AP STATISTICS

Summer Assignment

2019 -2020 School Year

South County High School

Please Note: ONLY PAGES 5, 8-15 NEED TO BE PRINTED!!!!

Congratulations on deciding to take statistics! Most students are surprised to find that statistics is very different from what they expected and that it is a very practical course. Though considered a math course, it is unlike any other you have ever taken. One element of the course work that surprises students is that solutions require good written communication, not just numerical answers. This is as much a writing course as it is a math course!

Brief Description of Summer Assignment: The summer assignment consists of 4 parts all contained in this packet. It introduces you to statistics and gives you practice on some basic statistical concepts.

Objectives of Summer Assignment: For students to…

- become excited about the study of statistics,
- learn that it is a course requiring lots of reading and writing,
- gain understanding of basic statistical topics, and
- review statistical concepts that you were exposed to during Math 7/8, Algebra I and Algebra 2.

Resources Necessary to Complete Assignment: Graphing calculator, Internet Access

Due Date: The THIRD DAY OF CLASS. It will count as a 30 point quiz grade for first quarter. There will be a content in class quiz on that day on parts 4 and 5.

Estimated Time to Complete: About 4 to 6 hours to do a good job on this assignment

Remember, this is an AP Course! Do not expect it to be "easy". Although it may not seem as difficult computationally as calculus, it requires a great deal of outside reading and homework, and a thorough understanding of many abstract concepts.

You are now an AP student. Lesson #1 – Do not procrastinate! Start the summer assignment early to allow for time for questions if necessary. If you have any questions or problems, you may contact me via e-mail at abaylor@fcps.edu. Please do not wait until the last minute to begin or to receive clarification about the assignment.

Enjoy your summer!

Mrs. Baylor

abaylor@fcps.edu

"Statistical thinking will one day be as necessary for effective citizenship as the ability to read and write." HG Wells

SUMMER ASSIGNMENT (4 parts)

- ☑ Part 1: Get the necessary materials!
- ☑ Part 2: Become excited about the study of statistics by viewing some videos and writing about them. (10 points)
- ☑ Part 3: Begin to think about some key statistical concepts by reading and writing about articles from the Washington Post. (5 points)
- ☑ Part 4: Learn important vocabulary (5 points).
- ☑ Part 5: Review statistical concepts that you learned in previous math classes and complete practice problems (10 points).

Quiz on Part 4 and 5 on third class period.

THE SCHS HONOR CODE APPLIES TO THIS ASSIGNMENT: DO YOUR OWN WORK. DO NOT COPY ANSWERS FROM YOUR CLASSMATES.

So let's begin ☺

Explaining in complete sentences is required on this assignment and throughout the course. You cannot just write down numbers and be done; you must use numbers in context – what they mean to that particular problem using appropriate units.

Part 1: Get Materials for Class

- **YOU MUST HAVE YOUR OWN GRAPHING CALCULATOR AND BRING IT TO CLASS EVERYDAY!!** We will begin using it on the first day of school. A TI-83 is the minimum calculator needed for this course. TI-84 or TI-84 + is better. TI-89s are allowed but have different menus and will NOT be demonstrated in class. If you choose to use the TI-89, you will be responsible to learn where to locate the functions we use in class.
- A 3 ring binder to keep class notes and graded work. Dividers are recommended for each chapter. You can easily fill a 2 inch binder this year!
- Paper and pencils for homework, of course.
- Recommended (not required): Purchase a copy of an AP review book. I recommend either **5 Steps to a 5 AP Statistics or Princeton Review AP Statistics.**

Part 2: Learn to love statistics.

Watch the following videos/read the article and write a summary with minimum 1 page of what you learned about statistics that you did not know before. Include examples from the videos in your write up. All videos must be referenced. The summary needs to be typed (double space). **Attach to part 4/5 to turn in.**

- http://www.ted.com/talks/lang/eng/arthur_benjamin_s_formula_for_changing_math_education.html
- http://www.gapminder.org/videos/the-joy-of-stats/
- http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html
- http://www.wired.com/magazine/2010/04/st_thompson_statistics/

Part 3: Reading and Writing.

Unlike your previous math classes, you will be expected to 1) READ (and understand) your AP Statistics textbook and 2) interpret your answer to problems by WRITING a sentences. These skills of reading for understanding and writing clear and concise sentences are skills you will need to master to be successful in your AP Statistics class and on the exam. The answers need to be typed. **Attach to part 4/5 to turn in.**

Read the two articles at the end of this packet in **Appendix 1** ("Overstating Aspirin's Role in Breast Cancer Prevention" and "Research Basics: Interpreting Change") from the Washington Post and then answer the following questions in complete sentences.

1. What was the story that the newspapers wrote after the research was published by the Journal of the American Medical Association?

2. What other information needed to be added to the story so that people could make decisions for themselves about the use of aspirin to prevent breast cancer?

3. How was the data collected to perform this study?

4. What type of study was performed?

5. Can this type of study be used to prove the aspirin prevents breast cancer?

6. What type of study must be done in order to 'prove' something?

7. What is the difference between 'cause' and 'association'?

8. You may have heard the statement "you can prove anything with statistics". Using what you have learned reading this article, explain what you think is meant by this statement.

Part 4: Learn important vocabulary and review the statistics you learned in previous math classes.

For help doing this part, use Appendix 2 and 3 at the end of the packet. They contain information on how to calculate numeric statistics, make graphs and calculator help. I also recommend the following websites for more help.

- http://www.stattrek.com/
- http://apstatsguy.com/
- http://www.khanacademy.org/math/statistics

"Super Duper Important Statistical Vocabulary"
Knowing these will be very important to your success in AP Statistics.
Start learning NOW!

Define each of the following. Type answers and attach to packet to turn in. Be sure to study these definitions as you have a QUIZ on September 8th/9th.

1. Individuals
2. Variable
3. Categorical variable (qualitative variable)
4. Quantitative variable
5. Distribution
6. Population
7. Sample
8. Parameter
9. Statistic
10. Continuous data
11. Discrete data
12. Bivariate data (two variables)
13. Univariate data (one- variable)
14. Resistant measure
15. 5 number summary of data
16. Measures of spread
17. measures of center
18. Standard Deviation

The following are important symbols you must know.

