of the value, and for $\chi^2_{.05}$, we will look up the value with an area of .95 to the left. The second piece of information is the degrees of freedom. So, for $\chi^2_{.95}$, we will use the equation IDF.CHISQ(.05,15), and for $\chi^2_{.05}$, we will use the equation IDF.CHISQ(.95,15).



So, we get $\chi^2_{.95}$=7.26094, and $\chi^2_{.05}$=24.99579.

►Constructing Confidence Intervals about a Population Standard Deviation

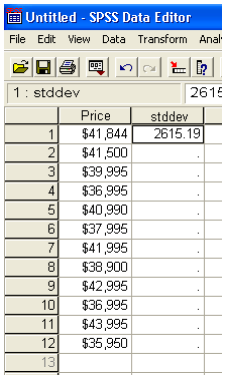
Example 2, Page 490

Calculating a confidence interval for the standard deviation in SPSS is similar to finding a confidence interval for μ when σ is known, or finding the confidence interval for a proportion. First, we can find the sample standard deviation.

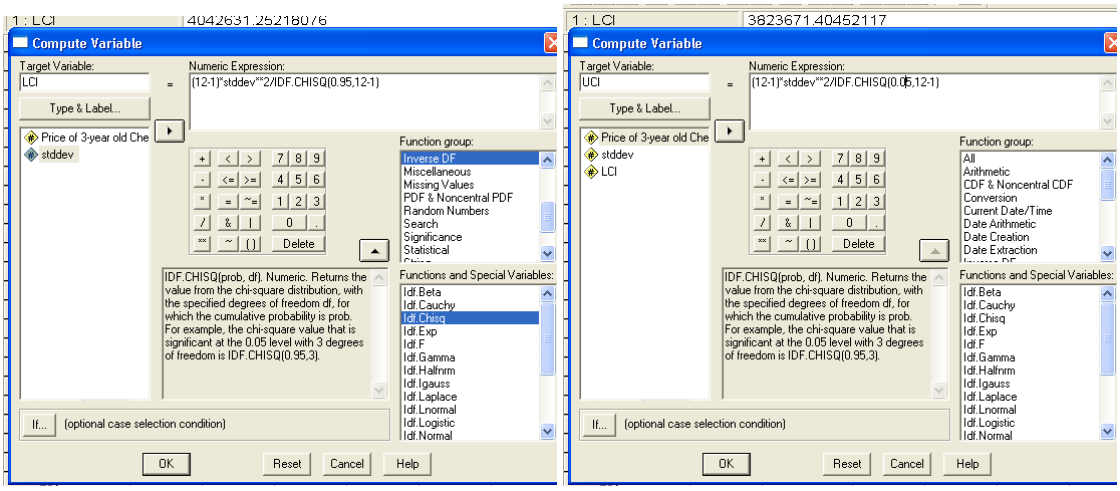
Descriptive Statistics

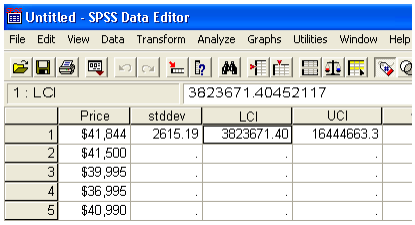
	N	Minimum	Maximum	Mean	Std. Deviation
Price of 3-year old Chevy	12	$35,950	$43,995	$40,012.42	$2,615.187
Valid N (listwise)	12				

We can then copy the value of the standard deviation into a new column.

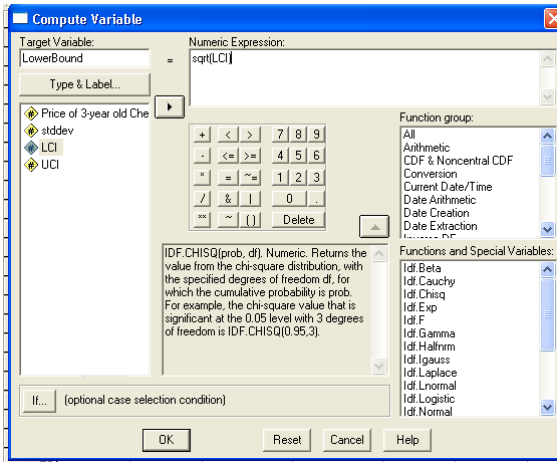


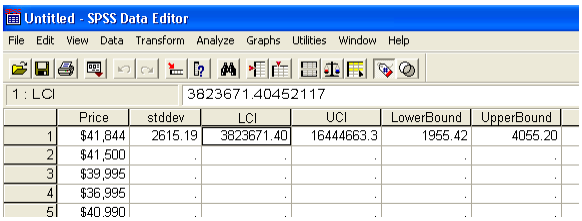
We then go to **Transform → Compute….** Create a name for the lower bound in the Target Variable box, and in the Numeric Expression box, use the equation (15-1)*stddev**2/ IDF.CHISQ (.95,14). The double asterisk represents a power, so stddev**2 is stddev^2. Remember that the tables in SPSS are to the left, where the ones in the book are to the right, so where the book has $\chi^2_{\alpha/2}$, SPSS wants $\chi^2_{1-\alpha/2}$.





We can then use the **Transform → Compute…** to take the square root of these values, to get the standard deviations.





	Price	stddev	LCI	UCI	LowerBound	UpperBound
1	$41,844	2615.19	3823671.40	16444663.3	1955.42	4055.20
2	$41,500	.	.	.	.	.
3	$39,995	.	.	.	.	.
4	$36,995	.	.	.	.	.
5	$40,990	.	.	.	.	.

So, we are 90% confident that the population standard deviation of the price of a three year old Chevy Corvette is between $1,955.42 and $4,055.20.

Chapter 10. Testing Claims Regarding a Parameter

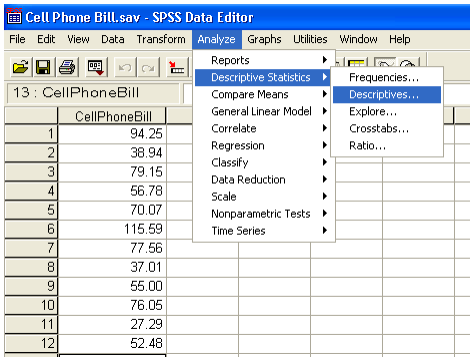
SPSS assumes that all data come from a sample. Because of this, all packaged methods assume σ to be unknown. Methods where σ is known must be done, at least in part, by hand.

Section 10.2 A Model for Testing Claims about a Population Mean

►Testing a hypothesis about μ, σ known

Example 5, Page 524

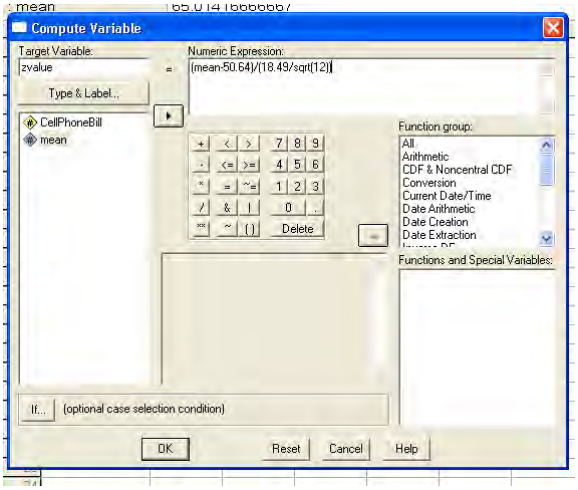
Since SPSS assumes that σ is unknown, the only pieces of the hypothesis test that SPSS will provide are the sample mean and the p-value. First, we must enter the data. Once the data have been entered, go to **Analyze → Descriptive Statistics → Descriptives…**, select the variable that you want the mean of, and push OK. The sample mean will be in the output window.



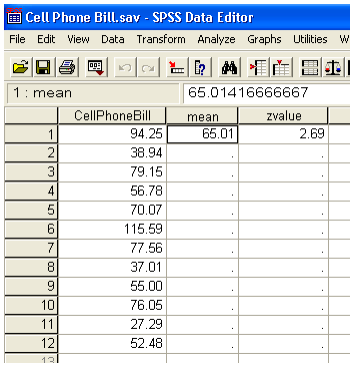
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CellPhoneBill	12	27.29	115.59	65.0142	25.45873
Valid N (listwise)	12				

If you click on the cell that contains the mean, until the mean is highlighted, you can copy the value of the sample mean into the clipboard, 65.01416666667. This will prevent rounding problems. You can then paste this into a new column. To calculate the z-value, go to **Transform → Compute…**. Once in the Compute Variable window, create a name for the z-value, and use the copied sample mean value in the z-equation.

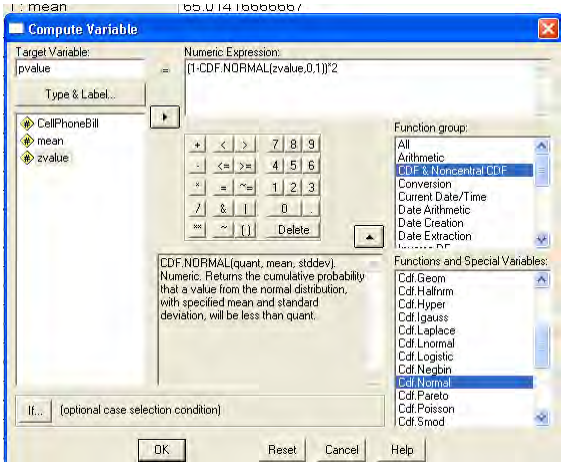


Once the equation is correct, push OK. Remember to get the parentheses in the correct places. Putting too many parentheses is better than not having enough to make sure the order of operations for the computer is what you believe they should be. This will create a new variable, with the value of z, 2.69.

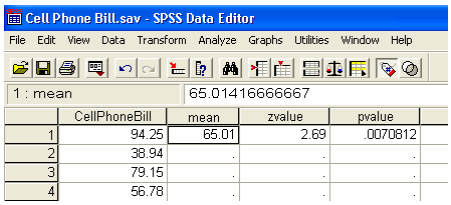


	CellPhoneBill	mean	zvalue
1	94.25	65.01	2.69
2	38.94	.	.
3	79.15	.	.
4	56.78	.	.
5	70.07	.	.
6	115.59	.	.
7	77.56	.	.
8	37.01	.	.
9	55.00	.	.
10	76.05	.	.
11	27.29	.	.
12	52.48	.	.

The test statistic is 2.69, or the sample mean is 2.69 standard errors above the hypothesized mean. To get the two-sided p-value, we will again use **Transform → Compute…**. Clear the z equation in the Compute Variable window, and click on CDF and Noncentral CDF in the Function Group box. In the Functions and Special Variables, select CDF.NORMAL. It must be remembered that this is the area to the left of the z-value. Since the CDF is the area to the left of z, and we want the tail areas, we will use 1-CDF.NORMAL, with a mean of 0 and standard deviation of 1, and since this is a two tailed test, we will multiply this value by 2.



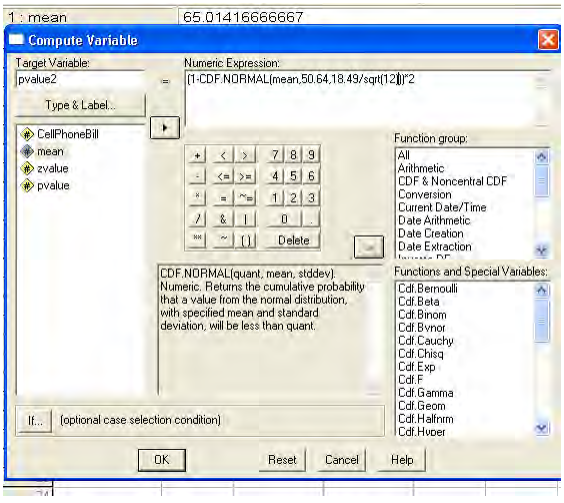
Once the equation is done, push OK. The p-value will be a new variable in the data set. You may need to go to Variable View to change the number of decimal places to see the actual p-value. The default is to show only two decimal places.



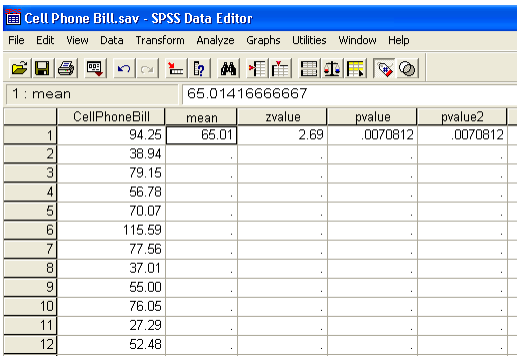
	CellPhoneBill	mean	zvalue	pvalue
1	94.25	65.01	2.69	.0070812
2	38.94	.	.	.
3	79.15	.	.	.
4	56.78	.	.	.

This value may be slightly different from what you would get by hand, as the z-value is not rounded to two decimal places. We get a p-value of .0071 which is less than .05, so we reject the null hypothesis.

If you don't need to know the z-value, and just want to calculate the p-value, you can skip the z-value calculation. We can calculate the tail probabilities directly for the sample mean.



Use the CDF.NORMAL, with the value of the sample mean as the first value, the hypothesized mean as the second value, and the standard deviation over the square root of the sample size as the third value. Here you will have to see if the sample mean is larger or smaller than the hypothesized mean in order to see if you want the CDF or 1-CDF to get the tail values.


Cell Phone Bill.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : mean 65.01416666667

	CellPhoneBill	mean	zvalue	pvalue	pvalue2
1	94.25	65.01	2.69	.0070812	.0070812
2	38.94	.	.	.	.
3	79.15	.	.	.	.
4	56.78	.	.	.	.
5	70.07	.	.	.	.
6	115.59	.	.	.	.
7	77.56	.	.	.	.
8	37.01	.	.	.	.
9	55.00	.	.	.	.
10	76.05	.	.	.	.
11	27.29	.	.	.	.
12	52.48	.	.	.	.

The p-value is the same, whether you do the test step-by-step, or whether you calculate the p-value directly.

Section 10.3 Testing Claims about a Population Mean in Practice

►Testing a hypothesis about μ, σ unknown

Example 3, Page 535

First, we must enter the data.

Normality Check

To create the box-plot and normal probability plot, see previous sections.

Step 1: Based on the quality control engineer's claim that the "fun size" Snickers bars **do not** weigh an average of 20.1 grams, we get

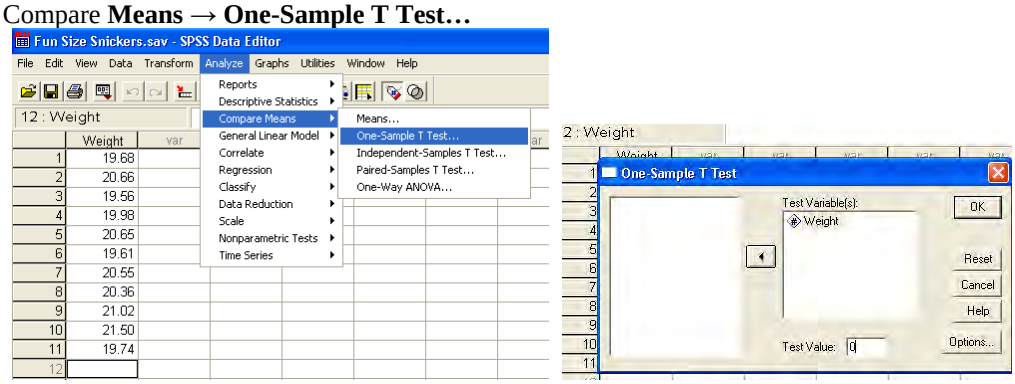
H_0: $\mu = 20.1$ -status quo (what M&M-Mars states on the bar)

H_1: $\mu \neq 20.1$ -quality control engineer's claim

Step 2: α=.01. If the p-value is less than .01 we will reject the null hypothesis. This is similar to finding a critical region, but since the computer calculates the p-value for us, there is no necessity in looking up values on the t-table.

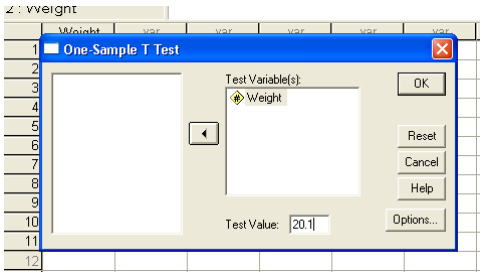
Step 3-4: done in SPSS

Go to Compare **Means → One-Sample T Test…**



Highlight the variable of interest, and push the ► button.

In the Test Value box, type in the value from the **Null Hypothesis**. (μ = 20.1)



Click OK. The output window will have the following:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Weight	11	20.3009	.64037	.19308

This provides the sample size, sample mean, sample standard deviation, and standard error, which is the sample standard deviation over the square root of the sample size. If you want to perform a test by hand, these are all the values you will need.

One-Sample Test

	Test Value = 20.1					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Weight	1.041	10	.323	.20091	-.2293	.6311

Note that the hypothesized mean appears above the test. In creating confidence intervals, this value should be 0. The value of t in the first box is the **Test Statistic** (calculated in Step 3). The sample mean is 1.041 standard errors (standard deviation of the sampling distribution) above the hypothesized mean.
The degrees of freedom in the second box are the degrees of freedom for the test.
The p-value is .323 which is larger than α (.323>.01) so we fail to reject the null hypothesis.
The confidence interval provided is the confidence interval for the difference between the sample mean and the hypothesized mean. We are 95% confident that the sample mean is 0.2293 below to 0.6311 grams above the hypothesized mean. Since 0 is in this interval, we will fail to reject the null hypothesis.

Note: SPSS always provides the p-value for a two-tailed test. If we want a 1-sided test, we can divide this in half, but we need to pay attention to the direction of the test and the sign of the test statistic. (0.323/2=0.1615), so p-value for a test with H_1: $\mu>20.1$, the p-value would be 0.1615. For a test with H_1: $\mu<20.1$, the p-value would be 1-0.1615, or 0.8385.

Step 5: We do not have sufficient evidence to show that the mean "fun size" Snickers bar weighs something other than 20.1 grams at the $\alpha=.01$ level of significance.

Interpreting the result in step 4 is more important than getting the result. If others cannot understand the results then we have wasted the time in performing the test. Make sure that the statement declares what was measured (not just that the mean has(n't) changed) and the level of significance that was used, or the p-value, so that future researchers can verify the results.

We can also use SPSS to get the t-value to use for the critical region; we use **Transform → Compute…**,

similar to finding the critical values for a confidence interval. Create a name for the column, t, and use IDF.T(p,df), where p is the area to the left, (so for an upper tail test, use 1- α, for a lower tail test, use α, and for a two tailed test, use α/2 and 1- α/2) and the correct degrees of freedom.

Section 10.4 Testing Claims about a Population Proportion

►Testing a Hypothesis about a Population Proportion

SPSS does not work directly with proportions. There is very little advantage to using SPSS over a calculator unless you are working with multiple proportions at the same time.

Example 1, Page 546

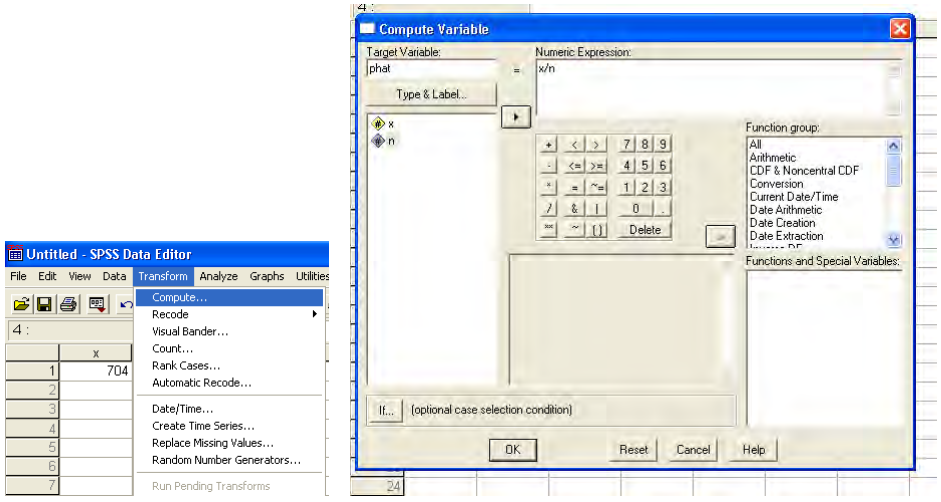
Step 1: The claim is that the proportion of adult Americans who thought that the death penalty was morally acceptable has increased since 2004.

H_0: p = 0.65

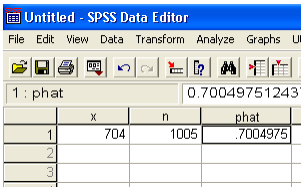
H_1: p > 0.65

Step 2: α=0.05.

Step 3: Compute the value of $\hat{p}$. In SPSS, create one variable for the value x, and another for the value n. Go to **Transform → Compute…**, create a new variable, called phat, by putting the name phat in the Target Variable box, and the equation x/n in the Numeric Expression box.

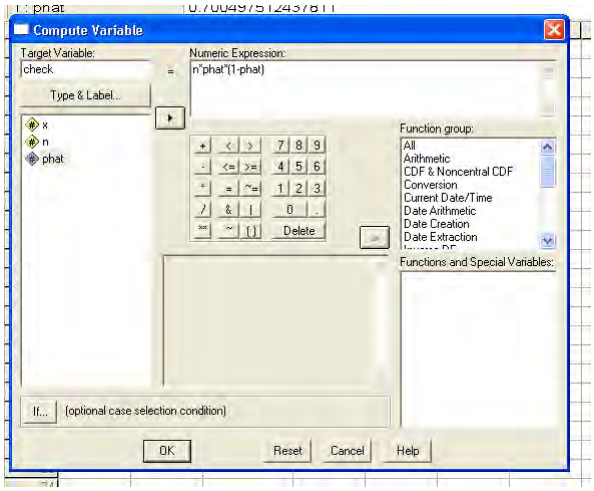


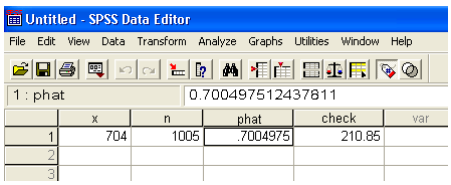
When the name and equation are ready, push OK.



The point estimate will appear in the Data View sheet, although you may need to change the number of decimal places, using the Variable View sheet to see it better. In this example, $\hat{p}$ is 0.7004975, or 0.70.

