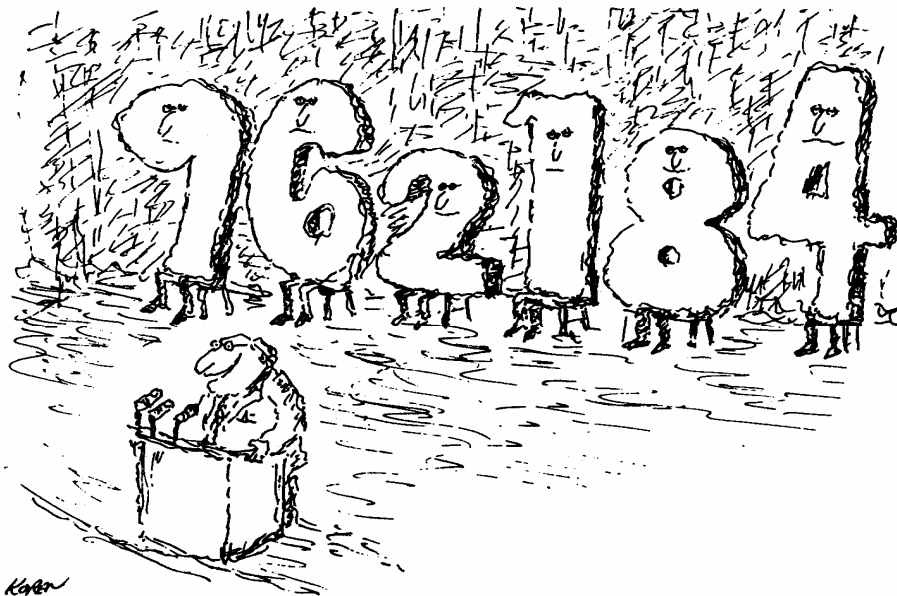


Data Analysis 1



Karen

"Tonight, we're going to let the statistics speak for themselves."

TJHSST

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.

Discussion 1: Measures of Central Tendency

Measures of central tendency are numbers that are most representative of the data contained in the set. There are three measures of central tendency: mean, median, and mode. The **mean** is the arithmetic average of a set of data. The **median** is the middle number of the set of data once it has been ordered from least to greatest. If there are an odd number of pieces of data, the median is the middle number; if there are an even number of pieces of data, the median is the average of the two middle numbers. The **mode** is the data point that occurs most often. There may be more than one mode; there may be no modes.

There are advantages and disadvantages for each of these three measures of central tendency. It is important to choose the most representative measure, so you must understand the pros and cons of choosing each:

The Mode

Advantages

- Gives most frequently occurring measure
- Easy to find once data is arrayed

Disadvantages

- May not be central
- May not be unique
- May change significantly with addition of new scores
- May not exist

The Median

Advantages

- Gives middle score
- Not easily influenced by extremes

Disadvantages

- May not be part of data set
- Data must be arrayed to identify it

The Mean

Advantages

- Most people are familiar with it
- Easy to define algebraically
- Gives information about total of scores
- Used in other statistical calculations

Disadvantages

- Can be greatly influenced by extremes
- May not be part of data set

Whichever measure of central tendency you decide to use, it is important to look at the reason you are choosing it. If you are a buyer for a clothes department, finding the mean or median size dress that is sold will not be helpful. In this instance, using the mode (the dress size most often sold) will be most appropriate. If you must identify the upper and lower half of a group of qualifying scores for a race, the median is the most appropriate. The numbers are used as standards of comparison and simple indicators of a population. Choose them in appropriate ways.

The graphing calculator will be used for the following example.

Example: The following are the weights in pounds of children in a fourth grade class:

64	71	57	67	74	65	59	62	60	72	84	60	68
72	91	55	69	71	69	75	59	60	70	76	62	

We will use the graphing calculator to enter data (1-variable data) into the calculator (the instructions are for the TI-83), and then use the calculator as an aid in determining the mean, median and mode.

Process: (1) To enter 1-variable data, press

[STAT] <ENTER>

This chooses the <EDIT> option that allows for data entry. The screen should read

L1 L2 L3 (See Note at bottom of page.)

Type in the first piece of data, press <ENTER>. Enter the second piece of data and continue until all the data is entered. Then press [STAT] <▶>

The screen should now be at CALC.

(2) To obtain the mean, median and mode, press <ENTER>

This selects the **1-Var Stats** option. Press [2nd] [1] for L1 <ENTER>

You will obtain a display of data where \bar{x} = MEAN.

To obtain the MEDIAN, move the cursor down until you see **MED =**. The MEDIAN can also be found by arranging the data in ascending order. To do so, press [STAT]

Choose **SORT A**(

Press [2nd] [1] <ENTER>

The data is now arranged in ascending order in L1. To view the sorted data, press

[STAT] <ENTER>

Move the cursor to the middle term to read the median (in this case, the thirteenth term). If there are an even number of terms, the median is the average of the two middle terms.

Scan the ordered data and find which data point occurs most often. This is the MODE.

Note: When data needs to be cleared to make room for new entries, there are several options,

(1) Press [STAT] [4] [2nd] [1] <ENTER>

The screen displays **DONE** when the data has been cleared and you are ready to begin again.

or

(2) Press [STAT] <ENTER> <▲> [CLEAR] <ENTER>

Using the previous results from above, answer the following questions:

1. What is the mean weight of these fourth graders?_____ What is the median weight?_____ What is the mode?_____
2. Which do you think is the most representative number for these weights, the mean, median or mode? Explain.

