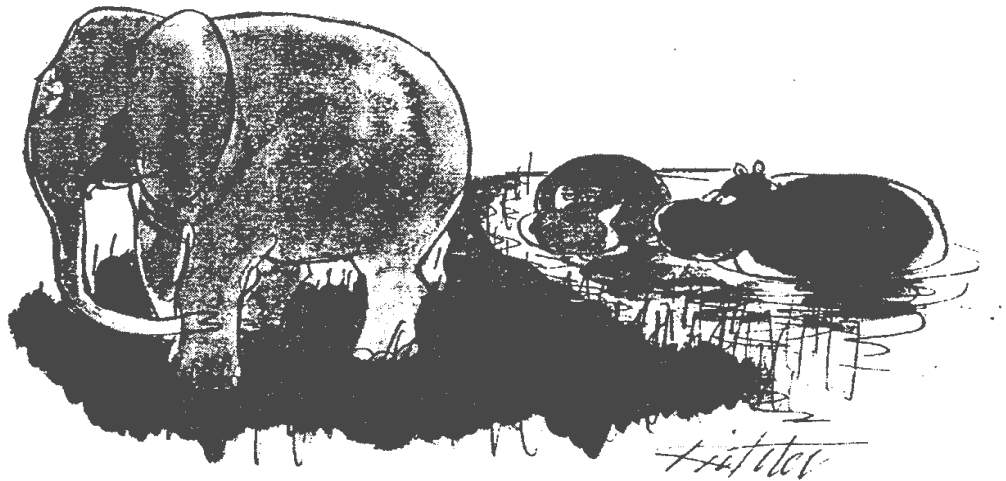


# Data Analysis 2



*"Oh, sure, he remembers everything. But show me one significant insight he's been able to draw from all that data."*

TJHSST

## Discussion 1: The Vocabulary of Statistics

**Statistics** is the study of the best way to collect, describe and draw conclusions from data. When we collect data, we use surveys or experiments. A survey is simply a gathering of data. An experiment differs from a survey because a "treatment" or "external force" is applied to the objects being studied. It is very important in an experiment to control conditions carefully so that the factor being studied is not confused with the effects of factors not of interest. When we describe data we use tables, graphs, charts and numerical measures. When we make conclusions from data or draw "inferences" we use the data we have collected in conjunction with probability standards to make predictions.

A **population** is a group of objects about which we wish to gain information. It is important that the population be well defined. We must avoid vague words such as child, good, bad, large, etc.

**Exercises:** Determine whether the following populations are well defined. Underline any vague terms.

1. The population of teams that comprise the Concorde District.
2. The population of residents in Northern Virginia.
3. The population of large colleges in the United States.
4. The population of economy cars manufactured in Japan.
5. The population of students currently enrolled at the University of Virginia.
6. The population of nutritious breakfast cereals on the market.
7. The population of men that have served as President of the U. S.
8. The population of expensive homes in Langley, Virginia.

Because it is often difficult and expensive to study an entire population, generally a **sample** or a portion of the population is used to gain information about the whole population. It is very important that this sample be representative of the entire population. A **random sample** is a sample in which each member of the population has an equal chance of being included. A **nonrandom sample** is a sample where subjective or arbitrary choices of members of the population are made so that biases are built in. Biases can be visible (obvious) or invisible (subtle).

**Exercises:** In the following examples, distinguish the random from the nonrandom samples. Discuss any visible or invisible biases that may have occurred.

1. A farmer wants to be certain that fewer than 2% of the tomatoes in a shipment are rotten, so he tests four tomatoes taken from the top of the load.
2. A pollster wishes to determine the proportion of voters who will vote for Candidate A in the next election. The pollster conducts a survey at the local Young Democrats Luncheon.
3. A school newspaper wants to survey the attitudes of students toward the college's drama program. The editors obtain a computer print-out of all students in school, and assign each student a number. These numbers are then put in a box and mixed well, and sample numbers are drawn out. The students who correspond to these numbers are interviewed.
4. A farmer is concerned about a skin infection in his pigs. He examines the first five pigs he can catch.

## Discussion 2: Manipulating Data

The difficulty in finding a truly random sample often leads to incorrect interpretation of data. The misinterpretation is often unintentional. There was a classic example of misusing a sample during the 1936 presidential election. Ten million people were surveyed and the results of the survey indicated that Alf Landon would easily defeat Franklin Roosevelt. Roosevelt won handily. This was a huge survey. What went wrong? The question to really ask is what was wrong with the sample? Was it truly random? The survey was done by *The Literary Digest*. Their target audience was their readership and a widespread telephone poll. In 1936, the height of

the depression, people who had a subscription to this magazine or even a telephone were not representative of the general population. It was a biased sample.

Surveyors can intentionally or unintentionally bias their samples by disregarding what may seem like unimportant information. The surveyors from *The Literary Digest* believed that because they had such a large sample, they had a random sample. They did not intentionally bias their sample. The bias was invisible. There are times that political pollsters, advertising executives and others purposely bias their samples to show their candidates or products or opinions in the best light.

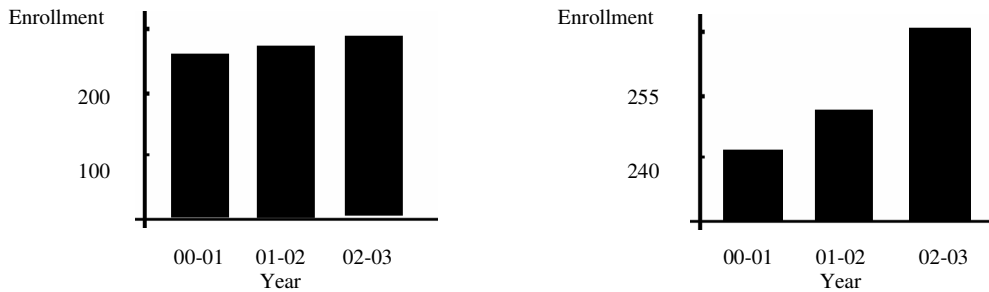
There are many other ways of misrepresenting data other than manipulating the sample. As visual aids, graphs are an efficient and effective means of organizing data. Because they are so deceptively simple, many readers of statistics scan the graphs rather than analyzing them carefully. A quick scan can lead the reader to misinterpret the data being presented.

