Freeport AP Statistics

Chapter 1: EXPLORING DATA

Intro Data Analysis: Making Sense of Data

DAY 1

Objectives for this section:

- ***Identify the individuals and variables of a set of data***
- ***Classify variables as categorical or quantitative***
- ***Identify the units of measurement for a quantitative variable.***

Data Analysis: Learning from the data; what is the data trying to communicate?

- How do we anayze?
  - Organize
  - Display
  - Summarize
  - Ask questions

Individuals –

Variable –

Key questions to ask when you encounter a set of data:

1) Who/what are the ________________ that are described by this data?
2) How many individuals are under study here?
3) What variables are present?
4) What are the units of these variables?
5) Do I need to do anything with units to make the data comparable?

*Other questions will arise later, but this is a good start.

There are different types of variables. We must be able to classify these variables properly.

- Proper classification leads us to the correct method of data analysis.
- Proper classification tells us what can be done with the data.

Classifications of Variables

Variables in a statistical study can be classified as either:

- _________________________ (_____________________) or
- _________________________ (_____________________)

Categorical Variable –

Quantitative Variable –

1. Quantitative variables are not always numerical. Give an example of a numerical data set that is in fact categorical.

Distribution –

2. **HIRING DISCRIMINATION – IT JUST WON'T FLY**

An airline has just finished training 25 pilots – 15 male and 10 female – to become captains. Unfortunately, only eight captain positions are available right now. Airline managers announce that they will use a lottery to determine which pilots will fill the available positions. The names of all 25 pilots will be written on identical slips of paper, which will be placed in a hat, mixed thoroughly, and drawn out one at a time until all eight captains have been identified.

A day later, managers announce the results of the lottery. Of the 8 captains chosen, 5 are female and 3 are male. Some of the male pilots who weren't selected suspect that the lottery was not carried out fairly. One of these pilots asks your statistics class for advice about whether to file a grievance with the pilots' union.

Could these results have happened just by chance?

a. We are going to use a deck of cards to perform a ___________________.

b. How many cards in the deck are we going to use? ___________________

c. What are the cards you are using going to represent?

d. How many cards do we select? When do we stop selecting cards?

e. Perform 5 trials and record the number of females chosen in each trial:

Trial	**1**	**2**	**3**	**4**	**5**
No. of females					

f. Let’s graph the results:

0 1 2 3 4 5 6 7 8 9 10 11

g. Does the lottery appear to be fair? What advice would you give the male pilot who contacted you?

3. Categorize each variable as either categorical or quantitative.
   a. Gender (female or male)

   b. Age (years)

   c. Color of M&M's

   d. Weight (in oz) of a bag of potato chips

   e. Zip Code

   f. Concentration of contaminants in water (in ppm – parts per million)

   g. Another name for categorical is _______________________. Another name for quantitative is ____________________.

4. Many people like to ride roller coasters. Amusement parks try to increase attendance by building exciting new coasters. The table below displays data on several roller coasters that were opened in 2009.

Roller Coaster	**Type**	**Height (ft)**	**Design**	**Speed (mph)**	**Duration (s)**
Wild Mouse	Steel	49.3	Sit Down	28	70
Terminator	Wood	95	Sit Down	50.1	180
Manta	Steel	140	Flying	56	155
Prowler	Wood	102.3	Sit Down	51.2	150
Diamondback	Steel	230	Sit Down	80	180

   a. What individuals does this data set describe?

   b. Clearly identify each of the variables. Which are quantitative? In what units are they measured?

   c. Describe the individual in the highlighted row?

5. You are preparing to study the television-viewing habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student. Give the units of measurement for the quantitative variables

Intro wrap-up:

Define Statistics:

**The entire group to be studied is called the ______________________________. An ________________________ is a ____________________ or ____________________ that is a member of the _____________________________ being studied. A ________________________ is a ____________________ of the population.

