Module 5: Statistics

This material corresponds to chapters 5, 6 & 7 of the textbook, *For All Practical Purposes*

STATISTICS
CHAPTERS 5, 6 AND 7

TIME FRAME: 17 days

ENDURING UNDERSTANDINGS:

There are various ways to describe data, including plots and numerical summaries, each of which has strengths and weaknesses in certain contexts. Choosing the appropriate way to express data is an important step in analyzing it.

ESSENTIAL (ASSESSMENT) QUESTIONS:

1. Explain the meaning of the statement "Association does not imply causation"
2. Give examples of multiple graphical ways to represent statistical data, and explain benefits and shortcomings of each.
3. Explain bias in sampling, and give several examples.

CRMS:

Connections 3.3
Probability/Statistics 6.2, 6.3, 6.4
Algebra 7.1, 7.3
Functions 8.2, 8.3

AT THE END OF THE MODULE STUDENTS WILL KNOW AND BE ABLE TO:

1. Calculate mean, median, quartiles, standard deviation.
2. Graph histrograms, boxplots, stemplots, scatterplots.
3. Give a 5-number summary.
4. Explain correlation and calculate it using a graphing calculator.
5. Identify bias in sampling methods.
6. Design an experiment that minimizes bias.

PREREQUISITE KNOWLEDGE/SKILLS: Graphing points in a Cartesian coordinate system

PRE-ASSESSMENT: None

ACTIVITIES

Gathering Data: Reaction Times
U.S. Average Incomes and the Forbes 400 Worksheet
Calculating a Line of Best Fit Worksheet ***(Optional)***
How Long Are The Strings?

POST-ASSESSMENTS:

Chapter 5 Exam
Linear Regression Take-Home Assessment

LITERACY STRATEGY INTRODUCED:

Finding information on the internet and citing sources.

RESOURCES

TI-84 Graphing Calculators
Centimeter Rulers (30cm minimum)
Access to the internet
Paper lunch bags
String cut in 2-, 4- and 6-inch lengths

DAILY PLAN

Pick at least two days each week to begin class with a warm-up question from the "Algebra Skills Review" found at the end of the Preface to this Teacher Resource Manual.

Day 1

- Activity: **Gathering Data: Reaction Times**
- Ch. 5 histograms and stemplots

HW: 7e p. 210 #3-8, read pp. 179-188
8e p. 175 #2-8, read pp. 150-157

Days 2-3

- Mean and Median
- Worksheet: **U.S. Average Incomes, and the Forbes 400** (students will need access to the internet for this activity)

HW: **7e** p. 213 #11-13, read pp. 188-192
8e p. 177 #13-15, read pp. 157-160

Day 4

- Quartiles and 5-number summary, boxplot

HW: 7e p. 214 #17-18 and 21-23, read pp. 192-195
8e p. 176 #19-21, 23, 25, read pp. 163-173

Days 5-6

- Standard deviation, normal distributions, the 68-95-99.7 rule

HW: 7e pp. 216-218 #29-31, 33, 38-40, 42, read pp. 195-207
8e pp. 179-181 #31-34, 38-40, 42 read pp. 163-173

Day 7

- Review Worksheet

Day 8

- Chapter 5 Exam (a sample exam is included for the instructor)

Day 9

- Scatterplots, regression lines

HW: 7e pp. 242-243 #2-5, 9, 13, read pp. 222-228
8e pp. 202-204 #2-5, 9, 13, read pp. 186-190

Days 10-11

- Correlation and causation

HW: 7e pp. 245-246 #15, 17, 18, 21, 22, read pp. 228-232
8e pp. 204-205 #15, 17, 18, 21, 22, read pp. 191-193

Day 12

- **Calculating a Line of Best Fit** worksheet **(optional)**

HW: *See handout*

Days 13-14

- Least-squares with formulas, spreadsheets and calculators
- Distribute: *TI-84 Calculator Reference Guide*
- Distribute: *Linear Regression Take-Home Assessment*

HW: **7e** pp. 246-247 #27, 28, 34, 35, read pp. 232-239
8e pp. 205-206 #27, 28, 33, 35, read pp. 194-200

Day 15

- Introductory activity about bias: **How long are the strings?**

HW: 7e p. 285, #1-4, read pp. 254-257
8e p. 237, #1-4, read pp. 211-214

Day 16

- Random Digits Table
- Terminology: sampling, bias, voluntary response sample, simple random samples, table of random digits

HW: 7e p.285-287, #7-9, 11-12, 15; read pp. 257-262
8e p.237-238, #7-9, 11-12, 15; read pp. 214-217

Day 17

- Terminology: undercoverage and nonresponse (students read ahead), confidence intervals
- Observation and experiment
- Statistical significance

HW: 7e pp. 287-290, #18-20, 28, 31, 33; read pp. 263-270
8e pp. 238-240, #18-20, 28, 31, 33; read pp. 218-225

GATHERING DATA: REACTION TIMES
INTRODUCTORY ACTIVITY FOR CHAPTER 5
Description

NOTE: This activity is adapted from "Real-Life Math Statistics", by Eric T. Olson and Tammy Perry Olson, 2000, Walch Publishing, Portland, Maine.