Exercises:

1. If you wanted to find the total amount spend on junk food for a week by your class, would you want to know the mean, median or mode amount spent by the class? Explain.
2. If you wanted to know if you read more or fewer books per month than most people in the class, would you want to know the mean, median or mode? Explain.
3. The Reston Town Center skating rink is ordering new skates. Which would be more useful to know, the mode, mean or median skate size? Explain.
4. You want to know which Virginia county has a large portion of people with low incomes. Which is most helpful to know for each county: the mean, mode or median income? Explain.
5. A manufacturing company boasts that they pay an average salary of \$30,000 to their employees. Study the chart below and answer the following questions:

Type of Job	Salary	Number Employed
President	\$183,000	1
Vice-President	\$90,000	2
Plant-Manager	\$50,000	3
Foreman	\$30,000	12
Skilled Operator	\$22,000	21
Unskilled Operator	\$18,000	36

- a. Is the company telling the truth? To help you decide, find each of the following:

mean salary_____ median salary_____ mode salary_____

Note: In this problem, all but the first salary must be entered into the list a multiple number of times, i.e., 90,000 must be entered twice, 50,000 three times, 30,000 twelve times, etc. We can use **L2** in order to enter the frequency. Press **[STAT]** **<ENTER>**

In **L1**, enter the salaries. In **L2**, enter the number employed.

Then press **[STAT]** **< ► >** (to CALC) **<ENTER>**

1-Var Statistics is now displayed. Type **[2nd]** **1** , **[2nd]** **2** for L1, L2. Press **<ENTER>**

- b. Which do you think is a more representative number for these salaries, the mean, median or mode? Explain.

(Taken from Statistics and Information Organization: Math Resource Program by University of Oregon)

6. The following is a set of final grades on a statistics semester exam:

72	78	63	87	97	90	97	85	64	73	73	79
79	60	83	70	83	71	78	79	76	77	80	19
93	54	72	58	74	82	79	56	60	77	81	71
58	62	95	54	80	98	63	82	81	84	88	48
93	69	78	90	70	92	85	91	53	92		

a. Find the mean, median and mode.

mean_____ median_____ mode_____

b. If the low score of 19 was dropped, what happens to the mean?

c. Which do you think is a more representative number for these test scores, the mean, median or mode? Explain.

Statistics Discussion #2: Stem-Leaf Plots and Box Plots

There are two major ways of organizing data. One is the ordered array and the second is the stem and leaf diagram. An **ordered array** is a set of data arranged in ascending order. Your calculator can perform this task. A **stem and leaf diagram** is a set of numbers arranged so that a numerical classification heads each row and the unit values are listed to the right of it. It is used to rank order data and provide an indication of the shape of the distribution. The procedure for drawing stem-leaf plots is as follows:

- (1) Identify the stems (leading digit(s)).
- (2) Place leaves with corresponding stems.
- (3) Order stem-leaf plot.

Consider the following set of data:

33 30 24 22 33 25 30 37 21 29 42

Initial (unordered) plot:

Stem	Leaves
2	4 2 5 1 9
3	3 0 3 0 7
4	2

Final (ordered) stem-leaf plot:

Stem	Leaves
2	1 2 4 5 9
3	0 0 3 3 7
4	2

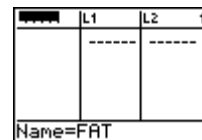
Exercises: The following data is taken from Fast Food Facts (1994):

Item	Calories	Fat(g)	Carbohydrates(g)	Sodium(mg)
Hamburgers				
Burger King Whopper	570	31	46	870
McDonald's Big Mac	500	26	42	890
Wendy's Single	440	23	36	850
Roast Beef Sandwiches				
Subway 6" Roast Beef	345	12	42	1140
Hardee's Roast Beef	380	18	29	1230
Arby's Roast Beef	383	18	35	936
Fish Sandwiches				
Hardee's Fisherman's Filet	480	21	50	1210
McDonald's Filet-O-Fish	370	18	38	730
Burger King Ocean Catch	450	28	33	760
Chicken Nuggets (6)				
Kentucky Fried Chik'n	284	18	15	865
McDonald's	270	15	17	580
Wendy's	280	20	12	600
Breakfast Sandwiches				
Hardee's Rise ' N Shine	320	18	34	740
Egg McMuffin	280	11	28	710
Burger King Bacon Croissant	353	23	19	780

1. Make a stem-leaf plot of the fat content.
2. Use this plot to help find the mean, median and mode. Confirm your answers by using your calculator. Because this data will be used again in a later exercise, we will store it in a named list that can be recalled at a later time. Press

[STAT] <ENTER>

Display the **Name=** prompt in the entry line by moving the cursor onto the list name in the column where you want to insert a list, and then press <2nd> <INS>.



Your screen should appear like the diagram at the right.

Type in a list name of up to five letters. In this case, calling the list FAT would be appropriate.

Press <F> <A> <T> <ENTER>

Now enter the data as before and answer the question posed.

3. Make a stem-leaf plot of the carbohydrate content. Store this data in a named list.
4. Use this plot to help find the mean, median and mode.

A **box plot** is a means for illustrating measures of central tendency and the range of data in an easy-to-read format. The procedure for drawing box plots is as follows:

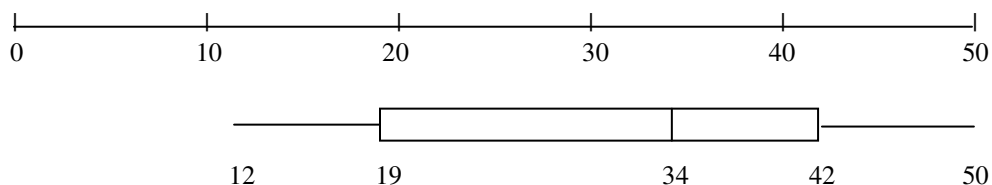
1. Put the numbers in an ordered array (by use of a stem-leaf plot or with the calculator).
2. Find the median and draw a line through it.
3. Find the median to the left of the line--this is the lower quartile.
4. Find the median to the right of the line--this is the upper quartile.
5. Find the lower extreme--the smallest data value.
6. Find the upper extreme--the largest data value.
7. Draw a reference line with a scale.
8. Below the reference line put a dot where the upper and lower extremes occur.
9. Put a small vertical segment where the upper and lower quartiles and median occur.
10. Draw a horizontal segment from the lower extreme to the lower quartile. Draw a box between the two quartiles. Draw a horizontal segment from the upper quartile to the upper extreme.