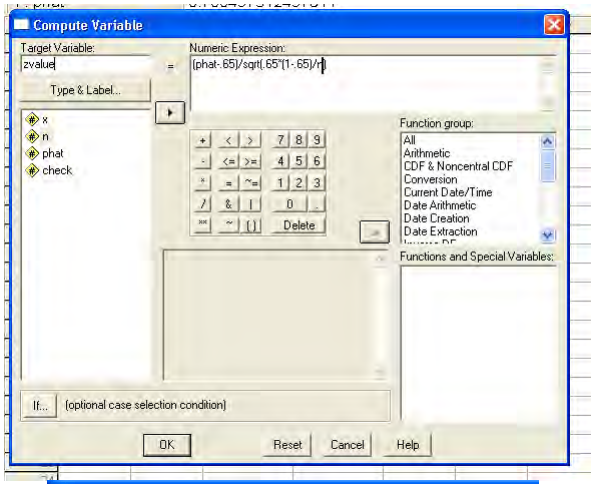
We can check the normality assumption using the **Transform → Compute…** as well. Create a new variable, check, by using the equation n*phat*(1-phat) in the numeric expression box.

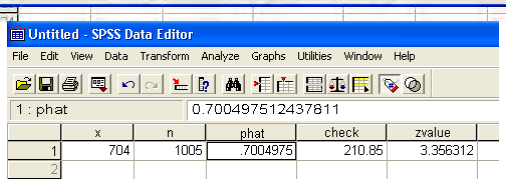




As the value is 210.85, which is greater than 10, we can proceed with the test. One advantage of using the computer is that you don't have to worry about rounding, but this may present slight differences from what you see in the book.

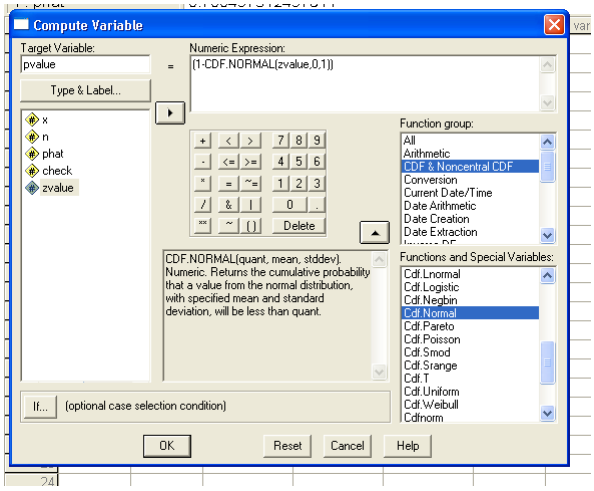
Returning to **Transform → Compute…**, we will calculate the z-value, similar to the test for μ with σ unknown. Type in the Z-equation, $z=(phat-p_0)/sqrt(p_0*(1-p_0)/n)$, substituting the value of p_0 into the equation.



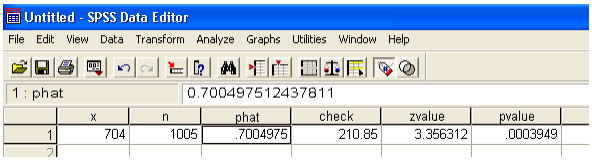


So, in this example, z = 3.36.

Step 5: We can then use SPSS to calculate the p-value. To get the one-sided p-value, we will again use **Transform → Compute…**. This is similar to what was done in chapter 7. Clear the z equation in the Compute Variable window, and click on CDF and Noncentral CDF in the Function Group box. In the Functions and Special Variables, select CDF.NORMAL. It must be remembered that this is the area to the left of the z-value.

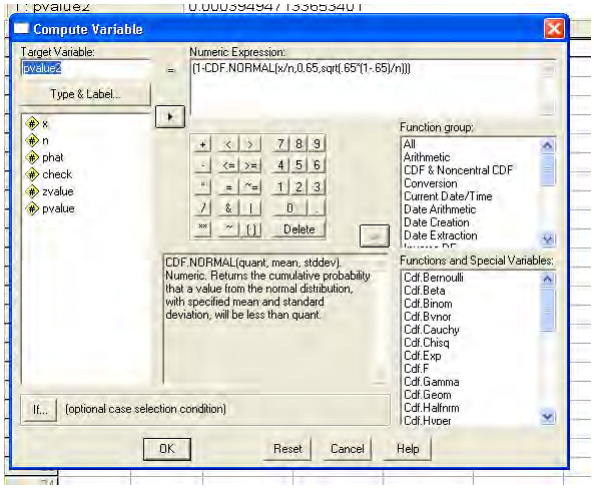


Once the equation is done, push OK. The p-value will be a new variable in the data set. Again, it will appear as many times as you had variables, and you may need to go to Variable View to change the number of decimal places.



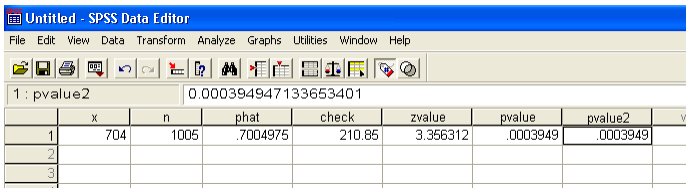
	x	n	phat	check	zvalue	pvalue
1	704	1005	.7004975	210.85	3.356312	.0003949

If you don't need to know the z-value, and just want to calculate the p-value, you can skip the z-value calculation. We can calculate the tail probabilities directly for the sample proportion, we just need to make sure to use the correct standard error.



Use the CDF.NORMAL, with the value of the x/n as the first value, the hypothesized proportion as the second value, and the standard error equation as the third value. Here you will have to see if the sample

proportion is larger or smaller than the hypothesized mean in order to see if you want the CDF or 1-CDF to get the tail values.



	x	n	phat	check	zvalue	pvalue	pvalue2
1	704	1005	.7004975	210.85	3.356312	.0003949	.0003949
2							
3							

The p-value is the same, whether you do the test step-by-step, or whether you calculate the p-value directly.

Section 10.5 Testing a Claim about a Population Standard Deviation

SPSS does not have a method to test standard deviations. You can use the CDF.CHISQ in **Transform → Compute…** to find the p-values for the Chi-Square distribution, but the rest must be done either by hand, or similar to the method for proportions.

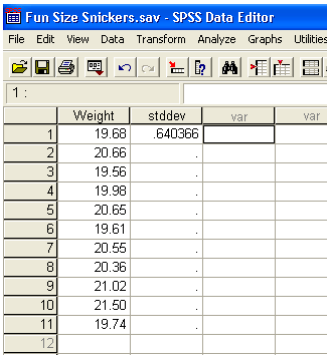
►Testing a Hypothesis about σ

Example 1, Page 555

Calculating a test statistic for the standard deviation in SPSS is similar to finding a test statistic for μ when σ is unknown, or finding the test statistic for a proportion. First, we create a variable with the sample standard deviation. This can be done by hand, or we can use **Analyze → Descriptive Statistics → Descriptives…** to calculate s, and copy it into a cell on the data page.

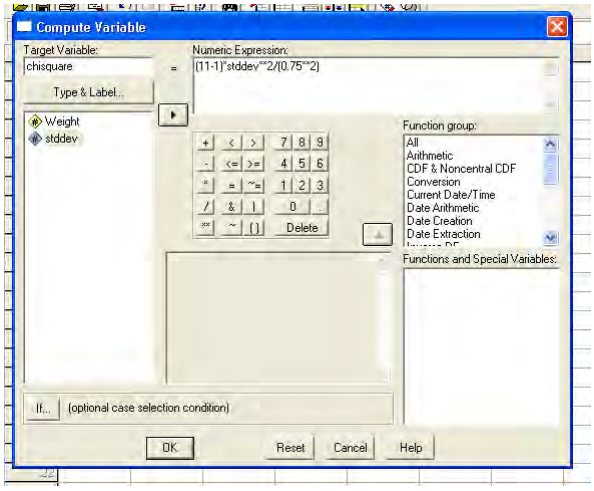
Descriptive Statistics

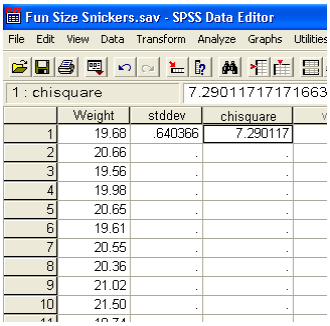
	N	Minimum	Maximum	Mean	Std. Deviation
Weight	11	19.56	21.50	20.3009	.64037
Valid N (listwise)	11				



	Weight	stddev	var	var
1	19.68	.640366		
2	20.66	.		
3	19.56	.		
4	19.98	.		
5	20.65	.		
6	19.61	.		
7	20.55	.		
8	20.36	.		
9	21.02	.		
10	21.50	.		
11	19.74	.		
12				

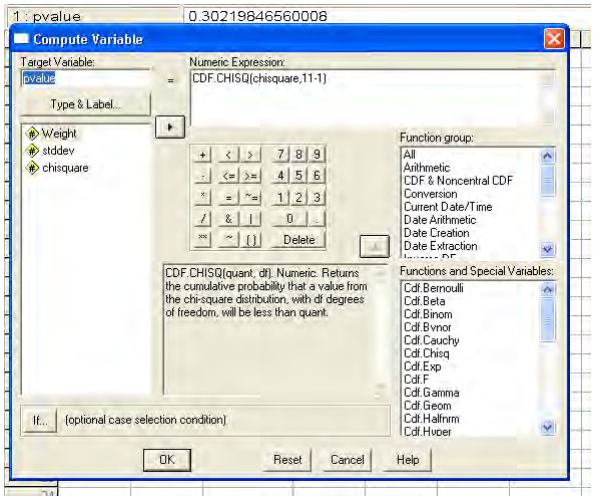
We then go to **Transform → Compute….** Create a name for the test statistic in the Target Variable box, and in the Numeric Expression box, use the equation (11-1)*s**2/hypothesized value**2. The double asterisk represents a power, so s**2 is s^2.



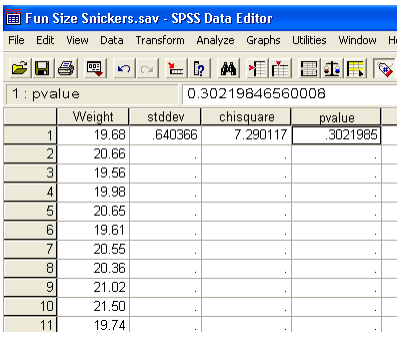


	Weight	stddev	chisquare
1	19.68	.640366	7.290117
2	20.66	.	.
3	19.56	.	.
4	19.98	.	.
5	20.65	.	.
6	19.61	.	.
7	20.55	.	.
8	20.36	.	.
9	21.02	.	.
10	21.50	.	.

We can then use CDF.CHISQUARE in **Transform → Compute…** to find the p-value.



Here, we must remember that the CDF is the area to the left. The CDF.CHISQ requires a chi-square value, and the degrees of freedom for the test. You cannot calculate the p-value without the chi-square value as you could with the normal distribution.



	Weight	stddev	chisquare	pvalue
1	19.68	.640366	7.290117	.3021985
2	20.66	.	.	.
3	19.56	.	.	.
4	19.98	.	.	.
5	20.65	.	.	.
6	19.61	.	.	.
7	20.55	.	.	.
8	20.36	.	.	.
9	21.02	.	.	.
10	21.50	.	.	.
11	19.74	.	.	.

The p-value is .3022 which is larger than .05, so we fail to reject the null hypothesis. There is not sufficient evidence to support the claim that the standard deviation is less than 0.75 miles at the α=.05 level of significance.

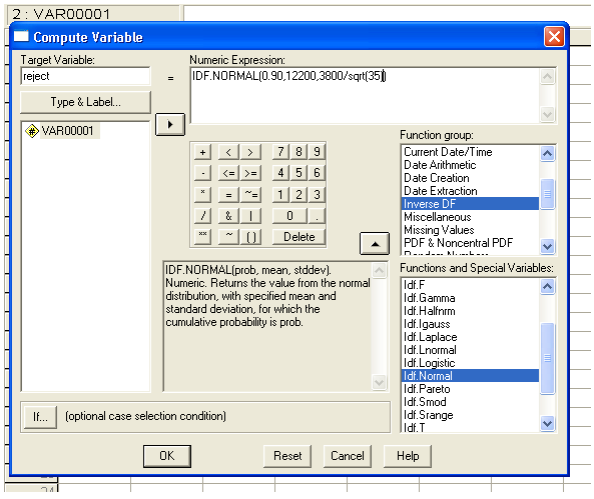
Section 10.7 The Probability of a Type II Error and the Power of the Test

►The probability of a type II error of the test

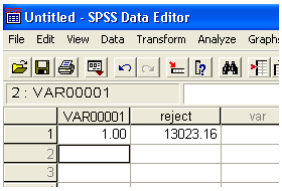
Finding the probability of a type II error in SPSS is similar to finding the p-value for the test, we just change the values to be those for the rejection region under the true distribution instead of the hypothesized distribution.

Example 1, Page 563

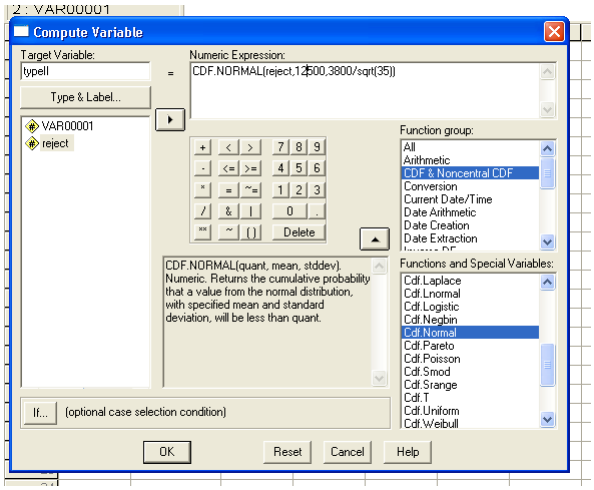
We had the following problem: H_0: μ=12,200, and H_1: μ>12,200. One of the problems with computing the probability of a type II error in practice is the fact that there are an infinite number of values under the alternate hypothesis. We had n=35, σ=3,800, and α=0.1. Based on this information, we know that we will reject the null hypothesis for any sample mean larger than $\mu_0+z_{0.1}*\sigma/\sqrt{n}$, we can find this using SPSS.



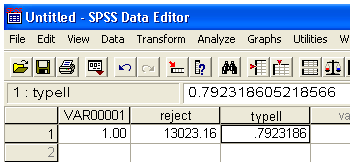
Using the IDF.NORMAL provides the value under the Normal curve for which the area to the left is 0.90 (which means the area to the right is α).



So, we will reject the null hypothesis for any sample mean larger than 13,023.16. Any sample mean less than this will lead us to fail to reject the null hypothesis. The probability of making a type I error is already known to us, but now we want to calculate the probability of a type II error. To calculate this, we are assuming that the true mean is 12,500. Since we had α be the area to the right of 13,023.16 under the Normal curve with the hypothesized mean, the probability of a type II error will be the area to the left (opposite direction of the type I error) of the same point, 13,023.16, with the true mean. So, now we will use the CDF.NORMAL to calculate the probability.



This should look the same as finding a p-value, except that we are using the critical value of the rejection region for x, and the new true mean instead of the hypothesized mean as the center.



There is a probability of 0.7923 of making a type II error if the true mean is 12,500 for a test with α=0.10.

Chapter 11. Inferences on Two Samples

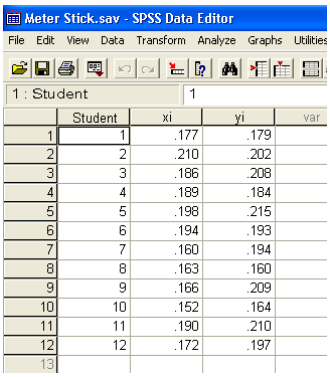
SPSS assumes that all data come from a sample. Because of this, all packaged methods assume σ to be unknown. Methods where σ is known must be done, at least in part, by hand.

Section 11.1 Inference about Two Means: Dependent Samples

►Inferences about Two Means: Dependent Samples

Example 2, Page 577

First, we type in the data. For dependent samples, each value is it's own variable, and it is paired by the observation.



	Student	xi	yi	var
1	1	.177	.179	
2	2	.210	.202	
3	3	.186	.208	
4	4	.189	.184	
5	5	.198	.215	
6	6	.194	.193	
7	7	.160	.194	
8	8	.163	.160	
9	9	.166	.209	
10	10	.152	.164	
11	11	.190	.210	
12	12	.172	.197	
13				

Normality Check

Use **Transform → Compute…** to calculate the differences. In the equation box, take the first variable minus the second variable. These are not needed for the test, but are necessary to check the normality of the differences. To create the box-plot and normal probability plot, see previous sections.

Step 1: Based on Professor Neill's claim that reaction time in the dominant hand is **less** than the reaction time in the non-dominant hand, we use

H_0: $\mu_d = 0$ -status quo (no difference in hands)

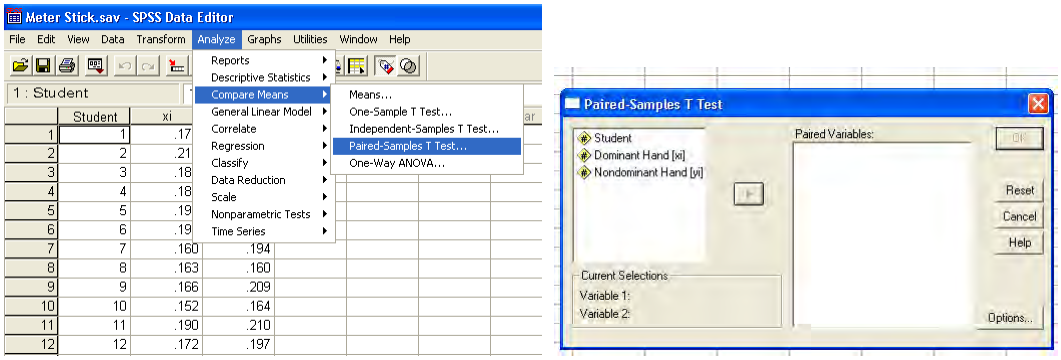
H_1: $\mu_d < 0$ -Professor's claim

assuming that the difference is Dominant-Non-dominant. If we took Non-dominant-Dominant we would expect the mean of the differences to be positive. This is important to be aware of because SPSS **always** takes the first column minus the second column. The test will be the same either way, only the sign on the test statistic will change.

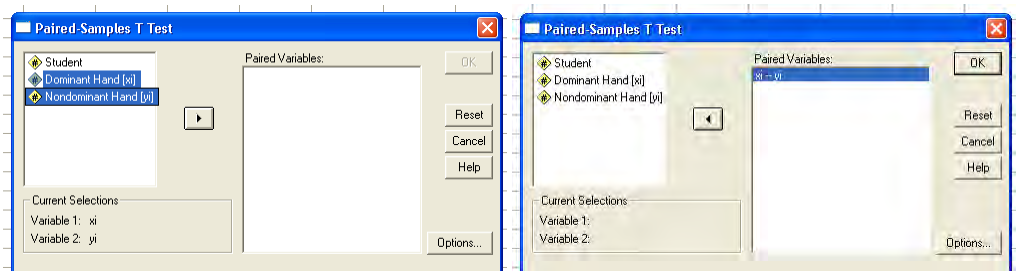
Step 2: α=0.05

Step 3-4: done in SPSS

Go to Compare **Means → Paired-Samples T Test…**



Highlight the variables of interest, and push the ► button. Note that the order in which you select the variables does not matter. Use the shift key when highlighting the variables if they are next to each other, or the Ctrl key if they are not next to each other so that both variables can be highlighted.



Using the Options button will allow you to set the confidence level for a confidence interval on the mean difference.
Click Continue, and OK. The output window will have the following:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Dominant Hand	.17975	12	.017525	.005059
	Nondominant Hand	.19292	12	.017987	.005192

This provides the sample size, sample means, sample standard deviations, and standard errors, which is the sample standard deviation over the square root of the sample size. Note that the label appears as the name, not the variable name. This way, you can keep the variable names simple, and still remember which value is which in the output.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Dominant Hand & Nondominant Hand	12	.572	.052

This provides a correlation coefficient for the Paired samples. SPSS tests this, seeing how the correlation should be fairly high (the p-value small), but it is the method of sampling used that should determine whether or not the paired test should be performed.

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	Dominant Hand - Nondominant Hand	-.01317	.016431	.004743	-.023606	-.002727	-2.776	11	.018

The mean of the differences, standard deviation of the differences, and standard error of the differences are provided here. Here you can calculate the test statistic by hand if you want to compare the by hand methods to SPSS.

The value of t in the seventh box is the **Test Statistic** (calculated in Step 3). The sample mean difference is 2.776 standard errors (standard deviation of the sampling distribution) below the hypothesized mean difference of 0.

The degrees of freedom, in the eighth box, are the degrees of freedom for the test, which is the number of pairs minus one. The Sig. (2-tailed) is the p-value for a two sided test. Since we want a 1-sided test, we divide this in half. (.018/2=.009) so the p-value is .009 which is smaller than α (.009<.05) so we reject the null hypothesis.

Note: The p-value calculated is **always** the tail values for a two-tailed test. Make sure that the p-value corresponds to the hypotheses that you created in step 1. If you are expecting a positive t-value (H_1: $\mu >$ 142.8) and the sample yields a negative t-value (sample mean was less than the hypothesized mean instead of larger), then the p-value would be one minus the p-value calculated above. P(t>Test Statistic)=1-P(t<Test Statistic). Also remember, p-values MUST be POSITIVE!!! Just because the test statistic is negative does not make the p-value negative.

Also provided is the confidence interval, as in example 4. We are 95% confident that the mean difference of reaction time between dominant and non-dominant hands is between -0.023606 and -0.002727, or the average reaction time for the dominant hand is 0.0027 to 0.0236 seconds faster than the reaction time for the non-dominant hand.

Note: The confidence interval does not have to agree with the results of the test, unless you are performing a two tailed test with the same level of confidence. One tailed tests can provide different results, as you are only looking at one tail instead of both.

Step 5: We have sufficient evidence to show that the reaction time for the dominant hand is less than the reaction time for non-dominant hand at the α=.05 level of significance.
We have sufficient evidence to show that the reaction time for the dominant hand is less than the reaction time for non-dominant hand (p-value=.009).

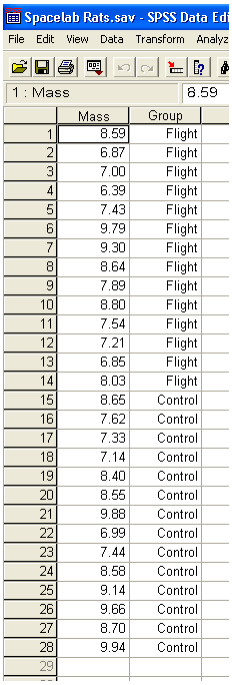
Interpreting the result in step 4 is more important than getting the result. If others cannot understand the results then we have wasted the time in performing the test. Make sure that the statement declares what was measured (not just that the mean has(n't) changed) and the level of significance that was used, or the p-value, so that future researchers can verify the results.

