First Test

Tuesday Sept. 18th

1.3

Have your graphing calculator handy

GDC

Describing Quantitative Data

with numbers

over 2 days

By the end of this section, you should be able to:

- ✓ CALCULATE measures of center (mean, median) for a distribution of quantitative data.
- ✓ CALCULATE and INTERPRET measures of variability (range, standard deviation, IQR) for a distribution of quantitative data.
- ✓ EXPLAIN how outliers and skewness affect measures of center and variability.
- ✓ IDENTIFY outliers using the 1.5 × IQR rule.
- ✓ MAKE and INTERPRET boxplots of quantitative data.
- ✓ Use boxplots and numerical summaries to COMPARE distributions of quantitative data.

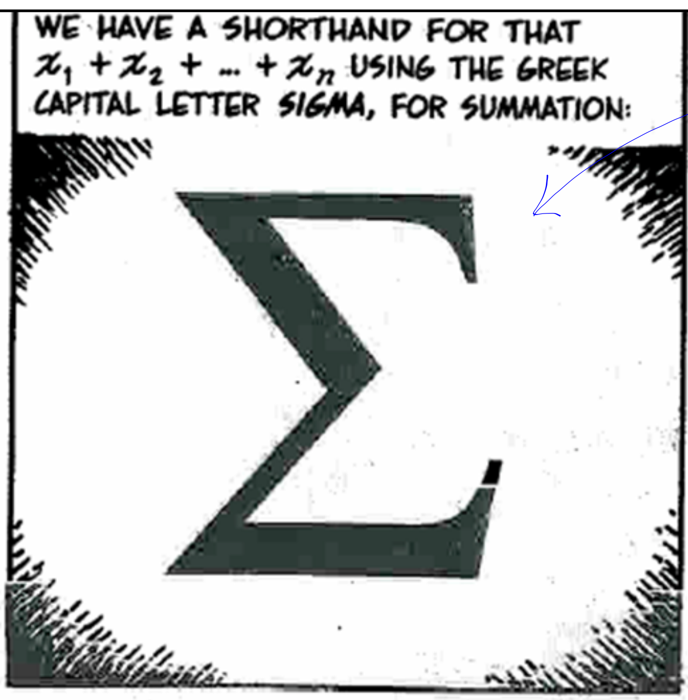
x 5 62 10 33 2

positions

1 2 3 4 5 ... n

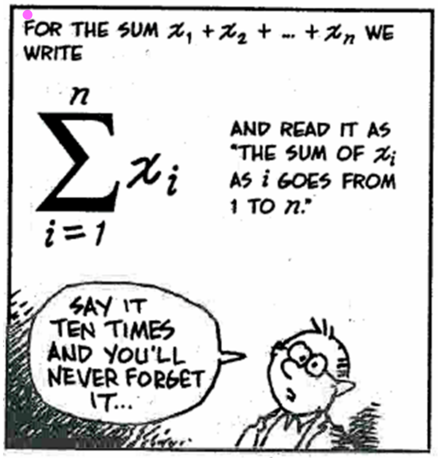
x 5 62 10 33 2 ... x_n

$x_2 = 62$ $x_5 = 2$



Means "Add Up"

σ



5 62 10 33 2

$$\sum_{i=1}^{4} x_i = 5 + 62 + 10 + 33$$
$$= 110$$

5 62 10 33 2 ... x_n

$$\sum_{i=1}^{n} x_i$$ add up all of the data

Write down

Mean

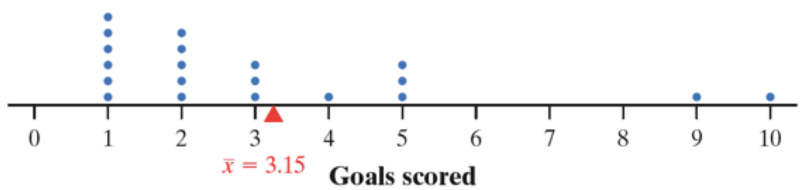
$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

$$\bar{x} = \frac{1+1+1+1+1+1+2+2+2+2+2+3+3+3+4+5+5+5+9+10}{20}$$

$\bar{x} = 3.15$ *goals*

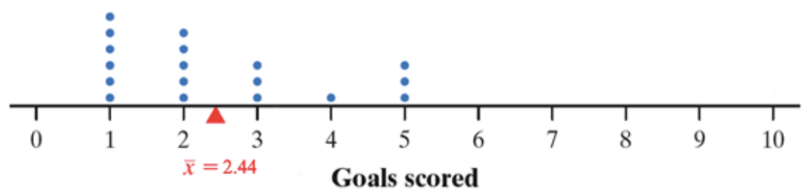


A statistical measure is **resistant** if it isn't sensitive to extreme values.