Example: Consider the stem-leaf plot for the carbohydrate content of fast foods in the previous exercise:

```

1      | 2 5 7 9
2      | 8 9
3      | 3 4 5 6 8
4      | 2 2 6
5      | 10
    
```

The median is 34, the lower quartile is 19; the upper quartile is 42. The lower extreme is 12; the upper extreme is 50. The following box plot illustrates this data:



Boxplots can also be found on the TI-83. Enter the data as described before. Then press

[2nd] [y=] <ENTER>

Move the cursor over **[On]** and press **<ENTER>**.

Move the cursor down to the **Type** line and over to the boxplot icon as shown at the right and press **<ENTER>** to choose the boxplot.



Move the cursor down to **Xlist**. Press **<2nd> <STAT>**, highlight the list you want and press **<ENTER>**.

Press **[ZOOM] <9>**. The calculator automatically chooses an appropriate scale.

Press **[TRACE]** and the cursor moves back and forth between the lower extreme (minX), lower quartile (Q1), median (Med), upper quartile (Q3), and upper extreme (maxX).

The information found above can be used to determine if there are any outliers in the data. An outlier is a data point that is very different from the other points and can consequently cause the mean of the data to be overly influenced in one direction. Recall the set of statistics test grades. The low score of 19 is an outlier. There is a formula for determining an outlier:

Steps for determining an outlier

1. Find the interquartile range (the difference between the upper and lower quartiles).
2. Multiply this by 1.5.
3. Add this number to the upper quartile; subtract this number from the lower quartile. If a data point falls above or below the resulting values, it is an outlier. Consider the carbohydrate content information. The interquartile range is $(42 - 19) = 23$. $1.5(23) = 34.5$. $42 + 34.5 = 76.5$. There are no data points above 76.5 so there are no upper outliers. $16 - 34.5 = -18.5$. There are no data points below -18.5, so there are no lower outliers.

These are standard box plots showing the two extremes, upper and lower quartiles and median.

Exercises:

1. Use the data from the previous set of exercises.
 - a. Draw a box plot for the fat content of fast foods. In order to recall this list, go into your lists, move the cursor to highlight the title of the list where you will recall the FAT list, and then press **<2nd> <STAT>**, highlight the list you want and press **<ENTER>**.
 - b. Are there any outliers for this data? Justify your answer using the method above.

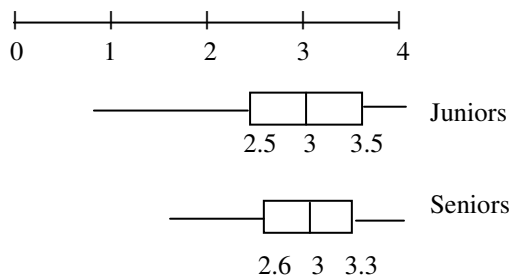
There is also a modified box plot that better indicates the outliers. Rather than drawing the whiskers to the extremes, the whiskers are drawn to the points where outliers begin to occur, if there are any. For this problem, the whiskers will still extend to 11 since that is the minimum value and there are no left hand outliers. But the right whisker would only extend to 30.5. The last value, the outlier, 31, will be plotted.

The TI-83 draws this modified box plot.

Choose the first type of box plot in the Stat Plots as shown in the first window and then replot the data. Reproduce this modified box plot in your work.



2. Refer to the plots below.



- a. Which class had the higher median?
- b. What was the interquartile range for each team?
- c. Estimate each of the classes' best and worst grade point averages. Are there any outliers? Explain.

Discussion #3: Qualitative Data and Frequency Tables

There are two types of data that can be generated in an experiment or survey. The first is qualitative data. **Qualitative** data are non-numeric measures or characteristics such as hair color, sex, political affiliation, religion, brand name, etc. The second is quantitative data. **Quantitative** data are numeric data such as height, weight, test scores, etc.

When collecting data it is important to tally the data in a convenient manner. With qualitative data, the classes occur naturally. For example, if you are studying hair color, the classes naturally become blonde, brunette, redhead. With quantitative data, the process of determining the class divisions is somewhat arbitrary and can vary from person to person.

The tally or count of the number of times a particular measure occurs is called the **frequency** (f). The **relative frequency** refers to the proportion of all given values that fall within an interval, i.e., the frequency divided by the total number of data points.

Frequently, statisticians organize data in a chart called a **frequency table**. The simplest frequency table must indicate classes, frequency and relative frequency.

Exercise: Eighteen college students were asked to rate the school food according to the following classifications:

- E--excellent
- VG--very good
- G--good
- F--fair
- P--poor

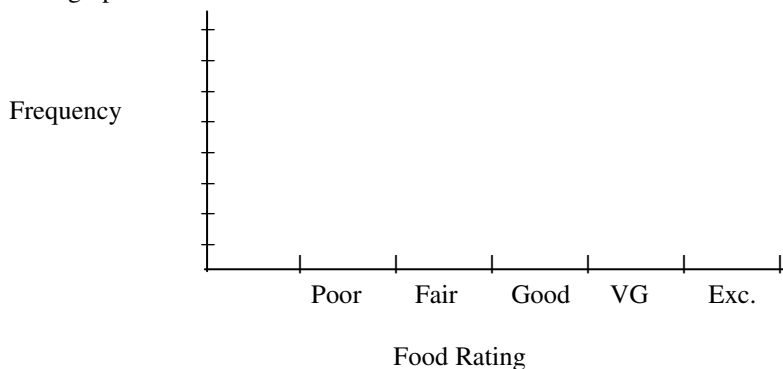
The results of the study are coded as follows:

G F VG E E P F G G G VG E
 F G VG E VG G

Set up a frequency table charting these results.

Interval	Frequency	Relative Frequency
Poor		
Fair		
Good		
Very Good		
Excellent		

It is simple to set up a bar graph illustrating this data. The intervals are preset. Be sure to label the axes appropriately when you draw the bar graph.