μ	Population mean
$\bar{x}$	Sample mean
p	Population proportion
$\hat{p}$	Sample proportion
σ	Standard deviation of the population
s	Standard deviation of the sample
n	Number of observations

r	Correlation coefficient
$\hat{y}$	Estimated value of y
b	Slope of regression
a	y-intercept of regression
P(A)	Probability of event A
P(A $\cap$ B)	Probability of events A and B
P (A | B)	Probability of A given that B has already occurred

IMPORTANT COMPARISONS –VOCABULARY PRACTICE

Categorical vs. Quantitative Data

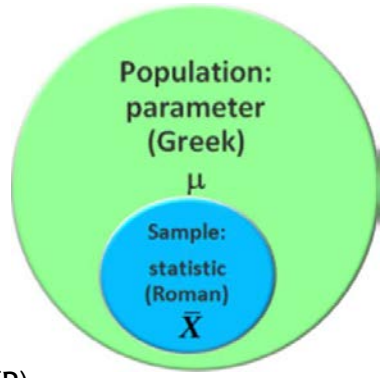
Categorical (or Qualitative) Variable: takes on values that are names and descriptions. Ex. Color
Quantitative (or Numeric) Variable: takes on numerical values, measureable quantities. Ex. Weight

Determine if the variables listed below are *quantitative* or *categorical*. Neatly print "Q" for quantitative and "C" for categorical.

_____ 1. Time it takes to get to school
_____ 2. Number of shoes owned
_____ 3. Hair color
_____ 4. Temperature of a cup of coffee
_____ 5. Teacher salaries
_____ 6. Gender
_____ 7. Facebook user
_____ 8. Height
_____ 9. Amount of oil spilled
_____ 10. Age of Oscar winners
_____ 11. Type of pain medication
_____ 12. Jellybean flavors
_____ 13. Country of origin
_____ 14. Type of meat

Population vs. Sample

Population: The entire group of individuals intended to be studied
Ex. Every individual living in Fairfax County
Sample: Part of a population that is examined in order to gather information
Ex. 200 individuals living in Fairfax County



Statistic vs. Parameter

Parameter: Number that describes a population (ex. Mean, proportion, max …)
Statistic: Number that describes a sample (ex. Mean, proportion, max …)

Identify each as Population (Pop), Sample (Sam), Statistic (S), or Parameter (P):

_____ 1. All dogs in Fairfax county
_____ 2. Students in one classroom of the school
_____ 3. True mean height of everyone living in California
_____ 4. All students in a school
_____ 5. Mean height of students in a class
_____ 6. Proportion of days with temperature above 50 in a given month
_____ 7. 50 dogs in a city
_____ 8. Proportion of test grades above 75 for 30 students in a class
_____ 9. Mean weight of cereal boxes at my house
_____ 10. Mean amount of liquid in 100 selected bottle of a certain juice
_____ 11. 80 families in a county
_____ 12. True mean number of family members in Montana
_____ 13. All shirts at Dulles Mall
_____ 14. True proportion of students wearing glasses in school

UNIVARIATE DISTRIBUTIONS – DESCRIBING, NUMERICAL SUMMARIES

★**Distribution**- of a variable tells us what values the variable takes and how often it takes these values

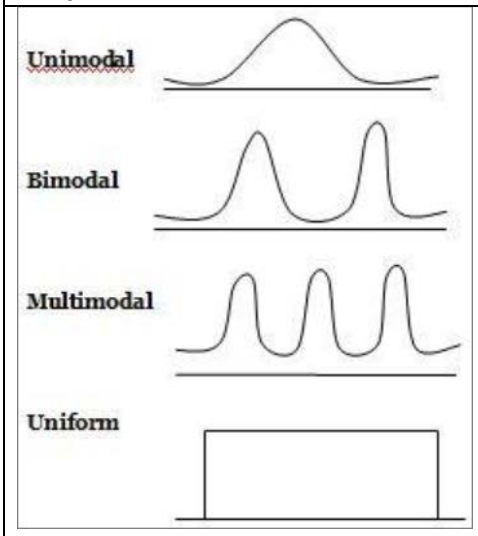
★**Univariate data (one- variable)** – data that describes a single characteristic of a population

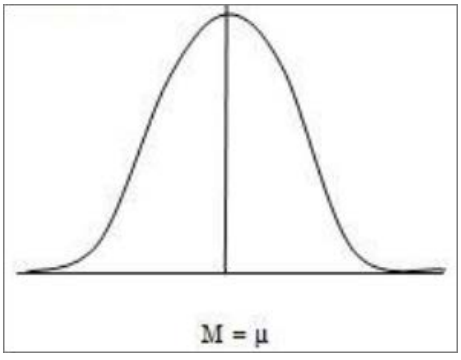
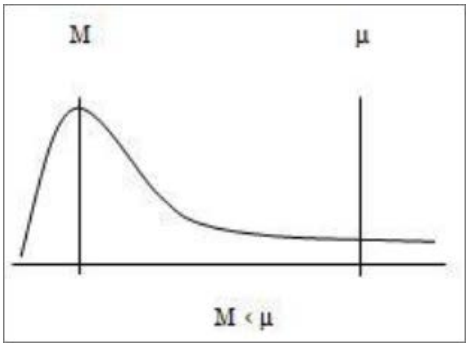
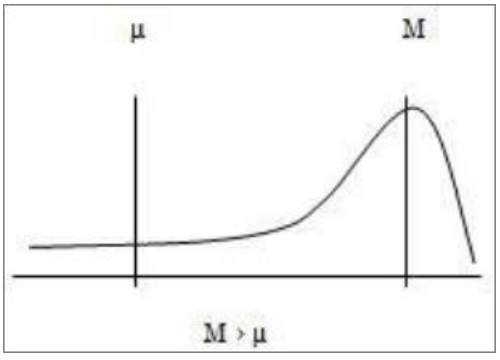
★**Resistant measure-** a measure that is not sensitive to extreme values

How do we describe univariate distributions? There are 4 characteristics to look for:

1. **Shape** – What form does the distribution take?
2. **Center** – Where is it centered? Which measure of center should you use?
3. **Spread** – How dispersed (spread out) is the data? Which measure of spread should you use?
4. **Unusual features** – identify any extreme values (outliers) and gaps or cluster (clumps) in the distribution

When you are asked to <u>describe a distribution</u>, be sure to identify the characteristics listed above, use sentences and write in context

Shape	**Measures of Center**	**Measures of Spread**
	★Median (M) – Resistant to outliers – Use if distribution is skewed – Calculate by placing observations in numerical order and find middle observation ★Mean (μ) – NOT resistant to outliers – Use if distribution is symmetric – Calculate by taking the average $(\Sigma x_i)/n$ x-bar ($\bar{x}$) is the SAMPLE mean (use with sample standard deviation, s) μ is the PARAMETER (population) mean (use with population standard deviation, σ)	★Interquartile Range (IQR) = Q3 – Q1 Q1 is the median of 1st half Q3 is the median of 2nd half IQR always goes with the median ★Standard deviation (s) – measures the spread or dispersion about the mean and should only be used when the mean is the chosen measure of center ★Variance is just the standard deviation squared s^2

	Skewness (part of SHAPE)	
★Symmetric	**★Right Skew**	**★Left Skew**
		
Median = mean	**Median < mean** **"Skewed right, the mean is MIGHT"**	**Median > mean** **"Skewed left, the mean is LESS"**

Example Problem (use Appendix 2 and 3 for help)

All answers are included right after the question, but try to see if you can get those answers yourself!