Section 11.2 Inference about Two Means: Independent Samples

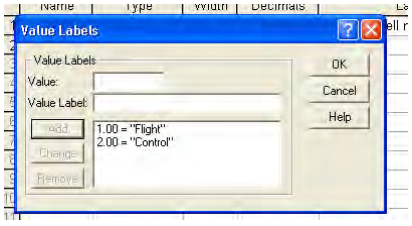
►Inferences about Two Means: Independent Samples

Example 1, Page 590

First we must type in the data. This is a little different from the paired samples, in that we can no longer have two variables to contain the values. Since SPSS is built around each observation being the same person, we must create two variables, one for the value that we want to use, and one for which group the value belongs to. The group should be a number (1 and 2) which can be labeled in the Variable View screen.



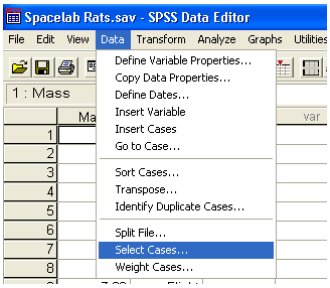
	Mass	Group
1	8.59	Flight
2	6.87	Flight
3	7.00	Flight
4	6.39	Flight
5	7.43	Flight
6	9.79	Flight
7	9.30	Flight
8	8.64	Flight
9	7.89	Flight
10	8.80	Flight
11	7.54	Flight
12	7.21	Flight
13	6.85	Flight
14	8.03	Flight
15	8.65	Control
16	7.62	Control
17	7.33	Control
18	7.14	Control
19	8.40	Control
20	8.55	Control
21	9.88	Control
22	6.99	Control
23	7.44	Control
24	8.58	Control
25	9.14	Control
26	9.66	Control
27	8.70	Control
28	9.94	Control
29		



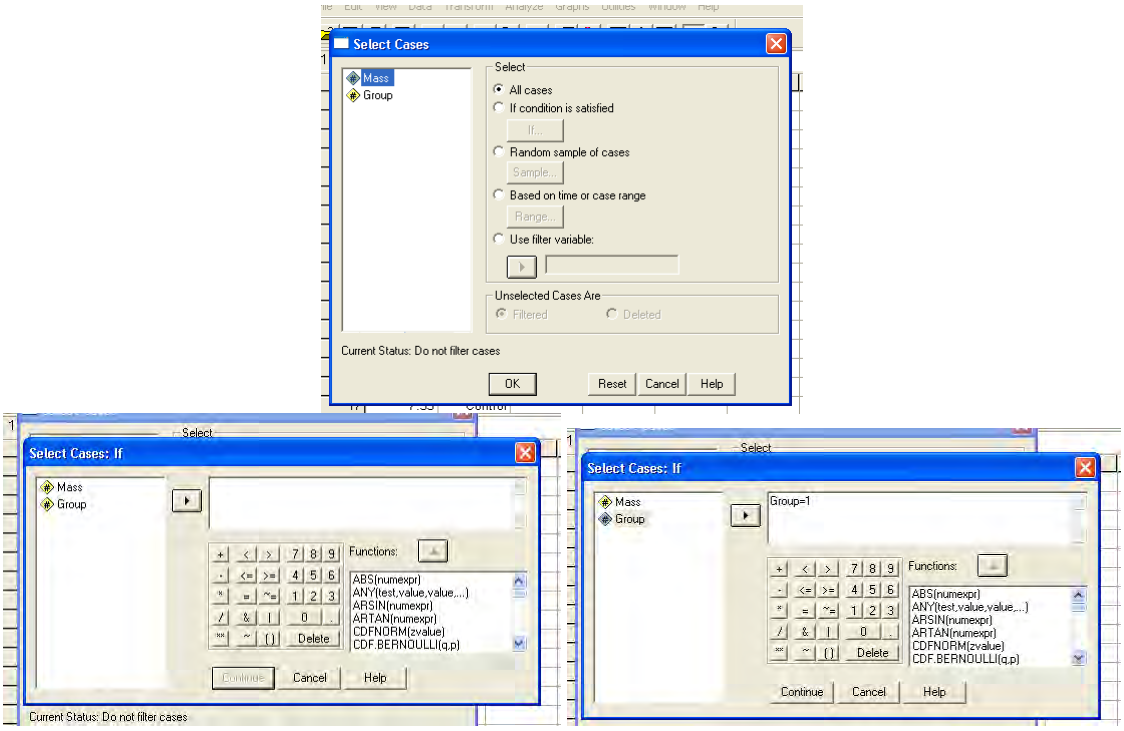
This allows for longer group descriptions as well as keeping SPSS happier.

Normality Check

To create the box-plot and normal probability plot, see previous sections. Each group should be independently normally distributed. In order to create the normal probability plots, we will have to separate out the groups. To do this, go to **Data → Select Cases**. This allows you to look at each group separately.



In the Select Cases menu, select If condition is satisfied, and click on the **If…** button.



You can now define which group, by name or number, you want to look at. You must remember to return to All cases before continuing the analysis.

Step 1: Based on the claim that flight animals have **different** red blood cell mass from the control animals, we use

H_0: $\mu_1-\mu_2 = 0$ -status quo ($\mu_1=\mu_2$, no difference in the means)

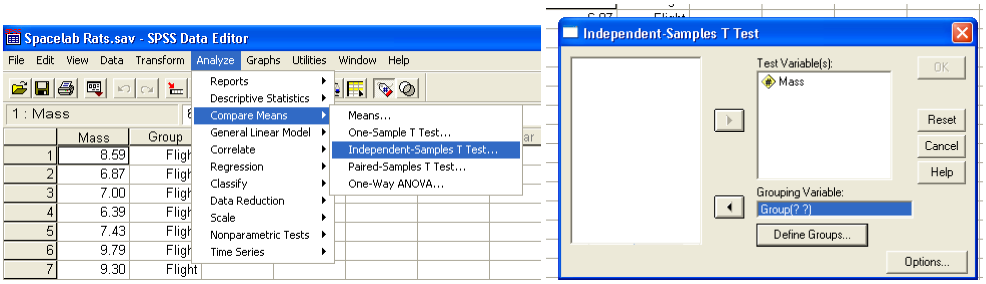
H_1: $\mu_1-\mu_2 \neq 0$ -claim-there is a difference in the means

Note that it doesn't matter which is the first group, and which is the second, as long as you keep in mind which is which and use the appropriate relation (<. >) for the difference. Switching groups will only switch the sign of the test statistic.

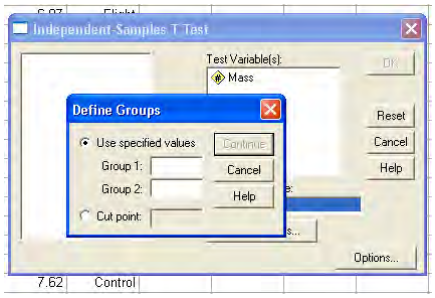
Step 2: α=0.05.

Step 3-4: done in SPSS

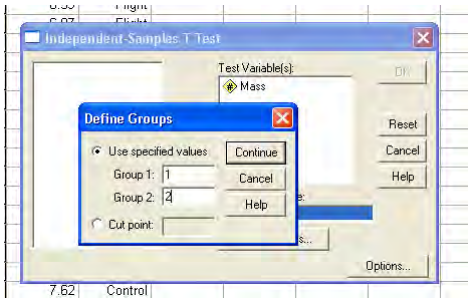
Go to **Compare Means → Independent-Samples T Test…**



Select the column which contains the values, and put that in the Test Variable(s) box. Select the column which contains the group value and put that in the Grouping Variable box. You then have to let SPSS know what groups to use, by clicking on the Define Groups button.



Type in the name or number for the two groups, (this is another reason short numbers are better than long names). SPSS will take the difference to be Group 1-Group 2, so use the group that corresponds to the direction you want the difference to be taken.



Using the Options button will allow you to set the confidence level for a confidence interval on the difference in the means.
Click Continue, and OK. The output window will have the following:

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Mass	Flight	14	7.8807	1.01745	.27193
	Control	14	8.4300	1.00547	.26872

This provides the sample size, sample means, sample standard deviations, and standard errors for each group.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Mass	Equal variances assumed	.023	.882	-1.437	26	.163
	Equal variances not assumed			-1.437	25.996	.163

Independent Samples Test

		t-test for Equality of Means			
		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
				Lower	Upper
Mass	Equal variances assumed	-.54929	.38230	-1.33512	.23655
	Equal variances not assumed	-.54929	.38230	-1.33513	.23655

The difference in the means, standard deviation of the differences, and standard error of the differences are provided here. Here you can calculate the test statistic by hand.

SPSS provides the output of two separate tests, one when equal variances are assumed (the pooled t-test), and one where equal variances are not assumed (known as Welch's t-test). The Levene test for equality of variances is a robust test, which allows you to better select which test you wish to use, and as the tests are both presented, you can see what differences there may be.

We want to use the Welch's t-test, which is the second row. So we have a t-value of -1.437, with 25.996 degrees of freedom, which is found using the formula on page 594, and the p-value for a two-tailed test of 0.163.

Note: The p-value calculated is the tail values for a two-tailed test. Make sure that the p-value corresponds to the hypotheses that you created in step 1. If you are expecting a positive t-value (H_1: $\mu_1 > \mu_2$) and the sample yields a negative t-value (sample mean was less than the hypothesized mean instead of larger), then the p-value would be one minus the p-value calculated above. P(t>Test Statistic)=1-P(t<Test Statistic). Also remember, p-values MUST be POSITIVE!!! Just because the test statistic is negative does not make the p-value negative.

Also provided is the confidence interval. We are 95% confident that the difference in the means between flight and control rats is between -1.3351 and 0.2365, or the average for rats in flight is 1.335 milliliters below to 0.237 milliliters above the average for control rats.

Note: The confidence interval does not have to agree with the results of the test, unless you are performing a two tailed test with the same level of confidence. One tailed tests can provide different results, as you are only looking at one tail instead of both.

Step 5: We do not have sufficient evidence to show that red blood cell mass in flight animals is different from red blood cell mass in control animals at the α=.05 level of significance.
We do not have sufficient evidence to show that red blood cell mass in flight animals is different from red blood cell mass in control animals (p-value=.1627).

Section 11.3 Inference about Two Population Proportions

►Inferences about Two Population Proportions

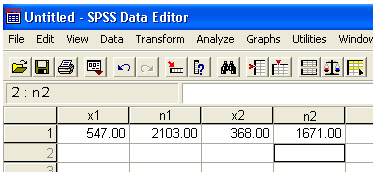
As we have seen before, SPSS does not work directly with proportions. We can use the SPSS calculator, and the SPSS z-table for the test, but there is little advantage to using SPSS for a proportion test over a calculator.

Example 1, Page 604

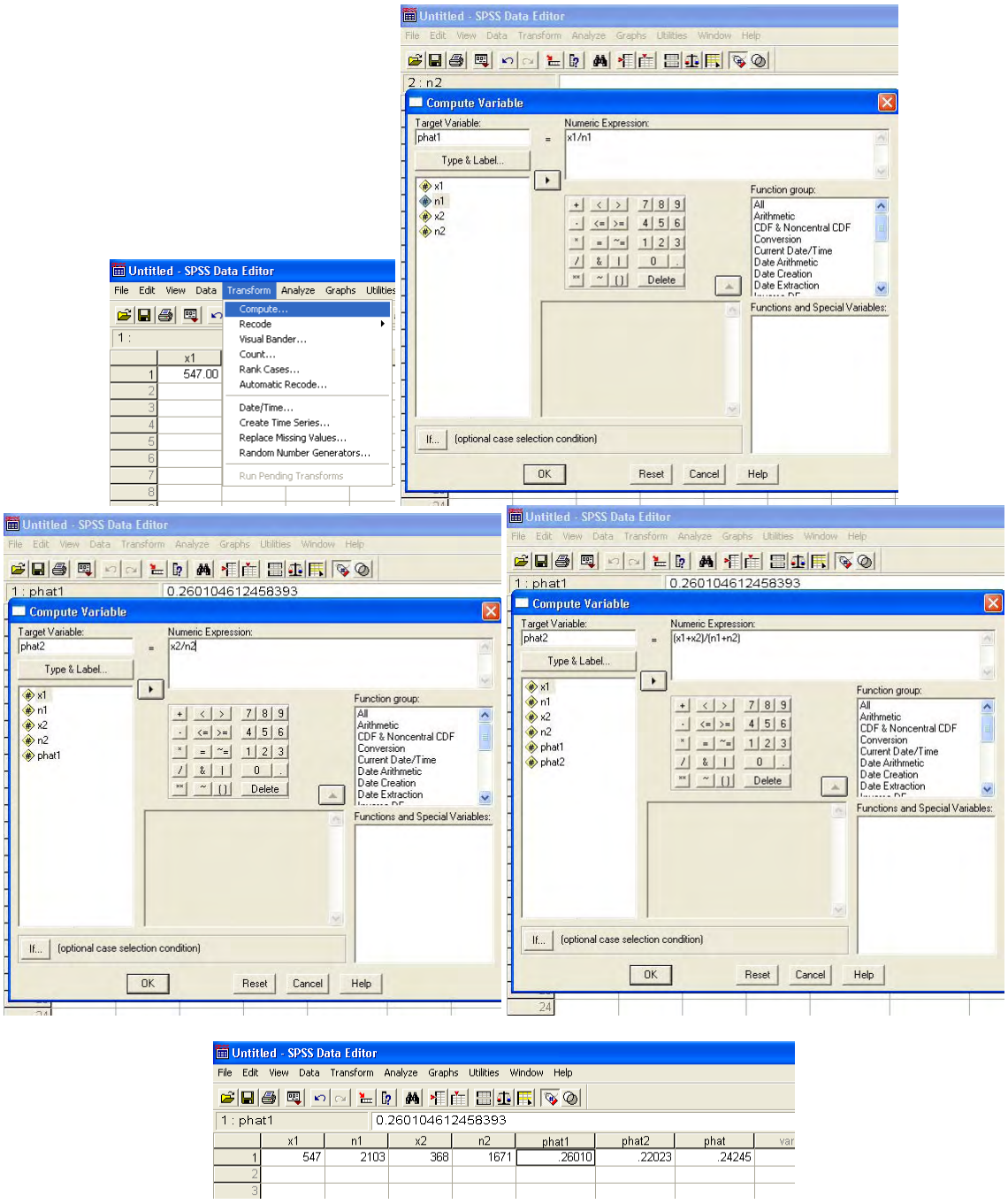
Step 1: The claim is that the proportion of Americans 18 years old or older who believe that men are more aggressive than women is less than 0.74.

H_0: $p_1 = p_2$
H_1: $p_1 > p_2$

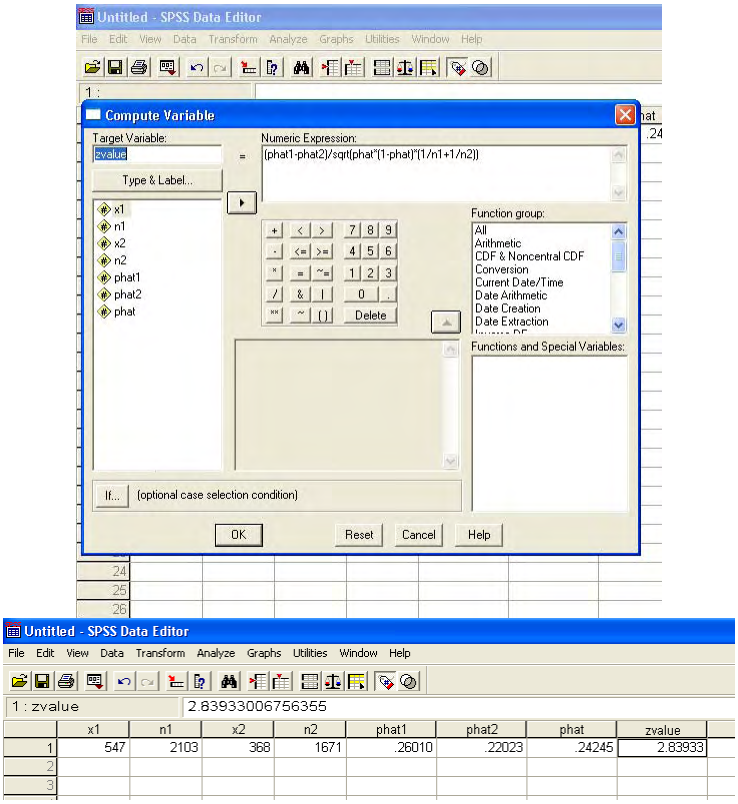
Step 2: We will use p-values instead of critical values.
Step 3: Compute the values of $\hat{p}$. In SPSS, create two variables for the value x, and another two for the values of n.



Go to **Transform → Compute…**, create a new variable, called phat1, by putting the name phat1 in the Target Variable box, and the equation x1/n1 in the Numeric Expression box. Similarly create phat2, and phat, which is the sum of the x values divided by the sum of the n values.

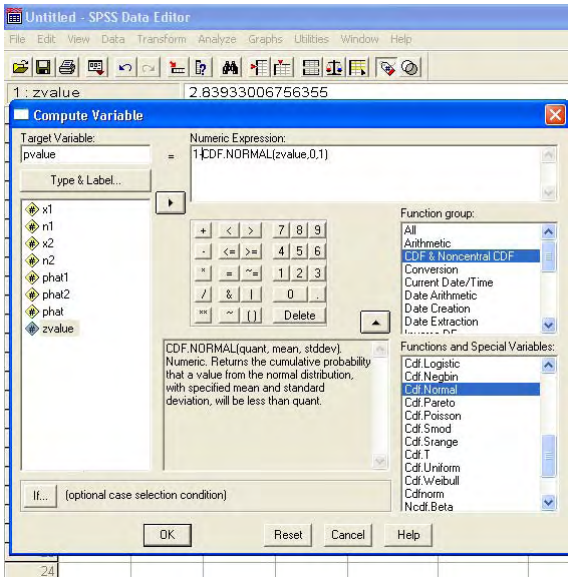


We can check the sample sizes, using phat1 and phat2, similar to chapter 9.
Returning to **Transform → Compute…**, we will calculate the z-value, similar to the test for μ with σ unknown. Type in the Z-equation, z=(phat1-phat2)/sqrt(phat*(1-phat)*(1/n1+1/n2)).

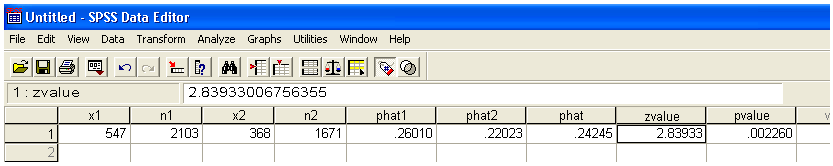


So, in this example, z=2.84.

Step 5: We can then use SPSS to calculate the p-value. To get the one-sided p-value, we will again use **Transform → Compute…**. This is similar to what was done in chapter 7. Clear the z equation in the Compute Variable window, and click on CDF and Noncentral CDF in the Function Group box. In the Functions and Special Variables, select CDF.NORMAL. It must be remembered that this is the area to the left of the z-value.

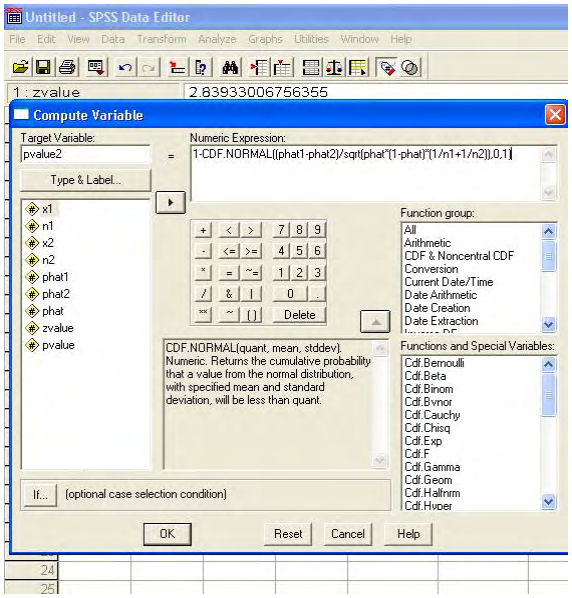


Once the equation is done, push OK. The p-value will be a new variable in the data set. Again, it will appear as many times as you had variables, and you may need to go to Variable View to change the number of decimal places.

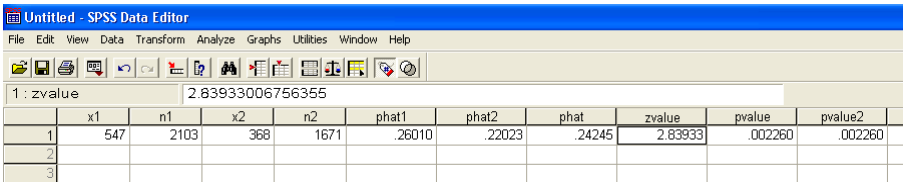


	x1	n1	x2	n2	phat1	phat2	phat	zvalue	pvalue
1	547	2103	368	1671	.26010	.22023	.24245	2.83933	.002260
2									

If you don't need to know the z-value, and just want to calculate the p-value, you can skip the z-value calculation. We can calculate the tail probabilities, similar to chapter 7, directly for the sample mean.



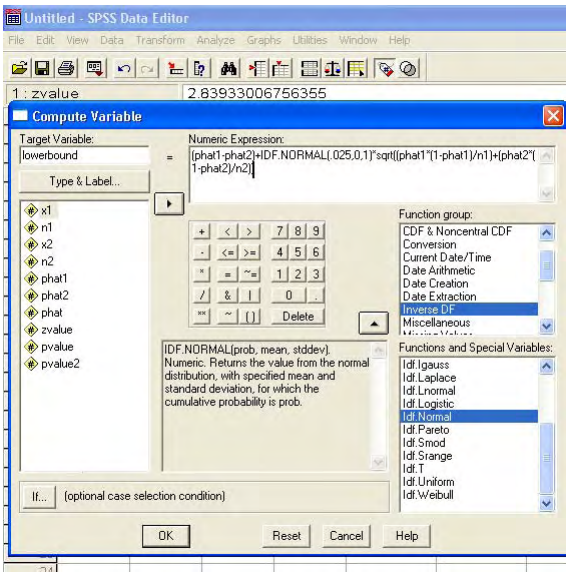
Use the CDF.NORMAL, with the value of the phat as the first value, the hypothesized proportion as the second value, and the standard error equation as the third value. Here you will have to see if the sample proportion is larger or smaller than the hypothesized mean in order to see if you want the CDF or 1-CDF to get the tail values.