Discussion #4: Quantitative Data and Frequency Tables

When you work with continuous data, class intervals do not occur naturally. In order to determine a convenient interval, you should first find the range. The **range** is the difference between the smallest and largest values in the set. After examining the range and the number of pieces of data you have, you must determine how many intervals would be convenient. We generally follow Sturges' Rule to determine the number of intervals:

Number of Values in a Set	Appropriate Number of Intervals
10 to 100	4 to 8
100 to 1000	8 to 11
1000 to 10000	11 to 14

Follow the procedure outlined below in setting up the frequency table:

1. Establish the class limits--the smallest and largest values that would be placed in a given class. Decide on the lowest limit (make it convenient) and work from there. This limit is often included in the given class.
2. Tally your data.
3. Find the frequency--the number of data points in a class. The symbol f is often used to represent frequency.
4. Find the relative frequency--the fractional part of the data points in a class. If n data points are tallied, the relative frequency is f/n .

Exercises: For the following set of data, determine an appropriate number of classes and set class limits. Then, set up a frequency table (as shown below) to organize the data.

1. The following are the specific gravities of 25 samples of magnetite from a local region:

3.0	3.1	3.7	4.3	5.7
2.4	4.0	5.6	2.6	3.9
3.4	4.4	3.7	3.7	4.6
3.9	5.0	3.6	2.7	4.6
5.1	3.8	5.1	4.3	6.2

Interval	Class Limits	Frequency	Relative Frequency

(You may not need to use all six interval spaces).

2. The following are the weights in pounds of children in a fourth grade class. Determine an appropriate number of classes, set class limits, find the frequency and relative frequency. Then, set up a frequency table to organize the data.

64	71	57	67	74	65	59	62	60	72	84	60	68
72	91	55	69	71	69	75	59	60	70	76	62	

Interval	Class Limits	Frequency	Relative Frequency

Discussion #5: Histograms and Frequency Polygons

When plotting a histogram for continuous data, remember to label your axes appropriately and follow the procedure below when drawing the histogram on your TI-83 calculator (use the fourth grade weights data from the previous exercises to complete the following example):

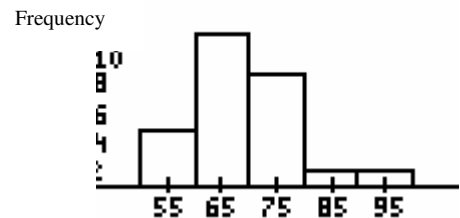
1. Clear any data in the calculator.
2. Press <WINDOW>.
3. Decide on the **Xmin** and **Xmax** that best suits the data (perhaps 50 and 100).
4. Decide on how wide you wish each bar to be. Your frequency table should aid in this decision. Enter this in **Xscl** (x-scale).
5. Enter 0 for **Ymin**. Decide on the **Ymax** and **Yscl** (perhaps 15 and 1, respectively).
6. Clear any previous graphs stored in your calculator ([Y=] [CLEAR])
7. Press [2nd] [Y=]. (This gives you **STAT PLOT**.) <ENTER>.
8. Turn on Plot 1. Press <▼> to Type and select histogram. Press <▼> to xlist and select the location of your data (L1, L2, or named list, etc.)
9. Press [GRAPH].

Generally you will be asked to reproduce this histogram on your own paper in order to get a frequency polygon. Be sure to

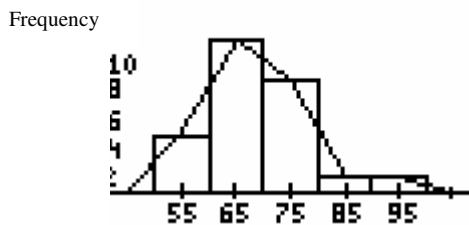
1. Label the axes, label the beginning and ending points of each bar, and label the scale on the y-axis.
2. Display the entire vertical axis (don't truncate).
3. Leave a space to the left and right of the histogram equal to the width of a bar. This is necessary in constructing a frequency polygon.

Example: Use the data set of fourth grade weights from the previous set of exercises (page 3) and use the procedure outlined above to find a histogram. Add labels and scales and compare your results to the histogram shown here.

A frequency polygon is easy to plot using your histogram. Find the midpoint of the top of each bar of your histogram and of the initial and terminal empty intervals. Connect these points. The frequency polygon is shown below.



Weights of 4th Graders in lb.

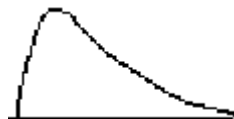


Weights of 4th Graders in lb.

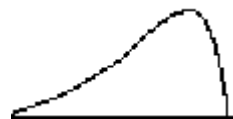
There are four commonly occurring shapes of smoothed frequency polygons:



Bell-Shaped



Right Skewed

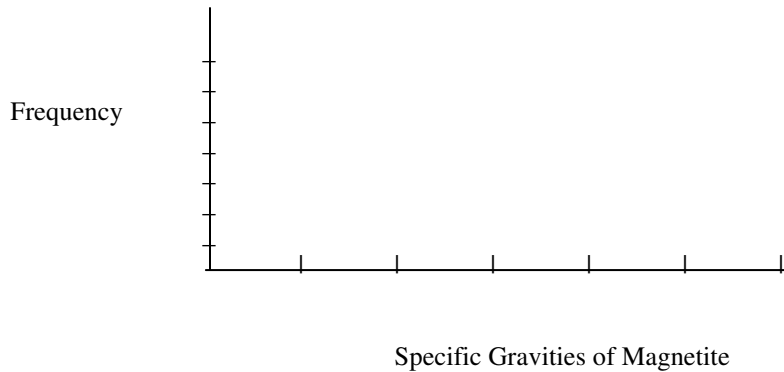


Left Skewed



Bimodal

Exercise: Using the data from the frequency table you constructed in Ex. 1 on p. 10, design a histogram and then a frequency polygon of the magnetite data.

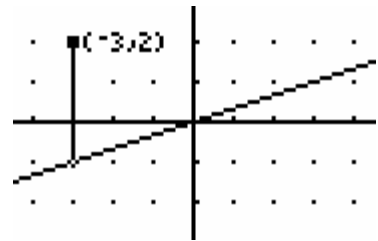


What shape does the frequency polygon appear to have? Explain.