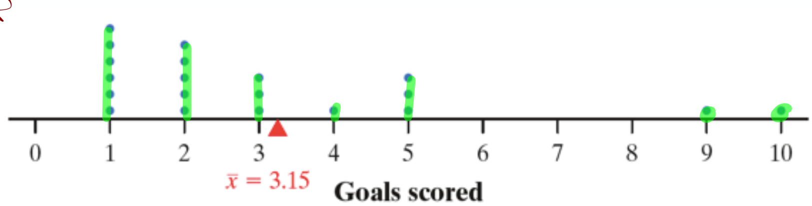
3. What is a resistant measure?

A statistic that does not dramatically change when an extreme value (low or high) gets added to the distribution.

Here is the mean number of goals scored by the 2016 U.S. women's soccer team, *if we exclude the games that are possible outliers* (when they scored 9 and 10 goals).

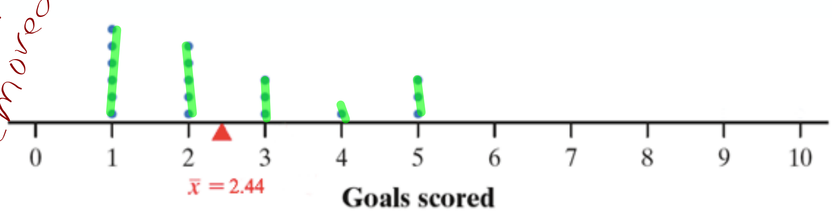


Original



$\bar{X} = 3.15$ goals

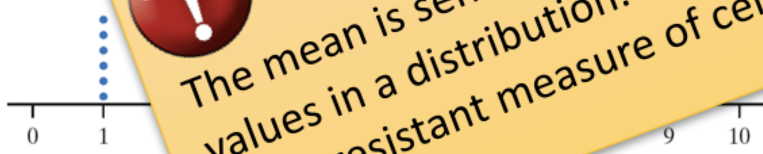
Outliers removed



$\bar{X} = 2.44$ goals

Here is the mean number of goals sco… U.S.
women's soccer team, if we ex… e
possible outliers (when t…

+ 5 + 5



CAUTION:

The mean is sensitive to extreme values in a distribution. The mean is ***not*** a resistant measure of center.

A statistical measure is **resistant** if it isn't sensitive to extreme values.

Speaking of Formulas

The official Formula Sheet ← can be used on both parts of AP exam

→ Appendix F-1

→ Your own copy you can use during quizzes and tests

- we may write other tidbits on this sheet later but....

4. **How many likes on Instagram for ASA?**

The American Statistical Association (www.amstat.org) has an Instagram account (@amstatnews) to post updates on new statistical publications and adorable normal distribution plushies. Here are the number of Instagram likes for 10 posts selected at random:

16	4	8	7	8
6	15	2	9	5

(a) Calculate the mean number of Instagram likes for these 10 posts. Show your work.

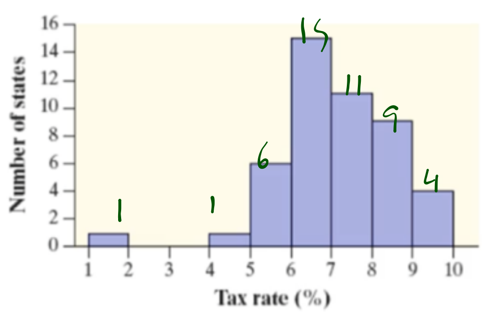
$$\bar{x} = \frac{\sum x}{n} = \frac{16+4+8+\cdots}{10} = \frac{80}{10} = 8 \text{ likes}$$

(b) The posts with 15 and 16 likes are possible outliers. Calculate the mean number of Instagram likes in the other 8 posts. What do you notice?

