Part II: Exploring and Understanding Data

Describing Distributions Numerically

All our measures in this section can be found in the 1-VAR STATS output for your TI-83 or TI-84. Make sure you know how to enter a list and use this feature – even if only to check your work you did by hand!

Know how to find all numeric measures.

	Shape of Distribution	
	Symmetric distribution	Skewed distribution or one with outliers
Measure of Center	mean	median
Measure of Spread	standard deviation	IQR

Expect to be asked which measure of center or spread to use for a described (or graphically presented) data set. Use the above chart to determine and pay close attention to if they're asking for center or spread.

Example Questions:

1.) The salaries of professional athletes are right-skewed with a number of high outliers. Which measure would be used to describe the spread of the salaries?

Hint: What measure are we interested in: center or spread?

A. Mean B. Median C. Standard Deviation D. IQR E. Range

2.) What measure of center would be appropriate for an approximately normally distributed variable?
   A. The mean, because it's always the best measure of center.
   B. The mean, because it's the best measure of center for a symmetric distribution.
   C. The mean, because it's the best measure of center for a skewed distribution.
   D. The median, because it's the best measure of center for a symmetric distribution.
   E. The median, because it's the best measure of center for a skewed distribution.

Measures of Center

All of these give us a measure of where the center of our data is.

- **Mean:** average of all data values
- **Median:** middle value of all data values
  - Remember to use the median location formula = (n+1)/2
  - Robust to outliers or skew since it doesn't matter at all what the values to either side of the median are.

Remember, **your mean will move in the direction of a tail (skew) or outliers**. Assume you will have at least one question giving you a mean and median and asking about the shape of the distribution.

- A mean larger than the median means the distribution is skewed right (positive skew)
- A mean smaller than the median means the distribution is skewed left (negative skew)
- A mean close to the median means the distribution is approximately symmetric

Example Questions:

1.) The salaries of professional athletes are right-skewed with a number of high outliers. Which of the following is true?
   A. The mean will be larger than the median.
   B. The median will be larger than the mean.
   C. The mean and median will be close to one another.
   D. It cannot be determined from the information given.

2.) A professor made his test fairly easy and the resulting mean was 90.2 with a median of 83. Which of the following can we conclude about the distribution of test scores for this exam?

A. The test scores will have a symmetric distribution.
B. The test scores will be positively skewed.
C. The test scores will be negatively skewed.
D. There is not enough information.

3.) True or False. If two sets have the same mean and median, then they must be identical.

Measures of Spread

All of these give us a measure of how spread out our data is (or how much our values vary from one another).

- **Standard Deviation (and Variance):** measures the spread or variability of the data.
  - $s \geq 0$
  - The closer s is to 0, the less spread out our data is
  - For s to be exactly 0, all data values must be identical
  - Will get very large with a skewed distribution or with outliers
- **Interquartile Range (IQR) = Q3 - Q1**
  - Memorize the formula, it won't be given to you
  - Robust to outliers or skew since it cuts off the 25% of data on either end
- **Range = max – min**
  - Unlikely you'll be asked for it and it's never the right choice when asked for the appropriate measure of spread, but know the formula in case

Example Questions:

1.) True or False. A standard deviation of zero indicates all the data values must be the same.

2.) True or False. If ten is added to each value in a data set, the IQR will remain unchanged.

3.) Calculate the standard deviation for the following data set:

7.6	7.8	8.1	8.1	8.7	9.2	9.5	9.5	10

Finding Quartiles

If you can't use the 1-VAR-STATS because it's a histogram question or because you have a scientific calculator, you'll need to follow the steps below. Know them!

Find both quartiles by **FIRST** determining how many values are in half of your data set.

- If you have an even number of values, then you can divide the number of values in half.
  - If there are 50 people in the sample, then there are 25 in each half.
- If you have an odd number of values, subtract 1 (because you won't include the median in your quartile calculations) from your total number before dividing in half.
  - If there are 21 people in your sample, then there are (21-1)/2 = 10 in each half.

SECOND, use the median location formula with our new sample size (our value for half the data), to find where Q1 and Q3 fall. For Q1, count over from the smallest data value to find the location. For Q3, count over from the largest data value to find the location.

- **First Quartile (Q1):** Middle value of the smallest half of the data, the 25^{th} percentile.
- **Third Quartile(Q3):** Middle value of the largest half of the data, the 75^{th} percentile.

Do not use the A+ Review method here! It's WRONG.

Five Number Summary

Consists of minimum, Q1, median, Q3, maximum. Remember to label them!

