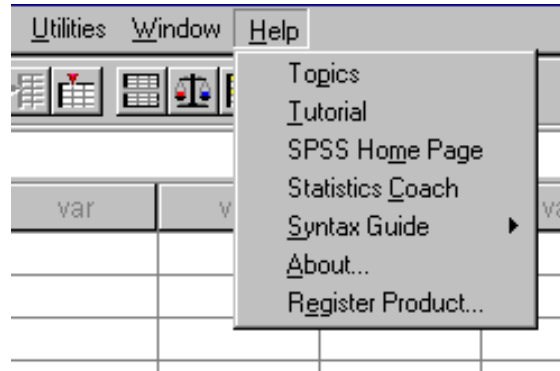


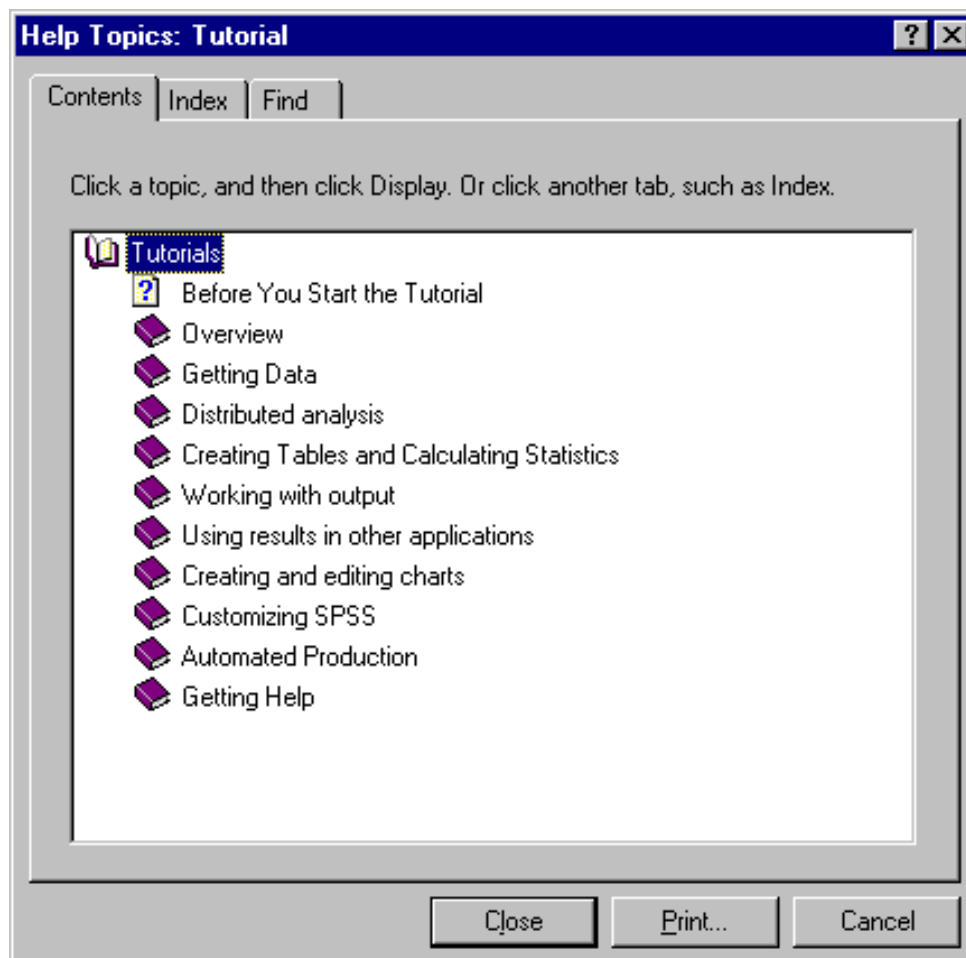
Using SPSS 11*

Descriptive Statistics

SPSS Help. SPSS has a good online help system. Once SPSS is up and running, you can find it by going to **Help>Topics** in the menu bar, i.e., click **Help** in the menu bar and then click **Topics** in the drop window that opens.



You will now be in the help contents window. Double click **Tutorials**.



You can then open any of the books comprising the tutorial by double clicking on it,

* Brother Walter Schreiner, FSC (June 19, 2002)

and bring up a topic by double clicking on it. Once a topic is open, you can just keep clicking on the **Next** button on the upper right to move through it page by page, or on **Contents** to go back to the contents page. I suggest going through the entire **Overview** booklet. Once you are working with a data set, and have an idea of what you want to do with the data, you can also use the **Statistics Coach** under the **Help** menu to help get the information you wish. It will lead you through the SPSS process.

Using the SPSS Data Editor. When you begin SPSS, you open up to the Data Editor. For our purposes right now, you can learn how to do this by going to **Tutorials>Getting Data** under Contents in **Help>Topics**, and reading the document **Basic Structure of an SPSS Data File** along with the topics in the booklet **Using the Data Editor**. For more advanced information, you can go to the **Data Management** booklet on the **Contents** page. The data we will use is given in the table below, with the numbers indicating total protein ($\mu\text{g}/\text{ml}$).

76.33	77.63	149.49	54.38	55.47	51.70
78.15	85.40	41.98	69.91	128.40	88.17
58.50	84.70	44.40	57.73	88.78	86.24
54.07	95.06	114.79	53.07	72.30	59.36
76.33	77.63	149.49	54.38	55.47	51.70
59.20	67.10	109.30	82.60	62.80	61.90
74.78	77.40	57.90	91.47	71.50	61.70
106.00	61.10	63.96	54.41	83.82	79.55
153.56	70.17	55.05	100.36	51.16	72.10
62.32	73.53	47.23	35.90	72.20	66.60
59.76	95.33	73.50	62.20	67.20	44.73
57.68					

For our data, double click on the **var** at the top of the first column or click on the **Variable View** tab at the bottom of the page, type in “protein” in the **Name** column, and hit **Enter**. Under the assumption that you are going to enter numerical data, the rest of the row is filled in.

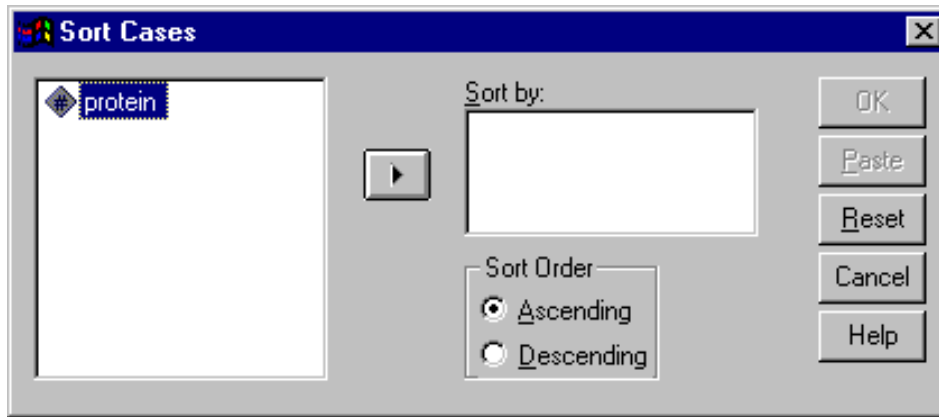
	Name	Type	Width	Decimals	Label	Values
1	protein	Numeric ...	8	2		None
2						

Changes in the type and display of the variable can be made by clicking in the appropriate cells and using any buttons given. Then hit the **Data View** tab and type in the data values, following each by **Enter**.

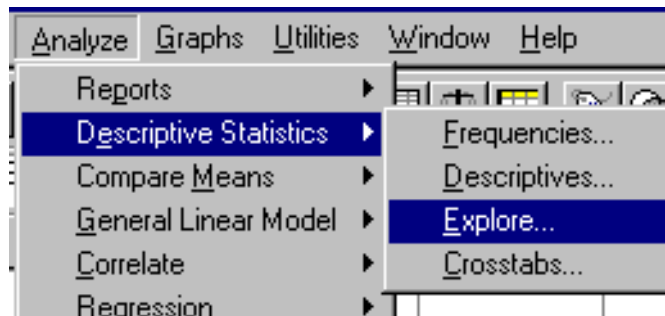
1 : protein		35.9
	protein	var
1	35.90	
2	41.98	
3	44.40	
4	44.73	

Save the file as usual where you wish under the name **protein.sav**. You just need type **protein**. The suffix is attached automatically.

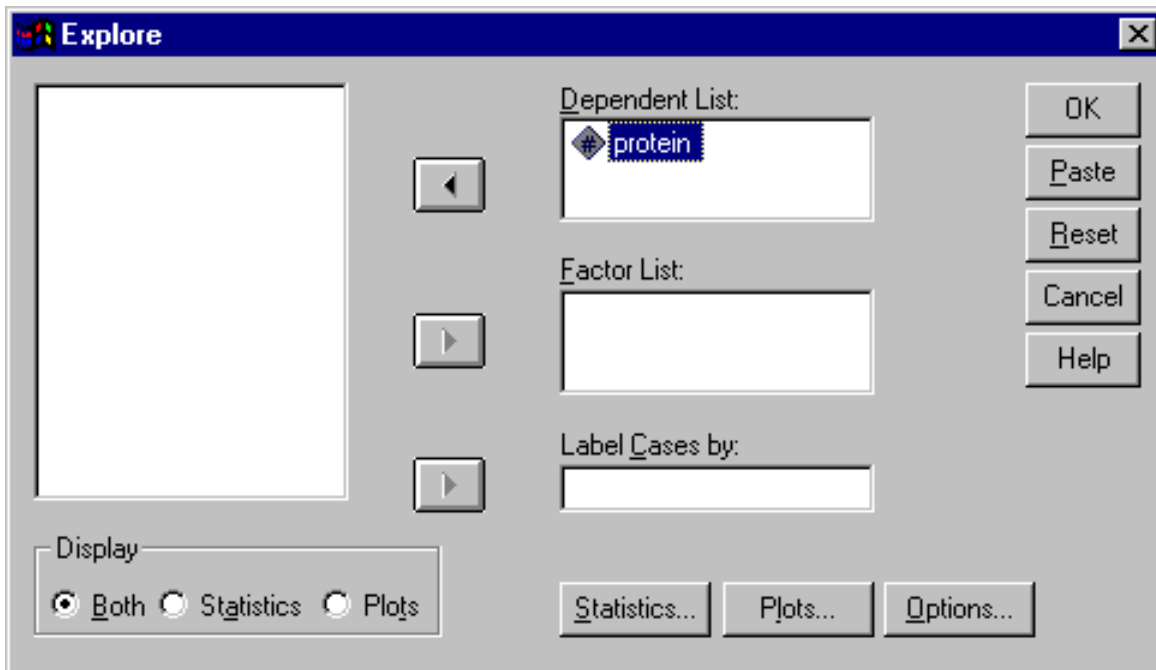
Sorting the Data. From the menu, choose **Data>Sort Cases...**, click the right arrow to move **protein** to the **Sort by** box, make sure **Ascending** is chosen, and click **OK**. Our data column is now in ascending order.



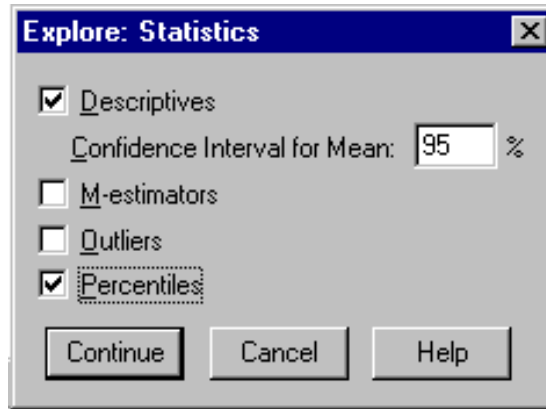
Obtaining the Descriptive Statistics. For an overview, you can go to **Creating Tables and Calculating Statistics** under **Tutorials**, or from the contents window check out **Descriptive Statistics** in the **Statistical Analysis** book, along with the **Graphical Analysis** and **Interactive Charts** books. For our purposes, go to **Analyze>Descriptive Statistics>Explore...**,



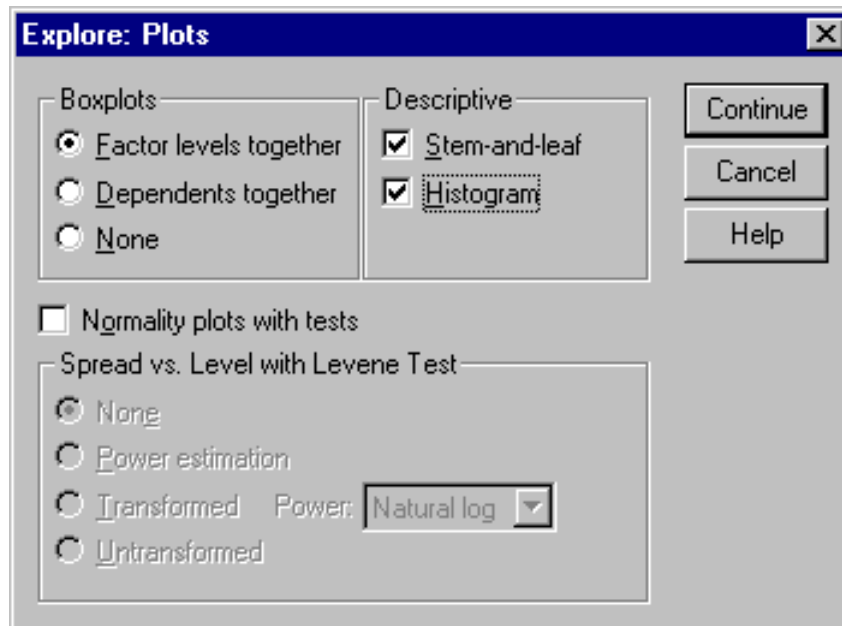
select **protein** from the box on the left, and then click the arrow for **Dependent List**. Make sure **Both** is checked under **Display**.



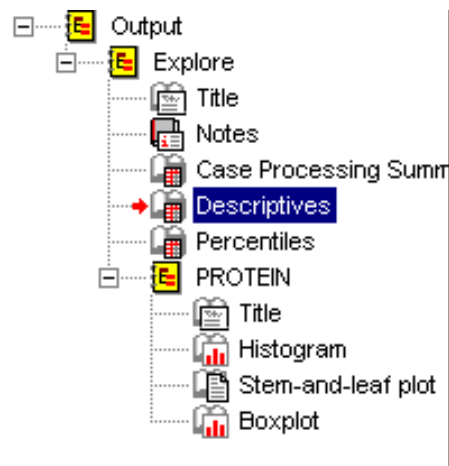
Click the **Statistics...** button, then make sure **Descriptives** and **Percentiles** are checked. We will use **95%** for **Confidence Interval for Mean**. Click **Continue**.



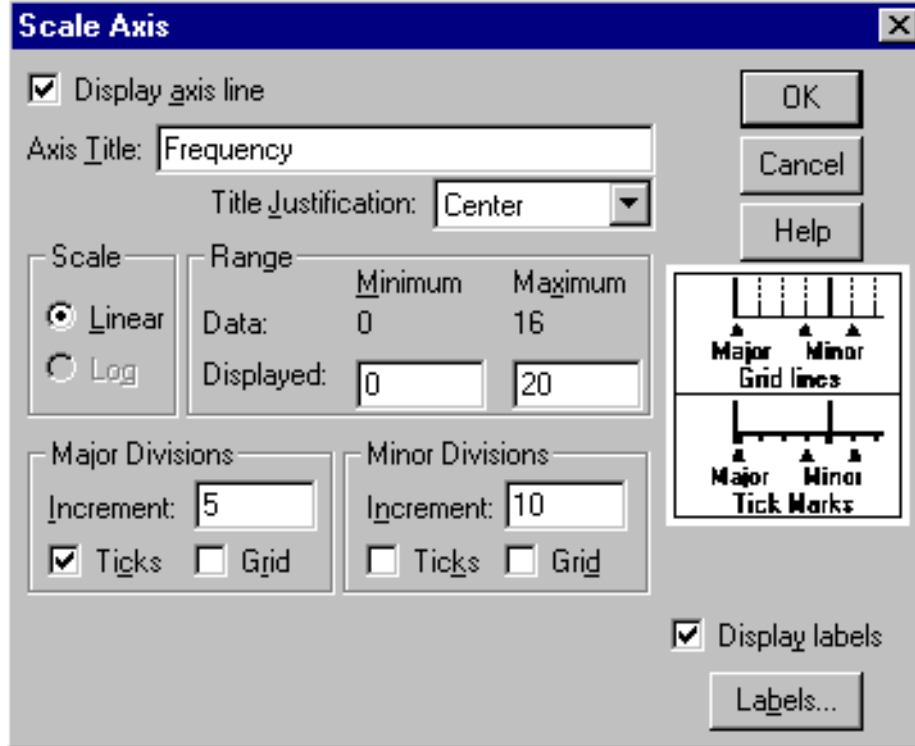
Then click **Plots...** Under **Boxplots**, select **Factor levels together**, and under **Descriptive**, choose both **Stem-and-leaf** and **Histogram**. Then click **Continue**.



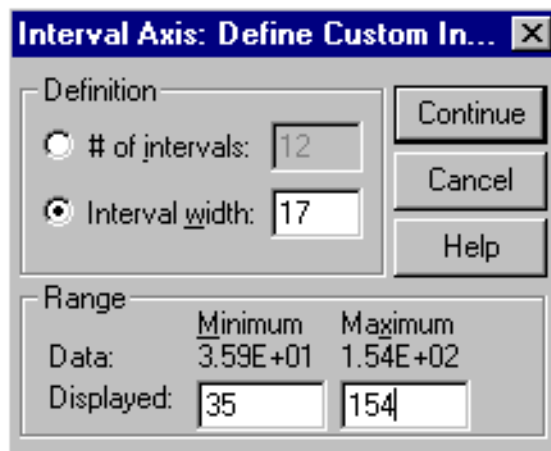
Then click **OK**. This opens an output window with two frames. The frame on the left contains an outline of the data on the right.



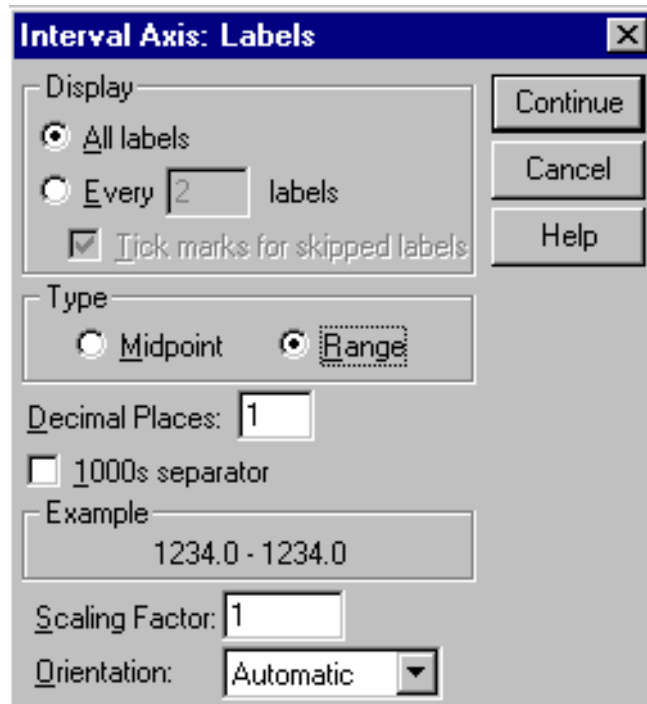
Clicking an item in either frame selects it, and allows you to copy it (and paste into a word processor), for instance. Double clicking an item in the left frame either shows or hides that item in the right frame. Double clicking an item in the right frame opens its editor, if it has one. Double click on the histogram. Once the chart editor opens, choose **Chart>Axis...** from the menu. With **Scale** chosen, which refers to the vertical axis, click **OK**. For **Title Justification**, choose **Center**. For **Increment**, under **Major Divisions**, type in **5**. Then click **OK**.