Statistics Discussion #6: Residuals

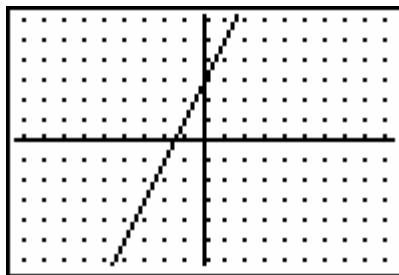
Residuals are a means of determining how far a point is from a line. Rather than finding the perpendicular (shortest) distance, it is often more convenient to find the vertical distance. This distance is the difference between the y-coordinates of two points and is called a residual. The vertical distance is used because of ease of calculation and because y is the dependent variable. In order to find the residual between a point and a line, draw a vertical segment from the point to the line so that the x-coordinates are the same. Then find the distance between the y-coordinates.

In the diagram at the right, the residual between the point $(-3, 2)$ and the pictured line is 3. Note that the residual is positive because the point lies above the line. If the point is positioned below the line, the residual is negative.



Exercises:

1. Given the line whose equation is $y = 2x + 3$ and the points $A = (0, 0)$, $B = (1, -5)$, $C = (2, 4)$, $D = (-1, -1)$, and $E = (-4, -1)$.
 - a. Plot the five points on the axes below and draw in the vertical lines representing the residuals.

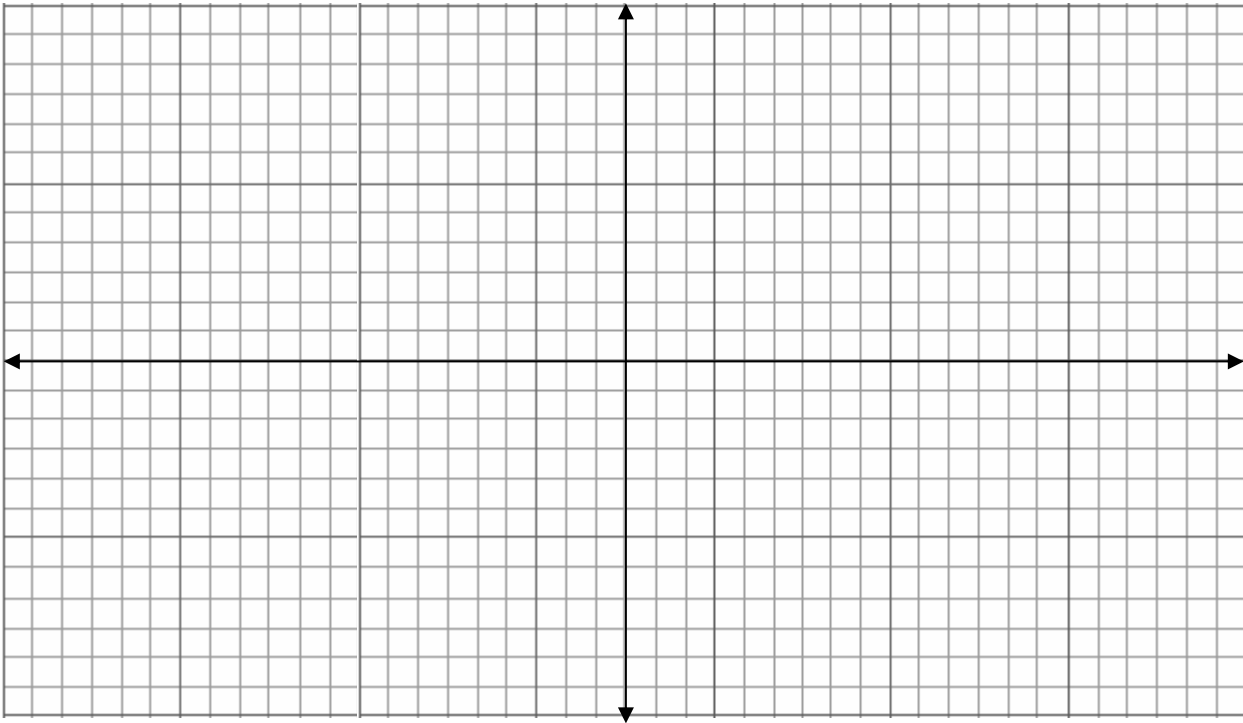


- b. Find the residuals for each of the five points:

A _____ B _____ C _____ D _____ E _____

- c. Which point is closest to the line? Which is farthest from the line? Why?

2. Plot a point P near the left margin of your graph paper.



a. Move two squares to the right and one square up and plot point Q. Use a ruler to draw a line through P and Q. Plot a point R that is 24 squares to the right of P and 12 squares up. If you extend your line, it should go through R. Why?

b. Let Q' and R' be the points on the line you draw that are 2 squares and 24 squares, respectively, to the right of P. Assuming that the line you draw actually goes through R, use the vertical separation between R and R' to calculate the vertical distance between Q and Q'.

3. Given $P = (1.35, 4.26)$, $Q = (5.81, 5.76)$, and $R = (19.93, 9.71)$.

a. Verify that P, Q and R are not collinear.

b. Given that $R' = (19.93, y)$ is on the line through P and Q, find y. Calculate the residual value $9.71 - y$.

c. Given that $Q' = (5.81, y)$ is on the line through P and R, find y. Calculate $5.76 - y$.