Consider the following fictitious data:

TJHSST BC Calculus Enrollment

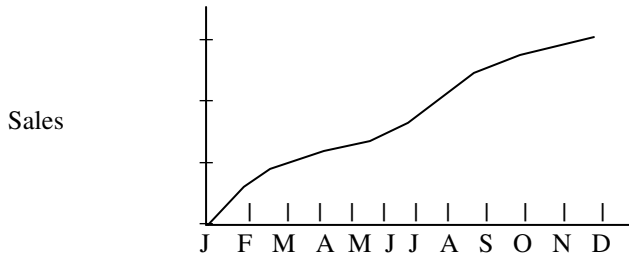
Year	Enrollment
2000-01	240
2001-02	250
2002-03	270

The data has been presented in two different graphs.

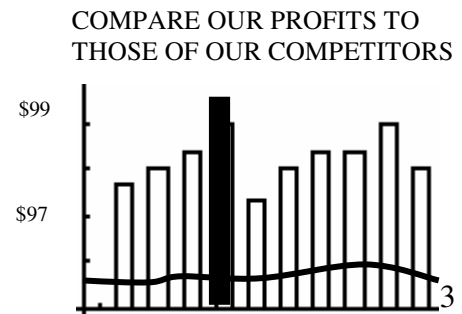


These graphs represent the same data, but the first graph is the least likely to be misinterpreted because the vertical axis has not been truncated. Truncating this axis is a very common method of emphasizing or de-emphasizing differences in data. Not only is the axis commonly truncated, it is not unusual for labels on either axis to be missing.

This graph is meaningless without numbers on the y-axis



Bar graphs like the one at the right are often seen in advertising.



Finally, you have to use common sense. For example, when an advertisement states that 3 out of 4 dentists recommend sugarless gum, there are several questions you should ask yourself: How many dentists were surveyed? Who did the survey? The dentists recommended sugarless gum; did they actually recommend the advertised brand?

**Exercise:** Collect three examples from newspapers and magazines of graphs or of other statistical claims that present data in a deceptive manner. Identify the source from which each example was taken. Explain briefly the ways in which each example might have been deceptively presented and then suggest ways the data might be presented more fairly or in a less distorted fashion. An original or photocopy of the graph or statistical claim must be included with the project. The examples must illustrate three different techniques that misrepresent data (i.e., you may not have three truncated graphs).

### Discussion #3: The Line of Best Fit

Once bivariate data has been graphed on a scatter plot, the statistician wishes to determine (1) if there is some type of relationship between the two pieces of information from each observation, (2) how strong this relationship is and (3) an equation expressing this relationship so that predictions can be made. Lines or curves can result from the plots. We will begin by looking at linear relationships and discuss the process for fitting a line to the data.

The technique studied in Data Analysis I involved the median-median line. A review of the technique is found below.

Finding Median-Median Summary Lines:

- (1) Separate the data into three groups of equal size (or as close to equal as possible) according to values of the horizontal coordinate.
- (2) Find a summary point for each group using the median x value and the median y value (found by first sorting the y values in each group).
- (3) Find the slope-intercept equation for the line through the left most and right most point.
- (4) Calculate the residual of the middle point, i.e., the vertical distance the middle point is from the equation found in #3.
- (5) Add one third of this residual to the y-intercept in order to find the best fit equation.

**Exercise:** The data below, from the Chicago Bulls' basketball team, describe the total points scored in the 1987-88 season vs. the number of minutes played during the season.

Player	Minutes Played	Total Points
Brown	591	197
Corzine	2328	804
Grant	1827	622
Jordan	3311	2868
Oakley	2816	1014
Paxon	1888	640
Pippen	1650	625
Sellers	2212	777
Sparrow	1044	260
Vincent	1501	573

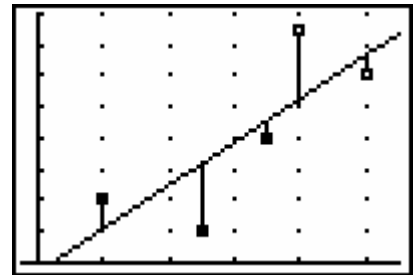
*1989 World Almanac*

- a. On your own paper, construct a graph of this information. Be sure to accurately label and scale the axes.
- b. Describe your graph. Are there any outliers?
- c. Find the median-median line for the data points.

- d. Predict the total points scored if Sellers played 2000 minutes.
- e. If one of the players scored 900 points, how many minutes did he play?
- f. What is the meaning of the slope of your median-median line?
- g. What is the meaning of the y-intercept?

The second commonly used "best fit" line is called the Least Squares Line. This line is very difficult to find without a computer or graphing calculator because it uses all of the points and it minimizes the sum of squared errors. Karl Friedrich Gauss developed this method of finding the "best fit" line. What he did was to find the vertical distances from each point in the scatter plot from a line drawn through the data (unlike the median-median line which only uses the middle representative distance). He found the average of the squares of these distances. The line that has the smallest average distance from each point in the scatter plot is the "best fit" line.

The points (2, 4), (5, 2), (7, 8), (8, 15), and (10,12) are graphed on the following grid. If you took a straight edge and tried to fit a line so that the residuals are minimized, it should look like the one pictured which has an equation of approximately  $y = (7/5)x - 2/3$ . When more than three points are involved it is virtually impossible to completely zero out the residuals, so we look for the line that gives us the smallest average of the squared distances. The reason that the average of the squared distances is used is that it is easy to define average and distance algebraically and program a calculator or computer to find the minimum value.



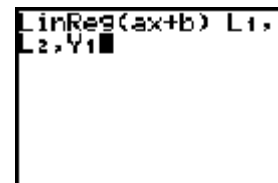
A later mathematician coined the phrase regression line, meaning that as the averages got closer to the "best fit" line, their values "regressed". The terms linear regression line and "best fit" line have become synonymous. Generally, this linear regression method has the best inferential properties, i.e., predictions based on this line have the smallest margin of error. However, just like the mean, it is very sensitive to a few extreme values.

