

POPULATION (TARGET POPULATION) - A population is a collection of all measurements of interest in a particular study. It consists of the totality of the observations with which we are concerned.

SAMPLE - A sample is a portion of the population selected to represent the whole population. It is a subset of the population.

RANDOM SAMPLE - A random sample is a sample that is chosen in such a way that each observation in the population have the same chance of being selected. *Note: if some elements of the population (even one element) have a greater likelihood of being sampled than other elements, then the sample **cannot** be a random sample. For example, the classical sampling technique where a reporter is sent to one or several street corners to interview typical citizens cannot be a random sample, because it is impossible to ensure that each citizen has an equal chance of being at the street corners when the reporter is conducting interviews.*

FRAME - A frame is a list from which the sample is selected. Often the frame will not match the target population exactly. If the frame differs considerable from the population targeted, any conclusions drawn may be inaccurate. Suppose that the target population was all doctors in Maryland and that the membership list of the Maryland Chapter of The American Medical Association (AMA) was used as the frame. Since some doctors may not be members of the AMA, this list may not include all doctors in the population. In other words, this frame may not match the population exactly.

PARAMETER - Any calculation that is made using all of the observations from the population is called a parameter. It is a numerical value describing a characteristic of the population. Most of the time letters from the Greek alphabet are used to represent parameters.

STATISTIC - Any calculation that is made using only a subset of the population (from a sample) is called a statistic. It is a numerical value describing a characteristic of the sample. A statistic is usually used to estimate a population parameter. Letters from the English alphabet are used to represent statistics. The symbols are illustrated in the chart below. *Note: Different symbols are used to distinguish quantities that are statistics (from a sample) from those that are parameters (from the population). Sometimes the formulas differ. Some examples are listed below.*

Quantity Calculated	Size	Mean	Median	Variance	Standard Deviation	Proportion
Statistic	n	\bar{x}	\tilde{x}	s^2	s	\hat{p}
Parameter	N	μ	$\tilde{\mu}$	σ^2	σ	p

Quantity	Formula for the Statistic	Formula for the Parameter
Mean	$\bar{x} = \frac{x}{n}$	$\mu = \frac{x}{N}$
Proportion	$\hat{p} = \frac{x}{n}$	$p = \frac{x}{N}$
Variance (Computational Formula)	$s^2 = \frac{n \sum x^2 - (\sum x)^2}{n(n-1)}$	$\sigma^2 = \frac{N \sum x^2 - (\sum x)^2}{N^2}$
Variance (Conceptual Formula)	$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$
Standard Deviation (Computational Formula)	$s = \sqrt{\frac{n \sum x^2 - (\sum x)^2}{n(n-1)}}$	$\sigma = \sqrt{\frac{N \sum x^2 - (\sum x)^2}{N^2}}$
Standard Deviation (Conceptual Formula)	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$

NOMINAL MEASUREMENT - The level of measurement is nominal when the measure assigned to an item is a label used to identify the item. The labels can be names or numbers. There is no natural order to the data. Ordinary arithmetic with nominal data is meaningless. **Examples:**

1. **Political affiliation:** democrat, republican, independent.
2. **Race:** African American, Caucasian, etc.
3. **Automobile:** Chevrolet, Honda, Toyota, etc.
4. **Numbers on ballplayer uniforms:** Although one could list the numbers on the uniforms of baseball players in "order", the ordering would be meaningless and arbitrary. It would change, every time a player got a new number (label).

ORDINAL MEASUREMENT - The level of measurement is ordinal when the measures assigned permit the item to be ordered with respect to some criterion. Ordinal data is similar to nominal data in that it assigns labels to the items, however, these labels can be ordered in some natural way. Ordinary arithmetic with ordinal data is meaningless. **Examples:**

1. **Condition of hospital patients:** good, fair, critical.
2. **Class ranking:** first, second, third, etc. Ordinary arithmetic with ordinal data is meaningless. For example you cannot say that the difference between two students who rank first and second in class is the same as that between two students who rank tenth and eleventh.
3. **Size of Automobile:** compact, mid-sized, full sized. If you change the labels on the automobile sizes to small, medium, and large, the ordering remains the same.
4. **Rating scales:** rate performance of gymnasts from 1 to 5, where 5=excellent, 4=good, 3=average, 2=poor, & 1=very poor. Here the numbers from 1 to 5 are just labels. A poor performance (2) and an average (3) do not "add" up to an excellent performance (5). Ordinary arithmetic rarely makes sense here. The units of measurement are not fixed. The difference between an excellent (5) and average (3) performance is not necessarily the same as that between a good (4) and poor (2).

INTERVAL MEASUREMENT - The level of measurement is interval when there is a fixed numerical unit of measurement and each measure assigned is expressed as a quantity of those units. Essentially interval data can take on all values in an interval on the number line. Interval data does not have a "true" zero point. If it contains a zero point, it is an arbitrary one. Addition and subtraction with interval data is meaningful. Ratios and proportions with interval data are not meaningful. **Examples:**

1. **Temperature measured on the Fahrenheit scale:** The zero is arbitrary because the value 0°F does not indicate the absence of heat. All values above -459.8°F are possible.
2. **Temperature measured on the Celsius scale:** The zero is arbitrary (the freezing point of ice), and all values above -273.2°C are possible. *(Note: It makes no sense to say that 80°F is twice as hot as 40°F because the amount of heat represented by 0°F is not referring to a condition of no heat. Also, on the Celsius scale the temperature of 40°F is 4.44°C , and 80°F is 26.67°C . The higher temperature on the Celsius scale is no longer twice the lower temperature. As one can see, the ratio of the two temperatures changes when you change the scale. This makes any interpretation of such a ratio meaningless.)*
3. **Clock Time and Calendar Time:** The zero is arbitrary.

RATIO MEASUREMENT - The level of measurement is ratio when there is a fixed unit of measure and the zero point is inherently defined on the scale of measurement. The only difference between the ratio scale and the interval scale is the presence of a true zero. As a result, ratios and proportions as well as addition and subtraction are meaningful on the ratio scale. **Examples:**

1. **Height (length)** has a true zero and all values above zero are possible. A person 6 feet tall is twice as tall as a person who is 3 feet tall. On the metric scale 6 feet corresponds to 1.8288 meters and 3 feet corresponds to .9144 meters. As you can see, the ratio of the larger height to the smaller equals 2 no matter which scale is used. This is because there is a true (absolute) zero point.
2. **Weight** is measured on the ratio scale. A weight of 0 pounds or 0 grams represents the absence of weight.
3. **Temperature measured on the Kelvin scale** does have an absolute zero. The temperature of 0°K represents the absence of heat. Temperature measured in degrees Kelvin is a ratio measurement.
4. **Time on task or completion time:** The zero here is a true zero. It indicates the task was never begun.

QUALITATIVE (CATEGORICAL) DATA - Nominal and ordinal data are classified as qualitative or categorical data. The distinguishing characteristic of such data is that mathematical operations such as addition and subtraction do not lead to meaningful results. The data is grouped into categories. There is a difference in quality between the categories and this difference has no numerical meaning.

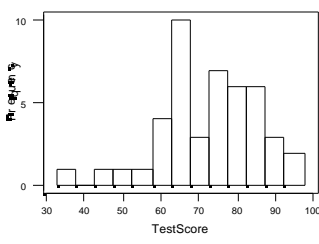
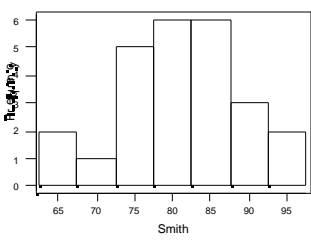
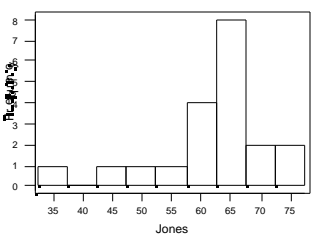

QUANTITATIVE (NUMERICAL) DATA - Interval and ratio data are classified as quantitative or numerical data. They share the distinguishing characteristic that the measures assigned are quantities defined in terms of fixed and equal units. Numerical differences between the quantities measured are meaningful.

The file labeled ‘TestScores.mtw’ will be used to illustrate the graphs. Assume that these are scores on a common test of 45 men and women who were taught by two different instructors (1 = Bill Smith & 2 = George Jones).