Freeport AP Statistics

Chapter 1: EXPLORING DATA

1.1 Analyzing Categorical Data

DAY 2

OBJECTIVE(S):

- Students will learn how to construct a bar graph of the distribution of a categorical variable or, in general, to compare related quantities.
- Students will learn to recognize when a pie chart can and cannot be used.
- Students will be able to identify what makes some graphs deceptive.
- Students will be able to answer questions involving marginal and conditional distributions from a two-way table of counts.
- Students will learn how to describe the relationship between two categorical variables by computing appropriate conditional distributions.
- Students will learn how to construct a bar graph to display the relationship between two categorical variables

Ex. A bag of M&M's contains 4 red, 8 green, 7 blue, 6 brown, 9 yellow, and 1 orange.

Frequency Table **– displays a _________________ (frequency) of the data**

Color	Frequency

Relative Frequency Table **– displays a ________________________ (rel. freq.) of the data.**

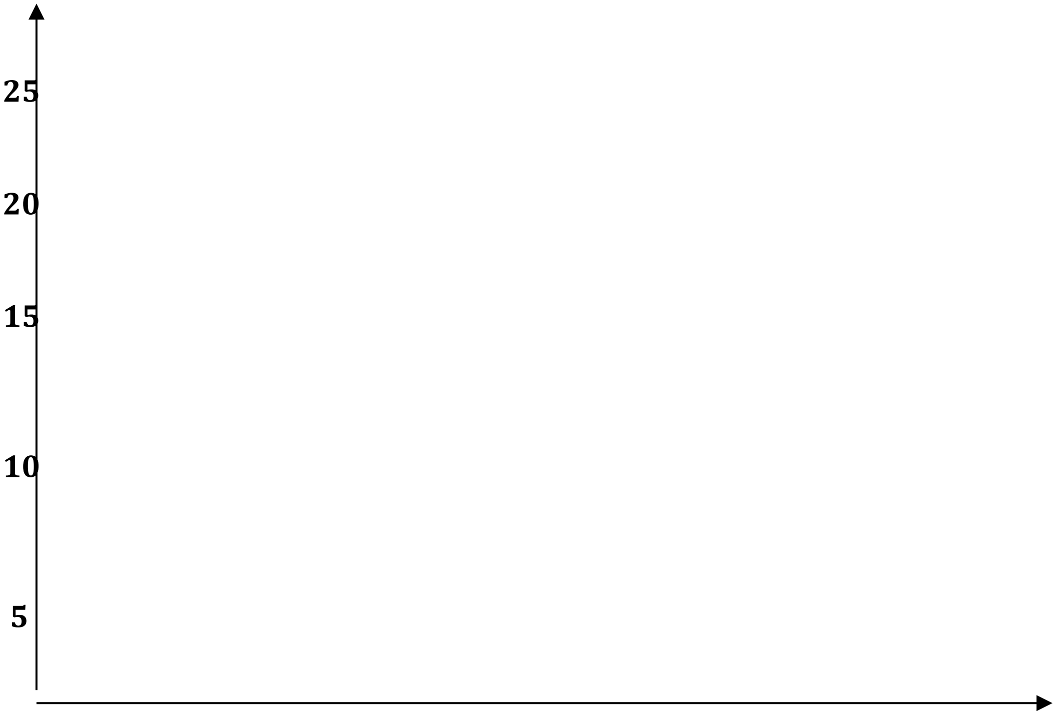
Color	Rel. Freq.