New York Yankee Roger Maris held the single-season home run record from 1961 until 1998. Here are Maris' home run counts for his 15 years in the American League:

14 28 16 39 61 33 23 26 8 13 15 14 19 10 16

Enter data into a list STAT → EDIT
STAT → CALC → 1-Var Stats
Information for both the 2- and 5-number summary is given!

1. a. What is the Mean?

Answer: 22.333

b. Median?

Answer: 16

c. Standard deviation? (Use 1-Var Stats in calculator)

Answer: 13.788

d. Variance?

Answer: 190.109

e. IQR?

Answer: 28 – 14 = 14

2. Looking at the box plot and histogram for this set of data,

a. Would it be more appropriate to use the mean or the median as a measure of center? Explain.

Answer: It would be more appropriate to use the median as the measure of center in this case. According to the box plot, this data set is skewed right, so we should use a measure of center that is more resistant. We don't want the value of the center to fluctuate significantly when there are extreme data points in the data set!

b. Would it be more appropriate to use the IQR or the standard deviation as a measure of spread? Explain.

Answer: It would be more appropriate to use the IQR as the measure of center in this case. We ALWAYS use the median and IQR in conjunction with each other. Since we have already established the use of the median as the measure of center, we should, without question, use the IQR as the measure of spread.

PRACTICE PROBLEMS NAME: _____

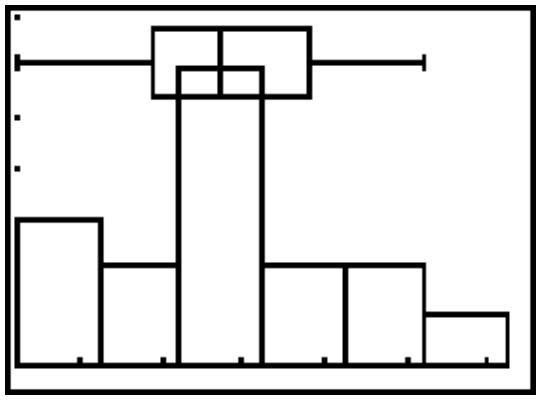
NUMERICAL SUMMARIES

Quantitative data has many different numerical summaries that can be calculated. Determine the following from the data below on the number of homeruns Mark McGuire has hit in each season from 1982 – 2001.

70	52	22	49	3	32	58	39
39	65	42	29	9	32	9	33

1. You should be able to calculate all these by hand (if you need help see Appendix 2) except standard deviation (from calculator – see Appendix 3 for help).

Mean	
Minimum	
Maximum	
Median	
Q1	
Q3	
Range	
IQR	
Standard deviation	
Variance	



2. Looking at the box plot and histogram for this set of data,

a. Would it be more appropriate to use the mean or the median as a measure of center? Explain.

b. Would it be more appropriate to use the IQR or the standard deviation as a measure of spread? Explain.

PRACTICE GRAPHING DISTRIBUTIONS (NUMERICAL VARIABLES)

See Appendix 2 if you need instuctions for these graphs.

★A **dotplot** is a type of graphic display used to compare frequency counts within categories or groups. As you might guess, a dotplot is made up of dots plotted on a graph. Here is how to interpret a dotplot. Each dot can represent a single observation from a set of data, or a specified number of observations from a set of data.

WEATHER!

3. The data below gives the number of hurricanes that happened each year from 1944 through 2000 as reported by *Science* magazine.

3	2	1	4	3	7	2	3	3	2	5	2	2	4	2	2	6	0	2	5	1	3	1	0
3	2	1	0	1	2	3	2	1	2	2	2	3	1	1	1	3	0	1	3	2	1	2	1
1	0	5	6	1	3	5	3																

a. Make a dotplot to display these data. Make sure you include appropriate labels, title, and scale.

b. Describe this distribution (see page 6 for what to include). Write in sentences and use context.

c. Find the 2 or 5 number summary, whichever is appropriate for this distribution. Hint: What is the shape? You may use the calculator (see Appendix 3)

★A **stemplot** is used to display quantitative data, generally from small data sets (50 or fewer observations).

SHOPPING SPREE!

4. A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (round to the nearest dollar), arranged in increasing order:

3	9	9	11	13	14	15	16	17	17
18	18	19	20	20	20	21	22	23	24
25	25	26	26	28	28	28	28	32	35
36	39	39	41	43	44	45	45	47	49
50	53	55	59	61	70	83	86	86	93

a. Make a stemplot using tens of dollars as the stem and dollars as the leaves. Make sure you include appropriate labels, title and key.

KEY

b. Describe this distribution (see page 6 for what to include). Write in sentences and use context.

★**Histograms are a way to display groups of quantitative data into bins (the bars). These bins have the same width and scale and are touching because the number line is continuous. To make a histogram you must first decide on an appropriate bin width and count how many observations are in each bin.**

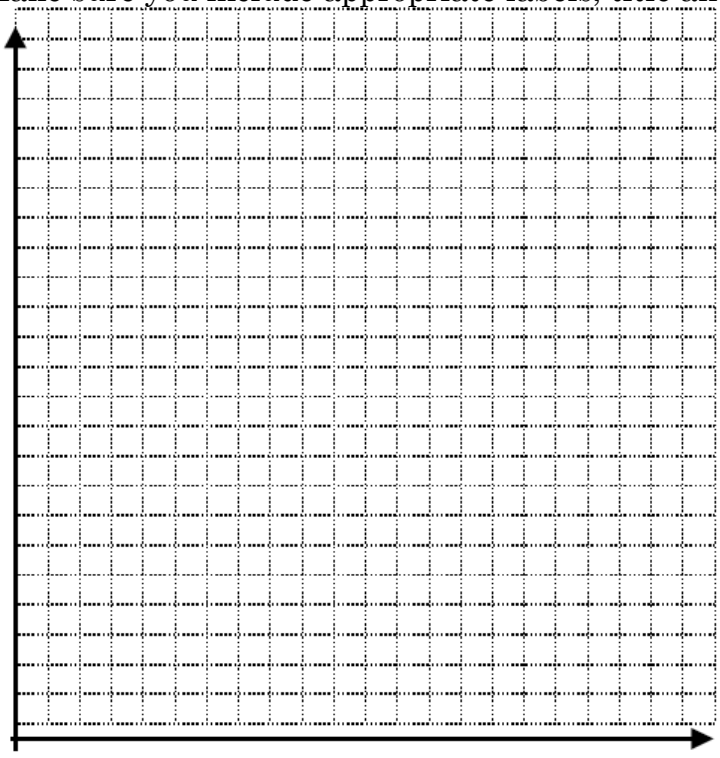
WHERE DO OLDER FOLKS LIVE?