Simulated Test Score Data

C1-T	C2	C3	C4	C1-T	C2	C3	C4	C1-T	C2	C3	C4
Student	Test Scores	Teach	Gender	Student	Test Scores	Teach	Gender	Student	Test Scores	Teach	Gender
Dennis	33	2	1	Peter	63	1	1	Dennis	33	2	1
Nate	44	2	1	Julia	66	1	2	Nate	44	2	1
Mary	52	2	2	Ted	71	1	1	William	60	2	1
Denise	57	2	2	Bill	75	1	1	Jim	61	2	1
William	60	2	1	Michael	76	1	1	Frank	61	2	1
Jim	61	2	1	Alice	76	1	2	Peter	63	1	1
Frank	61	2	1	Helen	76	1	2	Randy	63	2	1
Sharon	61	2	2	Millisa	77	1	2	Bob	66	2	1
Randy	63	2	1	Aaron	78	1	1	Otto	67	2	1
Peter	63	1	1	Ellen	78	1	2	Sam	69	2	1
Bridget	63	2	2	Jack	81	1	1	Ted	71	1	1
Natalie	64	2	2	Tom	81	1	1	Bill	75	1	1
Karen	65	2	2	Ken	82	1	1	Michael	76	1	1
Carol	65	2	2	Courtney	82	1	2	Brian	76	2	1
Bob	66	2	1	Sean	84	1	1	Aaron	78	1	1
Julia	66	1	2	Mandy	85	1	2	Jack	81	1	1
Otto	67	2	1	Eric	86	1	1	Tom	81	1	1
Bonnie	67	2	2	Mac	86	1	1	Ken	82	1	1
Sam	69	2	1	Amanda	86	1	2	Sean	84	1	1
Ted	71	1	1	Wilma	86	1	2	Eric	86	1	1
Laura	71	2	2	George	88	1	1	Mac	86	1	1
Sally	74	2	2	Jennifer	89	1	2	George	88	1	1
Bill	75	1	1	Steven	91	1	1	Steven	91	1	1
Brian	76	2	1	Harold	93	1	1	Harold	93	1	1
Michael	76	1	1	Alan	94	1	1	Alan	94	1	1
Alice	76	1	2	Dennis	33	2	1	Mary	52	2	2
Helen	76	1	2	Nate	44	2	1	Denise	57	2	2
Millisa	77	1	2	Mary	52	2	2	Sharon	61	2	2
Aaron	78	1	1	Denise	57	2	2	Bridget	63	2	2
Ellen	78	1	2	William	60	2	1	Natalie	64	2	2
Jack	81	1	1	Jim	61	2	1	Karen	65	2	2
Tom	81	1	1	Frank	61	2	1	Carol	65	2	2
Ken	82	1	1	Sharon	61	2	2	Julia	66	1	2
Courtney	82	1	2	Randy	63	2	1	Bonnie	67	2	2
Sean	84	1	1	Bridget	63	2	2	Laura	71	2	2
Mandy	85	1	2	Natalie	64	2	2	Sally	74	2	2
Eric	86	1	1	Karen	65	2	2	Alice	76	1	2
Mac	86	1	1	Carol	65	2	2	Helen	76	1	2
Amanda	86	1	2	Bob	66	2	1	Millisa	77	1	2
Wilma	86	1	2	Otto	67	2	1	Ellen	78	1	2
George	88	1	1	Bonnie	67	2	2	Courtney	82	1	2
Jennifer	89	1	2	Sam	69	2	1	Mandy	85	1	2
Steven	91	1	1	Laura	71	2	2	Amanda	86	1	2
Harold	93	1	1	Sally	74	2	2	Wilma	86	1	2
Alan	94	1	1	Brian	76	2	1	Jennifer	89	1	2

HISTOGRAMS

		
<p><i>Click on: Graph</i> <i>Click on: Histogram</i> <i>Choose for Graph 1: TestScore</i> <i>Click on: OK</i></p>	<p><i>Click on: Graph</i> <i>Click on: Histogram</i> <i>Choose for Graph 1: Smith</i> <i>Click on: OK</i></p>	<p><i>Click on: Graph</i> <i>Click on: Histogram</i> <i>Choose for Graph 1: Jones</i> <i>Click on: OK</i></p>
	<p><i>Note: The unstack command was used to get separate columns for The Smith and Jones test scores. This is illustrated in the window on the left.</i></p> <p><i>Click on: Manip</i> <i>Click on: Stack/Unstack</i> <i>Click on: Unstack One Column</i></p> <p><i>Note: You can also use cut and paste if you sort the columns properly.</i></p>	

Notice that in these examples Minitab choose the scale. Now we will choose the scale by specifying the midpoints of each class. The midpoints shall be put in column C5 labeled "Midpoints". Here is how we shall do this. Suppose that we want 7 classes of data. Then perform the following calculations:

$$\text{ClassWidth} = \frac{\text{Range}}{\text{Number of Classes Desired}} = \frac{94 - 33}{7} = \frac{61}{7} = 8.71 \approx 9$$

(Note: we rounded up to the next odd integer, which in this case is 9.)

Suppose we choose to start our first class limit at 32, then each of our next class limits will start at 9 units higher the previous one, until we have spanned the entire range of our data (32, 41, 50, 59, 68, 77, 86, 95). Here is what the classes will look like along with the frequency distribution:

Starting Point of Each Class	Distinct Class Boundaries	Class Limits	Class Midpoints	Frequency	Cumulative Frequency	Percent	Cumulative Percent
32	32 – 40	31.5 – 40.5	36	1	1	2.2%	2.2%
41	41 – 49	40.5 – 49.5	45	1	2	2.2%	4.4%
50	50 – 58	49.5 – 58.5	54	2	4	4.4%	8.9%
59	59 – 67	58.5 – 67.5	63	14	18	31.1%	40.0%
68	68 – 76	67.5 – 76.5	72	9	27	20.0%	60.0%
77	77 – 85	76.5 – 85.5	81	9	36	20.0%	80.0%
86	86 – 94	85.5 – 94.5	90	9	45	20.0%	100.0%
95				45		99.9%	

$$\text{ClassMidpoint} = \frac{\text{Class Limits}}{2} = \frac{315 + 405}{2} = \frac{72}{2} = 36$$

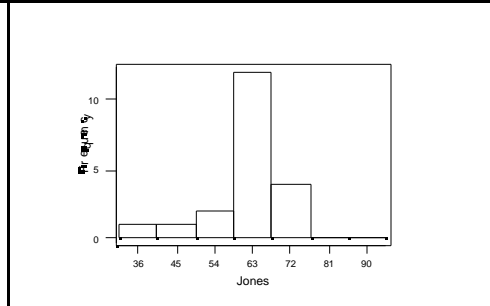
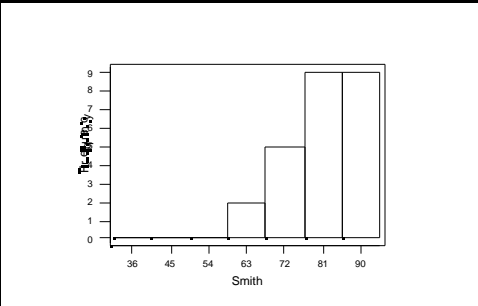
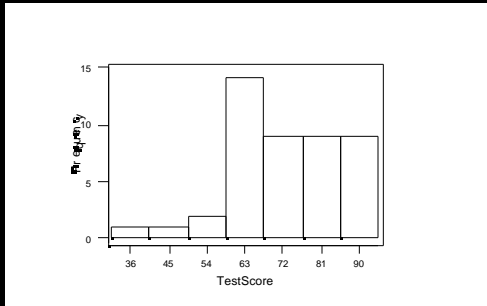
Note: you only have to calculate the first class midpoint this way, because each succeeding one is 9 units greater than the previous one.

In order to control the placing of the midpoints on the x-axis in Minitab, you must type in the midpoints you wish to use in their own Minitab column. Column C5 was used for that purpose here. Your Minitab worksheet should now essentially look as follows.

<i>C1-T</i>	<i>C2</i>	<i>C3</i>	<i>C4</i>	<i>C5</i>
Student	TestScore	Teacher	Gender	Midpoints

Dennis	33	2	1	36
Nate	44	2	1	45
Mary	52	2	2	54
Denise	57	2	2	63
William	60	2	1	72
Jim	61	2	1	81
Frank	61	2	1	90

Once we have specified the midpoints we want to use in column C5 we can now get the desired histograms. This is done under options from the histogram window.