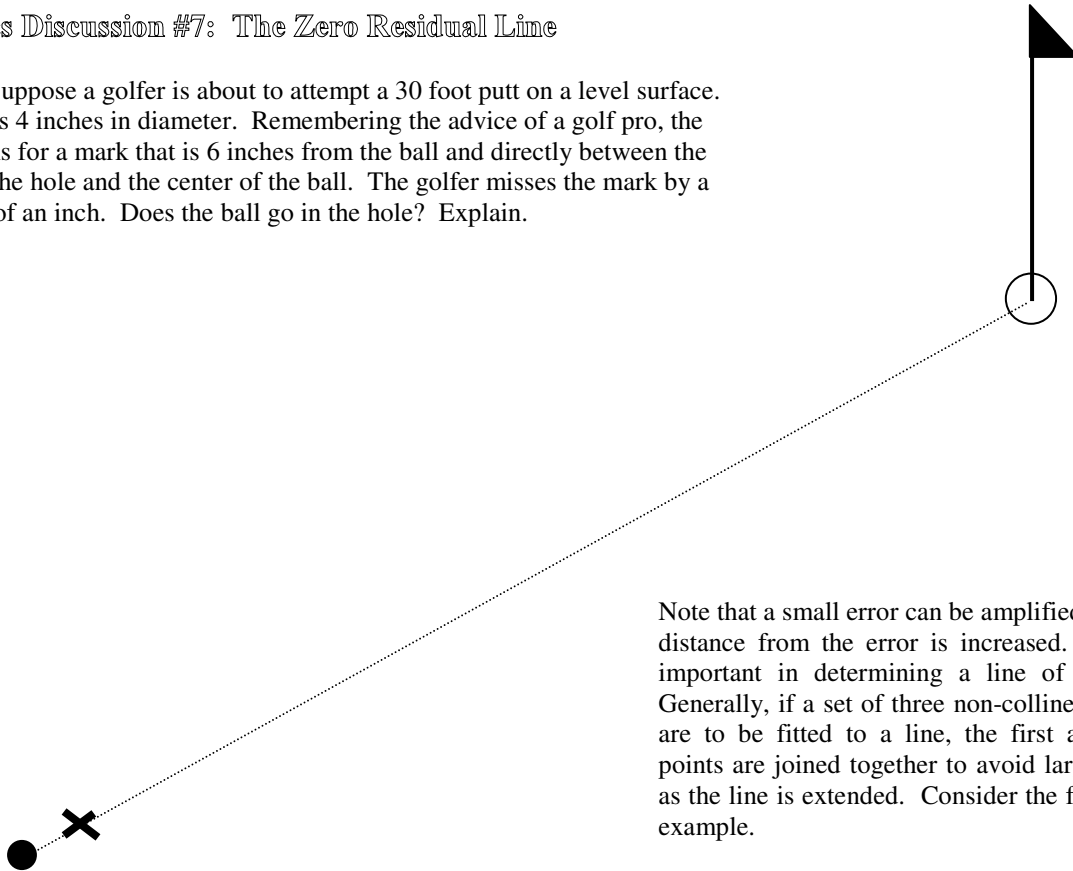
d. Given that $P' = (1.35, y)$ is on the line through Q and R, find y. Calculate $4.26 - y$.

e. Which of the three lines best fits the given data? Why do you think so?

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

Statistics Discussion #7: The Zero Residual Line

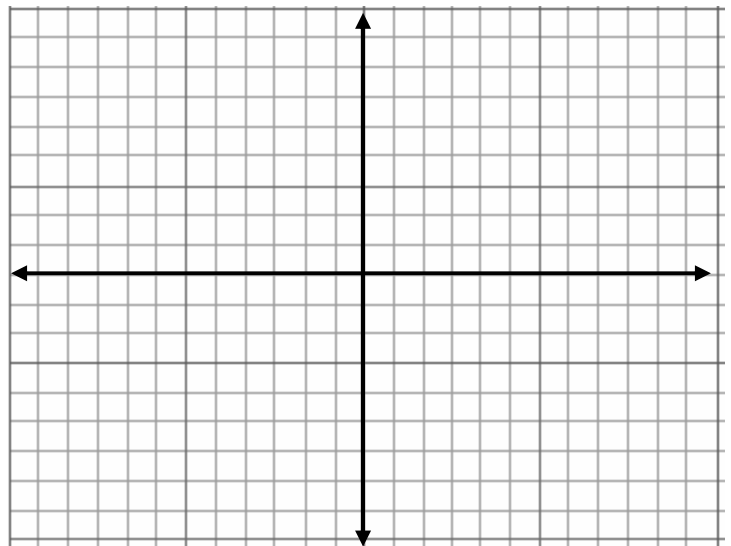
Suppose a golfer is about to attempt a 30 foot putt on a level surface. The hole is 4 inches in diameter. Remembering the advice of a golf pro, the golfer aims for a mark that is 6 inches from the ball and directly between the center of the hole and the center of the ball. The golfer misses the mark by a sixteenth of an inch. Does the ball go in the hole? Explain.



Note that a small error can be amplified as the distance from the error is increased. This is important in determining a line of best fit. Generally, if a set of three non-collinear points are to be fitted to a line, the first and third points are joined together to avoid large errors as the line is extended. Consider the following example.

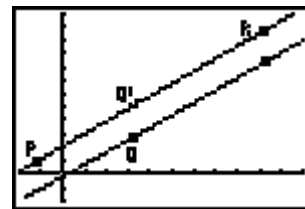
Example: Given $P = (-1.15, 0.97)$, $Q = (3.22, 2.75)$, and $R = (9.21, 10.68)$. Plot the points below.

- Verify that P , Q and R are not collinear. Explain your process.
- Find the equation of the line PR .
- Point Q is below this line. Given that $Q' = (3.22, y)$ is on the line through P and R , find y , and calculate the residual value $2.75 - y$.
- One might regard PR as the line that best fits the given data. The point Q has played no part in this decision, however! It therefore seems necessary to investigate other possibilities for a *line of best fit*. Find an equation for the line that is parallel to PR and that makes the sum of the three residuals zero.



If this line is below PR, the residual values of P and R will increase to positive values, while the residual value of Q will diminish in size (becoming less negative). The sum of the three residuals will therefore diminish in size, until finally it is zero. This zero residual line goes through a point Q'' that is between Q and Q'; the distance from Q'' to Q' equals the new residual values of both P and R.

The desired balance between positive and negative residuals is therefore accomplished by making QQ'' twice as large as Q'Q''. In other words, Q'Q'' must be one third of 2.3158, or 0.7719. This value also happens to be the difference between the y-intercepts of the two lines (see the diagram). To obtain an equation for the zero residual line, it is only necessary to subtract 0.7719 from the y-intercept of the equation for PR. Therefore, what is the equation of the zero residual line?



A summary of the procedure: Given three non-collinear points, to obtain an equation for the zero residual line, first find the slope-intercept equation for the line through the leftmost and right most point, then calculate the residual of the middle point, then add one third of this residual to the y-intercept.

Notice that the residual of the middle point is sometimes positive and sometimes negative. In either case, adding one third of this value moves the line PR towards Q.

