Chapter 1.3 Lecture Notes & Examples

Section 1.3 – Describing Quantitative Data with Numbers (pp. 50-74)

1.3.1 Measuring Center: The Mean

Mean - The arithmetic average. To find the mean (pronounced x bar) of a set of observations, add their values and divide by the number of observations. If the n observations are x1, x2,…, xn, their mean is:

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Or

$$\bar{x} = \frac{\sum x_i}{n}$$

Actually, the notation $\bar{x}$ refers to the mean of a sample. Most of the time, the data we'll encounter can be thought of as a sample from some larger population. When we need to refer to a population mean, we'll use the symbol μ (Greek letter mu, pronounced "mew"). If you have the entire population of data available, then you calculate μ in just the way you'd expect: add the values of all the observations, and divide by the number of observations.

Example – Travel Times to Work in North Carolina Calculating the mean

Below is data on travel times of 15 North Carolina residents.

0	5
1	000025
2	005
3	00
4	00
5	
6	0

Key: 2|5 is a NC worker who travels 25 minutes to work.

1) Find the mean travel time for all 15 workers

2) Calculate the mean again, this time excluding the person who reported a 60-minute travel time to work. What do you notice?

The previous example illustrates an important weakness of the mean as a measure of center: the mean is sensitive to the influence of extreme observations. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a resistant measure of center.

Resistant Measure - A statistic that is not affected very much by extreme observations.

1.3.2 Measuring Center: The Median

Median - The median M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.

2. If the number of observations n is odd, the median M is the center observation in the ordered list.

3. If the number of observations n is even, the median M is the average of the two center observations in the ordered list.

Example – Travel Times to Work in North Carolina
Finding the median when n is odd

What is the median travel time for our 15 North Carolina workers? Here are the data arranged in order:

5 10 10 10 10 12 15 **20** 20 25 30 30 40 40 60

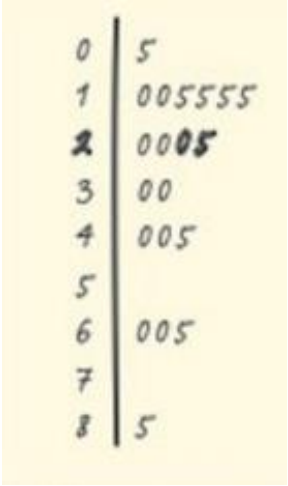
The count of observations n = 15 is odd. The bold **20** is the center observation in the ordered list, with 7 observations to its left and 7 to its right. This is the median, M = 20 minutes.

Example – Stuck in Traffic Finding the median when n is even

People say that it takes a long time to get to work in New York State due to the heavy traffic near big cities. What do the data say? Here are the travel times in minutes of 20 randomly chosen New York workers:

10 30 5 2 5 40 20 10 15 30 20 15 20 85 15 65 15 60 60 40 45

1. Make a stemplot of the data. Be sure to include a key.



2. Find an interpret the median.

3. Resistance:

1.3.3 Comparing the Mean and Median

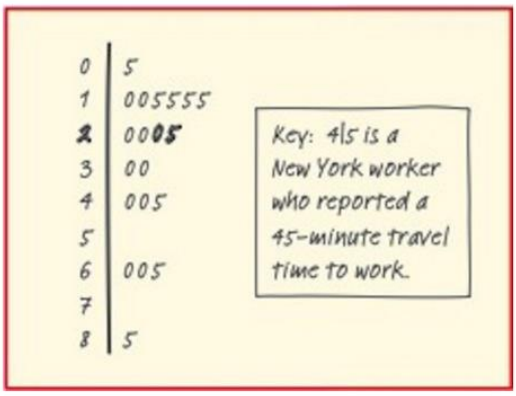
Our discussion of travel times to work in North Carolina illustrates an important difference between the mean and the median. The median travel time (the midpoint of the distribution) is 20 minutes. The mean travel time is higher, 22.5 minutes. The mean is pulled toward the right tail of this right-skewed distribution. The median, unlike the mean, is resistant. If the longest travel time were 600 minutes rather than 60 minutes, the mean would increase to more than 58 minutes but the median would not change at all. The outlier just counts as one observation above the center, no matter how far above the center it lies. The mean

- The mean and median of a *roughly symmetric* distribution will be close together.
- If the distribution is *exactly symmetric*, the mean and median will be exactly the same.
- In a *skewed* distribution, the mean is usually farther out in the long tail than the median.

