Holyoke Community College
2. STATISTICS MTH 142.07
3. Frost 143 TR 11-12:15
Fall 2018

1. Instructor: Catherine Dillard **Office:** Frost 376
Email: cdillard@hcc.edu **Phone**: 413-552-2432 (email is better than calling)
Office Hours: Tuesday 10-11 & 2-2:30, Wednesday 12:30-1:30, Thursday 2-2:30 and by appointment
Online office hour Monday 8-9pm. Use email or Skype ProfessorDillardHCC

4. Course Description:
This course will cover graphical presentation of data, measures of central tendency and variability, probability and probability distributions, the central limit theorem, estimation of parameters, hypothesis testing, regression and correlation, analysis of variance, as well as other topics at the instructor's discretion.
The prerequisite for this course is the completion of MTH085 or MTH099 or SM12 with a grade of C- or better, or an adequate score on the Mathematics Placement Examination.

5. No Required Text:
The textbook sections in the course refer to the free e-book Introductory Statistics from OpenStax
https://openstax.org/details/books/introductory-statistics
You can view the textbook online or you can download an electronic version. There is an option to buy a printed version. You are welcomed to this, but be aware it will probably be a different edition/version so the exact pages numbers may not coincide, but the topics should follow the same general order.
You are not required to use this book, it is available to you as a resource and I recommend you use it as needed.

Required Material:
1. You are required to have a calculator for this course.
2. You will need some basic spreadsheet software for graphs and statistical computations.
   All data will be given in Google Sheets and some homework assigments will require the use of Google Sheets – It is available through your HCC email
   MS Excel – Is available on HCC computers and if you are more comfortable on this, you are welcome to use it. Data from Google sheets can be copied into MS Excel.

Course Website: http://sites.google.com/a/hcc.edu/dillard/math142 (NOTE: this is not Moodle!)

6. Student Learning Outcomes
A student enrolled in this course should be developing the following abilities:
1. The ability to understand the methods of descriptive and inferential statistics.
2. The ability to assess the factors that affect the variability of data.
3. The ability to approach statistical experimentation in a clear, concise and logical manner.
4. The ability to solve problems of a statistical nature, including problems that require the application of learned knowledge in new and different situations.

These abilities will be applied to:
1. Thinking critically about others' use of statistical data
2. Using statistical data to support claims
3. Using statistical data to draw conclusions

7. Teaching Procedures:
Methods of instruction include small group in-class work, lecturing, demonstrating, question posing, large group in-class work, student presentations, guided discovery, projects, etc.

8. Course Topics:

Gathering Data: experiments, variables, populations, samples and biases

Analyzing Data: frequency distributions, central tendencies, range, variance, standard deviation, quartiles, percentiles, expected value, z-scores, probability fallacy, law of large numbers, probability density functions, common distributions, sampling distributions, estimation, confidence intervals, hypothesis testing, type I and II errors, p-value, central limit theorem, regression and correlation

Presenting Results: persuasive visual representations and persuasive written explanations of results

9. Tentative Test dates are midterm on Tuesday October 16th and final exam during final exam week TBA

10. Grading:

30% Homework Assignment

Homework assignments will be assigned in class and will be due the following class period unless otherwise stated.

Homework can be corrected and resubmitted at any time.

You may work with classmates or others but make sure you understand the material and each turn in your own work.

20% Papers

Papers can be corrected and resubmitted at any time.

25 % Midterm Exam

The midterm exam is scheduled for Tuesday March 6th. If you are going to miss that day you must notify me prior to the day of the test. Make up tests will be dealt with case by case.

25% Final Exam

This will be comprehensive. Date will be during exam week.

11. Attendance Policy:

Attendance will be taken but will not be a part of your grade. I reserve the right to Administratively Withdraw (AW) any student who misses 4 classes throughout the semester. Students who fail to attend class are not assured an AW and may receive an F grade. Communication with me is critical if you are regularly missing class.

12. Accessibility Accommodations

HCC is committed to full inclusion of all students. Students with a documented disability should contact the Office for Students with Disabilities and Deaf Services at 413-552-2417 or 413-552-5502 (Video phone for Deaf/Hard-of-hearing students) to request an intake appointment and discuss accommodations. Collaboration between students, OSDDS, and instructors is essential for timely accommodations, so students are urged to provide their accommodation agreement to instructors in the first week of the term or immediately after receiving the approved accommodation from OSDDS.

Academic Dishonesty:

Please see the student handbook for definition. I reserve the right to respond to academic dishonesty as I see fit, including but not limited to reporting the incident to the Dean of Students, giving a zero on the assignment and all like assignments, and giving the student an F for the semester.

Main Points:

1- If you are struggling, come see me! I am happy to find a time outside of office hours to meet with you if my office hours do not work for you.

2- Ask questions, ask questions, ask questions!

General Outline for the course leading up to the midterm exam

Sept 4 TEXT Ch1 and Ch 2 Introduction to statistics (pages 5-8) HW: Finish introduction to statistics (pages 5-8) Finish bias in surveys (pages 9-12) Definitions (pages 13-20)	Sept 6 TEXT 2.1-2.4 Discuss definitions (pages 13-20) Using graphs (pages 21-25) HW: Graphing homework (page 24 and online)
Sept 11 TEXT 2.5 and 2.7 Comparing data sets notes (pages 25-30) mean, standard deviation, median, 'iles HW: Comparing two sets of data homework (pages 31-35) Calculator practice (page 36)	Sept 13 TEXT 6.1 Coefficient of variation (pages 37-38) Start exploring z-scores worksheet (pages 43-48) HW: Coefficient of variation and spreadsheet homework (pages 39-42) Finish z-scores worksheet (pages 43-48)
Sept 18 Stats in the news (pages 49-54) HW: News article homework (page 49)	Sept 20 TEXT Ch 3 Basic Probability (pages 55-56) HW: Probability homework (pages 57-58)
Sept 25 TEXT 4.1, 6.1 and 5.2 Probability Density Function worksheet (pages 59-64) HW: finish Probability Density Function worksheet (pages 59-64) Normal and Uniform density function (pages 65-66)	Sept 27 TEXT 6.1 and 6.2 Normal data and z-scores (pages 67-70) HW: z-scores and probabilities homework (pages 73-74)
Oct 2 Catch up day HW: Midterm review I (pages 75-77)	Oct 4 Take questions on topics on midterm review I HW: Midterm review II (pages 78-82)
Oct 9 No Class Monday Classes Today	Oct 11 Take questions on topics on midterm review II HW: study for midterm
Oct 16 Midterm Exam	Oct 18 Go over the midterm exam

Note: TEXT refers to the online textbook.

If you are using other resources, I would be happy to help you locate topics within those resources.

Grouping

I am number _____

Shapes

Square	1	2	3	4
Triangle	5	6	7	8
Hexagon	9	10	11	12
Octagon	13	14	15	16
Rectangle	17	18	19	20
Rhombus	21	22	23	24
Diamond	25	26	27	28

Trucks

Dump	1	5	10	14
Mixer	3	7	12	16
Recycling	2	8	9	13
Semi	4	11	15	18
Pick-up	6	17	21	26
Logging	19	22	25	28
Bucket	20	23	24	27

Kayak Words

Kayak	1	7	22	28
Paddle	3	9	14	18
Tandem	2	5	11	20
Rudder	4	10	13	27
Roof rack	6	12	15	23
Life vest	16	17	25	26
Dry bag	8	19	21	24

Veggies

Broccoli	1	6	9	16
Peppers	3	5	22	26
Spinach	7	11	14	20
Lettuce	12	13	17	18
Carrots	2	15	24	27
Asparagus	4	10	21	25
Eggplant	8	19	23	28

Fruits

Apple	1	8	11	16
Orange	3	6	13	22
Kiwi	4	5	24	25
Pear	7	10	15	26
Mango	2	14	17	23
Peach	12	18	19	27
Banana	9	20	21	28

Pets

Hamster	1	13	20	26
Snake	3	6	11	18
Hermit crab	5	9	12	22
Bird	7	17	24	28
Fish	2	14	15	19
Cat	4	8	21	27
Dog	10	16	23	25

Trees

Elm	1	2	6	25
Oak	3	14	27	28
Maple	4	5	23	26
Pine	7	9	18	24
Dogwood	8	10	11	17
Aspen	12	15	20	21
Willow	13	16	19	22

Olympic Sports

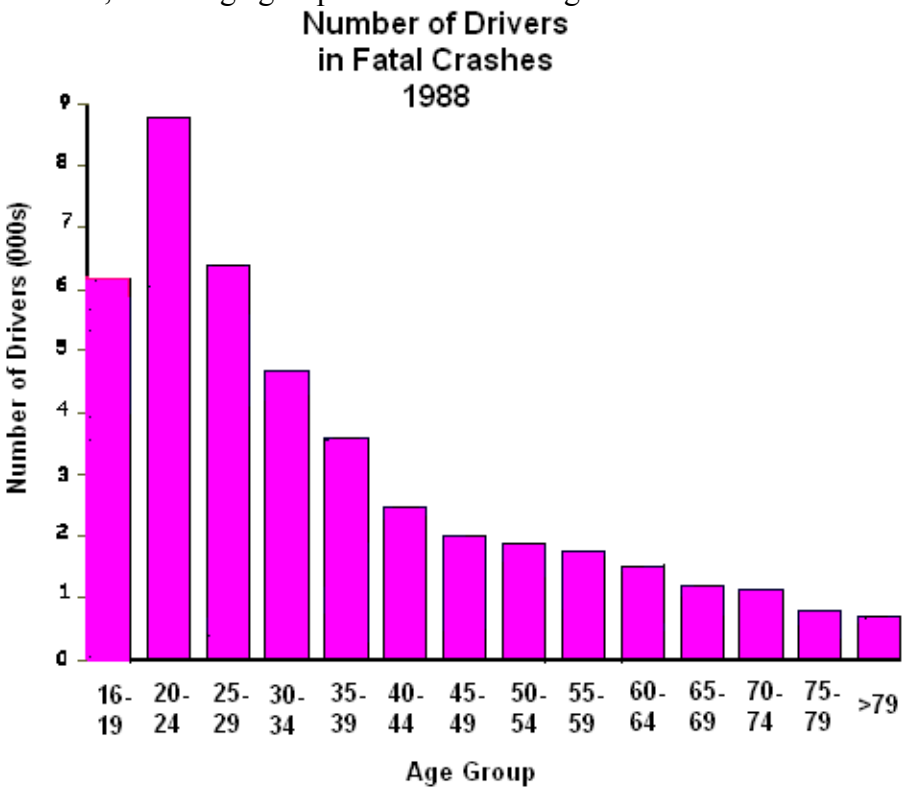
Bobsled	1	12	26	27
Archery	3	8	10	15
Curling	5	16	17	20
Handball	2	4	7	19
Ice dancing	6	9	23	28
Steeple chase	11	22	24	25
Biathlon	13	14	18	21

Insects

Ant	1	15	24	28
Bee	2	3	21	26
Caterpillar	5	6	10	19
Dragonfly	7	18	20	23
Earwig	4	9	14	17
Firefly	11	16	22	27
Grasshopper	8	12	13	25

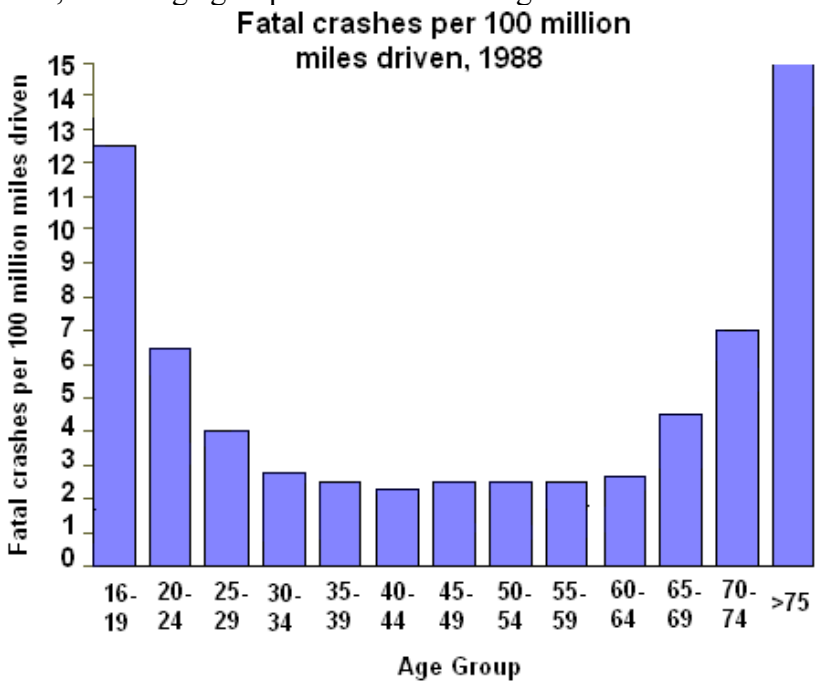
Introduction to Statistics

1. According to this data, which age groups are the most dangerous?



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

2. According to this data, which age groups are the most dangerous?



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

3. According to this data, rank these airlines from worst to best
Complaints per paying passenger from Nov 2000 (rounded)

United Airlines	American Airlines	Delta Airlines	Continental Airlines	Southwest Airlines	Alaska Airlines
0.040	0.030	0.010	0.030	0.005	0.010

4. According to this data, rank these airlines from worst to best
US News and World Report on February 5, 2001: Complaints Nov 2000

Most complaints			Fewest Complaints		
United Airlines	American Airlines	Delta Airlines	Continental Airlines	Southwest Airlines	Alaska Airlines
252	162	119	60	22	13

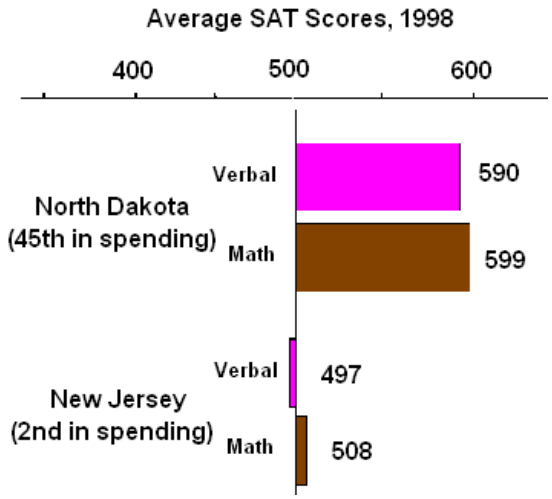
5. The following statistics about motorcycle helmet use seem to suggest that helmets cause more injuries and fatalities. Is it safer to go without helmets?

	Registrations	Reported Accidents	Fatalities	Accidents Per 10,000 Registrations	Fatalities Per 100 Accidents
Mandatory Helmet Use	2,352,293	52,270	1,557	222.21	2.98
Voluntary Helmet Use	1,497,923	29,062	844	194.02	2.9
Total	3,850,216	81,331	2,401	211.24	2.95

Source: Motorcycle Statistical Annual, Motorcycle Industry Council, Inc., 1994, as reported on http://www.bikersrights.com/statistics/stats.html.