5. This table gives the percentage of residents aged 65 of older in each of the 50 states.

State	Percent	State	Percent	State	Percent
Alabama	13.1	Louisiana	11.5	Ohio	13.4
Alaska	5.5	Maine	14.1	Oklahoma	13.4
Arizona	13.2	Maryland	11.5	Oregon	13.2
Arkansas	14.3	Massachusetts	14.0	Pennsylvania	15.9
California	11.1	Michigan	12.5	Rhode Island	15.6
Colorado	10.1	Minnesota	12.3	South Carolina	12.2
Connecticut	14.3	Mississippi	12.2	South Dakota	14.3
Delaware	13.0	Missouri	13.7	Tennessee	12.5
Florida	18.3	Montana	13.3	Texas	10.1
Georgia	9.9	Nebraska	13.8	Utah	8.8
Hawaii	13.3	Nevada	11.5	Vermont	12.3
Idaho	11.3	New Hampshire	12.0	Virginia	11.3
Illinois	12.4	New Jersey	13.6	Washington	11.5
Indiana	12.5	New Mexico	11.4	West Virginia	15.2
Iowa	15.1	New York	13.3	Wisconsin	13.2
Kansas	13.5	North Carolina	12.5	Wyoming	11.5
Kentucky	12.5	North Dakota	14.4		

a. Finish the chart of Bin widths (started for you below) and then create a histogram using those bins on the grid below. Make sure you include appropriate labels, title and scale.

Bin Widths	Frequency
4 to < 6	1
6 to < 8	
8 to < 10	



b. Describe this distribution (see page 6 for what to include). Write in sentences and use context.

★**Boxplot,** sometimes called a box and whisker plot, is a type of graph used to display patterns of quantitative data. A boxplot splits the data set into quartiles. The body of the boxplot consists of a "box" (hence, the name), which goes from the first quartile (Q1) to the third quartile (Q3).

★**5 number summary of data** - minimum, maximum, Q1, Q3, and median (use when skewed and outlier resistant)

SSHA Scores

6. Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for 18 first-year college women:

154 109 137 115 152 140 154 178 101 103 126 126 137 165 165 129 200 148

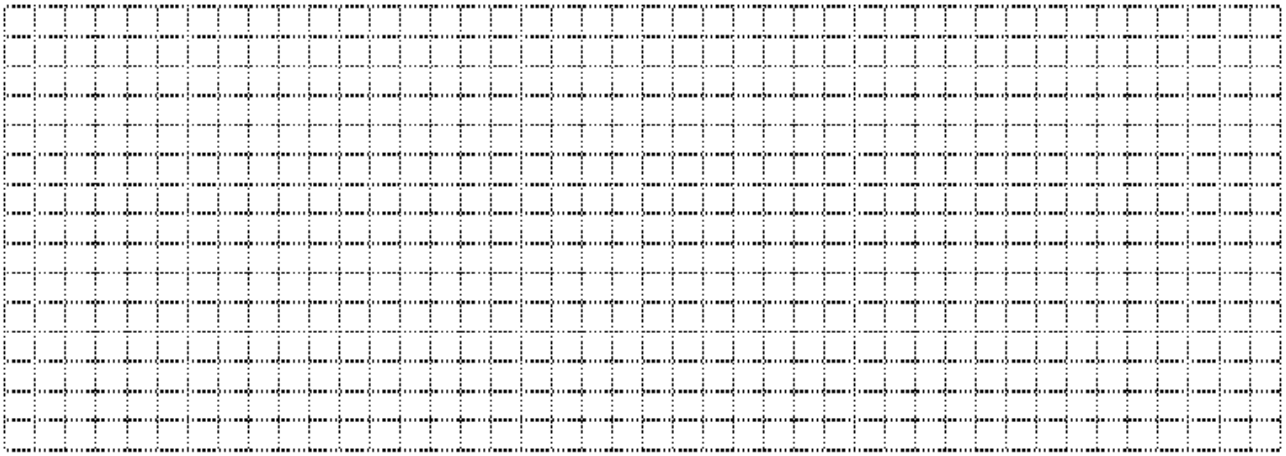
and for 20 first-year college men:

108 140 114 91 180 115 126 92 169 146 109 132 75 88 113 151 70 115 187 104

a. Put the data values in order for each gender. Compute numeral summaries for each gender. (Need help? See Appendix 2)

Women			Men	
Mean			Mean	
Minimum			Minimum	
Q1			Q1	
Median			Median	
Q3			Q3	
Maximum			Maximum	
Range			Range	
IQR			IQR	

b. **Using the minimum, Q1, Median, Q3, and Maximum from each gender, make parallel boxplots to compare the distributions.** Be sure to include appropriate labels, title and scale.



c. Comparison the scores for males and females. Disscuss the same topics as describing a distribution with the addition of comparison words.

PRACTICE GRAPHING DISTRIBUTIONS (CATEGORICAL VARIABLES)

ACCIDENTAL DEATHS

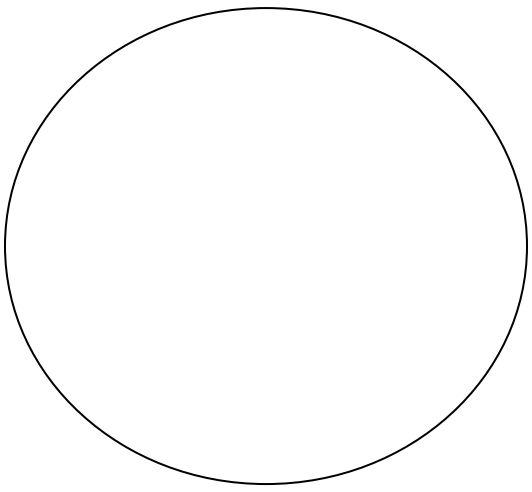
7. In 1997 there were 92,353 deaths from accidents in the United States. Among these were 42,340 deaths from motor vehicle accidents, 11,858 from falls, 10,163 from poisoning, 4051 from drowning, and 3601 from fires. The rest were listed as "other" causes.

a. Find the percent of accidental deaths from each of these causes, rounded to the nearest percent.

b. What percent of accidental deaths were from "other" causes?

c. NEATLY create a well-labeled **bar graph** of the distribution of causes of accidental deaths. Be sure to include an "other causes" bar. Label axes, scale and title.

d. A pie chart is another graphical display used to show all the categories in a categorical variable relative to each other. By hand, create a pie chart for the accidental death percentages. Label appropriately.