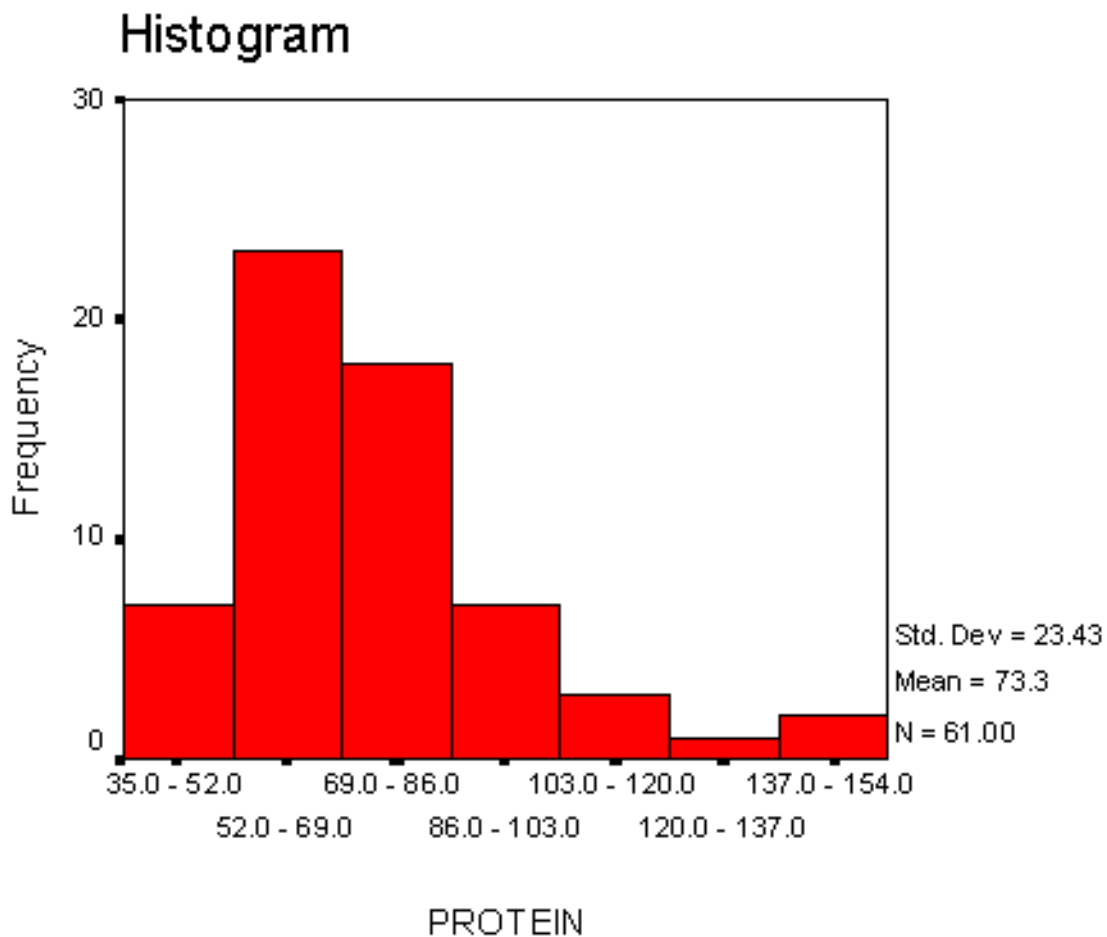
Choose **Chart>Axis...** again, but this time check **Interval** before clicking **OK**. Again choose **Center** for **Title Justification**. Choose **Custom** under **Intervals**, and then click **Define...** Choose **Interval Width** and type in **17**. Change the **Displayed Maximum** to **154**. Then click **Continue**.



Now click **Labels...** and under **Type**, click **Range**. This displays the cut points rather than the midpoints on the horizontal axis of the histogram.



Again click **Continue**, followed by **OK**. We get the histogram below.



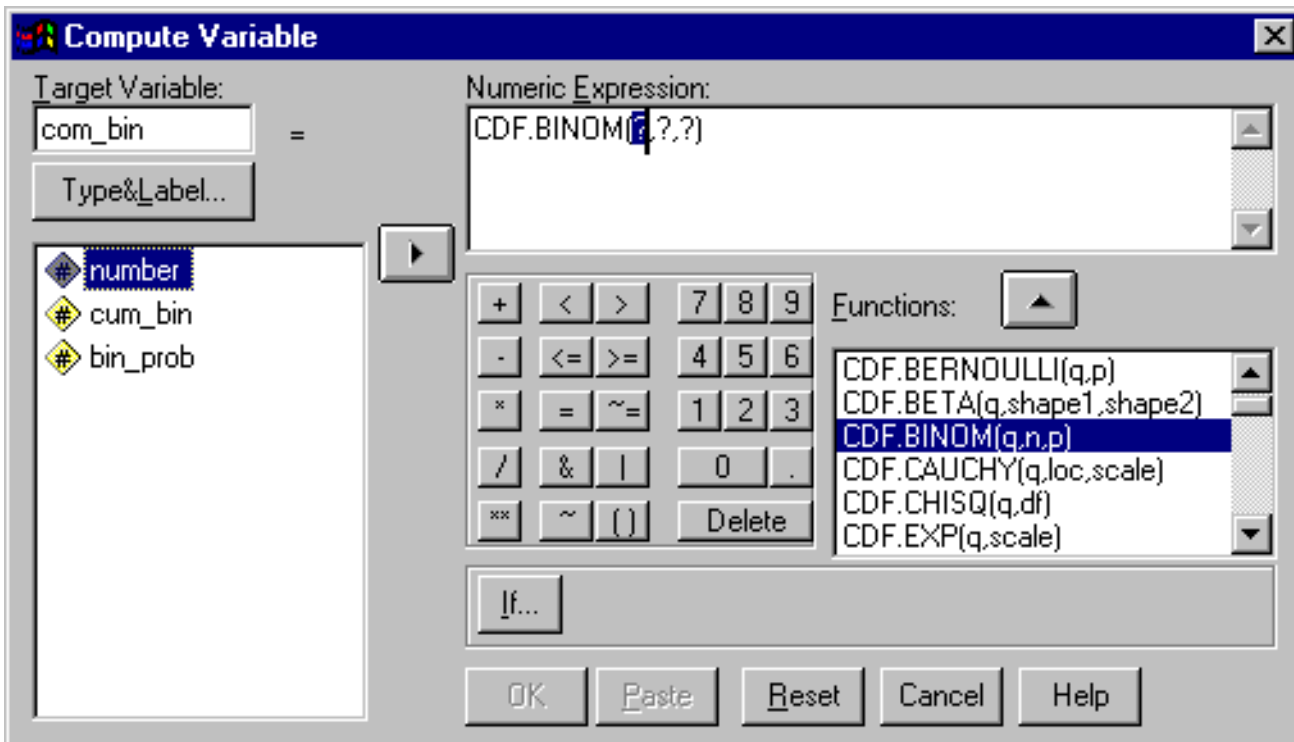
Copying Output to Word (for instance). When copying output, I suggest choosing **Edit>Copy objects**, because then when you paste into Word, it is treated as a picture and seems to place correctly. For graphics, you can either use **Edit>Copy objects** or **Edit>Copy**.

Probability Distributions

Binomial Distribution. We shall assume that $n=15$ and $p=.75$. We will first find $P(X \leq x | 15, .75)$ for $x = 0, \dots, 15$, i.e., the cumulative probabilities. First put the numbers 0 through 15 in a column of a worksheet. (Actually, you only need to enter the numbers whose cumulative probability you desire.) Then click **Variable View**, type in **number** (the name you choose is optional) under **Name**, and I suggest putting in 0 for **Decimal**. Still in **Variable View**, put the names **cum_bin** and **bin_prob** in new rows under **Name**, and set **Width** to 12, **Decimal** to 10, and **Columns** to 12 for each of these.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	number	Numeric	8	0		None	None	8	Right
2	cum_bin	Numeric	12	10		None	None	12	Right
3	bin_prob	Numeric	12	10		None	None	12	Right
4									

Then click back to **Data View**. From the menu, choose **Transform>Compute...** When the **Compute Variable** window comes up, click **Reset**, and type **cum_bin** in the box labeled **Target Variable**. Scroll down the **Functions** window to **CDF.BINOM(q,n,p)** to select it and press the up arrow. We need to fill in the three arguments indicated by question marks. The first is the x . This is given by the **number** column. At this point, the first question mark should be highlighted. Click on **number** in the box on the left to highlight it,



then hit the right arrow to the right of that box. Now highlight the second question mark and type in 15 (our n), and then highlight the third question mark and type in .75 (our p). Then hit **OK**. If you get a message about changing the existing variable, hit **OK** for that too. The cumulative binomial probabilities are now found in the column **cum_bin**. Now we want to put the individual binomial probabilities into the column **bin_prob**. Do basically the same as the above, except make the **Target Variable** “**bin_prob**,” and the **Numeric Expression** “**CDF.BINOM(number,15,.75) - CDF.BINOM(number-1,15,.75)**.” The Data

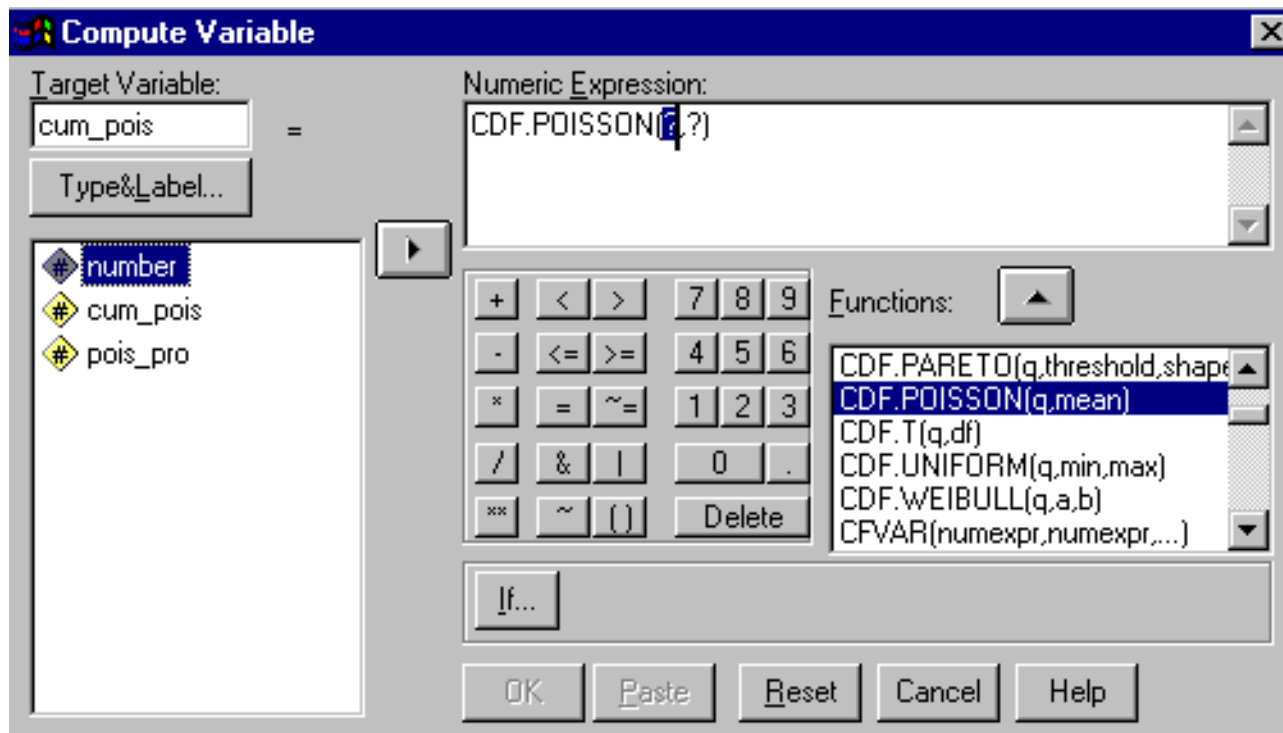
View now looks like the table below, with the cumulative binomial probabilities in the second column and the individual binomial probabilities in the third column.

	number	cum_bin	bin_prob
1	0	.0000000009	.0000000009
2	1	.0000000428	.0000000419
3	2	.0000009229	.0000008801
4	3	.0000123642	.0000114413
5	4	.0001153359	.0001029717
6	5	.0007949490	.0006796131
7	6	.0041930145	.0033980655
8	7	.0172998384	.0131068239
9	8	.0566203101	.0393204717
10	9	.1483680774	.0917477673
11	10	.3135140585	.1651459811
12	11	.5387131236	.2251990652
13	12	.7639121888	.2251990652
14	13	.9198192339	.1559070451
15	14	.9866365390	.0668173051
16	15	1.0000000000	.0133634610

Poisson Distribution. Let us assume that $\lambda = .5$. We will first find $P(X \leq x|.5)$ for $x = 0, \dots, 15$, i.e., the cumulative probabilities. First put the numbers **0** through **15** in a column of a worksheet. (Actually, you only need to enter the numbers whose cumulative probability you desire.) Then click **Variable View**, type in **number** (the name you choose is optional) under **Name**, and I suggest putting in **0** for Decimal. Still in **Variable View**, put the names **cum_pois** and **pois_pro** in new rows under **Name**, and set **Width** to **12**, **Decimal** to **10**, and **Columns** to **12** for each of these.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns
1	number	Numeric	8	0		None	None	8
2	cum_pois	Numeric	12	10		None	None	12
3	pois_pro	Numeric	12	10		None	None	12

Then click back to **Data View**. From the menu, choose **Transform>Compute...** When the **Compute Variable** window comes up, click **Reset**, then type **cum_pois** in the box labeled **Target Variable**. Then scroll down the **Functions** window to **CDF.POISSon(q,mean)** to select it and press the up arrow. We need to fill in the two arguments indicated by question marks. The first is the x . That is given by the **number** column. At this point, the first question mark should be highlighted. Click on **number** in the box on the left to highlight it,



then hit the right arrow to the right of that box. Now highlight the second question mark and type in `.5` (our λ). Then hit **OK**. If you get a message about changing the existing variable, hit **OK** for that too. The cumulative Poisson probabilities are now found in the column `cum_pois`.

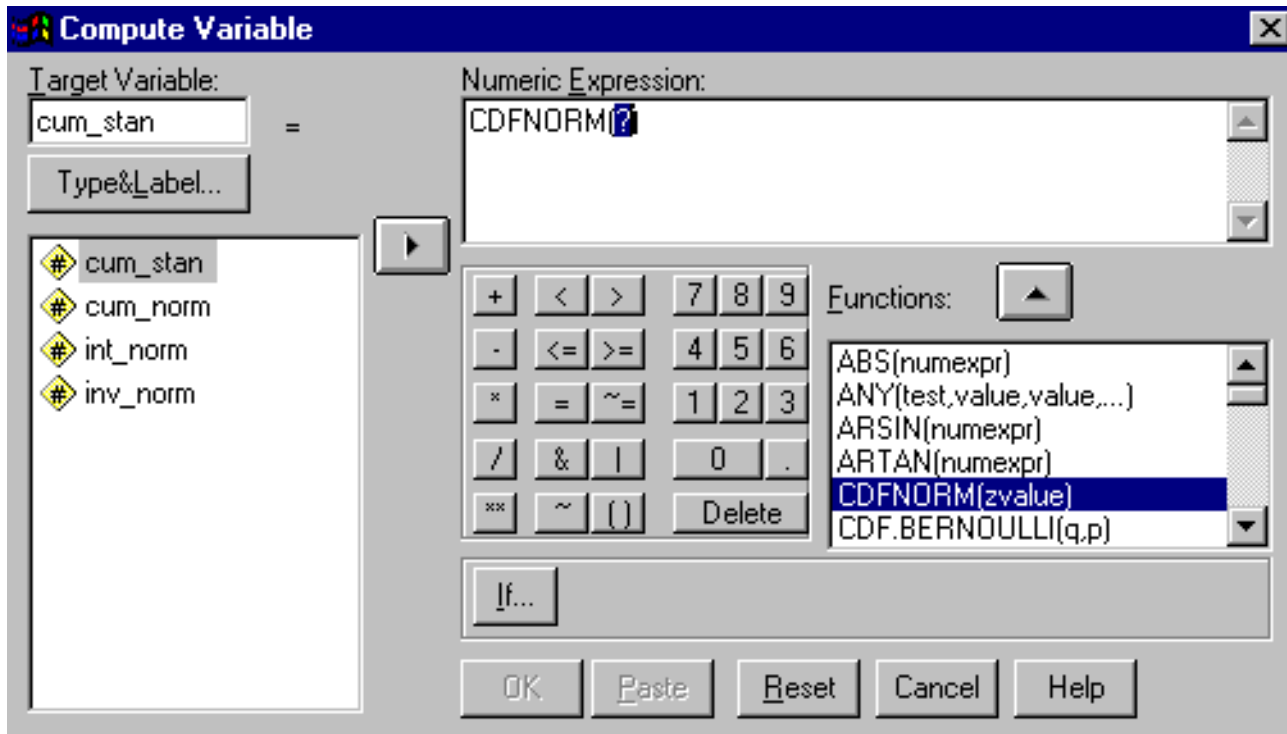
Now we want to put the individual Poisson probabilities into the column `pois_pro`. Do basically the same as up above, except make the **Target Variable** “`pois_pro`”, and the **Numeric Expression** “`CDF.POISSON(number,.5) - CDF.POISSON(number-1,.5)`.” The **Data View** now looks like the table below, with the cumulative Poisson probabilities in the second column and the individual Poisson probabilities in the third column.

	number	cum_pois	pois_pro
1	0	.6065306597	.6065306597
2	1	.9097959896	.3032653299
3	2	.9856123220	.0758163325
4	3	.9982483774	.0126360554
5	4	.9998278844	.0015795069
6	5	.9999858351	.0001579507
7	6	.9999989976	.0000131626
8	7	.9999999378	.0000009402
9	8	.9999999966	.0000000588
10	9	.9999999998	.0000000033
11	10	1.0000000000	.0000000002
12	11	1.0000000000	.0000000000

Normal Distribution. Suppose we wish to find $P(Z \leq 1.5)$. Here we are assuming the standard normal distribution. Start a new **Data Editor** sheet, and just type 0 in the first row of the first column and then hit **Enter**. Then click **Variable View**, put the names **cum_stan**, **cum_norm**, **int_norm**, and **inv_norm** in new rows under **Name**, and set **Decimal** to 4 for each of these.

	Name	Type	Width	Decimals
1	cum_stan	Numeric	8	4
2	cum_norm	Numeric	8	4
3	int_norm	Numeric	8	4
4	inv_norm	Numeric	8	4

Then click back to **Data View**. From the menu, choose **Transform>Compute...** When the **Compute Variable** window comes up, click **Reset**, then type **cum_stan** in the box labeled **Target Variable**. Then scroll down the **Functions** window to **CDFNorm(zvalue)** to select it and press the up arrow. We need to fill in the argument indicated by question mark.



Type in 1.5 (our z). Then hit **OK**. If you get a message about changing the existing variable, hit **OK** for that too. The cumulative standard normal probability is now found in the column **cum_stan**.

Now suppose we are using a normal distribution with mean 100 and standard deviation 20 and we wish to find $P(X \leq 135)$. Do as above except make the **Target Variable** “**cum_norm**,” and the **Numeric Expression** “**CDF.NORMAL(135,100,20)**.” The probability is now found in the column **cum_norm**.

Staying with the normal distribution with mean 100 and standard deviation 20, suppose we wish to find $P(90 \leq X \leq 135)$. Do as above except make the **Target Variable** “**int_norm**,” and the **Numeric Expression** “**CDF.NORMAL(135,100,20) - CDF.NORMAL(90,100,20)**.” The probability is now found in the column **int_norm**.

Continuing to use a normal distribution with mean 100 and standard deviation 20, suppose we wish to find x such that $P(X \leq x) = .6523$. Again, do as above except make the **Target Variable** “inv_norm,” and the **Numeric Expression** “IDF.NORMAL(.6523,100,20).” The x -value is now found in the column inv_norm. From the table below we see that for the standard normal distribution, $P(X \leq 1.5) = .9332$, and for the normal distribution with mean 100 and standard deviation 20, $P(X \leq 135) = .9599$ and $P(90 \leq X \leq 135) = .6514$. Finally, for this latter distribution, if $P(X \leq x) = .6523$ then $x = 107.8307$.

	cum_stan	cum_norm	int_norm	inv_norm
1	.9332	.9599	.6514	107.8307

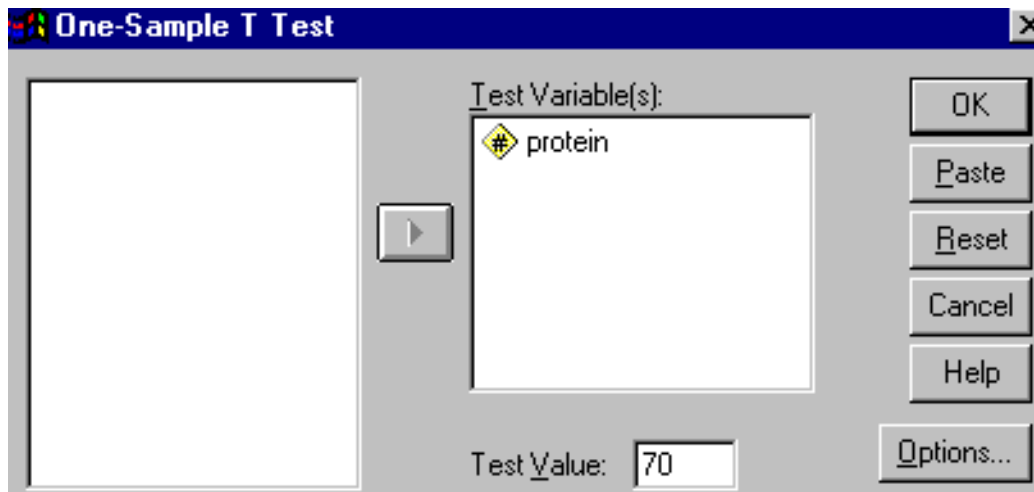
Confidence Intervals and Hypothesis Testing Using t

A Single Population Mean. We found earlier that the sample mean of the data given on page 2, which you may have saved under the name **protein.sav**, is 73.3 to one decimal place. We wish to test whether the mean of the population from which the sample came is 70 as opposed to a true mean greater than 70. We test

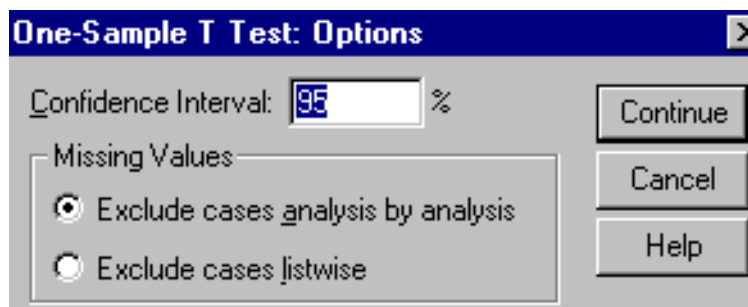
$$H_0 : \mu = 70$$

$$H_a : \mu > 70.$$

From the menu, choose **Analyze>Compare Means>One-Sample T Test**. Select **protein** from the left-hand window and click the right arrow to move it to the **Test Variable(s)** window. Set the **Test Value** to 70.