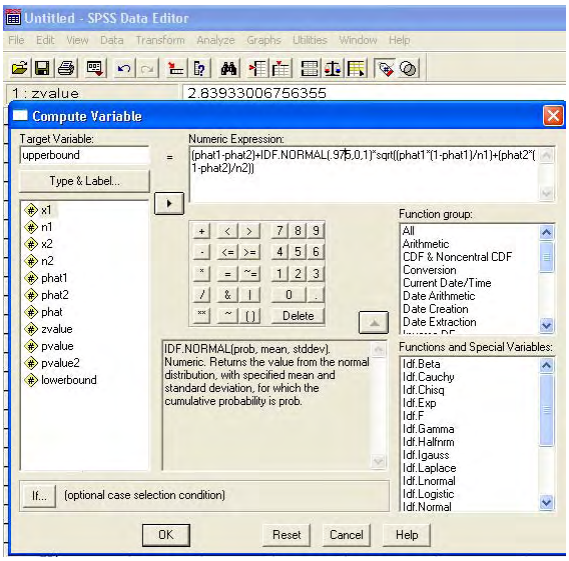


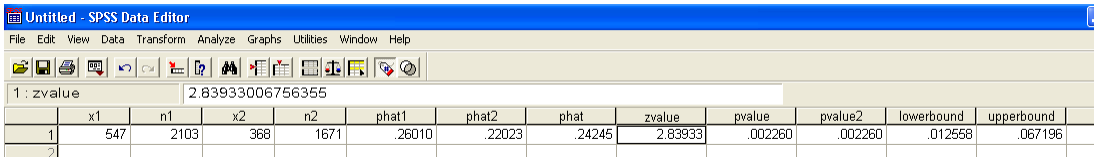
	x1	n1	x2	n2	phat1	phat2	phat	zvalue	pvalue	pvalue2
1	547	2103	368	1671	.26010	.22023	.24245	2.83933	.002260	.002260
2										
3										

The p-value is the same, whether you do the test step-by-step, or whether you calculate the p-value directly. Since the p-value is less than 0.05, we reject the null hypothesis. There is enough evidence to show that the proportion of individuals 12 years and older taking 200 mcg of Nasonex who experience headaches is greater than the proportion of individuals 12 and older taking a placebo who experience headaches.

Confidence intervals can be similarly produced, using the IDF.NORMAL function, as in chapter 7. For a 95% confidence interval we would create a lower bound using the area to the left as .025 and the upper bound using the area to the left as .975.







	x1	n1	x2	n2	phat1	phat2	phat	zvalue	pvalue	pvalue2	lowerbound	upperbound
1	547	2103	368	1671	.26010	.22023	.24245	2.83933	.002260	.002260	.012558	.067196

So we are 95% confident that the proportion of individuals 12 and older taking 200 mcg of Nasonex who claim a headache as a side effect is 1.3% to 6.7% more than individuals 12 and older taking a placebo.

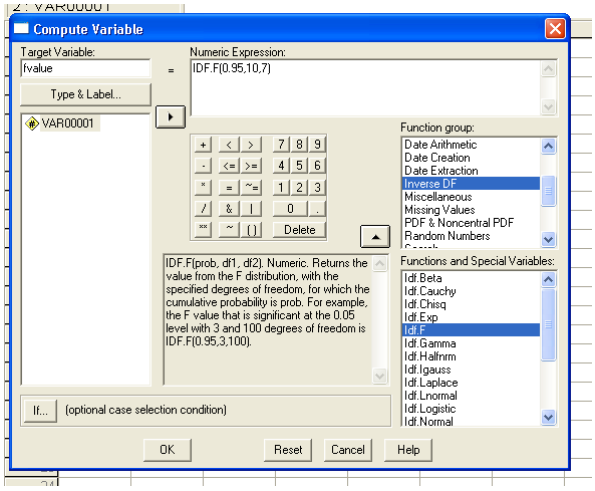
Section 11.4 Inference about Two Population Standard Deviations

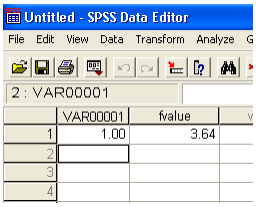
Instead of using the F-test, which is not robust, SPSS uses the Levene test, which is robust. This is calculated at the same time as performing a two-sample t-test for independent samples.

► Finding Critical Values for the F-Distribution

Example 1, page 614

To find critical values under the F-distribution, we turn to the SPSS calculator. Go to **Transform → Compute**. We find the function group Inverse DF and find IDF.F. The F distribution requires three values, which are the probability to the left (which is complement of what you see in the book), and the degrees of freedom for the numerator and denominator. So, for a right-tailed test with $\alpha=0.05$, degrees of freedom in the numerator =10, and degrees of freedom in the denominator =7, we would have 0.95 as the probability to the left, and would use IDF.F(0.95,10,7).



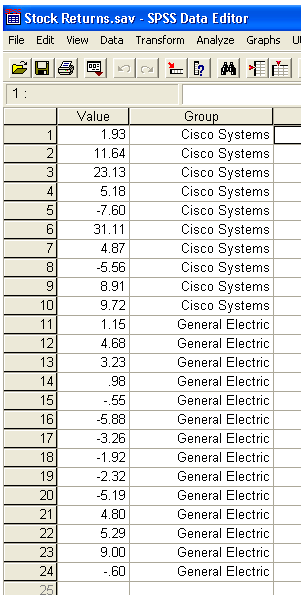


So, for this test, we would have a critical F value of 3.64.

►Inferences about Two Population Standard Deviations

Example 2, page 616

We enter the data as for an independent sample t-test in Section 11.2. It must be remembered that all values are in one column, and which group they belong to in another column.



Step 1: Based on the claim that Cisco systems is more volatile than General Electric stock

H_0: $\sigma_1 = \sigma_2$ -status quo (no difference in the variation)

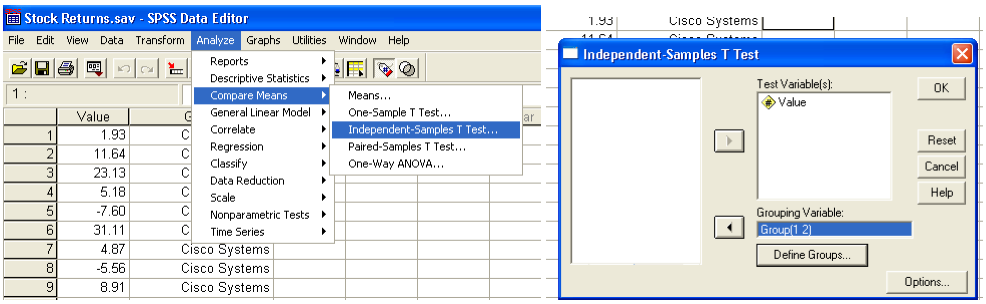
H_1: $\sigma_1 > \sigma_2$ -claim- Cisco systems is more volatile

Note that it doesn't matter which is the first group, and which is the second, as long as you keep in mind which is which and use the appropriate relation (<. >) for the difference.

Step 2: We will use the p-value instead of a critical value.

Step 3-4: done in SPSS

Go to **Compare Means → Independent-Samples T Test…**



Select the column which contains the values, and put that in the Test Variable(s) box. Select the column which contains the group value and put that in the Grouping Variable box. You then have to let SPSS know what groups to use, by clicking on the Define Groups button.

Type in the name or number for the two groups, (this is another reason short numbers are better than long names). SPSS will take the difference to be Group 1-Group 2, so use the group that corresponds to the direction you want the difference to be taken.

Using the Options button will allow you to set the confidence level for a confidence interval on the difference in the means.
Click Continue, and OK. The output window will have the following:

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Value	Cisco Systems	10	8.3330	11.83565	3.74276
	General Electric	14	.6721	4.31622	1.15356

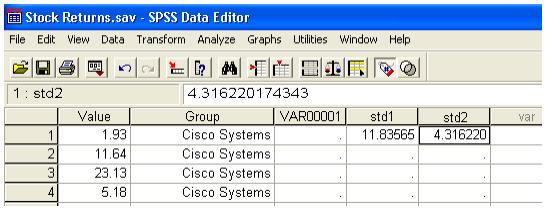
This provides the sample size, sample means, sample standard deviations, and standard errors for each group.

Independent Samples Test

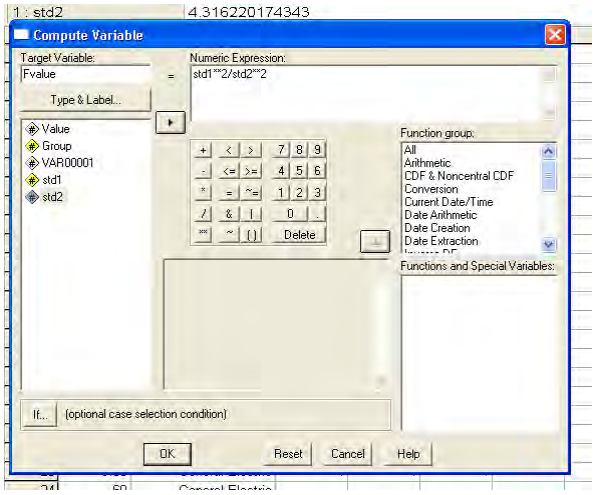
		Levene's Test for Equality of Variances	
		F	Sig.
Value	Equal variances assumed	5.536	.028

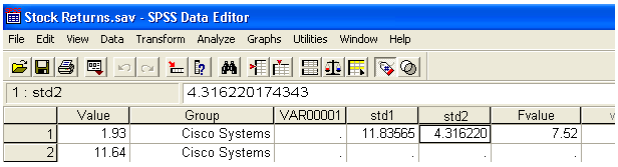
So, with an F-value of 5.536, and a p-value of 0.028, we would reject the null hypothesis. There is enough evidence to show that Cisco systems is more volatile than General Electric stock.

This is a different test than provided in the book. To perform the non-robust F test, we need to do some work by hand. We can copy the standard deviation values from the group statistics and put them into two separate columns.



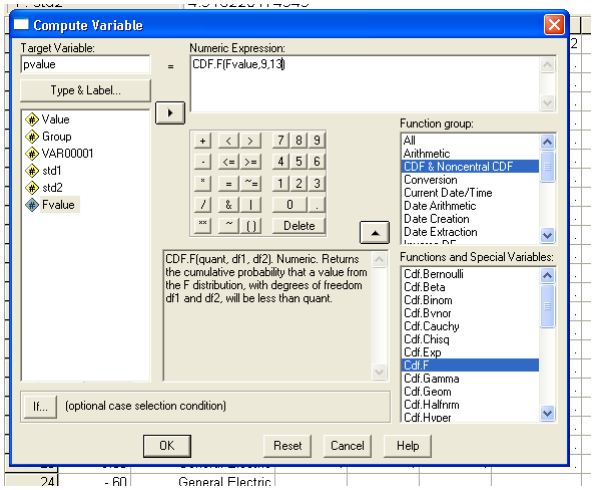
We then turn to **Transform → Compute** to calculate F for us. We can use **2 to square the standard deviations.

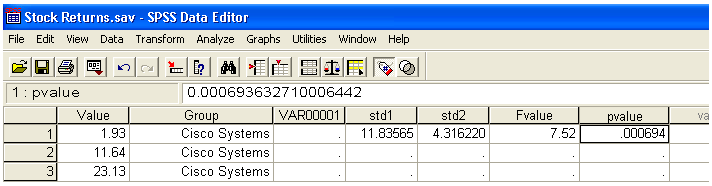




Now we need to calculate the p-value for the test. We again turn to **Transform → Compute**, and use the function group CDF & Noncentral CDF. We find the CDF.F which requires the F-value, and the

numerator and denominator degrees of freedom. Because we want the right tail value, we have to use the complement rule to find the p-vlaue.





	Value	Group	VAR00001	std1	std2	Fvalue	pvalue
1	1.93	Cisco Systems	.	11.83565	4.316220	7.52	.000694
2	11.64	Cisco Systems	.	.	.	.	.
3	23.13	Cisco Systems	.	.	.	.	.

We get a p-value of 0.0007 which is less than 0.05, so we reject the null hypothesis.

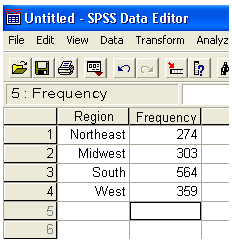
Chapter 12. Inference on Categorical Data

Section 12.1 Goodness of Fit Test

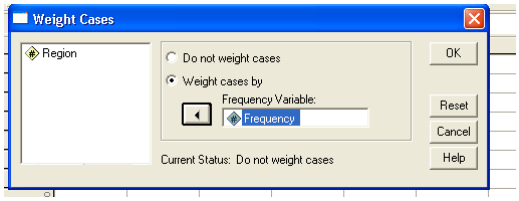
►Goodness-of-Fit

Example 2, Page 634

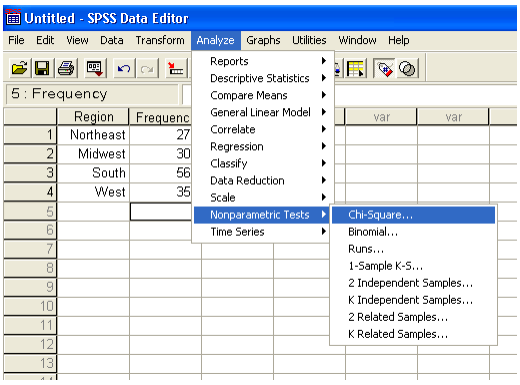
First we must type in the data. It is very important to know the exact order of the data, so it is easiest to type the categories as numbers and use the label options in the Variable View page. We need to create one variable with the category, and one with the observed frequency.



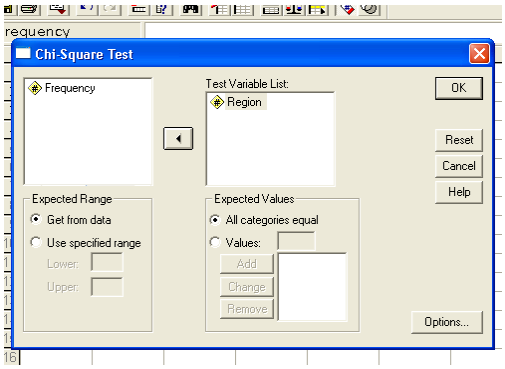
Once the data have been entered, we go to Data → **Weight Cases…** to tell SPSS how many times each category was observed. Weigh the cases by frequency.



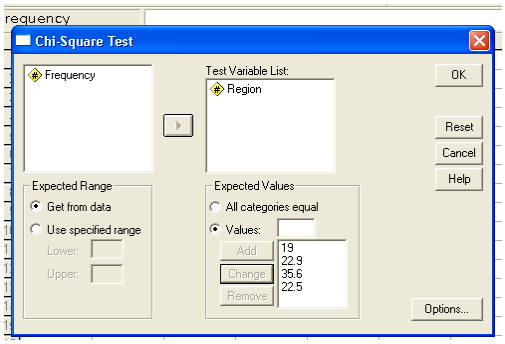
Once SPSS can see the data the way we want it to, go to **Analyze → Nonparametric Tests → Chi-Square…**. Note that this Chi-Square test only works for the goodness-of-fit test. We will use another method for table data.



Once in the Chi-Square test window, select the category as the Test Variable List.



We now need to let SPSS know what the null hypothesis is. This is done in the Expected Values section. If the null hypothesis is that all categories are equal, we can use the default. In this case, the proportions are different, depending on the area. Northeast is 19.0%, Midwest is 22.9%, South is 35.6% and West is 22.5%. This is where we need to know the order. It must be remembered that SPSS likes to reorder things alphabetically, so if you used the actual names, you will have to figure out the alphabetical order. Click on the Values button. In order, from first to last, type in the hypothesized proportion, clicking the Add button between each value.



If you get these out of order, you will get different expected values, and, thus, a different chi-square value.

Once you have the values you want, in the correct order, press OK. In the output window, you will see the following:

Region

	Observed N	Expected N	Residual
Northeast	274	285.0	-11.0
Midwest	303	343.5	-40.5
South	564	534.0	30.0
West	359	337.5	21.5
Total	1500		

This is the summary table. This provides the observed value, expected value, and residual value, which is the difference between the observed and the expected. This is a good way to check yourself and make sure that the order you typed the values in was the order you expected.

Test Statistics

	Region
Chi-Square[a]	8.255
df	3
Asymp. Sig.	.041

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 285.0.

The second table contains the test information. We get a chi-square value of 8.255, with 3 degrees of freedom, and a p-value of .041. Since .041 is less than .05, we reject the null hypothesis. There is sufficient evidence to show that the proportions have changed since 2000. Note that SPSS also provides the information needed to check for Normality. 0% cells have expected frequencies less than 5, so we have less than 20%, and the minimum expected cell frequency is 285 which is greater than 1, so the conditions have been met.

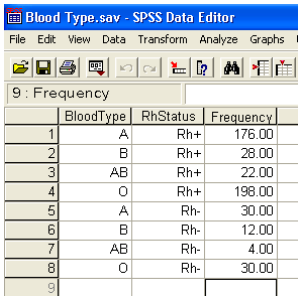
Although this method will provide the correct p-value for a single sample test on proportions (see problems 23 and 24), it will not provide confidence intervals.

Section 12.3 Tests for Independence and Homogeneity of Proportions

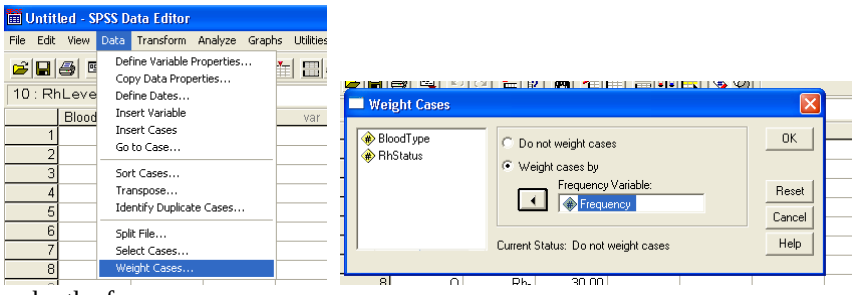
►Contingency Tables and Independence Tests

Example 2, Page 655

Similar to the independent sample t-test in chapter 10, we have to enter the data into SPSS in a very specific manner. We will need to create three variables, one with the value of category 1, one with the value of category 2, and one with the observed frequency.



To make the table look the way you expect, use numbers and labels, otherwise, SPSS will rearrange the table in alphabetical order. Once the data have been entered, go to **Data → Weight Cases**.



Weigh the cases by the frequency.

Step 1: H_0: Blood Type and Rh-status are Independent
OR H_0: Proportion of Rh^+ and A is $P(Rh^+ \text{ and } A)=P(Rh^+)*P(A)$; Proportion of Rh^+ and B is $P(Rh^+ \text{ and } B)=P(Rh^+)P(B)$; etc.

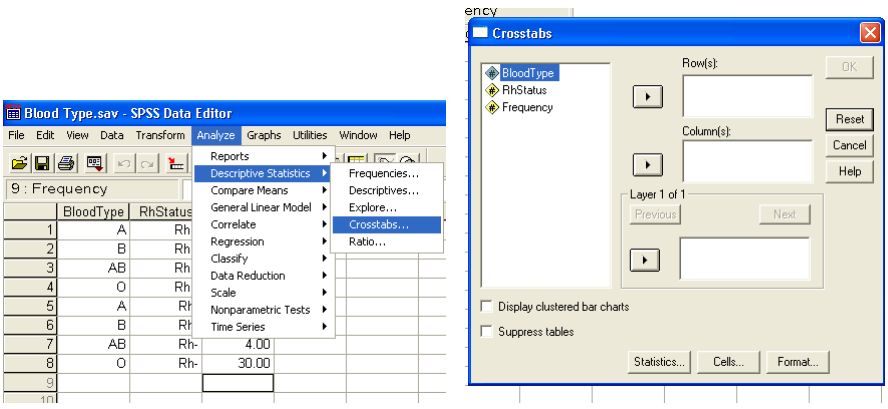
Note: although we write the hypothesis in words, the test is based on the probability definition of independence. This comes in how we calculate the expected values. Unlike the Goodness of Fit test, we don't need to worry about writing these out, as the computer can calculate them for us.
The alternate hypothesis would be:

H_1: Blood Type and Rh-status are Dependent
OR H_1: at least one proportion is different.

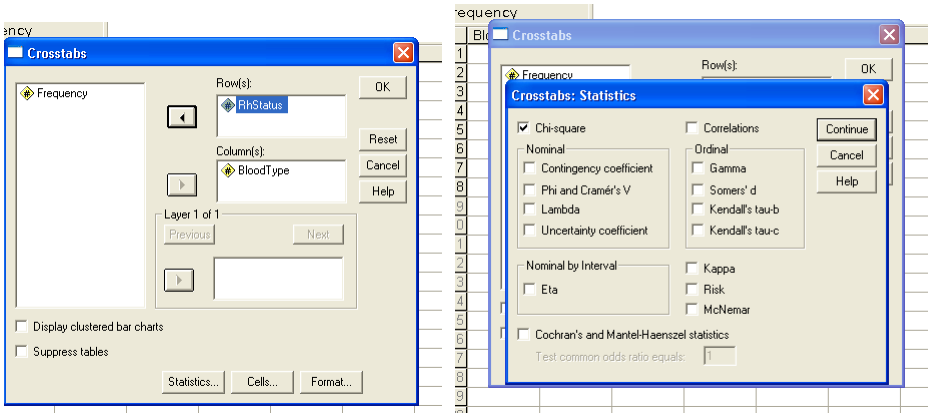
Step 4: we will use the p-value instead of a critical value.

Step 2, 3, 5: done in SPSS

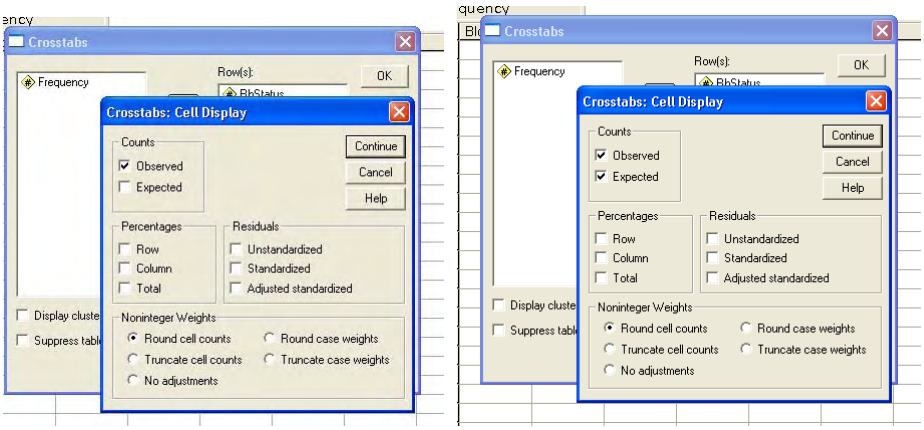
Go to **Analyze → Descriptive Statistics → Crosstabs…**



Place the Row variable in the Row(s) box, and the Column variable in the Column(s) box. Then click on the Statistics box.