6. Email spam is the curse of the Internet. Here is a compilation of the most common types of spam?

Type of spam	Percent
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

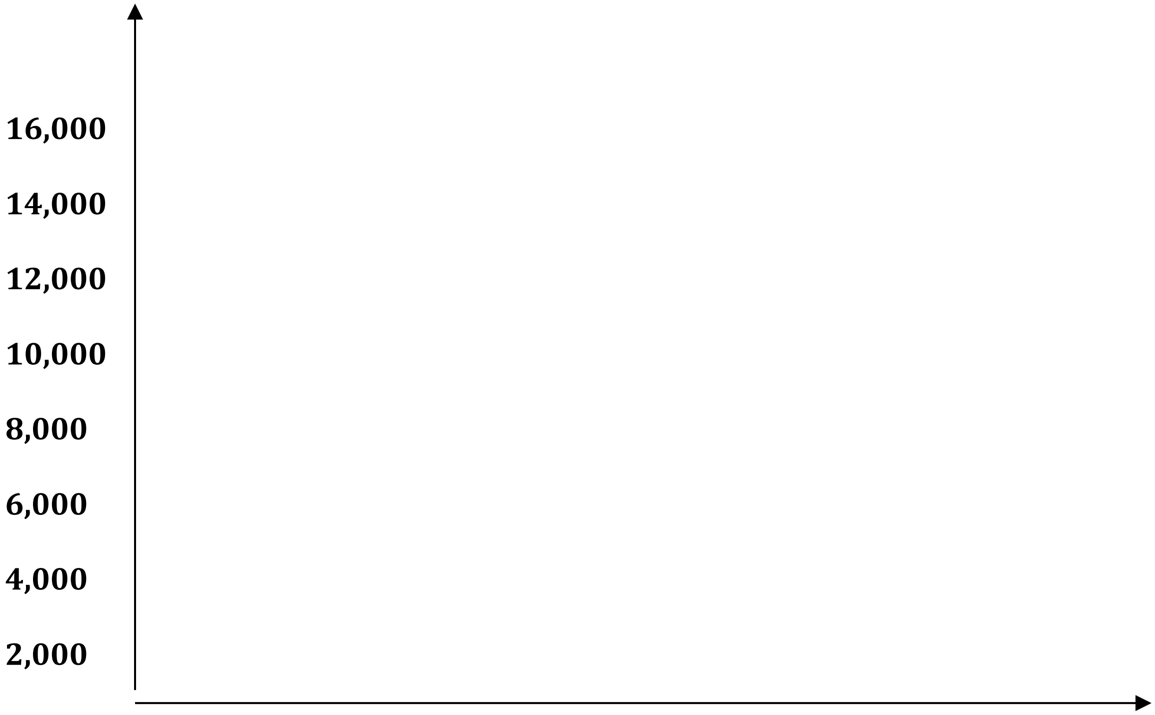
a. What percent of spam would fall in the "Other" category?

b. Display these data in a bar graph. Be sure to label your axes and title your graph.



c. Would it be appropriate to make a pie chart of these data? Explain.

7. Among persons aged 15 to 24 years in the United States, the leading causes of death and number of deaths in a recent year were as follows: accidents, 15,567; homicide, 5359; suicide, 4139; cancer, 1717; heart disease, 1067; congenital defects, 483.
    a. Make a bar graph to display these data.



    b. To make a pie chart, you need one additional piece of information. What is it?