Click on **Options**. Set the **Confidence Interval** to 95% (or any other value you desire).



Then click **Continue** followed by **OK**. You get the following output.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
PROTEIN	61	73.3292	23.4295	2.9998

One-Sample Test

	Test Value = 70					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
PROTEIN	1.110	60	.272	3.3292	-2.6714	9.3298

SPSS gives us the basic descriptives in the first table. In the second table, we are given that the t -value for our test is **1.110**. The p -value (or **Sig. (2-tailed)**) is given as **.272**. Thus the p -value for our one-tailed test is one-half of that or **.136**. Based on this test statistic, we would not reject the null hypothesis, for instance, for a value of $\alpha = .05$. SPSS also gives us the **95% Confidence Interval of the Difference** between our data scores and the hypothesized mean of **70**, namely **(-2.6714, 9.3298)**. Adding the hypothesized value of **70** to both numbers gives us a 95% confidence interval for the mean of **(67.3286, 79.3298)**. If you are only interested in the confidence interval from the beginning, you can just set the **Test Value** to **0** instead of **70**.

The Difference Between Two Population means. For a data set, we are going to look at a distribution of 32 cadmium level readings from the placenta tissue of mothers, 14 of whom were smokers. The scores are as follows:

non-smokers

10.0 8.4 12.8 25.0 11.8 9.8 12.5 15.4 23.5 9.4 25.1 19.5 25.5 9.8 7.5 11.8 12.2 15.0

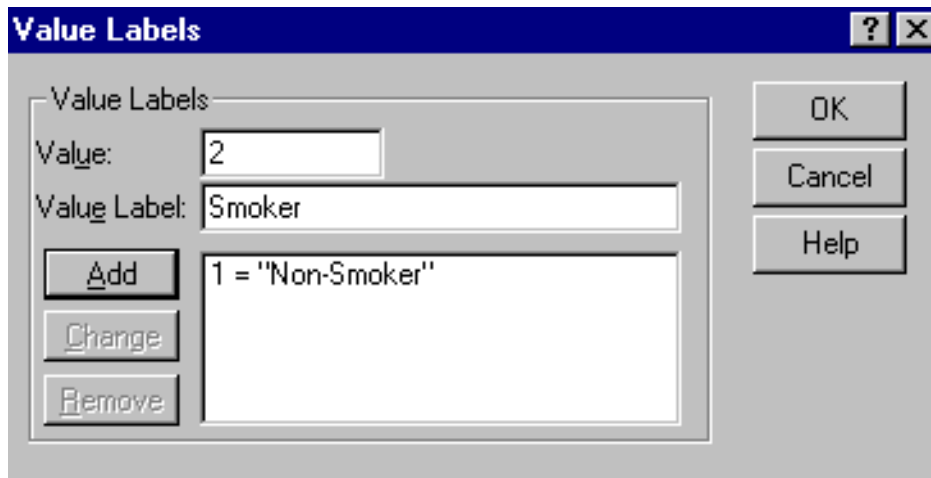
smokers

30.0 30.1 15.0 24.1 30.5 17.8 16.8 14.8 13.4 28.5 17.5 14.4 12.5 20.4

We enter this data in two columns of the **Data Editor**. The first column, which is labeled **s_ns**, contains a **1** for each non-smoking score and a **2** for each smoking score. The scores are contained in the second column, which is labeled **cadmium**. Clicking **Variable View**, we put **s_ns** for the name of the first column, change **Decimals** to **0**, and type in **Smoker** for **Label**. Double-click on the three dots following **None**,

	Name	Type	Width	Decimals	Label	Values	Missing	C
1	s_ns	Numeric	8	0	Smoker	None ...	None	8

and in the window that opens, type **1** for **Value**, **Non-Smoker** for **Value Label**, and then press **Add**. Then type **2** for **Value**, **Smoker** for **Value Label**,



and again press **Add**. Then hit **OK** and complete the **Variable View** as follows.

	Name	Type	Width	Decimals	Label	Values	Missing
1	s_ns	Numeric	8	0	Smoker	{1, Non-Smoker}	None
2	cadmium	Numeric	8	1		None	None

Returning to **Data View** gives a window whose beginning looks like that below.

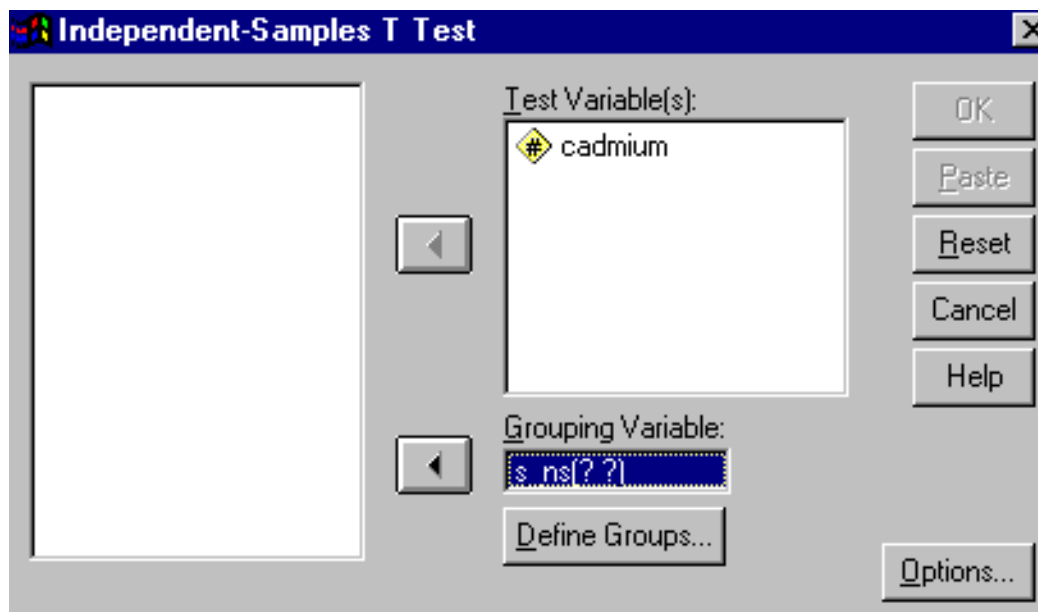
	s_ns	cadmium
1	1	10.0
2	1	8.4
3	1	12.8

Now we wish to test the hypotheses

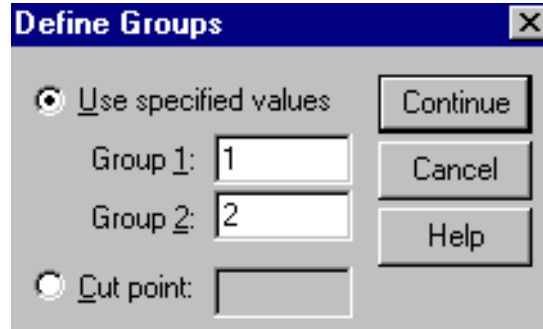
$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

where μ_1 refers to the population mean for the non-smokers and μ_2 refers to the population mean for the smokers. From the menu, choose **Analyze>Compare Means>Independent-Samples T Test**, and in the window that comes up, move **cadmium** to the **Test Variable(s)** window, and **s_ns** into the **Grouping Variable** window.



Notice the two question marks that appear. Click on **Define Groups...**, put in 1 for Group 1 and 2 for Group 2.



Then click **Continue**. As before, click **Options...**, enter **95** (or any other number) for **Confidence Interval**, and again click **Continue** followed by **OK**. The first table of output gives the descriptives.

Group Statistics

		N	Mean	Std. Deviation	Std. Error Mean
CADMIUM	Nonsmokers	18	14.722	6.199	1.461
	Smokers	14	20.414	6.814	1.821

To get the second table as it appears here, I first double-clicked on the **Independent Samples Test** table, giving it a fuzzy border and bringing us into the table editor, and then chose **Pivot>Transpose Rows and Columns** from the menu.

Independent Samples Test

		CADMIUM	
		Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F	.461	
	Sig.	.502	
t-test for Equality of Means	t	-2.468	-2.438
	df	30	26.671
	Sig. (2-tailed)	.020	.022
	Mean Difference	-5.692	-5.692
	Std. Error Difference	2.306	2.335
	95% Confidence Interval of the Difference	Lower	-10.403
	Upper	-.982	-.899

In interpreting the data, the first thing we need to determine is whether we are assuming equal variances. **Levene's Test for Equality of Variances** is an aid in this regard. Since the p -value of **Levene's test** is $p = .502$ for a null hypothesis of all variances equal, in the absence of other information we have no strong evidence to discount this hypothesis, so we will take our results from the **Equal Variances Assumed** column. We see that, with 30 degrees of freedom, we have that $t = -2.468$ and $p = .020$, so we would reject the null

hypothesis $H_0 : \mu_1 - \mu_2 = 0$ at the $\alpha = .05$ level of significance. That we would reject this null hypothesis can also be seen in that the **95% Confidence Interval of the Difference** of $(-10.403, -.982)$ does not contain 0. However, we would not reject the null hypothesis at the $\alpha = .01$ level of significance and, correspondingly, the **99% Confidence Interval of the Difference**, had we chosen that level, would contain 0.

Paired Comparisons. We consider the weights (in kg) of 9 women before and after 12 weeks on a special diet, with the goal of determining whether the diet aids in weight reduction. The paired data is given below.

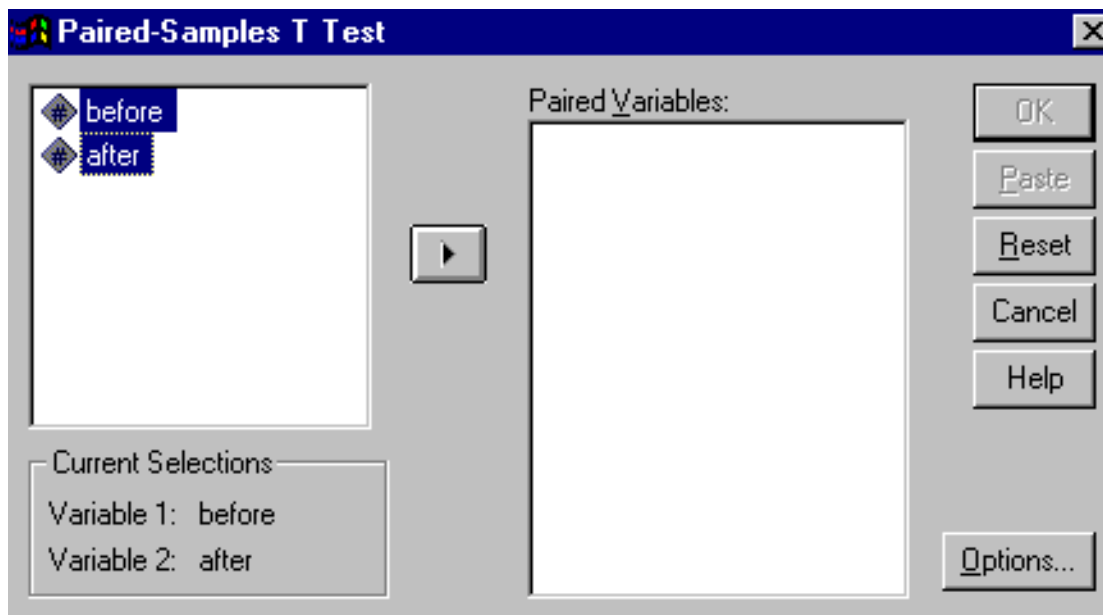
Before	117.3	111.4	98.6	104.3	105.4	100.4	81.7	89.5	78.2
After	83.3	85.9	75.8	82.9	82.3	77.7	62.7	69.0	63.9

We place the **Before** data in the first column of our worksheet and the **After** data in the second column. We wish to test the hypotheses

$$H_0 : \mu_{B-A} = 0$$

$$H_a : \mu_{B-A} > 0$$

with one-sided alternative. From the menu, choose **Analyze>Compare Means>Paired-Samples T Test**. In the window that opens, first click **before** to make it **Variable 1** and then **after** to make it **Variable 2**.



Next click the right arrow to move the pair into the big window to the right, followed by clicking **Options...** to set **Confidence Interval** to **99%**. Then click **Continue** to close the **Options...** window followed by **OK** to get the output.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair	BEFORE	98.533	9	13.134	4.378
1	AFTER	75.944	9	8.759	2.920

The first output table gives the descriptives and a second (not shown here) gives a correlation coefficient. From the third table, which has been pivoted to interchange rows and columns,

Paired Samples Test

		Pair 1
		BEFORE - AFTER
Paired Differences	Mean	22.589
	Std. Deviation	5.319
	Std. Error Mean	1.773
	99% Confidence Interval	Lower
	of the Difference	Upper
t		12.740
df		8
Sig. (2-tailed)		.000

we see that we have a t -score of 12.740. The fact that **Sig.(2-tailed)** is given as .000 really means that it is less than .001. Thus, for our one-sided test, we can conclude that $p < .0005$, so that in almost any situation we would reject the null hypothesis. We also see that the mean of the weight losses for the sample is 22.589, with a **99% Confidence Interval of the Difference** (the mean weight loss for the population from which the sample was drawn) being (16.639, 28.538).

One-Way ANOVA

For data, we will use percent predicted residual volume measurements as categorized by smoking history.

Never 35, 120, 90, 109, 82, 40, 68, 84, 124, 77, 140, 127, 58, 110, 42, 57, 93
 Former 62, 73, 60, 77, 52, 115, 82, 52, 105, 143, 80, 78, 47, 85, 105, 46, 66, 95, 82, 141, 64, 124, 65, 42, 53, 67, 95, 99, 69, 118, 131, 76, 69, 69
 Current 96, 107, 63, 134, 140, 103, 158

We will place the volume measurements in the first column and the second column will be coded by 1 = "Never," 2 = "Former," and 3 = "Current." The **Variable View** looks as below.

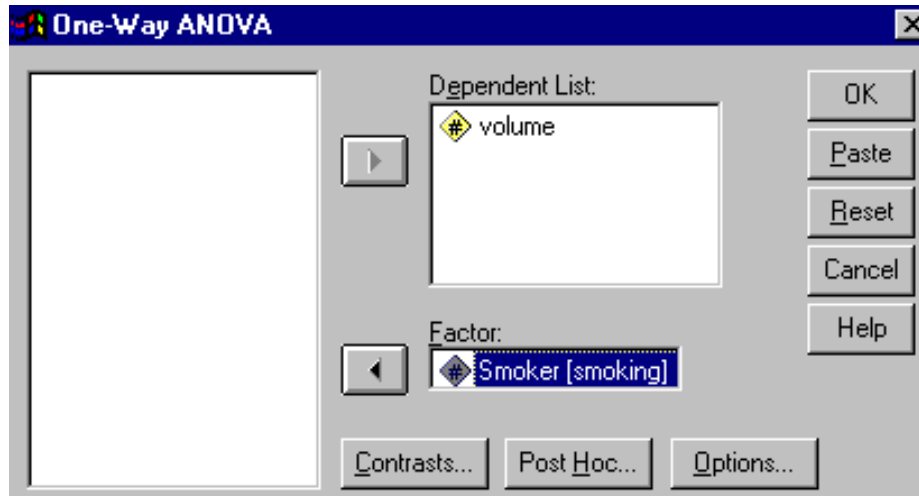
	Name	Type	Width	Decimals	Label	Values
1	volume	Numeric	8	0		None
2	smoking	Numeric	8	0	Smoker	{1, Never}...

We test to see if there is a difference among the population means from which the samples have been drawn. We use the hypotheses

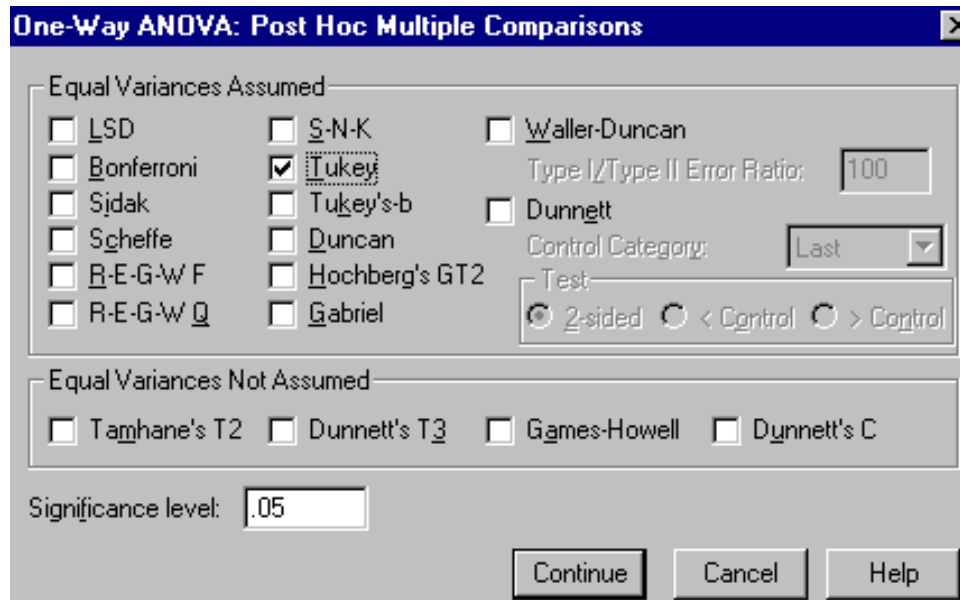
$$H_0 : \mu_N = \mu_F = \mu_C$$

$$H_a : \text{Not all of } \mu_N, \mu_F, \text{ and } \mu_C \text{ are equal.}$$

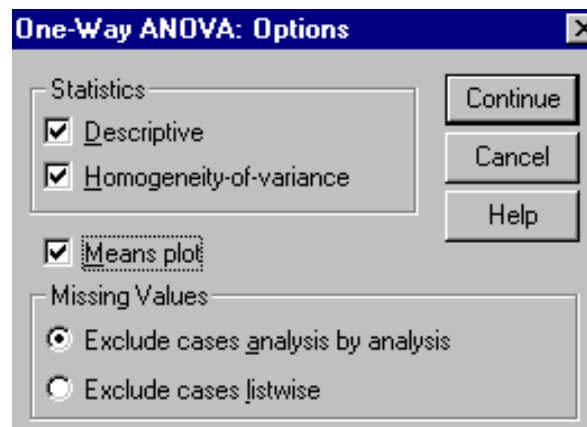
From the menu we choose **Analyze>Compare Means>One-Way ANOVA...** In the window that opens, place **volume** under **Dependent List** and **Smoker[smoking]** under **Factor**.



Then click **Post Hoc...** For a post-hoc test, we will only choose **Tukey** (Tukey's HSD test) with **Significance Level .05**, and then click **Continue**.



Then we click options and choose all three of **Descriptive**, **Homogeneity-of-variance**, and **Means plot**.



Then we click **Continue** followed by **OK** to get our output.

Descriptives

Dependent Variable		VOLUME			
		Never	Former	Current	Total
N		21	44	7	72
Mean		82.14	84.25	114.43	86.57
Std. Deviation		30.44	29.30	31.90	30.86
Std. Error		6.64	4.42	12.06	3.64
95% Confidence Interval for Mean	Lower Bound	68.29	75.34	84.93	79.32
	Upper Bound	96.00	93.16	143.93	93.82
Minimum		35	40	63	35
Maximum		140	151	158	158

A first impression from the **Descriptives** is that the mean of the **Current** smokers differs significantly from those who **Never** smoked and the **Former** smokers, the latter two means being pretty much the same.

Test of Homogeneity of Variances

VOLUME

Levene Statistic	df1	df2	Sig.
.026	2	69	.974

The results of the **Test of Homogeneity of Variances** is nonsignificant since we have a p -value of **.974**, showing that there is no reason to believe that the variances of the three groups are different from one another. This is reassuring since both ANOVA and Tukey's HSD have equal variance assumptions. Without this reassurance, interpretation of the results would be difficult.