6.
Columnist George Will wrote in the Washington Post in 1993 that "... the 10 states with the lowest per pupil spending included four — North Dakota, South Dakota, Tennessee, Utah — among the 10 states with the top SAT scores ... New Jersey has the highest per pupil expenditures, an astonishing $10,561… [Its] rank regarding SAT scores? Thirty-ninth."

Does this mean that spending more on education makes students worse off?



7. “[P]eople in fluoridated areas have a higher cancer death rate than those in non-fluoridated areas.” –Dr. John Yianmouyiannis. Thus, fluoride consumption by human beings increases the general cancer death rate. Should fluoridation be prohibited?

8. Except from: “The Dark Side of Illegal Immigration: Nearly One Million Sex Crimes Committed by Illegal Immigrants in the United States.” By Dr. Deborah Schurman-Kauflin

> *After conducting a 12 month in-depth study of illegal immigrants who committed sex crimes and murders for the time period of January 1999 through April 2006 , it is clear that the U.S. public faces a dangerous threat from sex predators who cross the U.S. borders illegally.*
>
> *There were 1500 cases analyzed in depth. They included: serial rapes, serial murders, sexual homicides, and child molestation committed by illegal immigrants. Police reports, public records, interviews with police, and media accounts were all included. Offenders were located in 36 states, but it is clear, that the most of the offenders were located in states with the highest numbers of illegal immigrants. California was number one, followed by Texas, Arizona, New Jersey, New York, and Florida.*
>
> *Based on population numbers of 12,000,000 illegal immigrants and the fact that young males make up more of this population than the general U.S. population, sex offenders in the illegal immigrant group make up a higher percentage. When examining ICE reports and public records, it is consistent to find sex offenders comprising 2% of illegals apprehended. Based on this 2% figure, which is conservative, there are approximately 240,000 illegal immigrant sex offenders in the United States.*
>
> *This translates to 93 sex offenders and 12 serial sexual offenders coming across U.S. borders illegally per day. The 1500 offenders in this study had a total of 5,999 victims. Each sex offender averaged 4 victims. This places the estimate for victimization numbers around 960,000 for the 88 months examined in this study.*

Find some flaw(s) in her logic.

A non-statistical example of two conclusions drawn from the same data

http://theinconvenientskeptic.com/2011/04/reaching-different-conclusions-from-the-same-data/

Excerpt by John Kehr

The Data:
In 1969 palaeontologist John Ostrom described the Deinonychus (velociraptor variant) as a human sized predator that was quick fast and otherwise a very nasty predator. He found that the remnants of at least three of these raptors around the body of a Tenantosaurus which is much, much bigger than a raptor. Many times that tenantosaurus remains were found, raptor teeth were also found. This is evidence that the raptors fed on tenantosaurus. However a raptor was not large enough to hunt such a large creature on it's own. A tenantosaurus falls into the size of creature that it doesn't need spikes to defend itself.

Theory #1:
This leads to first theory which was proposed by Dr. Ostrom in his initial paper. This theory is that raptors hunted in packs and they used pack behavior to take down much larger prey. Much like wolves are effective at taking down prey they could not tackle alone.
This theory is popular, partly because it is cool and exciting. It has highly popularized by the movie Jurassic Park. This requires intelligence and the ability to coordinate actions between individuals, but it remains only a theory.

Theory #2:
In 2007 Brian Roach and Daniel Brinkman argued that the evidence does not support that they were pack hunters, but that they were scavengers that fought over carcasses that were killed by larger predators. So a T-Rex would kill a tenantosaurus, it would eat its fill and go away. Then the smaller scavengers would show up and fight over the remains.
This would also explain why the raptor remains were around the bodies of tenantosaurus. Komodo dragons behave in a similar manner when there is a successful hunt and it could be argued that raptors have more in common with komodo dragons than they do with wolves.

Possible Bias is Survey Questions

In one study of the effect of wording in survey results researchers showed college students films of an automobile accident, after which they asked them a series of questions. One group was asked the question, "About how fast were the cars going when they contacted each other?" The average response was 31.8 miles per hour. Another group was asked, "About how fast were the cars going when they collided with each other?" In that group, the average response was 40.8 miles per hour. Simply changing the word *contacted* to the word *collided* increased the estimates speed by 9 miles per hour, or 28%, even though the respondents had witnessed the same film.

1. a) Give another example of wording that might increase the average speed estimates

b) Give another example of wording that might decrease the average speed estimates

Sometimes, if a survey is being conducted to support a certain cause, questions are deliberately worded in a biased manner. *Be careful about survey questions that begin with phrases like "Do you agree that…?" most people want to be agreeable and will be inclined to answer "yes" unless they have strong feelings the other way.* For example, suppose an antiabortion group and a pro-choice group each wanted to conduct a survey in which it would find the best possible agreement with its position. Here are the two questions that would each produce different estimates of the proportion of people who think abortion should be illegal.

Do you agree that abortion, the killing of innocent beings, should be outlawed?

Do you agree that there are circumstances under which abortion should be legal to protect the rights and well-being of the mother?

2. Reword the abortion question to remove most, if not all, of the bias.

Sometimes questions are worded in such a way that the meaning is misinterpreted by a large percentage of the respondents. For example, if you were to ask people whether or not they use drugs, you would need to specify if you mean prescription drugs, illegal drugs, over-the-counter drugs, or common substances like caffeine.

If you were to ask people to recall the most important date in their life, you would need to clarify if you meant the most important calendar date or the most important social engagement with a potential partner. (It is unlikely that anyone would mistake the question as being about the shriveled fruit, but you can see that the same word can have multiple meanings.)

3. Give another example of a question that could be interpreted in several very different ways.

Most survey respondents have a desire to please the person who is asking the question. They tend to understate their responses about undesirable social habits and opinion, and vice versa. For example, in recent years estimates of the prevalence of cigarette smoking based on surveys do not match those based on cigarette sales. Either people are not being completely truthful, or lots of cigarettes are ending up in the garbage.

4. Give another example of an "undesirable" social habit researchers might be interested in learning about.

If questions are to be understood, they must be kept simple. A question like "Shouldn't former drug dealers not be allowed to work in hospitals after they are released from prison?" is sure to lead to confusion. Does a "yes" answer mean they should or should not be allowed to work in hospitals? It would take a few readings to figure that out.

Another way in which a question can be unnecessarily complex is by actually asking more than one question at once. An example would be a question like "Do you support the President's Health Care Plan, since it would ensure that all Americans receive health coverage?" If you agree with the idea that all Americans should receive health coverage but disagree with the remainder of the plan, do you answer "yes" or "no"? Or what if you support the President's plan, but not for that reason?

5. Write down another example of a slightly confusing question.

Sometimes words mean different things to different people. When you read about survey results, you should get a precise definition of what was actually asked or measured.

6. Write an example of a question that could be interpreted different ways.

If one question requires respondents to think about something that they may not have otherwise considered, then the order in which questions are presented can change the results.
For example, suppose a survey were to ask, "To what extent do you think teenagers today worry about peer pressure related to drinking alcohol?" and then ask, "Name the top five pressures you think face teenagers today." It is quite likely that the respondents would use the idea they had just been given and name peer pressure related to drinking alcohol as one of the five choices.

People do not like to admit that they don't know what you are talking about when you ask them a question. One author gives an example: "When the American Jewish Committee studied Americans' attitudes toward various ethnic groups, almost 30% of the respondents had an opinion about the fictional Wisians, rating them in social standing above a half-dozen other real groups, including Mexicans, Vietnamese and African blacks." Political pollsters, who are interested in surveying only those who will actually vote, learned long ago that it is useless to simply ask people if they plan to vote. Most of them will say "yes." Instead, they ask questions to establish a history of voting, such as "Where did you go to vote in the last election?"

People will often answer questions differently based on the degree to which they believe they are anonymous. Because researchers often need to perform follow-up surveys, it is easier to try to ensure confidentiality than anonymity.
In ensuring **confidentiality**, the researchers promise not to release identifying information about respondents.
In an **anonymous** survey, the researcher does not know the identity of the respondents.
Questions on issues like sexual behavior and income are particularly difficult because people consider those to be private matters. A variety of techniques have been developed to help ensure confidentiality, but surveys on such issues are hard to conduct accurately.

7. So do all these potential sources of bias make surveys useless? _____

8. How does being aware of these potential issues help us as statisticians and researchers?

9. How does being aware of these potential issues help us as members of the general public?

Interesting Assignment from a Calvin University Stats Class

Students were asked to design a pair of questions that related to the same issue, but for which one or both would bias answers in a particular direction. In order to test out how well their biased questions "worked", they asked the questions to 20 people, randomly deciding who got which question. The questions and results are shown below:

- Gun Control
    1. With all the gang killings and domestic disputes ending with gun fire, do you think there should be legislation passed to hinder gun ownership? (6 yes; 4 no)
    2. The second amendment guarantees citizens the right to bear arms, should this right be compromised by legislation in the twenty-first century? (1 Yes; 9 No)

- Drinking
    1. Do you agree that drinking is ok as long as you are of age and that you don't get drunk? (yes=8; no=2)
    2. Do you think drinking is good or bad? (good=6; bad=4)

- Abortion
    1. Do you think that it is ok to get an abortion if it is the only way to save the mother's life? (yes=9; no=1)
    2. Do you think abortion is good or bad? (good=3; bad=7)

- Smoking
    1. Should people have the right to smoke? (yes=9; no=1)
    2. Since cigarettes are dangerous and have deadly side effects such as cancer wouldn't you agree that smoking should be controlled to save the lives of many? (yes=7; no=3)

- Amount of Homework
    1. Do you agree or disagree that Calvin's professors assign an average amount of homework? (Agree=2; Disagree=8)
    2. Do you agree with the National College Study Program which says that Calvin assigns an average amount of homework for the college student? (Agree=6; Disagree=4)

- Bathroom Size
    1. Do you agree or disagree that Calvin has big enough bathrooms? (Agree=5; Disagree=5)
    2. Do you agree with the National Students Rights Poll which says that Calvin has big enough bathrooms? (Agree=9; Disagree=1)

9. Circle the two that were the most interesting or surprising to you

Definitions

You may look this up online or several of them are in the textbook (use the index).

Any definition marked with an * is not in your text.

Define these words as used in a statistical context and be able to do it in your own words

1. Define experiment (Text 3.1)

2. Give an example of a statistical experiment

3. Define variable (Text 1.1)

4. What is the variable in the statistical experiment in #2?

5. Define parameter (Text 1.1)

6. Define population (Text 1.1)

7. Define sample (Text 1.1)

8. Define sample space (Text 3.1)

9. Define random sample (Text 1.2)

10. Define representative sample (Text 1.1)

11. Define sampling error (Text 1.2)

12. Define outlier (Text 2.3)

13-16. Go to: www.youtube.com/watch?v=t3DrD_XDUhM The video is titled: Square One TV 142
This video is 27 minutes long but you only need to watch the following clips (they are three parts to the same clip) but feel free to watch and enjoy the entire show. Direct link on Google site for this course.

Watch 0:37 to 2:43
6:14 to 8:20
11:31 to 13:10

13. Use the video to explain random sampling

14. How did the pollster make the final sample more random than the first two tries?

15. How could the pollster have made the final sample even more random?

16. Some people believe it is impossible to get a totally random, totally unbiased sample. And those who believe you can, generally admit that it is very, very difficult. So part of being a good statistician is being aware of potential biases and non-randomness in any sample.
List some things about the pollster in the video that might influence how people responded to his poll.

17. What is the difference between truncating and rounding?*

18. Give an example where truncation and rounding of the same number yield the same answer

19. Give an example where truncation and rounding of the same number yield different answers

20. Define the mean of a set of data (Text 2.5)

21. Give an example of fewer than 10 numbers where the mean is a good description of the data

22. Give an example of fewer than 10 numbers where the mean is not a good description of the data

23. Define the median of a set of data (Text 2.5)

24. Give an example of fewer than 10 numbers where the median is a good description of the data

25. Give an example of fewer than 10 numbers where the median is not a good description of the data

26. Define the mode of a set of data (Text 2.5)

27. Give an example of fewer than 10 numbers where the mode is a good description of the data

28. Give an example of fewer than 10 numbers where the mode is not a good description of the data

29. Define range as it pertains to a set of numerical data*

30. Define descriptive statistics (Text chapter 2 intro)

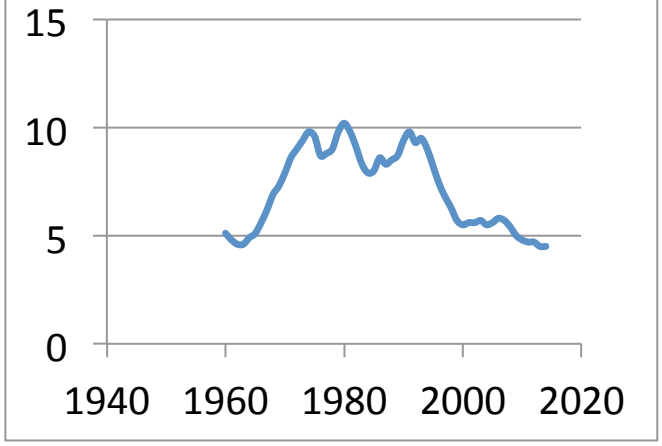
31. Define inferential statistics (Text chapter 8 intro)

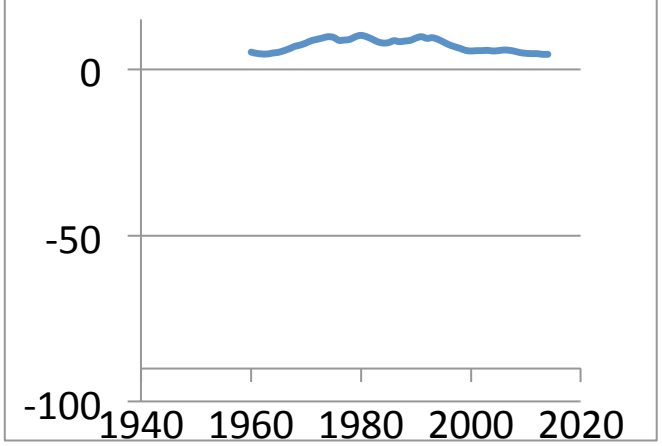
32. Define qualitative data (Text 1.2)

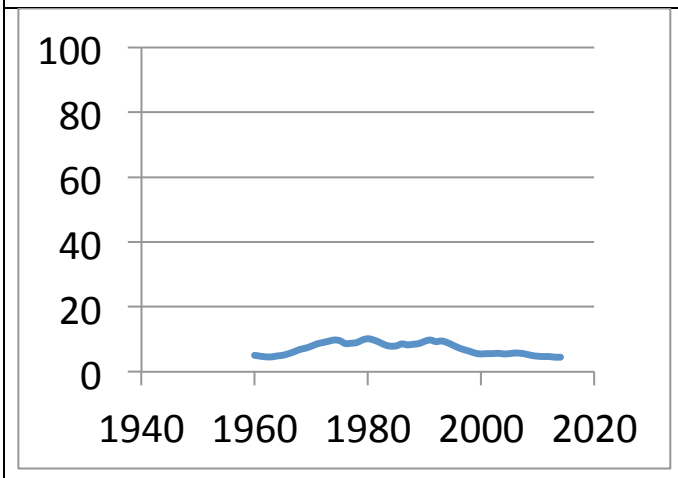
33. Define quantitative data (Text 1.2)