8. **TEXTBOOK p. 21 #14**

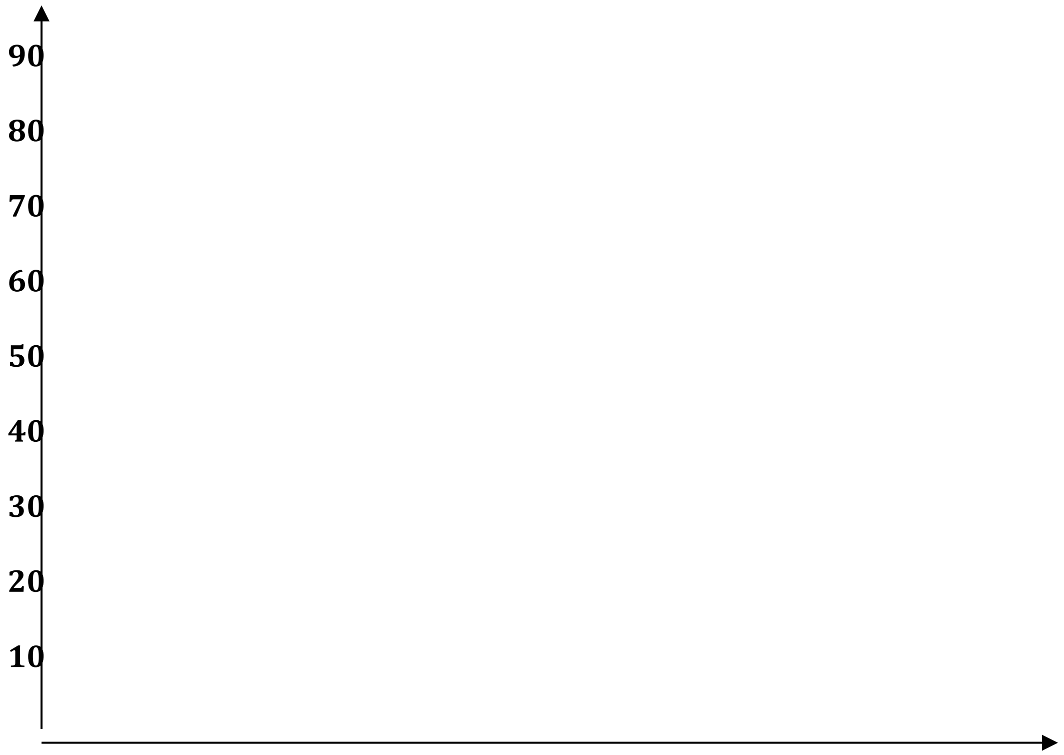
    Business ____

    Social Science ____

9. Here are data on the percent of people in several age groups who attended a movie in the past 12 months:

Age group	**Movie attendance**
18 to 24 years	83%
25 to 34 years	73%
35 to 44 years	68%
45 to 54 years	60%
55 to 64 years	47%
65 to 74 years	32%
75 years and over	20%

a. Display these data in a bar graph? Describe the relationship between the two variables.



b. Would it be correct to make a pie chart of these data? Why or why not?

c. A movie studio wants to know what percent of the total audience for movies is 18 to 24 years old. Explain why these data do not answer this question

DAY 3

Two-way Table –

Marginal Distribution –

Conditional Distribution –

Segmented Bar Graph –

***Association* -**

10. Here are data from a survey conducted at eight high schools on smoking among students and their parents:

	Neither Parent Smokes	**One Parent Smokes**	**Both Parents Smoke**	
Student does not smoke	1168	1823	1380	
Student smokes	188	416	400	

a. How many students are described in the two-way table? What percent of these students smoke?

b. Give the marginal distributions of parents' smoking behavior, both in counts and in percents.

	Neither Parent Smokes	**One Parent Smokes**	**Both Parents Smoke**	
Student does not smoke	1168	1823	1380	
Student smokes	188	416	400	

c. Calculate 6 conditional distributions of students' smoking behavior: one for each of the three parental smoking categories. Describe the relationship between the smoking behaviors of students and their parents in a few sentences.