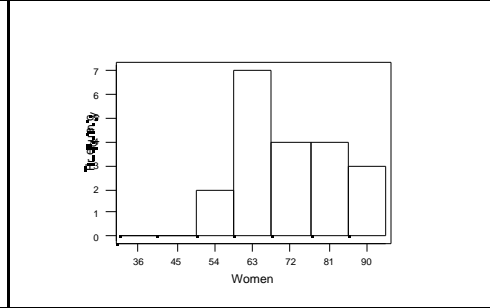
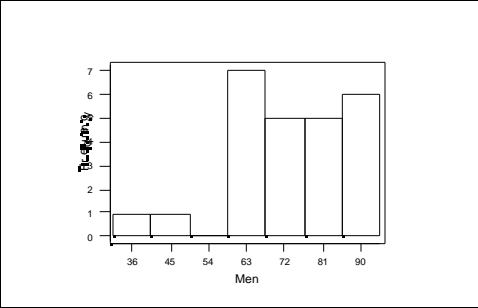


Click on: Graph
 Click on: Histogram
 Choose for Graph 1: TestScore
 Click on: OK
 Under Options: Choose Midpoints
 Select Column
 'Midpoints'

Click on: Graph
 Click on: Histogram
 Choose for Graph 2: Smith
 Click on: OK
 Under Options: Choose Midpoints
 Select Column
 'Midpoints'

Click on: Graph
 Click on: Histogram
 Choose for Graph 3: Jones
 Click on: OK
 Under Options: Choose Midpoints
 Select Column
 'Midpoints'

Here are Similar Graphs for the scores of the Men and Women.



STEM & LEAF DIAGRAMS

<p>Minitab Chooses Increment</p>	<p>You Choose Increment</p>
<p>Character Stem-and-Leaf Display Stem-and-leaf of TestScor N = 45 Leaf Unit = 1.0</p> <pre> 1 3 3 1 3 2 4 4 2 4 3 5 2 4 5 7 12 6 01113334 19 6 5566779 22 7 114 (8) 7 56666788 15 8 11224 10 8 5666689 3 9 134 </pre>	<p>Character Stem-and-Leaf Display Stem-and-leaf of TestScor N = 45 Leaf Unit = 1.0</p> <pre> 1 3 3 2 4 4 4 5 27 19 6 011133345566779 (11) 7 11456666788 15 8 112245666689 3 9 134 </pre>
<p><i>Click on: Graph</i> <i>Click on: Character Graphs</i> <i>Click on: Stem and leaf</i> <i>Choose for Variables: TestScore</i> <i>Click on :OK</i></p>	<p><i>Click on: Graph</i> <i>Click on: Character Graphs</i> <i>Click on: Stem and leaf</i> <i>Choose for Variables: TestScore</i> <i>Put in for Increment: 10</i> <i>Click on :OK</i></p>
<p>Here are Stem & Leaf Diagrams for the Smith & Jones, and Men & Women Separately.</p>	
<p>Character Stem-and-Leaf Display Stem-and-leaf of Smith N = 25 Leaf Unit = 1.0</p> <pre> 2 6 36 10 7 15666788 (12) 8 112245666689 3 9 134 </pre>	<p>Character Stem-and-Leaf Display Stem-and-leaf of Jones N = 20 Leaf Unit = 1.0</p> <pre> 1 3 3 2 4 4 4 5 27 (13) 6 01113334556779 3 7 146 </pre>
<p>Character Stem-and-Leaf Display Stem-and-leaf of Men N = 25 Leaf Unit = 1.0</p> <pre> 1 3 3 2 4 4 2 5 10 6 01133679 (5) 7 15668 10 8 1124668 3 9 134 </pre>	<p>Character Stem-and-Leaf Display Stem-and-leaf of Men N = 25 Leaf Unit = 1.0 Stem-and-leaf of Women N = 20 Leaf Unit = 1.0</p> <pre> 2 5 27 9 6 1345567 (6) 7 146678 5 8 25669 </pre>

Back to Back Stem & Leaf Diagram

These must be done by hand, since Minitab cannot do them this way.

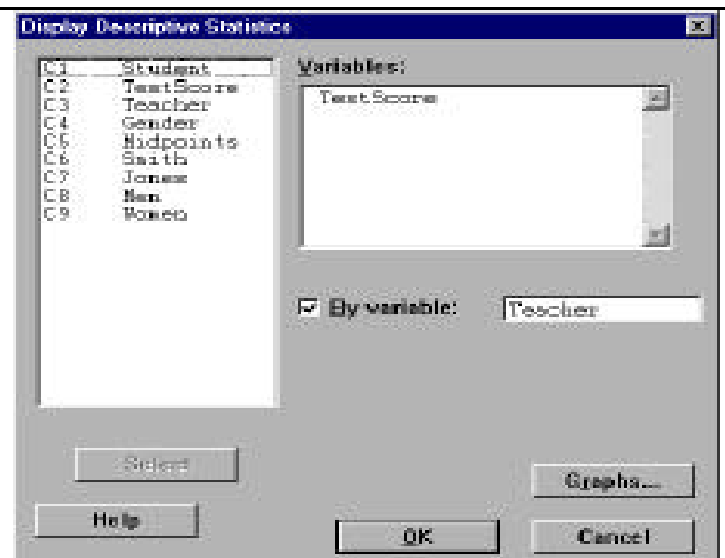
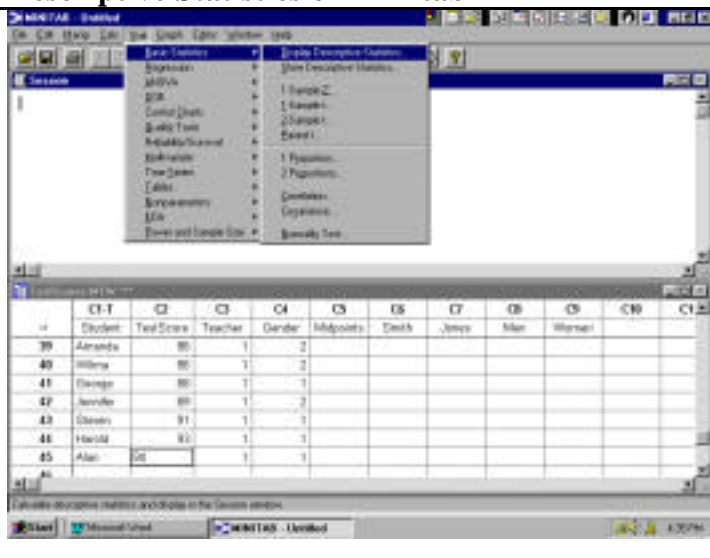
Leaves Smith	Stem	Leaves Jones
	3	3
	3	
	4	4
	4	
	5	2
	5	7
3	6	0111334
6	6	556779
1	7	14
8876665	7	6
42211	8	
9866665	8	
431	9	

Leaves Smith	Stem	Leaves Jones
	3	3
	4	4
	5	27
63	6	0111334556779
88766651	7	146
986666542211	8	
431	9	

Leaves Men	Stem	Leaves Women
3	3	
	3	
4	4	
	4	
	5	2
	5	7
33110	6	134
976	6	5567
1	7	14
8665	7	6678
4211	8	2
866	8	5699
431	9	

Leaves Men	Stem	Leaves Women
3	3	
4	4	
	5	27
98633110	6	13455677
86651	7	146678
8664211	8	25669
431	9	

Descriptive Statistics on Minitab



Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE
Mean						
TestScor	45	72.64	75.00	73.29	12.99	
1.94						

We can also separate the test scores by Teacher or Gender.

Variable	Minimum	Maximum	Q1	Q3
TestScor	33.00	94.00	63.50	83.00

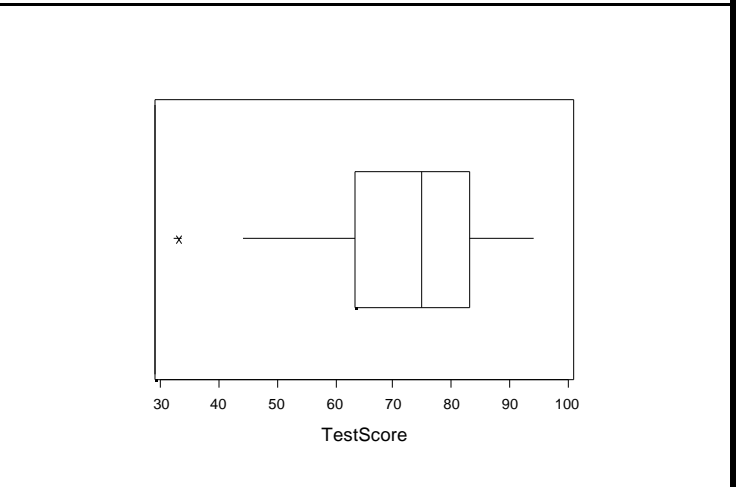
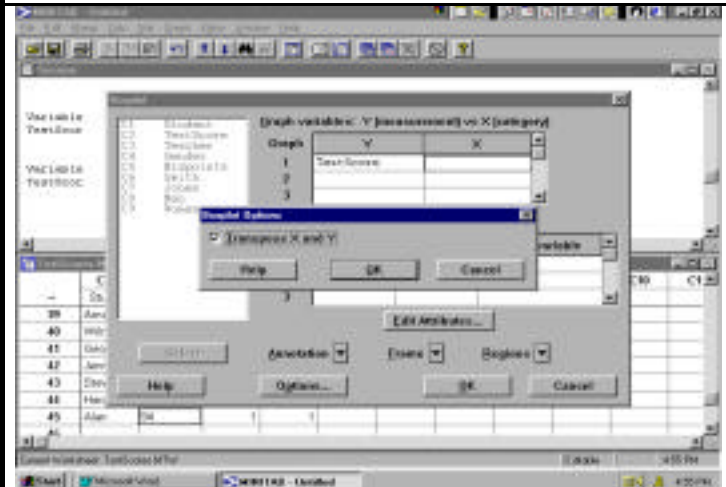
Descriptive Statistics					
Variable	Teacher	N	Mean	Median	TrMean
TestScor	1	25	81.20	82.00	81.43
	2	20	61.95	63.50	62.78
			7.76		
			9.89		