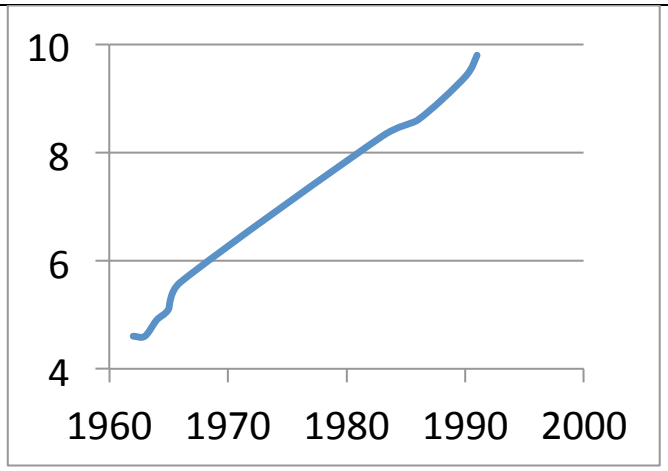
Using Graphs

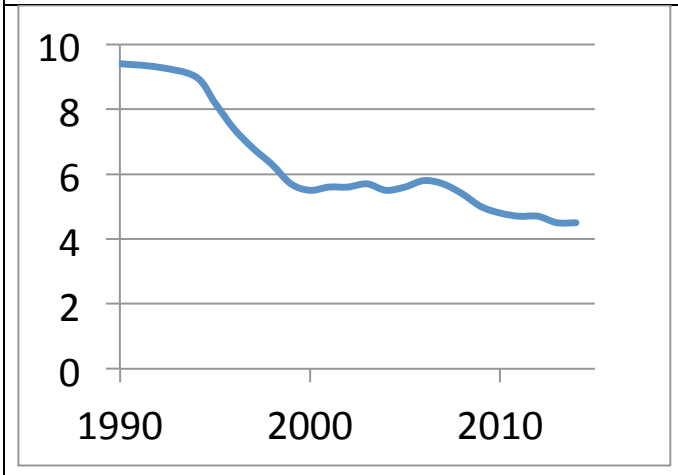
For each of the following graphs, what conclusion is the author hoping you will draw about murder rates in the United States?

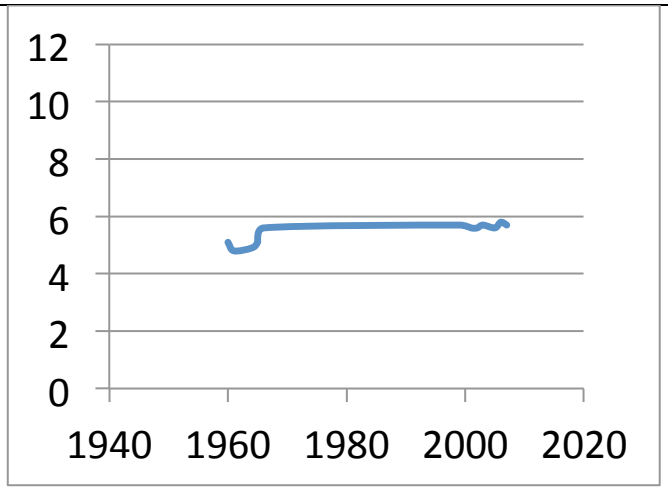




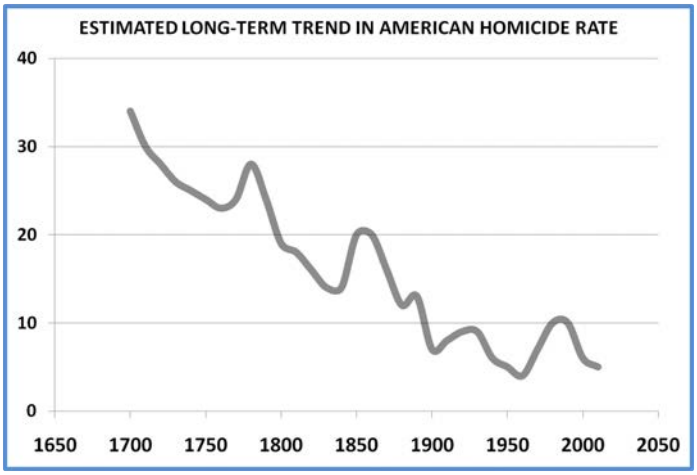






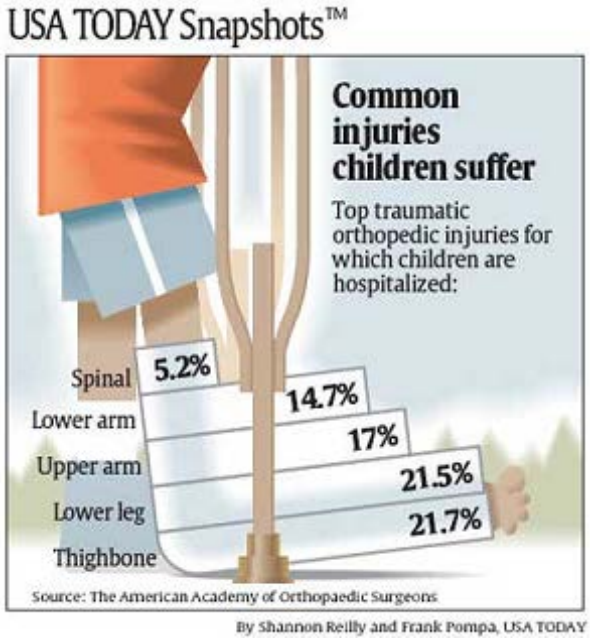


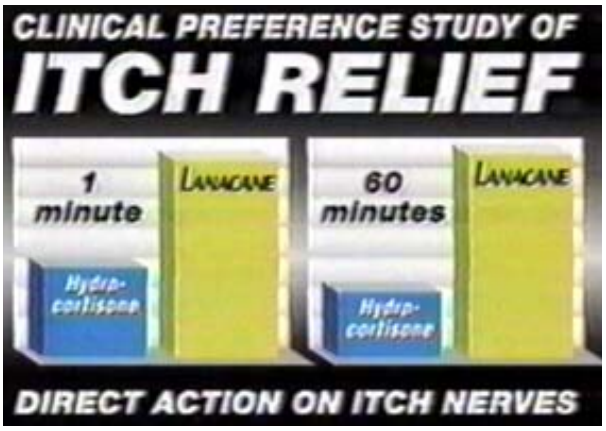
All of the data on the previous page is from the exact same data set. The graphs are not "lying" in that they use actual data points, but the conclusions implied are very different.
Annual murder rates for the US per 100,000 people from the Uniform Crime Report data compiled by the FBI. The UCR data has its own issues as well.
Just to add more information, this graph is another estimate of murder rates dating back to 1700

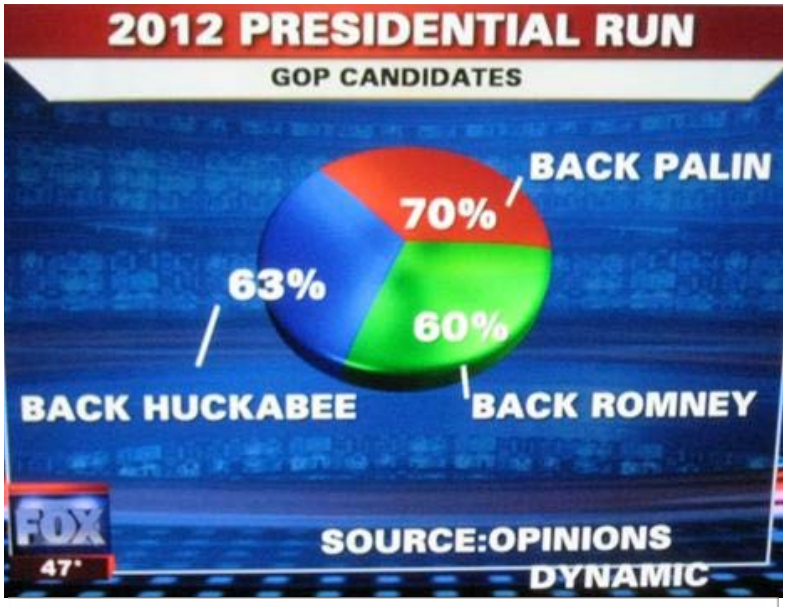


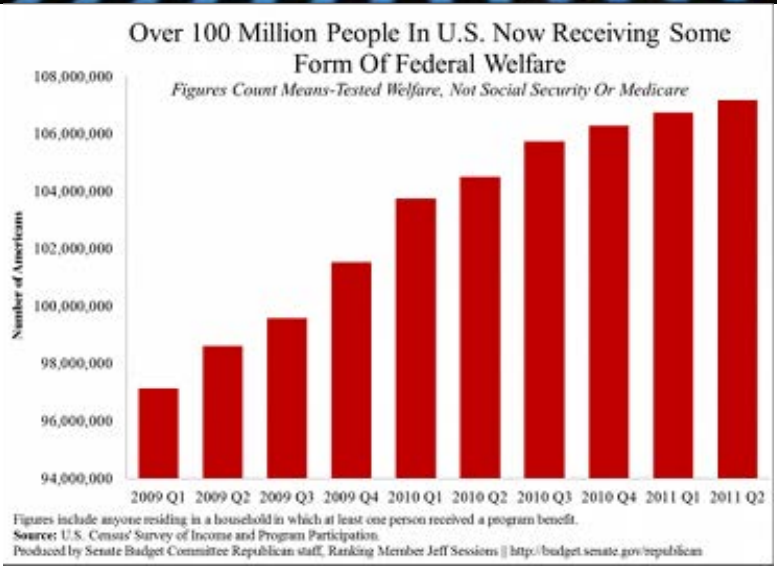
The graphs on the previous page used selective data points, zooming and changing the axes to make the data look different. There are also other graphing methods to help be deceptive or make your visual argument more effective.

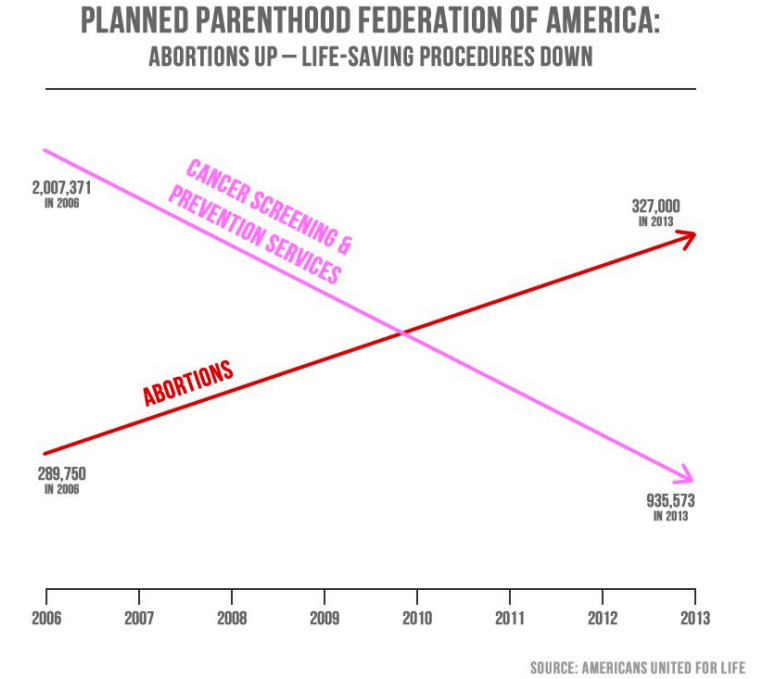
Here are a few examples of misleading graphs

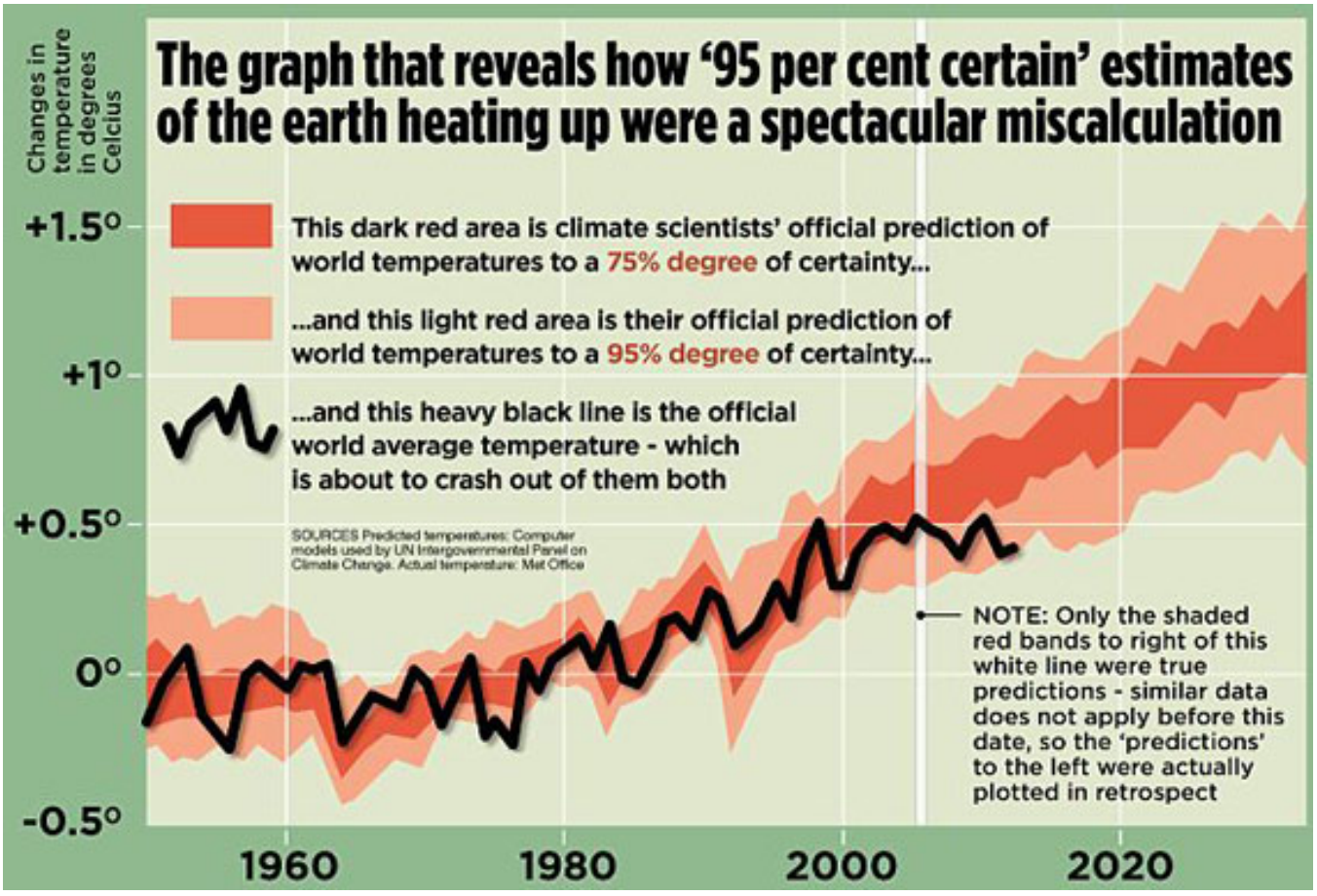





2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47°
SOURCE:OPINIONS
DYNAMIC