Check the Chi-square box and push Continue.



Pushing the Cells button will bring up the Cell Display window. If you want SPSS to calculate the Expected Values for you, check the Expected box. SPSS can also give the proportions for each cell, as well as the residuals. Residuals are useful for checking which cell contributes the most to a rejected null hypothesis, and the standardized is the easiest to interpret for this.
Push Continue, and OK

The output will provide the following tables:

RhStatus * BloodType Crosstabulation

			BloodType				
			A	B	AB	O	Total
RhStatus	Rh+	Count	176	28	22	198	424
		Expected Count	174.7	33.9	22.0	193.3	424.0
	Rh-	Count	30	12	4	30	76
		Expected Count	31.3	6.1	4.0	34.7	76.0
Total		Count	206	40	26	228	500
		Expected Count	206.0	40.0	26.0	228.0	500.0

Here we can see the marginal frequencies, as well as the observed and expected frequencies for each cell.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.601[a]	3	.055
Likelihood Ratio	6.410	3	.093
Linear-by-Linear Association	.494	1	.482
N of Valid Cases	500		

a. 1 cells (12.5%) have expected count less than 5. The minimum expected count is 3.95.

The test statistic (7.601) is provided in the next box, as well as the check for normality, or Step 3, SPSS will verify that no more than 20% of the cells are less than 5 and that all the expected values are greater than or equal to 1. If either of these is not met we should not perform the test.

Step 6. Because the p-value (.055) is not less than α (.05), we fail to reject the null hypothesis.
Step 7. There is insufficient evidence at the α=.05 level of significance to show that blood type and Rh-status are dependent.

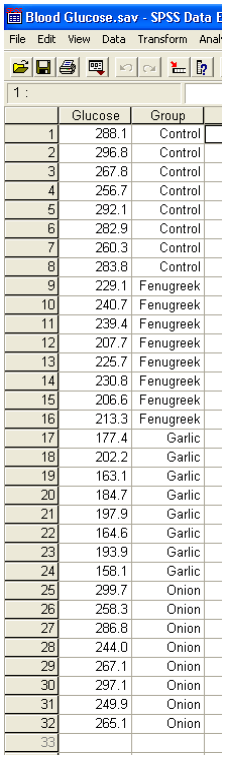
It is difficult based on a global test, such as the Chi-Square test, or ANOVA test, to show where the differences really are once a difference has been found. In the case of ANOVA, post-hoc tests can be run to see where the differences are. In the case of Chi-Square, the cell with the largest individual chi-square value is seen to be the one contributing the most to the rejection, and is thus the one that causes the difference. All other differences may not be significant, and should be tested separately. To find the largest contributor to the Chi-square value, find the cell with the largest absolute standardized residual.

Chapter 13. Comparing Three or More Means

Section 13.1 Comparing Three or More Means (One-way Analysis of Variance)

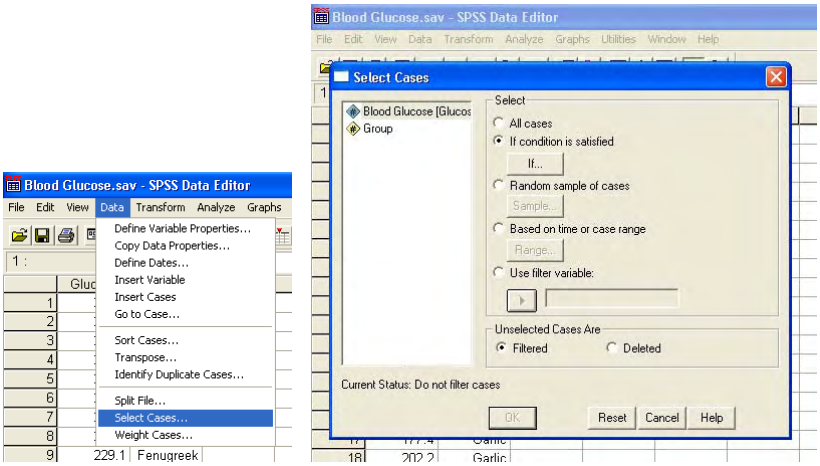
►One-Way Analysis of Variance

Example 1, Page 679
First, we must type in the data. This is done similar to the independent sample t-test in chapter 11, with the value of interest in one column, and the group in another. There can now be more than two groups. The groups must be numerical, but you can use labels, which will help in the output.

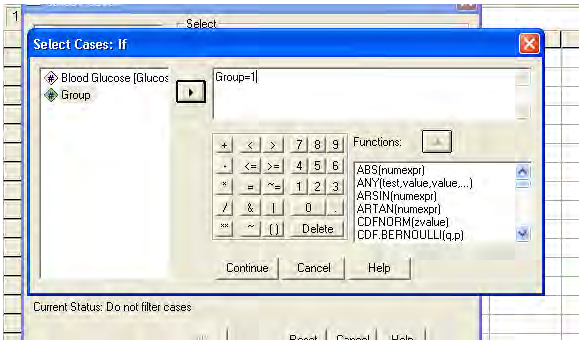


	Glucose	Group
1	288.1	Control
2	296.8	Control
3	267.8	Control
4	256.7	Control
5	292.1	Control
6	282.9	Control
7	260.3	Control
8	283.8	Control
9	229.1	Fenugreek
10	240.7	Fenugreek
11	239.4	Fenugreek
12	207.7	Fenugreek
13	225.7	Fenugreek
14	230.8	Fenugreek
15	206.6	Fenugreek
16	213.3	Fenugreek
17	177.4	Garlic
18	202.2	Garlic
19	163.1	Garlic
20	184.7	Garlic
21	197.9	Garlic
22	164.6	Garlic
23	193.9	Garlic
24	158.1	Garlic
25	299.7	Onion
26	258.3	Onion
27	286.8	Onion
28	244.0	Onion
29	267.1	Onion
30	297.1	Onion
31	249.9	Onion
32	265.1	Onion
33		

The normality must be tested on each group separately. To do this, you can use the **Data → Select Cases** option.



In the Select Cases window, you can select If condition is satisfied, click on the **if** button, and create a condition that separates out a specific group to test the normality on, for example, Group=1 for the first group.



Once you have selected the group, you can test for normality the same as in earlier chapters. After checking the normality of each group, you must remember to go back and select all cases before going to the Analysis of Variance.

The claim is that the mean blood glucose levels are the same for all diets.

H_0: $\mu_{Control} = \mu_{Fenugreek} = \mu_{Garlic} = \mu_{Onion}$

OR H_0: mean blood glucose level is independent of diet

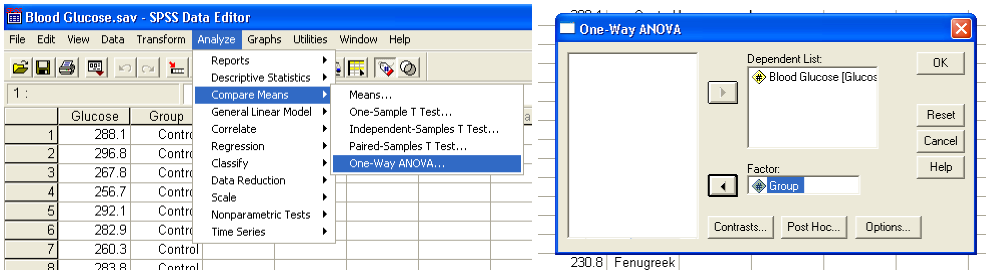
The second statement shows that the test is similar to the chi-square tests, but because we are looking at means, we no longer have the restrictions that everything adds up to 100%. Note: if there were only two means, this would be the same as the independent samples t-test hypothesis from chapter 11.

The alternate hypothesis would be:

H_1: at least one mean is different

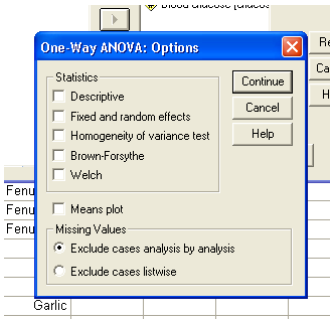
We set $\alpha=0.05$.

Go to **Analyze → Compare Means → One-Way ANOVA…**



Use the column with the values for the Dependent list, and the variable containing the groups as the factor. This should remind you of the method used in chapter 11, for the independent t-test, except now we do not have to restrict to two groups, so we don't need to define which two groups to use. Also, since this is a two tailed test, the order in which the groups are taken doesn't matter, as it does in the independent samples t-test.

Click on the Options box



Checking Descriptive will provide you with descriptive statistics for each group. The Fixed and random effects will provide the different results based on whether the effects (groups) are fixed or random. This is for a design of experiments class, and is beyond the scope of this class. The Homogeneity of Variances test will provide a check to see that the assumption of equal variances is met. This is the Levene's test, as used in chapter 11. The Brown-Forsythe, and Welch tests are better tests than the F test if the assumption of equal variances doesn't hold. They also tests the equality of means. The Means Plot provides a graphical way to look at the means to see where a difference occurs, if one exists.

Once you have the options you want to use, click continue.

Clicking on the Contrasts button will open up a contrasts option window. This is also for a design of experiments class, and is beyond what we will discuss here.

Clicking the Post-Hoc Test button provides the possible Post-Hoc (after this) tests. These are tests to show, if a difference was found, where the differences are. We will discuss these later.

Once you have selected the options you want to try, push continue and OK.

The output (using only default values) will provide the following table.

ANOVA

Blood Glucose

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	50090.691	3	16696.897	58.209	.000
Within Groups	8031.616	28	286.843		
Total	58122.307	31			

Here we see a similar table to that provided by MINITAB on page 681.

Adding the descriptive statistics provides the following table

Descriptives

Blood Glucose

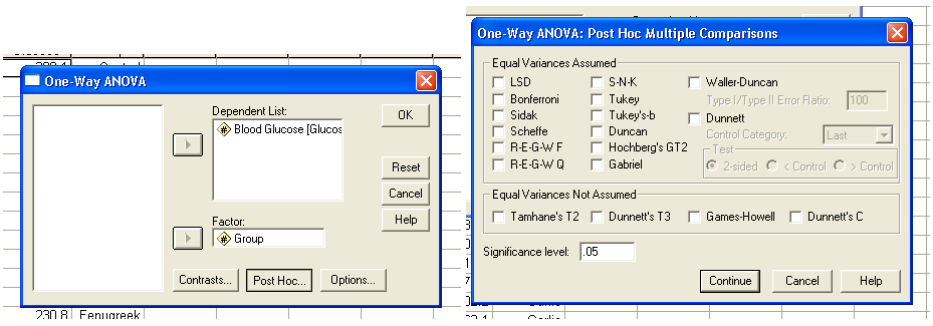
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Control	8	278.563	15.0257	5.3124	266.001	291.124	256.7	296.8
Fenugreek	8	224.163	13.4903	4.7695	212.884	235.441	206.6	240.7
Garlic	8	180.238	17.0597	6.0315	165.975	194.500	158.1	202.2
Onion	8	271.000	21.1797	7.4882	253.293	288.707	244.0	299.7
Total	32	238.491	43.3003	7.6545	222.879	254.102	158.1	299.7

Based on the fact that we reject the null hypothesis, so we know that at least two means are different, we can use this table to find where a difference may be. In this example, the most extreme means are the Control, with a mean of 278.56 and the Garlic, with a mean of 180.24. Because we reject the null hypothesis, we can show that these two are different. We cannot, however, look at the other values without some further analysis.

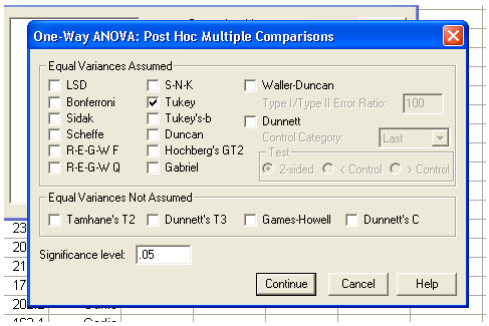
Section 13.2 Post-Hoc Tests on One-Way Analysis of Variance

►Performing Post-Hoc Tests

We use post-hoc tests after finding a difference in the means to show us where those means might be. When running the one-way ANOVA test, you can select post-hoc tests by clicking on the post-hoc button.



This provides a list of different post-hoc tests. Each test has its own way of conserving the α value that we have selected, but some are more conservative (will find less differences) than others. You can experiment with the different tests to see the differences, but we will present the results of the Tukey test.



We select the Tukey test, and the Significance level (α). This will add the following tables to the output.

Multiple Comparisons

Dependent Variable: Blood Glucose

Tukey HSD

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Control	Fenugreek	54.4000*	8.4682	.000	31.279	77.521
	Garlic	98.3250*	8.4682	.000	75.204	121.446
	Onion	7.5625	8.4682	.808	-15.558	30.683
Fenugreek	Control	-54.4000*	8.4682	.000	-77.521	-31.279
	Garlic	43.9250*	8.4682	.000	20.804	67.046
	Onion	-46.8375*	8.4682	.000	-69.958	-23.717
Garlic	Control	-98.3250*	8.4682	.000	-121.446	-75.204
	Fenugreek	-43.9250*	8.4682	.000	-67.046	-20.804
	Onion	-90.7625*	8.4682	.000	-113.883	-67.642
Onion	Control	-7.5625	8.4682	.808	-30.683	15.558
	Fenugreek	46.8375*	8.4682	.000	23.717	69.958
	Garlic	90.7625*	8.4682	.000	67.642	113.883

*. The mean difference is significant at the .05 level.

Blood Glucose

Tukey HSD[a]

Group	N	Subset for alpha = .05		
		1	2	3
Garlic	8	180.238		
Fenugreek	8		224.163	
Onion	8			271.000
Control	8			278.563
Sig.		1.000	1.000	.808

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 8.000.

The first table provides the actual difference in the means, the p-value (Sig.) for the Tukey test, and a confidence interval for the difference. Each mean that is significantly different at α is marked by an *. We can see in this example that Control is significantly different from Fenugreek and Garlic, but not significantly different from Onion. We can also use the confidence intervals to show that Control is 31.279 to 77.521 mg/dl higher than Fenugreek. This is because when we look at the row where Control comes first, the difference is positive, and the upper and lower bounds of the confidence interval are both positive. If we look at the row with Fenugreek in the first column, then when we look at control, the mean difference is negative, as are the upper and lower bounds of the confidence interval. That tells us that Fenugreek had a lower mean than Control. The standard error column provides the value for

$$\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

which is similar to what we calculate by hand. SPSS uses a table that incorporates the constant $\sqrt{2}$ in the denominator.

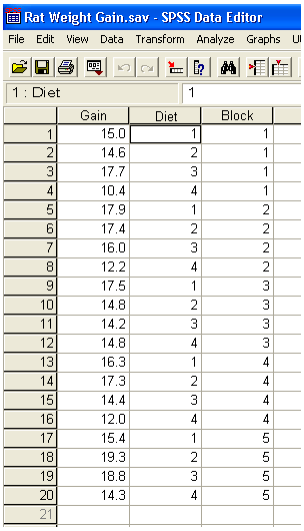
Although the first table provides the most information, for most students, the second table is the easier table to follow. Here we place the diets into separate groups, where we can see a difference between groups but not within groups. Garlic is in its own group, and has the lowest mean. This means that Onion is significantly lower than the other diets. Fenugreek is also different than the other diets. It has a higher mean than Garlic, but a lower mean than Onion or Control. Since Onion and Control are in the same group, we cannot show a significant difference between them.

Section 13.3 The Randomized Complete Block Design

►Performing an ANOVA for random block designs

Example 2 page 704

For a two-way ANOVA, we need to enter the data similarly to a one-way ANOVA, but we will have an extra column. So, we need one column for the response variable, one column for the first treatment, and a second column for the second treatment, or the block.

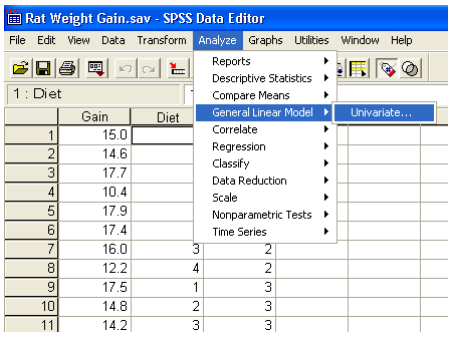

Rat Weight Gain.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs U

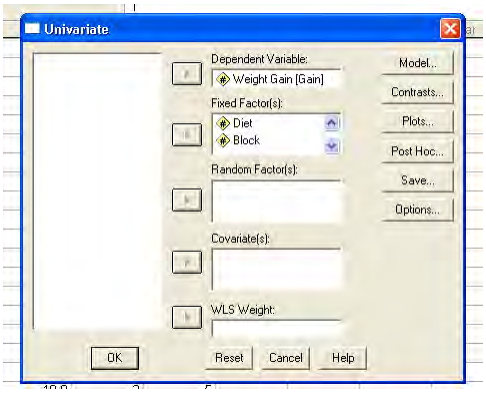
1 : Diet 1

	Gain	Diet	Block
1	15.0	1	1
2	14.6	2	1
3	17.7	3	1
4	10.4	4	1
5	17.9	1	2
6	17.4	2	2
7	16.0	3	2
8	12.2	4	2
9	17.5	1	3
10	14.8	2	3
11	14.2	3	3
12	14.8	4	3
13	16.3	1	4
14	17.3	2	4
15	14.4	3	4
16	12.0	4	4
17	15.4	1	5
18	19.3	2	5
19	18.8	3	5
20	14.3	4	5
21			

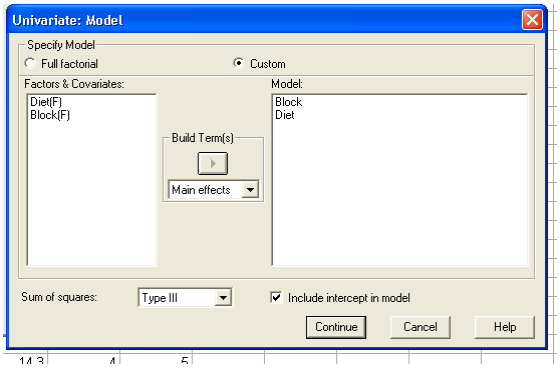
To run the two-way ANOVA, we go to **Analyze → General Linear Model → Univariate**.



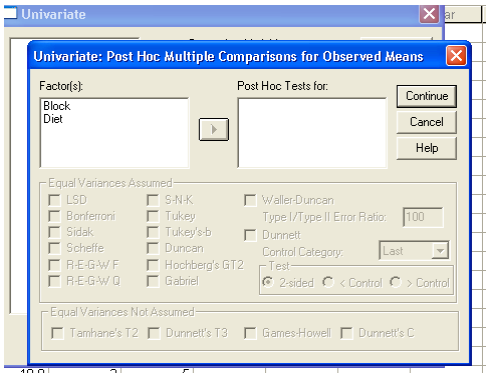
The General Linear Model can be used for more complicated models. In this example, we are going to treat both treatments as fixed.



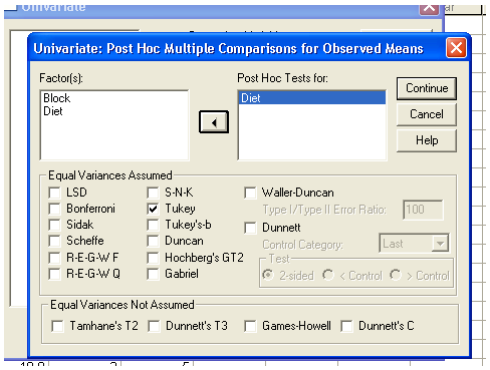
We put the response variable into the Dependent Variable box. We place both the treatment of interest, and the blocking variable into the Fixed Factor(s) box. We then have to define what the model looks like. To do this, click on the **Model** button. We are not performing a full factorial design, so we need to make a custom model.



Select each factor and put them in the model, and select only Main effects under the Build Term(s). This will provide the ANOVA that we want. You can either keep the intercept in the model, or take it out. Neither method affects the portion of the ANOVA table that we want to look at.
Back in the main window, we can also add a post-hoc test, by selecting the post-hoc button.



We are not really interested in differences among the blocks, so we will select Diet. Once we have selected a variable, the different post-hoc tests become available. We will use the Tukey.



The results of the test should look similar to the following:

Tests of Between-Subjects Effects

Dependent Variable: Weight Gain

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	66.583[a]	7	9.512	3.117	.041
Intercept	4814.305	1	4814.305	1577.469	.000
Block	14.713	4	3.678	1.205	.359
Diet	51.870	3	17.290	5.665	.012
Error	36.623	12	3.052		
Total	4917.510	20			
Corrected Total	103.206	19			

a. R Squared = .645 (Adjusted R Squared = .438)

This is the ANOVA with the intercept included. We can ignore the first two rows. The rows with Block, Diet, Error, and Corrected Total are all we actually require, and are the same as provided by MINITAB on page 705. If you opted to not have the Intercept, the output will look like

Tests of Between-Subjects Effects

Dependent Variable: Weight Gain

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	4880.887[a]	8	610.111	199.911	.000
Block	14.713	4	3.678	1.205	.359
Diet	51.870	3	17.290	5.665	.012
Error	36.623	12	3.052		
Total	4917.510	20			

a. R Squared = .993 (Adjusted R Squared = .988)

The output is the same, except the values in the intercept line have now been added to the Model line. We will still ignore the Model and Total rows.

We are not interested in differences in the blocking variable, so we look only at the diet. Diet has a p-value of 0.012, which is less than 0.05, so we reject the null hypothesis. This means we can show a difference in the means of the diets, so we should look at the tukey test to see where those differences are. The output for the Tukey test are similar to those in the previous section.

Multiple Comparisons

Dependent Variable: Weight Gain

Tukey HSD

(I) Diet	(J) Diet	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.260	1.1049	.995	-3.540	3.020
	3	.200	1.1049	.998	-3.080	3.480
	4	3.680*	1.1049	.027	.400	6.960
2	1	.260	1.1049	.995	-3.020	3.540
	3	.460	1.1049	.975	-2.820	3.740
	4	3.940*	1.1049	.018	.660	7.220
3	1	-.200	1.1049	.998	-3.480	3.080
	2	-.460	1.1049	.975	-3.740	2.820
	4	3.480*	1.1049	.037	.200	6.760
4	1	-3.680*	1.1049	.027	-6.960	-.400
	2	-3.940*	1.1049	.018	-7.220	-.660
	3	-3.480*	1.1049	.037	-6.760	-.200

Based on observed means.

*. The mean difference is significant at the .05 level.

Weight Gain

Tukey HSD[a,b]

Diet	N	Subset 1	Subset 2
4	5	12.740	
3	5		16.220
1	5		16.420
2	5		16.680
Sig.		1.000	.975

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 3.052.

a. Uses Harmonic Mean Sample Size = 5.000.

b. Alpha = .05.