Variable	Teacher	SE Mean	Minimum	Maximum	Q1	Q3
TestScor	1	1.55	63.00	94.00	76.00	86.00
	2	2.21	33.00	76.00	60.25	67.00

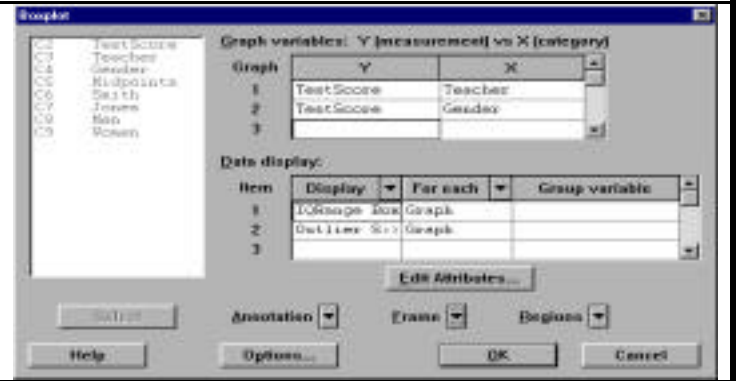
Descriptive Statistics					
Variable	Gender	N	Mean	Median	TrMean
TestScor	1	25	73.16	76.00	74.00
	2	20	72.00	72.50	72.17
			14.87		
			10.51		

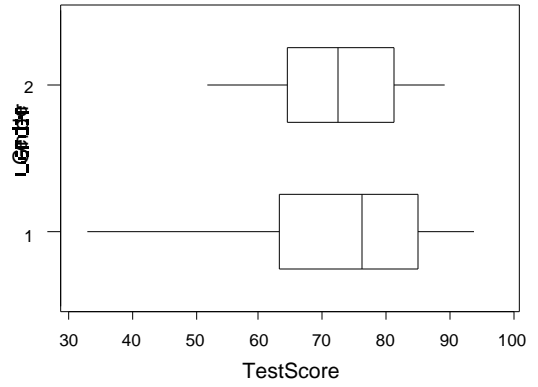
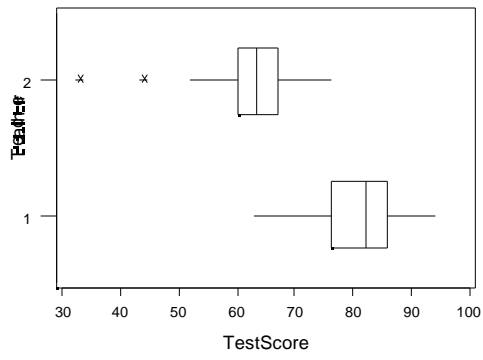
Variable	Gender	SE Mean	Minimum	Maximum	Q1	Q3
TestScor	1	2.97	33.00	94.00	63.00	85.00
	2	2.35	52.00	89.00	64.25	81.00

Box Plots in Minitab.



We can separate the Box Plots by Teacher or Gender as shown in the window displayed to the right.





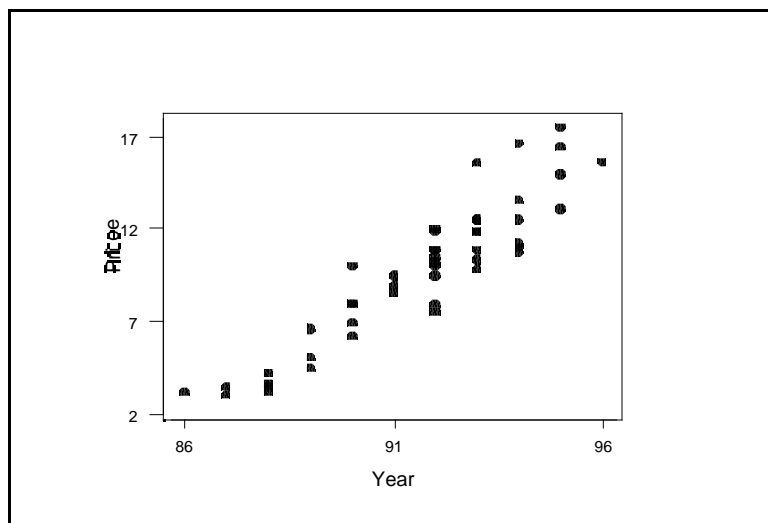
Data Set 1: Prices in Thousands of Dollars of Used Honda Accords in 1997.

Source: 1 = February 14, 1997 edition of “The Philadelphia Inquirer”
 2 = February 15, 1997 edition of “The News Journal”.

Paper	Year	Price	Paper	Year	Price	Paper	Year	Price	Paper	Year	Price	Paper	Year	Price
1	96	15.6	1	93	11.9	1	91	8.5	2	89	6.6	2	92	7.9
1	95	14.9	1	93	11.8	1	90	10.0	2	94	11.0	2	94	16.6
1	95	13.0	1	92	10.9	1	90	7.9	2	93	12.5	2	91	8.9
1	95	16.4	1	92	11.9	1	89	5.0	2	87	3.5	2	92	10.5
1	93	15.5	1	92	10.2	1	89	4.5	2	90	6.9	2	93	10.3
1	94	12.5	1	92	10.0	1	88	3.6	2	90	8.0	2	94	11.2
1	94	10.7	1	92	12.0	1	88	3.5	2	92	10.9	2	93	10.9
1	94	13.5	1	92	9.5	1	88	3.2	2	95	17.5			
1	93	9.8	1	92	7.5	1	87	3.0	2	88	4.2			
1	93	12.5	1	91	9.5	1	86	3.2	2	90	6.2			

Scatter Plot:

Choose: Graph
 Choose: Plot
 Select: y & x Variables



Regression Lines:

Statistics has a way of fitting the ‘best’ line through the data points called regression analysis. To get the regression equation in Minitab, one has to use the regression command. Two options of the regression command box, one called ‘Fits’ and the other called ‘Residuals’, must be checked off to get all of the plots needed. Here is how it work:

Choose: Stat
 Choose: Regression
 Choose: Regression (Again in the box on the right)
 Select: ‘Price’ for the Response Variable and ‘Year’ for the Predictor Variable

Regression Analysis

The regression equation is
 Price = - 124 + 1.46 Year

Predictor	Coef	StDev	T	P
Constant	-124.401	8.332	-14.93	0.000
Year	1.46364	0.09091	16.10	0.000

S = 1.513 R-Sq = 85.2% R-Sq(adj) = 84.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	593.63	593.63	259.21	0.000
Residual Error	45	103.06	2.29		
Total	46	696.69			

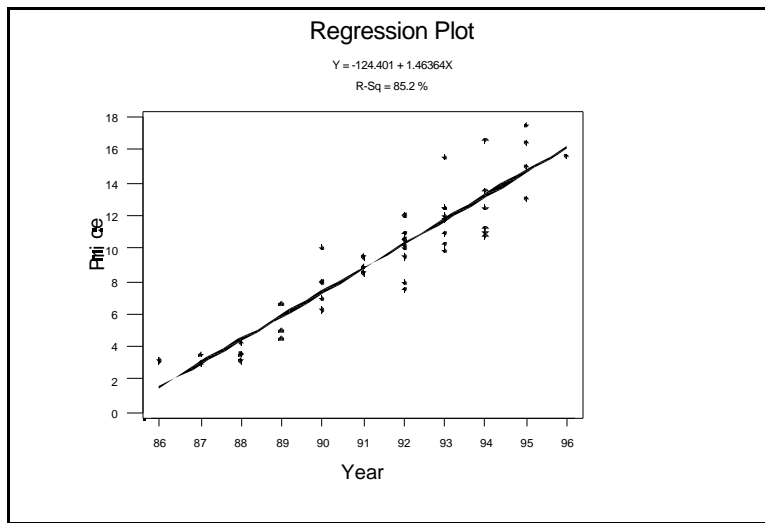
Unusual Observations

Obs	Year	Price	Fit	StDev Fit	Residual	St Resid
5	93.0	15.500	11.718	0.254	3.782	2.54R
30	86.0	3.200	1.472	0.556	1.728	1.23 X
42	94.0	16.600	13.181	0.309	3.419	2.31R