Over 100 Million People In U.S. Now Receiving Some Form Of Federal Welfare
Figures Count Means-Tested Welfare, Not Social Security Or Medicare
Number of Americans
108,000,000
106,000,000
104,000,000
102,000,000
100,000,000
98,000,000
96,000,000
94,000,000
2009 Q1 2009 Q2 2009 Q3 2009 Q4 2010 Q1 2010 Q2 2010 Q3 2010 Q4 2011 Q1 2011 Q2
Figures include anyone residing in a household in which at least one person received a program benefit.
Source: U.S. Census' Survey of Income and Program Participation
Produced by Senate Budget Committee Republican staff, Ranking Member Jeff Sessions || http://budget.senate.gov/republican


PLANNED PARENTHOOD FEDERATION OF AMERICA:
ABORTIONS UP – LIFE-SAVING PROCEDURES DOWN
CANCER SCREENING & PREVENTION SERVICES
2,007,371
IN 2006
327,000
IN 2013
ABORTIONS
289,750
IN 2006
935,573
IN 2013
2006
2007
2008
2009
2010
2011
2012
2013
SOURCE: AMERICANS UNITED FOR LIFE



Graphing Homework: (direct link on Google site for this class)

https://docs.google.com/spreadsheets/d/1kTk1Gt8q9Hx2_071tELA7AW7PBSF5v9AmFOnQHIVcU0/edit?usp=sharing

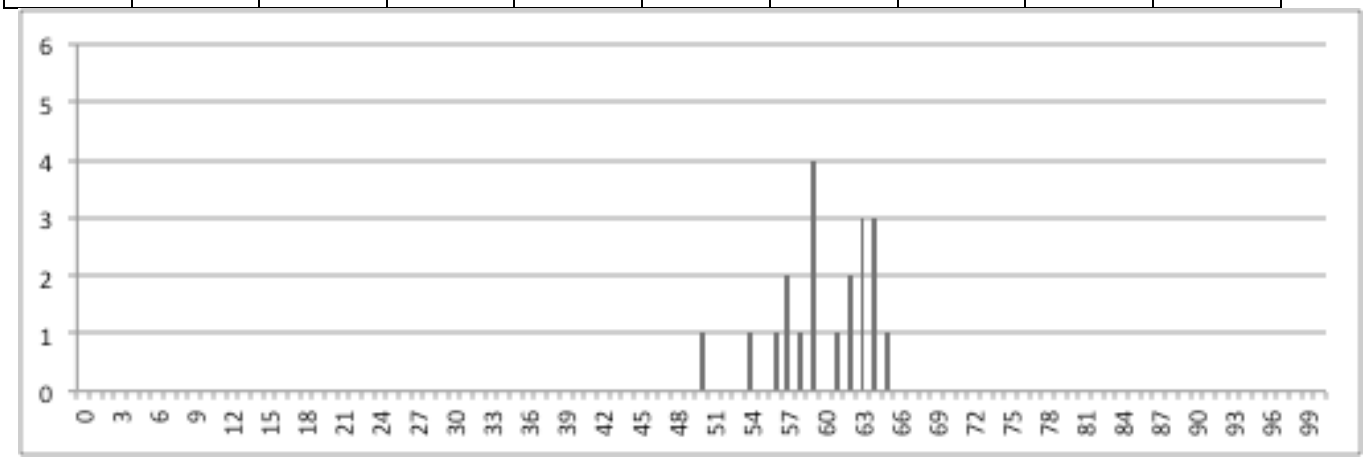
Note: there are five sets of data and you are going to email me a google doc of your graphs.

Comparing Data Sets Notes

Blood samples were collected from 20 people at three different locations in Holyoke. The researchers were measuring the percentage of the blood cells that appeared to contain Gloxum.

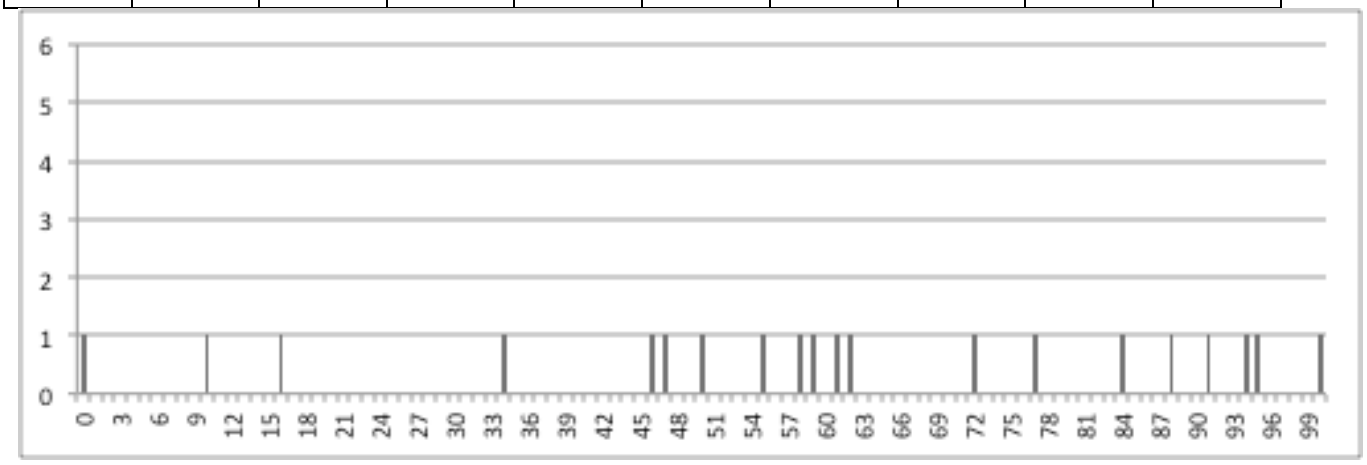
Region 1:

50	54	56	57	57	58	59	59	59	59
61	62	62	63	63	63	64	64	64	65



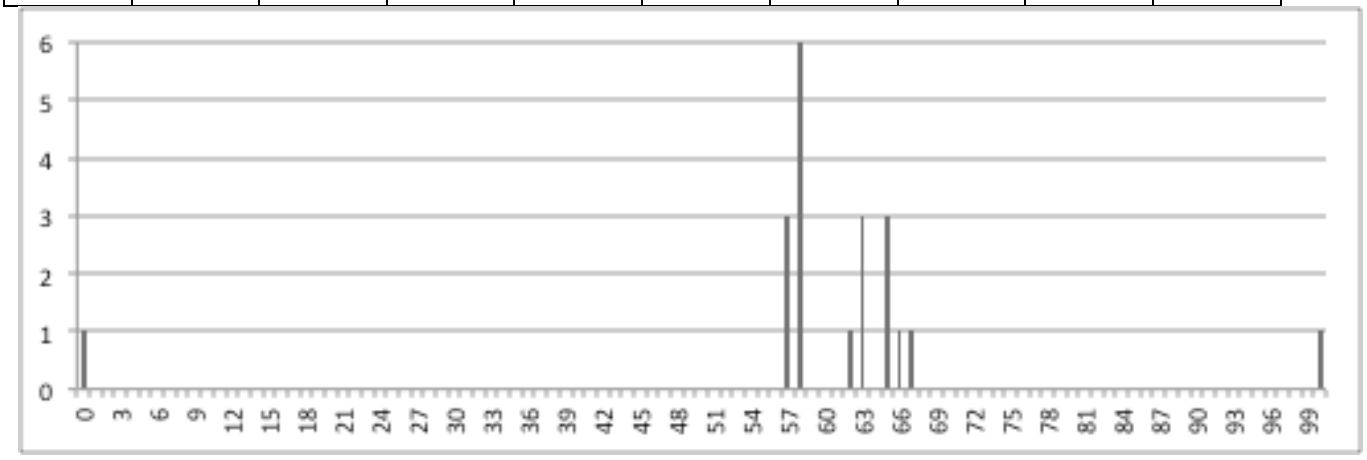
Region 2:

0	10	16	34	46	47	50	55	58	59
61	62	72	77	84	88	91	94	95	100



Region 3:

0	57	57	57	58	58	58	58	58	58
62	63	63	63	65	65	65	66	67	100



Compare the results from these 3 regions:

Describing center and spread with median and ‘iles

The measure of central tendency used: median

One benefit of the median over the mean is that it is not as affected by a few outliers.

i.e. 1, 1, 1, 2, 2, 2, 3, 3, 3, 1998 the mean=201.6 the median=2

The measure to show the spread of the data: range or ‘iles,

namely quartiles, percentiles and deciles (there are others)

Split an ordered data set into equal sized groups and state the splits.

2 groups

Region 1: minimum= middle= max=

This is just the median and the range so let’s try more groups.

Quartiles

As the name suggests we are breaking the data into how many equal parts?_____

Region 1:

50 54 56 57 57 58 59 59 59 59 61 62 62 63 63 63 64 64 64 65

Split into 4 equal sized groups and list the splits.

Region 1	Region 2	Region 3
Minimum=	Minimum= 0	Minimum= 0
Quartile 1 split=	Quartile 1 split= 46.5	Quartile 1 split= 58
Quartile 2 split/Median=	Quartile 2 split/Median= 60	Quartile 2 split/Median= 60
Quartile 3 split=	Quartile 3 split= 86	Quartile 3 split= 65
Maximum=	Maximum= 100	Maximum= 100

Compare the three regions:

Describing center and spread with mean and standard deviation

The measure of central tendency used: mean

The measure to show the spread of the data: standard deviation

Deviation: To stray from a standard, principle, or topic. To depart from an established course or norm. (Merriam-Webster dictionary)

In statistics, to stray from the norm is to quantify how far the data points are from the mean.

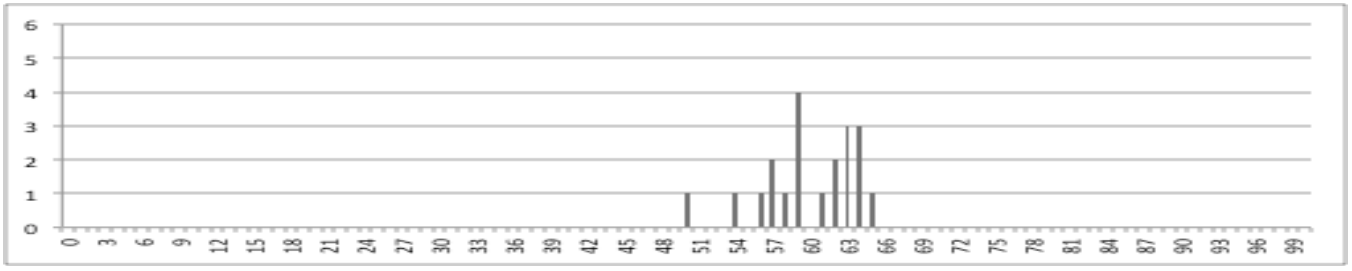
It defines how clustered around the mean the data is. More cluster, the less standard deviation.

It defines how spread out data is. The less spread out, the less standard deviation.

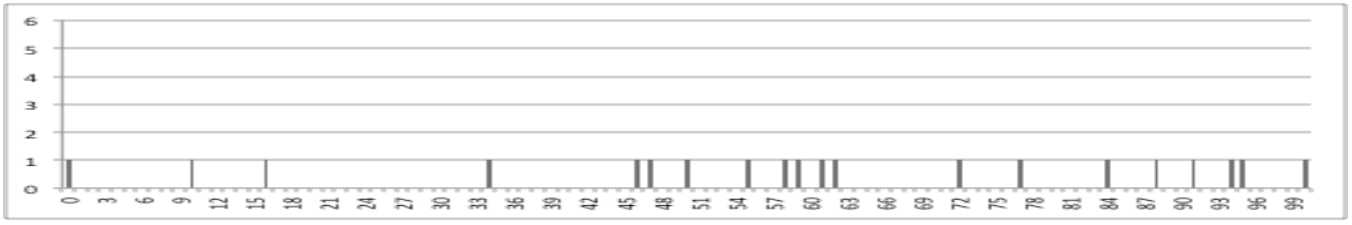
Think of the standard deviation as an "average" of the distances each point is from the mean

Look back at the three regions.

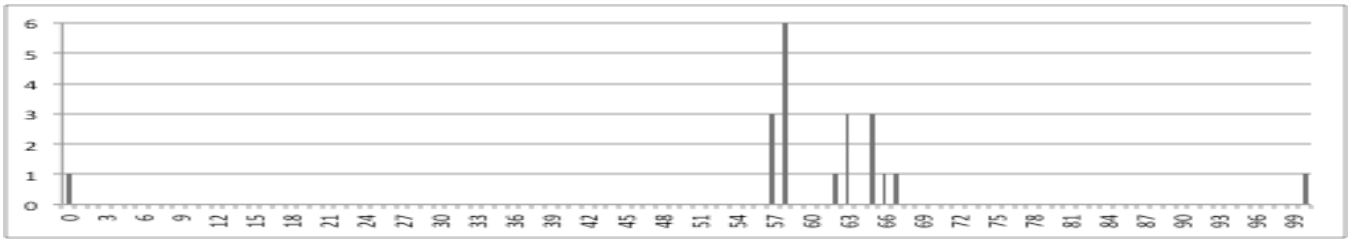
Region 1:



Region 2:



Region 3:



Which region has the lowest standard deviation? Why?

Which region has the highest standard deviation? Why?

Now let's look at the formula for standard deviation. Note μ is the mean.

Standard Deviation = $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2}{n-1}}$ where the x's are each data point.

(Note: this is the formula for standard deviation of a sample. There is another, slightly different, formula for the standard deviation of a population, but we only use that when we have gathered data for the entire population of our experiment. For the types of experiments we are discussing and doing in class, we will not be gathering data from an entire population. But you need to know it exists.)

Why isn't it a simple average of the distances from the mean?

Use the formula to explain why including a point much, much higher than the mean increases the standard deviation.

Use the formula to explain why having points all very closer to the mean results in a small standard deviation.

Now let's find the standard deviations of the three regions (done using a spreadsheet)

Region 1	Region 2	Region 3
σ= 3.89 percentage points	σ= 28.97 percentage points	σ= 16.92 percentage points

It makes sense to directly compare standard deviations in this case because the three data sets are looking at similar information. But that is not always the case.

Ways to describe the center of a set of data			Ways to describe the spread of the data	
Median	Mode	Mean μ	Range	Standard deviation σ
The middle number in the list when the data is ordered from least to greatest	The most common number appearing in a data set	The average of all the numbers in the data set	The highest number and the lowest number in a set of data OR The highest number minus the lowest number in a set of data	A measure of how clustered the data points are around the mean

Which is better: the median, the mode, or the mean? Depends on the data and the purpose of the analysis!

Which is better: the range or the standard deviation? Depends on the data and the purpose of the analysis!

Which is better: a small standard deviation or a large one? Depends on the data and the purpose of the analysis!