DESCRIPTION

Student will gather data about reaction times by working in pairs and taking turns either catching or dropping a ruler: by measuring the distance the ruler falls before it is caught, they will be able to determine how much time passed after the ruler was dropped before it was caught.

Repeated experimentation with this will allow students to gather data that will be used to illustrate histograms and stemplots; the data will continue to be useful as a reference for discussions throughout this module.

MATERIALS

Rulers (with centimeter units; should be 30cm long) – 1 for each pair of students
Copies of activity sheet on next page – 1 for each student

ESTIMATED TIME

5 minutes for explanation and demonstration, and to distribute activity sheets
20 minutes for group work
Remaining part of class used for drawing and discussing histograms

DEMONSTRATION

Begin the activity by demonstrating, with the help of one student, how to hold and drop the ruler so that the ruler, but not the hand, is visible to the seated student, and so that the seated student is correctly lining up his or her hand at the bottom of the ruler. Also make a point of the fact that the ruler should be held with 0cm at the bottom so that mistakes are not made in the data-recording.

DISCUSSION

On the board or overhead projector, demonstrate to students how to plot the data on a histogram on the back side of their activity sheets. After they do this (allow them 5 to 10 minutes), ask them questions about what they think their average reaction time was, and whether any of their measurements seemed to be unusual compared to the majority of their individual results. These ideas will be useful later as a reference when you discuss mean, median and outliers in class.

TIP

You might collect all the students' data before they draw histograms and then redistribute them randomly – this might help students to focus less on their own performance or on their peers'.

GATHERING DATA: REACTION TIMES

THE EXPERIMENT

One person sits in a chair, and a second person stands behind her. The standing person drops a ruler, and the first person catches it. By measuring how far the ruler falls, you will be able to determine how long it took for the seated person to catch it.

The standing person holds the ruler a few inches in front of the seated person's face so that the ruler, but not the hand holding it, is visible. The seated person starts with her thumb and forefinger on either side of the ruler, about 1 inch apart, aligned with the very bottom of the ruler. When the ruler is dropped, the seated person closes her fingers to catch it. They then record in the table below the distance the ruler falls.

After performing 20 trials of this experiment and recording all the results, switch roles and repeat the process.

CALCULATION: CHANGING DISTANCES INTO TIMES

Because the ruler will be in a free fall accelerated by gravity, there is a formula from physics that tells us how to determine from the distance traveled how long the ruler was falling. If time is measured in seconds and distance is measured in centimeters, then the conversion formula is

$$t \approx (0.072)\sqrt{d}.$$

Use this formula to fill in the second column in the table below.

Trial	Change in Position	Reaction Time	Trial	Change in Position	Reaction Time
1			11		
2			12		
3			13		
4			14		
5			15		
6			16		
7			17		
8			18		
9			19		
10			20		

Results

In the next few lines, describe your results in words. Include a discussion of any factors that you think affected the results.

____

____

____

____

____

____

____

Histogram

Use the space below and a ruler to *carefully* draw a histogram that illustrates the reaction time data in the table. You may want to sketch the histogram first on a separate paper.

U.S. AVERAGE INCOMES AND THE FORBES 400

Description

The following activity is designed to illustrate the effects of outliers on means. Students will need to look up certain information on the internet to complete the activity. It is hoped that using current real-world data makes the activity more interesting since it is otherwise a straight-forward application of the definitions of mean and median.

ESTIMATED TIME: 35 minutes (provided students have ready access to the internet)

If students do not have immediate access to the internet in the classroom, the teacher may wish to assign the activity as homework, or to ask them to gather the relevant data as homework in advance of the activity or teachers can print out a data sheet so in groups students can work with the data.

ABOUT THE REQUIRED DATA

Students are asked to look up the following information for this activity. Here are some answers found at the time of this writing and their sources. Note that other answers are possible, but they should not differ significantly.

Mean U.S. Household Net Worth: $136,171
Source: *www.census.gov, search for "mean net worth"*

Number of U.S. Households: 113,146,000
Source: *Wikipedia article "Household Income in the United States"*

Total Net Worth of Forbes 400 richest people: $1.54 trillion
Source: *www.forbes.com*

WORKSHEET: U. S. AVERAGE NET WORTH AND THE FORBES 400

Use the internet (or another current resource) to find answers to the following questions. For each question, indicate the source you used to find the answer (for example, include the URLs of any web sites you used).

(1) What is the (estimated) number of households in the United States?

(2) What is the mean net worth of households in the United States?

Based on your answers to the two previous questions, calculate the total net worth of all households in the United States.

Here's another question to look up on the internet (remember to cite your source):

(3) What is the total net worth of the Forbes 400 richest people in the United States?

Imagine those 400 people lose 90% of their assets somehow. What will the new total net worth of U.S. households be? And what will be the new mean net worth of U.S. households?