ANOVA

VOLUME

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6081.117	2	3040.559	3.409	.039
Within Groups	61542.536	69	891.921		
Total	67623.653	71			

Now we look at the results of the ANOVA itself. The **Sum of Squares Between Groups** is the SSA, the **Sum of Squares Within Groups** is the SSW, the **Total Sum of Squares** is the SST, the **Mean Square Between Groups** is the MSA, the **Mean Square Within Groups** is the MSW, and the **F** value of **3.409** is the Variance Ratio. Since the p -value is **.039**, we will reject the null hypothesis at the $\alpha = .05$ level of significance, concluding that all three population means are not the same, but would not reject it at the $\alpha = .01$ level of significance.

So now the question becomes which of the means significantly differ from the others. For this we look to post-hoc tests. One option which was not chosen was **LSD** (least significant difference) since this simply does a t test on each pair. Here, with three groups we would

test three pairs. But if you have 7 groups, for instance, that is 21 separate t tests, and at an $\alpha = .05$ level of significance, even if all the means are the same, you can expect on the average to get one Type I error where you reject a true null hypothesis for every 20 tests. In other words, while the t test is useful in testing whether two means are the same, it is not the test to use for checking multiple means. That is why we chose ANOVA in the first place. We have chosen Tukey's HSD because it offers adequate protection from Type I errors and is widely used

Multiple Comparisons

Dependent Variable: VOLUME

Tukey HSD

(I) Smoker	(J) Smoker	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Never	Former	-2.11	7.92	.962	-21.08	16.87
	Current	-32.29*	13.03	.041	-63.51	-1.06
Former	Never	2.11	7.92	.962	-16.87	21.08
	Current	-30.18*	12.15	.040	-59.29	-1.07
Current	Never	32.29*	13.03	.041	1.06	63.51
	Former	30.18*	12.15	.040	1.07	59.29

Looking at all of the p -values (Sig.) in the Multiple Comparisons table, we see that **Current** differs significantly ($\alpha = .05$) from **Never** and **Former**, with no significant difference detected between **Never** and **Former**. The second table for Tukey's HSD, seen below, divides the groups into homogeneous subsets and gives the mean for each group.

VOLUME

Tukey HSD^{a,b}

Smoker	N	Subset for alpha = .05	
		1	2
Never	21	82.14	
Former	44	84.25	
Current	7		114.43
Sig.		.981	1.000

Means for groups in homogeneous subsets are displayed.

- a. Uses Harmonic Mean Sample Size = 14.071.
- b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Simple Linear Regression and Correlation

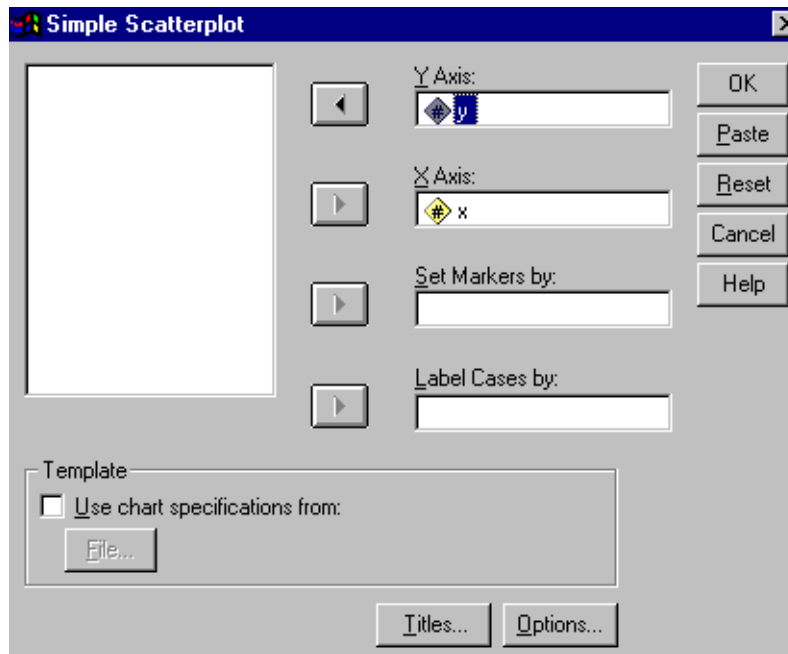
We will use the following 109 x - y data pairs for simple linear regression and correlation.

	x	y		x	y		x	y
1	74.75	25.72	23	79.90	35.43	45	83.00	96.54
2	72.60	25.89	24	89.20	60.09	46	107.10	118.00
3	81.80	42.60	25	82.00	45.84	47	94.30	107.00
4	83.95	42.80	26	92.00	70.40	48	94.50	123.00
5	74.65	29.84	27	86.60	83.45	49	79.70	65.92
6	71.85	21.68	28	80.50	84.30	50	79.30	81.29
7	80.90	29.08	29	86.00	78.89	51	89.80	111.00
8	83.40	32.98	30	82.50	64.75	52	83.80	90.73
9	63.50	11.44	31	83.50	72.56	53	85.20	133.00
10	73.20	32.22	32	88.10	89.31	54	75.50	41.90
11	71.90	28.32	33	90.80	78.94	55	78.40	41.71
12	75.00	43.86	34	89.40	83.55	56	78.60	58.16
13	73.10	38.21	35	102.00	127.00	57	87.80	88.85
14	79.00	42.48	36	94.50	121.00	58	86.30	155.00
15	77.00	30.96	37	91.00	107.00	59	85.50	70.77
16	68.85	55.78	38	103.00	129.00	60	83.70	75.08
17	75.95	43.78	39	80.00	74.02	61	77.60	57.05
18	74.15	33.41	40	79.00	55.48	62	84.90	99.73
19	73.80	43.35	41	83.50	73.13	63	79.80	27.96
20	75.90	29.31	42	76.00	50.50	64	108.30	123.00
21	76.85	36.60	43	80.50	50.88	65	119.60	90.41
22	80.90	40.25	44	86.50	140.00	66	119.90	106.00

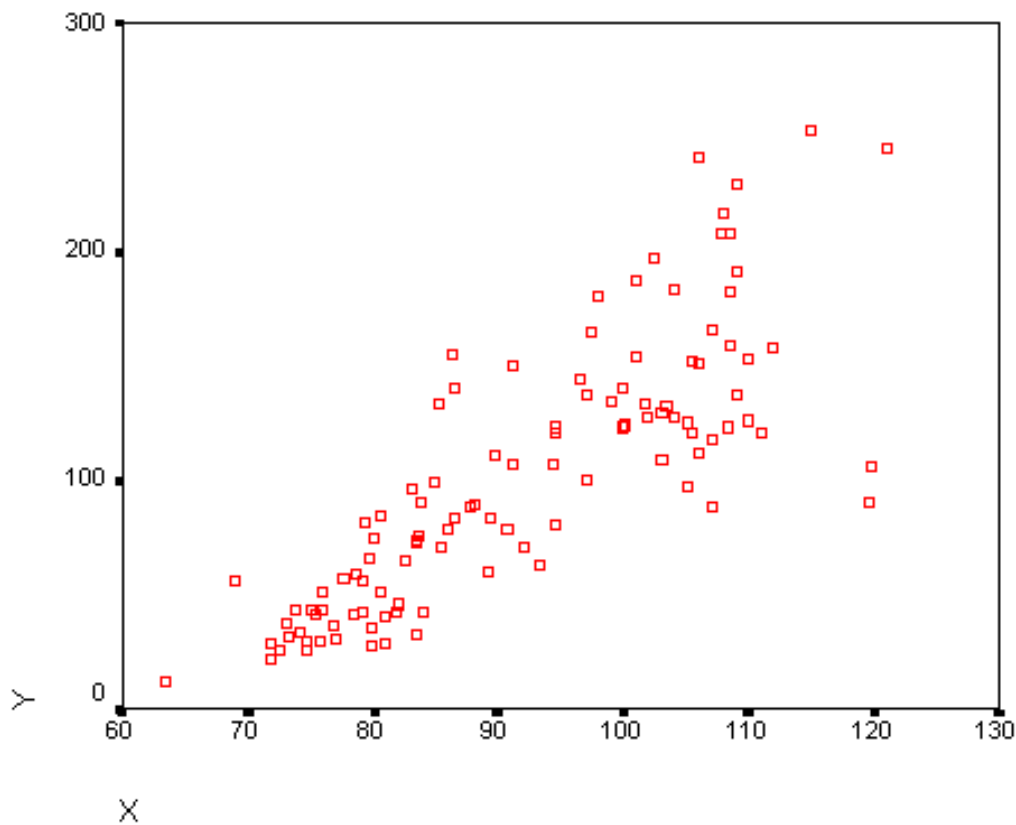
	x	y		x	y
67	96.50	144.00	89	121.00	245.00
68	105.50	121.00	90	109.00	137.00
69	105.00	97.13	91	97.50	165.00
70	107.00	166.00	92	105.50	152.00
71	107.00	87.99	93	98.00	181.00
72	101.00	154.00	94	94.50	80.95
73	97.00	100.00	95	97.00	137.00
74	100.00	123.00	96	105.00	125.00
75	108.00	217.00	97	106.00	241.00
76	100.00	140.00	98	99.00	134.00
77	103.00	109.00	99	91.00	150.00
78	104.00	127.00	100	102.50	198.00
79	106.00	112.00	101	106.00	151.00
80	109.00	192.00	102	109.10	229.00
81	103.50	132.00	103	115.00	253.00
82	110.00	126.00	104	101.00	188.00
83	110.00	153.00	105	100.10	124.00
84	112.00	158.00	106	93.30	62.20
85	108.50	183.00	107	101.80	133.00
86	104.00	184.00	108	107.90	208.00
87	111.00	121.00	109	108.50	208.00
88	108.50	159.00	110		

The x 's are waist circumferences (cm) and the y 's are measurements of deep abdominal adipose tissue gathered by CAT scans. Since CAT scans are expensive, the goal is to find a predictive equation.

First we wish to take a look at the scatter plot of the data, so we choose **Graphs>Scatter...** from the menu. In the **Scatterplot** window that opens, click on **Simple**, followed by **Define**.

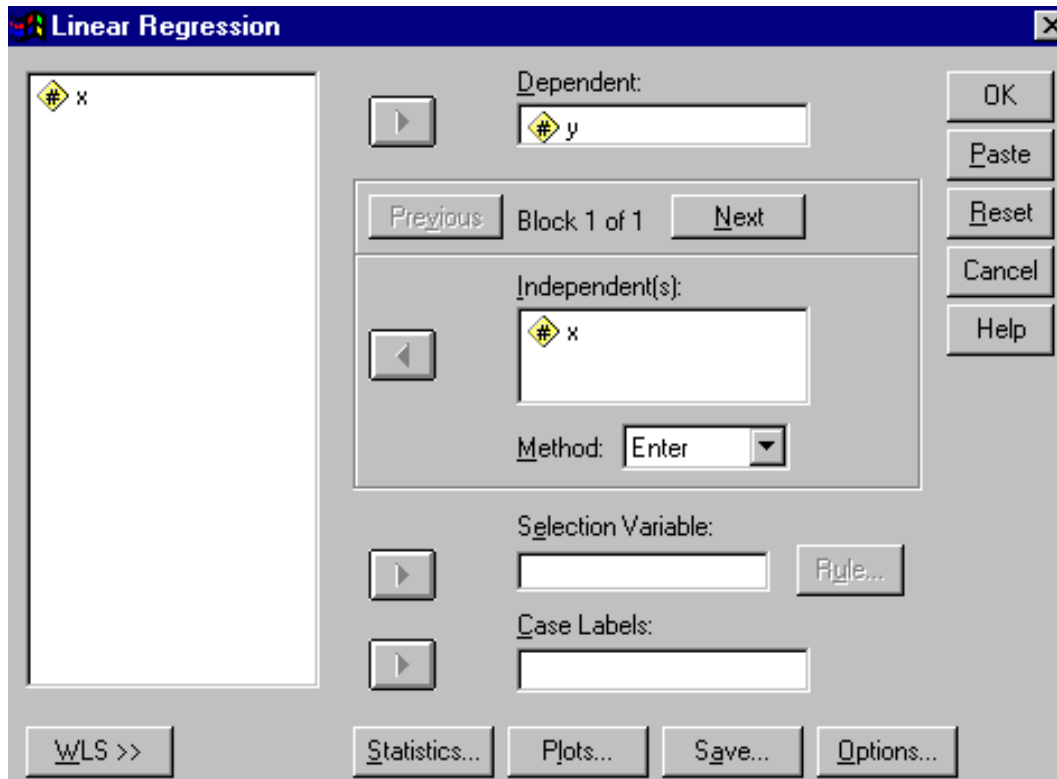


In the **Simple Scatterplot** window that now opens, use selection and the arrows to move y under **Y Axis** and x under **X Axis**. Then click **OK** to get the following scatter plot,

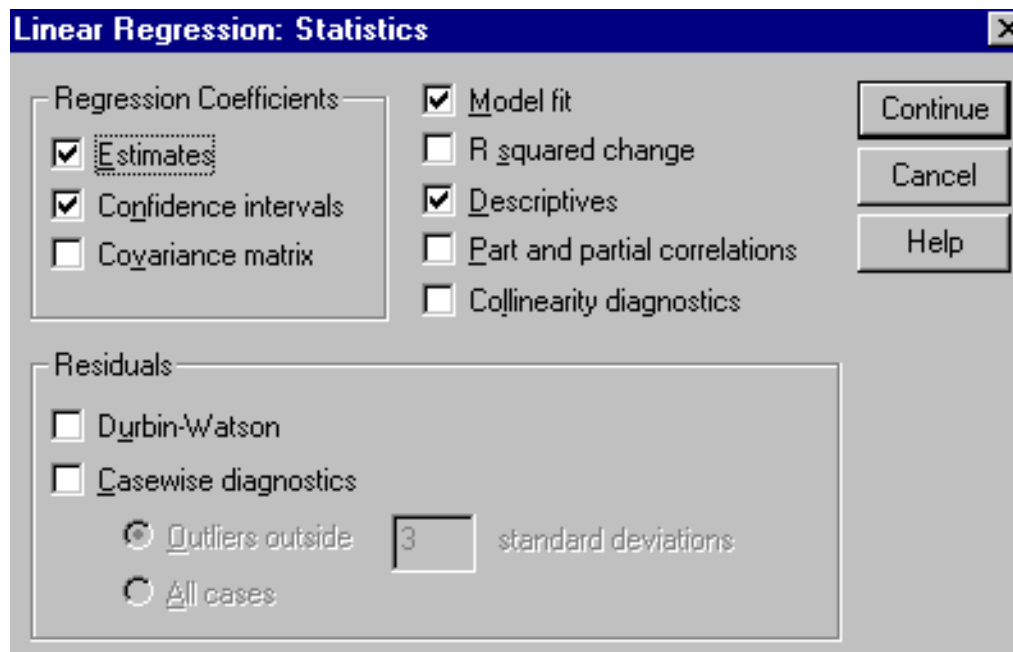


which leads us to suspect that there is a significant linear relationship.

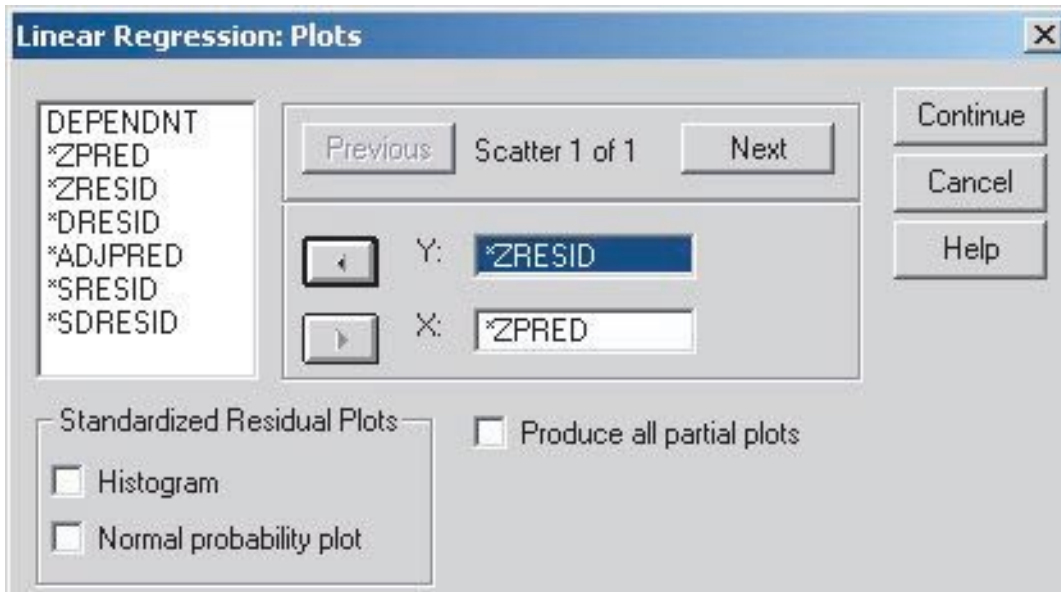
Regression. To explore this relationship, choose **Analyze>Regression>Linear** from the menu, select and move **y** under **Dependent** and **x** under **Independent(s)**.



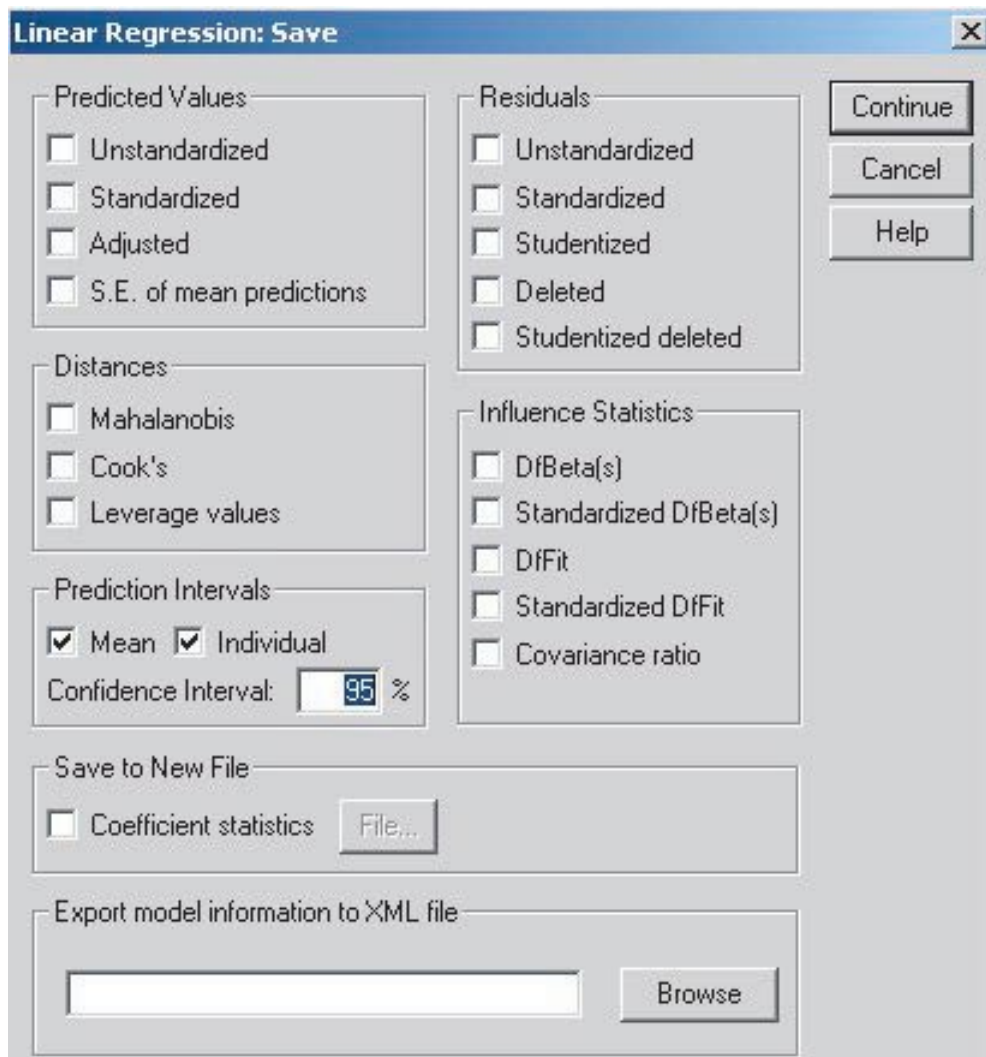
Then click **Statistics...**, and in the window that opens with **Estimates** and **Model fit** already checked, also check **Confidence intervals** and **Descriptives**.



Then click **Continue**. Next click **Plots...** In the window that opens, enter ***ZRESID** for **Y** and ***ZPRED** for **X** to get a graph of the standardized residuals as a function of the standardized predicted values.



After clicking **Continue**, next click **Save...** In the window that opens, check **Mean** and **Individual** under **Prediction Intervals** with **95%** for **Confidence Intervals**. This will add four columns to our data window that give the 95% confidence intervals for the mean values $\mu_{y|x}$ and individual values y_I for each x in our set of data pairs.



Then click **Continue** followed by **OK** to get the output.