PRACTICE NORMAL DISTRIBUTIONS (FROM ALGEBRA 2)

(See appendix #2 for assistance)

8. Suppose your Statistics professor reports test grades as z-scores, and you got a score of 2.20 on an exam. Write a sentence explaining what that means.

9. One of the authors has an adopted grandson whose birth family members are very short. After examining him at his 2-year check-up, the boy's pediatrician said that the z-score for his height relative to American 2-year olds was -1.88. Write a sentence explaining what that means.

10. The mean score on the Statistics exam was 75 points with a standard deviation of 5 points, and Gregor's z-score was -2. How many points did he score?

11. A town's January high temperatures average 36^0 F with a standard deviation of 10^o, while in July the mean high temperature is 74^0 and standard deviation is 8^0. In which month is it more unusual to have a day with a high temperature of 55^0? Explain.
Hint: calculate the z-scores!

12. Environmental Protection Agency fuel economy estimates for automobile models tested recently predicted a mean of 24.8 mpg and a standard deviation of 6.2 mpg for highway driving. Assume that a Normal model can be applied.

    a. Draw the normal model for auto fuel economy. Clearly label it, showing what the 68-95-99.7 rule predicts.

    b. In what interval would you expect the central 68% of the autos to be found?

    c. About want percent of cars should get more than 31 mpg?

    d. About what percent of cars should get between 31 and 37.2 mpg?

    e. Describe the gas mileage of the worst 2.5% of all cars.

    f. What mpg would you consider unusually high? Explain.

SUMMER ASSIGNMENT CHECK-LIST

Have you completed the following....

- ☑ Part 1: Get the necessary materials!
- ☑ Part 2: Writing Assignment about Videos (15 points)
- ☑ Part 3: Writing Assignment about Article (15 points)
- ☑ Part 4: Vocabulary definitions (10 points)
- ☑ Part 5: Complete the practice problems (pages 5 and 8 – 14)

Estimate how long did it take you to complete this summer assignment? Thank you!

_____hours

Bring completed assignment on the THIRD day of class!!

I cannot wait to meet you and begin this awesome course!

Mrs. Baylor

Appendix 1: Articles for Part 3

The Washington Post

Overstating Aspirin's Role in Breast Cancer Prevention

How Medical Research Was Misinterpreted to Suggest Scientists Know More Than They Do

By Lisa M. Schwartz, Steven Woloshin and H. Gilbert Welch
Special to The Washington Post
Tuesday, May 10, 2005

Medical research often becomes news. But sometimes the news is made to appear more definitive and dramatic than the research warrants. This series dissects health news to highlight some common study interpretation problems we see as physician researchers and show how the research community, medical journals and the media can do better.

Preventing breast cancer is arguably one of the most important priorities for women's health. So when the Journal of the American Medical Association published research a year ago suggesting that aspirin might lower breast cancer risk, it was understandably big news. The story received extensive coverage in top U.S. newspapers, including The Washington Post, the Wall Street Journal, the New York Times and USA Today, and the major television networks. The headlines were compelling: "Aspirin May Avert Breast Cancer" (The Post), "Aspirin Is Seen as Preventing Breast Tumors" (the Times).

In each story, the media highlighted the change in risk associated with aspirin -- noting prominently something to the effect that aspirin users had a "20 percent lower risk" compared with nonusers. The implied message in many of the stories was that women should consider taking aspirin to avoid breast cancer.

But the media message probably misled readers about both the size and certainty of the benefit of aspirin in preventing breast cancer. That's because the reporting left key questions unanswered:

· Just how big is the potential benefit of aspirin?

· Is it big enough to outweigh the known harms?

· Does aspirin really prevent breast cancer, or is there some other difference between women who take aspirin regularly and those who don't that could account for the difference in cancer rates?

This article offers a look at how the message got distorted, what the findings really signify--and some broader lessons about interpreting medical research.

How Big a Benefit?

Just how big is the potential benefit of aspirin?

The 20 percent reduction in risk certainly sounds impressive. But to really understand what this statistic means, you need to ask, "20 percent lower than what?" In other words, you need to know the chance of breast cancer for people who do not use aspirin. Unfortunately, this information did not appear in any of the media reports. While it might be tempting to fault journalists for sloppy, incomplete reporting, it is hard to blame them when the information was missing from the journal article itself.

In the study, Columbia University researchers asked approximately 3,000 women with and without breast cancer about their use of aspirin in the past. The typical woman in this study was between the ages of 55 and 64. According to the National Cancer Institute, about 20 out of 1,000 women in this age group will develop breast cancer in the next five years. Therefore, the "20 percent lower chance" would translate into a change in risk from 20 per 1,000 women to 16 per 1,000 -- or four fewer breast cancers per 1,000 women over five years.

For people who prefer to look at percentages, this translates as meaning that 2 percent develop breast cancer without aspirin, while 1.6 percent develop it with aspirin, for an absolute risk reduction of 0.4 percent over five years.

Another way to present these results would be to say that a woman's chance of being free from breast cancer over the next five years was 98.4 percent if she used aspirin and 98 percent if she did not. Seeing the actual risks leaves a very different impression than a statement like "aspirin lowers breast cancer risk by 20 percent." (See "Research Basics: How Big Is the Difference?")

Against What Size Harms?

Is the potential benefit of aspirin big enough to outweigh its known harms?

Unfortunately, aspirin, like most drugs, can have side effects. These, according to the U.S. Preventive Services Task Force, include a small risk of serious (and possibly fatal) bleeding in the stomach or intestine, or strokes from bleeding in the brain -- harms briefly noted but not quantified in the original study or in most media reports. To decide whether aspirin is worth taking, women need to know how the potential size of aspirin's benefit in reducing breast cancer compares with the drug's potential harms.

Sound medical practice dictates doing the same kind of calculation -- of potential benefits against potential harms -- anytime you consider taking a drug.

We provide the relevant information in the "Aspirin Study Facts," below. The first column shows the health outcome being considered (e.g., getting breast cancer, having a major bleeding event). The second column shows the chance of the outcome over five years for women *not* taking aspirin. The third column shows the corresponding chance for women taking aspirin. And the fourth column shows the difference -- the possible effect of aspirin.

As the table shows, the size of the known risk for stomach bleeding to a woman taking aspirin daily nearly matches the size of the still-hypothetical benefit in terms of breast cancer protection. That kind of comparison might lead some women to conclude that the tradeoff doesn't warrant the risk.

While it may take you some time to become familiar with this table, we think this sort of presentation would be helpful in many situations; for example, whenever people are deciding about taking a new medication or undergoing elective surgery.

Is It Really Aspirin?