You can use the TI-83 calculator to find both the median-median and linear regression line. The following procedure outlines the process to find the linear regression line. It is easily adaptable to finding the median-median line.

To find the slope and y-intercept of the linear regression line, press [STAT] < ► > (to highlight **CALC**)

Move the cursor down to **4:LinReg (ax +b)**. Press <ENTER>.

**LinReg(ax + b)** will appear on the home screen. You must indicate which two lists you are working with, and if you wish, you may also automatically place the equation of the line into graphing screen so that you can plot the line easily. Your screen should appear as follows once you select your lists and the location of your equation.



Then press <ENTER>. The values for a (slope) and b (y-intercept) will appear on the screen. If you press <Y=>, you will find that the equation has been pasted into Y1.

**Example:** Enter the Bulls basketball data from p 4 into List 1 and List 2.

- a. Use your calculator to find the equation of the linear regression line, record it here and store it in Y1.
- b. Now, use the calculator to find the median-median line. Follow the previous instructions choosing **3:Med-Med**. Record it here and store the equation of the line in Y2 so that you have both equations available for graphing.
- c. According to the linear regression line, if someone on the team played 1750 minutes, how many points did they score?
- d. According to the median-median line, how many minutes must a player play, before he begins to score points?
- e. According to the linear regression line, if someone on the team played 2000 minutes how many points did he score?
- f. According to the median-median line, if someone on the team played 2000 minutes how many points did he score?
- g. Examine the differences between the two lines. Which do you think is a better model for this data, the linear regression line or the median-median line? Explain.

**Exercise:**

In the last two decades there has been a great deal of concern about the possible extinction of the manatee that lives in the coastal waters of Florida. Many environmentalists have blamed the increasing number of power boats for the rising death rate for manatees. When the bodies of manatees are found washed up on shore, they are often marked with propeller cuts and scraps. The Florida Department of Natural Resources collected the following data:

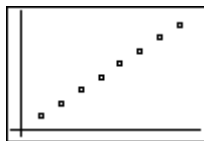
Year	Power Boat Registrations (in thousands)	Manatees Killed
1977	447	13
1978	460	21
1979	481	24
1980	498	16
1981	513	24
1982	512	20
1983	526	15
1984	559	34
1985	585	33
1986	614	33
1987	645	39
1988	675	43
1989	711	50
1990	719	47
1991	716	53
1992	716	38
1993	716	35
1994	735	49

- Plot the data and find the linear regression line for the data (Save the data with names; i.e., BOAT and MANT. You will use it again.)
- What is the real world meaning of the slope of the line? What is the real world meaning of the y-intercept for the line?
- If the number of power boat registrations increases to 750 thousand, what do you predict the number of manatees killed will be?
- Does the information support the environmentalists assertion that the increased number of power boats accounts for the rising death rate for manatees? Explain.

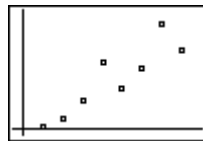
(Adapted from Technology and the Mathematics Curriculum through Functions sponsored by the Woodrow Wilson National Fellowship Foundation, Summer 1994)

#### Discussion #4: Correlation

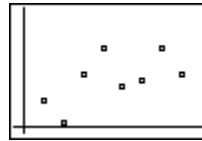
Correlation, represented by the symbol  $r$ , measures how strong the linear association is between two variables, on a scale from -1 to +1. The closer the value of  $r$  is to +1 or to -1, the higher the correlation. So, a correlation of 0 means that there is no relationship between the two variables. Below are eight diagrams illustrating various correlations. If there is a "high" correlation between two items, this means that the items are closely related; that is, they have an effect on each other. If most of the data points of a scatter plot lie close to the regression line, there is a high correlation.



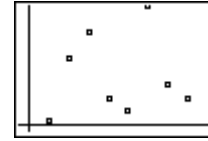
$r = +1$



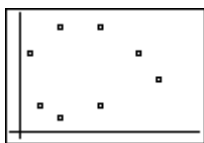
$r = +0.9$



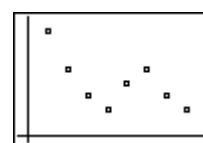
$r = +0.6$



$r = 0$



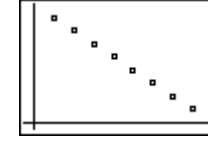
$r = 0$



$r = -0.6$



$r = -0.9$



$r = -1$

There is a formula for finding the correlation coefficient, called  $r$ , of a set of data. The calculator is programmed to find this value for you. In order to have the calculator display  $r$  when you calculate regression lines, press <CATALOG> and scroll down to **DiagnosticOn**. When the cursor is pointing to **DiagnosticOn**, press <ENTER>.

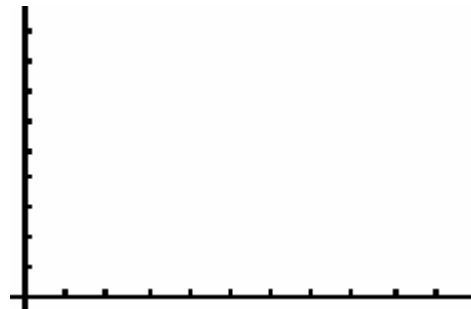
Example: Honda Resale. The following data indicates the resale value of 13 Honda Accords and the age of the car at the time of resale.

x: Age	y: Price
5	11995
5	9900
6	9000
6	8000
7	6900
7	6000
7	8700
7	5650
7	7300
7	7000
8	6999
9	5500
10	3500

1. Graph the scatter plot and reproduce it below clearly indicating the scale.

2. Before finding the correlation coefficient, predict what you think  $r$  will be and tell why.

3. Find the linear regression line for this data.



4. Find the correlation coefficient,  $r$ . Is there a high linear correlation between these two variables? Explain.

Recall that the information you just found is for the line of "best fit". But is a line the best shape to predict the distance given the time? Perhaps a logarithmic, exponential or power function would better fit this data. The TI 83 has a long list of regression curves which are available for approximating scattered  $x$ - $y$  data to some "best fit" curve. The list below may be displayed by pressing **<STAT>** **<CALC>**