Descriptive Statistics

	Mean	Std. Deviation	N
Y	101.8940	57.2948	109
X	91.9018	13.5591	109

We first see the mean and the standard deviation for the two variables in the **Descriptive Statistics**.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.819 ^a	.670	.667	33.0649

a. Predictors: (Constant), X

In the **Model Summary**, we see that the bivariate correlation coefficient r (**R**) is .819, indicating a strong positive linear relationship between the two variables. The coefficient of determination r^2 (**R Square**) of .670 indicates that, for the sample, 67% of the variation of y can be explained by the variation in x . But this may be an overestimate for the population from which the sample is drawn, so we use the **Adjusted R Square** as a better estimate for the population. Finally, the **Standard Error of the Estimate** is 33.0649.

Coefficients^a

		Model	
		1	
		(Constant)	X
Unstandardized Coefficients	B	-215.981	3.459
	Std. Error	21.796	.235
Standardized Coefficients	Beta		.819
t		-9.909	14.740
Sig.		.000	.000
95% Confidence Interval for B	Lower Bound	-259.190	2.994
	Upper Bound	-172.773	3.924

a. Dependent Variable: Y

We use the sample regression (least squares) equation $\hat{y} = a + bx$ to approximate the population regression equation $\mu_{y|x} = \alpha + \beta x$. From the **Coefficients** table, a is **-215.981** and b is **3.459** from the first row of numbers (rows and columns transposed from the output), so the sample regression equation is $\hat{y} = -215.981 + 3.459x$. From the last two rows of numbers in the table, one gets that 95% confidence intervals for α and β are **(-259.190, -172.773)** and **(2.994, 3.924)**, respectively.

The t test is used for testing the null hypothesis $\beta = 0$, for if $\beta = 0$, the sample regression equation will have little value for prediction and estimation. It can be used similarly to test the null hypothesis $\alpha = 0$, but this is of much less interest. In this case, we read from the above table that for $H_0 : \beta = 0$, $H_a : \beta \neq 0$, we have $t = 14.740$. Since the p -value (**Sig. = .000**) for that t test is less than .001, we can reject the null hypothesis of $\beta = 0$.

Although the ANOVA table is more properly used in multiple regression for testing the null hypothesis $\beta_1 = \beta_2 = \dots = \beta_n = 0$ with an alternative hypothesis of not all $\beta_i = 0$, it can also be used to test $\beta = 0$ in simple linear regression. In the table below, the **Regression Sum of Squares** (SSR) is the variation explained by regression, and the **Residual Sum of Squares** (SSE) is the variation not explained by regression (the “E” stands for error). The **Mean Square Regression** and the **Mean Square Residual** are MSR and MSE respectively, with the **F** value of **217.279** being their quotient. Since the *p*-value (**Sig. = .000**) is less than .001, we can reject the null hypothesis of $\beta = 0$.

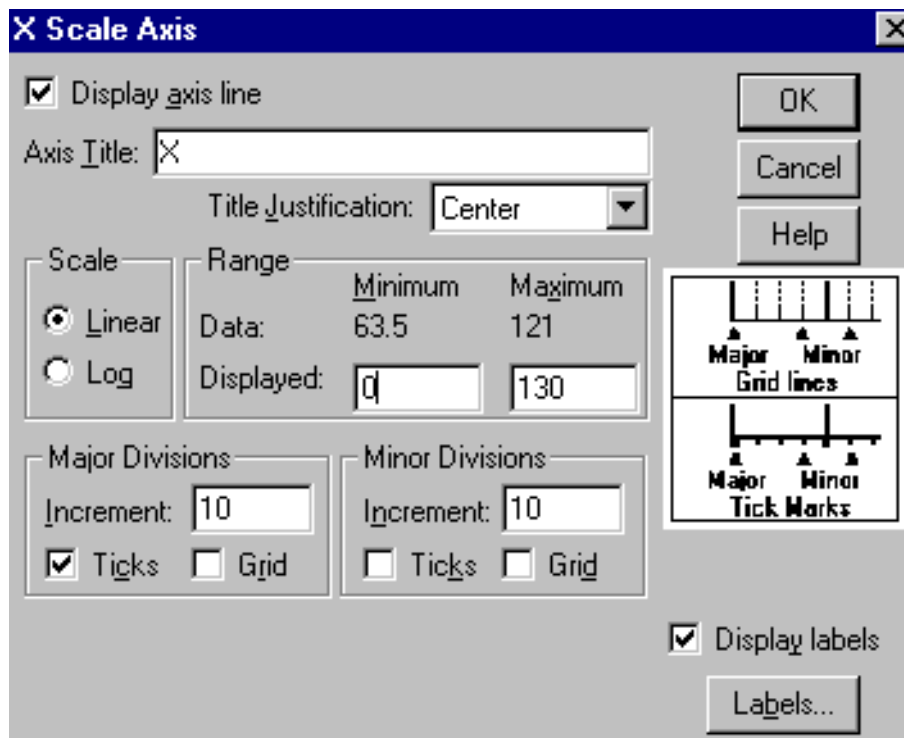
ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	237548.5	1	237548.516	217.279	.000 ^a
	Residual	116982.0	107	1093.290		
	Total	354530.5	108			

a. Predictors: (Constant), X

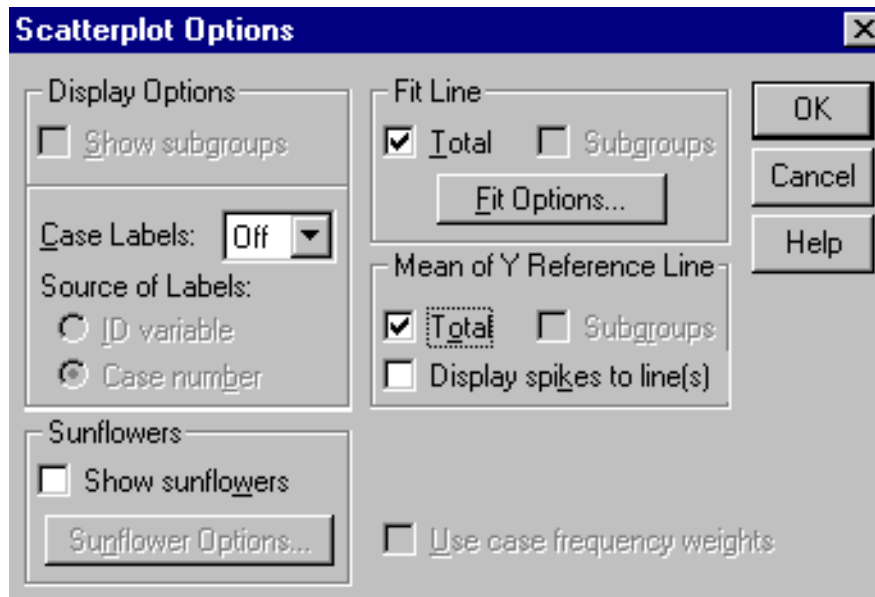
b. Dependent Variable: Y

We now return to the scatter plot. Double click on the plot to bring up the **Chart Editor** and choose **Chart>Axis...** from the menu. In the window that opens, select **X scale** and click **OK**.

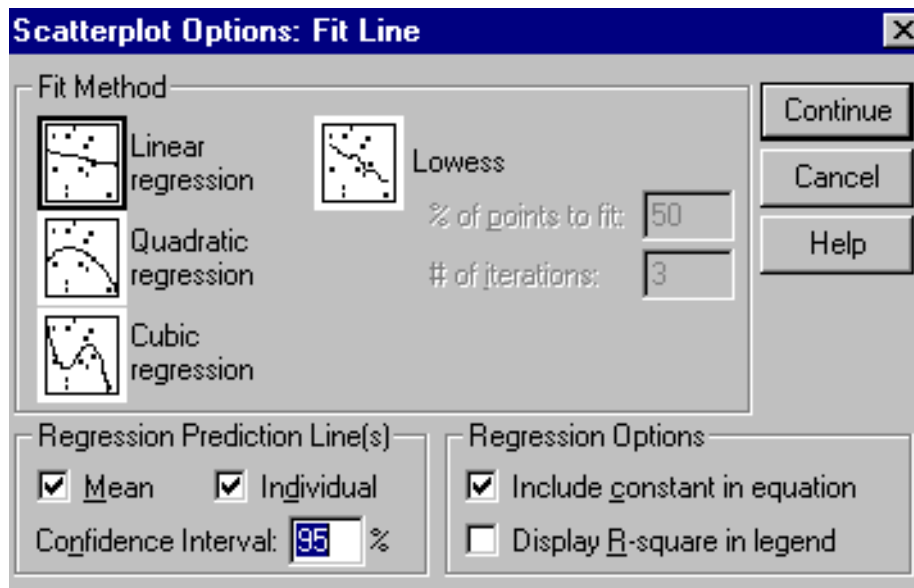


Choose **Center** for **Title Justification** and change the **Displayed Minimum** to 0. Then click **OK**. Repeat the same for the **Y scale** axis, except change the **Displayed Minimum** to -300 and the **Major Divisions Increment** to 40. Click **OK** to close this window.

Next select Chart>Options... from the menu and, in the window that appears,

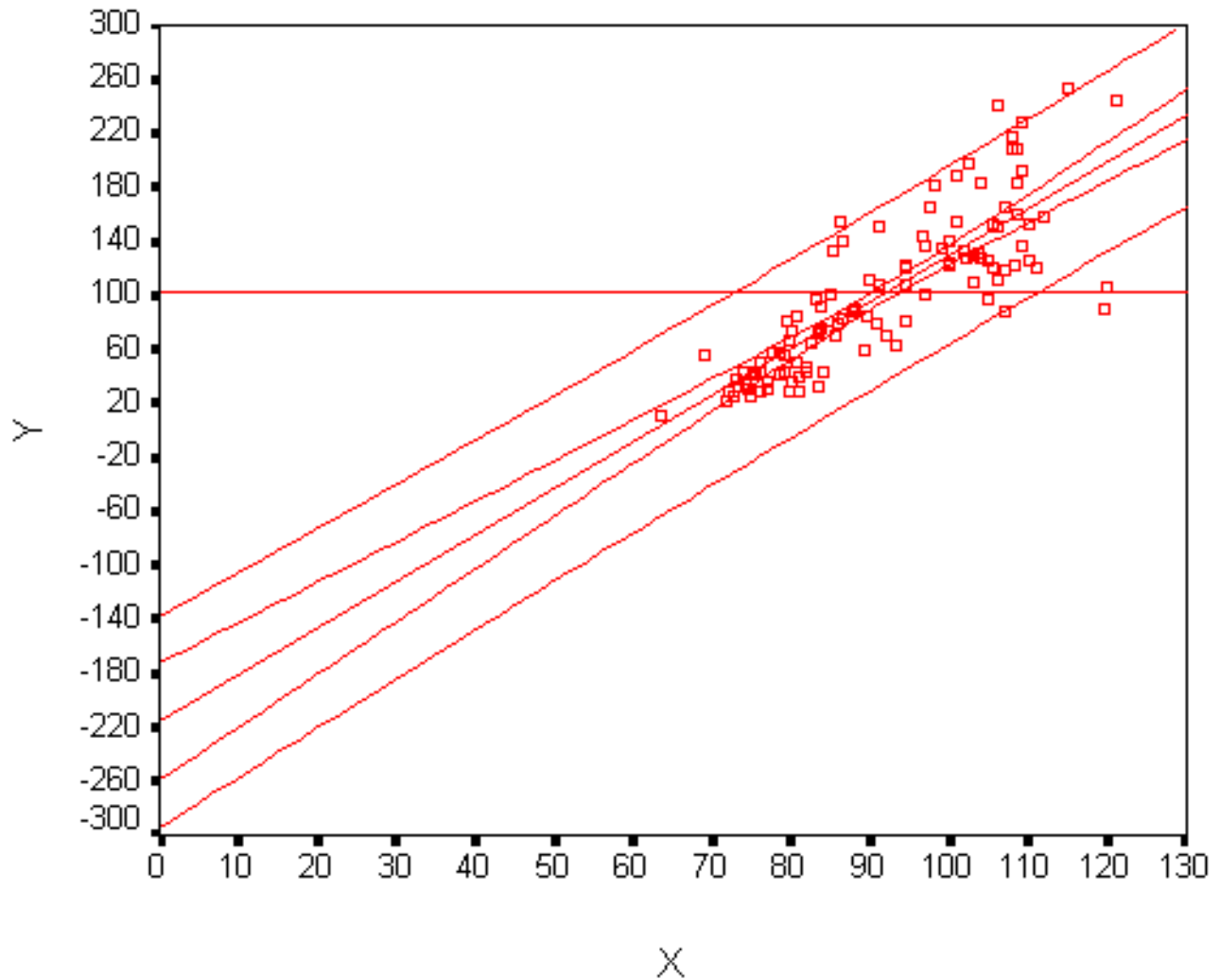


check Total under both Fit Line and Mean of Y Reference Line. Then click on Fit Options....



In the new window that opens, make sure **Linear regression** is selected. If we were doing **Quadratic regression** or **Cubic regression**, we would check those instead. This places the regression line or curve on the scatter plot. Now click **Mean** and **Individual** under **Regression Prediction Lines**, at the same time making sure **Confidence Interval** is set to **95%**. This relates to the prediction intervals saved earlier. Then click **Continue** followed by **OK** to get the edited scatter plot, found on the next page.

In this graph, the horizontal line shows the mean of the y -values, 101.894. We see that the scatter about the regression line is much less than the scatter about the mean line, which is as it should be when the null hypothesis $\beta = 0$ has been rejected. The inner bands about the regression line give the 95% confidence interval for the mean values $\mu_{y|x}$ for each x , or from another point of view, the probability is .95 that the population regression line $\mu_{y|x} = \alpha + \beta x$ lies within these bands. The outer bands give the 95% confidence interval for the individual y_I for each value of x .

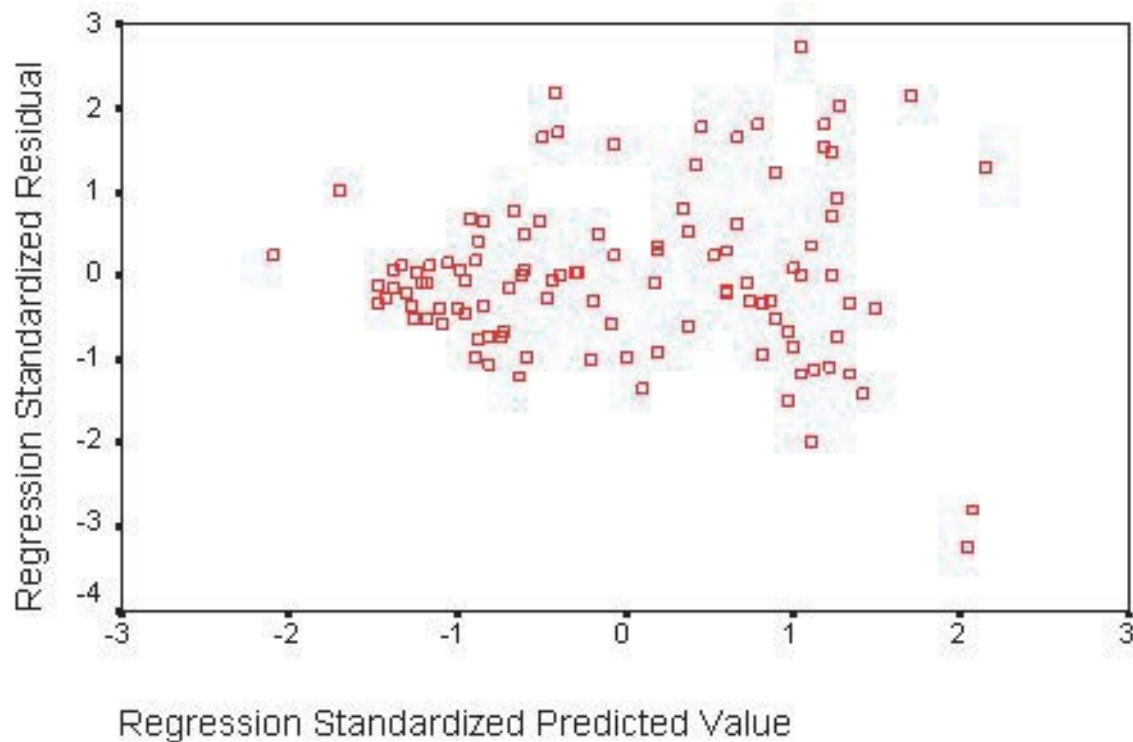


The confidence bands in the above scatter plot relate to the four new columns in our data window, a portion of which is shown below. We interpret the first row of data. For $x = 74.5$, the 95% confidence interval for the mean value $\mu_{y|74.5}$ is (32.41572, 52.72078), corresponding to the limits of the inner bands at $x = 74.5$ in the scatter plot, and the 95% confidence interval for the individual value $y_I(74.5)$ is (-23.7607, 108.8972), corresponding to the limits of the outer bands at $x = 74.5$. The first pair of acronyms **lmci** and **umci** stand for “lower mean confidence interval” and “upper mean confidence interval,” respectively, with the **i** in the second pair standing for “individual.”

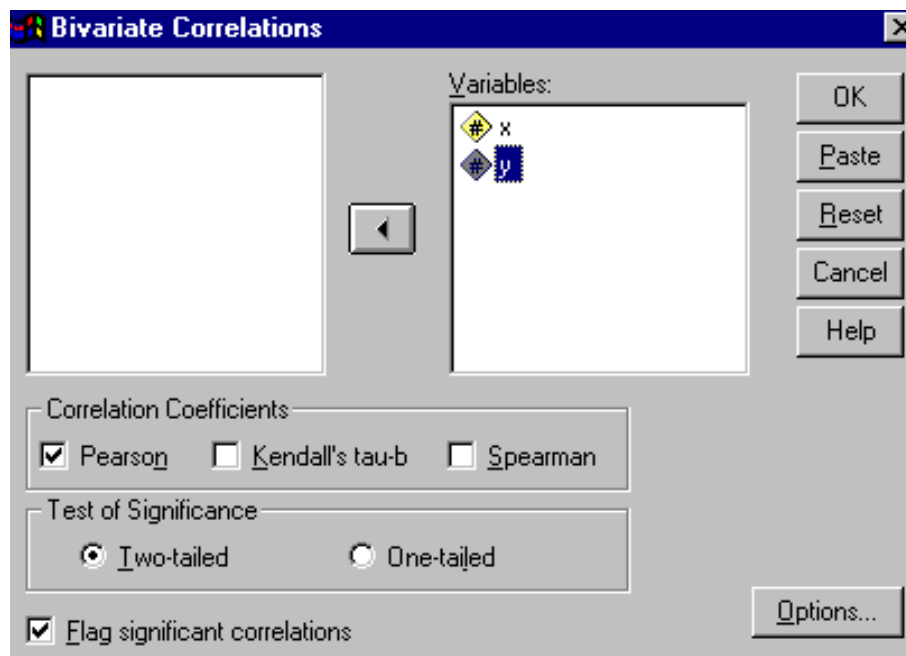
	x	y	lmci_1	umci_1	lici_1	uici_1
1	74.75	25.72	32.41572	52.72078	-23.7607	108.8972
2	72.60	25.89	24.17575	46.08766	-31.3250	101.5884
3	81.80	42.60	59.11112	74.79530	.93840	132.9680
4	83.95	42.80	67.10283	81.67669	8.43859	140.3409

Finally, consider the residual plot at the top of the next page. On the horizontal axis are the standardized y values from the data pairs, and on the vertical axis are the standardized residuals for each such y . If all the regression assumptions were met for our data set, we would expect to see random scattering about the horizontal line at level 0 with no noticeable

patterns. However, here we see more spread for the larger values of y , bringing into question whether the assumption regarding equal standard deviations for each y population is met.



Correlation. Choose **Analyze>Correlate>Bivariate...** from the menu to study the correlation of the two variables x and y .



In the window that opens, move both x and y to the **Variables** window and make sure **Pearson** is selected. The other two choices are for nonparametric correlations. We will choose **Two-tailed** here since we already have the results of the **One-tailed** option in the **Correlation** table in the regression output. In general, you choose **One-tailed** if you know

the direction of correlation (positive or negative), and **Two-tailed** if you do not. Clicking **OK** gives the results.