Explain carefully why the same scenario would not change the median U.S. household net worth at all. (You may assume that, even after the 400 richest people lose 90% of their wealth, they will still be very wealthy.)

REVIEW WORKSHEET

Description

The worksheet on the next few pages is meant to be copied and distributed to your students to help them prepare for the exam on Chapter 5 material. It is imagined that students will work in groups to answer all the questions and help each other understand the methods while the teacher moves around the room giving hints and helping student to get unstuck.

ESTIMATED TIME: 30 minutes

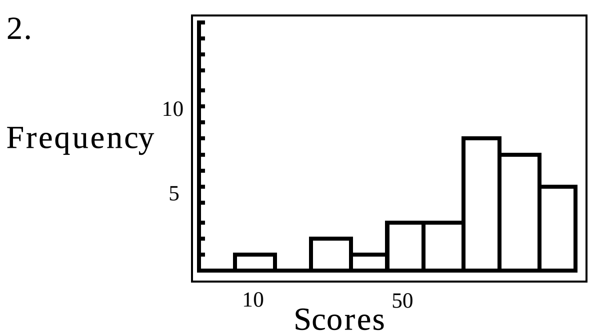
Answers to the worksheet questions:

1.

0	
1	7
2	
3	68
4	5
5	156
6	379
7	01335677
8	1257799
9	23689

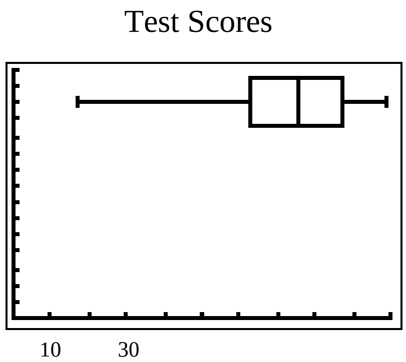
"Test Scores"

2.



3. $72.2\overline{3}$
4. 75.5
5. 63 and 87
6. 17 63 75.5 87 99

7.



8. Variance: 42.38
   Standard Deviation: 6.519
9. 57, 114 and 228
10. 25
11. more than 25.25 hours

Review Worksheet for Chapter 5

Questions 1 through 7 pertain to the following set of scores on an exam:

89	76	93	96	63	87	38	17
99	92	56	87	45	98	77	77
36	69	89	85	81	73	55	67
51	73	75	82	71	70		

1) Make a stemplot of the data.

2) Draw a histogram of the data, using intervals of width 10.

3) Find the mean of the data.

4) Find the median of the data.

5) Find the first and third quartiles for the data.

6) Write a five-number summary for the data.

7) Accurately draw a boxplot for the data. Be sure to have a number line below the graph with units and title labeled.

8) Find the variance and the standard deviation (round to 3 decimal places) for the following data:

3, 6, 7, 9, 20

9) The average of three numbers is 133. The middle number is twice the smallest number, and the largest number is 4 times the smallest number. What are the three numbers?

10) The distribution of scores on a standardized test is normal with a mean of 150 and a standard deviation of 20. A thousand students took the test. Find the number of students who scored above 190.

11) The average length of time, per week, that students at a certain university spend on homework is normally distributed with a mean of 24.3 hours and a standard deviation of 1.4 hours. Jane tells her parents that she spends more time studying than 75% of the students on campus. How many hours per week must Jane spend on homework for this to be true?

SAMPLE EXAM

1. Make a stemplot of the following data:

17	31	25	30	21
25	28	19	32	25
32	27	21	26	21

2. The following data represents exam scores for 26 students in an English course. Make a histogram of the scores.

84	77	67	94	90	77	79
81	56	89	77	88	72	93
74	76	28	80	58	94	
66	77	89	81	78	93	

3. Find the median of the exam scores in question 2.
4. Find the mean of the exam scores in question 2.
5. Below are the ages of students attending an art exhibit. Find the first and third quartiles of the data.
   11, 11, 12, 12, 13, 13, 13, 13, 13, 14, 14, 15, 15, 16, 16, 17, 17, 18
6. Find the five-number summary for the data in question 5.
7. Draw a boxplot for the data in question 5.
8. Find the variance of the following set of data.
   5, 7, 17, 31, 47, 68, 85, 96, 99
9. Find the standard deviation for the data in question 8.
10. Two towns both have a mean income for their residents of $30,000. The standard deviation of income of residents in town A is $2600 and the standard deviation of incomes of residents in town B is $25,000. Explain what this says about the difference in the distribution of incomes in the two towns.
11. In the list of five measurements shown below, one of the numbers is blurred. What must this number be if the mean of the five measurements is 7?
    8 5 10 3 ??
12. In the list of five measurements shown below, one is blurred. Supply a fifth value that would make the median 8.
13. The scores on a standardized test form a normal distribution with a mean of 300 and a standard deviation of 40. Two thousand students took the test. Find the number of students who scored above 380.
14. The distribution of the scores on a standardized exam is approximately normal with mean 400 and standard deviation 35. What percentage of scores lies between 435 and 470?
15. The distribution of scores on a standardized exam is approximately normal with mean 400 and standard deviation 35. Between what two values do the middle 50% of scores lie?