Exercises:

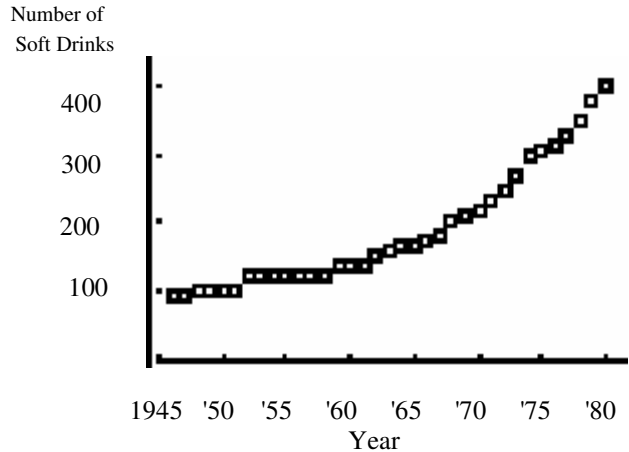
1. Given T = (0.98, 7.53), U = (6.03, 4.69) and V = (8.16, 1.44), find an equation for the zero residual line.
2. Is there a line that contains (7, 3), (6, 5), and (2, 8)? If so, find an equation; if not, find the zero residual line.
3. Is there a line that contains (-3, 16), (1, 10), and (9, -3)? If so, find an equation; if not, find the zero residual line.

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

Statistics Discussion #8: The Scatter Plot

Generally, more than three data points are collected in surveys and experiments. The data is generally collected as ordered pairs. A scatter plot is a graphic display of data points in a two-dimensional plane (xy plane). Each data point on a scatter plot represents two pieces of data for a single unit of observation. The most common plots are time plots--the time is always put along the horizontal axis.

Exercise: Soft Drinks. The following is a plot over time showing how many 12 ounce soft drinks the average person in the U. S. drank each year from 1945 to 1980. The following problem was taken from the Exploring Data packet by James M. Landwehr and Ann E. Watkins prepared for the American Statistical Association and National Council of Teachers of Mathematics Joint Committee on the Curriculum in Statistics and Probability.



1. About how many soft drinks did the average person drink in 1950? in 1970?
2. About how many six-packs of soft drinks did the average person drink in 1980?
3. About how many soft drinks did the average person drink per week in 1950? in 1980?
4. If the trend in the plot continued, about how many 12 ounce soft drinks did the average person drink in the year 2000?
5. In what year did soft drink consumption start to "take off"? Can you think of any possible reason for this phenomenon?

The graphing calculator may be used to find a scatter plot. Consider the following data relating a substance's temperature to its volume:

Temperature (C)	Volume(cc)
0	10.8
100	15.5
20	12.2
50	13.2
60	13.5
10	11.5
90	15.2
30	12.4
70	14.2
40	12.5

In order to find the scatter plot on the calculator, use the same procedure to enter the data as with univariate data:

First, clear old data from memory. Then, press **[STAT]** **<ENTER>**

Enter the first set of data under **L1** and the second set of data under **L2**.

Be sure that old graphs are cleared.

Now, set the range by pressing [WINDOW]. Examine the data to determine appropriate entries. In this case $0 \leq x \leq 100$ and $10 \leq y \leq 16$ appears appropriate. An x scale of 10 and a y scale of 1 are appropriate.

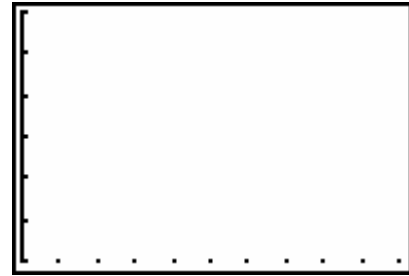
To draw the scatter plot, press [2nd] [y=] <ENTER>.

To turn on Plot 1, move the cursor to <ON> and press <ENTER>.

Select scatter plot and press <ENTER>. Let Xlist be L1 and Ylist be L2. Select the type of mark you would like on your graph. Press [GRAPH].

Draw this plot at the right. Label axes and scales clearly.

If you were to draw a line through these points, would the slope of the line be positive or negative?



Determining the slope and equation of this line can be very important in making predictions from your data. Press [2nd] [QUIT] to leave the graphing window.

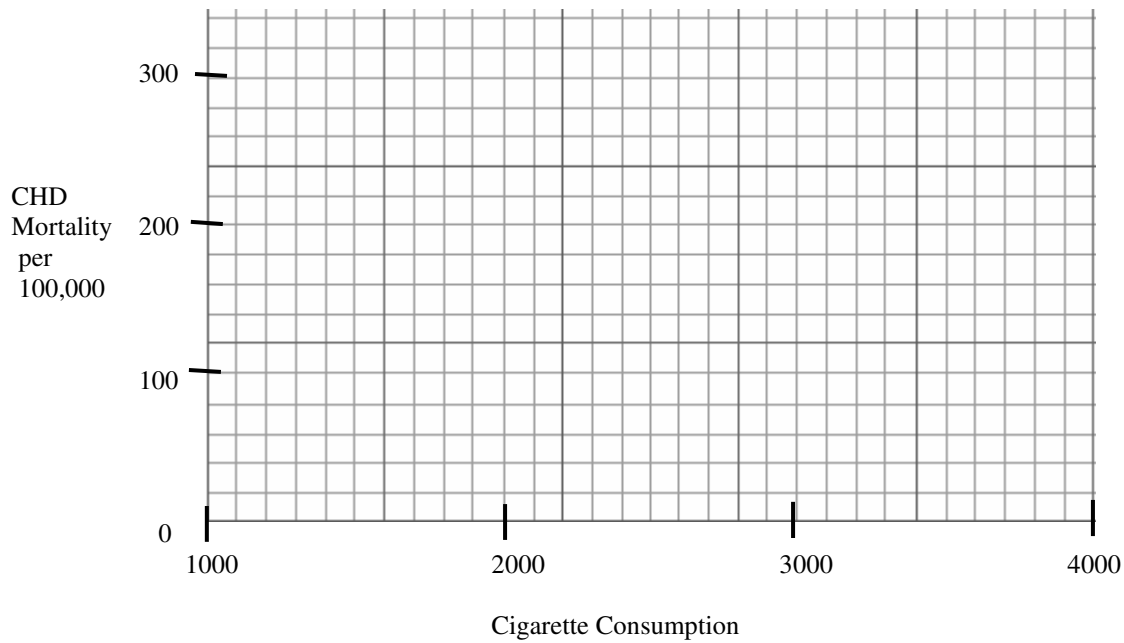
Statistics Discussion #9: The Median-Median Line

In order to determine the slope and equation of a line through more than three data points, the zero residual line technique can be extended. This process is known as finding the median-median line. Consider the following example.