Correlations

		X	Y
X	Pearson Correlation	1.000	.819**
	Sig. (2-tailed)	.	.000
	N	109	109
Y	Pearson Correlation	.819**	1.000
	Sig. (2-tailed)	.000	.
	N	109	109

** . Correlation is significant at the 0.01 level

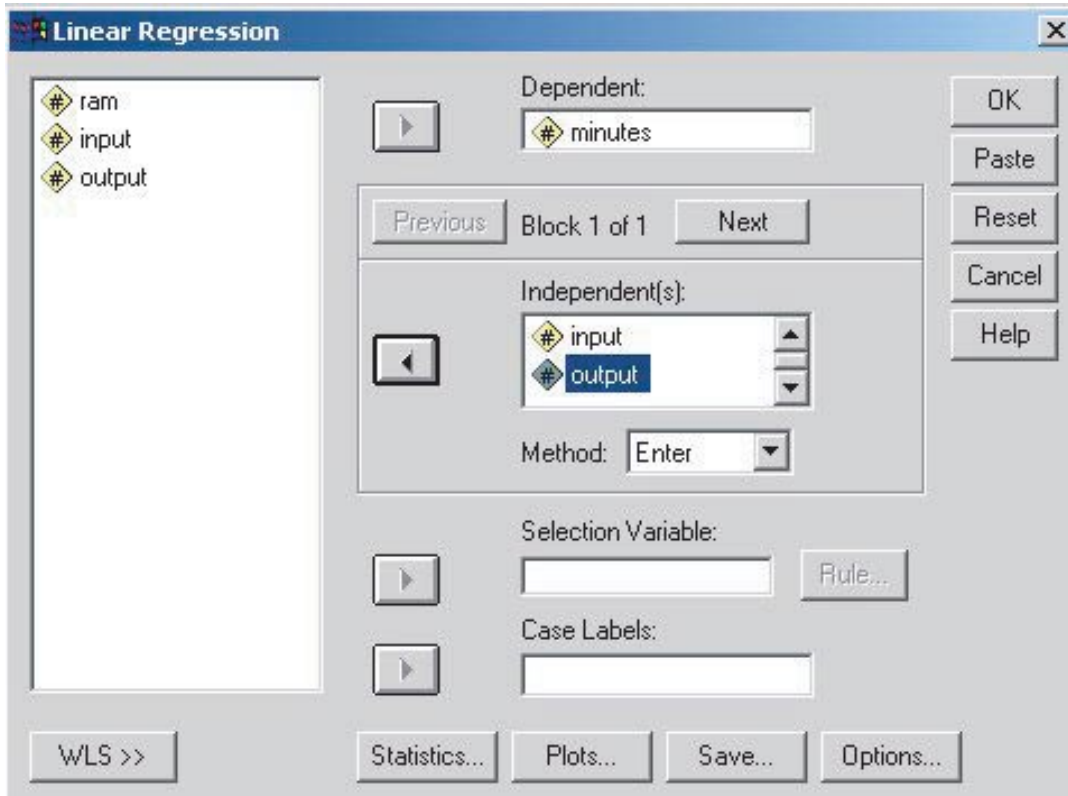
We see again that the **Pearson Correlation** r is .819, and from the **Sig.** of .000, we know that the p -value is less than .001 and so we would reject a null hypothesis of $r = 0$.

Multiple Regression

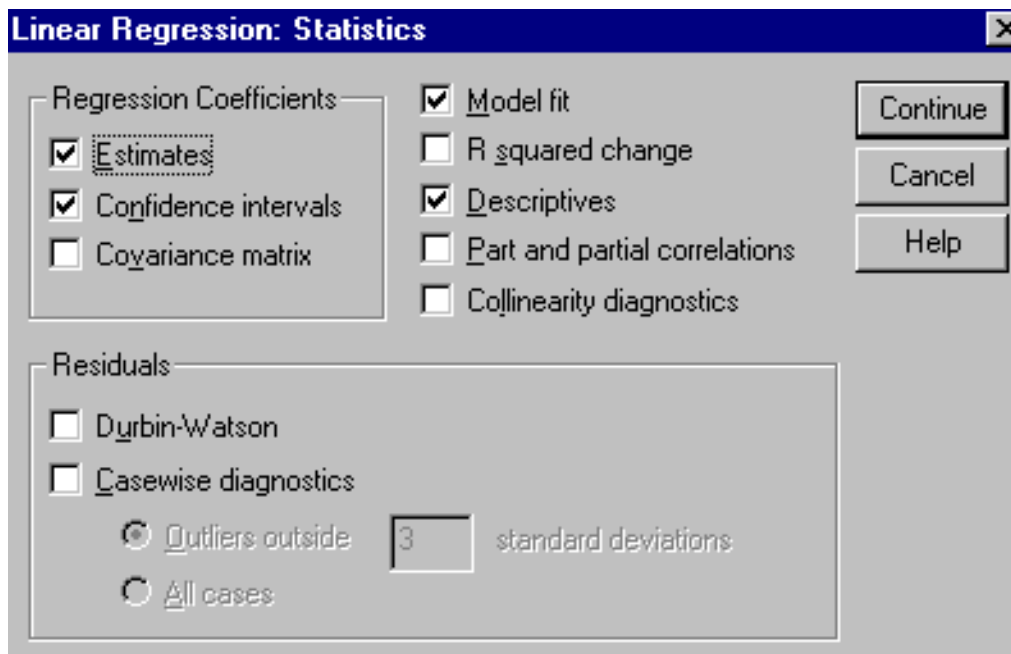
We will use the following data set for multiple linear regression. In this data set, required **ram**, amount of **input**, and amount of **output**, all in kilobytes, are used to predict **minutes** of processing time for a given task. From left to right, we will use the variables y , x_1 , x_2 , and x_3 . Overall, the process used parallels that of simple linear regression.

	minutes	ram	input	output
1	5.2	19	5	1
2	17.3	105	10	2
3	15.5	70	15	5
4	23.4	80	20	8
5	15.4	24	12	10
6	9.5	15	2	5
7	6.2	22	3	4
8	10.0	35	10	3
9	7.7	42	5	2
10	6.3	15	2	2
11	7.2	8	4	5
12	8.5	7	5	6
13	8.9	12	10	3
14	5.6	15	7	2
15	4.1	17	4	1
16	9.7	18	3	6
17	13.4	24	8	5
18	11.7	25	8	4
19	8.4	32	10	3
20	12.1	79	12	2

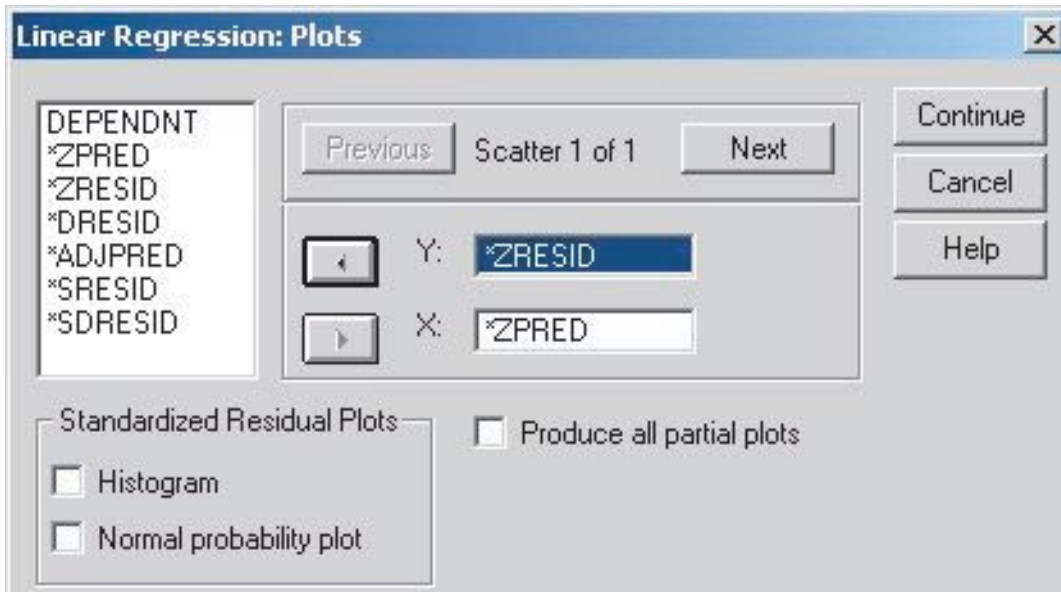
Regression. Choose Analyze>Regression>Linear from the menu, select and move minutes under Dependent and ram, input, and output, in that order, under Independent(s).



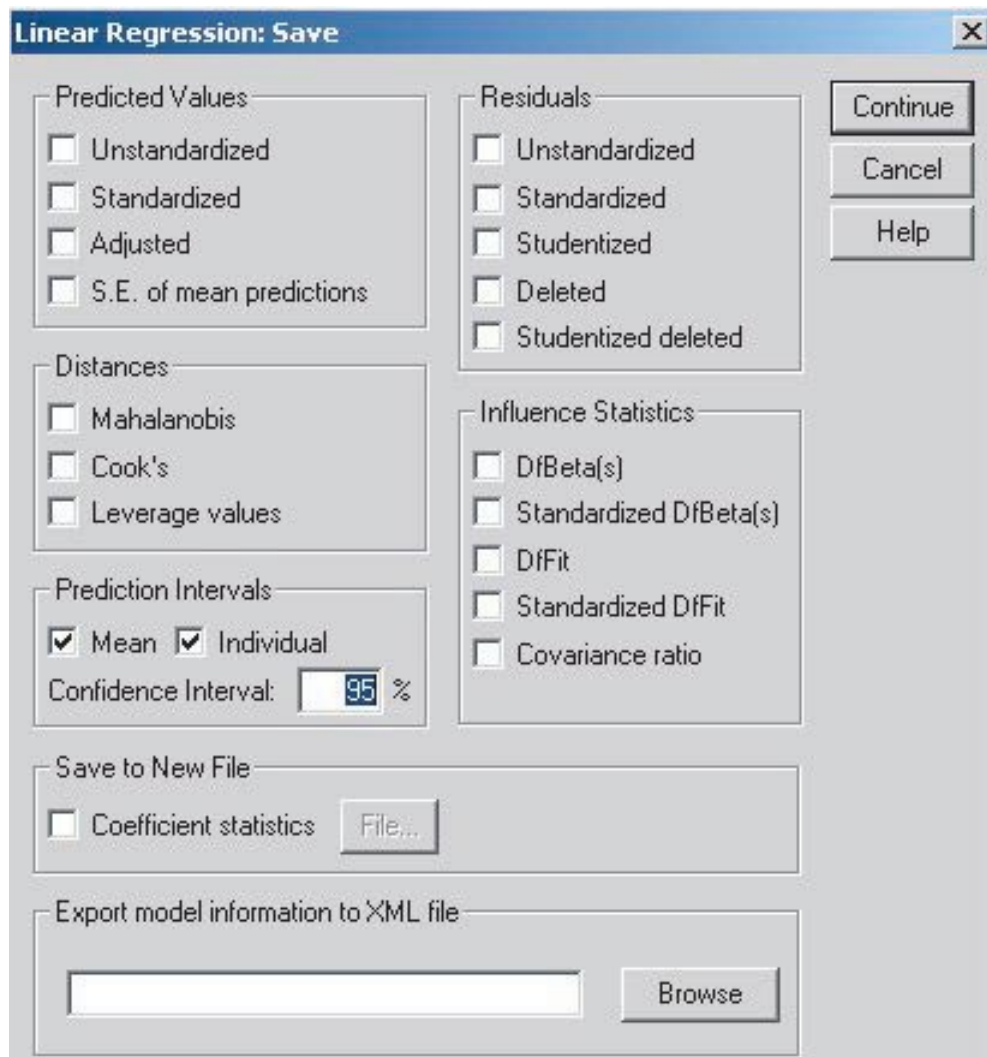
Then click **Statistics...**, and in the window that opens with **Estimates** and **Model fit** already checked, also check **Confidence intervals** and **Descriptives**.



Then click **Continue**. Next click **Plots...** In the window that opens, enter *ZRESID for Y and *ZPRED for X to get a graph of the standardized residuals as a function of the standardized predicted values.



After clicking **Continue**, next click **Save...** In the window that opens, check **Mean** and **Individual** under **Prediction Intervals** with **95%** for **Confidence Intervals**. This will add four columns to our data window that give the 95% confidence intervals for the mean values $\mu_{y|(x_1, x_2, x_3)}$ and individual values y_I for each (x_1, x_2, x_3) in our set of data points.



Then click **Continue** followed by **OK** to get the output.

Descriptive Statistics

	Mean	Std. Deviation	N
MINUTES	10.305	4.7798	20
RAM	33.20	27.816	20
INPUT	7.75	4.711	20
OUTPUT	3.95	2.350	20

We first see the mean and the standard deviation for all of the variables in the **Descriptive Statistics**.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959 ^a	.920	.904	1.4773

a. Predictors: (Constant), OUTPUT, RAM, INPUT

b. Dependent Variable: MINUTES

In the **Model Summary**, we see that the coefficient of multiple correlation r (**R**) is .959, indicating a strong positive linear relationship between the predictors and the dependent variable. The coefficient of determination r^2 (**R Square**) of .920 indicates that, for the sample, 92% of the variation of y can be explained by the variation in x_1 , x_2 , and x_3 . But this may be an overestimate for the population from which the sample is drawn, so we use the **Adjusted R Square** as a better estimate for the population. Finally, the **Standard Error of the Estimate** is 1.4773.

Coefficients^a

		Model			
		1			
		(Constant)	RAM	INPUT	OUTPUT
Unstandardized Coefficients	B	.975	9.937E-02	.243	1.049
	Std. Error	.787	.018	.115	.169
Standardized Coefficients	Beta		.578	.240	.516
t		1.239	5.469	2.116	6.221
Sig.		.233	.000	.050	.000
95% Confidence Interval for B	Lower Bound	-.694	.061	.000	.692
	Upper Bound	2.645	.138	.487	1.407

a. Dependent Variable: MINUTES

We use the sample regression (least squares) equation $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$ to approximate the population regression equation $\mu_{y|(x_1,x_2,x_3)} = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$. From the **Coefficients** table, a is .975, b_1 is .09937, b_2 is .243, and b_3 is 1.049 from the first row of numbers (rows and columns transposed from the output), so the sample regression equation is $\hat{y} = .975 + .09937x_1 + .243x_2 + 1.049x_3$. From the last two rows of numbers in the table, one gets that 95% confidence intervals are (-.694,2.645) for α , (.061,.138) for β_1 , (.000,.487) for β_2 , and (.692,1.407) for β_3 .

The t test is used for testing the various null hypotheses $\beta_i = 0$. It can be used similarly to test the null hypothesis $\alpha = 0$, but this is of much less interest. In this case, we read from the above table that, as an example, for $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$, we have $t = 5.469$. Since the p -value (**Sig.** = .000) for that t test is less than .001, we can reject the null hypothesis of $\beta_1 = 0$. Notice that at the $\alpha = .05$ level, we would accept the null hypothesis $\beta_2 = 0$ since $p = .05$. Also, notice that 0 is in the 95% confidence interval for β_2 (barely). But if using these t tests, keep in mind the dangers of using multiple hypothesis tests and/or finding multiple confidence intervals on the same set of data.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	399.169	3	133.056	60.965	.000 ^a
	Residual	34.920	16	2.183		
	Total	434.090	19			

a. Predictors: (Constant), OUTPUT, RAM, INPUT

b. Dependent Variable: MINUTES

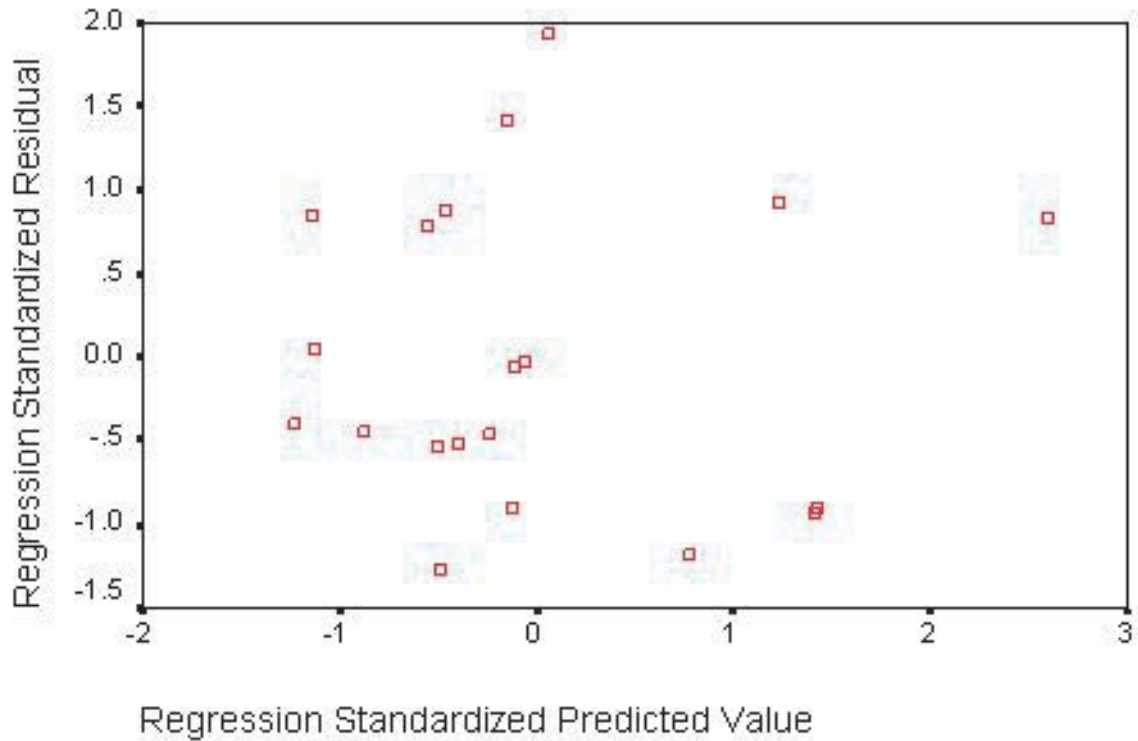
Preferably, we use the ANOVA table for testing the null hypothesis $\beta_1 = \beta_2 = \beta_3 = 0$ with an alternative hypothesis of not all $\beta_i = 0$. In the ANOVA table, the **Regression Sum of Squares** (SSR) is the variation explained by regression, and the **Residual Sum of Squares** (SSE) is the variation not explained by regression (the “E” stands for error). The **Mean Square Regression** and the **Mean Square Residual** are MSR and MSE respectively, with the **F** value of **60.965** being their quotient. Since the p -value (**Sig.** = .000) is less than .001, we can reject the null hypothesis of $\beta_1 = \beta_2 = \beta_3 = 0$, inferring indeed that there is a regression effect.

The mean value $\mu_{y|(x_1, x_2, x_3)}$ and individual y_I confidence intervals for each data point relate to the four new columns in our data window, a portion of which is shown below. We interpret the first row of data. For the predictor triple $(x_1, x_2, x_3) = (19, 5, 1)$, the 95% confidence interval for the mean value $\mu_{y|(19, 5, 1)}$ is (3.88934, 6.36905) and the 95% confidence interval for the individual value $y_I(19, 5, 1)$ is (1.76090, 8.49749). The first pair of acronyms **lmci** and **umci** stand for “lower mean confidence interval” and “upper mean confidence interval,” respectively, with the **i** in the second pair standing for “individual.”

	minutes	ram	input	output	lmci_1	umci_1	lici_1	uici_1
1	5.2	19	5	1	3.88934	6.36905	1.76090	8.49749
2	17.3	105	10	2	13.58944	18.29218	12.02455	19.85707
3	15.5	70	15	5	15.49101	18.16387	13.42241	20.23247

Finally, consider the residual plot at the top of the next page. On the horizontal axis are the standardized y values from the data points, and on the vertical axis are the standardized residuals for each such y . If all the regression assumptions were met for our data set, we would expect to see random scattering about the horizontal line at level 0 with no noticeable patterns. In fact, that is exactly what we see here.

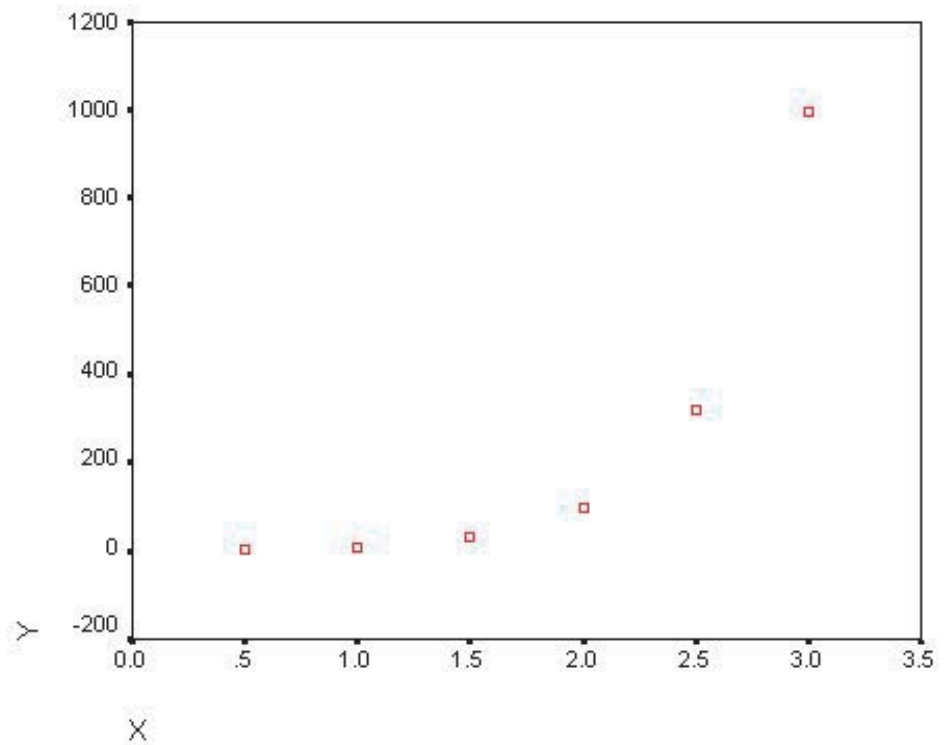
Dependent Variable: MINUTES



Nonlinear Regression

We will use the data set below for nonlinear regression. The fact that the data is nonlinear is made clear by the scatter plot, which was obtained by methods indicated in the section on Simple Linear Regression and Correlation.