CALCULATING A LINE OF BEST FIT WORKSHEET

Description

This activity asks students to find a least-squares linear regression by hand. It requires them to recall the point-slope formula for a line and the formula for the x-coordinate of the vertex of a parabola. It is intended that students will complete this activity before being introduced to the procedure for finding the regression with a graphing calculator.

This activity is optional! Although it requires only second-year algebra in terms of the tools used, the level of mathematical maturity required is rather high. You should feel free to skip this activity if you feel your students will not benefit from it enough to justify the time spent. If you do skip the worksheet, then you should spend some class time talking about the idea of a least-squares regression and drawing a couple pictures so that your students will know what it is that a linear regression on their calculator gives them.

ESTIMATED TIME: 40 minutes

KEY IDEA

The **least-squares** linear regression is the line that fits the data with the smallest total of the sums of the squares of the vertical distances from data points to the line.

WORKSHEET: CALCULATING A LINE OF BEST FIT

Answer Key

1. Three pairs of shoes were collected to have their eyelets counted and the lengths of their shoelaces measured. The results are recorded in the table at right.

Number of Eyelets	Length of Shoelace
8	39 in
12	48 in
16	63 in

(a) In the following space, plot as carefully as you can these three data points on an *xy*-plane, letting *x* be the number of eyelets and *y* be the length of the shoelace. Notice that the points do not all lie on a straight line, but you can draw a line that comes rather *close* to all the points. Use a ruler or straight-edge to draw a line on the graph that comes close to all four of the plotted points.

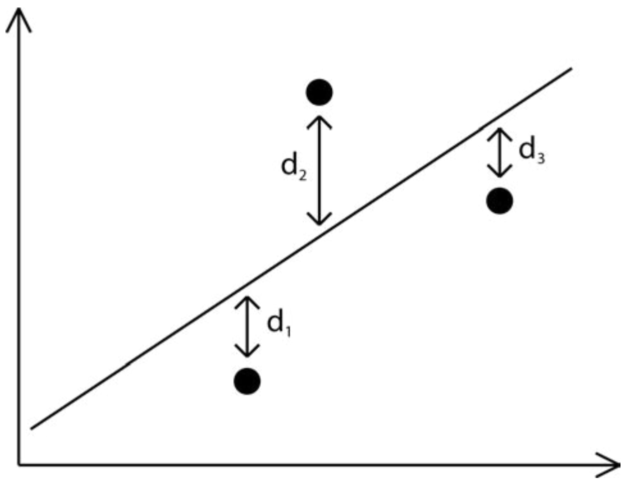
Answers will vary.

(b) Use the line you drew to estimate, as precisely as you can, the length of a shoelace that would fit a shoe with 10 eyelets.

Answers will vary.

The answer you got to part (b) on the previous page depended on the line you drew, and that line will vary a little bit with each person who draws it. When we're trying to make predictions like this, it would be nice if there was a mathematical way of prescribing the line that fits the data so that we could avoid that kind of ambiguity.

Fortunately, there is such a method. It's called the **least-squares line of best fit**; the reason for that name is explained by the next figure of a line being fit to some given data points:



The line drawn in the figure might be any line of the form **y=mx+b**; the *least-squares* line of best fit is the one that makes the following quantity as small as possible:

$$d_1^{\,2} + d_2^{\,2} + d_3^{\,2}$$

The least-squares best fit line doesn't necessarily go through all the data points -- it might not even go through any of them! But the idea is that, in some sense, this line will fit the data "on average".

In fact, it turns out that the least-squares line of best fit passes through the point $(\bar{x}, \bar{y})$, where $\bar{x}$ is the average of the x-values and $\bar{y}$ is the average of the y-values in your data set.

2.

(a) Let x be the number of eyelets in a shoe and let y be the length of the corresponding shoelace. Use the data for question (1) to find $\bar{x}$ and $\bar{y}$.

$\bar{x} = \frac{8+12+16}{3} = 12$ **and** $\bar{y} = \frac{39+48+63}{3} = 50$

(b) Write down an equation the point slope form of the line with slope m that passes through the point $(\bar{x}, \bar{y})$ that you found in 2(a). *(Your answer will have the unknown slope m in it -- we won't know the value of m for a few more steps.)*

$y - 50 = m(x - 12)$ or $y = mx - 12m + 50$

(c) Use the table of data from question (1) to find d_1, d_2, and d_3. *(Your answers will have the unknown slope m in them again.)*

$$d_1 = |-4m + 11|$$

$$d_2 = |m + 2|$$

$$d_3 = |4m - 13|$$

(d) Now write down $d_1^2 + d_2^2 + d_3^2$ and simplify it as much as possible. *(Again, you will have the unknown slope m in your answer, but you should be able to combine like terms and simplify a lot.)*

$$(-4m+11)^2 + (m+2)^2 + (4m-13)^2 = (16m^2 - 88m + 121) + (m^2 + 4m + 4) + (16m^2 - 52m + 169)$$
$$= 33m^2 - 136m + 294$$

(e) The answer to (e) is a function of m. What kind of function is it?