Which is better: a small range or a large range? Depends on the data and the purpose of the analysis!

Comparing two sets of data homework

Deciles

1. As the name suggests we are breaking the data into how many equal parts?_____

2. Take this ordered data and mark the deciles and the median

-6	-5	-3.4	-2.7	-0.4	-0.3	0.8	1	2	2.1	3.06	3.85	4.1	4.5	5
5.02	5.03	5.6	5.9	5.97	6	6.02	6.7	7.2	8	8.4	8.56	9.05	9.2	9.9

3. For any set of data

a) If a data point N is in the first decile, then that point is above at least _____% of the data points and below at least _____% of the data points.

b) If a data point S is in the fifth decile, then that point is above at least _____% of the data points and below at least _____% of the data points.

c) If a data point H is in the ninth decile, then that point is above at least _____% of the data points and below at least _____% of the data points.

d) If a data point C is in the fourth decile, then that point is above at least _____% of the data points and below at least _____% of the data points.

Percentiles

4. As the name suggests we are breaking the data into how many equal parts?_____

5. Explain the meaning of the following and make sure to include the context.

a) Jo scored in the 75^{th} percentile on the SAT

b) Kai is in the 23^{rd} percentile for head circumference in 18 month old girls

c) Eliza's scoring average is in the 94^{th} percentile for high school basketball players

d) Ezra's earned run average (ERA) is in the 7^{th} percentile for college baseball pitchers

Box plot

Look up what a box plot is. Note: it shows the median and each quartile.

6. Make a box plot for each of the data sets provided below.

Make sure you use and label an accurate scale

Data Set A

4.08
10.35
10.76
15.43
17.01
17.17
17.41
18.06
19.02
19.13
19.96
20.2
20.71
21.83
21.9
23.11
23.22
23.67
24.13
27.56
27.6
27.9
28.15
30.96
31.6
33.07
33.27
35.54
36.8
37.6
38.11
38.96
42.99
44.44
44.44
45.17
45.35
49.27
49.37
51.9

Box Plot for data set A

Box Plot for Data Set B

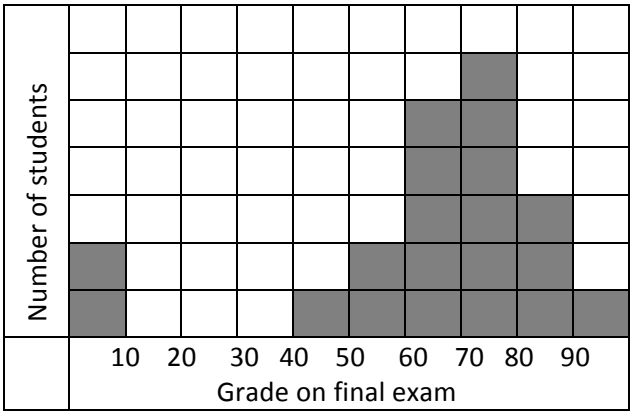
Data Set B

-11
-8
-7
-5
-5
-4
-3
-2
-2
-1
-1
-1
-1
0
0
1
1
1
2
2
3
4
4
5
6
10
13
13
14
20
21

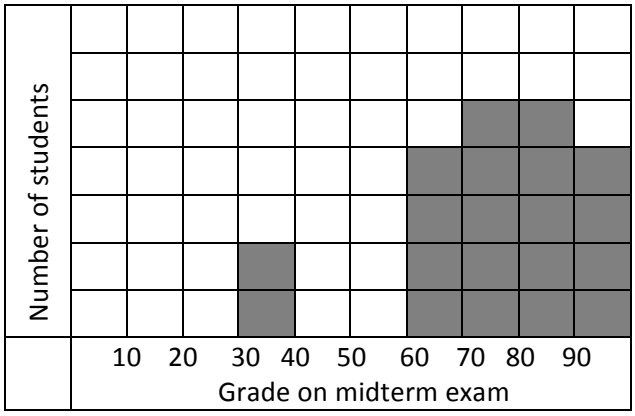
#7-#9 For each pairing of graphical data, label which one has a HIGHER standard deviation. Briefly state why.

7.

a. mean is 64

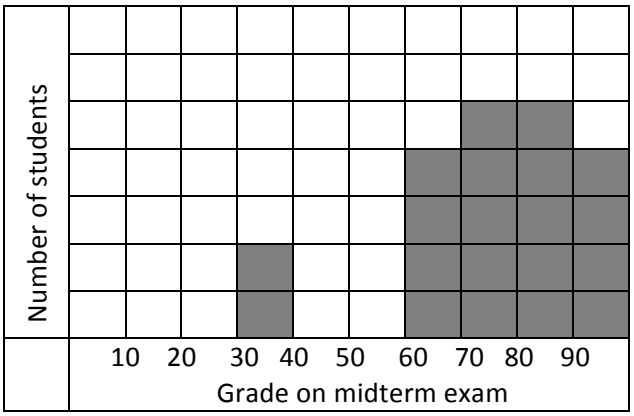


b. mean is 76

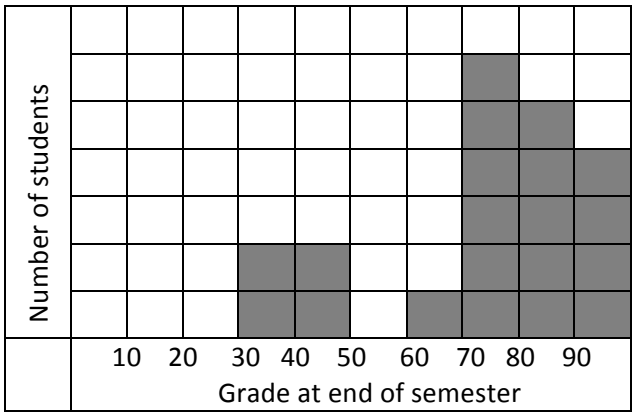


8.

a. mean is 76

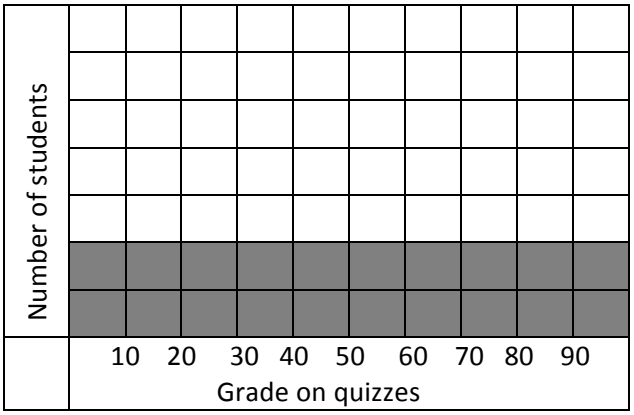


b. mean is 74

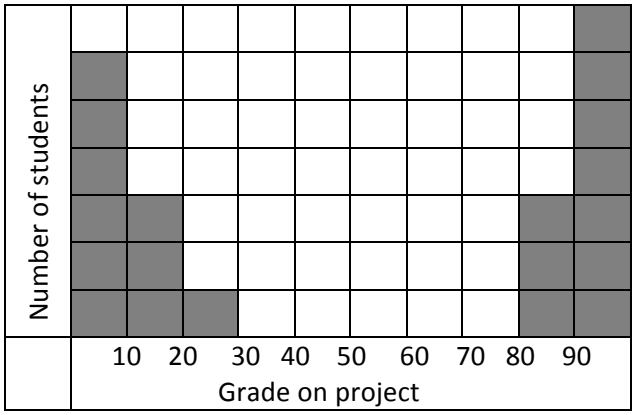


9.

a. mean is 50

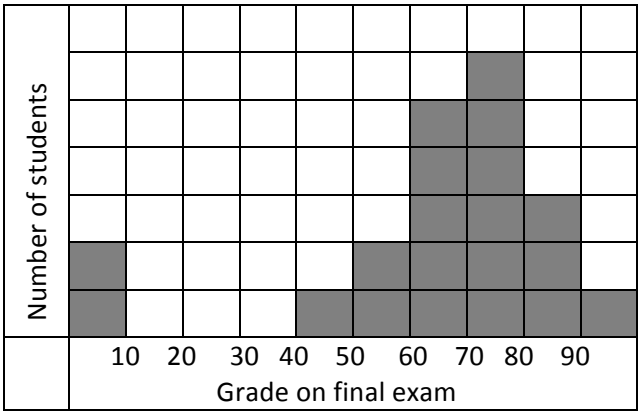


b. mean is 50



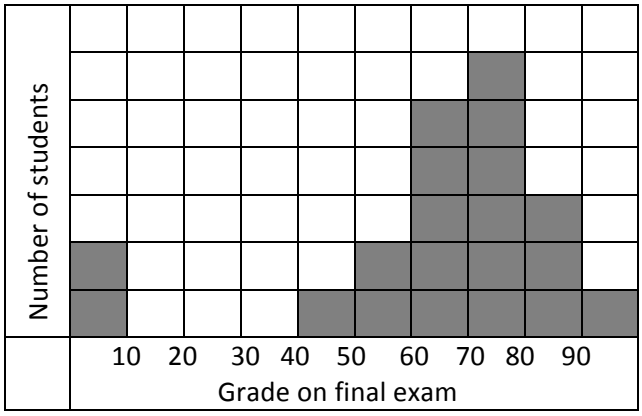
#10-#14 Color in a box on the graph that would cause the desired result. The mean for each graph is 64.

10.
Add a data point that would cause the mean to increase and the standard deviation to decrease.

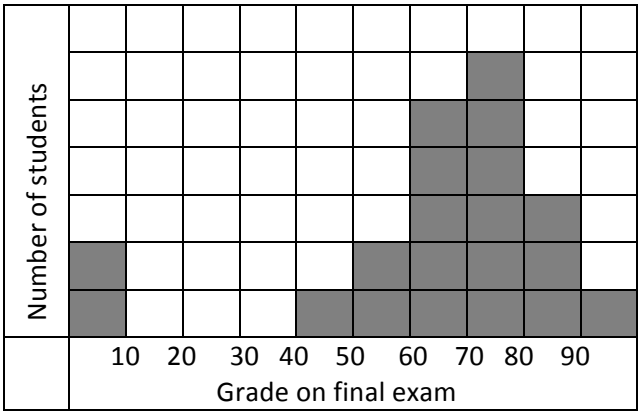


.

11.
Add a data point that would cause the mean to increase and the standard deviation to increase.

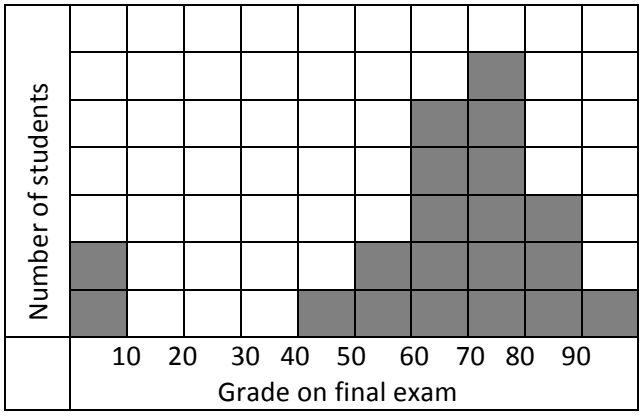


12.
Add a data point that would cause the mean to decrease and the standard deviation to decrease.

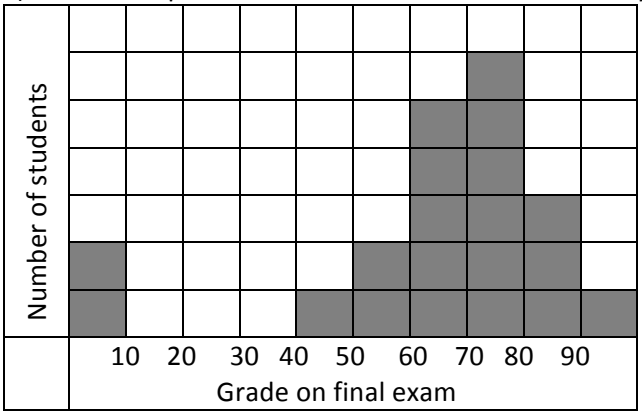


.

13.
Add a data point that would cause the mean to decrease and the standard deviation to increase.



14.
a) Add a data point that would cause the mean to stay the same.



b) Did the standard deviation increase, decrease or stay the same?

Calculator Practice Homework

1	$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$	$z_{\alpha/2} = 2.46$ p=0.856 n=45	$E =$
2	$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$	$t_{\alpha/2} = 1.043$ s=4.72 n=2096	$E =$
3	$L = \sqrt{\frac{(n-1)\cdot s^2}{\chi_R^2}}$	n=42 s=0.31 $\chi_R^2 = 19.24$	$L =$
4	$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$	p=0.702 $\hat{p} = 0.63$ n=481	$z =$
5	$t = \frac{\hat{x} - \mu}{\frac{s}{\sqrt{n}}}$	$\hat{x} = 74$ s=0.342 n=28	$t =$
6	$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$	n=531 s=23.93 $\sigma = 13.6$	$\chi^2 =$
7	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$	$\hat{p}_1 = 0.23$ $\hat{p}_2 = 0.24$ $\bar{p} = 0.236$ $n_1 = 69$ $n_2 = 72$ $z_{\alpha/2} = 2.957$	$z =$
8	$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$		$E =$
9	$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$\overline{x_1} = 34.2$ $\overline{x_2} = 35.1$ $s_1 = 54.2$ $s_2 = 50$ $n_1 = 346$ $n_2 = 553$ $t_{\alpha/2} = 2.97$	$t =$
10	$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$		$E =$
11	$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$	$\bar{d} = 2.08$ $s_d = 3$ n=57 $t_{\alpha/2} = 4.06$	$t =$
12	$E = t_{\alpha/2}\frac{s_d}{\sqrt{n}}$		$E =$

Comparing the standard deviation between data sets numerically

Data set A has a standard deviation of $4.27
Data set B has a standard deviation of 19 inches

1. Which data set has more variation?

2. What if data set A was price per gallon of gasoline in Holyoke, MA and data set B was the height of skyscrapers in downtown Boston? Which data set has more variation? Why?

3. What if data set A was the price of a beachfront condo unit in Hilton Head, SC and data set B was the height of kitchen ceilings for homes built in 1923? Which data set has more variation? Why?

To deal with the issue of context, statisticians decided they needed a way to compare the standard deviations of two totally different samples. So they figured out how to turn variation into a percentage.

4. Using percentages to compare different data sets is not a new idea. If you survey 20 people and find 5 love chocolate, and I survey 500 people and find 100 love chocolate, which sample has a higher percentage of chocolate lovers?

Comparing standard deviations from different contexts is done by taking the standard deviation as a percentage of the mean. This is called the **coefficient of variation**.

5. What if data set A was price per gallon of gasoline in Holyoke, MA and data set B was the height of skyscrapers in downtown Boston? Which data set has more variation? Why?

Average price of gasoline is $2.54 and the average height of skyscrapers is 495 ft (5940 inches)
(the mean and the standard deviation for these data sets are made up)

Find the coefficient of variation for each data set.