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

Fitted Line Plot:

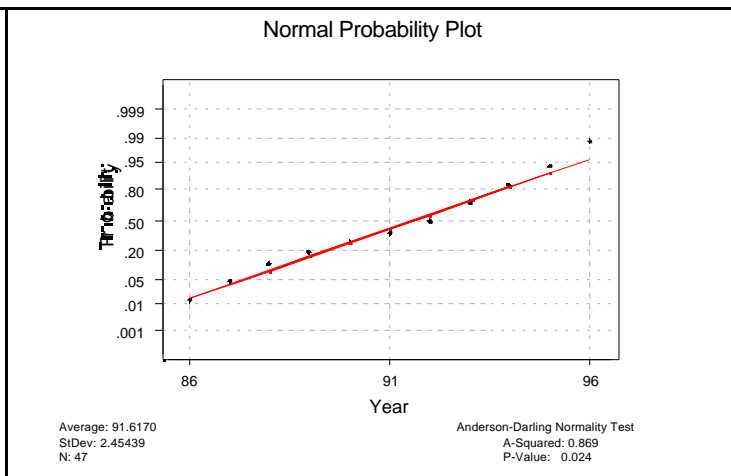
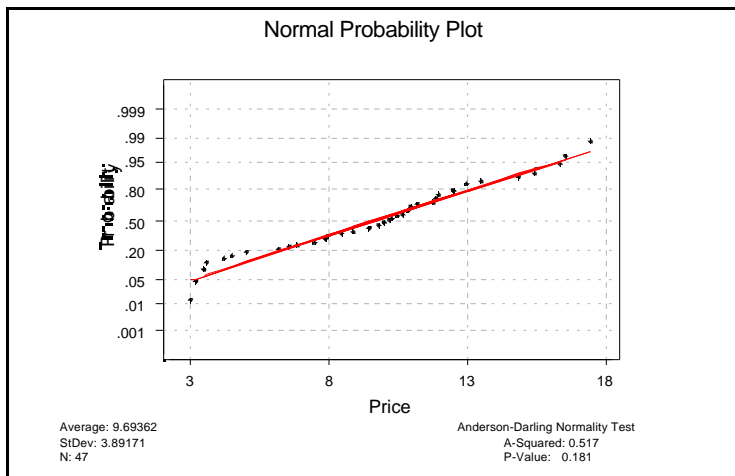
Choose: Stat
 Choose: Regression
 Choose: Fitted Line Plot
 Select: 'Price' for the Response Variable
 and 'Year' for the Predictor Variable



Normal Plots:

An important question in statistics is whether or not the data follows the pattern of the Empirical Rule. In other words, does the data follow the pattern of the normal distribution or is it "normal"? There are several ways to get normal probability plots in Minitab. The One describe here can be reproduced as follows:

Choose: Stat
 Choose: Basic Statistics
 Choose: Normality Test
 Select: the column for your variable



If the data is normal then this plot should be roughly a straight line and the P-value listed in the bottom right hand corner should not be very small. Typically we say that a P-value that is less than 0.05 is too small. The first plot above does appear to be a straight line and its P-value = 0.179 which is not less than 0.05 ($0.179 > 0.05$), so we conclude that the prices of the Honda Accords does follow a normal distribution. The second plot above also appears to be linear; however, its P-value = 0.02 which is less than 0.05 ($0.02 < 0.05$), so we conclude that the prices of the Honda Accords does not follow a normal distribution.

Alternative Fitted Line Plot:

There is another way to get the fitted plot line using the “Plot” command. To do this, you must first store the fitted \hat{y} values in a Minitab column. Storing the fitted values is simple under Minitab. Here is how you do it:

Choose: Stat

Choose: Regression

Choose: Regression (In the box on the right)

Select: ‘Price’ for the Response Variable and ‘Year’ for the Predictor Variable

Click on: Storage

Choose: Fits

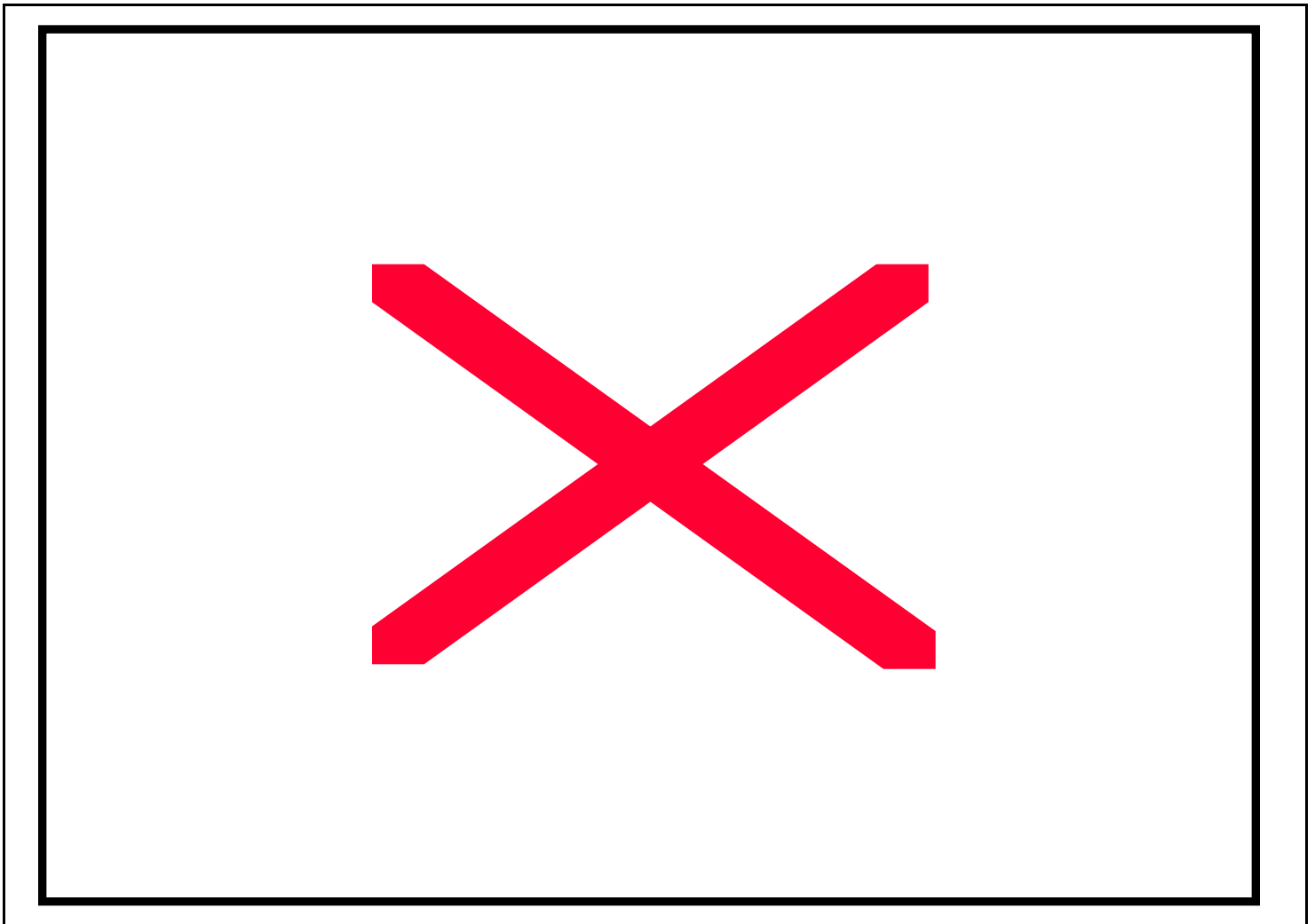
Let’s discuss the ‘Fits’ column. The ‘Fits’ are the y values that are the result of substituting the x values in the regression equation. For example the first few fit are the y value that results from replacing x with the years (96, 95, 94, etc.) in the regression equation $\hat{y} = -124 + 146x$ as shown:

$$\hat{y} = -124 + 14(96) = 1616$$

$$\hat{y} = -124 + 14(95) = 1470$$

$$\hat{y} = -124 + 14(94) = 1324$$

Here is what the Minitab window looks like:

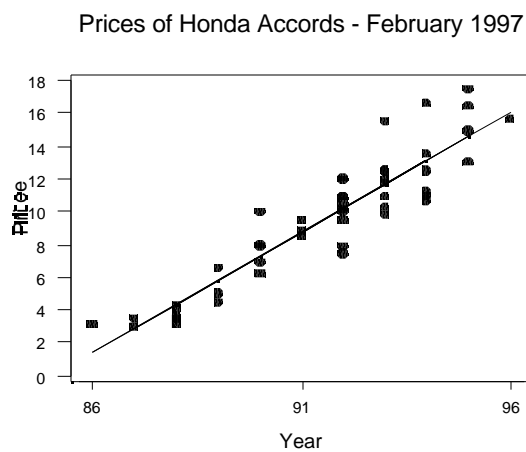


To get a scatter plot of the data along with the regression line one has to do a scatter plot, but under the “Annotation” option one has to choose the “Line” option. In the first box of the first line you have to write in the column names (Year, Fits1) or the column number (C1, C6). Basically you are telling Minitab which columns contain the x & y coordinates for the line. The boxes should be as follows:

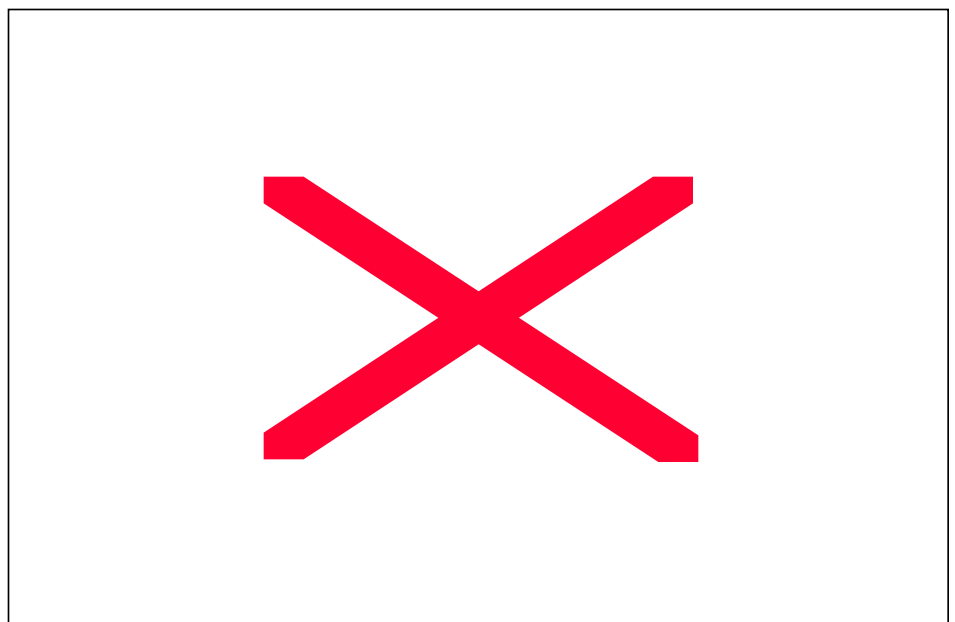
	Points	Type	Color	Size
1	Year Fits1	Solid	Black	1
2				
3				
4				

Here is the result of such a plot:

Note Under “Annotation” the “Title” option was used to get the title shown in the graph.



Here is what the screen looked like in the above plot.



We can have our plot use different symbols for the points from each newspaper by altering the previous plot screen as follows:

Note: the differences from before. We now have:
 “Group” in the “For Each” column & we are using “Paper” for the “Group Variables”.

Plot

C1 Year
 C2 Price
 C3 Model
 C4 Mileage
 C5 Paper
 C6 FITS1

Graph variables:

Graph	Y	X
1	Price	Year
2		
3		

Data display:

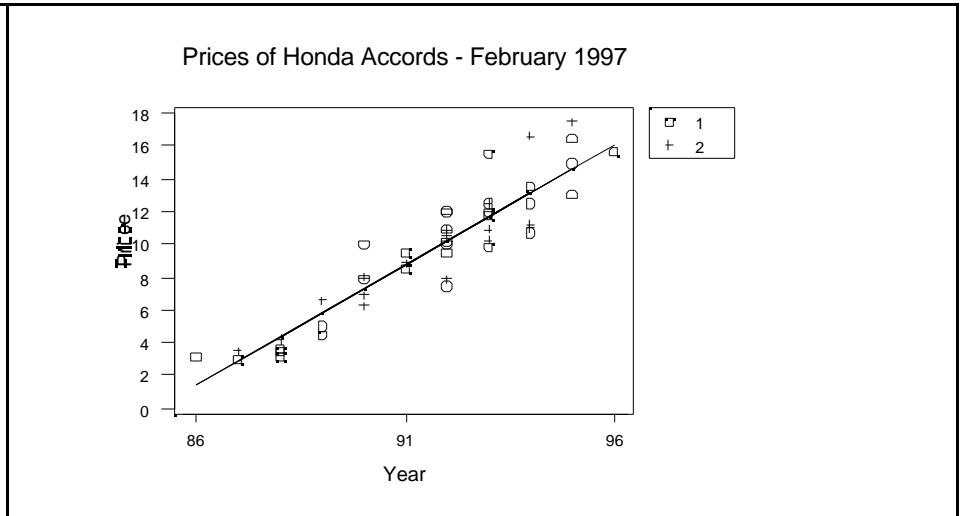
Item	Display	For each	Group variables
1	Symbol	Group	Paper
2			
3			

Edit Attributes...

Select Annotation Frame Regions

Help Options... OK Cancel

Here is what the resulting plot looks like:



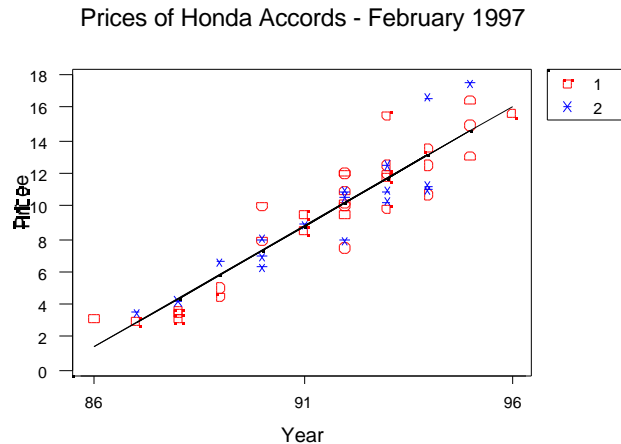
We can even change the color and the symbol used for each point by choosing the “Edit Attributes” box in the graph window. Here is what this menu looks like. Here blue and red have been chosen for the colors and the “plus” sign for the points from the second newspaper has been changed to the “asterisk” sign.

Symbol

Paper	Type	Color	Size
1	Circle	Red	1.0
2	Asterisk	Blue	1.0

Help Defaults OK Cancel

Here is the resulting graph. You can see that the asterisk is used instead of the plus sign, but of course black and white copies cannot show the color.



Extrapolating our linear equation to other years:

Our linear equation, $\hat{y} = -124 + 146x$, was based on model years from 1986 to 1996. Using values of x either above or below the range of values that generated our linear equation is called extrapolation. We often use values of x in our equation to predict values of y . For example we may be interested here in the price we would expect to pay for a 1984 Honda Accord. We could calculate that easily by substituting 85 into our equation.

$$\hat{y} = -124 + 146(85) = 01$$

Based on our equation, a 1985 Honda Accord should cost \$100. Well if a 1985 Honda Accord “should” sell for \$100, then what should we expect to pay for a 1984 Honda Accord. Here is what our linear regression equation predicts.

$$\hat{y} = -124 + 146(84) = -136$$

Does that mean that the seller would have to pay us \$1,360 to “buy” the seller’s 1984 Honda Accord? Clearly this is absurd. By using the years 1985 and 1984 in our linear equation, we are assuming that the linear relationship observed extends to model years that were not in our original data set. Sometimes this assumption may be reasonable, but in this case it is clearly not possible. Eventually our y values will have to get negative and that just cannot happen. The linear relationship we observed appears to hold for the model years of our data, 1986 – 1996, but it should not be extrapolated to other model years.

Data Set 2: Prices in Thousands of Dollars of Used Honda Accords in 1999.

Source: 1 = February 7, 1999 edition of “The Baltimore Sun”
 2 = February 7, 1999 edition of “The News Journal”.