Quadratic.

(f) Find the value of m that minimizes the expression you found in 3(d). *(Hint: What do you remember about the vertex of a parabola?)*

The vertex of a parabola $y = ax^2 + bx + c$ *occurs at* $x = \frac{-b}{2a}$. *So the vertex here occurs at*

$$m = \frac{-(-136)}{2(33)} = \frac{68}{33} = 2.\overline{06}$$

(g) Plug your answer from 3(f) into your equation from 3(b) to get a complete equation for this line of best fit.

$$y = \frac{68}{33}(x - 12) + 50$$

(h) Use the answer to 3(g) to predict the length of a shoelace that will fit a shoe with 10 eyelets.

$$y = \frac{68}{33}(10 - 12) + 50 = 50 - \frac{134}{33} = 45.\overline{93}\, inches$$

Do it on your own: You operate a small stand at the beach selling sunglasses. By experimenting, you discover that when you charge $30 for each pair of sunglasses, you sell 10 pairs per day; when you charge $20, you sell 18 pairs; and when you charge $15, you sell 28 pairs. Use a least-squares line of best fit to predict what price will allow you to sell 25 pairs per day.

WORKSHEET: CALCULATING A LINE OF BEST FIT

1. Three pairs of shoes were collected to have their eyelets counted and the lengths of their shoelaces measured. The results are recorded in the table at right.

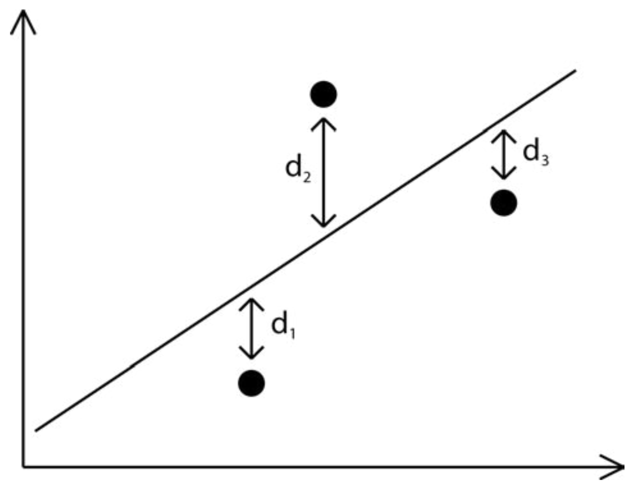
Number of Eyelets	**Length of Shoelace**
8	39 in
12	48 in
16	63 in

(a) In the following space, plot as carefully as you can these three data points on the coordinate system, letting *x* be the number of eyelets and *y* be the length of the shoelace. Notice that the points do not all lie on a straight line, but you can draw a line that comes rather *close* to all the points. Use a ruler or straight-edge to draw a line on the graph that comes close to all three of the plotted points.

(b) Use the line you drew to estimate, as precisely as you can, the length of a shoelace that would fit a shoe with 10 eyelets.

The answer you got to part (b) on the previous page depended on the line you drew, and that line will vary a little bit with each person who draws it. When we're trying to make predictions like this, it would be nice if there was a mathematical way of prescribing the line that fits the data so that we could avoid that kind of ambiguity.

Fortunately, there is such a method. It's called the **least-squares line of best fit**; the reason for that name is explained by the next figure of a line being fit to some given data points:



The line drawn in the figure might be any line of the form **y=mx+b**; the *least-squares* line of best fit is the one that makes the following quantity as small as possible:

$$d_1^{\ 2} + d_2^{\ 2} + d_3^{\ 2}$$

The least-squares best fit line doesn't necessarily go through all the data points -- it might not even go through any of them! But the idea is that, in some sense, this line will fit the data "on average".

In fact, it turns out that the least-squares line of best fit passes through the point $(\bar{x}, \bar{y})$, where $\bar{x}$ is the average of the x-values and $\bar{y}$ is the average of the y-values in your data set.

2.

(a) Let x be the number of eyelets in a shoe and let y be the length of the corresponding shoelace. Use the data for question (1) to find $\bar{x}$ and $\bar{y}$.

(b) Write down an equation for a line with slope m that passes through the point $(\bar{x}, \bar{y})$ that you found in 2(a). *(Your answer will have the unknown slope m in it -- we won't know the value of m for a few more steps.)*

(c) Use the table of data from question (1) to find d_1, d_2, and d_3. *(Your answers will have the unknown slope m in them again.)*

$d_1 =$

$d_2 =$

$d_3 =$

(d) Now write down $d_1^{\ 2} + d_2^{\ 2} + d_3^{\ 2}$ and simplify it as much as possible. *(Again, you will have the unknown slope m in your answer, but you should be able to combine like terms and simplify a lot.)*

(e) The answer to (e) is a function of m. What kind of function is it?