Example: Smoking and CHD. The following table (Landwehr and Watson, 1987) lists 21 countries, along with the cigarette consumption per adult per year and the number of deaths per 100,000 people from coronary heart disease (CHD). From this table, can you discern a relationship between the yearly per-adult cigarette consumption of a country and the death rate due to CHD? How would you describe this relationship? How might the data in this table be displayed so that the relationship could be seen more clearly?

Country	Cigarette Consumption per Adult per Year	CHD Mortality per 100,000 (ages 35-64)
United States	3900	257
Canada	3350	212
Australia	3220	238
New Zealand	3220	212
United Kingdom	2790	194
Switzerland	2780	125
Ireland	2770	187
Iceland	2290	111
Finland	2160	233
West Germany	1890	150
Netherlands	1810	125
Greece	1800	41
Austria	1770	182
Belgium	1700	118
Mexico	1680	32
Italy	1510	114
Denmark	1500	145
France	1410	60
Sweden	1270	127
Spain	1200	44
Norway	1090	136

Create a scatter plot of the data in your graphing calculator. Now carefully graph the points on the following grid in order to prepare to find the median-median line.



The graph gives a visual sense of the trend in the data: a country that has a high rate of cigarette consumption generally has a high CHD mortality rate, and vice versa. We will use a median-median line to produce a best fit line. Its ease of use and resistance to the influence of outliers may make it the best descriptive approach. The method for finding median-median lines is outlined:

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. (The left and right groups should be equal in size.)
- (2) Find a summary point for each group using the median x value and the median y value.
- (3) Use these three summary points and find the zero residual line through them. This is the median-median line.

The "Smoking and CHD" data is easily divided into three groups:

First Group:

United States	3900	257
Canada	3350	212
Australia	3220	238
New Zealand	3220	212
United Kingdom	2790	194
Switzerland	2780	125
Ireland	2770	187

What is the median x value of the first group?

What is the median y value of the first group?

Middle Group

Iceland	2290	111
Finland	2160	233
West Germany	1890	150
Netherlands	1810	125
Greece	1800	41
Austria	1770	182
Belgium	1700	118

What is the median x value of the middle group?

What is the median y value of the middle group?

Third Group

Mexico	1680	32
Italy	1510	114
Denmark	1500	145
France	1410	60
Sweden	1270	127
Spain	1200	44
Norway	1090	136

What is the median x value of the third group?

What is the median y value of the third group?

Plot these three points on your graph and find the zero residual line for these three points. This is the median-median line for the data. What is the equation of this line? Graph this line on your plot.

Exercises:

1. Plot the following nine ordered pairs: $(0, 1)$, $(1, 2)$, $(2, 2.7)$, $(3, 4)$, $(4, 3)$, $(5, 4.6)$, $(6, 6.2)$, $(7, 8)$, $(8, 8.5)$.

a. Find the median-median for these nine points.

b. If the number of data points is not divisible by three, the three groups cannot have the same number of points. In such cases, it is customary to arrange the group sizes in a symmetric fashion. For instance, enlarge the data set to include a tenth point, $(9, 9.5)$, and then separate the ten points into groups, of sizes 3-4-3 points, reading from left to right. Calculate the summary points for these three groups and find the median line for the set of ten points.

c. Enlarge the data set again to include an eleventh point, $(10, 10.5)$. There are many ways to divide the points, 3-5-3 or 4-3-4, for example. Choose the one that you believe will give the best results. Find the median-median line and explain your reasoning.

2. If you are camping in the woods, can you tell what the temperature is by how quickly the crickets chirp?
Examine the data below:

Temperature (°C)	Chirps/min
18	110
19	110
20	130
21	135
23	154
24	158
26	179
29	201
31	210
32	230

- Use graph paper to construct an accurate graph of this data. Be sure to accurately label and scale the axes.
 - Find the median-median line for the data points.
 - In your model, what is the rate of change of the number of chirps with temperature?
 - According to your model, at what temperature will the crickets be quiet?
 - At 27°C, what would you expect the number of chirps/min. to be?
3. In a research article published in 1965, data were published that compared the death rate due to cancer and the proximity to the Hanford Nuclear Reservation in Washington state. The Hanford site produced nuclear weapons for the government from soon after World War II until it was closed recently for safety reasons. The table below lists eight counties or cities along the Columbia River, into which radioactive waste has been seeping from the Hanford site. For each of these counties or cities, the table lists the death rate due to cancer for each 100,000 residents. It also lists an exposure index, which takes into account both the amount of river frontage and the distance from the Hanford site. The higher the index number, the greater the exposure rate. It is easier to work with ordered data. If you enter the data in your calculator in L1 and L2, sort the pairs by entering

[STAT] [2] L1,L2 <ENTER>

This keeps the ordered pairs correctly matched by sorting according to L1 and carrying the appropriate L2 entries along.

County/City	Exposure Index	Cancer Deaths
Umatilla	2.5	147
Morrow	2.6	130
Gilliam	3.4	130
Sherman	1.3	114
Wasco	1.6	138
Hood River	3.8	162
Portland	11.6	208
Columbia	6.4	178
Clatsop	8.3	210

- a. On graph paper, construct an accurate graph of this information . Be sure to label and scale the axes.
- b. Find the median-median line for the data points.
- c. What would be the death rate due to cancer in a county or city that has an exposure index of zero? Explain.
- d. Which city or county has an actual death rate farthest from the death rate predicted by your line?
- e. What is the meaning of the slope of your median-median line?
- f. Use your equation to predict the death rate due to cancer for a county or city with an exposure index of 4.0.

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