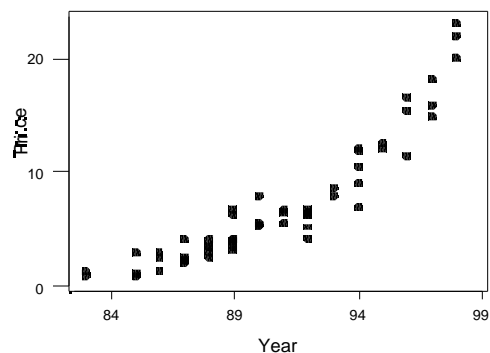
Paper	Year	Price	Paper	Year	Price	Paper	Year	Price	Paper	Year	Price	Paper	Year	Price
1	98	23.0	1	94	9.0	1	90	5.4	1	87	2.2	2	90	5.4
1	98	21.9	1	93	8.7	1	92	5.2	1	87	2.1	2	88	3.5
1	97	18.0	1	90	8.0	1	92	4.2	1	86	1.4	2	88	3.5
1	96	16.5	1	93	8.0	1	87	4.0	1	83	1.3	2	89	3.5
1	97	15.8	1	94	6.9	1	88	4.0	1	85	1.2	2	88	3.0
1	96	15.4	1	92	6.8	1	89	4.0	1	83	0.9	2	88	2.5
1	97	14.8	1	91	6.7	1	89	4.0	2	98	20.0	2	86	2.5
1	95	12.5	1	91	6.5	1	89	3.8	2	94	11.9	2	87	2.3
1	94	12.0	1	92	6.5	1	89	3.1	2	89	6.8	2	85	0.9
1	95	12.0	1	92	6.4	1	85	2.9	2	92	6.5			
1	96	11.4	1	90	5.5	1	86	2.9	2	89	6.3			
1	94	10.5	1	91	5.5	1	87	2.5	2	92	6.2			

Question 1: What should we do first?

Answer 1: Do a scatter plot of the data!

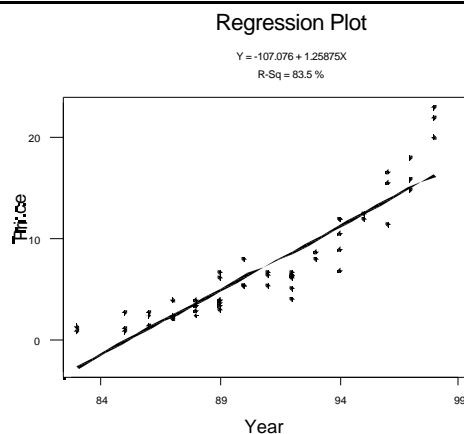
Question 2: Does the graph look linear?

Answer 2: No, it appears to have a curve to it. In other words, it is curvilinear.



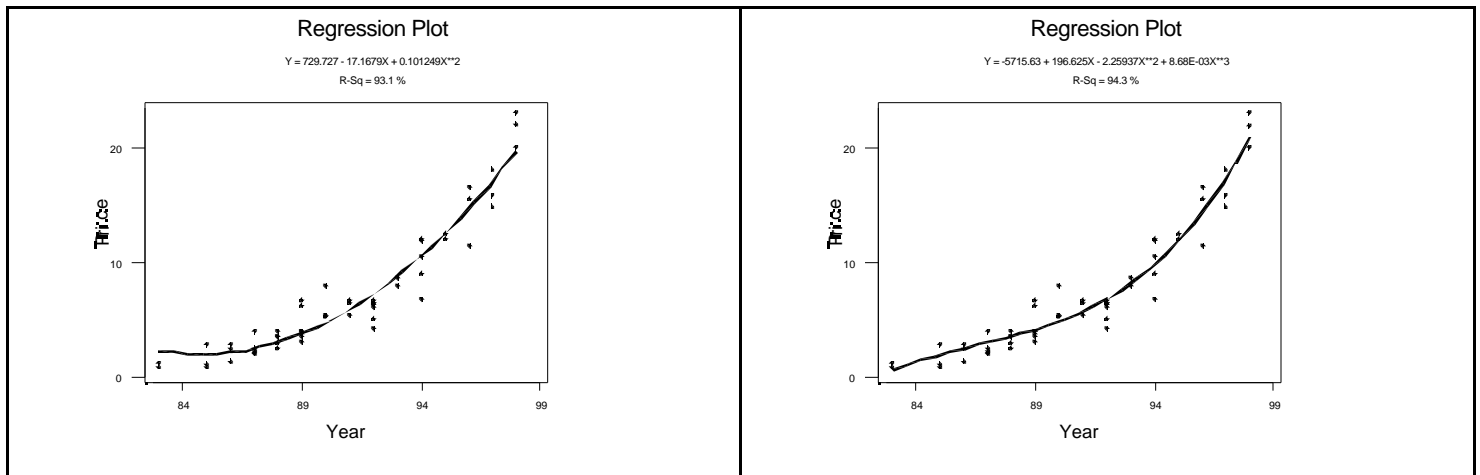
Question 3: What happens if you fit a straight line to this data?

Answer 3: Minitab will do what you tell it. Here is a fitted line plot for this data. Notice how the points seem to be above the line, then below it and finally above it again. One should not see such a pattern around the line. If the two variables were linear, the points should be distributed randomly around the line.



Question 4: What curve should we fit to the data?

Answer 4: One could try a quadratic equation or a cubic equation. As one can see in the graphs below, both of these graphs fits the data better than a straight line; however both of these curves have properties that make them unsatisfactory choices. Note: these two graphs were plotted selecting the “quadratic” and “cubic” choices respectively for the “fitted Line Plot”.



Question 5: Why is this data not Quadratic?

Answer 5: A quadratic equation takes the shape of a parabola, so if we were to continue this curve to the left we would see it eventually rise. But that would mean that much older cars would start to rise in price, and that is not what we would expect.

Question 6: Why is this data not Cubic?

Answer 6: If one looks at the cubic graph above, we can see that the graph takes a downward turn on the left side. This would mean the price of much older cars would eventually be negative. How can the price of an older Honda be negative? Would the owner have to pay us to “buy” it?

Question 7: So what is the correct shape of this graph?

Answer 7: There is an exponential relationship between these two variables! Here is the mathematics of this relationship.

Exponential Function:

$Y = c \cdot a^X$, where a & c are constant.

If we take the log of both sides of this equation and apply the laws of logarithms we get:

$\log(Y) = \log(c \cdot a^X)$

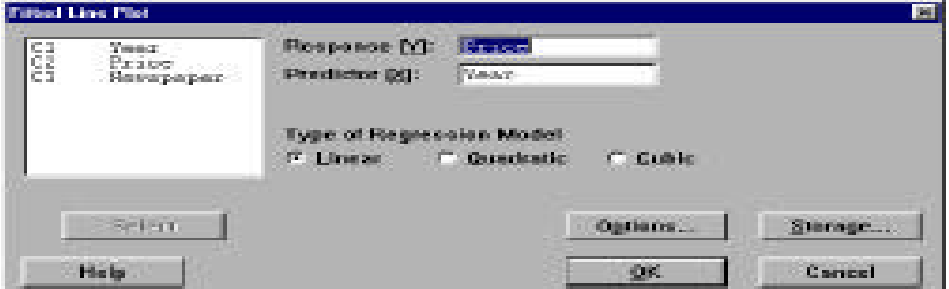
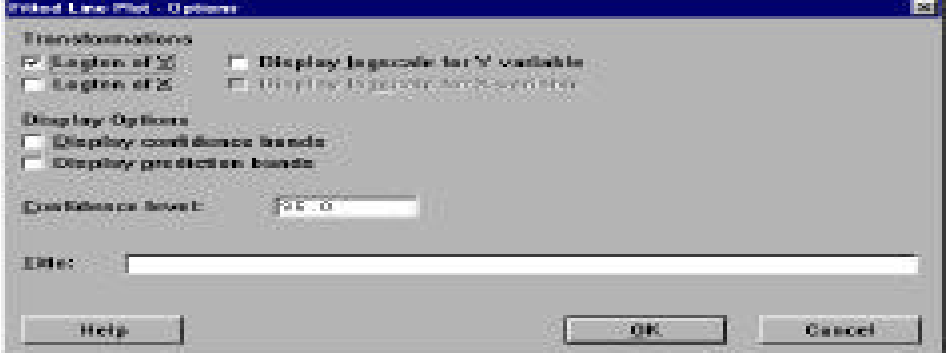
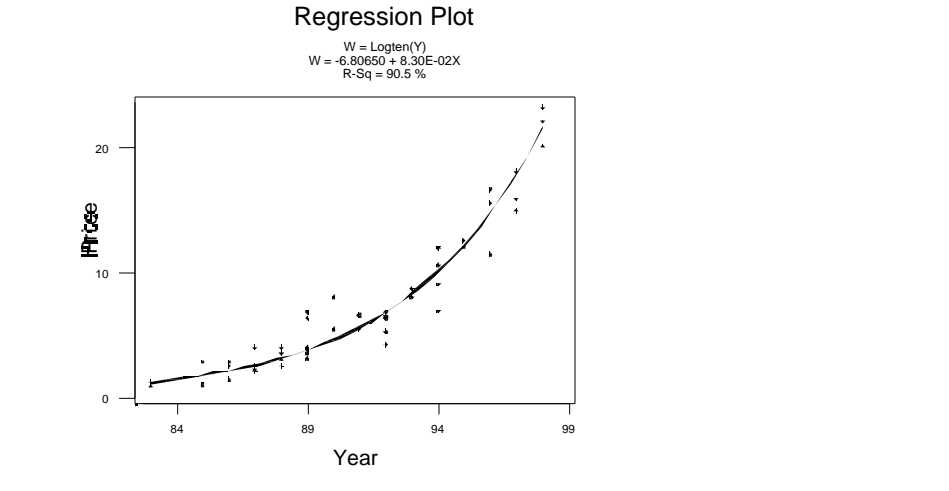
$\log(Y) = \log(c) + \log(a^X)$