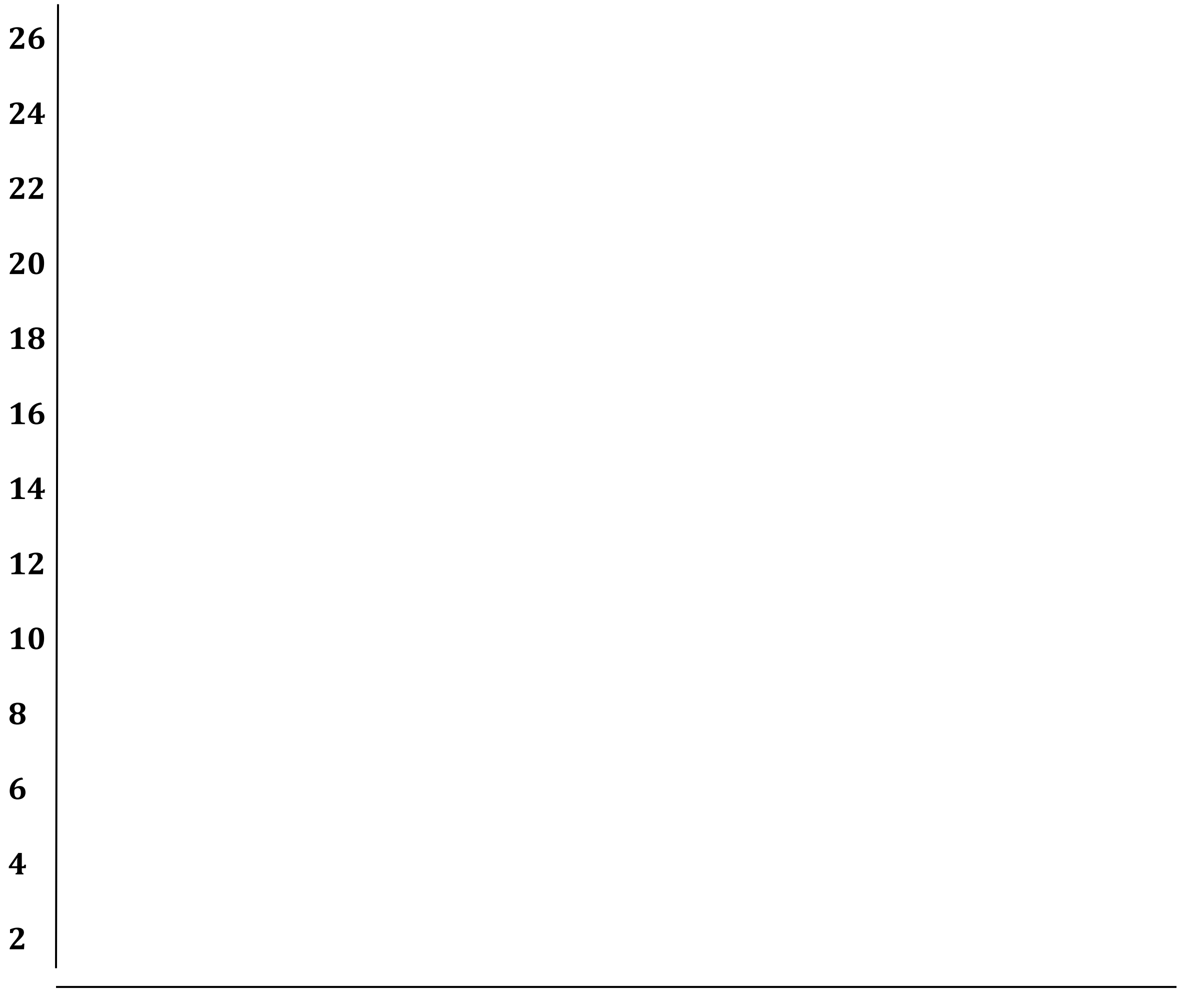
	Neither Parent Smokes	**One Parent Smokes**	**Both Parents Smoke**	
Student does not smoke	1168	1823	1380	
Student smokes	188	416	400	

11.

Favorite vehicle colors may differ among types of vehicle. Here are data on the most popular colors in 2008 for luxury cars and for SUVs, trucks, and vans.

Color	**Luxury cars (%)**	**SUVs, trucks, vans (%)**
Black	22	13
Silver	16	16
White pearl	14	1
Gray	12	13
White	11	25
Blue	7	10
Red	7	11
Yellow/gold	6	1
Green	3	4
Beige/brown	2	6

a. Make a graph to compare colors by vehicle type.



b. Write a few sentences comparing the color preference between Luxury cars and SUV’s, trucks, and vans.

12. People who get angry easily tend to have more heart disease. That's the conclusion of a study that followed a random sample of 12,986 people from three locations for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data for the 8474 people in the sample who had normal blood pressure. CHD stands for "coronary heart disease." This includes people who had heart attacks and those who needed medical treatment for heart disease.

	Low anger	**Moderate anger**	**High anger**	**Total**
CHD	53	110	27	190
No CHD	3057	4621	606	8284
Total	3110	4731	633	8474

Do these data support the study's conclusion about the relationship between anger and heart disease?