$$\bar{x} = 6.125 \text{ likes}$$

The mean number of likes dropped from 8 to 6.125

5. How can you estimate the mean from a histogram or dot plot?



$$\frac{1 \cdot 1.5 + 1 \cdot 4.5 + 6 \cdot 5.6 + \cdots}{47}$$

Definition of the Median is on page 57

if n is odd

if n is even

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

Raw data

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

Sorted data

1 1 1 1 1 1 2 2 2 2 2 3 3 3 4 5 5 5 9 10

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

Raw data

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

Sorted data

1 1 1 1 1 1 2 2 2 2 2 3 3 3 4 5 5 5 9 10

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

Raw data

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

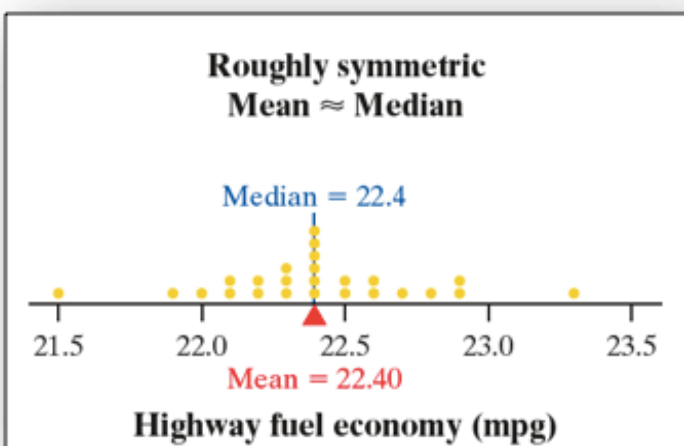
Sorted data

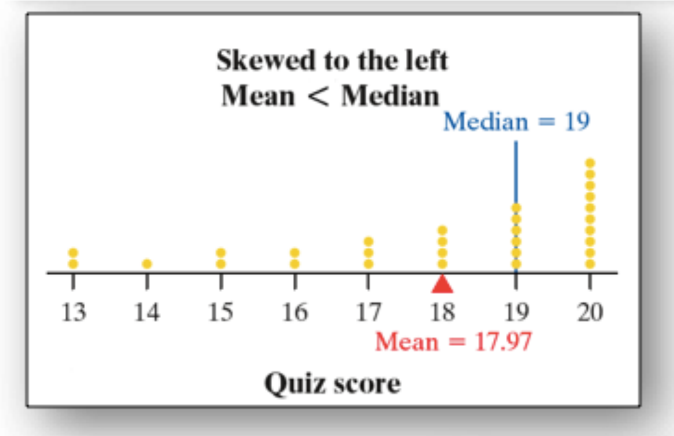
1 1 1 1 1 1 2 2 2 2 | 2 3 3 3 4 5 5 5 9 ~~10~~

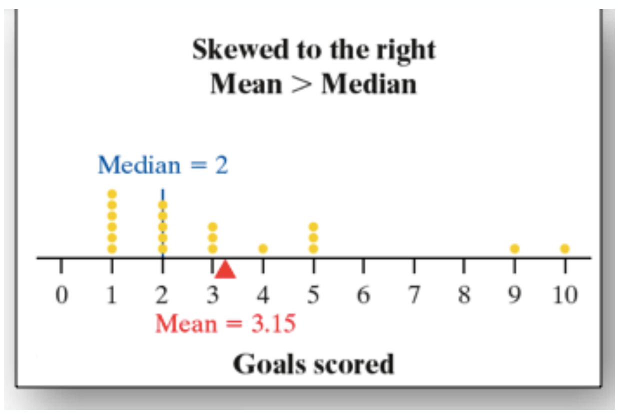
Median = $\frac{2+2}{2} = 2$

Comparing Mean and Median

Applet




Skewed to the left
Mean < Median
Median = 19
13
14
15
16
17
18
19
20
Mean = 17.97
Quiz score


Skewed to the right
Mean > Median
Median = 2
0
1
2
3
4
5
6
7
8
9
10
Mean = 3.15
Goals scored

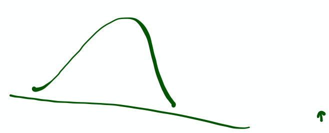
Effect of Skewness and Outliers on Measures of Center

- If a distribution of quantitative data is roughly symmetric and has no outliers, the mean and median will be similar.
- If the distribution is strongly skewed, the mean will be pulled in the direction of the skewness but the median won't. For a right-skewed distribution, we expect the mean to be greater than the median. For a left-skewed distribution, we expect the mean to be less than the median.
- The median is resistant to outliers but the mean isn't. ✓