(f) Find the value of m that minimizes the expression you found in 3(d). *(Hint: What do you remember about the vertex of a parabola?)*

(g) Plug your answer from 3(f) into your equation from 3(b) to get a complete equation for this line of best fit.

(h) Use the answer to 3(g) to predict the length of a shoelace that will fit a shoe with 10 eyelets.

Do it on your own: You operate a small stand at the beach selling sunglasses. By experimenting, you discover that when you charge $30 for each pair of sunglasses, you sell 10 pairs per day; when you charge $20, you sell 18 pairs; and when you charge $15, you sell 28 pairs. Use a least-squares line of best fit to predict what price will allow you to sell 25 pairs per day.

TI-84 CALCULATOR REFERENCE GUIDE FOR STATISTICS

The following four pages are meant to be copied and distributed to your students so that they can have as a reference while using the graphing calculators to find linear regressions and perform other statistical calculations.

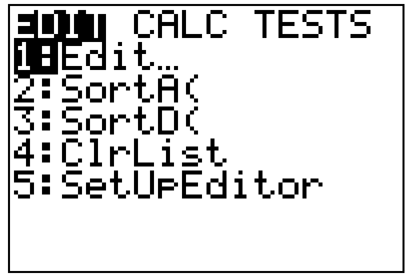
TI-84 Calculator Reference Guide for Statistics

ENTERING LISTS

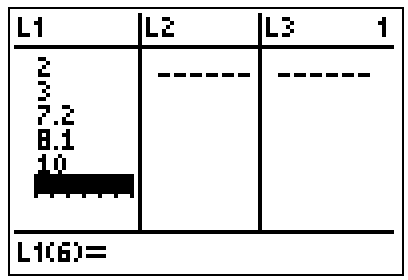
x	y
2	5
3	8
7.2	9
8.1	9
10	8

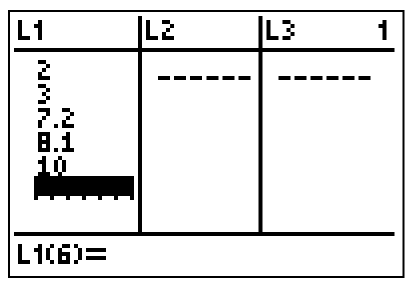
Most statistical calculations involve working with tables of data, so we need to be able to input such tables into the graphing calculator. In the following example, we will enter the data in the table at right. The same data will be used in the later examples in this guide.

1. Select the STAT menu by pressing:

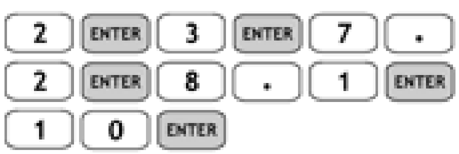


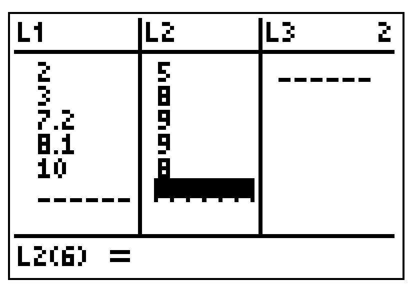
2. Select the first item in order to edit the calculator's lists:





3. Enter the x-values in the first column, labeled L1, as follows:





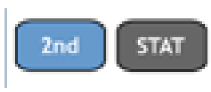
4. Use the right arrow key to move the cursor to the first row of the second column (labeled L2); then enter the y-values as follows:

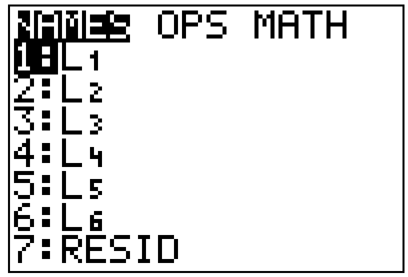
TI-84 Calculator Reference Guide for Statistics

CALCULATIONS WITH LISTS

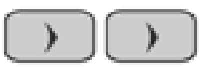
Once you have entered data lists, you can have the calculator compute standard statistical quantities from that data, like the mean and median. The following example will illustrate how to calculate the mean of data already entered into the list L1.

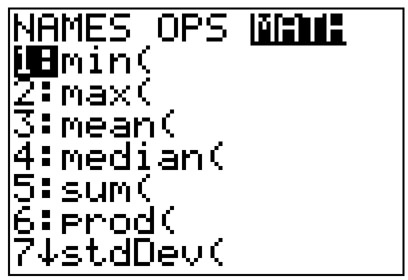
1. From the main calculation screen (the one use to multiply numbers, for example), go to the LIST menu by pressing:



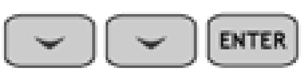


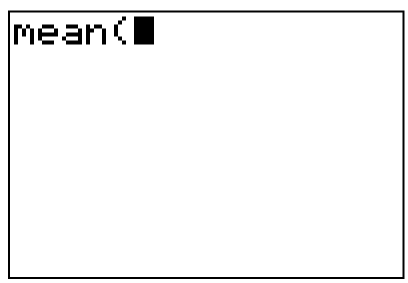
2. Use the right arrow key to scroll to the MATH menu:



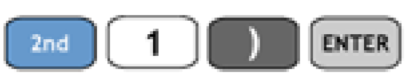


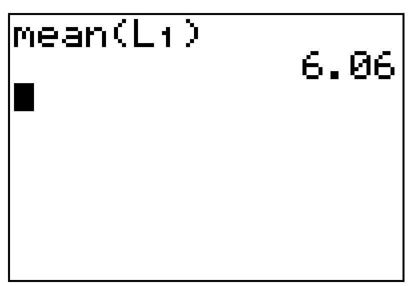
3. Select **mean** from the menu by pressing 3 or scrolling down to the item with the arrow keys and then pressing enter.





4. Tell the calculator to use the data in the list L1, then close the parentheses and press enter to view the result.

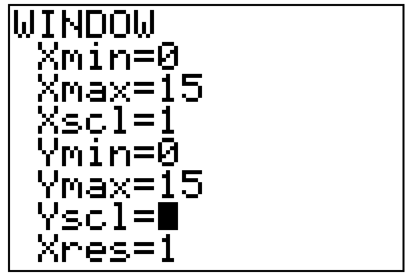




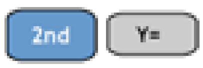
SCATTERPLOTS

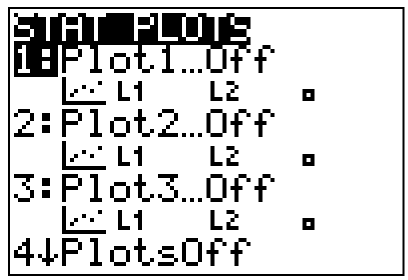
You can plot individual points on an xy-plane by performing a scatterplot.

1. First, set the window as you would when graphing a function. To got to the menu for this, press:

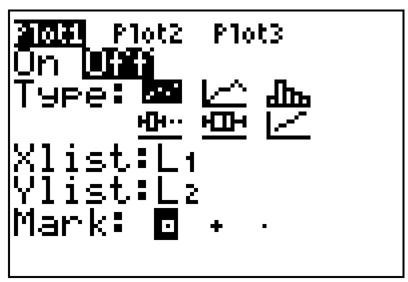


2. Go to the STAT PLOT menu by pressing:

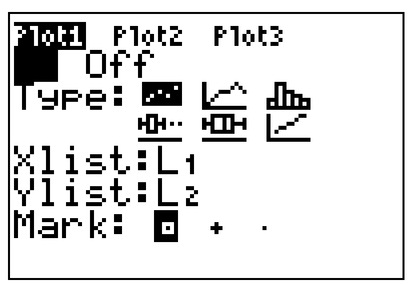




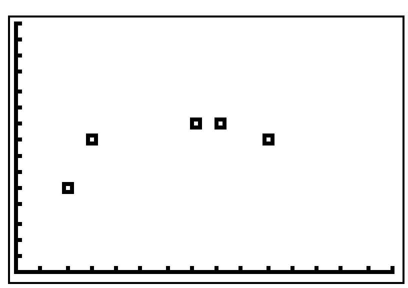
3. Select the first plot by pressing 1 or enter:



4. Select **On** by pressing enter:



5. View your plot by pressing:

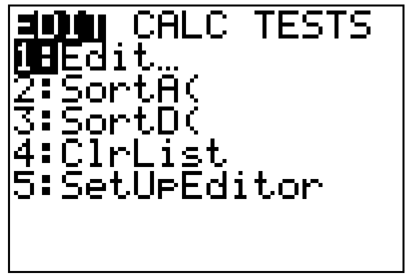


TI-84 Calculator Reference Guide for Statistics

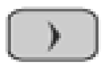
REGRESSIONS

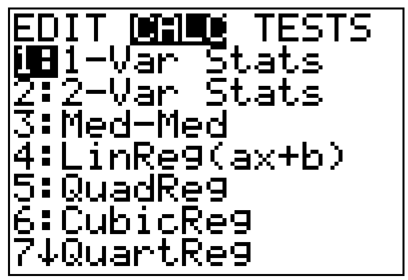
You can fit a line or curve to a set of data points by performing a regression with the graphing calculator. In this example, we find a line of best fit using a linear regression.

1. Go to the STAT menu by pressing:

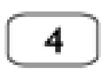


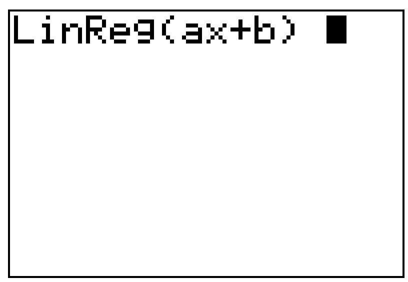
2. Use the right arrow key to select the CALC menu:



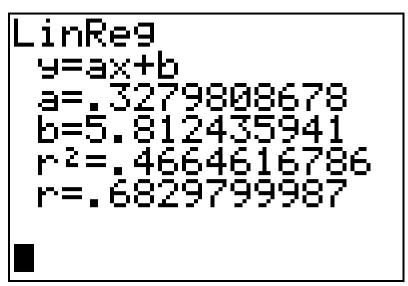


3. Select **LinReg(ax+b)** to perform a linear regression on the data in lists L1 (for x-values) and L2(for y-values):





4. Calculate the coefficient for the regression by pressing:



5. The resulting screen gives you the coefficients for the regression, as well as the correlation coefficient r and its square.