Freeport AP Statistics

Chapter 1: EXPLORING DATA

1.2 Displaying Quantitative Data With Graphs

DAY 4

OBJECTIVE(S):

- Students will learn how to construct a dotplot or stemplot to display small sets of data.
- Students will learn how to describe the overall pattern (shape, center, and spread) of a distribution and identify any major departures from the pattern (like outliers).
- Students will learn how to make a histogram with a reasonable choice of classes.
- Students will learn how to identify the shape of a distribution from a dotplot, stemplot, or histogram as roughly symmetric or skewed and identify the number of modes.
- Students will learn how to interpret histograms.

13. What does it mean that our data is **UNIVARIATE** (as opposed to **BIVARIATE**)?

When examining **UNIVARIATE** data, you need to address **SOCS**.

S

Symmetric –

Skewed to the right –

Skewed to the left –

Bimodal –

O

C

S

14. THE GAME OF GREED

DAY 4

15. The following table displays the total number of gold medals won by a sample of countries in the 2008 Summer Olympic Games in China.

Country	Gold Medals	Country	Gold Medals
Sri Lanka	0	Thailand	2
China	51	Kuwait	0
Vietnam	0	Bahamas	0
Great Britain	19	Kenya	5
Norway	3	Trinidad and Tobago	0
Romania	4	Greece	0
Switzerland	2	Mozambique	0
Armenia	0	Kazakhstan	2
Netherlands	7	Denmark	2
India	0	Latvia	1
Georgia	3	Czech Republic	3
Kyrgyzstan	0	Hungary	3
Costa Rica	0	Sweden	0
Brazil	3	Uruguay	0
Uzbekistan	1	United States	36

a. What are the individuals in this sample?

b. How many variables are we measuring for each individual?________
Ergo, this is ________________________________ data.

c. Make a dotplot to display the data.

d. Describe the distribution of Gold Medals.

e. Overall, 204 countries participated in the 2008 Summer Olympics, of which 55 won at least one gold medal. Do you believe that the sample of countries listed in the table is representative of this larger population? Why or why not?

16. **TEXTBOOK p. 41 #40**

17. **TEXTBOOK p. 44 #44**

18. Here are the scores of games played in the California Division I-AAA high school basketball playoffs:

71-38	52-47	55-53	76-65	77-63	65-63	68-54	64-62
87-47	64-56	78-64	58-51	97-74	71-41	67-62	106-46

On the same day, the final scores of games in Division V-AA were

98-45	67-44	74-60	96-54	92-72	93-46
98-67	62-37	37-36	69-44	86-66	66-58

a. Construct a back-to-back stemplot to compare the points scored by the 32 teams in Division I-AAA playoffs and the 24 teams in the Division V-AA playoffs.

b. Write a few sentences comparing the two distributions.

19. In 1846, a published paper provided chest measurements (in inches) of 5738 Scottish militiamen. The below summarizes the data.

Chest size	Count	Chest size	Count
33	3	41	934
34	18	42	658
35	81	43	370
36	185	44	92
37	420	45	50
38	749	46	21
39	1073	47	4
40	1079	48	1

a. Make a histogram.

b. Describe the shape, center, and spread of the chest measurements distribution. Why might this information be useful?

Freeport AP Statistics

Chapter 1: EXPLORING DATA

1.3 Describing Quantitative Data with Numbers

DAY 5

OBJECTIVE(S):

- Students will learn how to calculate and interpret measures of center (mean, median).
- Students will learn how to calculate and interpret measures of spread (IQR, standard deviation).
- Students will learn how to identify outliers using the 1.5 x *IQR* rule.
- Students will learn how to make a boxplot.
- Students will learn how to select an appropriate measures of center and spread.
- Students will learn how to use appropriate graphs and numerical summaries to compare distributions of quantitative variables.

Mean $\overline{x}$ ("x-bar") **-**

What's the difference between $\overline{x}$ and μ ?

Median, (M) **–**

1.

2.

3.