- You can use the five number summary to create a boxplot.
- Can be found in the 1-VAR-STATS by scrolling to the second page.

Example Questions:

1.) If a data set has a low outlier, how will it change the measures of center and spread if we remove it? Will they increase, decrease or stay the same?

Mean: Median: Standard Deviation: IQR:

2.) If we add or subtract 5 from each observation in our data set, how will this change our measures of center and spread?

Mean: Median: Standard Deviation: IQR:

3.) If we divide every observation in our data set by 2, how will this change our measures of center and spread?

Mean: Median: Standard Deviation: IQR:

4.) If 75% of values in our distribution are at least 200 and the IQR is 50, which of the following is true?

A. 50% of the values are between 200 and 250.
B. The median of our distribution is 225.
C. 50% of the values are between 150 and 200.
D. Both A and B are true.
E. Both B and C are true.

Displaying Categorical Data

One Categorical Variable: Bar Graphs and Pie Charts
- Just be aware these are the only two displays we learned for categorical data

Two Categorical Variables: Contingency Tables
- Most of these questions can also be answered using what we learned in probability, make sure to read carefully for what is being asked (see more in the Part IV review notes)
- Know how to make a **conditional contingency table**, whatever is **conditioned on** or **conditioned by** will become the denominator for every cell in that row or column

Example Questions:

1.) Which of the following is an appropriate display for a data set containing ratings (G, PG, PG-13, or R) of a sample of 90 movies?

A. Pie Chart B. Box plot C. Bar Chart D. Both A and C

2.) 250 elementary school children in a certain city were randomly selected and then give a test to measure math ability.

	Private School	Public School	Home School
Passed	60	40	20
Failed	25	75	30

What percent of the students failed?

What percent of the students passed or went to private school?

What percent of public school students failed?

Of the students who went to home school, what percent failed?

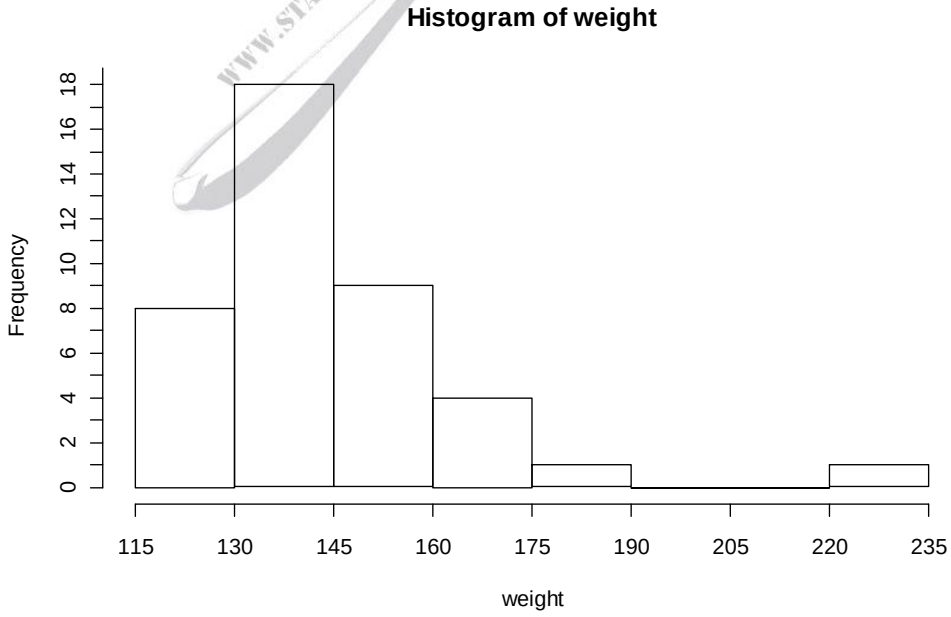
Create a conditional distribution for test performance, conditioning on where the child attends school.

Displaying Quantitative Data

One Quantitative Variable: Histograms, Stem plots, Boxplots and Timeplots

- Histograms
  - Remember, first step should be filling any missing information into your histogram – putting any missing x values down and writing how many observations are in each bar at the top.
  - Know how to find median, Q1, and Q3 intervals for a histogram
  - Know how to find probabilities using a given histogram.

Example Question:



Identify the intervals for Q1, median and Q3.

What is the appropriate choice of center and spread for the above distribution of weights?