6. What if data set A was the price of a beachfront condo unit in Hilton Head, SC and data set B was the height of kitchen ceilings for homes built in 1923? Which data set has more variation? Why?

Average price of a beachfront condo in Hilton Head, SC is $565,000 and the average height of kitchen ceilings for homes built in 1923 is 7'3" (87 inches)
(the mean and the standard deviation for these data sets are made up)

Find the coefficient of variation for each data set.

Coefficient of Variation and Spreadsheet Homework

To answer questions #18-#21 Go to the link below (a direct link is on the Google site for this class)
https://docs.google.com/spreadsheets/d/1VFtaMB-WC6gdthZDBWPTljuxeIc7ZmReLfCsSortljM/edit?usp=sharing
Click on File, Save a Copy, edit and then Save this version. You can share this document with me by using the blue Share button, but you do not need to share this document with me unless you have questions about the data or what formula you used.
Google sheet formulas: Mean: =average(startcell:endcell) Median: =median(startcell:endcell) Mode: =mode(startcell:endcell) Standard Deviation: =stdev(startcell:endcell)
Data also provided on pages 43-44.

1. Find the mean for movie length and describe in a sentence what that means.

2. Find the median for movie length and describe in a sentence what that means.

3. Find the mode for movie length and describe in a sentence what that means.

4. Find the standard deviation for movie length.

5. For each of the four data sets provided, what percentage of the data points fall within 1 standard deviations? Within 2 standard deviations? NOTE: use the ordered lists on pages 43-44.

	Movie Length	Tobacco Use	Alcohol Use	NASA flight lengths
Percent of data points within 1 standard deviation of the mean				
Percent of data points within 2 standard deviations of the mean				

6. Looking at the percentages above

a) the majority of the data points fall within _____ deviation.

b) the VAST majority of data points fall within _____ deviations.

While these two statements are not true for absolutely every data set, they are true for most of them.

7. a) Find the coefficient of variation for the two data sets given.

	Alcohol Use in Children's movies	NASA flight lengths
Coefficient of variation		

b) Which shows the greatest variation?

c) Does that make sense looking at the data? Explain.

Movie	Company	Length (min)	Tobacco Use (sec)	Alcohol Use (sec)		Length in order	Tobacco Use in order	Alcohol Use in order
Snow White	Disney	83	0	0		64	0	0
Pinnocchio	Disney	88	223	80		64	0	0
Fantasia	Disney	120	0	0		69	0	0
Dumbo	Disney	64	176	88		70	0	0
Bambi	Disney	69	0	0		71	0	0
Three Caballeros	Disney	71	548	8		71	0	0
Fun and Fancy Free	Disney	76	0	4		71	0	0
Cinderella	Disney	74	37	0		72	0	0
Alice in Wonderland	Disney	75	158	0		73	0	0
Peter Pan	Disney	76	51	33		73	0	0
Lady and the Tramp	Disney	75	0	0		74	0	0
Sleeping Beauty	Disney	75	0	113		74	0	0
101 Dalmations	Disney	79	299	51		74	0	0
Sword and the Stone	Disney	80	37	20		74	0	0
Jungle Book	Disney	78	0	0		75	0	0
Aristocats	Disney	78	11	142		75	0	0
Robin Hood	Disney	83	0	39		75	0	0
Rescuers	Disney	77	0	0		75	0	0
Winnie the Pooh	Disney	71	0	0		75	0	0
Fox and the Hound	Disney	83	0	0		75	0	0
Black Cauldron	Disney	80	0	34		76	0	0
Great Mouse Detective	Disney	73	165	414		76	0	0
Oliver and Company	Disney	72	74	0		76	1	0
Little Mermaid	Disney	82	9	0		77	2	0
Rescuers Down Under	Disney	74	0	76		77	5	0
Beauty and the Beast	Disney	84	0	123		78	6	3
Aladdin	Disney	90	2	3		78	6	4
Lion King	Disney	89	0	0		79	9	5
Pocahontas	Disney	81	6	7		79	9	7
Toy Story	Disney	81	0	0		80	11	8
Hunchback of Notre Dame	Disney	90	23	46		80	17	13
James and the Giant Peach	Disney	79	206	38		81	23	20
Hercules	Disney	92	9	13		81	24	28
Secret of NIMH	MGM	82	0	0		81	37	33
All Dogs Go to Heaven	MGM	89	205	73		82	37	34
All Dogs Go to Heaven 2	MGM	82	162	72		82	51	38
Babes in Toyland	MGM	74	0	0		82	55	39
Thumbelina	Warner Bros	86	6	5		83	74	39
Troll in Central Park	Warner Bros	76	1	0		83	91	46
Space Jam	Warner Bros	81	117	0		83	117	51
Pippi Longstocking	Warner Bros	75	5	0		84	155	72
Cats Don't Dance	Warner Bros	75	91	0		86	158	73
An American Tail	Universal	77	155	74		88	162	74
Land Before Time	Universal	70	0	0		89	165	76
Fievel Goes West	Universal	75	24	28		89	176	80
We're Back: Dinosaur Story	Universal	64	55	0		90	205	88
Land Before Time 2	Universal	73	0	0		90	206	113
Balto	Universal	74	0	0		92	223	123
Once Upon a Forest	20th Century Fox	71	0	0		94	299	142
Anastasia	20th Century Fox	94	17	39		120	548	414
	mean		57.44	32.46			57.44	32.46
	st. dev. (s)		103.996	66.3399			103.996	66.3399

NASA Space Transport System - Shuttle Flight Duration in hours

Length in hours - chronological

54	236	259
54	175	309
192	143	307
169	222	285
122	239	306
120	239	285
146	236	283
145	336	262
247	259	259
191	174	332
167	335	259
144	269	330
197	353	382
191	262	333
73	269	306
167	262	
168	198	mean
169	399	220.1478261
190	235	
170	214	st. dev. (s)
97	260	85.96625648
168	381	
165	196	
146	214	
0	377	
97	221	
105	240	
119	405	
96	243	
121	423	
119	244	
120	239	
261	95	
106	221	
121	376	
98	284	
117	259	
215	376	
143	211	
199	381	
218	235	
213	213	
128	283	
166	235	
193	118	
214	191	
213	269	
331	237	
191	283	
190	309	

Length in hours - least to greatest

0	211	331
54	213	332
54	213	333
73	213	335
95	214	336
96	214	353
97	214	376
97	215	376
98	218	377
105	221	381
106	221	381
117	222	382
118	235	399
119	235	405
119	235	423
120	236	
120	236	
121	237	
121	239	
122	239	
128	239	
143	240	
143	243	
144	244	
145	247	
146	259	
146	259	
165	259	
166	259	
167	259	
167	260	
168	261	
168	262	
169	262	
169	262	
170	269	
174	269	
175	269	
190	283	
190	283	
191	283	
191	284	
191	285	
191	285	
192	306	
193	306	
196	307	
197	309	
198	309	
199	330	

Exploring Z-scores

60
60
60
64
64
64
64
68
68
68
68
68
72
72
72
72
72
72
72
72
76
76
76
76
76
76
80
80
80
80
80
80
88
88
88
88
88
96
104
124

The data to the left is a list of heart rates of *a sample of adults in Boston MA*. Looking at the data answer the following questions:

1. *In the city of Boston*,

a) Is an adult heart rate of 80 common or rare? Why?

b) Is an adult heart rate of 74 common or rare? Why?

c) Is an adult heart rate of 113 common or rare? Why?

d) Is an adult heart rate of 50 common or rare? Why?

2. In the city of Boston, which is more rare for an adult's heart rate 75 or 80? Why?

The mean (μ =76.3) and standard deviation (σ =12.50) of a sample of adult heart rates in Boston MA. Using only the mean and standard deviation answer the following questions:

3. In the city of Boston,

a) Is an adult heart rate of 80 common or rare? Why?

b) Is an adult heart rate of 74 common or rare? Why?

c) Is an adult heart rate of 113 common or rare? Why?

d) Is an adult heart rate of 50 common or rare? Why?

4. In the city of Boston, which is more rare for an adult's heart rate 75 or 80? Why?

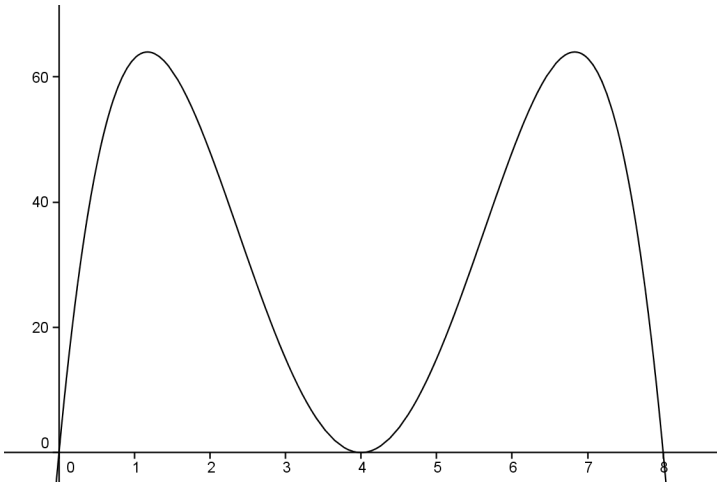
Let's come up with a way to quantify how rare or common a given value is.

5) Define quantify:

Hopefully in number four you talked about distance from the mean.
If the data is nice, then the further a value is from the mean the less likely it is.

If the data is not nice, then anything goes.

As an example:



The mean is 4

Close to the mean is actually less common than a little further away

(i.e. 4.2 is less common or more rare than 7)

So for the rest of the worksheet we will **assume nice data**.

We want to quantify commonness and rareness. Well, if in number four we talked about distance from the mean, what do we already have an concept that refers to distance from the mean and that is the standard deviation.

So let's find the commonness or rareness of a given value in terms of how many standard deviations they are away from the mean.

6. Given a mean of 45 and a standard deviation of 8, find the number of standard deviations each of the following points are from the mean. BE EXACT!

a) 53

b) 29

c) 57

d) 43

e) 63.26

7. a) For each of the values in number six, are they above or below the mean?

a	b	c	d	e
Above/Below	Above/Below	Above/Below	Above/Below	Above/Below

b) How might we show that a data point is above or below the mean using the standard deviation?

Congratulations, you have found some z-scores.

8. In your own words, define z-score

9. Write down the formula for finding a z-score

10. Check to see that the formula gives you the same answers as in #6 a) and #6 d)

Note: You do not need to use this formula to find z-scores. You can find the z-score however you wish, as long as it is correct.

12. Interpret the following z-scores

a) Kyle's height in centimeters compared to the height of all adults in Holyoke has a z-score of -1.2 What conclusion can you draw about Kyle's height?

b) Maria graduated with a 4.0 GPA from a very selective and challenging university. What would be a good approximation of the z-score associated with her GPA? Why?

c) Abdullah's salary has a z-score of 0.04 when compared to other police dispatchers around the country. What conclusion can you draw about Abdullah's salary?

Stats in the News

In class we are going to go over the article Regular marijuana users have more sex, study says, but for homework you are going to analyze the article Racism in Boston.

News Article Homework:

Read the article Racism in Boston in this packet. (pages 55-56)

On a separate sheet of paper answer the following questions.

1. What is the main point of the article? Or of what is the author trying to convince the reader?

2. For each of the statistics *that are in italics and underlined* in the article, what is exactly being said? Include who is included in the stat and what the statistic means in context, be very specific.

3. Where did they get the data? How was it gathered? What were the limitations to this method of gathering?

4. Do the statistics support the argument the author is trying to make?

5. What questions should be asked about this article and the statistics found within it?

6. In what ways can the conclusion or main point of this article be refuted or questioned?

Regular marijuana users have more sex, study says

By Jen Christensen, CNN Updated 8:21 AM ET, Fri October 27, 2017

Dr. Michael Eisenberg, an assistant professor of urology, sees a lot of patients at the Stanford University Medical Center who have problems performing in the bedroom. To determine what the problem is, they'll go through a laundry list of regular activities. Often, patients will ask whether they need to smoke less marijuana. There isn't a lot of research on the topic. However, with marijuana becoming legal in a growing number of states, Eisenberg thought it'd be worth exploring.

What he found surprised him. "Usually, people assume the more frequently you smoke, the worse it could be when it came to sex, but in fact, we learned the opposite was true," Eisenberg said. His study was published in this week's Journal of Sexual Medicine.

The study looked at data from the US government's National Survey of Family Growth. It asked more than 28,000 women and nearly 23,000 men how often they had sex in the four weeks prior to the survey and how frequently they used marijuana in the past year. Women who didn't use marijuana reported having sex six times on average during the past four weeks. Women who used marijuana daily had sex 7.1 times on average. The trend was similar for men. Men who abstained from marijuana said they had sex an average of 5.6 times in the four weeks before the survey, compared with the daily marijuana users who reported having sex 6.9 times, on average. "We were surprised to see the positive association between users," Eisenberg said. "This was across the board: marital status, race, none of that mattered." The study focused on heterosexual sex, and it didn't explain why there might be a connection between sex and marijuana. Eisenberg said past research on human and rodent models has shown that marijuana use may generally increase arousal. However, studies have also shown that too much marijuana use can decrease sperm count, and while men may want to have sex more, orgasm may be a challenge.

"It can have a different impact on different people," said Joseph Palamar, an associate professor in the Department of Population Health at New York University, who is not connected with the current study.

He thought it was a "cool epidemiological paper" that "did the best it could with the data," but it did have limitations. "It's unclear from the data if people had marijuana in their system before or during sex," Palamar

said. Someone could smoke in the morning but not have sex until the evening, when it wouldn't be in their system any more, for example. He added he'd like to see a study that could show more of a direct effect on frequency.

Palamar authored a small study comparing the sexual experience of people who are under the influence of alcohol versus marijuana. Studying 24 adults, his research found that people under either influence had increased feelings of self-attractiveness, but alcohol seemed to make people more social and bold and helped them make more connections with potential partners, compared with those people using marijuana. It showed that drinkers typically have more regrets about who they slept with and are less choosey, whereas marijuana users tended to be more selective.

Because marijuana is still illegal in the majority of places, Palamar found that most people have to smoke in private, and that could lead to more opportunities to initiate intimacy, compared with people who drink, since alcohol is everywhere.

Marijuana may also have increased some people's sensitivity during the act itself, although some reported getting so "lost in their own heads," they weren't paying as much attention to their partners, and they did not enjoy sex as much.

"And if marijuana makes you paranoid, as it does with some people, it could really, pardon the pun, screw your ability to have an orgasm," Palamar said. Some women also reported vaginal dryness when they smoked pot, and that too can limit sexual pleasure. Both scholars hoped these studies will encourage other researchers to dive deeper into the topic. In the meantime, Eisenberg said that if a patient asks whether his frequent marijuana use is getting in the way of his sex life, he will tell them that "it may not be the culprit."

Regular marijuana use can have other impacts on your health. Research in adults is still limited, but what we know is that smoking can irritate your lungs, and studies have shown it can raise your heart rate, making you more vulnerable to a heart attack. "For most people, we tell them instead to go to the gym and lose 20 pounds," Eisenberg said. Being overweight can give men arousal problems. "We always talk about anything that can be good for your heart can be good for your penis," he said. "For a lot of guys, hearing that is an amazing motivator."