Does aspirin really prevent breast cancer, or is there some other difference between women in the study that could account for the difference in cancer rates?

Can we be sure that aspirin was responsible for the "20 percent fewer" breast cancers that the Columbia researchers found among aspirin users compared with nonusers?

To understand why not, it is necessary to know some of the details about how the study was conducted.

The researchers collected information from all of the women in New York's Nassau and Suffolk counties on Long Island, who were diagnosed with breast cancer in 1996 and 1997. For comparison, they matched these women with others who did not have breast cancer, but who were about the same age and from the same counties. The researchers asked all the women about their use of aspirin.

They found that aspirin use was more common among the women without breast cancer. While the researchers were careful to report that the use of aspirin was "associated" with reduced risk of breast cancer, the media used stronger language, suggesting aspirin played a role in preventing breast tumors.

Unfortunately, this kind of study -- an observational study -- cannot prove that it was the aspirin that lowered breast cancer risk. Strictly speaking, the researchers demonstrated only that there is an association between aspirin and breast cancer.

Consider how an association between aspirin and breast cancer could exist even if aspirin has no effect on breast cancer.

It could be that women who use aspirin regularly are already at a lower risk of breast cancer. Imagine, for example, there was a gene that protected against breast cancer but also made people more susceptible to pain. Women who carried this gene would be more apt to use aspirin for pain relief. The lower breast cancer risk in aspirin users might simply reflect the fact that they had this gene. In other words, aspirin might have nothing to do with the findings. To really know if aspirin lowers breast cancer risk would require a different kind of study -- a randomized trial. (See "Research Basics: Cause or Association?")

Nonetheless, observational studies are important (and often crucial) in building the case for doing a randomized trial. In this instance, the researchers had a theory for how aspirin might prevent breast cancers. They predicted that it would only be true for certain kinds of cancers (so-called hormone receptor positive cancers, the most dangerous kind, which account for about 60 percent of all breast cancers). And that is just what they observed: The association between aspirin and breast cancer was not seen in hormone receptor negative cancers. That the researchers' prediction was correct supports (but does not prove) the idea that aspirin reduces risk. The next logical step would be a randomized trial.

The difference between "cause" and "association" may seem subtle, but it is actually profound. Even so, people -- like the headline writers in this case -- often go beyond the evidence at hand and assume that an association is causal. Readers should know that many associations do not reflect cause and effect.

The Bottom Line

In a large observational study, researchers found slightly fewer breast cancers among women who took aspirin regularly compared with women who did not. Because aspirin's benefit in reducing breast cancer (assuming it can be proven) was small, it may not outweigh the drug's known harms. While it is possible that aspirin itself reduces the risk of breast cancer, we cannot be sure from this study. It would take a randomized trial to be certain. Fortunately, one has just been completed by researchers at Harvard Medical School, and the results are expected in the very near future. Until then, it is too soon to recommend taking aspirin to prevent breast cancer. ·

Lisa Schwartz, Steven Woloshin and Gilbert Welch are physician researchers in the VA Outcomes Group in White River Junction, Vt., and faculty members at the Dartmouth Medical School. They conduct regular seminars on how to interpret medical studies. (Seehttp://www.vaoutcomes.org.) The views expressed do not necessarily represent the views of the Department of Veterans Affairs or the United States Government.

Research Basics: Interpreting Change

Tuesday, May 10, 2005

How Big Is the Difference?

Many medical studies end up concluding that two groups have different health outcomes -- death rates, heart attack rates, cholesterol levels and so forth. This difference is typically expressed as a *relative change* , as in the statement: "The treatment group had 50 percent fewer cases of eye cancer than the control group." The problem with this comparison is that it provides no information about how common eye cancer is in either group.

Thinking about relative changes in risk is like deciding when to use a coupon at a store. Imagine you have a coupon that says "50 percent off any one purchase." You go to the store to buy a pack of gum for 50 cents and a large Thanksgiving turkey for $35. Will you use the coupon for the gum or the turkey? Most people would use it for the turkey.

Why? Because paring half the price off $35 reaps a bigger savings --$17.50 --than cutting half off 50 cents -- or $0.25.

The analogy in health is that "50 percent fewer cases" is a very different number when applied to eye cancer -- a rare problem accounting for about 2,000 new cases in the U.S. each year -- than when applied to heart attacks -- a common problem accounting for about 800,000 new cases annually.

To really understand how big a difference is, you need to find out the *starting* and *ending points* -- sometimes called " *absolute risks* ." In the coupon example, the start and end points are the regular and the sales price. In a study about medical treatment, the start and end points are the chances of something happening in the untreated and treated groups.

Presenting the starting and ending point requires a few more words than presenting relative changes. For example, "In a year, two of 100,000 untreated people developed eye cancer; in contrast, one of 100,000 treated people developed eye cancer." For the price of a few more words you gain perspective: The chance of developing eye cancer is small.

Cause or Association?

Many important insights into human health come from *observational studies* -- studies in which the researcher simply records what happens to people in different situations, without intervening. Such studies first linked cigarette smoking to lung cancer and high cholesterol to heart disease. But not all observed associations represent cause and effect. And problems can occur when this key point is overlooked.

An example may help make the distinction clear. A man thought his rooster made the sun rise. Why? Because each morning when he woke up while it was still dark, he would hear his rooster crow as the sun rose. He confused association with causation until the day his rooster died, when the sun rose without any help.

A more serious example involves the long-held belief that most women should take estrogen after menopause. That idea, only recently discredited, also came from observational studies. The observation -- shown in more than 40 studies involving hundreds of thousands women -- was that women who took estrogen supplements also had less heart disease. But it turned out that estrogen was not the reason why this was the case. Instead, women taking estrogen tended to be healthier and wealthier. Their health and wealth -- not their estrogen supplements -- were responsible for the lower risk of heart disease.

The only way to reliably distinguish a cause from an association is to conduct a true experiment -- a *randomized trial* . In this type of study, patients are assigned randomly --that is, by chance--to receive a therapy or not receive it. This study design is the best way to construct two groups that are similar in every way except one -- whether they get the therapy being studied. That means any differences observed afterward must be caused by the therapy. In the case of estrogen and heart disease, such a study showed that the long-held beliefs were wrong.

Unfortunately, it is not always possible to do a randomized trial. For example, it is extremely unlikely that we could get people to agree to be randomly assigned to either eating only fast food or only organic food every day for a year (and that they would actually adhere to the diet if they did agree to be randomized). In such cases, scientists have to rely on observational studies. But when new tests or treatments are proposed, randomized trials ought to be conducted prior to their widespread use. Doctors prescribed estrogen to millions of women for many years until the randomized trial showed that intuition and dozens of observational studies were wrong.