- Stem plots
  - On back-to-back stemplots, always make sure you read away from the stem and that you're answering for the group the question is about!
- Boxplots
  - **Know how to make a boxplot!** Assume this will be a free response question.
  - Memorize your fence formulae!
    - Mild lower fence: Q1 – 1.5(IQR)
    - Mild upper fence: Q3 + 1.5(IQR)
    - Extreme lower fence: Q1 – 3(IQR)
    - Extreme upper fence: Q3 + 3(IQR)
  - Know how to use your fences to determine if a value is an outlier
  - Each portion of the boxplot represents 25% of the data, the box itself represents the interquartile range (IQR, middle 50%) of the data.

Note: We also learned two other types of graphs: Timeplots which can be appropriate for a single quantitative variable if we are interested in how a continuous variable varies over TIME. And scatterplots which require TWO quantitative variables.

Example Questions:

1.) A class of 5th graders took a spelling test. The five-number summary is below:

Minimum	Q1	Median	Q3	Maximum
2.0	3.9	4.3	4.9	9.0

Is the minimum score a mild outlier?

Is the maximum score a mild outlier, an extreme outlier or neither?

2.) A back-to-back stem plot was created for a data set and is given below:

Group A		Group B
	1	7
	2	2 2 6 8
5 5 4 4 2	3	3 4 5 9
9 7 3 0	4	0 1 2 6
6 2 1	5	9
5	6	
	7	
1	8	

For the data above, identify and calculate the appropriate measures of center and spread:

Group A

Measure of center: _____

Calculated value: _____

Measure of spread: _____

Calculated value:_____

Group B

Measure of center: _____

Calculated value: _____

Measure of spread: _____

Calculated value:_____

3.) Below are the numbers of text message sent in one week by the cell phone users on one floor of a college dormitory.

90	105	108	109	112	112	113	114	116	117
118	119	122	122	123	125	127	130	131	139

Construct a split-stem plot.

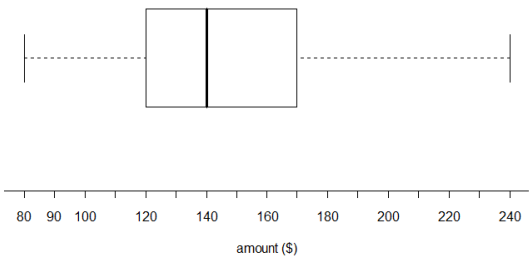
Calculate (and clearly label) the five-number summary for the data.

Check the data for outliers, using the criterion discussed in class.

Use the information above to construct a boxplot for the data.

90 100 110 120 130 140

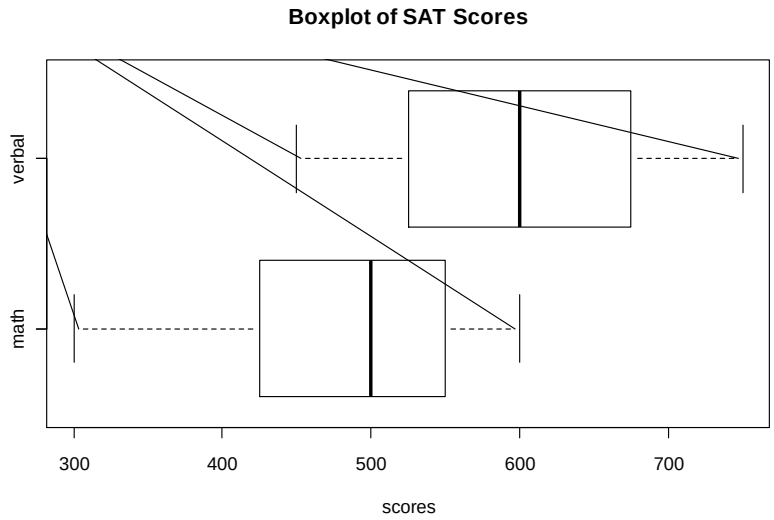
4.) A chain of sports shops in Lake Tahoe wants to study how much a beginning skier spends on his or her initial purchase of ski equipment. Data was collected from 48 sales and is displayed below.



75% of sales receipt totals were at least what amount? _____

25% of sales receipt totals were less than what amount? _____

5.) A box plot below summarizes the distribution of SAT verbal and math scores for students at an Arizona high school



Which of the following statements is FALSE?

A. The interquartile range for the math scores is smaller than the interquartile range for the verbal scores.
B. The range of the math scores equals the range of the verbal scores.
C. The highest math score equals the median verbal score.
D. The verbal scores are roughly symmetric while the math scores are skewed to the right.

Graph Features

The quantitative graphs listed above can help you see the following aspects of a distribution:

- Modes (unimodal, bimodal, multimodal)
- Symmetry (symmetric, skewed to left, skewed to right)
- Usual Features (outliers, gaps)