$\log(Y) = \log(c) + X \log(a)$

Note: the resulting equation has two variables X and $\log(Y)$ and that the right side is linear with a slope of $\log(a)$ and y-intercept of $\log(c)$.

Question 8: Does one have to know this mathematics to have Minitab fit an exponential function to this data?

Answer 8: No, it can be done quite easily as follows.

<p>Choose: <i>linear</i></p> <p>Click on: <i>Options Box</i></p>	
<p>This is the screen that you get when you click on the <i>options</i> box.</p> <p>Simply check off the “Logten of Y” box.</p>	
<p>Here is the graph you get. Observe that our regression equation is: $W = -68065 + .0830X$</p>	

Question 9: How do we use this equation?

Answer 9: Here is how to use it.

Suppose we wish to use this equation to predict the price of a 1988 Honda Accord.

Step1: Substitute 88 into this equation: $W = -68065 + .0830(88) = 04975$. This is telling us that $\log(y) = 04975$.

Step2: To find the value of y, undo the log function as follows: $y = 10^{04975} = 314$. Thus a 1988 Honda Accord should cost about \$3,140 which is close to the price of the 1988 Honda Accords in our data set.

Question 10: Can we extrapolate this exponential equation to earlier model year Hondas?

Answer 10: Yes, we will not get negative values for price with this model. Here are a couple of examples.

Model Year	Value of W	Value of Y	Cost of the Honda Accord
1982	$W = -68065 + .0830(82) = -.00005$	$Y = 10^W = 10^{-000005} = .999$	\$999
1981	$W = -68065 + .0830(81) = -.0835$	$Y = 10^W = 10^{-00835} = 825$	\$825
1980	$W = -68065 + .0830(80) = -.1665$	$Y = 10^W = 10^{-01665} = 682$	\$682
1979	$W = -68065 + .0830(79) = -.2495$	$Y = 10^W = 10^{-02495} = .563$	\$563

As you can see, extrapolating the exponential model to earlier years give reasonable answers with this model; whereas extrapolating the linear model of the linear model gave unsatisfactory results.

```
MTB > Retrieve 'C:\MTBW\IN\DATA\Mat127\Bank-rob.mtw'.
Retrieving worksheet from file: C:\MTBW\IN\DATA\Mat127\Bank-rob.mtw
Worksheet was saved on 10/ 2/1994
```

```
MTB > Name c5 = 'FITS1'
MTB > Regress 'Robbery' 1 'Unemploy';
SUBC> Fits 'FITS1';
SUBC> Constant.
```

Regression Analysis

The regression equation is
 Robbery = - 71.5 + 1.64 Unemploy

Predictor	Coef	StDev	T	P
Constant	-71.46	37.01	-1.93	0.059
Unemploy	1.6409	0.1908	8.60	0.000

S = 198.5 R-Sq = 60.1% R-Sq(adj) = 59.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2914033	2914033	73.94	0.000
Error	49	1931179	39412		
Total	50	4845212			

Unusual Observations

Obs	Unemploy	Robbery	Fit	StDev Fit	Residual	St Resid
10	359	117.0	517.6	52.1	-400.6	-2.09R
38	567	208.0	858.9	88.3	-650.9	-3.66RX
48	737	2161.0	1137.9	119.5	1023.1	6.45RX

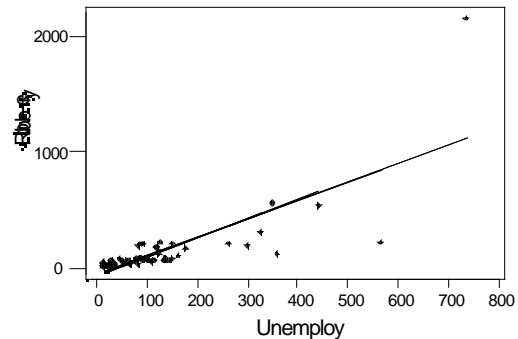
R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

```
MTB > Plot 'Robbery'*'Unemploy';
SUBC> Symbol;
SUBC> Line 'Unemploy' 'Fits1'.
```

Graph Plot
 Choose: Annotation
 Line For Points Choose: Unemploy Fits1

```
MTB > Correlation 'Robbery' 'Unemploy'.
Correlations (Pearson)
Correlation of Robbery and Unemploy = 0.776
```

Note: $r = 0.776$ and $r^2 = (0.776)^2 = 0.602176 = 60.2\%$ which is the same as the Coefficient of Determination.



Was our sample a random sample? (Explain!) _____

List potential Biases (if any). _____

Find a 95% confidence interval for the mean height of Cecil Community College students.

_____ Is the sample size large enough for us to use the central limit theorem here? (Why?)

_____ What assumption must we make about the distribution of student's height?

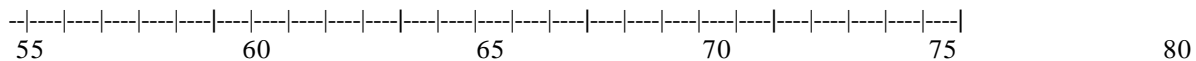
_____ Give the proper formula to use for k (**give complete standard deviation**).

_____ What is the correct critical t or z value you should use?

_____ What is the numerical value of k (rounded to the nearest **tenth**)?

_____ < < _____ (2 pt.) Give the confidence interval (rounded as above for k).
Note: Place proper symbol between inequality signs.

Draw a box plot of the student heights on the grid below. Place the sample mean in your box plot using a large dot. Use parentheses on the number line given to mark off the left and right endpoints of your confidence interval.



How large of a sample would we need to estimate the height of Cecil students, if we wanted to be 95% confident that our error was no more than _ inch?

_____ Give the proper formula to use to calculate n.

_____ What is the critical t or z value you should use here?

_____ Give the value of n (**rounded up to the next whole number**).

Problem 2: Are peoples' arm spans equal to their heights?

What is our target population? _____

Was our sample a random sample? (Explain!) _____

Do you feel that our sample is representative of the population? _____

Find a 95% confidence interval for the mean difference in height and arm length.

_____ What assumption must we make about the distribution of these differences?

_____ Give the proper formula to use for k (**give complete standard deviation**).

_____ What is the correct critical t or z value you should use?

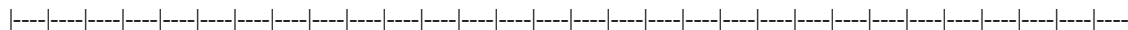
_____ What is the numerical value of k (rounded to the nearest **tenth**)?

_____ < _____ (2 pt.) Give the confidence interval (rounded as above for k).
Note: Place proper symbol between inequality signs.

_____ If the student is correct, what would we expect the true mean difference (μ_D) to be?

Based on the confidence interval, explain whether or not you agree with the students claim. _____

Draw a box plot of the differences on the grid below. Place the sample mean in your box plot using a large dot. Use parentheses on the number line given to mark off the left and right endpoints of your confidence interval.



Hypothesis Testing: Is there sufficient evidence to conclude that there is no difference between peoples' heights and arm spans? Use a level of significance of 0.05. Summarize what is known about the sample and the population, and based on them, give the results for the critical criteria (sample size, normality, and known variance) needed to choose the correct statistical formula.

Sample:

Population:

Critical Criteria Summary:

H₀: _____

H₁: _____

α = _____

Critical Region (Rejection Zone): _____

Computational Formula: _____

Value of Computational Formula: _____

Decision: _____