ASSESSMENT PROJECT: LINEAR REGRESSION

Description

The activity on the following two pages is meant as a take-home assessment for students to apply what they've learned in from Chapter 6 about finding linear regressions and correlation coefficients.

Part of the project asks students to gather information online regarding used car prices, and then to use that information to predict the price of a car. You should not plan to grade this piece for absolute correctness since each student will likely have different answers, but you should make sure that they uses a line of best fit to make the prediction (rather than simply looking up the price that they are asked to predict).

The remainder of the project asks students to graph some given data as a scatterplot and to find the correlation coefficient using a calculator, then to answer some questions about the data. The answers are given below.

LINEAR REGRESSION – TAKE-HOME ASSESSMENT

1. Select a model of used car that you would consider purchasing:

Make:

Model:

2. Go to www.kbb.com (Kelly Blue Book online) to find the current suggested retail value of the car (in good condition) you selected from each of the following years:

1998:

1999:

2002:

2003:

2005:

3. Use your calculator to find a least-squares linear regression for a linear function $y=mx+b$ that predicts the suggested retail value for your model car made in the year x.

4. Use the function you found in part 2 to predict the suggested retail value of a model from each of the following years:

2004:

2001:

1990:

5. Do each of the predictions your model gave seem reasonable to you? Explain.

6. Each line in the following table of data gives information about a Major League baseball team for one season: that team's number of hits (H) and number of wins (W) throughout the year.

Hits	Wins
1415	78
1415	79
1441	79
1414	90
1333	76
1636	96

Generate a scatterplot of the data from this table.

7. Use your calculator to find a least-squares linear regression for the data in question 6, and add the graph of that line to the scatterplot above.

Equation of line:

8. What is the correlation coefficient for the linear regression you found in question 7?

r=

9. Use the data from question 6 to argue in support of or against the following statement: "A baseball team's number of wins is correlated with how many hits it has in a season." Use what you learned in chapter 6 to support your answer. (Remember to consider the effects of outliers.)

SAMPLING ACTIVITY: HOW LONG ARE THE STRINGS?

Description

MATERIALS REQUIRED

- Brown paper lunch bags (1 for each pair of students)
- Pre-cut lengths of string (need 2 ft of string for each pair of students, cut into smaller segments as described below)
- Optional: Glue (to keep the ends of string segments from fraying)

DESCRIPTION

Students work in pairs to try to figure out, via sampling the average length of the string segments in their bags. Because they will be more likely to pull out a longer string than a shorter one, sampling bias will be illustrated.

ADVANCE SET-UP

This activity requires some preparation by the teacher. Each pair of students will receive a brown paper lunch bag with six segments of string: two 6-inch segments, two 4-inch segments, and two 2-inch segments. Teachers will have to cut these length of string in advance; it is also strongly advised to dip each end of each cut string segment with glue (otherwise the ends fray and the entire strings can unravel in the paper bags during the activity).

INSTRUCTIONS TO STUDENTS

Each team of two has been given a paper bag with several short lengths of string in it. Do not look inside the bag!

Each of the string segments is either 2-inches long, 4-inches long, or 6-inches long. Your team's job is to try to figure out the average (that is to say, the mean) length of string in your bag by `taking a poll": you will reach into your bag (without looking), pull out the first segment of string that you touch, write down on a table what length it was, and then put the string back into the bag. Shake the bag to shuffle the strings around, and repeat the whole thing. Do this a total of 30 times. When you're done, calculate the average length of all the string lengths you recorded. Then record your answer in a table on the board.

FURTHER DESCRIPTION

Thirty trials is enough that all teams should see an average length of something around 4.7 inches, with perhaps a few outliers. After there seems to be agreement on this point, have students dump the strings out of the bag and determine exactly what the average length is. They will see that the average length is 4 inches. At the point, the teacher should ask students to discuss why they think there was a tendency to overestimate the average.

The answer is that what they really computed was a weighted average: a 4-inch string was twice as likely to be grabbed as a 2-inch string because there is twice as much of it to touch; similarly, a 6-inch string was 3 times as likely to be the first one touched. Because not all string were equally likely to be grabbed, this is an example of what statisticians call bias – a tendency to favor a certain subgroup of the population being polled.

If time permits, the teacher should encourage students to brainstorm other instances where such bias might show up. A few examples are:

- door-to-door polling neglects the homeless entirely (this is an important census issue)
- telephone polling (younger people are less likely to be polled because they are less like to have a land-line, and pollsters are prohibited from calling cell phones)