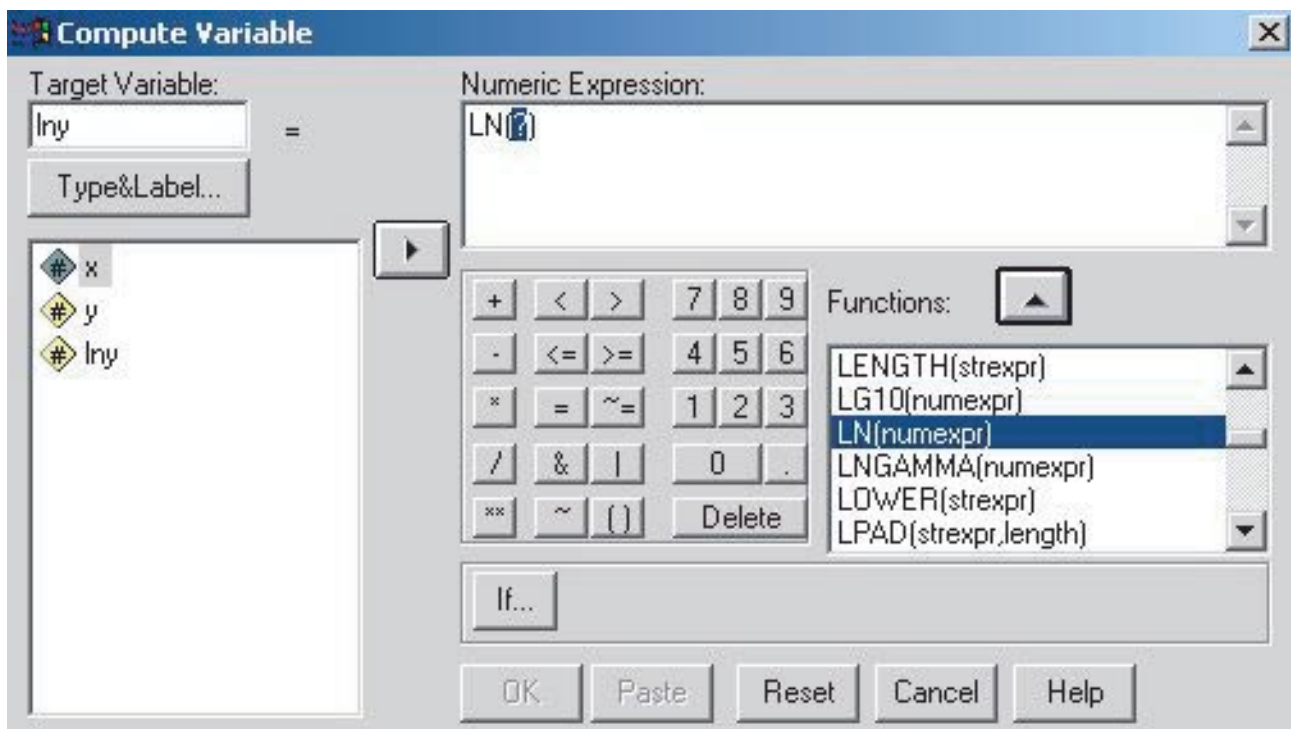
	x	y
1	.5	3.2
2	1.0	9.8
3	1.5	31.8
4	2.0	98.8
5	2.5	321.5
6	3.0	995.0



Transformation of Variables to Get a Linear Relationship. In this case we take the natural logarithm of the dependent variable y to see if x and $\ln y$ are linearly related. First return to **Variable View** in the **Data Editor**, and in the third row enter $\ln y$ under **Name** and 4 for **Decimals**, as shown below.

	Name	Type	Width	Decimals
1	x	Numeric	8	1
2	y	Numeric	8	1
3	lny	Numeric	8	4

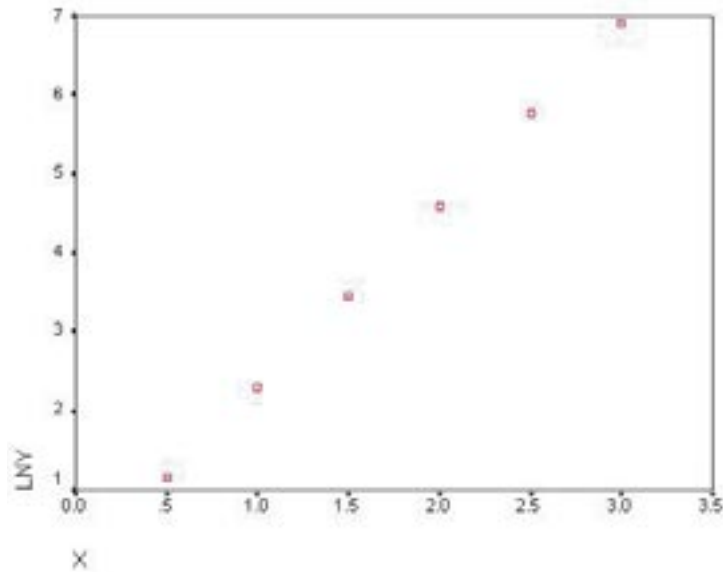
Then click back to **Data View**. From the menu, choose **Transform>Compute....** When the **Compute Variable** window comes up, click **Reset**, then type $\ln y$ in the box labeled **Target Variable**. Then scroll down the **Functions** window to **LN(numexpr)** to select it and press the up arrow.



To fill in the argument indicated by question mark, click on y in the box on the left to highlight it, then hit the right arrow to the right of that box. Then hit **OK**. If you get a message about changing the existing variable, hit **OK** for that too. The natural logarithm for each y are now found in the column $\ln y$, as seen below.

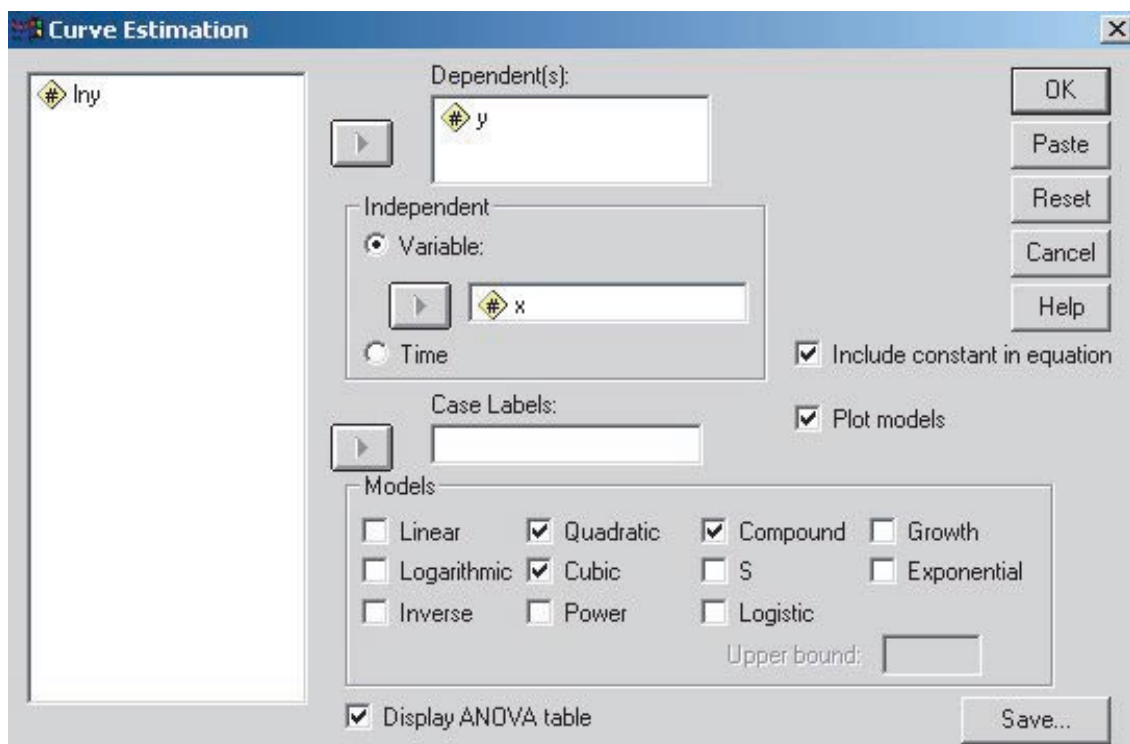
	x	y	lny
1	.5	3.2	1.1632
2	1.0	9.8	2.2824
3	1.5	31.8	3.4595
4	2.0	98.8	4.5931
5	2.5	321.5	5.7730
6	3.0	995.0	6.9027

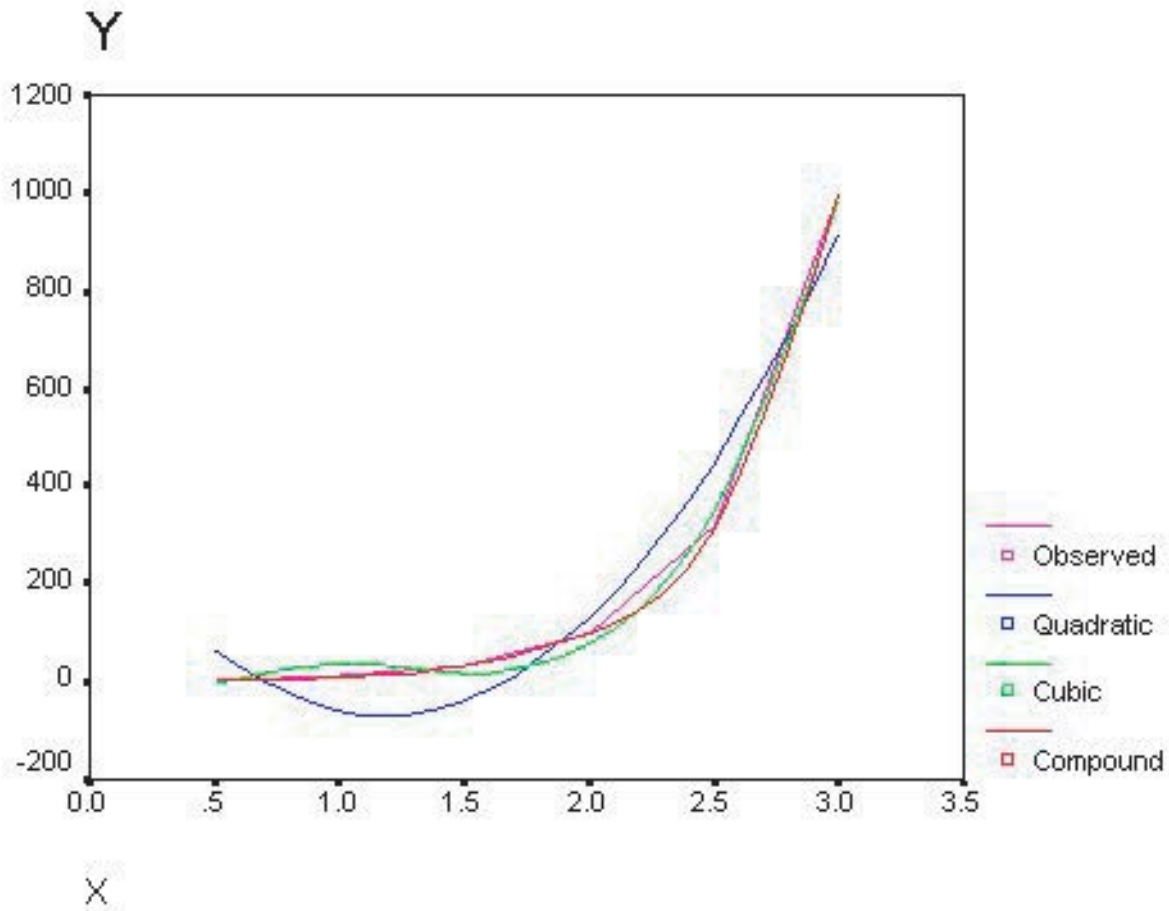
From the scatter plot that follows, it seems clear that x and $\ln y$ are linearly related.



Doing a linear regression with x as the independent variable and $\ln y$ as the dependent variable as described in the section **Simple Linear Regression and Correlation**, we get the regression equation $\ln y = .00137 + 2.303x$ with a **Standard Error of the Estimate** of .0159804. This is equivalent to the exponential regression equation $\hat{y} = .9986(10.0047)^x$.

Choosing a Model using Curve Estimation. To find an appropriate model for a given data set, such as the one in the previous section, choose **Analyze>Regression>Curve Estimation...** In the **Curve Estimation** window that opens, enter y under **Dependent(s)**, x under **Independent** with **Variable** selected, and make sure **Include constant in equation**, **Plot models**, and **Display ANOVA table** are all checked. Under **Models**, for this example check **Quadratic**, **Cubic**, and **Compound**.





Chi-Square Test of Independence

For data, we will use a survey of a sample of 300 adults in a certain metropolitan area where they indicated which of three policies they favored with respect to smoking in public places.

Highest education level	Policy Favored				Total
	No restrictions on smoking	Smoking allowed in designated areas only	No smoking at all	No opinion	
College graduate	5	44	23	3	75
High school graduate	15	100	30	5	150
Grade school graduate	15	40	10	10	75
Total	35	184	63	18	300

We wish to test if there is a relationship between education level and attitude toward smoking in public places. We test the hypotheses

H_0 : Education level and policy favored are independent

H_a : The two variables are not independent

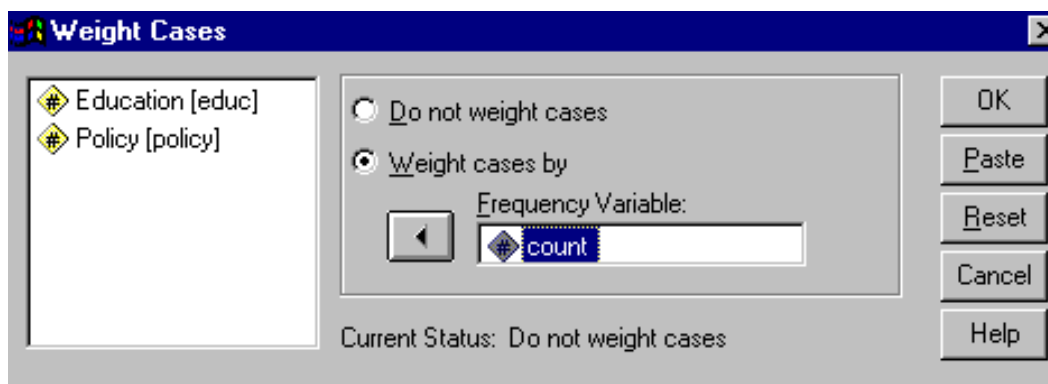
Ignoring the Total row and column, we enter the data from the table into the first column of the **Data View**, reading across the rows from left to right. In the second column we list the row the data came from, and in the third column the column the data came from. This is seen below.

	count	educ	policy
1	5	1	1
2	44	1	2
3	23	1	3
4	3	1	4
5	15	2	1
6	100	2	2
7	30	2	3
8	5	2	4
9	15	3	1
10	40	3	2
11	10	3	3
12	10	3	4

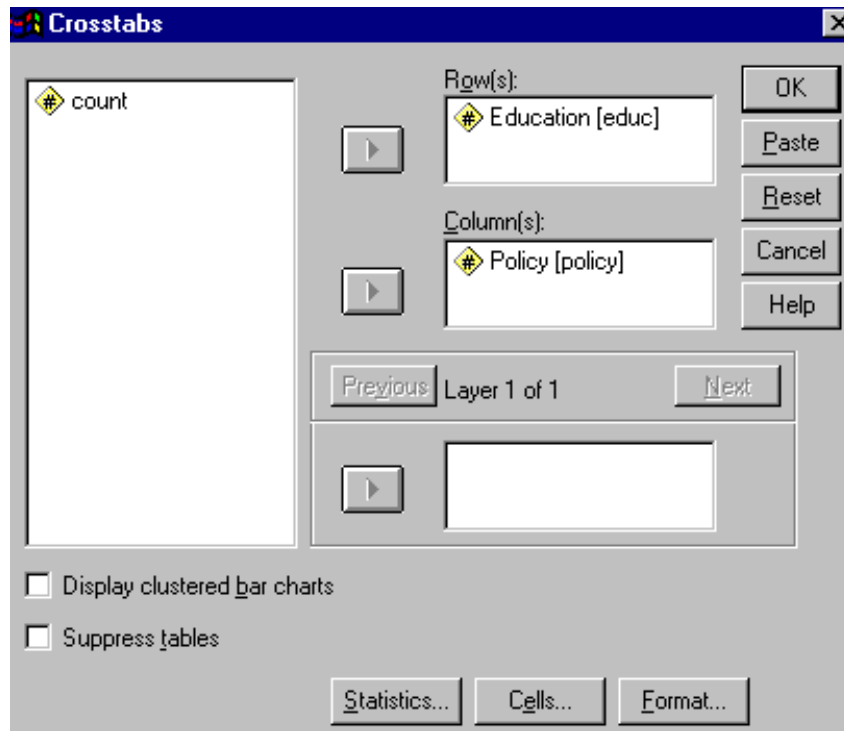
In the **Variable View** below, for the **Label "Education,"** we use the **Values "1 = College," "2 = High School,"** and **"3 = Grade School."** For the **"Label Policy,"** we use **"1 = No restrictions," "2 = designated areas," "3 = No smoking,"** and **"4 = No opinion."**

	Name	Type	Width	Decimals	Label	Values
1	count	Numeric	8	0		None
2	educ	Numeric	8	0	Education	{1, College}...
3	policy	Numeric	8	0	Policy	{1, No restri ...

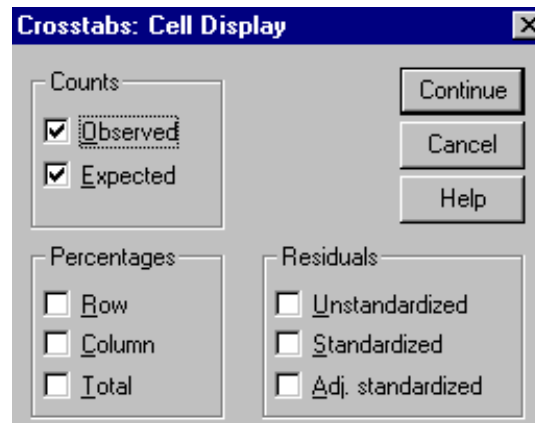
This is not very well documented, but the first thing we need to do for χ^2 is to tell SPSS which column contains the frequency counts. Choose **Data>Weight Cases...** from the menu, and in the window that opens,



choose **Weight cases by** and move the variable **count** under **Frequency Variable**. Then click **OK**. Now choose **Analyze>Descriptive Statistics>Crosstabs...** from the menu.



In the window that opens, move **Education[educ]** under **Row(s)** and **Policy[policy]** under **Column(s)**. Next click **Statistics...**, and in the window that opens, check only **Chi-square**, and then click **Continue**. Next click **Cells...**



Check **Observed** and **Expected** under **Counts**, followed by **Continue** and **OK**.

Education * Policy Crosstabulation

			Policy				Total
			No restrictions	Designated areas	No smoking	No opinion	
Education	College	Count	5	44	23	3	75
		Expected Count	8.8	46.0	15.8	4.5	75.0
	High School	Count	15	100	30	5	150
		Expected Count	17.5	92.0	31.5	9.0	150.0
	grade School	Count	15	40	10	10	75
		Expected Count	8.8	46.0	15.8	4.5	75.0
Total		Count	35	184	63	18	300
		Expected Count	35.0	184.0	63.0	18.0	300.0

The first table of output simply provides a table of the **Counts** and the **Expected Counts**

if the variables are independent.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.502 ^a	6	.001
Likelihood Ratio	20.598	6	.002
Linear-by-Linear Association	1.033	1	.310
N of Valid Cases	300		

a. 2 cells (16.7%) have expected count less than 5. The minimum expected count is 4.50.

From the second table, the **Pearson Chi-Square** statistic is 22.502 with a p -value (**Asymp. Sig. (2-sided)**) of .001. Thus, for instance, we would reject the null hypothesis at the $\alpha = .01$ level of significance. Notice the note that 16.7% of the cells have expected counts less than 5 and the minimum expected count is 4.5. Typically, we need no more than 20% of the expected counts less than 5 with a minimum expected count of at least 1.