- 3: Med-Med**
- 4: LinReg(ax+b)**
- 5: QuadReg**
- 6: CubicReg**
- 7: QuartReg**
- 8: LinReg(a+bx)**
- 9: LnReg**
- 0: ExpReg**
- A: PwrReg**
- B: Logistic**
- C: SinReg**

Most of the regression analyses will give a value for the correlation coefficient  $r$ , a measure of the *linear relationship* between the data sets. For example, when you tell the calculator to perform LnReg, it is taking the logarithm of the data in order to straighten it. The  $r$  reports the measure of the linear relationship between the linearized data sets. This is how the calculator finds the  $r$  and the regression equation; by performing a transformation on the data to linearize it. The closer the absolute value of  $r$  is to 1, the better the regression curve "fits" the data stored. Of course, if  $|r| = 1$ , then the data exactly lies on the selected curve AND then can be used to exactly predict other points on the curve. Before looking at further examples, we will define what each of these curves means.

**3: Median-Median Line:** A line which has slope determined by median values in the first and third sectors of the graph. The line is parallel to and one-third of the vertical distance from the line through these two median points and the median point in the middle sector. The correlation coefficient,  $r$ , is not provided.

**4: Linear regression:** A line of the form  $y = ax + b$ , this is the regression line of best fit found by using the least sum of the squares of the deviations.

**5: Quadratic regression:** This is a parabola of the form  $y = ax^2 + bx + c$ .

**6: Cubic regression:** This is a cubic polynomial of the form  $y = ax^3 + bx^2 + cx + d$ .

**7: Quartic regression:** This is a quartic polynomial of the form  $y = ax^4 + bx^3 + cx^2 + dx + e$ .

**8: Linear regression:** This is the same as #4 but the line is  $y = a + bx$ . This was the traditional way of expressing the linear regression line in statistics but students found it confusing because  $b$  was the slope rather than the intercept, so #4 is also provided.

**9: Logarithmic regression:** This is a logarithmic curve,  $y = a + b(\ln)x$  that uses the natural logarithm (base  $e$ ).

**0: Exponential regression:** This is an exponential curve of the form  $y = ab^x$ .

**A: Power regression:** This is a power curve of the form  $y = ax^b$ .

**B: Logistic regression.** This is a curve of the form  $y = \frac{c}{1 + ae^{-bx}}$  commonly used to describe population patterns and utilizing on the natural exponential base  $e$ .

**C: Sinusoidal regression.** This is a sine curve of the form  $y = a \sin(bx + c) + d$ .

#### Discussion #5: Residuals and Re-Expression of Data

Employment of hourly workers at GM has shrunk to just slightly more than half the level of 15 years ago. The Los Angeles Times (9/25/93) reported that

*"In the midst of difficult labor negotiations, General Motors Corporation is seeking to slash its hourly work force by as many as 50,000 more jobs in the next three years, sources said. The cuts would be in addition to 54,000 blue-collar jobs the company previously announced it would trim by the mid-1990's. About 39,000 of those jobs have already been eliminated.*

*The downsizing is part of a process that began in 1991, when the company said it would lay off 74,000 blue-and-white collar workers and close nearly two dozen plants. The effort has intensified in the last year under Chief Executive John F. Smith.*

*GM spokesman Jack Harned said reports of the new job cuts, first published Friday in the Detroit News, were "speculation". But he added, "We plan to have our work force at competitive levels with our competition." Analysts said that GM needs to reduce its work force to 200,000 - 220,000 hourly workers from the current 265,000 if it is to regain profitability. The auto maker has lost \$17 billion in North America in the last three years."*

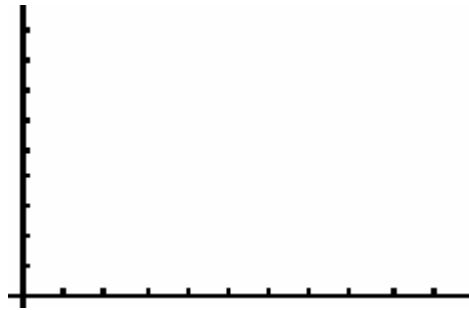
This article reflects a nation-wide trend toward a decrease in the number of hourly workers in the U. S. Consider the following.

Year End	U. S. Hourly Workers
1978	520,000
1979	510,000
1980	514,000
1981	476,000
1982	446,000
1983	426,000
1984	424,000
1985	435,000
1986	421,000
1987	399,000
1988	373,000
1989	345,000
1990	329,000
1991	304,000
1992	288,000

**Example:** Enter the data in List 1 and List 2 on your calculator. Index the numbers so that 1983 = "83" and 465,000 = "465".

Plot the data on your calculator. Then sketch it at the right, Be sure to include labels and scales.

There are two requirements in finding a model which best fits given data: (1) to find a correlation coefficient ("r-value") closest to 1, and, (2) to look for randomly scattered, relatively small residual values. First, use your calculator and find the equation and correlation coefficient for a linear regression (round to each value to three decimal places):



Equation of Line \_\_\_\_\_

r \_\_\_\_\_

Linear Regression \_\_\_\_\_

Plot this line on your calculator and on the graph on this page.

There is a second requirement of testing the fit of the line involves examining the residuals. A residual is a measure of how far each point is from the line you found. In other words, residual = actual y value - predicted y value from the line you found. So, the residual for 1981 is  $476 - 479.5$  (the value you get when you substitute 81 into Y1) = -3.5.

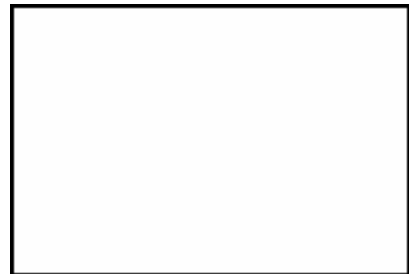
Clear List 3.