Check Your Understanding

Questions 1 through 4 refer to the following setting. Here, once again, is the stemplot of travel times to work for 20 randomly selected New Yorkers. Earlier, we found that the median was 22.5 minutes.

1. Based only on the stemplot, would you expect the mean travel time to be less than, about the same as, or larger than the median? Why?



2. Use your calculator to find the mean travel time. Was your answer to Question 1 correct?

3. Interpret your result from Question 2 in context without using the words “mean” or “average.”

4. Would the mean or the median be a more appropriate summary of the center of this distribution of drive times? Justify your answer.

1.3.4 Measuring Spread: The Interquartile Range (IQR)

A useful numerical description of a distribution requires both a measure of center and a measure of spread.

How to Calculate Quartiles Q1 | M | Q3

1. Arrange the observations in increasing order and locate the median **M** in the ordered list of observations. (Median cuts off 50%; sometimes the median is called Q2, the second quartile)

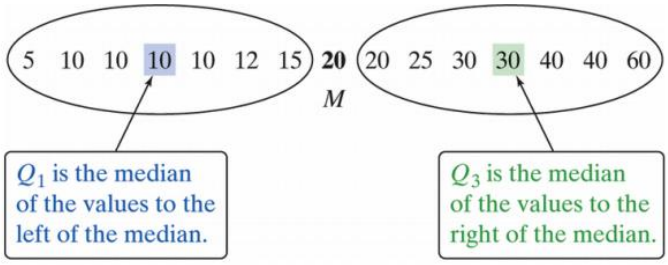
2. **The first quartile Q1** is the median of the observations whose position in the ordered list is to the left of the median. (Cuts off 25%)

3. **The third quartile Q3** is the median of the observations whose position in the ordered list is to the right of the median. (Cuts off 75%)

4. **Interquartile Range – IQR = Q3 - Q1**

Example – Travel Times to Work in North Carolina Calculating quartiles

Our North Carolina sample of 15 workers' travel times, arranged in increasing order, is



There is an odd number of observations, so the median is the middle one, the bold **20** in the list. The first quartile is the median of the 7 observations to the left of the median. This is the 4th of these 7 observations, so Q1 = 10 minutes (shown in blue). The third quartile is the median of the 7 observations to the right of the median, Q3 = 30 minutes (shown in green).

Find and Interpret the Interquartile Range (IQR) of the travel times to work in North Carolina.

Resistance: The quartiles and the interquartile range are resistant because they are not affected by a few extreme observations

1.3.5 Identifying Outliers

In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

Identifying Outliers – An observation that falls more than 1.5 x IQR above Q_3 or below Q_1 is considered an *outlier*.

Example – Travel Times to work in New York Identifying Outliers using the 1.5*IQR rule

Identify any outliers in the data from the stemplot.

Q1 = 15 minutes
Q3 = 42.5 minutes
IQR = 27.5 minutes

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 is a New York worker who reported a 45-minute travel time to work.

Example – Travel Times to Work in North Carolina Identifying Outliers

Determine if the travel time of 60 minutes in the sample of 15 North Carolina workers is an outlier.

Q1 = 10 minutes
Q3 = 30 minutes
IQR = 20 minutes

0	5
1	000025
2	005
3	00
4	00
5	
6	0

Key: 2|5 is a NC worker who travels 25 minutes to work.

