

## 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  $x_1, x_2, \dots, x_n$ , their mean is:

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

Or

$$\bar{x} = \frac{\sum x_i}{n} \leftarrow \text{sample size}$$

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  $\mu$  (Greek letter mu, pronounced "mew"). If you have the entire population of data available, then you calculate  $\mu$  in just the way you'd expect: add the values of all the observations, and divide by the number of observations.

$$\mu = \frac{\sum x_i}{N} \leftarrow \text{population size}$$

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

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

- 1) Find the mean travel time for all 15 workers

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

$$= 5 + 10 + 10 + 10 + 10 + 12 + 15 + 20 + 20 + 25 + 30 + 30 + 40 + 40 + 60$$

$$= 337 = \sum x_i$$

$$\bar{x} = \frac{337}{15} \approx 22.5 \text{ min}$$

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

Key: 25 is a NC worker who travels 25 minutes to work

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

$$\sum x_i = 277 \quad \bar{x} = \frac{277}{14} \approx 19.8 \text{ min}$$

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  
1 2 3 4 5 6 7 8

$\frac{15}{2} = 7.5$  so 8th term is median

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.

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

Key: 4|5 is a NY worker who reported a 45 min travel time to work.

$$\text{Med} = \frac{(20 + 25)}{2} = \frac{45}{2} = 22.5 \text{ min}$$

$n = 20$

median has 10 #'s above and 10 below

2. Find and interpret the median.

In sample of NY workers, about  $\frac{1}{2}$  of people reported a travel time less than 22.5 min and  $\frac{1}{2}$  reported a time greater than 22.5 min.

\* 3. Resistance: Median is more resistant than the mean.



### 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

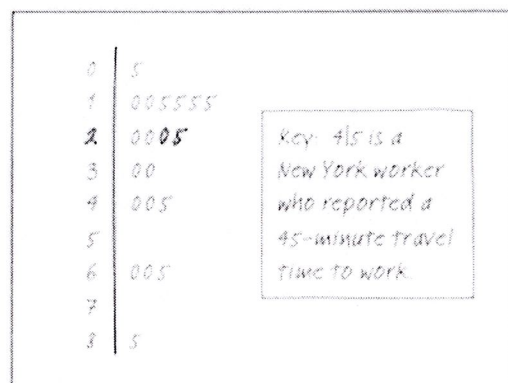
- ✱ • The mean and median of a roughly symmetric distribution will be close together.
- ✱ • If the distribution is exactly symmetric, the mean and median will be exactly the same.
- In a skewed distribution, the mean is usually farther out in the long tail than the median.

#### 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?

Distribution is skewed right, so we would expect mean to be larger than the median.



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

$\bar{x} = 31.25$  minutes which is greater than mean of 22.5 minutes  
 $\bar{x} = (5 + 10 + 10 + 15 + \dots + 85) / 20$

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

If we divided the travel time up evenly among all 20 people, each would have a 31.25 minute travel time.

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

Since distribution is skewed, the median is a better measure of the center.

### 1.3.4 Measuring Spread: The Interquartile Range (IQR)

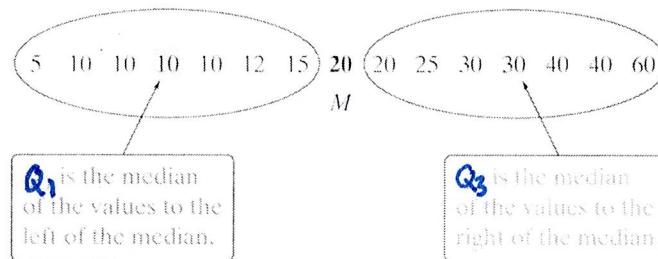
A useful numerical description of a distribution requires both a measure of center and a measure of spread.

#### 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.

$$IQR = 30 - 10 = 20 \text{ minutes}$$

The range of the middle half of travel times is 20 minutes.

**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  $1.5 \times \text{IQR}$  above  $Q_3$  or below  $Q_1$  is considered an outlier.

#### Example – Travel Times to work in New York Identifying Outliers using the $1.5 \times \text{IQR}$ rule

Identify any outliers in the data from the stemplot.

$Q_1 = 15$  minutes

$Q_3 = 42.5$  minutes

$\text{IQR} = 27.5$  minutes

$$1.5 \times \text{IQR} = 1.5(27.5) = 41.25$$

Outliers any values not falling between  
 $Q_1 - 1.5(\text{IQR}) = 15 - 41.25 = -26.25$   
 and  
 $Q_3 + 1.5(\text{IQR}) = 42.5 + 41.25 = 83.75$

$$Q_3 + 1.5(\text{IQR}) = 42.5 + 41.25 = 83.75$$

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.