AP Exam Tip

B ← → A

If students are asked to choose between the mean and median as a measure of center, be sure they justify their choice ***based on the shape of the distribution*** and **whether there are any possible outliers**



Measuring

Variability

-Range

-Standard Deviation

-IQR (Inter quartile range)

The **range** of a distribution is the distance between the minimum value and the maximum value. That is,

Range = Maximum – Minimum

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

$\text{Range} = 10 - 1 = 9 \text{ goals}$

caution

CAUTION:

- The range of a data set is a single number.
- The range is ***not*** a resistant measure of variability.

Measuring Variability: The Standard Deviation

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

or

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

The **standard deviation** measures the typical distance of the values in a distribution from the mean.

$$\frac{6x^2}{7}$$

$$\frac{1}{7} \cdot 6x^2$$

Measuring Variability: The Standard Deviation

How to calculate standard deviation, s_x:

1) Find the mean of the distribution.
2) Calculate the *deviation* of each value from the mean:
   deviation = value – mean.
3) Square each of the deviations.
4) Add all the squared deviations, divide by n – 1, and take the square root.

The **standard deviation** measures the typical distance of the values in a distribution from the mean.

Measuring Variability: The Standard Deviation

How to calculate standard deviation, s_x:

1) Find the mean of the distribution.
2) Calculate the *deviation* of each value from the mean:
   deviation = value – mean.
3) Square each of the deviations.
4) Add all the squared deviations, divide by n – 1, and take the square root.

The **standard deviation** measures the typical distance of the values in a distribution from the mean.

$$s_x = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

How to calculate standard deviation, s_x:

1) Find the mean of the distribution.
2) Calculate the *deviation* of each value from the mean: deviation = value – mean.
3) Square each of the deviations.
4) Add all the squared deviations, divide by $n - 1$, and take the square root.

The value obtained before taking the square root in the standard deviation calculation is known as the **variance**.

n high school students were asked how many "close" friends they have.
are their responses: 1 2 2 2 3 3 3 3 4 4 6

alculate standard , s_x:

l the mean of the
ribution.
ulate the *deviation* of
h value from the mean:
iation = value – mean.
are each of the
iations.
all the squared
iations, divide by n – 1,
take the square root.

$$s_x = \sqrt{\frac{18}{11 - 1}} = 1.34 \text{ close friends}$$

The value obtained before taking the square root in the standard deviation calculation is known as the **variance**.

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{18}{11 - 1}$$

= 1.80 *squared close friends*

6. How many likes on Instagram for ASA?

Here are the number of Instagram likes for 10 posts selected at random:

2	4	5	6	7
8	8	9	15	16

Calculate the standard deviation. Interpret this value in context.

table or follow sequence

2	4	5	6	7
8	8	9	15	16

$\bar{x} = 8$

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

x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	$2-8=-6$	$(-6)^2 = 36$
4	$4-8=-4$	$(-4)^2 = 16$
5	$5-8=-3$	$(-3)^2 = 9$
6	$6-8=-2$	$(-2)^2 = 4$
7	$7-8=-1$	$(-1)^2 = 1$
8	$8-8=0$	$(0)^2 = 0$
8	$8-8=0$	$(0)^2 = 0$
9	$9-8=1$	$1^2 = 1$
15	$15-8=7$	$7^2 = 49$
16	$16-8=8$	$8^2 = 64$

$\sum (x_i - \bar{x})^2 = 180$

$s = \sqrt{\dfrac{180}{10-1}}$

$= \sqrt{20}$

≈ 4.47 likes

Interpretation

The number of instagram likes for each ASA post typically varies by about 4.47 likes from the mean of 8 likes.

Properties of Standard Deviation

- s_x is always greater than or equal to 0.
- Larger values of s_x indicate greater variation.
- s_x is not a resistant measure of variability.
- s_x measures variation about the mean.