1.3.6 The Five-Number Summary and Boxplots

Five-Number Summary – Consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

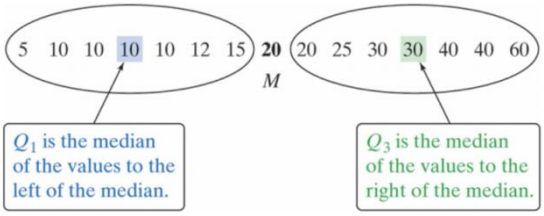
Minimum Q1 M Q3 Maximum

These five numbers divide each distribution roughly into quarters. About 25% of the data values fall between the minimum and Q1, about 25% are between Q1 and the median, about 25% are between the median and Q3, and about 25% are between Q3 and the maximum. The five-number summary of a distribution leads to a new graph, the boxplot (aka box and whisker plot).

How to Make a Boxplot

1. A central box is drawn from the first quartile (Q1) to the third quartile (Q3).
2. A line in the box marks the median.
3. Lines (called whiskers) extend from the box out to the smallest and largest observations that are not outliers.

Example: Make a boxplot for the data about the Travel Times to Work North Carolina



Calculator Procedures

TECHNOLOGY CORNER Making calculator boxplots

The TI-83/84 and TI-89 can plot up to three boxplots in the same viewing window. Let's use the calculator to make side-by-side boxplots of the travel time to work data for the samples from North Carolina and New York.

1. Enter the travel time data for North Carolina in L1/list1 and for New York in L2/list2.

2. Set up two statistics plots: Plot1 to show a boxplot of the North Carolina data and Plot2 to show a boxplot of the New York data.

TI-83/84

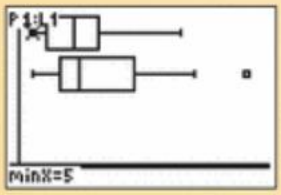
TI-89

***Note:* The calculator offers two types of boxplots: a "modified" boxplot that shows outliers and a standard boxplot that doesn't. We'll always use the modified boxplot.**

3. Use the calculator's Zoom feature to display the side-by-side boxplots. Then Trace to view the five-number summary.

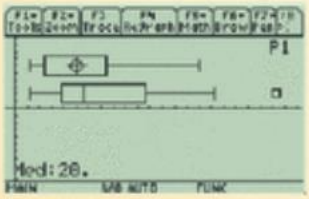
TI-83/84

- Press ZOOM and select 9 : ZoomStat.
- Press TRACE.



TI-89

- Press F5 (ZoomData).
- Press F3 (Trace).

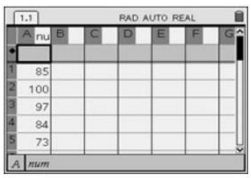


Travel times for a random sample for 15 adults in North Carolina in minutes: 5,10,10,10,10,12,15,20, 20,25, 30,30,40,40,60

Directions for TI Inspire

1. Enter the data on a *List & Spreadsheet Page.* This list is named "*num*".

(See Lists and Spreadsheets for entering data.)



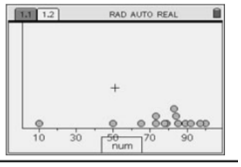
2. Create a *Data & Statistics Page*

Press (home) #5Data & Statistics.

At first, a generic statistics graph will appear.

3. Move the cursor to hover over the *Click to add variable* at the bottom of the screen. Click. Choose *num* from the choices.

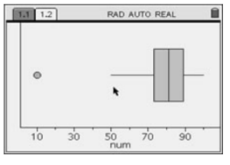
A dot plot will appear.



4. Press ctrl menu and change the graph type to Box Plot.

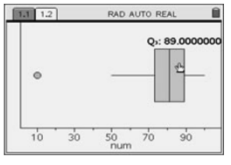
This is a **modified box-and-whisker plot** with the outlier separated from the rest of the data display.

Note: If there are no points in your data outside of 1.5 * Interquartile Range, no outliers will be displayed.



5. As you move the cursor to hover over the box-and-whisker plot, you will see the values for the minimum (not including outlier), first quartile, median, third quartile, and the maximum.