Nonparametric Tests

The Wilcoxon Matched-Pairs Signed-Rank Test. For data, we use cardiac output (liters/minute) of 15 postcardiac surgical patients. The data is as follows:

4.91	4.10	6.74	7.27	7.42	7.50	6.56	4.64
5.98	3.14	3.23	5.80	6.17	5.39	5.77	

We want to test the hypotheses

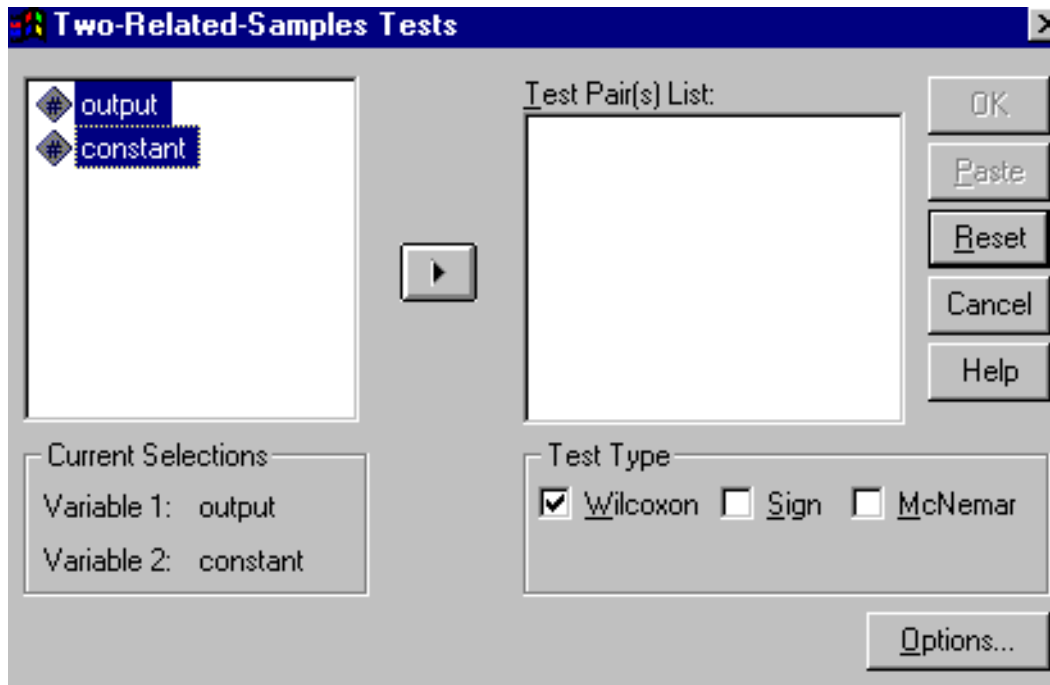
$$H_0 : \mu = 5.05$$

$$H_a : \mu \neq 5.05.$$

We enter the data by putting the numbers above in the first column, labeled **output**. Because we are using a matched-pairs test, we create the matched pairs by entering the test value **5.05** fifteen times in the second column, labeled **constant**. The **Data View** looks as below.

	output	constant
1	4.91	5.05
2	5.98	5.05
3	4.10	5.05
4	3.14	5.05
5	6.74	5.05
6	3.23	5.05

From the menu, choose **Analyze>Nonparametric Tests>2 Related Samples....**



In the window that opens, first click **output** to make it **Variable 1**, then **constant** to make it **Variable 2**. Then click the right arrow to move **output-constant** under **Test Pair(s) List**. Make sure **Wilcoxon** is checked. If you want descriptive statistics and/or quartiles, you can choose those under **Options....** Then click **OK** to get the output.

Ranks

		N	Mean Rank	Sum of Ranks
CONSTANT - OUTPUT	Negative Ranks	10 ^a	8.60	86.00
	Positive Ranks	5 ^b	6.80	34.00
	Ties	0 ^c		
	Total	15		

- a. CONSTANT < OUTPUT
- b. CONSTANT > OUTPUT
- c. OUTPUT = CONSTANT

The first table of output gives the number of the 15 comparisons that are **Negative** (rank of **constant** < rank of **output**), **Positive** (rank of **constant** > rank of **output**), and **Ties** (rank of **constant** = rank of **output**). We are also given the **Mean Rank** and **Sum of Ranks** for all of the **Negative Ranks** and the **Positive Ranks**. The test statistic is the smaller of the **Sum of Ranks**.

Test Statistics^b

	CONSTANT - OUTPUT
Z	-1.477 ^a
Asymp. Sig. (2-tailed)	.140

- a. Based on positive ranks.
- b. Wilcoxon Signed Ranks Test

The **Z** in the second table is the standardized normal approximation to the test statistic, and the **Asymp. Sig (2-tailed)** of .140, which we will use as our p -value, is estimated from the normal approximation. Because of the size of this p -value, we will not reject the null hypothesis at any of the usual levels of significance.

The Mann-Whitney Rank-Sum Test For data, we will look at hemoglobin determination (grams) for 25 laboratory animals, 15 of whom have been exposed to prolonged inhalation of cadmium oxide.

Exposed 14.4, 14.2, 13.8, 16.5, 14.1, 16.6, 15.9, 15.6, 14.1, 15.2, 15.7, 16.7, 13.7,
15.3, 14.0
Unexposed 17.4, 16.2, 17.1, 17.5, 15.0, 16.0, 16.9, 15.0, 16.3, 16.8,

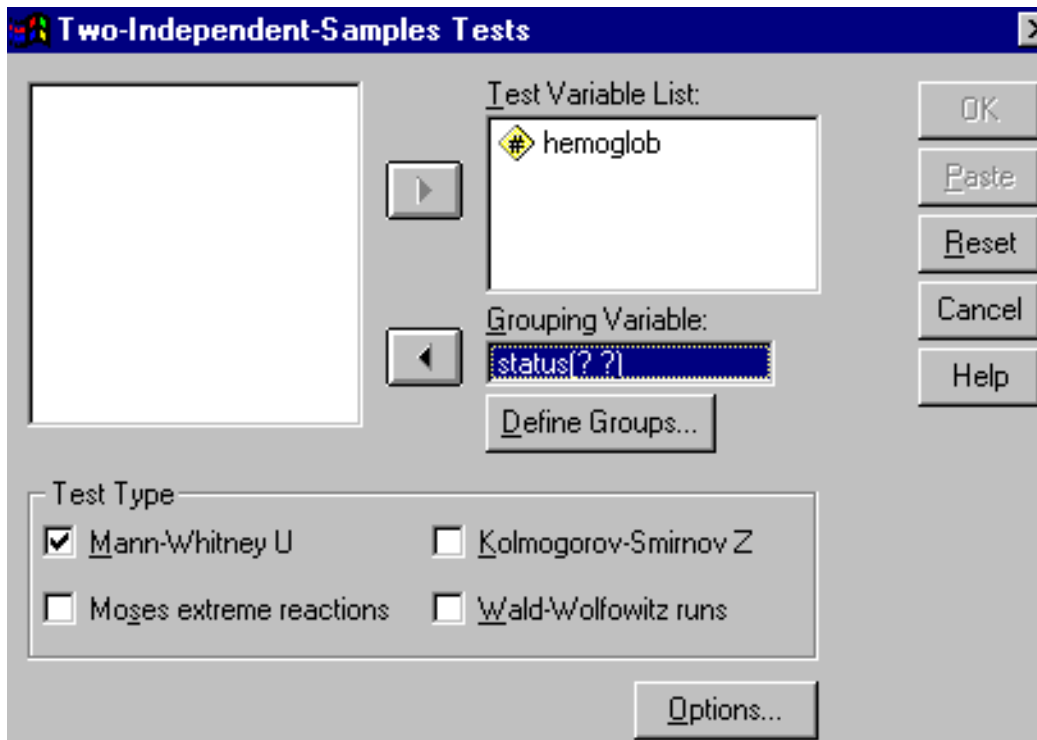
We want to test the hypotheses

$$H_0 : \mu_{\text{exposed}} = \mu_{\text{unexposed}}$$

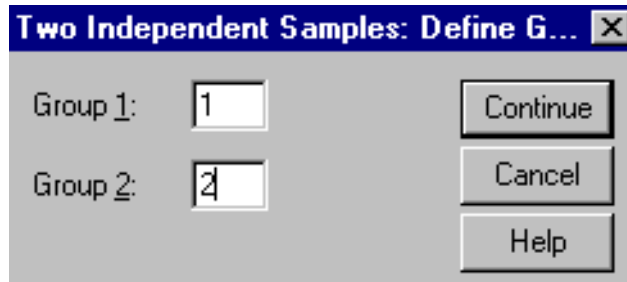
$$H_a : \mu_{\text{exposed}} > \mu_{\text{unexposed}}$$

As for the t test earlier, we enter the 25 hemoglobin readings in column one of the **Data View** and label the column **hemoglob**. In the second column, labeled **status**, we use “1 = Exposed” and “2 = Unexposed”, which are also listed under **Values** for **status** in the **Variable View**.

To do the test, choose **Analyze>Nonparametric Tests>Two Independent Samples...** from the menu.



In the window that opens, first check **Mann-Whitney U** under **Test Type**, then move the variable **hemoglob** to the **Test Variable List** box and the variable **status** to the **Grouping Variable** box. Then click **Define Groups...**



Put 1 in the box for **Group 1** and 2 in the box for **Group 2**. Then click **Continue**. You may click **Options...** if you want the output to include descriptive statistics and/or quartiles. Finally, click **OK** to get the output.

Ranks

	STATUS	N	Mean Rank	Sum of Ranks
HEMOGLOB	Exposed	15	9.67	145.00
	Unexposed	10	18.00	180.00
	Total	25		

We see from the first table, after ranking the **hemoglobin** values from least to greatest, the **Mean Rank** and **Sum of Ranks** for each **status** category.

Test Statistics^b

	HEMOGLOB
Mann-Whitney U	25.000
Wilcoxon W	145.000
Z	-2.775
Asymp. Sig. (2-tailed)	.006
Exact Sig. [2*(1-tailed Sig.)]	.004 ^a

a. Not corrected for ties.

b. Grouping Variable: STATUS

The **Mann-Whitney U**, calculated by counting the number of times a value from the smaller group (here **Unexposed**) is less than a value from the larger group (here **Exposed**), is **25.000**. This is equivalent to the **Wilcoxon W**, which is the **Sum of Ranks** of the smaller group. The **Z** in the second table is again the standardized normal approximation to the test statistic, and the **Asymp. Sig (2-tailed)** of **.006** is estimated from the normal approximation. Because we are using a 1-tailed test, we will take one-half of this number, **.003** as our *p*-value, causing us to reject the null hypothesis at all of the usual levels of significance.

Control Charts

Control Charts for the Mean. To illustrate control charts for the mean, we use the following sample yield data in grams/liter which have been obtained for each of five suc-

cessive days, with all samples of size 7. Let us also assume that the process has a specified mean $\mu_0 = 50$ and specified standard deviation $\sigma_0 = 1$.

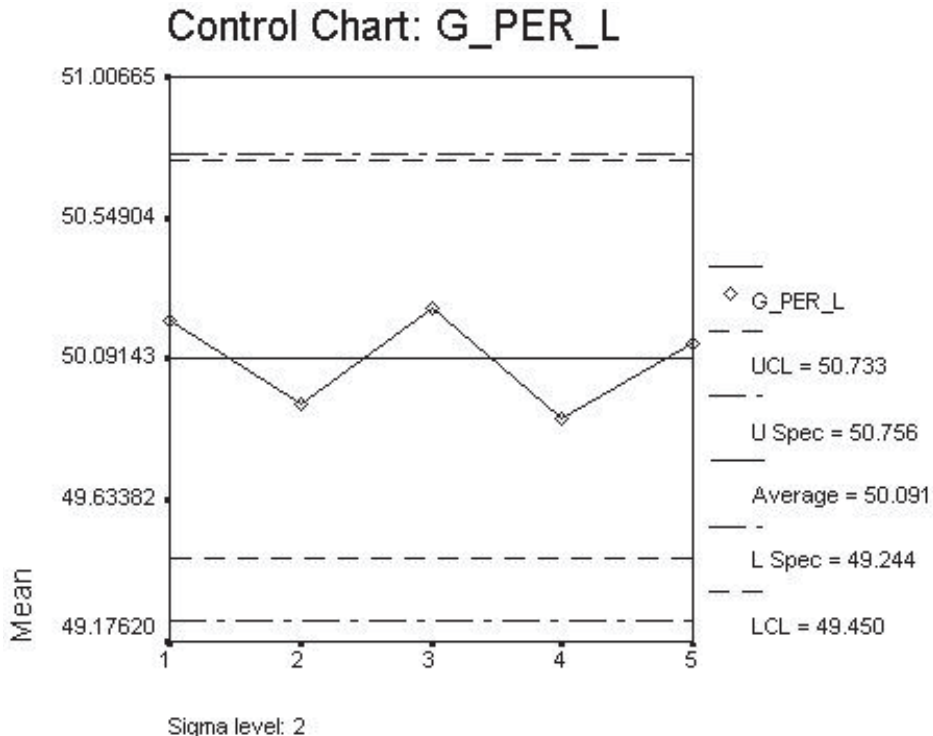
Day 1:	49.5	49.9	50.5	50.2	50.5	49.8	51.1
Day 2:	48.5	52.3	48.2	51.2	50.1	49.3	50.0
Day 3:	50.5	51.7	49.5	51.2	48.3	50.2	50.4
Day 4:	49.8	49.7	50.2	50.6	50.3	49.4	49.3
Day 5:	50.5	50.9	49.5	50.2	49.8	49.8	50.3

In entering the data in the **Data Editor**, put the 35 sample values in the first column, labeled **g_per_l**, with 1 decimal place, and put the day number in the second column, labeled **day**, with no decimal places. A portion of this **Data Editor** window is shown below.

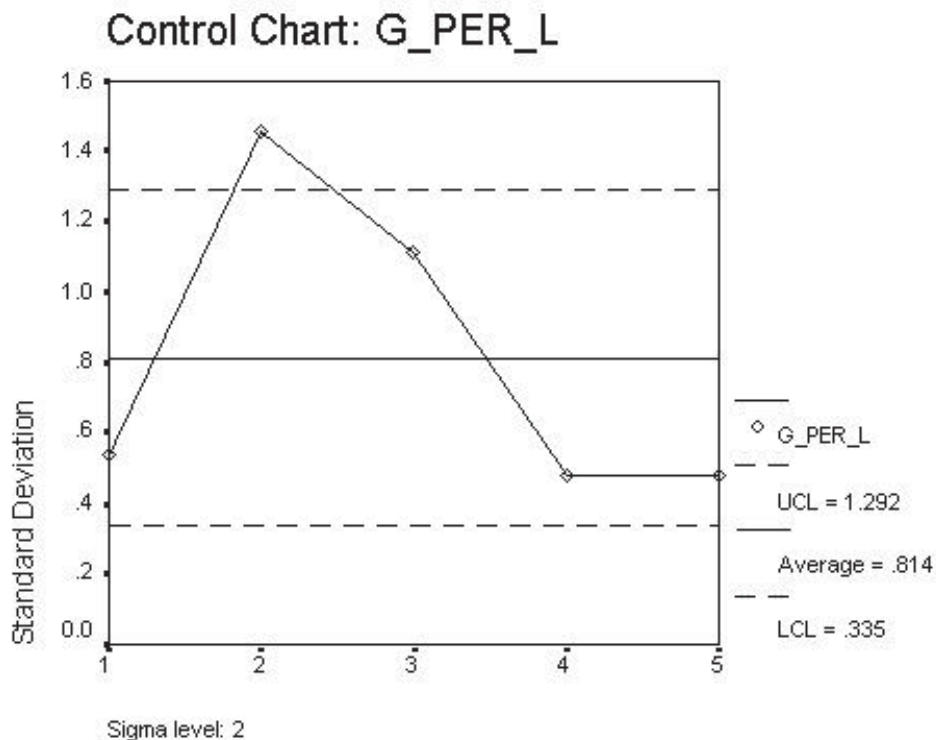
	g_per_l	day
1	49.5	1
2	49.9	1
3	50.5	1
4	50.2	1
5	50.5	1
6	49.8	1
7	51.1	1
8	48.5	2
9	52.3	2
10	48.2	2
11	51.2	2
12	50.1	2
13	49.3	2
14	50.0	2
15	50.5	3
16	51.7	3
17	49.5	3

To create the control chart(s), click **Graphs>Control...** from the menu bar, and in the window that opens, select **X-Bar, R, s** and make sure **Cases are units** is checked under **Data Organization**. Then click **Define**, and in the new window that opens, move **g_per_l** under **Process Measurement** and **day** under **Subgroups Defined by**. Under **Charts**, we will select **X-Bar and standard deviation**, with the other option being **X-Bar and range**. Click **Options**, and enter **2** for **Number of Sigmas**. After clicking **Continue**, since we have specifications for the mean, we click **Statistics...**, and in the window that opens, based on our specified mean and standard deviation, enter **50.756** for **Upper** and **49.244** under **Lower** for **Specification Limits**, then select **Estimate using S-Bar** under **Capability Sigma**. Finally, click **Continue** followed by **OK** to get the control charts.

The first control chart given as output is the chart for the mean. This chart, which is pretty much self-explanatory, clearly shows the daily means along with the unspecified (UCL and LCL) and specified (USpec and LSpec) control limits. It is clear that the process is always in control.



The second control chart is for the standard deviation, and it is clear that, as far as standard deviation is concerned, the process is out of control on Day 2.



In the event X-Bar and range had been chosen, the second chart would be a range chart.

Control Charts for the Proportion. To illustrate control charts for the proportion, we use the number of defectives in samples of size 100 from a production process for twenty days in August.

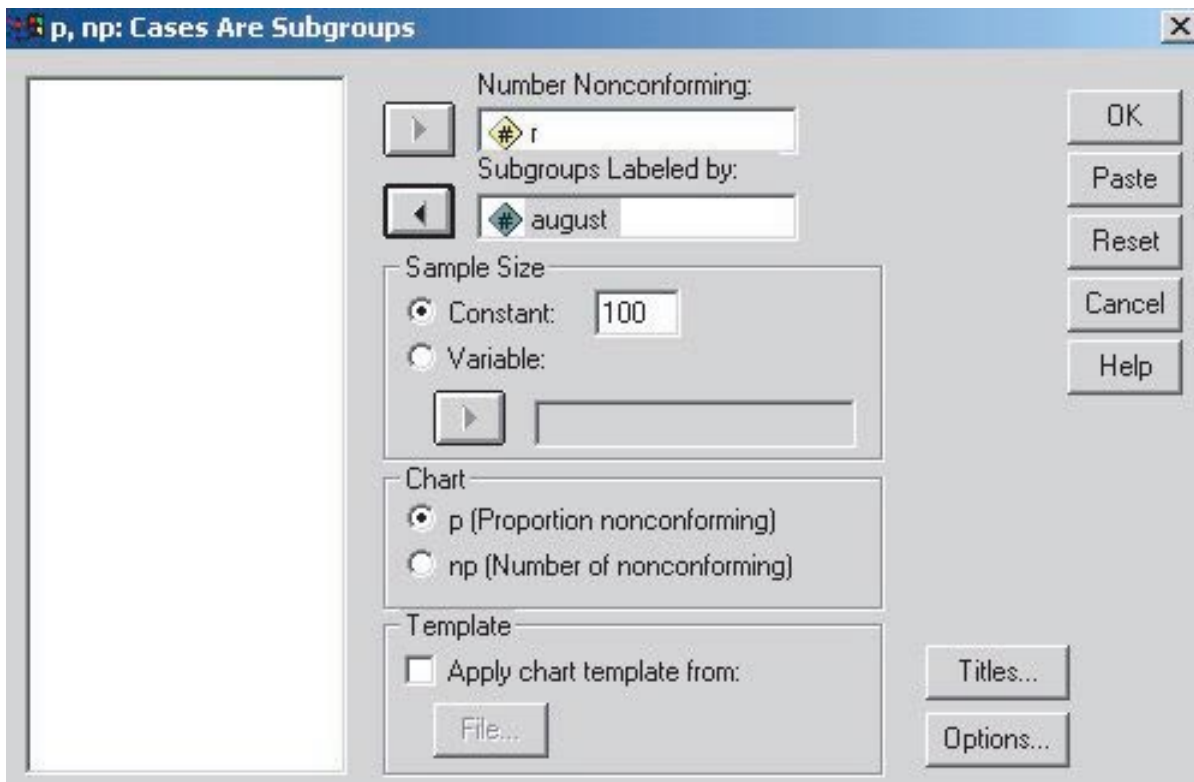
August: 6 7 8 9 10 11 12 13 14 15
 Defectives: 8 15 12 19 7 12 3 9 14 10

August: 16 17 18 19 20 21 22 23 24 25
 Defectives: 22 13 10 15 18 11 7 15 24 2

In entering the data in the **Data Editor**, put the 20 numbers of defectives (from each sample of size 100) in the first column, labeled **r**, and put the corresponding date in August in the second column, labeled **august**, both with no decimal places, as shown below.

	r	august
1	8	6
2	15	7
3	12	8
4	19	9
5	7	10

To create the control chart, click **Graphs>Control...** from the menu bar, and in the window that opens, select **p, np** and make sure **Cases are subgroups** is checked under **Data Organization**. Then click **Define**, and in the new window that opens, move **r** under **Number Nonconforming**, move **august** under **Subgroups Labeled by**, select **Constant** for **Sample Size**, and enter **100** in the following box. Under **Charts**, we will select **p (Proportion nonconforming)**, with the other option being **np (Number nonconforming)**.



Now click **Options**, and enter **3** for **Number of Sigmas**. Then click **Continue** followed by **OK** to get the control chart, which is again pretty much self-explanatory. We see that the

process is out of control on August 24 and 25, although it is hard to call too few defectives out of control.