20. What does it mean when we say that the median is a resistant calculation, and mean is not a resistant calculation?

21. Where is mean located in relation to the median for the following three types of distributions?

    a. Symmetrical

    b. Skewed Right

    c. Skewed Left

Range –

Interquartile Range -

First Quartile (Q_1 ***) –***

Third Quartile (Q_3 ***) -***

22. When measuring spread, IQR is preferred over Range, but why?

<u>(1.5 X IQR) Rule for Outliers</u> **–**

23. The 2009 roster of the Dallas Cowboys professional football team included 7 defensive linemen. Their weights (in pounds) were 306, 305, 315, 303, 318, 309, and 285.
    a. Calculate the mean and interpret your result in context.

    b. Calculate the median and interpret your result in context.

    c. Suppose the last player had weighed 245 pounds instead of 285 pounds. How would this change affect the mean and the median? What property of measures of center does this illustrate?

    d. Find and interpret the interquartile range (IQR).

    e. Determine whether there are any outliers in this distribution.

24. The mean and median selling prices of existing single-family homes sold in November 2009 were \$216,400 and \$172,600. Which of these numbers is the mean and which is the median? Explain your reasoning.

25. Last year a small accounting firm paid each of its five clerks \$22,000, two junior accountants \$50,000 each, and the firm's owner \$270,000.
    a. What is the mean salary paid at this firm?
    b. How many of the employees earn less than the mean?
    c. What is the median salary?
    d. Describe how an unethical recruiter could use statistics to mislead prospective employees?

26. **TEXTBOOK P. 69 #88**
    a. Median =
    b. Q_1 =
    c. Q_3 =
    d. Approximate the mean and explain your methodology.

DAY 6

27. What is the five number summary?

28. Each of the five numbers breaks your data up into what percents?

29. What is the most common measure of spread we use?

Standard Deviation –

Variance -

30. We have three ways to measure spread: Standard Deviation, IQR, and range. Rank these from most resistant to least resistant.

31. We now can address center with either mean or median, but never use both. When should we use mean and when should we use median?

32. We now can address spread with either IQR or standard deviation, but never use both. When should we use IQR and when should we use standard deviation?

33. Here are the scores of Mr. McGrady's students on their first statistics test over the last 4 years:

84	94	83	82	95	65	97	84	93	90	79
85	93	85	71	93	79	83	83	91	80	74
52	97	99	48							

a. Make a boxplot of the test score data.

b. How did the students do on the first test over the past years?

34. **TEXTBOOK p. 70 #94**

a. Make a boxplot of these data.

b. Describe this distribution.

35. **TEXTBOOK p. 71 #96**

36. The first four students to arrive for a first-period statistics class were asked how much sleep (to the nearest hour) they got last night. Their responses were 7, 7, 9, and 9.

    a. Find the standard deviation.

    b. Interpret the standard deviation.

    c. Do you think it's safe to conclude that the mean amount of sleep for all 30 students in this class is close to 8 hours? Why or why not.

37. Do male doctors perform more cesarean sections (C-sections) than female doctors? A study in Switzerland examined the number of cesarean sections (surgical deliveries of babies) performed in a year by samples of male and female doctors. Here are summary statistics for the two distributions.

	$\bar{x}$	s	Min.	Q_1	M	Q_3	Max.	IQR
Male doctors	41.333	20.607	20	27	34	50	86	23
Female doctors	19.1	10.126	5	10	18.5	29	33	19

a. Which distribution would you guess has a more symmetrical shape? Explain.

b. Explain how the *IQR*s of these two distributions can be fairly similar even though the standard deviations are quite different.

c. Does it appear that males perform more C-sections? Justify your answer.

38. **TEXTBOOK p. 73 #102**

CONCLUSION: The most important part of Chapter 1 is describing/comparing ________________________ distributions. You need to address all **4** components of _____________, but first you need to know the __________________. The table below summarizes what to summarize based on the shape of the distribution.

	SHAPE	
COMPONENTS	**Symmetrical**	**Skewed**
***S**hape*		
***O**utliers*		
***C**enter*		
***S**pread*		

Also, when comparing two or more distributions, it is not enough to just state SOCS for both distributions. You must also **COMPARE** them! i.e. $s_x = 2.7$ inches and $s_y = 5.0$ inches. Therefore the measurements of distribution y are more variable than x.

DAY 7