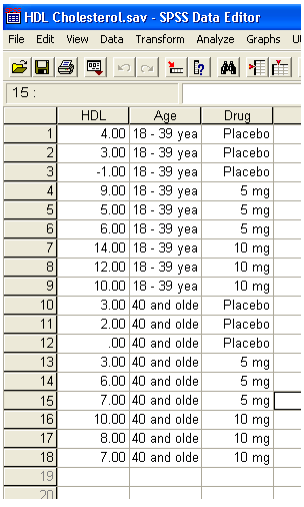
We can show that Diet 4 is different from the other three diets, but we cannot show a difference in the other three diets at all. Diet 4 results in the least amount of weight gain.

Section 13.4 Two-way Analysis of Variance

►Performing a two-way ANOVA

Example 3 page 716

Performing a two-way ANOVA is similar to the random block design, but now we are interested in both treatments. We put the data into SPSS like we do for the random block design.

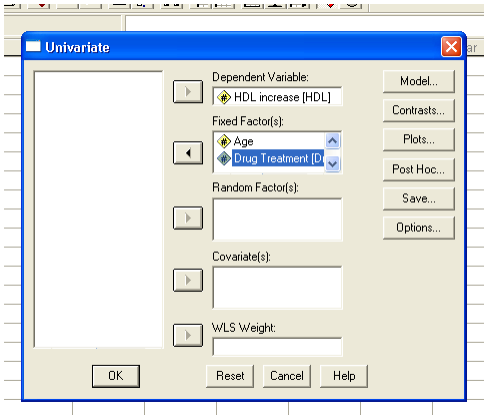

HDL Cholesterol.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs

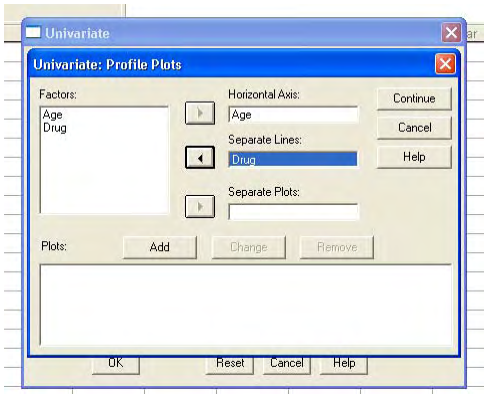
15 :

	HDL	Age	Drug
1	4.00	18 - 39 yea	Placebo
2	3.00	18 - 39 yea	Placebo
3	-1.00	18 - 39 yea	Placebo
4	9.00	18 - 39 yea	5 mg
5	5.00	18 - 39 yea	5 mg
6	6.00	18 - 39 yea	5 mg
7	14.00	18 - 39 yea	10 mg
8	12.00	18 - 39 yea	10 mg
9	10.00	18 - 39 yea	10 mg
10	3.00	40 and olde	Placebo
11	2.00	40 and olde	Placebo
12	.00	40 and olde	Placebo
13	3.00	40 and olde	5 mg
14	6.00	40 and olde	5 mg
15	7.00	40 and olde	5 mg
16	10.00	40 and olde	10 mg
17	8.00	40 and olde	10 mg
18	7.00	40 and olde	10 mg
19			
20			

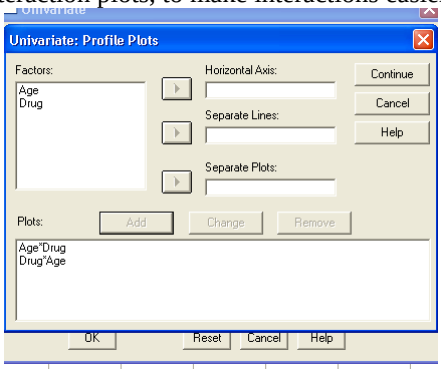
Once the data have been entered, we go to **Analyze → General Linear Model → Univariate**. We enter the variables similarly to the random block design.



Now comes the difference between the random block design and the two-way ANOVA. When we go to the Model menu, we want the full factorial. We do not have to use a custom design. The full factorial design will include both variables as well as the interaction between the two variables. We may want an interaction plot, so we will click on the Plots menu.



We will most likely want both interaction plots, to make interactions easier to see, if an interaction exists.



We can then continue to the output.

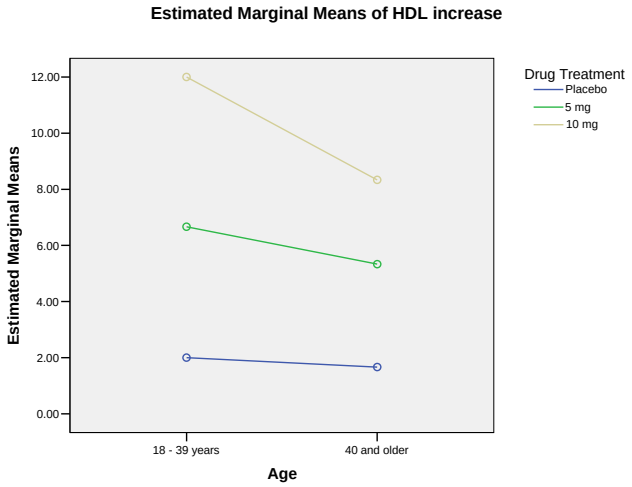
Tests of Between-Subjects Effects

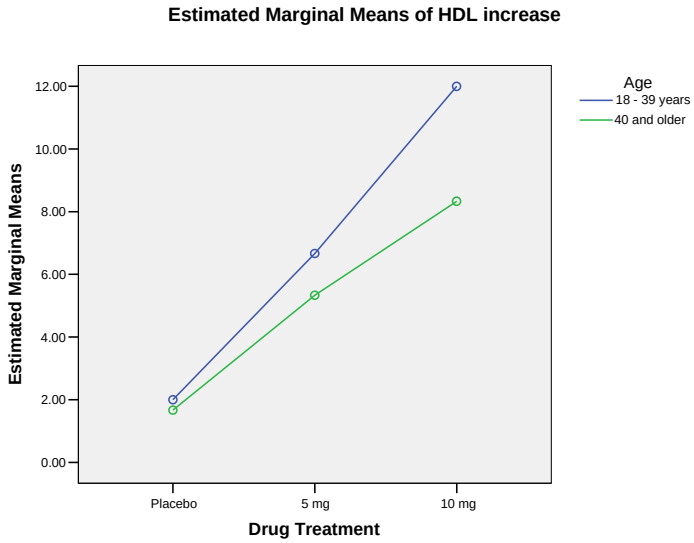
Dependent Variable: HDL increase

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	231.333[a]	5	46.267	11.408	.000
Intercept	648.000	1	648.000	159.781	.000
Age	14.222	1	14.222	3.507	.086
Drug	208.333	2	104.167	25.685	.000
Age * Drug	8.778	2	4.389	1.082	.370
Error	48.667	12	4.056		
Total	928.000	18			
Corrected Total	280.000	17			

a. R Squared = .826 (Adjusted R Squared = .754)

Again, we can ignore the Corrected Model, Intercept, and Total rows. (If you didn't include the intercept, ignore the Model and Total rows). We do not see a significant interaction in this model, but there is a significant difference in the means for the different drugs. We could do a post-hoc test to see where the differences were. We can still look at the interaction plots, although they aren't significant.





If there was an interaction, we would see a crossing, or non-parallelism in the lines that were outside the error of the model. We cannot actually interpret these plots, because there was no significant interaction.

Chapter 14. Inference on the Least-Squares Regression Model and Multiple Regression

Section 14.1 Testing the Significance of the Least-Squares Regression Model

►Testing the Least-Squares Regression Model

Example 1, Page 737

Much of what we will do for regression in chapter 12 was already done in chapter 4. The only difference is that we will now look at the tests as well as the coefficients. For this test, the data is entered in two columns, one with the predictor variable and the next with the response variable. Both must be numerical.

The claim is that the distribution of residents is the same today as it was in 1999. The null hypothesis is:

H_0: There is no linear relation between Age and Total Cholesterol

OR H_0: $\rho=0$

OR H_0: $\beta_1=0$

All three hypotheses mean the same thing. The first is the hypothesis in words, the second using correlation, and the third using the slope. The second and third are useful in calculation, while the first is useful in interpretation.

The alternate hypothesis would be:

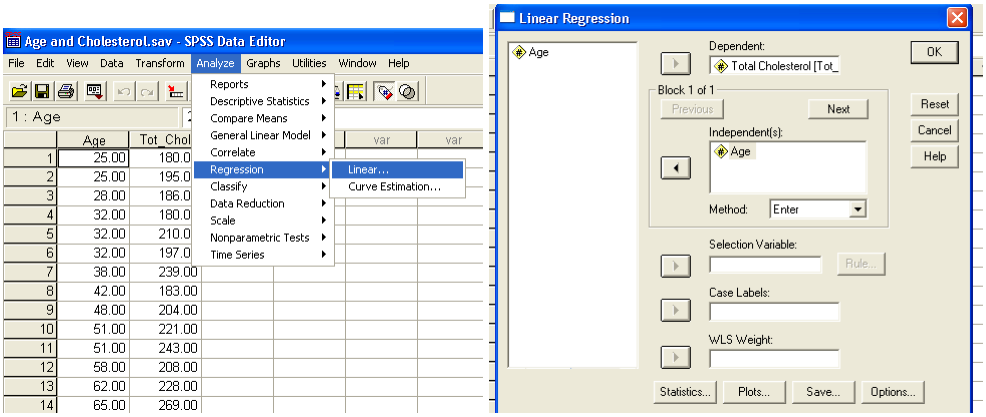
H_1: There is a linear relation between Age and Total Cholesterol

The alternate hypotheses can be one or two tailed by specifying a positive or negative linear relation. For this example we will use

H_1: $\rho\neq0$

OR H_1: $\beta_1\neq0$ or the two-tailed test.

Go to **Analyze → Regression → Linear…**



Use the Dependent, or Y, or response variable in the Dependent box. The Independent, or X, or predictor variable goes in the Independent(s) box. When you have selected the variables the way you want, push OK.

The output will provide the following table.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.718[a]	.515	.475	19.48054

a. Predictors: (Constant), Age

This provides the correlation, coefficient of determination, coefficient of determination adjusted for number of predictor variables (which is of more use later, when doing more than simple linear regression), and the **standard error of the estimate**, s_e, similar to example 2, page 740.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4840.462	1	4840.462	12.755	.004[a]
	Residual	4553.895	12	379.491		
	Total	9394.357	13			

a. Predictors: (Constant), Age

b. Dependent Variable: Total Cholesterol

The ANOVA table provides values used in both the hypothesis test. This is the test that at least one predictor variable has a linear relationship with the response. With a simple linear regression, the F-value is t^2, and the p-values are the same. The Mean Square for the Residual Row is s_e^2, or the variance of the estimate.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	151.354	17.284		8.757	.000
	Age	1.399	.392	.718	3.571	.004

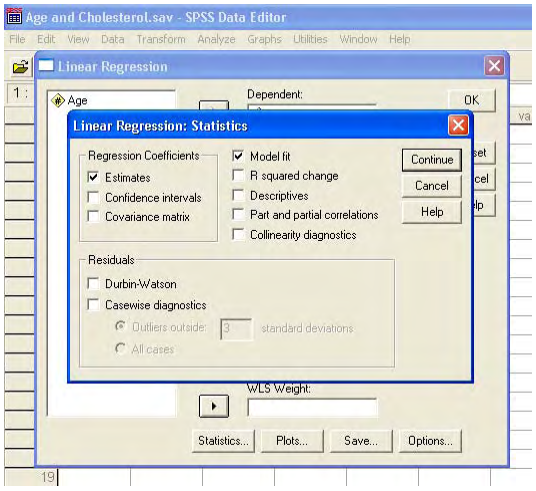
a. Dependent Variable: Total Cholesterol

For the Coefficients table, the (Constant) row relates to the Intercept, and the Age row relates to the slope. The B column is the estimated coefficients; the intercept and the slope. In this example, the regression line is Total Cholesterol=151.354+1.399*Age. The standard error column is the standard error for that estimate, so the standard error for the slope is .392. The t column is the t-test testing whether the population coefficient is 0. We don't usually worry about the intercept being 0, unless x=0 is part of the data, since we don't worry about values outside the scope of the sample. The significance (Sig.) value is the p-value for the test.

Because the p-value (.004) is less than α (.05), we reject the null hypothesis (See example 5 page 744). There is sufficient evidence at the α=.05 level of significance to show that there is a linear relationship between Age and Total Cholesterol.

It must be remembered that relation is NOT the same as causation. Just because we can show a relationship, even if the correlation is 1 or -1, that does not mean we know what causes the values of either variable.

To create a confidence interval for the slope (see Example 7, page 747), when in the **Linear Regression** window, click on the **Statistics** button, and check the Confidence Intervals option. This will create a 95% confidence interval for the intercept and slope. You cannot change the level of confidence. For a different confidence level, you will have to do it by hand.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	151.354	17.284		8.757	.000	113.696	189.012
	Age	1.399	.392	.718	3.571	.004	.546	2.253

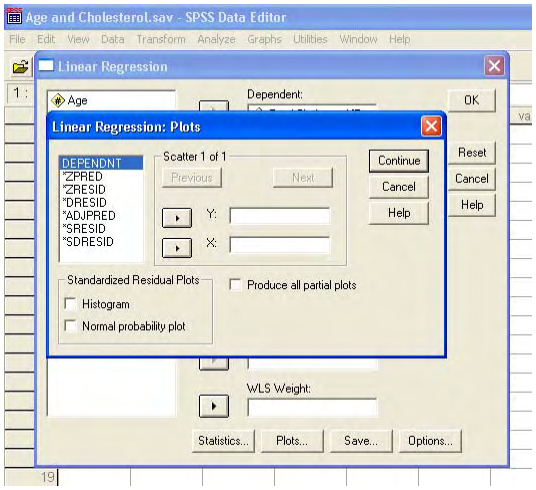
a. Dependent Variable: Total Cholesterol

We are 95% confident that the slope is between 0.546 and 2.253. Again, we normally do not worry about the intercept, unless it is in the scope of the data.

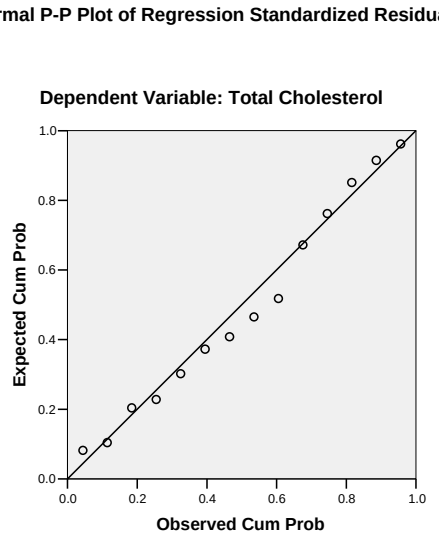
►Verifying that the Residuals are Normally Distributed

Example 4, page 742

To verify that the residuals are normally distributed, we go to the plots menu in the linear regression window.



In the bottom left hand corner, we see the Standardized Residual Plots. We can get a Normal probability plot by checking the box. Also, if wanted, we could also get a histogram of the residuals.

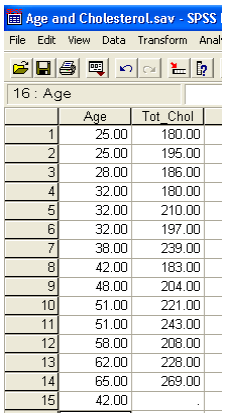


Section 14.2 Confidence and Prediction Intervals

►Creating Confidence and Prediction Intervals

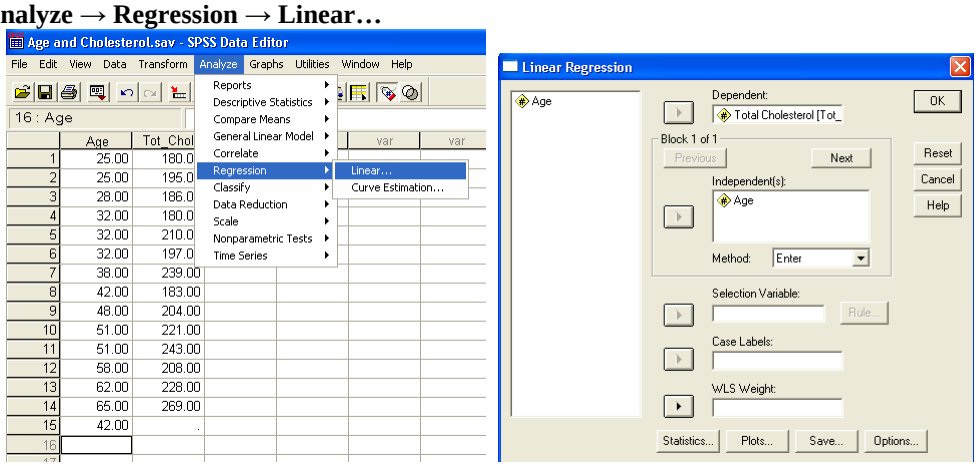
Example 1, Page 754

To obtain a prediction or confidence interval, add the value of x that you want to predict for (if it is not already a part of the list), leaving a missing value for the y.



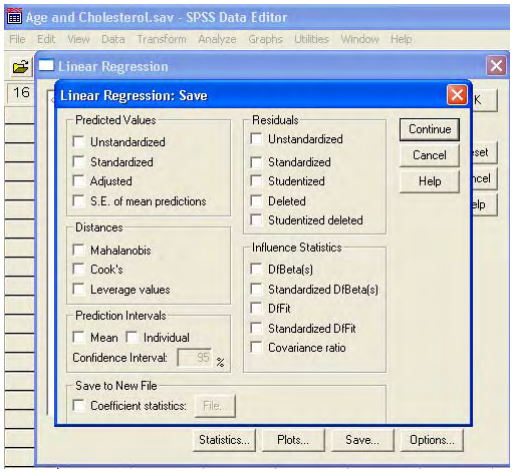
	Age	Tot_Chol
1	25.00	180.00
2	25.00	195.00
3	28.00	186.00
4	32.00	180.00
5	32.00	210.00
6	32.00	197.00
7	38.00	239.00
8	42.00	183.00
9	48.00	204.00
10	51.00	221.00
11	51.00	243.00
12	58.00	208.00
13	62.00	228.00
14	65.00	269.00
15	42.00	.

Go to **Analyze → Regression → Linear…**

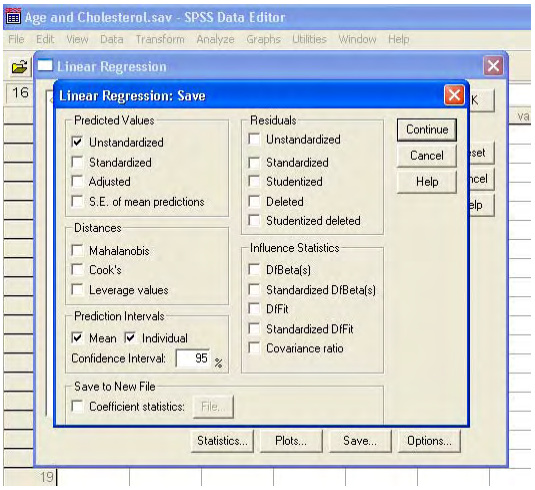


Use the Dependent, or Y, or response variable in the Dependent box. The Independent, or X, or predictor variable goes in the Independent(s) box.

Before going to OK, push the **Save** Button.

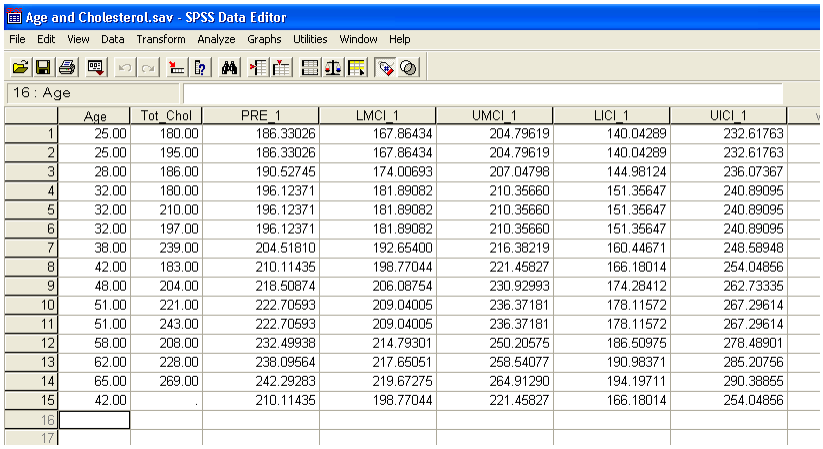


Select the options Unstandardized Predicted Value, which will give you $\hat{y}$, the predicted value for each x value, and click the Mean (confidence interval) and Individual (prediction interval) boxes in the Prediction Intervals section, and select the level of confidence.



Push Continue, and OK

The output will look the same, with only one table added, which is a table of residual statistics. The predicted values, and intervals are back in the data table.



	Age	Tot_Chol	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	25.00	180.00	186.33026	167.86434	204.79619	140.04289	232.61763
2	25.00	195.00	186.33026	167.86434	204.79619	140.04289	232.61763
3	28.00	186.00	190.52745	174.00693	207.04798	144.98124	236.07367
4	32.00	180.00	196.12371	181.89082	210.35660	151.35647	240.89095
5	32.00	210.00	196.12371	181.89082	210.35660	151.35647	240.89095
6	32.00	197.00	196.12371	181.89082	210.35660	151.35647	240.89095
7	38.00	239.00	204.51810	192.65400	216.38219	160.44671	248.58948
8	42.00	183.00	210.11435	198.77044	221.45827	166.18014	254.04856
9	48.00	204.00	218.50874	206.08754	230.92993	174.28412	262.73335
10	51.00	221.00	222.70593	209.04005	236.37181	178.11572	267.29614
11	51.00	243.00	222.70593	209.04005	236.37181	178.11572	267.29614
12	58.00	208.00	232.49938	214.79301	250.20575	186.50975	278.48901
13	62.00	228.00	238.09564	217.65051	258.54077	190.98371	285.20756
14	65.00	269.00	242.29283	219.67275	264.91290	194.19711	290.38855
15	42.00	.	210.11435	198.77044	221.45827	166.18014	254.04856
16							
17							

Here, we have 5 new variables. The first, *Pre_1* is the predicted value, so the predicted value when age is 42 is 210.11435. The next two, *LMCI_1* and *UMCI_1* are the lower and upper (L and U) limits for the **Mean** Confidence Interval (MCI). This is the 95% confidence interval for the mean of a group of people all aged 42. *LICI_1*, and *UICI_1* are the lower and upper limits for the **Individual** Confidence Interval (ICI), or the prediction interval. This is the 95% prediction interval for an individual aged 42.
We are 95% confident that the mean total cholesterol of all 42-year-old females is between 198.77044 and 221.45827. (198.77 and 221.46).
We are 95% confident that the total cholesterol for a randomly selected 42-year-old female is between 166.18014 and 254.04856 (166.18 and 254.05).