Arrow up so that L3 is highlighted.

Enter  $L2 - Y1(L1)$  by pressing

**<2nd>**      **<2>**      **<->**      **< VARS>**      **<Y-VARS>**  
**<1:Function>**      **<1:Y1>**      **<->**      **<2nd>**  
**<1>**      **<->**      **<Enter>**

Plot the residuals (L1 and L3) and examine them. If they are randomly scattered and relatively small, that indicates a good fit. Reproduce the graph at the right. Indicate scale.



This confirms that the linear regression line is a good fit. First, look at the size of the residuals. Residuals that are small relative to observed y-values provide evidence of good fit. For example, with y values ranging from 288 to 520 residuals from 0 - 40 are relatively small as opposed to residuals as large as 100. Obviously, many large residuals would cause us to question how good the model is, because the predictions based on our models would not be accurate. If there are many large residuals, we may need to find a different function to model the data. If there are only one or two large residuals, we may note them as outliers and determine why they exist. If the outliers result from errors in measurement, the data points should be corrected or excluded from the analysis. If all the points are correct, a large residual may provide interesting and useful information. If it can be explained we can often gain other useful information.

We must also see if the residuals follow any trend or pattern. Are the residuals in the middle all positive or all negative? Do the residuals at one end of the graph tend to be larger than at the other end? If there is a pattern, it might indicate an incorrect model. Notice that in our example, the residuals are both positive and negative and are scattered throughout the graph and that the residuals at each end are of similar magnitude, so there appears to be no problem. Our model is a good one.

The scatter plot of  $x_i$  vs.  $r_i$  (in our example, L1 vs. L3) is called a **residual plot**. If the model is a good fit for the original data, the residual plot should show points scattered randomly within a horizontal band about the

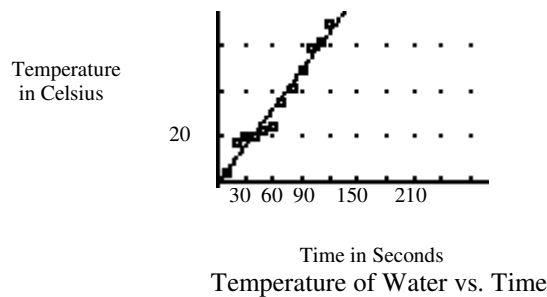
horizontal axis. The plot above shows that the residuals are reasonably small and do not follow a trend or pattern; therefore, our linear model is a good one.

The method you just used to find residuals requires that you write a formula to find the residuals, i.e., the differences between actual values and predicted values. The TI-83 finds residuals automatically. Recall the manatee data (p 6). Calculate the Linear Regression equation. Now go to an empty list. Press **<2nd>** **<INS>** **<2ND>** **<STAT>**. Note that there is a list called **RESID** in the named lists. Load this into your list. You now have the residuals for the manatee data based on the linear regression line. Find the residual plot and reproduce it at the right. Be sure to indicate the scale



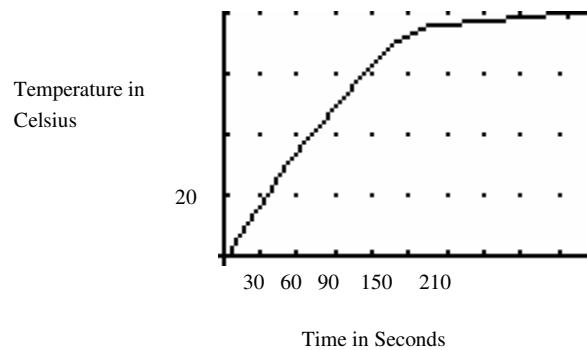
Based on this information, is the linear regression line a good predictor of the manatee data? Explain. If you were asked to use the line of best fit for the manatee data to predict the number of manatees killed in 2000, what problems could that lead to?

When asked to use the best fit line to predict manatee deaths in 2000, you are required to use the equation because the requested information was outside of the viewing window. When and how far can curves be extended to help in making predictions? This is tricky. In extending the curve beyond the available data we must be sure we take into consideration all the information we have available. We must realize that we are making predictions and they are at best only fallible estimates. Consider the following example. Suppose you put a tea kettle of ice water on your stove and turn on the heat. If you measure the temperature every 10 seconds, you get the following graph.



- a. What will the temperature be after 65 seconds?
- b. What will the temperature be after 3 minutes?
- c. What will the temperature be after 4 minutes?
- d. What will the temperature be after 5 minutes? Is this possible?

The temperature never goes over 100 degrees unless the water is under pressure. In an ordinary tea kettle, it would be impossible to get 105 degrees. So the trend line in the previous graph is wrong. The points do not lie on a straight line but near a curve that looks like the adjacent figure .



Consider the following example where a linear model is not a good fit. However, the residual plot yields information that leads to a best-fit curve model.

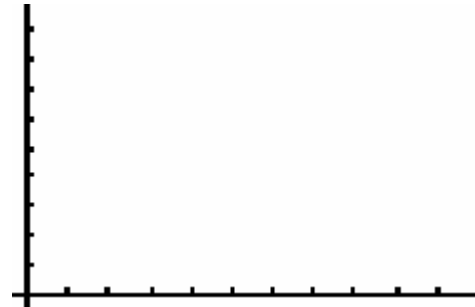
Free-Fall: An object is dropped from a high platform, and measurements are made of the distance it has fallen after certain times. The table below summarizes these measurements. Enter the following data into L1 and L2.

Time (sec)	Distance (cm.)
0.16	12.13
0.24	29.81
0.25	32.65
0.30	44.17
0.30	42.82
0.32	55.76
0.36	63.51
0.36	65.06
0.50	124.6
0.50	129.7
0.57	150.2
0.61	189.4
0.61	182.2
0.68	220.4
0.72	261.0
0.72	254.0
0.83	334.6
0.88	375.5
0.89	399.0

Plot L1 vs. L2. Then sketch it below. Include labels and scales.

Find the linear regression line for this data:

Plot the line in your window on the calculator and on the plot above.