-- Lisa M. Schwartz, Steven Woloshin and H. Gilbert Welch

Appendix 2: "Quick Reference" of Statistical Basics

Numerical Descriptions of Quantitative Data

Measures of Center

Mean: The sum of all the data values divided by the number (n) of data values.

Example

Data: 4, 36, 10, 22, 9 Mean = $\bar{x} = \sum \frac{x_i}{n} = \frac{4+36+10+22+9}{5} = \frac{81}{5} = 16.2$

Median: The middle element of an ordered set of data.

Examples

Data: 4, 36, 10, 22, 9 = 4 9 **10** 22 36 ⟶ Median = 10

Data: 4, 36, 10, 22, 9, 43 = 4 9 10 | 22 36 43 ⟶ Median = $\frac{10+22}{2} = 16$

Measures of Spread:

Range: Maximum value – Minimum value

Example

Data: 4, 36, 10, 22, 9 = 4 9 10 22 36

Range = Max. – Min. = 36 – 4 = **32**

Interquartile Range (IQR): The difference between the 75th percentile (Q_3) and the 25th percentile (Q_1). This is $\mathbf{Q_3 - Q_1}$. Q_1 is the median of the lower half of the data and Q_3 is the median of the upper half. In neither case is the median of the data included in these calculations.

The IQR contains 50% of the data. Each quartile contains 25% of the data.

Examples

1. Data: 4, 36, 10, 22, 9 = 4 9 **10** 22 36

$Q_1 = 6.5$ $Q_3 = 29$

So, the IQR = 29 – 6.5 = 22.5

2. Data: 4 **9** 10 | 22 **36** 43

Q_1 Q_3

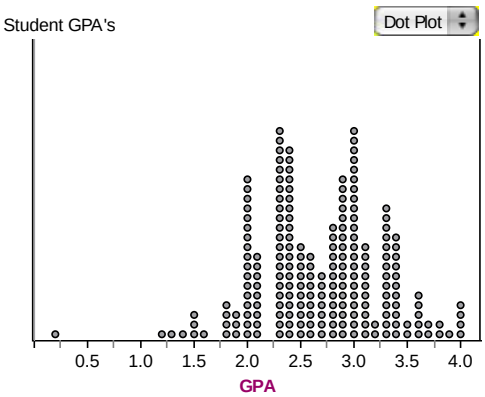
So, the IQR = 36 – 9 = 27

Five-number summary: consists of Minimum, Q_1, Median, Q_3, and Maximum. To find these statistics, enter the data you have into your calculator using the list function:
STAT → ENTER → type the data into L_1. If you make a mistake, you can go to the error and **DELETE**. If you forget an item, you can go to the line below where it is supposed to be and press **2nd DEL** to insert it. To find the each value of the five-number summary, go to 2nd STAT → MATH → 5 and then type in L_1 by typing 2nd → 1

NOTE: If the lists you are using already have numbers in them before you start, you can clear them this way: Arrow up (↑) to the line where L_1 is shown. Press CLEAR, then the down arrow (↓).

Graphical Displays of Univariate (one variable) Data

• Dotplot • Boxplot (Box and Whiskers) • Stemplot (Stem and Leaf) • Histogram



To make a Dotplot:

1. Draw and label a number line so that all the values in your dataset will fit.
2. Graph each of the data values with a dot. Be sure to line the dots up vertically as well as horizontally so that you can really see the shape of the graph.

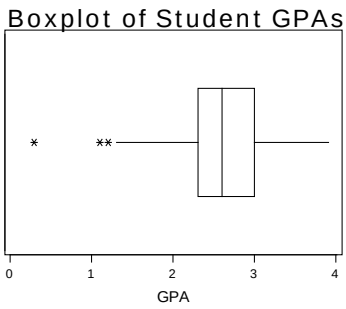
Stemplot of Student GPAs

1	23
1	444
1	67
1	88888999
2	00000000000000000111111111
2	333333333333333333333
2	444444444444444445555555555
2	66666666666677777
2	88888888889999999999999
3	00000000000000000000111111111
3	2233333333333333
3	44444444455
3	6666677
3	889

Key: 3 | 4 = 3.4

TO MAKE A STEMPLOT:

1. Put the data in ascending order. Make a key!
2. Use only the last digit of the number as a leaf (see the numbers to the right of the line –each digit is the last digit of a larger number).
3. Use one, two, or more digits as the stem. (Sometimes, you can truncate data when there are too many digits in each data value – i.e. the number 20, 578 would become 20 | 5, where the "20" is in thousands. Note that this is **different** from rounding.)
4. Place the "stem" digit(s) to the left of the line and the leaf digit to the right of the line. Do this for each data value. You should then arrange the "leaves" in ascending order.
5. Sometimes, there are many numbers with the same "stem." In this situation it might be useful to break the numbers with the same stem into either two distinct groups (each on a separate line; say, "leaves" from 0 – 4 on the first line and 5 – 9 on the second.) or into five distinct groups as is shown in the graph to the right. Here, the first line for each stem contains all the 0 – 1 leaves, the next line contains the 2 – 3 leaves and so on. This technique is called "splitting the stems." It is useful in some cases in order to show the shape of the data more clearly.

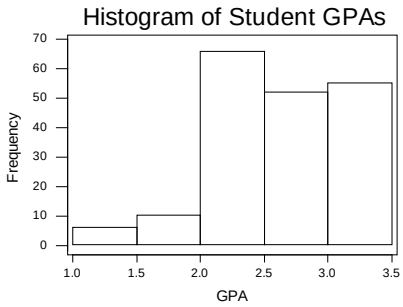


To make a Boxplot:

1. **Draw and label a number line** that includes the minimum and the maximum values for the set of data.
2. Calculate the five-number summary and make a dot for each of these summary numbers above the number line.
3. Draw a line between the 1st and 2nd dot, showing the "lower quartile"; and then draw a line from the 4th to the 5th dot to show the "upper quartile." These are commonly called the "whiskers."
4. Draw a rectangular box from the 2nd to the 4th dot and draw a line through the box on the middle dot – the median.

NOTE: In AP Statistics, a "modified boxplot" is used. This shows any "outliers." An outlier is a data point that does not fit the pattern of the rest of the data. When your calculator or computer software graphs a modified boxplot, an algorithm is used to determine what it takes to "not fit the pattern of the rest of the data." This algorithm is:

1.5*(IQR) away from the "box" part of the graph. (above and below the box). These outliers are shown with dots or stars, or any other small symbol.