America's middle class: Poorer than you think

http://money.cnn.com/2014/06/11/news/economy/middle-class-wealth/index.html
by Tami Luhby @Luhby August 5, 2014: 3:25 PM ET

Rich Americans. That's our global reputation. The numbers seem to back it up. Americans' average wealth tops $301,000 per adult, enough to rank us fourth on the latest Credit Suisse Global Wealth report. But that figure doesn't tell you how the middle class American is doing. Americans' median wealth is a mere $44,900 per adult -- half have more, half have less. That's only good enough for 19th place, below Japan, Canada, Australia and much of Western Europe. "Americans tend to think of their middle class as being the richest in the world, but it turns out, in terms of wealth, they rank fairly low among major industrialized countries," said Edward Wolff, a New York University economics professor who studies net worth.

Why is there such a big difference between the two measures?

Super rich Americans skew average wealth upwards. The U.S. has 42% of the world's millionaires, and 49% of those with more than $50 million in assets. This schism secures us the top rank in one net worth measure -- wealth inequality. There's one main reason why the average Spaniard or Italian has more to his name than the typical American: real estate. Home ownership rates are higher in many European countries than in the U.S., giving Joe European more assets to his name than his American counterpart. Plus, it's easier for Americans to borrow money, which eats away at their net worth, said Jim Davies, an economics professor at Western University in Ontario, Canada, and co-author of the Credit Suisse report. Middle class Americans were also hurt greatly by the housing collapse at the end of the last decade. The median wealth of families was $77,300 in 2010, a nearly 40% drop from 2007, according to Federal Reserve statistics. "Changes in home prices have a big effect on the wealth in the middle," Davies said.

Middle class Australians, by comparison, are leading the pack. The country's residents have the highest median net worth, coming in at $219,500. Australia also has low wealth inequality. This is in part because Australians have a strong tradition of home ownership, though escalating prices have made it tougher for young adults to secure the Australian Dream. Those down under also have a mandatory retirement savings program, where they must squirrel away more than 9% of their income for their Golden Years, and they carry relatively low credit card and student loan debt. Americans, meanwhile, are having trouble building wealth because wages have stagnated for more than a decade. Median household income was $51,017 in 2012, compared to $56,080 in 1999, according to the Census Bureau's most recent statistics. There are many reasons why middle class incomes are suffering, including the decline of unions' power, the shift of jobs overseas and the increasing use of technology in the workplace, said Kenneth Thomas, professor of political science at University of Missouri, St. Louis. Also, Americans have to pay more out of pocket for basics, such as health care and higher education, reducing their ability to build their nest egg. "Middle class families haven't been able to save anything," Wolff said.

Racism in Boston: African-Americans Have a Median Net Worth of $8, New Report Shows

By Carlos Ballesteros On 12/11/17 at 2:22 PM

http://www.newsweek.com/boston-african-americans-poverty-wealth-inequality-744108

Boston's reputation as a racist city was put to the test by The Boston Globe's Spotlight investigative team of reporters. They analyzed data, surveys and conducted hundreds of interviews to determine whether the reputation is warranted.

The first part of their investigation was released Sunday. Spotlight concluded that despite important stopgaps, racism is alive and well in the cradle of liberty.

"Here in Boston, a city known as a liberal bastion, we have deluded ourselves into believing we've made more progress than we have," the article reads. "Racism certainly is not as loud and violent as it once was, and the city overall is a more tolerant place. But inequities of wealth and power persist, and racist attitudes remain powerful, even if in more subtle forms. They affect what we do—and what we don't do."

One of Spotlight's most daunting findings is the economic disparity between whites and blacks in Boston. Using data from the Federal Reserve of Boston, Spotlight found that non-immigrant African-Americans in the Boston area had a *median net worth of $8*. "That means they owe almost as much as the combined value of what they own, be it a car, or house, or savings," Spotlight writes.

White households in Boston, on the other hand, *average a net worth of $247,500*, or nearly 31,000 times more than African-American Bostonians.

Anti-black racism in Boston has been studied and quantified for decades. Tens of thousands of black migrants and refugees escaping Jim Crow in the South during the early 20th century made their way to Boston, where they were often met with hostility from the city's white population.

These tensions climaxed in the 1970s when the city implemented a school busing program to lower the levels of segregation in Boston's public schools. Whites protested the program and oftentimes prevented black students from entering predominantly white schools.

In the 40 years since the busing crisis, overt racism isn't as visible in Boston. But Spotlight's reporting shows that systemic racism still prevents blacks in Boston from achieving equitable levels of economic prosperity.

The team compared their latest investigation with a series of articles from 1983 written by Globe reporters at the time. They found that the number of black officials and managers in Boston had only risen by 0.1 percent in nearly a quarter-century. Spotlight also found that, just as in 1983, blacks in Boston have *double the rate of unemployment* than white workers.

For James Jennings, professor emeritus of race, politics and urban policy at Tufts University, these numbers show just how far Boston has to go.

“A lot of times when Boston engages in looking at itself around race, it focuses on attitudes and prejudices,” Jennings told the Globe. "With that, Boston certainly has made a lot of progress, but Boston needs to start looking at structural inequality—racial hierarchy, poverty, academic achievement—to move the needle forward.”

Spotlight's investigation comes at a time when poverty in America has gained newfound attention.

A special envoy of the United Nations is currently investigating poverty and human rights abuses in the United States. Last week, the U.N.'s lead investigator said that parts of rural Alabama displayed some of the worst levels of poverty in the developed world.

Furthermore, a new report by the People's Policy Project, a left-leaning economic think tank, analyzing data from the Survey of Consumer Finances, found that black wealth after the Great Recession and during the Obama administration dissipated.

"Between 2007 and 2016, *the average wealth of the bottom 99 percent decreased by $4,500*. This decline was particularly concentrated among the housing wealth of African-Americans. Outside of home equity, black wealth recovered its 2007 level by 2016. But *average black home equity was still $16,700 less*. Meanwhile, over the same period, the *average wealth of the top 1% increased by $4.9 million*," the report concluded.

Basic Probability

We are going to discuss three types of probability: fair games, past performance and survey data.

Fair games (theoretical probability)

1. What is the probability that when I flip a coin, I will get heads?	2. What is the probability that when I flip a coin 3 times, I will get exactly 2 heads?
3. What is the probability that when I flip a coin 4 times, I will get at least 3 heads?	4. What is the probability that when I flip a coin 4 times, I will get at most 1 head?
5. If I draw 1 card from a standard deck of cards, what is the probability that the card is a jack?	6. What is the probability that I flip a coin and get heads, then draw a card and get a jack?
7. What is the probability that I draw a jack and then without putting that card back, draw a 7?	8. What is the probability that I draw a jack and then without replacement draw a red?

Past performance (predictive probability)

Casey got A's on 3 of his last 4 tests.

9. What is the probability he will make an A on the next one?

10. What is the probability that he gets an A on the next two tests?

In the last box of clocks, 13 out of 100 were defective.

11. What is the probability that the next clock I choose will be defective?

12. What is the probability that the next two clocks are defective?

Women have about a 12% chance of being diagnosed with breast cancer whereas 80% of women with BRCA1 or BRCA2 will be diagnosed in their lifetime.

13. What is the difference?

Survey Data (predictive probability) Note: this data is made up

Average daily screen time	Less than 2 hours (D)	2 to 6 hours (E)	Over 6 hours (F)
Under 18 (A)	5	6	4
18-24 (B)	5	3	12
25 and over (C)	0	8	7

14.

P(A)	P($\bar{C}$)
P(A or B)	P($\overline{C \text{ or } F}$)
P(A or E)	P(E|B)
P(A and F)	P(C|F)
P(A and C)	P(E|F)

15. Describe P(A) in a sentence

16. Describe P(E|B) in a sentence

17. Describe P(A or E) in a sentence

The Law of Large Numbers states that, as the number of trials in an experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero. In other words, the theoretical probability and the predictive probability will get closer and closer together. When applied to survey data, the law of large numbers stats that's as the sample size increases and gets closer and closer to the population size, the characteristics of the sample will approach the characteristics of the population.

Probability homework

The game for #1-#10 is to roll two six sided dice, one red and one blue, then find the sum.
I recommend writing out the sample space for this game, maybe using the table provided.

		Red die					
		1	2	3	4	5	6
Blue die	1						
	2						
	3						
	4						
	5						
	6						

1. P(red=1)

2. P(red=1 and blue=2)

3. P(red=1 or blue=2)

4. P(red=3 | blue=2)

5. What sum has the highest probability?

6. P(sum=4)

7. P(red=1 or sum=5)

8. P(red=1 and sum=5)

9. P(red=1 | sum=5)

10. P(sum=5 | red=1)

11. Rory made 7 of her last 10 free throw shots.
a) Will she make the next free throw shot she takes?

b) What is the probability that she makes the next shot?

c) What is the probability that she makes the next 5 shots?

#12-#19 Use the survey data below

Smoking_Cat	Weight_Cat			
	Underweight	Normal	Overweight	Total
Non-smoker	53	543	1903	2499
Light (1-5)	32	171	375	578
Moderate (6-15)	35	234	306	575
Heavy (16-25)	47	381	618	1046
Very Heavy (> 25)	11	133	325	469
Total	178	1462	3527	5167

Heart data set that comes with SAS statistical software package.

12. P(non-smoker)	16. P($\overline{non-smoker}$)
13. P(light smoker and normal weight)	17. P(very heavy smoker | normal weight)
14. P(heavy smoker and light smoker)	18. P(overweight | light smoker)
15. P(moderate smoker or overweight)	19. P($\overline{non-smoker\ or\ normal\ weight}$)

Probability Density Functions

Discrete variables with a "reasonable" number of outcomes

It is a display of each outcome and the probability associated with each outcome.

The net winning/losing of playing a carnival game

Outcome	Probabil ty
-5	0.2
-3	0.1
-2	0.05
0	0.15
1	0.1
3	0.2
4	0.05
6	0.1
9	0.05

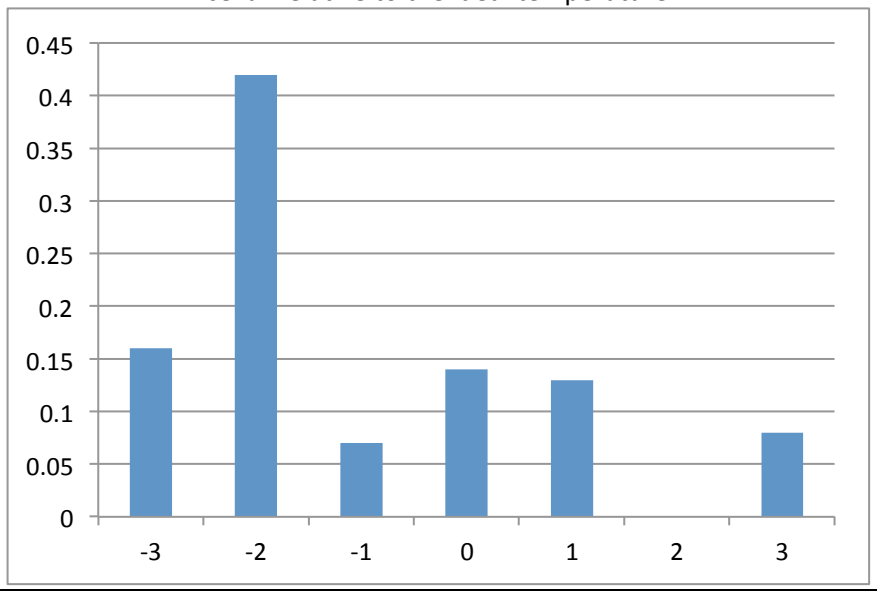
1. Create a histogram to display the data in the table (note the scale)

2. Fill in the table to display the information in the histogram

Degrees away from ideal	Probability

The ideal temperature for a red wine is 52.5°

This is the distribution for the temperature of the red wine in my wine cellar relative to the ideal temperature



3. The sum of all the probabilities must be _____.

4. Each probability must lie between _____and_____.

5. What is the probability you lose money playing the carnival game?

6. If my snobby wine friend is coming for dinner, what is the probability that the red wine I serve is within one degree of ideal?

Continuous variables or discrete variables with "too many" outcomes

Key: Think intervals!!

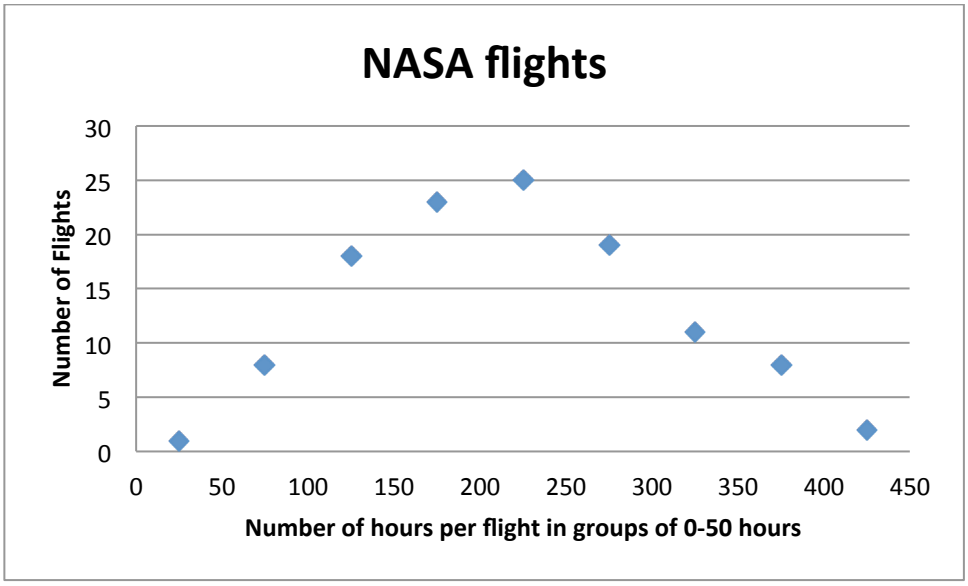
First way: Create a Histograms using intervals (very similar to previous examples)

Fall 2017 HCC enrollment data

Age	Percent of students as probability
Under 20	0.32
20-24	0.35
25-29	0.13
30-44	0.14
45 and over	0.06