Now, find the residuals and the residual plot. Then sketch it to the left, noting the scale:

The scatter plot of these points shows a decidedly non-linear relationship. What is the relationship between these data values?

The calculator has found the line that fits these data best, but it is not a very good fit. Is it possible to re-express the data so that there is a linear relationship? Can we perform an operation on the y-values that will straighten out the scatter plot? This operation must have the effect of decreasing the rate of growth of the y-values; it must reduce larger values more than smaller ones, or else we will just shift the curve down without straightening it. If we take the square root of the distance, this may provide a better fit. Clear L3 of the residuals and enter the square roots of the data in L2. Do this by highlighting L3 and entering  $\sqrt{L2}$ . Now plot L1 vs. L3. Fit a line to this data.

What is the equation? \_\_\_\_\_ What is r? \_\_\_\_\_

Now, find the residuals and residual plot of the re-expressed data. Then sketch it at the right, noting the scale:



Describe the residual plot. Is there a strong linear relationship between time and the square root of distance?

There is, therefore, a linear relationship between time and the square root of distance. By squaring both sides of the equation of the best fit line, thereby undoing the re-expression, we can find the relationship that summarizes the original data.

$$\begin{aligned}\sqrt{y} &= ax + b \\ (\sqrt{y})^2 &= (ax + b)^2 \\ y &= a^2x^2 + 2abx + b^2\end{aligned}$$

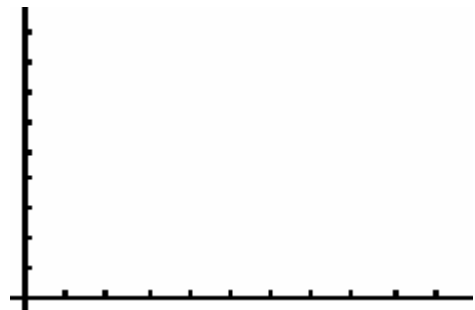
Find the specific equation for this data.

Plot this equation over the original data. This equation fits the data very nicely.

**Exercises:** Use the techniques in the previous two sections to analyze the following data.

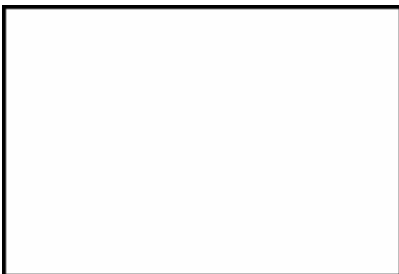
1. Many TJ graduates spend more than 4 years at college, often going on immediately for a masters or doctorate degree. One such student studied her tuition bills over an eight year period and recorded the following data:

Year	One Semester's Tuition
1	4680
2	4960
3	5510
4	6060
5	6845
6	7600
7	8435
8	9250



a. Construct the scatter plot for the data. Indicate the scale and labels.

b. What is the linear regression equation?



c. Now, find the residuals and the residual plot. Then, sketch it to the left, noting the scale.

d. What relationship does the residual plot indicate?

e. Re-express the data by finding  $\sqrt{\text{Tuition}}$ . What is the linear regression equation for  $\sqrt{\text{Tuition}}$  in terms of years?

f. What is the correlation coefficient for this line?

g. Find the residual plot for this re-expressed data. Then sketch it to the right, noting the scale.

h. Now, find an equation for Tuition in terms of years, based on your work in part e.



i. If this student remains in school 2 more years, what should she anticipate that her tuition will be?

2. Using the free-fall data from the class example, re-express the data in each of the following ways.

a. Square the values of time and examine a scatter plot of the ordered pairs  $(t^2, d)$ . Find the equation of the linear regression line through these points.

b. What is the correlation coefficient?

c. Examine the residual plot of the re-expressed data. Based on the collected information, is this a better or worse re-expression of the data? Why?

d. Now, divide the distance values by the corresponding time values to create ordered pairs  $(t, d/t)$ . Find the equation of the linear regression line through these points.

e. What is the correlation coefficient?

f. Examine the residual plot of the re-expressed data. Based on the collected information, is this a better or worse re-expression of the data? Why?

(taken from The Mathematics of Change: An Institute for Teachers of High School Mathematics.)

## Discussion #6: Measures of Variability

Data points may cluster about the mean or be spread out. A measure of variability is a single number that represents the spread or amount of dispersion in a set of data. The three most common measures of variability are range, variance and standard deviation.

The range of a set of numbers is the difference between the largest and smallest numbers in the set. For example, the range of the set of data 2, 3, 3, 5, 5, 5, 8, 10, 12 is 10.

A deviation is the distance a single measurement of a set is from the mean of the set ( $x - \mu$ ). It can be positive, negative, or zero. For example, for the set  $P = \{1,3,4,6,7,9\}$ , the mean is 5. The deviation for each data point is as follows:

x (data point)	$x - \mu$	deviation
1	1-5	-4
3	3-5	-2
4	4-5	-1
6	6-5	1
7	7-5	2
9	9-5	4

We use the squared deviations in calculation of the variance. Variance is the sum of the squared distances of the values from their mean, divided by the number of values in the population; the symbol for population variance is  $s^2$ .

For the six entries above 
$$\sigma^2 = \frac{(-4)^2 + (-2)^2 + (-1)^2 + (1)^2 + (2)^2 + (4)^2}{6} = \frac{42}{6} = 7$$

Expressed as a formula, 
$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

The standard deviation of a population is the square root of its variance so the symbol used is  $\sigma$ . The formula for standard deviation is  $\sigma = \sqrt{\sigma^2}$

For the data above, the standard deviation is approximately 2.65.

In today's world, calculators and computers make the computation of variance and standard deviation simple. Using the population  $\{0, 3, 4, 4, 6, 9, 12, 13, 15, 21, 23\}$  proceed through the following steps to familiarize yourself with the use of the calculator:

Enter the data as before and then press

**[STAT]**      **< ▸ >** (to CALC).      **<ENTER>**      **1-Var Stats L1**

This means that for this population of  $N = 11$  data points the standard deviation is approximately 7.2 and the variance is approximately 51.5. Variance was used in the past as an intermediate step in determining the standard deviation. It does not have the practical use today that it once had.