To make a histogram:

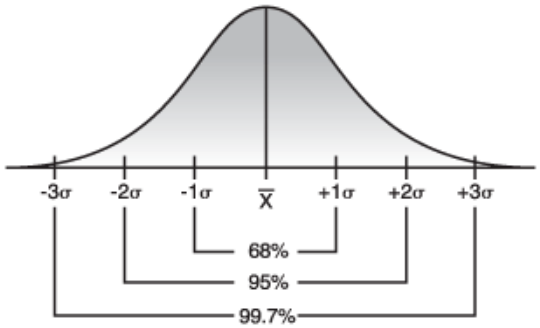
1. Put the data into ascending order.
2. Decide upon evenly spaced intervals into which to divide the set of data (such as 0, 10, 20, 30, etc.) and then count the number of values that fall within each interval. This number is called the "frequency." If you divide each of these frequencies by the size of the data set, *n*, making percents, then you have what are called "relative frequencies."
3. Draw and **label** a 1st quadrant graph using scales appropriate for the data. Be sure to include a title for the x- and for the y-axes.
4. Graph the frequencies that you calculated in step 2.

Categorical Data:

- Bar Graph
- Circle Graph (Pie Chart)

I'm assuming that you already know how to make these two types of graphs.

Review of the Normal Curve from Algebra 2



★Properties of a normal distribution curve

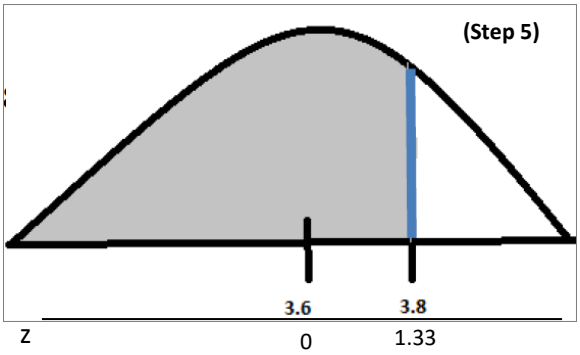
a. Mean and median are the same
b. Bell Shape (perfectly symmetrical) and follows the *empirical rule:*
   68% of everything in the population is within 1 Standard deviation
   95% of everything in the population is within 2 Standard deviations
   99.7% of everything in the population is within 3 Standard deviations
c. The total area under the curve is 1 or 100%

Notation:

For normal distributions, a short notation is helpful. We abbreviate the normal distribution with mean and standard deviation as ~N(μ,б). For example, the distribution of young women's heights is ~N(64.5,2.5). This means that the average heights of young women are 64 inches with a standard deviation of 2.5 inches.

To show the proportion or probability of a certain data falling below, above, or at a specific value can be depicted through your choice of notation. For example, P(X ≤ μ), means, "the probability that the sample value, X, is less than or equal to the population mean, μ. It is extremely important to use correct *probability notation*.

Example: The duration of a flight between 2 cities is normally distributed with a mean of 3.6 hours and a standard deviation of .15 hour. What proportions of flights will be less than 3.8 hours long?



1) NOTATION: ~N(3.6, .15)

***2)* P(X ≤ 3.8) =** ?

3) First calculate the z-score: $z = \frac{x - \mu}{\sigma}$ $z = \frac{3.8 - 3.6}{.15} = 1.33$

This means that the value 3.8 hours is 1.33 standard deviations above the mean.

4) Shaded Area = ***P(z < 1.33) .*** *You can look this up on a z-table or use your calculator. This probability is the area under the standard normal distribution*:

normalcdf(-1E99, 1.33) = .9087

6) Write a contextual statement:

The proportion of flights that will take less than 3.8 hours, when the mean flight length is 3.6 hours, is 90.9%.

Appendix 3: Calculator Help

YOUR CALCULATOR IS YOUR FRIEND!

STAT

- STAT - EDIT has 3 commands

SKILL: Enter data in a list.

2	2	3	4	4	4	6

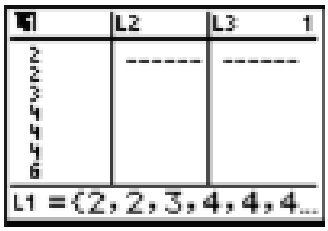
- Press Enter after each value.
- Highlight L1 to view screen at right.

- To exit to the home screen:
- Press 2nd MODE (QUIT in yellow).

- Press ENTER.

EDIT

- Accesses list editing and modifying

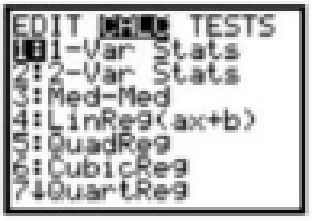


1- Var Stat is a calculator function that calculates statistics for one set of data.

This list of symbols at right will help you use this function.

Memorize it! ☺

1. Press STAT then scroll to the right to CALC
2. Select the first option: 1-Var Stats
3. The syntax is : 1-Var Stats [*(name of the list that contain the data)*]
4. Hit ENTER



Use your up and down arrow keys to see the statistics

$\bar{x}$	the mean or average
$\sum x$	the sum of all data
$\sum x^2$	the sum of the squares of the data
σ_x	the sample standard deviation
σx	the population standard deviation
n	the number of elements
$\min X$	the minimum element
Q_1	the first quartile
Med	the median of the data
Q_3	the third quartile
$\max X$	the maximum element

Calculating normal probabilities using a TI-83 and TI-84 graphing calculator

Skill: 

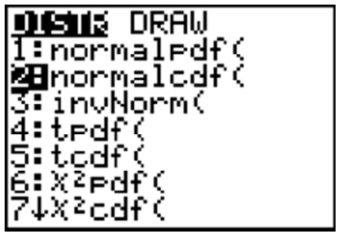
2:normalcdf(

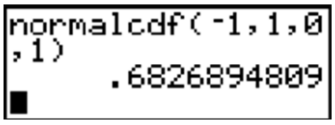
Calculate area under any normal curve

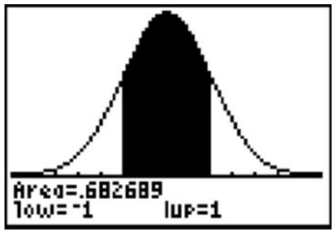
DISTR 2:normalcdf(

- The input for the command is the minimum value for area, the maximum value, the mean, μ, the standard deviation, σ.
- The keystrokes below calculate the area between −1 and 1 for a normal distribution with μ=0 and σ=1.
  - (the STANDARD normal distribution)

- Press ENTER.
- The shaded image will not appear, on the screen, but is the sketch that should accompany the solution.







Skill: 

3:invNorm(

Find a value that corresponds to an area.

DISTR 3:invNorm(

- The input for the command is the area as a decimal, the mean, μ, the standard deviation, σ.
- The syntax shown calculates the observation with an area 0.025 or 2.5% below its value.
  - First a z-score for $N(\mu=0,\sigma=1)$ the STANDARD normal distribution.