So only outlier is 85.

#### 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.

$Q_1 = 10$  minutes

$Q_3 = 30$  minutes

$\text{IQR} = 20$  minutes

$$1.5(20) = 30$$

$$Q_3 + 1.5(\text{IQR}) = 30 + 30 = 60$$

$$Q_1 - 1.5(\text{IQR}) = 10 - 30 = -20$$

$$-20 \text{ to } 60$$

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.

6|0 ✓ ok

so no outliers in this data set

### 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

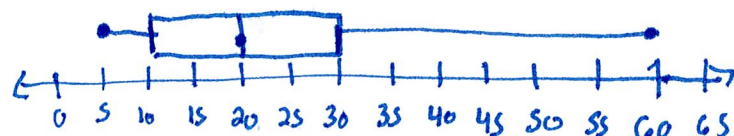
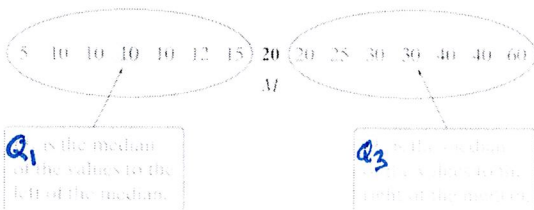
**Minimum Q1 M Q3 Maximum**

These five numbers divide each distribution roughly into quarters. About 25% of the data values fall between the minimum and  $Q_1$ , about 25% are between  $Q_1$  and the median, about 25% are between the median and  $Q_3$ , and about 25% are between  $Q_3$  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 ( $Q_1$ ) to the third quartile ( $Q_3$ ).
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



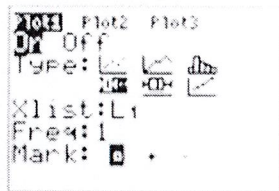
Travel Times to Work in N.C.

## TECHNOLOGY CORNER Making calculator boxplots

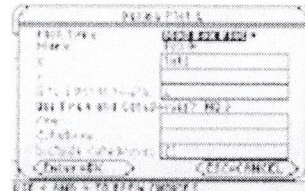
The TI-83/84 and TI-89 can plot up to three boxplots in the same viewing window. Let's use the calculator to make side-by-side boxplots of the travel time to work data for the samples from North Carolina and New York.

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



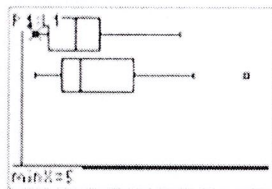
Travel times for a random sample for 15 adults in North Carolina in minutes:  
5,10,10,10,10,12,15,20,20,25,30,30,40,40,60

**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.

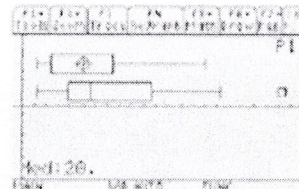
TI-83/84

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



TI-89

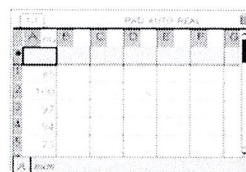
- Press **F5** (ZoomData).
- Press **F3** (Trace).



### Directions for TI Inspire

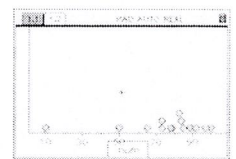
1. Enter the data on a *List & Spreadsheet Page*. This list is named "num".

(See [Lists and Spreadsheets](#) for entering data.)



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

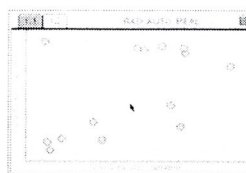
A dot plot will appear.



2. Create a *Data & Statistics Page*.

Press **2nd** **Home** #5Data & Statistics.

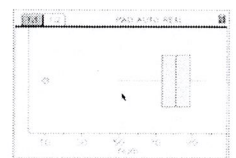
At first, a generic statistics graph will appear.



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

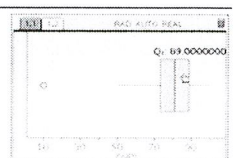
This is a **modified box-and-whisker plot** with the outlier separated from the rest of the data display.

**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.





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

$$Max = 353$$

Weight (Pounds)

300 310 320 330 340 350 360 370

outlier

\* Do not extend whisker to outlier. Let the next highest # be the max (so in this case 338)

Leave a point to mark outlier \*