For these problems, we will be working with populations and use  $\sigma$ .  $S$  gives the sample standard deviation. For example, if you wished to find the average GPA of all of the students in the school it would be very time consuming, so you might take a sample of 100 students and estimate the GPA and resulting standard deviation based on these students. There will be an error due to the fact that you are only working with a portion of the population.  $S$  has a correction factor built in to try to minimize this error.

**Exercises:**

1. The following grades were obtained on a statistics test:

86	73	97	60	78	83	68	92	70	63	72
65	90	85	76	72	65	69	62	74	70	

- a. Find the mean.
- b. Find the range.
- c. Find the standard deviation.
- d. Find the variance.

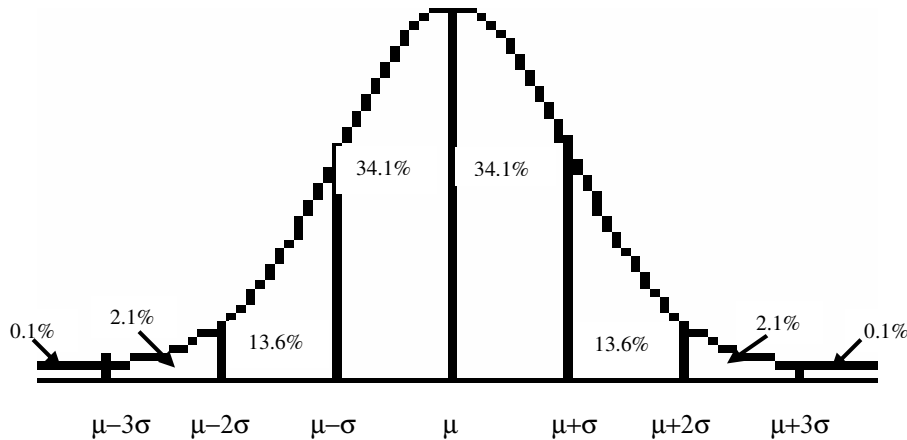
2. The following 45 numbers are given as the high temperatures on a given day:

57	54	37	59	50	51	52	53	57	52	43	47
47	46	63	40	46	40	56	44	88	53	48	46
70	51	53	84	44	45	61	53	55	44	52	76
46	61	62	48	54	58	51	53	56			

- a. Find the mean.
- b. Find the range.
- c. Find the standard deviation.
- d. Find the variance.
- e. Find  $\mu + \sigma$ ,  $\mu - \sigma$ .

*Discussion 7: The Normal Distribution*

The empirical rule states: If the population of measurements is symmetrical and bell-shaped, then approximately 68% of all the measurements in the set fall within the interval from  $\mu - \sigma$  to  $\mu + \sigma$ ; approximately 95% of all the measurements fall within the interval from  $\mu - 2\sigma$  to  $\mu + 2\sigma$ ; and, approximately 100% of all measurements fall within the interval from  $\mu - 3\sigma$  to  $\mu + 3\sigma$ . Study the diagram below:



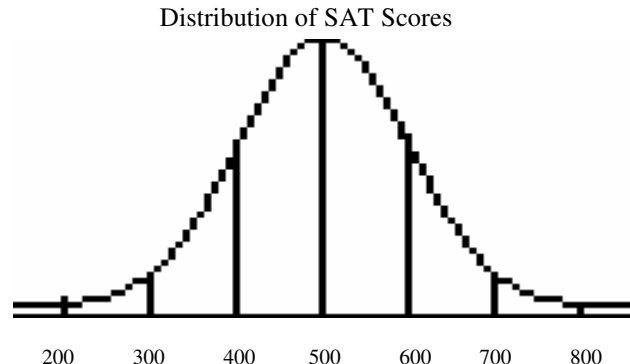
Notice that outliers occur beyond three standard deviations from the mean and represent less than 0.2% of the population.

**Exercises:** Consider the normal distribution while answering the following problems:

1. The SAT's are designed so that the distribution of scores would appear as shown:

- a. What is the mean score?
- b. What is the standard deviation of the scores?
- c. What percentage of scores were between

- (1) 500 and 600
- (2) 500 and 700
- (3) 500 and 800



- d. Some colleges do not admit applicants whose scores are less than 600. According to this distribution, what percentage of students would be expected to have scores of 600 or more?
- e. The most competitive colleges require scores over 700. What percentage of students would be considered by these colleges?

For each of the following problems apply the empirical rule to answer the questions. Draw a curve for each problem.

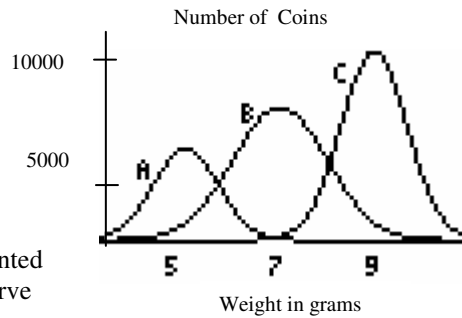
2. A catsup company has fixed the weight of a bottle at 16 oz., with a standard deviation of 0.5 oz. The curve depicting the weights is bell-shaped. Approximately what percentage of boxes will be

- a. greater than 15 oz.?
- b. greater than 17 oz.?
- c. less than 14 oz.?
- d. less than 13 oz.?
- e. between 15 and 17 oz.?

3. Assume that an insurance company's financial losses due to automobile accidents involving card three years old and older yield a symmetrical bell-shaped curve, with a mean of \$600 and a standard deviation of \$200. What proportion of such accidents will result in losses

- a. exceeding \$1000?
- b. exceeding \$800?
- c. Why would automobile insurance companies want to keep records of such information?

4. The curves in the figure at the right show the variations in the weights of quarters that were minted and put into circulation at the same time.\* One curve shows the weight distribution when the coins were new and the other two show the distributions when they had been in circulation for five years and for ten years.