7. Create a histogram to display the data in the table

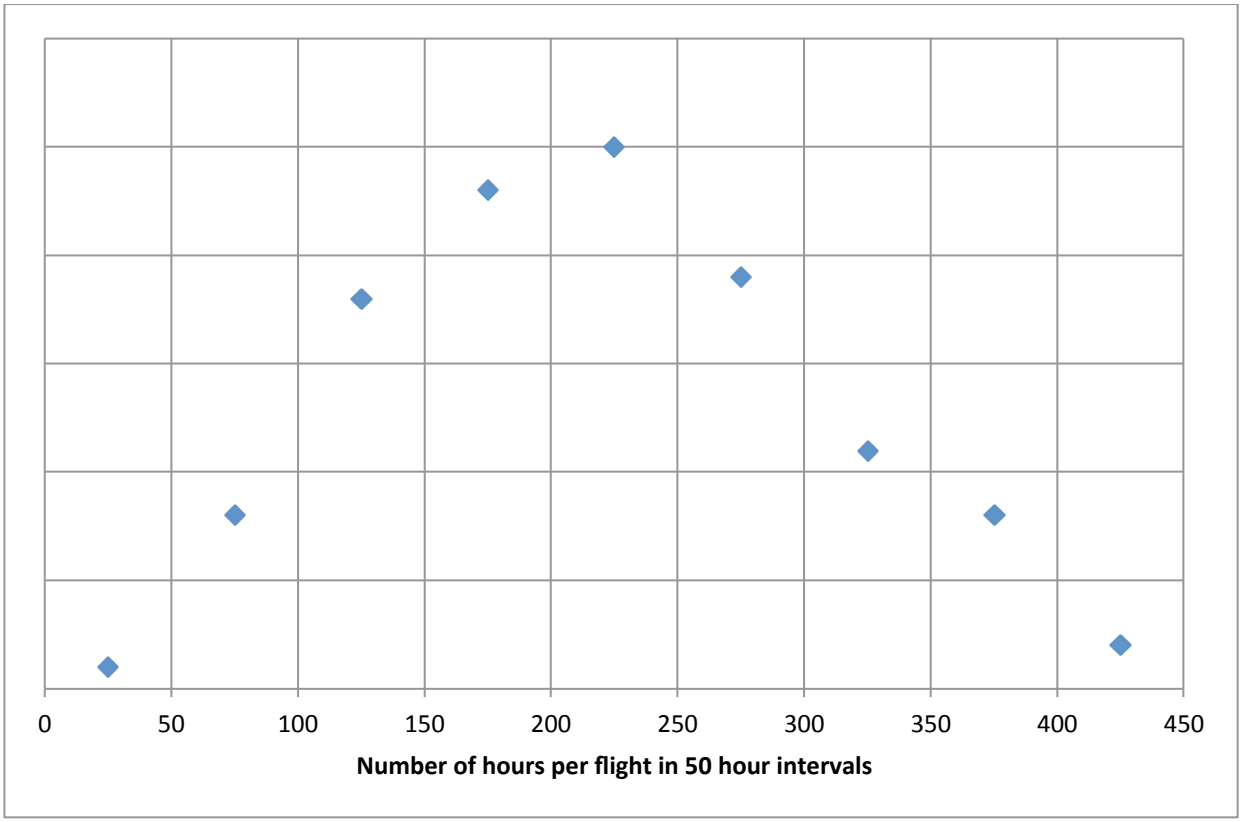
Second way: Plot the data in a scatter plot using intervals and approximate a smooth curve to fit the data



8. Draw a smooth curve on the above scatter plot to approximate all the data on a continuous scale. Be as close as you can to all data points, but you do not need to go through every data point.

9. This could be a probability density function if we change the scale so that one very important thing is true. With the histograms, the sum of the probabilities has to be 1. That same principle shows up with smooth curves by making *the area under the curve* equal to 1. Important: when make the area 1, the y-axis scale and label do not make sense, so cross them out!

Example:



This is again the NASA flight duration data.
Let's figure out how to turn this into a probability density function.

10. Draw a curve for the flight data.

11. Think of this as a probability density function (use the information from #9)
Notice we lost the y-axis label and scale. Now think about how to use this graph to answer the questions 12-14.

This will require some thought and wrong turns. Call me over and ask questions to help you and to make sure you have the right idea when your group thinks it knows how to proceed.
Don't give up on me here…

12. What is the probability that a flight will last between 50 and 75 hours?

13. What is the probability that a flight will last between 150 and 250 hours?

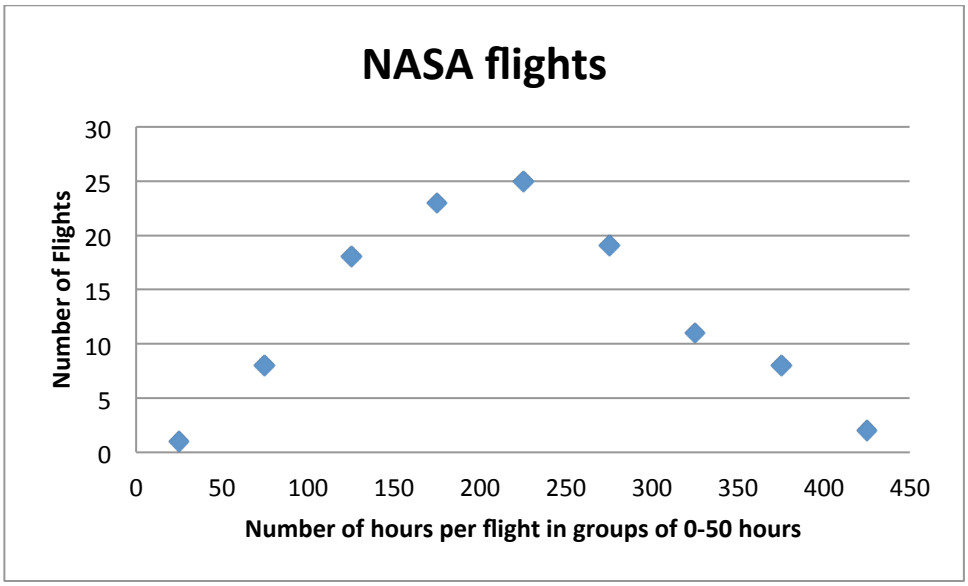
14. What is the probability that a flight will last between 0 and 450 hours?
This is why we have the rule stated in #9 of this worksheet

Note: not all sets of data will lend themselves to easy curve fitting.

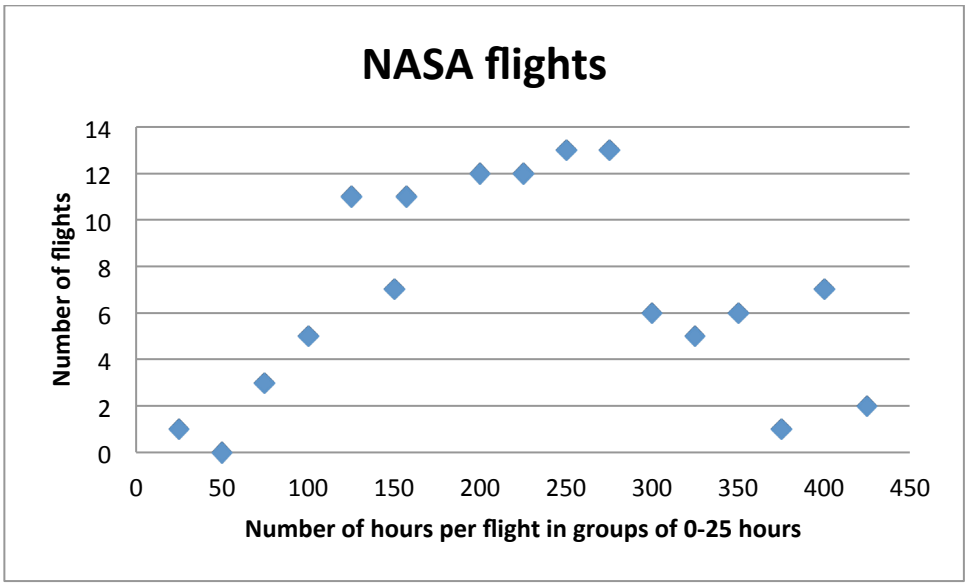
At some point, you might learn more about curve fitting and doing it with the help of technology, but for now, we will do it by hand and do the best we can.

Remember, smooth curves are best.

Example where changing the intervals changes the graph:

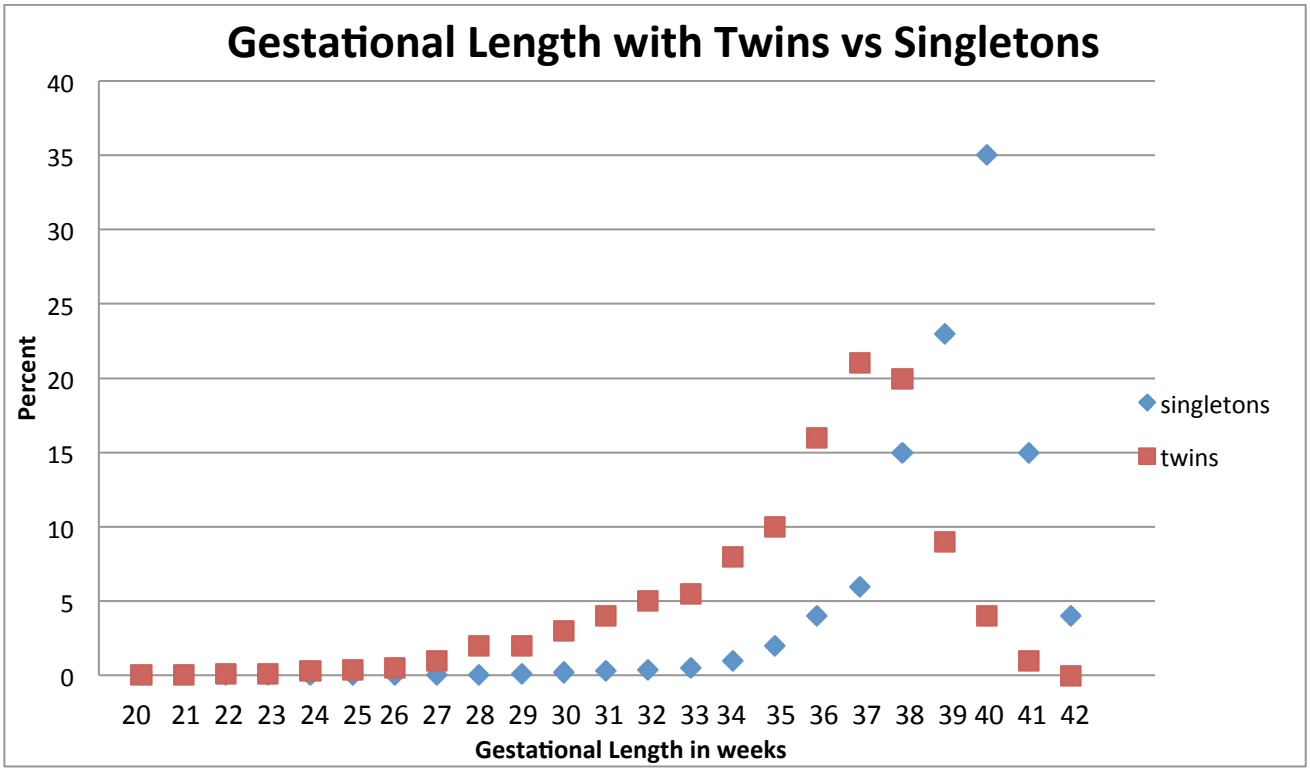


This one below is much harder to fit, and it is the same data. It doesn't mean that one graph is better than the other in all situations. It just means the data may not be quite as smooth as the above graph implies.



15. Fit a curve to the data above. Note: you will not hit every point, just get the best fit you can.

Example:



16. Fit two curves to the above graph- one for each type of pregnancy

Let's figure out how to turn this into a probability density function. Let's focus on the twins curve.

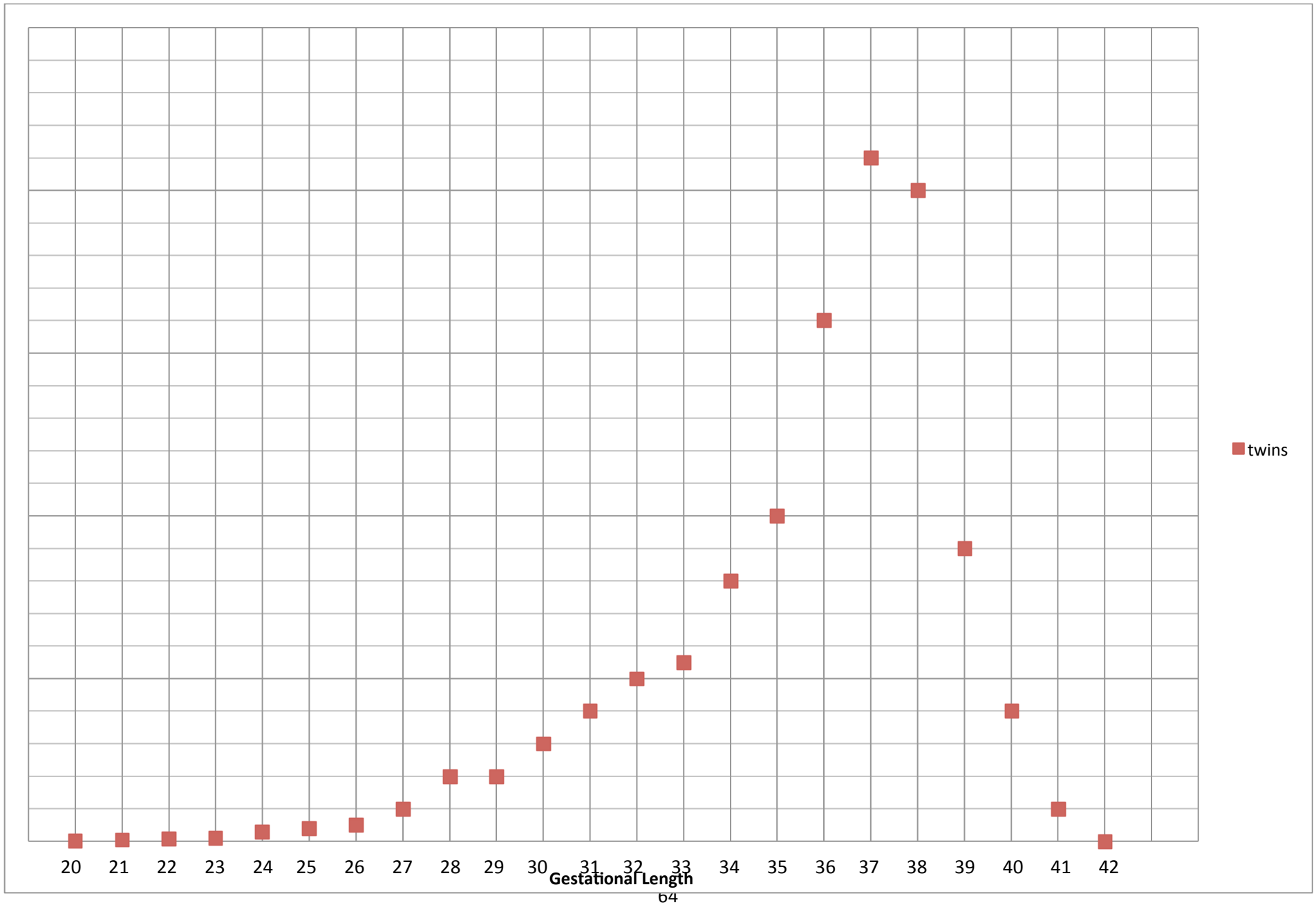
Use the copy of this graph that is on graph paper on the next page.

17. Draw a curve for the twins data only. Ignore the singleton data.

18. Think of this as a probability density function Look back at #11 if you get stuck

19. What is the probability a women pregnant with twins will deliver between 30.4 weeks and 33.6 weeks?

20. What is the probability a women pregnant with twins will deliver between 39 and 50 weeks?


20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
Gestational Length
twins