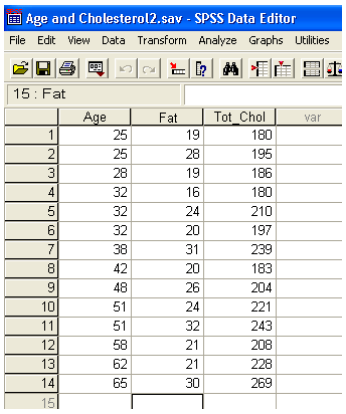
You can add as many values as is necessary to predict for, just keep in mind that they should be in the scope of the data.

Section 14.3 Multiple Linear Regression

►Obtaining the Correlation Matrix

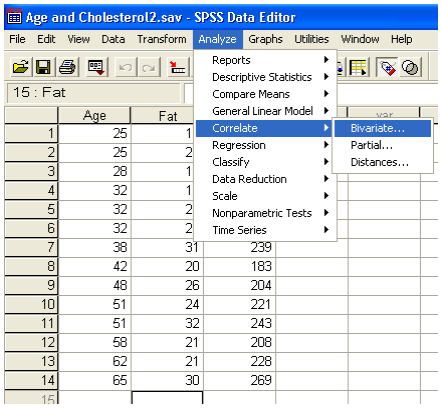
Example 1, Page 760

To find the correlation matrix, we follow many of the same steps as in chapter 4. First, we enter the data into an SPSS worksheet.



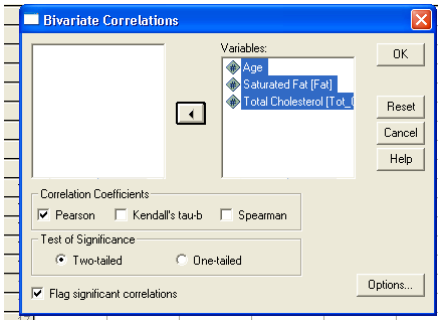
	Age	Fat	Tot_Chol	var
1	25	19	180	
2	25	28	195	
3	28	19	186	
4	32	16	180	
5	32	24	210	
6	32	20	197	
7	38	31	239	
8	42	20	183	
9	48	26	204	
10	51	24	221	
11	51	32	243	
12	58	21	208	
13	62	21	228	
14	65	30	269	
15				

Once the data have been entered, we go to **Analyze → Correlate → Bivariate**.



	Age	Fat	Tot_Chol
1	25	1	
2	25	2	
3	28	1	
4	32	1	
5	32	2	
6	32	2	
7	38	31	239
8	42	20	183
9	48	26	204
10	51	24	221
11	51	32	243
12	58	21	208
13	62	21	228
14	65	30	269
15			

Select all the variables that you want in the correlation matrix.



We are using Pearson Correlations, so make sure the Pearson box is checked. Push OK, and the output will provide you with the following table:

Correlations

		Age	Saturated Fat	Total Cholesterol
Age	Pearson Correlation	1	.324	.718**
	Sig. (2-tailed)		.258	.004
	N	14	14	14
Saturated Fat	Pearson Correlation	.324	1	.778**
	Sig. (2-tailed)	.258		.001
	N	14	14	14
Total Cholesterol	Pearson Correlation	.718**	.778**	1
	Sig. (2-tailed)	.004	.001	
	N	14	14	14

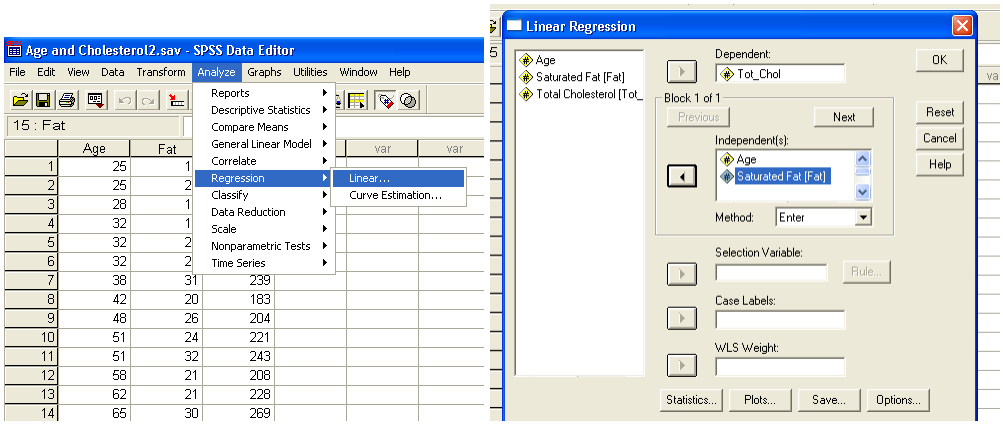
**. Correlation is significant at the 0.01 level (2-tailed).

The first value in each cell is the correlation coefficient. So the correlation between Saturated Fat and Age is 0.324, which has a p-value for the test of being 0 of 0.258, so we cannot show that age and saturated fat is correlated. The values with the ** by them are significant at α=0.01, and should be not be used together in the multiple regression for the same response. The response should be significantly correlated with the predictors in order to have a significant model. Since age and saturated fat are not significantly correlated, we do not have to worry about multicollinearity.

►Obtaining the Multiple Regression Equation

Example 2, Page 761

To find the multiple regression equation, we go to Analyze → Regression → Linear, just like in the simple linear regression. In the simple linear regression, we only picked one predictor variable, now we want to select both predictor variables.



We can follow the same steps as in the simple linear regression for plots, confidence intervals, etc. The output should be as follows:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.921[a]	.847	.820	11.418

a. Predictors: (Constant), Saturated Fat, Age

This provides us with the model correlation, the model R^2 value, and the R^2 value after adjusting for having two predictors and only 14 observations. The adjusted R^2 value will penalize the model for having too many predictors, so that you don't get a model that just connects the dots with little or no meaning. (See example 4, page 765) We also get the standard error for the model.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7960.297	2	3980.148	30.530	.000[a]
	Residual	1434.060	11	130.369		
	Total	9394.357	13			

a. Predictors: (Constant), Saturated Fat, Age

b. Dependent Variable: Total Cholesterol

The ANOVA table tests whether there is something significant in the model or not. This is a global test of the full model, or the F-test for Lack of Fit (see example 5, page 766). If the ANOVA test is not significant, we can throw out all of the predictor variables as being insignificant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	90.842	15.989		5.682	.000
	Age	1.014	.243	.520	4.179	.002
	Saturated Fat	3.244	.663	.609	4.892	.000

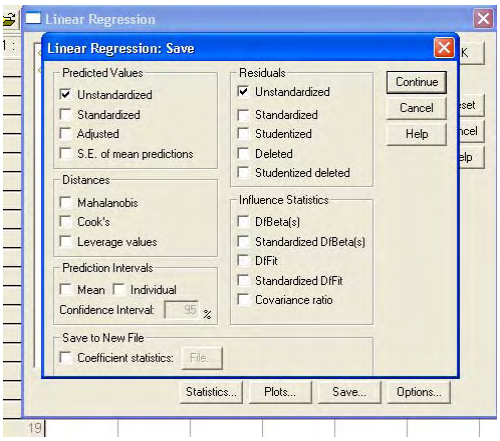
a. Dependent Variable: Total Cholesterol

If we do not throw out the whole model, we look at the Coefficients table to tell us what the model would be. Here, we have the model $\hat{y} = 90.842 + 1.014\text{Age} + 3.244\text{Saturated Fat}$. We can also test each coefficient using the t-tests. These test whether the coefficient is significant, provided the other coefficient is still in the model. In this example, both t-tests have p-values less than .05, so both coefficients are significant at the α=0.05 level. We still do not need to worry about the intercept unless 0 is a valid point for both Age and Saturated Fat. We need to stay within the scope of the model with both variables.

►Obtaining the Residual Plots

Example 2, Page 761

To obtain the residual plots, we must first find the residuals. While you are in the linear regression window, click on the **Save** button. In the save menu, you can obtain the predicted values, as we did with simple linear regression, and in the right hand column, you can obtain the residual values.



There are different types of residuals. The first on the list is the Unstandardized residuals. These are the residuals that you would calculate by hand, or the observed value-the expected value. The Standardized residuals are the residuals divided by an estimate of the standard error of the residuals. Since the standardized residuals are similar to z-values, we can look for outliers by looking for standardized values larger than 3, or less than -3. The Studentized residuals are similar to the standardized residuals, except the residual is divided by a standard error that is based on the location, similar to the confidence and prediction intervals. These three are common for residual plots.

The deleted and studentized deleted residuals are used for finding influential points. The Deleted residual is the residual value for a model that is made without the observation. These help to find influential points by omitting the value from the model, and seeing how well the observation fits with the predicted values for the rest of the observations.

We will select unstandardized residuals to match what we would get by hand. We also want the unstandardized predicted (fitted) values for use in the residual plots. Using the save menu does not add anything to the output window. It adds values to the data set. In the data, you will see a new column, RES_1.

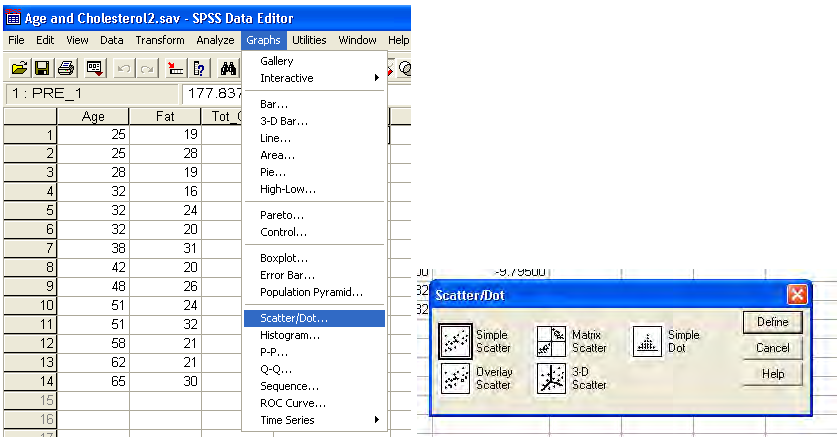
Age and Cholesterol2.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

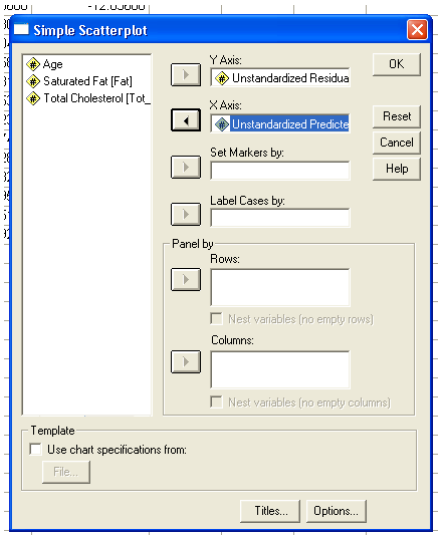
1 : PRE_1 177.837691353626

	Age	Fat	Tot_Chol	PRE_1	RES_1
1	25	19	180	177.83769	2.16231
2	25	28	195	207.03608	-12.03608
3	28	19	186	180.88031	5.11969
4	32	16	180	175.20433	4.79567
5	32	24	210	201.15846	8.84154
6	32	20	197	188.18139	8.81861
7	38	31	239	229.95355	9.04645
8	42	20	183	198.32345	-15.32345
9	48	26	204	223.87427	-19.87427
10	51	24	221	220.42836	.57164
11	51	32	243	246.38248	-3.38248
12	58	21	208	217.79500	-9.79500
13	62	21	228	221.85182	6.14818
14	65	30	269	254.09282	14.90718
15					

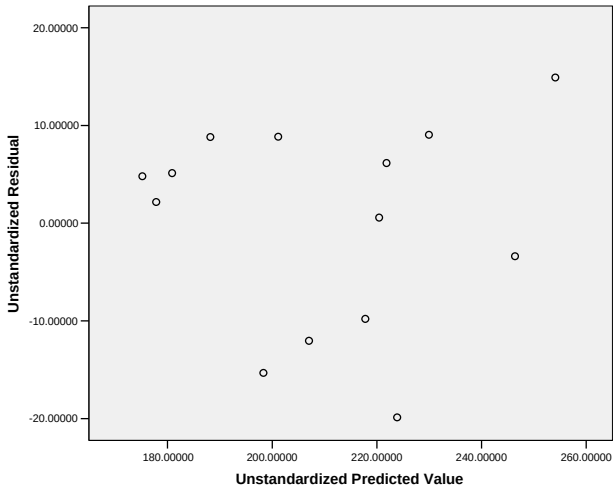
These are the residuals. We can now create plots. To create the plots, go to **Graphs → Scatter/Dot**, and select Simple Scatter.



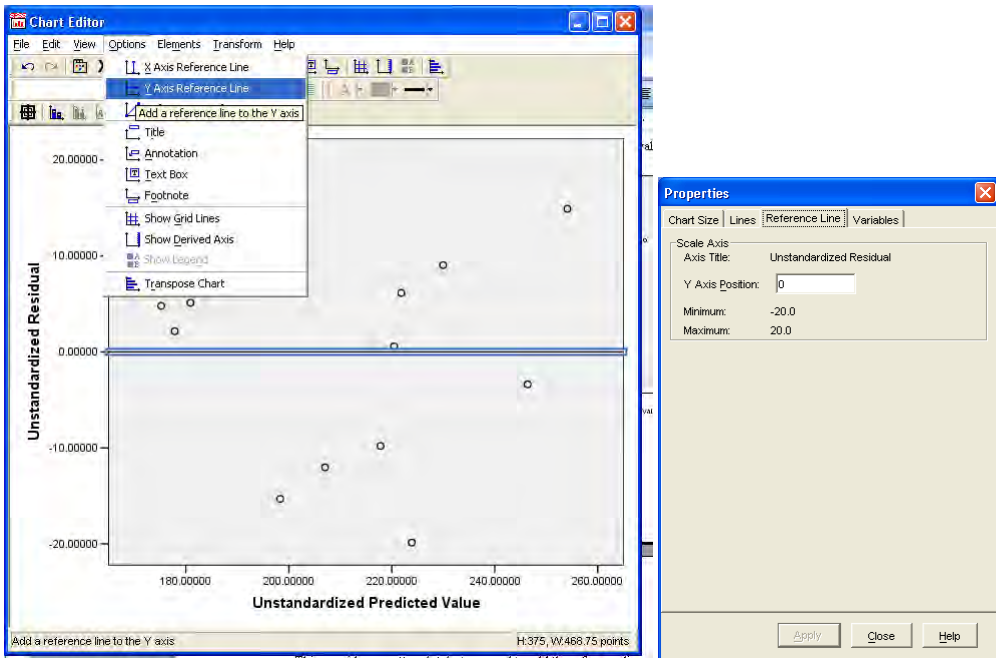
We now need to choose which plot we want to look at first. We will first look at the Residuals versus the fitted values plot.



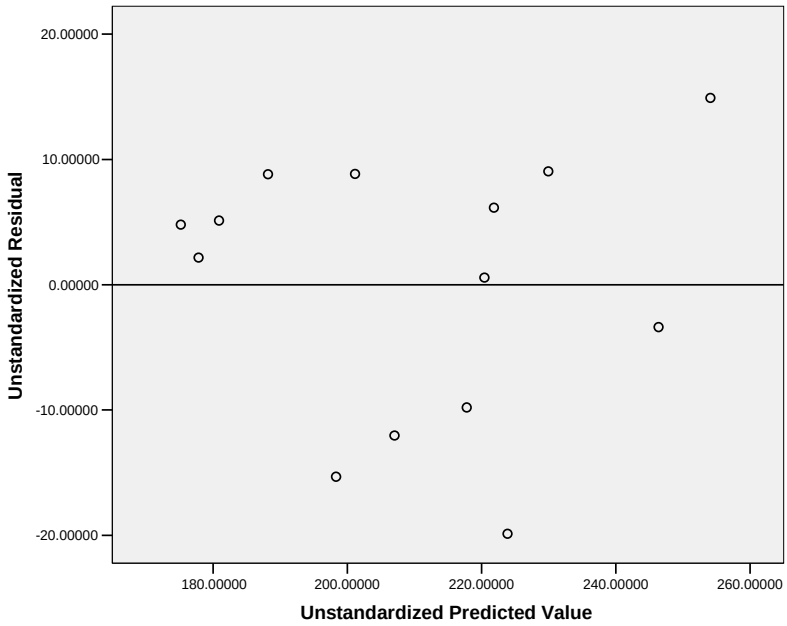
We select the residuals as the Y-Axis, and the predicted (fitted) values for the X-Axis. We then push OK.



This provides a scatter plot, but we need to add the reference line at 0. To do this, go to the chart editor. In the chart editor, find **Options → Y Axis Reference Line**.



Since we want a line at the mean (0), we put 0 in the Y Axis Position, and push the **Close** button. We can then close the chart editor, and we have a residual plot.

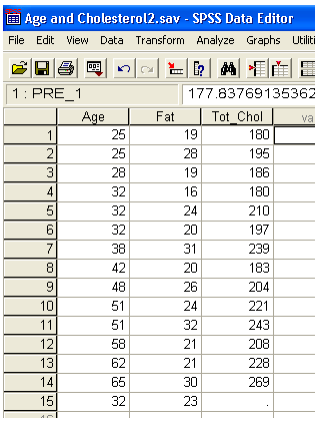


We can create the other residual plots in the same way, selecting different X-Axes.

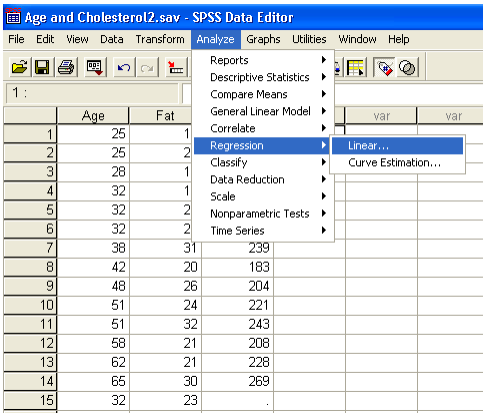
►Creating Confidence and Prediction Intervals

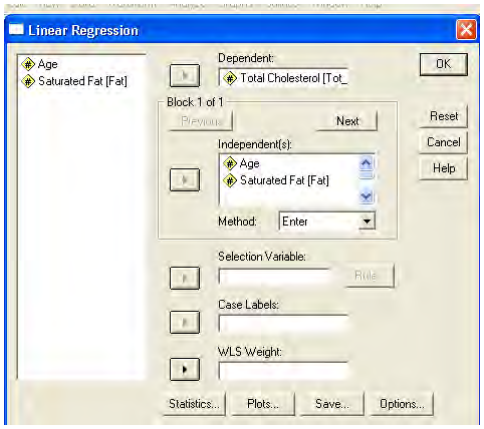
Obtaining a confidence and prediction interval is similar to the method used for the simple linear regression, except now we have to have two x values for each prediction.
Example 7, Page 768

To obtain a prediction or confidence interval, add the values of x that you want to predict for (if it is not already a part of the list), leaving a missing value for the y.



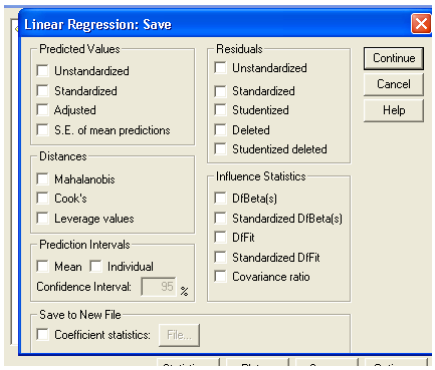
Go to **Analyze → Regression → Linear…**



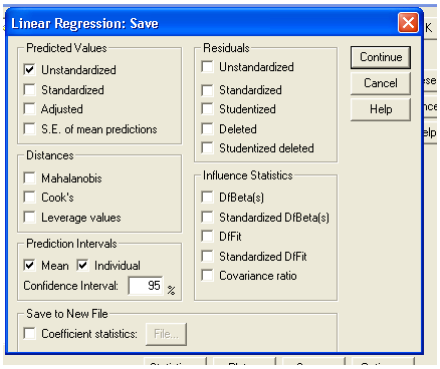


Put in the model that you want to use for the estimates.

Before going to OK, push the **Save** Button.



Select the options Unstandardized Predicted Value, which will give you $\hat{y}$, the predicted value for each x value, and click the Mean (confidence interval) and Individual (prediction interval) boxes in the Prediction Intervals section, and select the level of confidence.



Push Continue, and OK

The output will look the same, with only one table added, which is a table of residual statistics. The predicted values, and intervals are back in the data table.

Age and Cholesterol2.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : PRE_1 177.837691353626

	Age	Fat	Tot_Chol	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
1	25	19	180	177.83769	166.25585	189.41953	150.16657	205.50881
2	25	28	195	207.03608	192.67198	221.40018	178.08994	235.98222
3	28	19	186	180.88031	170.17898	191.58163	153.56602	208.19459
4	32	16	180	175.20433	162.57090	187.83776	147.07685	203.33182
5	32	24	210	201.15846	192.43223	209.88468	174.55585	227.76106
6	32	20	197	188.18139	179.02798	197.33481	161.43561	214.92717
7	38	31	239	229.95355	216.52561	243.38148	201.46037	258.44672
8	42	20	183	198.32345	189.76452	206.88237	171.77525	224.87165
9	48	26	204	223.87427	216.13380	231.61474	197.57852	250.17002
10	51	24	221	220.42836	212.27237	228.58434	194.00731	246.84941
11	51	32	243	246.38248	233.00531	259.75965	217.91319	274.85177
12	58	21	208	217.79500	205.39836	230.19164	189.77307	245.81692
13	62	21	228	221.85182	207.71142	235.99221	193.01603	250.68760
14	65	30	269	254.09282	239.68592	268.49972	225.12542	283.06023
15	32	23	.	197.91419	189.44870	206.37968	171.39596	224.43242

Here, we have 5 new variables. The first, *Pre_1* is the predicted value, so the predicted value when age is 32 and saturated fat is 23 grams is 197.91419. The next two, *LMCI_1* and *UMCI_1* are the lower and upper (L and U) limits for the **Mean** Confidence Interval (MCI). This is the 95% confidence interval for the mean of a group of people all aged 42. *LICI_1*, and *UICI_1* are the lower and upper limits for the **Individual** Confidence Interval (ICI), or the prediction interval. This is the 95% prediction interval for an individual aged 42.
We are 95% confident that the mean total cholesterol of all 32-year-old females who consume 23 grams of saturated fat daily is between 189.44870 and 206.37968 (189.45, 206.38).
We are 95% confident that the total cholesterol for a randomly selected 32-year-old females who consume 23 grams of saturated fat daily is between 171.39596 and 224.4342 (171.40 and 224.43).