If you move your cursor on the outlier and press Click, you will see its value.



Team Work: Complete Check Your Understanding
The 2009 roster of the Dallas Cowboys professional football team included 10 offensive linemen. Their weights (in pounds) were 338 318 353 313 318 326 307 317 311 311

1. Find the five-number summary for these data by hand. Show your work.

2. Calculate the IQR. **Interpret this value in context.**

3. Determine whether there are any outliers using the 1.5 × IQR rule.

4. Draw a boxplot of the data.

1.3.7 Measuring Spread: The Standard Deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the **mean to measure center** and the **standard deviation to measure spread**.

The ***standard deviation*** s_x measures the *average* distance of the observations from their mean.

Procedure: The *standard deviation* is calculated by finding the average of the squared distances and then taking the square root. The average *squared difference* is called the ***variance***.

Symbols/Formula:

$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

Example: These are the foot lengths (in cm) for a random sample of seven 14-year-olds from the United Kingdom:
25 22 20 25 24 24 28
The mean foot length is 24 cm.

x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
25		
22		
20		
25		
24		
24		
28		

Many calculators report **two standard deviations**, giving you a choice of **dividing by n or by n − 1**. The former is usually labeled σ_x, the symbol for the **standard deviation of a population**. If your data set consists of the entire population, then it's appropriate to use σ_x. **More often, the data we're examining come from a sample. In that case, we should use s_x.** More important than the details of calculating s_x are the properties that determine the usefulness of the standard deviation:

Properties of the Standard Deviation

- **s_x** measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- **s_x is always greater than or equal to 0**. **s_x = 0 only when there is no variability.** This happens only when all observations have the same value. Otherwise, $s_x > 0$. **As the observations become more spread out about their mean, s_x gets larger.**
- **s_x has the same units of measurement as the original observations**. For example, if you measure metabolic rates in calories, both the mean X and the standard deviation s_x are also in calories. This is one reason to prefer s_x to the variance, which is in squared calories.
- Like the mean X, s_x is not resistant. A few outliers can make s_x very large.

The use of squared deviations makes **s_x** even more sensitive than X to a few extreme observations.

Check your understanding, p. 64.

The heights (in inches) of the five starters on a basketball team are 67, 72, 76,76, and 84.

1. Find and interpret the mean.

2. Make a table that shows, for each value, its deviation from the mean and its squared deviation from the mean.

3. Show how to calculate the variance and standard deviation from the values in your table.

4. Interpret the meaning of the standard deviation in this setting.

1.3.9 Choosing Measure of Center and Spread

We now have a choice between **two descriptions of the center and spread of a distribution**:

1) the median and IQR,

or

2) $\bar{x}$ and s_x.

Because $\bar{x}$ and s_x are sensitive to extreme observations, they can be misleading when a distribution is strongly skewed or has outliers. In these cases, the median and IQR, which are both resistant to extreme values, provide a better summary. We'll see in the next chapter that the mean and standard deviation are the natural measures of center and spread for a very important class of symmetric distributions, the Normal distributions.

Choosing Measures of Center and Spread

- Skewed Distributions:
- Distributions with strong outliers:
- Reasonably symmetric distributions:

*****Resistance*****

- Median:
- IQR:
- IQR:

Analyzing Data Sets

From this point on, whenever you are analyzing data sets, in the "Do" step you should:

- Plot the distribution
- Create a numerical summary which includes:
    - Mean
    - Standard deviation
    - 5-Number Summary (min, Q1, median, Q3, max)

Example - Who Texts More—Males or Females? Pulling it all together

For their final project, a group of AP Statistics students investigated their belief that females text more than males. They asked a random sample of students from their school to record the number of text messages sent and received over a two-day period. Here are their data:

Males:	127	44	28	83	0	6	78	6	5	213	73	20	214	28	11	
Females:	112	203	102	54	379	305	179	24	127	65	41	27	298	6	130	0

What conclusion should the students draw? Give appropriate evidence to support your answer. (State, Plan, Do, Conclude)