- Which curve do you think shows the weights of the newly minted quarters, which curve the coins after five years, and which curve after ten years?
- What happens to the average weight of the coins as time passes?
- What happens to the standard deviation of the weight of the coins as time passes?

\*Adapted from a graph in the article "Scientific Numismatics" by D. D. Kosambi, *Scientific American*, February 1966.

### Discussion 8: Standard Scores

The standard score or z score is the number of standard deviations that a given value is above or below the mean, and it is found using the following formula:  $z = \frac{x - \mu}{\sigma}$

For example, which is better: a score of 65 on Test A or a score of 29 on Test B? The class statistics for the two tests are as follows:

<u>Test A</u>	<u>Test B</u>
$\mu = 50$	$\mu = 20$
$\sigma = 10$	$\sigma = 5$

For the score of 65 on Test A we get a z score 1.5, but for the score of 29 on Test B we get a z score of 1.8. That is, a score of 65 on Test A is 1.5 standard deviations above the mean, while a score of 29 on Test B is 1.8 standard deviations above the mean. This implies that the 29 on Test B is the better score. While 29 is below 65, it has a better *relative* position when considered in the context of the other test results.

### Exercises:

1. Transfer students to a new high school are sometimes given a standardized test with a mean of 100 and a standard deviation of 20. To two decimal places, convert the raw scores of the following students to z scores:

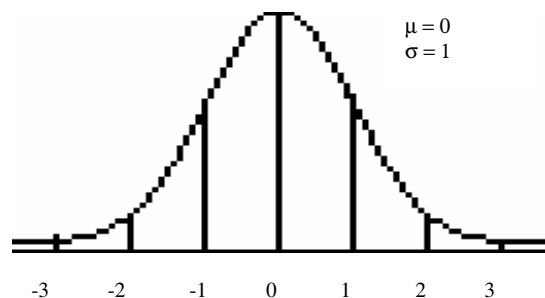
Alice--105      Bob--72      Carol--142      David--133      Elliott--95

2. John weighs 220 pounds; his dog Fido weighs 90 pounds. If human males weight an average of 160 pounds with a standard deviation of 20 pounds, and all dogs of Fido's breed have an average weight of 80 pounds with a standard deviation of 5 pounds, how do John and Fido compare, relative to their populations, with respect to weight?

## Discussion 9: The Standard Normal or Z Distribution

A normal distribution curve (called a z-distribution) is a symmetrical bell curve with the following characteristics:

1. The mean is 0.
2. The variance is 1.
3. The standard deviation is 1.
4. Approximately 68% of the data falls between  $z = \pm 1$ .
5. Approximately 95% of the data falls between  $z = \pm 2$ .
6. Approximately 99.7% of the data falls between  $z = \pm 3$ .
7. Total area under the curve is 1.0



We cannot talk about area under the curve at a specific point, but we can talk about area of a specific region. For example, the area left of the mean is 0.50. We can interpret that as  $z < 0$  has an area of 0.50. The probability that  $z < 0$  [written  $P(Z < 0)$ ] = 0.50.

**Examples:** From what you know about normal distribution, estimate:

1.  $P(Z < -1)$
2.  $P(Z < 2)$
3.  $P(Z > 1)$
4.  $P(-1 < Z < 2)$

You can use your calculator to find the probability for the proportion of data that falls within an interval that is not defined by a standard 1, 2, or 3 standard deviations. Press **<2nd> <VAR> <2>normalcdf**. There are four parameters for normalcdf(lower bound, upper bound, mean, standard deviation). So, in order to find  $P(Z < -2.13)$ , enter **<2nd> <VAR> <2>normalcdf(-1EE10, -2.13, 0, 1) (ENTER)**. You should get approximately 0.017. This means there is a 1.7% probability that the z-score is less than -2.13.

Use your calculator to find each of the following:

5.  $P(Z < -0.45)$
6.  $P(Z > 1.62)$
7.  $P(-1.40 < Z < -0.40)$
8.  $P(-1.50 < Z < 2.50)$

To reverse the process and find z, do the following **<2nd> <VAR> <3>invNorm**

There are three parameters for invNorm(area, mean, standard deviation). So, in order to find  $P(Z < z) < 0.217$ , enter **<2nd> <VAR> <3>invNorm(0.217, 0, 1) (ENTER)**. You should get approximately -0.782. This means there is a 21.7% probability that the z-score is less than -0.782.

Use your calculator to find each of the following:

9.  $P(Z < z) = 0.1469$ , let  $z = -1.05$ .
10.  $P(Z > z) = 0.0485$ . let  $z = 1.66$ .

**Exercises:** Find  $z$  if

1.  $P(Z < z) = 0.0668$

2.  $P(Z > z) = 0.9861$

We may find the probability of attaining specific data when we have a normal distribution and the mean ( $\mu$ ) and standard deviations ( $\sigma$ ) are not 0 and 1, respectively. We adjust our data, using the formula for  $z$  score from the previous section.

3.  $P(X < 3.5)$  when  $\mu = 5$ ,  $\sigma = 1$ .

4.  $P(X > 130)$  when  $\mu = 110$ ,  $\sigma = 25$

5.  $P(14.2 < x < 15.0)$  when  $\mu = 14$ ,  $\sigma = 0.5$ .

6. One part of a test administered to adults is an exercise in manual dexterity. The average time for the test is 165 seconds, with a standard deviation of 21 seconds. Assume the relative frequency distribution of the times needed to complete the test is approximately normal. What proportion of the adult population can complete the test in

a. more than 190 seconds

b. between 140 and 160 seconds

We may reverse the previous process, i.e., probabilities may be used to find corresponding values of data points.

7. If the mean is 10 and the standard deviation is 2, for what values of  $y$  is it true that  $P(Y < y) = 0.025$ ?

8. If the mean is 20.5 and the standard deviation is 0.2, find  $x$  such that  $P(X > x) = 0.7995$ .

9. If  $X$  is a continuous random variable that can be modeled as normal with a mean of 100 and a standard deviation of 10, find  $x$  so that

a.  $P(X > x) = 0.0049$

b.  $P(X < x) = 0.9332$