You can add as many values as is necessary to predict for, just keep in mind that they should be in the scope of the data.

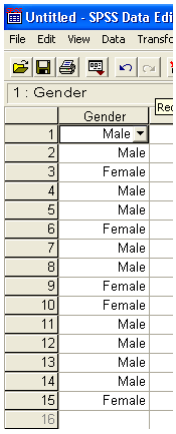
Chapter 15. Nonparametric Statistics

Section 15.2 Runs Test for Randomness

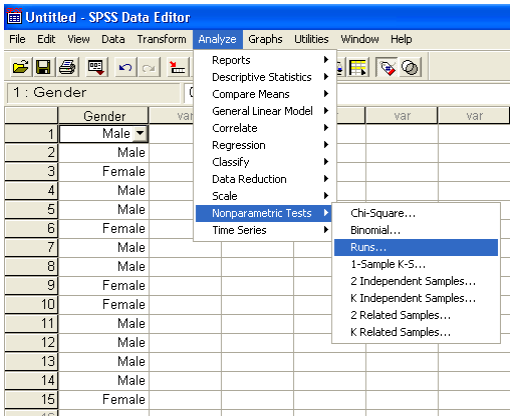
►Runs Test for Randomness

Example 3, Page 15-7

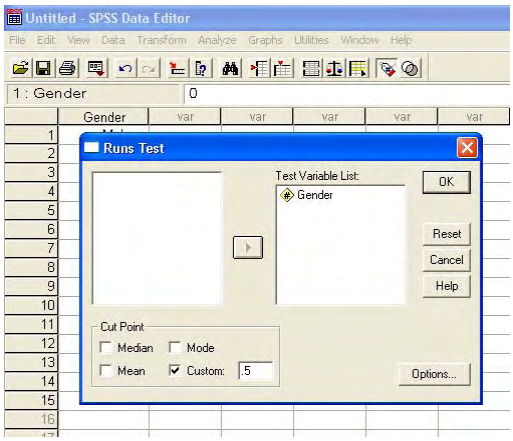
Enter the data in to SPSS. The values must be numerical, but labels can be attached. For this example, 0 was used to represent males, and 1 was used to represent females.



Once the data have been entered, go to **Analyze → Nonparametric Tests → Runs…**



In the Runs Test window, select the variable you wish to run the test on, and click on the ► button. There are four options you can use to separate the different groups. Observations with value less than the cut point are placed in one group, while observations with value higher than the cut point are placed in the second group. Two options are the median and mean. One problem with these choices, is that the median is most likely a part of one of the two groups. The median, and mean should be used when looking at more continuous values. The mode will be part of a group, and can also provide different answers. The easiest to use is the Custom. Here, type in a value that is between the two groups (this is why the values must be numerical). For this example, we use the value of 0.5, since that is between 0 and 1.



Once the variables and cut point are correct, push OK. The following will appear in the output window.

Runs Test

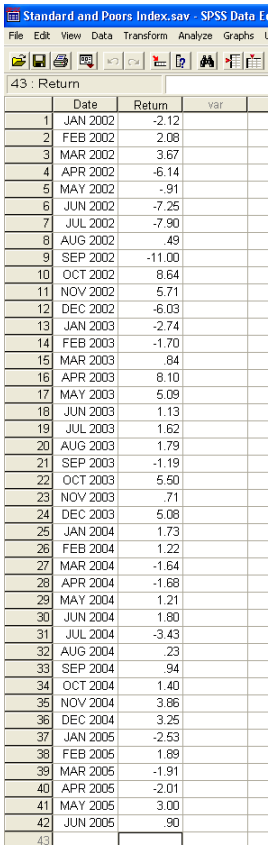
	Gender
Test Value[a]	.5000
Total Cases	15
Number of Runs	8
Z	.000
Asymp. Sig. (2-tailed)	1.000

a. User-specified.

Here, we see that we had a total of 15 observations, with 8 runs. We can use this in the small sample case, and look up 8 in the table. Because 8 is not less than or equal to the lower critical value, 3, and 8 is not greater than or equal to the upper critical value, 12, we cannot reject the null hypothesis. The p-value (Asymp. Sig. (2-tailed)) should only be used in large sample cases.

Example 4, page 15-7

Example 4 is a large sample case, so we can see how the z-score of SPSS correlates with what we would find by hand. First, we enter the data. This can be done by using the Ps and Ns as in the book, but in SPSS we can work with the original values, so there is no need to convert them.



	Date	Return	var
1	JAN 2002	-2.12	
2	FEB 2002	2.08	
3	MAR 2002	3.67	
4	APR 2002	-6.14	
5	MAY 2002	-.91	
6	JUN 2002	-7.25	
7	JUL 2002	-7.90	
8	AUG 2002	.49	
9	SEP 2002	-11.00	
10	OCT 2002	8.64	
11	NOV 2002	5.71	
12	DEC 2002	-6.03	
13	JAN 2003	-2.74	
14	FEB 2003	-1.70	
15	MAR 2003	.84	
16	APR 2003	8.10	
17	MAY 2003	5.09	
18	JUN 2003	1.13	
19	JUL 2003	1.62	
20	AUG 2003	1.79	
21	SEP 2003	-1.19	
22	OCT 2003	5.50	
23	NOV 2003	.71	
24	DEC 2003	5.08	
25	JAN 2004	1.73	
26	FEB 2004	1.22	
27	MAR 2004	-1.64	
28	APR 2004	-1.68	
29	MAY 2004	1.21	
30	JUN 2004	1.80	
31	JUL 2004	-3.43	
32	AUG 2004	.23	
33	SEP 2004	.94	
34	OCT 2004	1.40	
35	NOV 2004	3.86	
36	DEC 2004	3.25	
37	JAN 2005	-2.53	
38	FEB 2005	1.89	
39	MAR 2005	-1.91	
40	APR 2005	-2.01	
41	MAY 2005	3.00	
42	JUN 2005	.90	
43			

Once the data have been entered, go to **Analyze → Nonparametric Tests → Runs…**. We follow the same steps as for a small data set. We can set the cut point to 0, so that positive values would be in one group, and negative values in the second group. The output is then

Runs Test

	Return
Test Value[a]	.0000
Total Cases	42
Number of Runs	18
Z	-.889
Asymp. Sig. (2-tailed)	.374

a. User-specified.

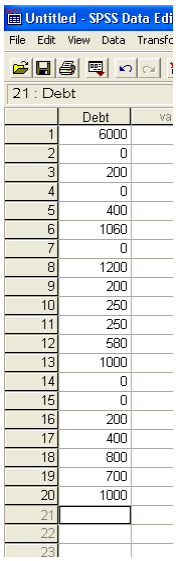
We see 42 observations, with 18 runs. We get a z-value of -0.889, which is different from what we get by hand. SPSS uses a correction for sample sizes less than 50, of $z = \begin{cases} (R - \mu_r + 0.5)/\sigma, \text{if } R - \mu \le 0.5 \\ (R - \mu_r - 0.5)/\sigma, \text{if } R - \mu \ge 0.5 \end{cases}$. We get a two-tailed p-value of 0.374, which is larger than 0.05, so we fail to reject the null hypothesis.

Section 15.3 Inferences about Measures of Central Tendency

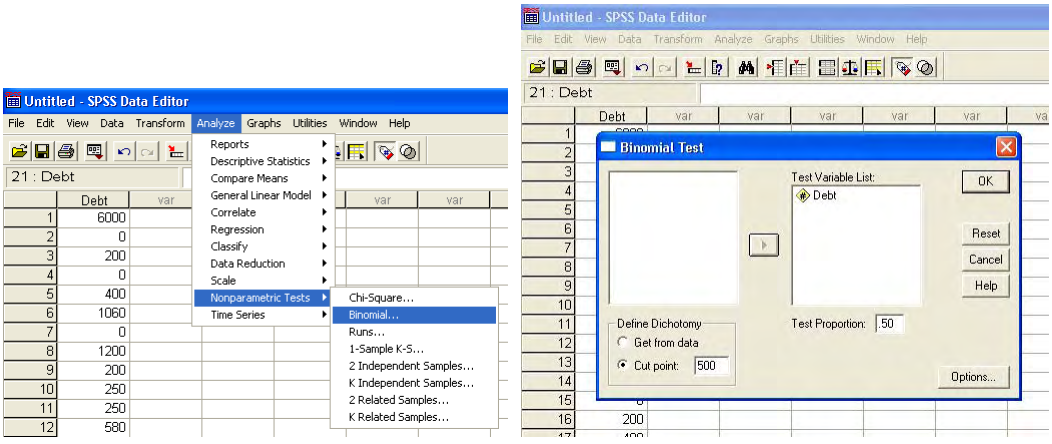
►One-Sample Sign Test

Example 1, Page 15-13

Type the data into SPSS.



Once the data have been entered, go to **Analyze → Nonparametric Tests → Binomial…**.



Select the variable of interest as the Test Variable List, and your hypothesized median as the Cut Point. Keep the Test Proportion at .50 (the median). When the values have been entered, push OK.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Debt	Group 1	<= 500	12	.60	.50	.503
	Group 2	> 500	8	.40		
	Total		20	1.00		

The output shows how many observations were less than the hypothesized median (12), and how many were above the hypothesized median (8). It also provides the p-value. We do not, in this case, have enough evidence to show that the median is not 500.

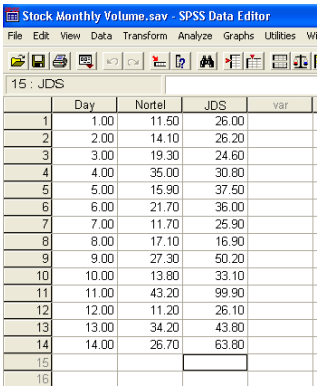
One note: SPSS does not take out ties. It groups the less than or equal to in the same group instead of removing those values which are equal to the hypothesized median.

Section 15.4 Inferences about the Difference between Two Measures of Central Tendency: Dependent Samples

►Wilcoxon Matched-Pairs Signed Rank Test

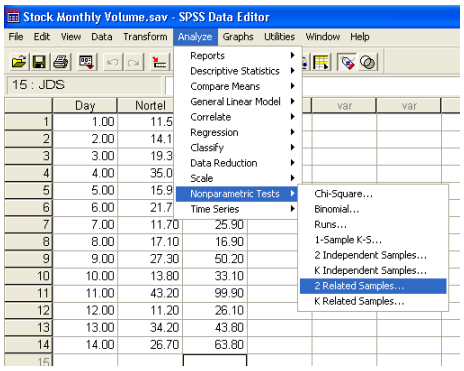
Example 1, Page 15-21

We start by typing in the data. This is similar to the paired t-test. We create two variables.

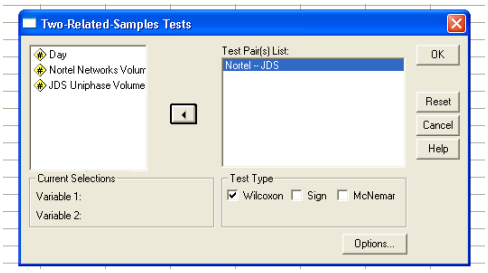


	Day	Nortel	JDS	var
1	1.00	11.50	26.00	
2	2.00	14.10	26.20	
3	3.00	19.30	24.60	
4	4.00	35.00	30.80	
5	5.00	15.90	37.50	
6	6.00	21.70	36.00	
7	7.00	11.70	25.90	
8	8.00	17.10	16.90	
9	9.00	27.30	50.20	
10	10.00	13.80	33.10	
11	11.00	43.20	99.90	
12	12.00	11.20	26.10	
13	13.00	34.20	43.80	
14	14.00	26.70	63.80	
15				
16				

Once the data have been entered, go to **Analyze → Nonparametric Tests → 2 Related Samples…**.



Similar to the dependent sample t-test, select the two variables. Again, the test will be run in the order of the columns.



Make sure that the Wilcoxon option is checked, and press OK.

Ranks

		N	Mean Rank	Sum of Ranks
JDS Uniphase Volume - Nortel Networks Volume	Negative Ranks	2[a]	1.50	3.00
	Positive Ranks	12[b]	8.50	102.00
	Ties	0[c]		
	Total	14		

a. JDS Uniphase Volume < Nortel Networks Volume

b. JDS Uniphase Volume > Nortel Networks Volume

c. JDS Uniphase Volume = Nortel Networks Volume

Test Statistics[b]

	JDS Uniphase Volume - Nortel Networks Volume
Z	-3.107[a]
Asymp. Sig. (2-tailed)	.002

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

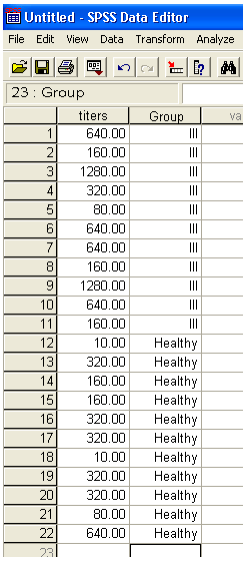
The output provides the mean (1.50) and sum (2) of the negative ranks, the mean (8.50) and sum (12) of the positive ranks, how many ties (0) there were, and the p-value for a two tailed test (0.002). We can get the p-value for the one-tailed test by dividing this by 2.

Section 15.5 Inferences about the Difference between Two Measures of Central Tendency: Independent Samples

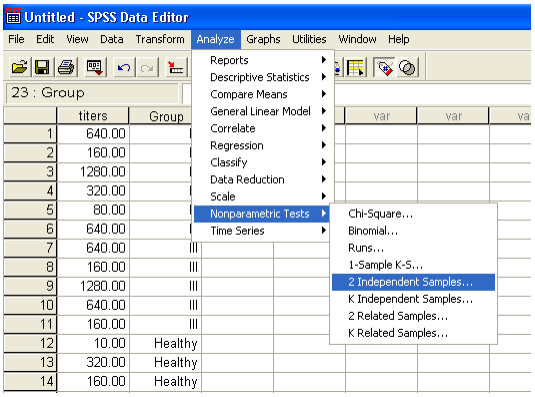
►Mann-Whitney Test

Example 1, Page 15-30

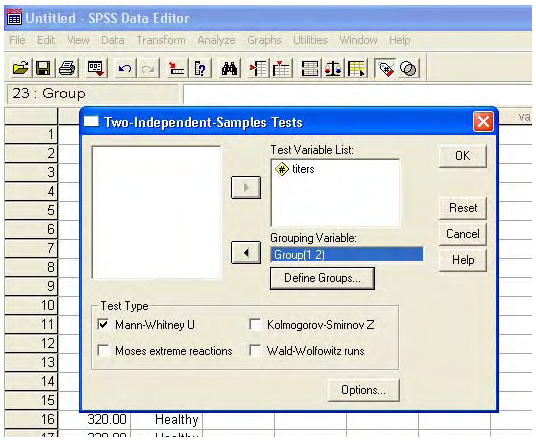
We start by typing in the data. This is similar to the independent t-test. We create two variables, one for the values and one for the group.



Once the data have been entered, go to **Analyze → Nonparametric Tests → 2 Independent Samples…**.



In the Two-Independent Samples Test window, we follow the basic steps of the independent t-test. We select the variable with the values for the Test Variable List, and the variable defining the group as the Grouping Variable. We must also define what the two groups are.



Make sure that Mann-Whitney U is checked, and push OK.

Ranks

	Group	N	Mean Rank	Sum of Ranks
titers	Ill	11	13.82	152.00
	Healthy	11	9.18	101.00
	Total	22		

Test Statistics[b]

	titers
Mann-Whitney U	35.000
Wilcoxon W	101.000
Z	-1.713
Asymp. Sig. (2-tailed)	.087
Exact Sig. [2*(1-tailed Sig.)]	.101[a]

a. Not corrected for ties.

b. Grouping Variable: Group

The output provides the mean and sum of the ranks for each group, along with the Mann-Whitney U statistic, and p-values. The p-value is for the two-tailed test, so for a one-tailed test we divide this in half, or 0.0507 in this example. Since the p-value is smaller than .1, we reject the null hypothesis.

Example 2, Page 15-32

Following the same steps as Example 1, we get the following output:

Ranks

	State	N	Mean Rank	Sum of Ranks
pH	Texas	22	16.09	354.00
	Montana	20	27.45	549.00
	Total	42		

Test Statistics[a]

	pH
Mann-Whitney U	101.000
Wilcoxon W	354.000
Z	-2.997
Asymp. Sig. (2-tailed)	.003

a. Grouping Variable: State

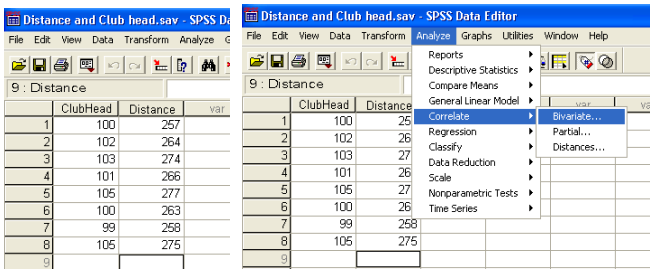
This shows the test statistic of 101 for the Mann-Whitney test, as well as the z-value of -2.997. Note that the Exact Sig. value (Exact Significance is based on the Mann-Whitney table) is no longer part of the output. Only the Asymptotic p-value, for the z-test is available.

Section 15.6 Spearman's Rank-Correlation Test

►Spearman's Rank-Correlation Test

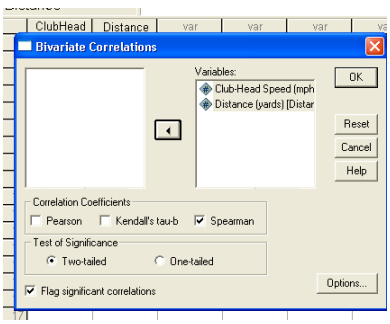
Example 1, Page 15-39

This test follows the steps for the parametric (Pearson Correlation) test almost exactly in SPSS. We begin by typing in the data. After entering the data into SPSS, go to **Analyze → Correlate → Bivariate...** (two variables)



Highlight the variables you wish to find the correlation between and push the ► button.

Make sure that Spearman is selected. Pearson may be selected for the parametric correlation tests.



Push OK

Correlations

			Club-Head Speed (mph)	Distance (yards)
Spearman's rho	Club-Head Speed (mph)	Correlation Coefficient	1.000	.928**
		Sig. (2-tailed)	.	.001
		N	8	8
	Distance (yards)	Correlation Coefficient	.928**	1.000
		Sig. (2-tailed)	.001	.
		N	8	8

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

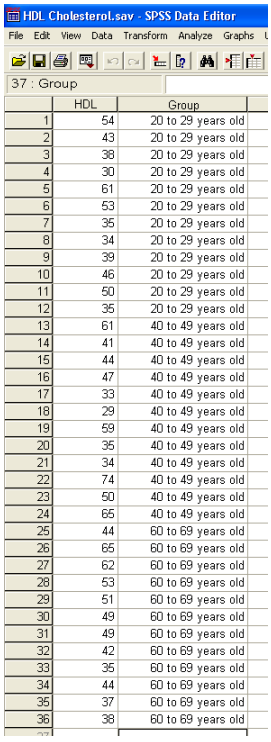
Here we have a correlation value of 0.928, with a p-value of 0.001, which shows that there is significant evidence to support the claim that GDP and Life Expectancy are associated.

Section 15.7 Kruskal-Wallis Test of One-Way Analysis of Variance

►Kruskal-Wallis One-Way Analysis of Variance Test

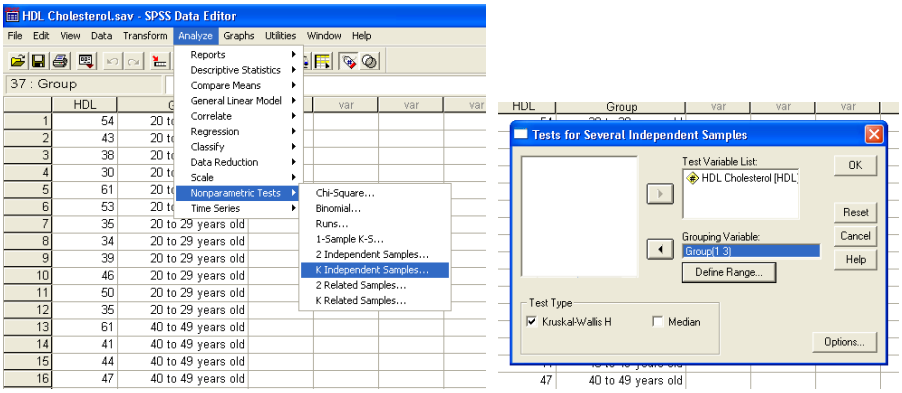
Example 1, Page 15-45

First, we must type in the data. This is done similar to the regular ANOVA test, or the Mann-Whitney test, with the value of interest in one column, and the group in another. The groups **must** be numerical, but you can use labels, which will help in the output.



	HDL	Group
1	54	20 to 29 years old
2	43	20 to 29 years old
3	38	20 to 29 years old
4	30	20 to 29 years old
5	61	20 to 29 years old
6	53	20 to 29 years old
7	35	20 to 29 years old
8	34	20 to 29 years old
9	39	20 to 29 years old
10	46	20 to 29 years old
11	50	20 to 29 years old
12	35	20 to 29 years old
13	61	40 to 49 years old
14	41	40 to 49 years old
15	44	40 to 49 years old
16	47	40 to 49 years old
17	33	40 to 49 years old
18	29	40 to 49 years old
19	59	40 to 49 years old
20	35	40 to 49 years old
21	34	40 to 49 years old
22	74	40 to 49 years old
23	50	40 to 49 years old
24	65	40 to 49 years old
25	44	60 to 69 years old
26	65	60 to 69 years old
27	62	60 to 69 years old
28	53	60 to 69 years old
29	51	60 to 69 years old
30	49	60 to 69 years old
31	49	60 to 69 years old
32	42	60 to 69 years old
33	35	60 to 69 years old
34	44	60 to 69 years old
35	37	60 to 69 years old
36	38	60 to 69 years old

Once the data have been entered, go to **Analyze → Correlate → K Independent Samples...**



Similar to the one-way ANOVA, we select the values as the Test Variable List, and the variable defining the groups as the Grouping Variable. Unlike the one-way ANOVA, we must define which groups we want to use, in a range.

Ranks

	Age Group	N	Mean Rank
HDL Cholesterol	20 to 29 years old	12	16.21
	40 to 49 years old	12	18.79
	60 to 69 years old	12	20.50
	Total	36	

Test Statistics[a,b]

	HDL Cholesterol
Chi-Square	1.012
df	2
Asymp. Sig.	.603

a. Kruskal Wallis Test

b. Grouping Variable: Age Group

The output provides the mean rank for each group, the Kruskal-Wallis test statistic, and the p-value (Asymp. Sig). Because the p-value is not less than 0.05, we fail to reject the null hypothesis. There is not sufficient evidence to show that the distributions of HDL cholesterol for the three age groups are different.