7. The value before taking the square root is known as the....

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

Std. Dev $s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$

IQR

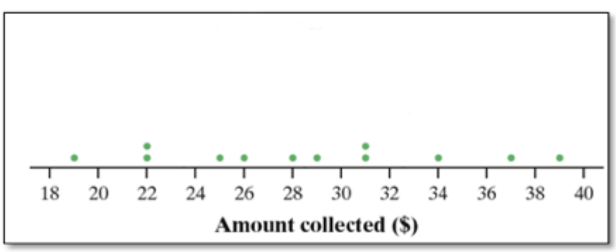
be sure to read the details on quartiles

pp. 63 to 65

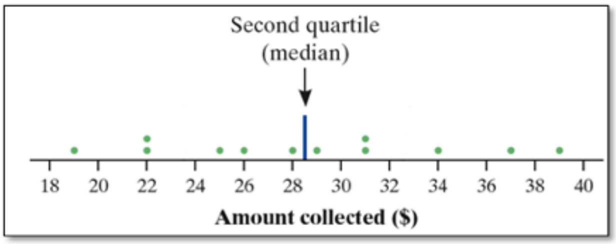
Measuring Variability: The Interquartile Range (IQR)

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.



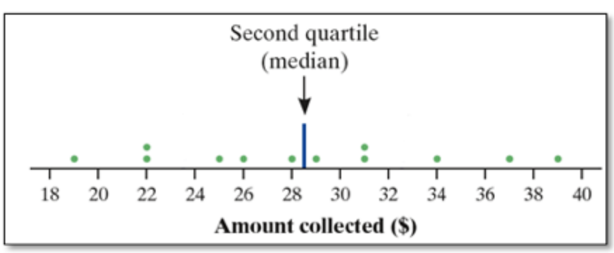
The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.



The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

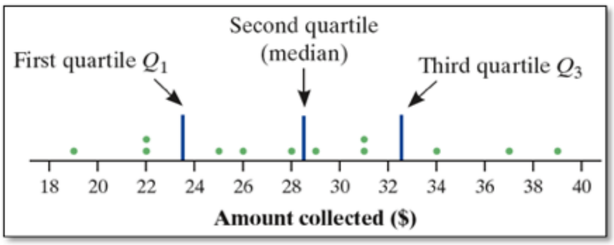
The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.



The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

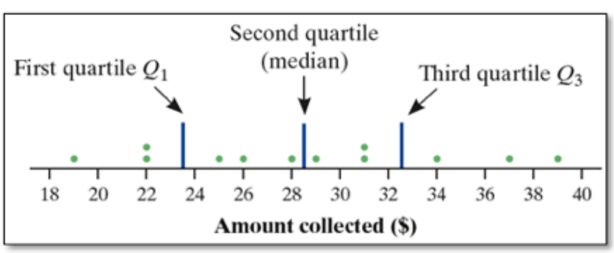
The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.



The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

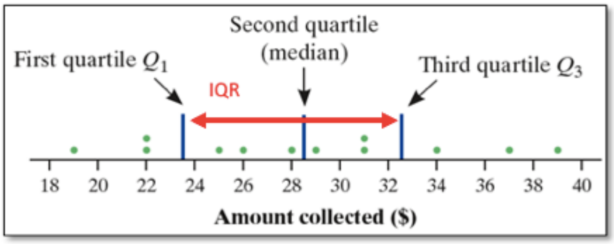
The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.



The **interquartile range (IQR)** is the distance between the first and third quartiles of a distribution.
In symbols:
$IQR = Q_3 - Q_1$

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.



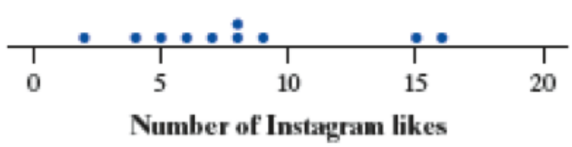
The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.

The **interquartile range (IQR)** is the distance between the first and third quartiles of a distribution. In symbols:
$IQR = Q_3 - Q_1$

8. Find the Interquartile Range

2	4	5	6	7
8	8	9	15	16



2 4 5 6 7 , 8 8 9 15 16 80

$Q_1 = 5$

Q_2
$\frac{7+8}{2}$
7.5

$Q_3 = 9$

$IQR = Q_3 - Q_1 = 9 - 5 = 4$ likes

Assignment

$\boxed{1.3}$ ... 87, 89, 91, 95, 97, 101, 103, 105, 121

hey mr. C
got it right
this time

Travel times for 20 New Yorkers:

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

Travel times for 20 New Yorkers:

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	15	15	15	20	20	20	25	30	30	40	40	45	60	60	65	85

Travel times for 20 New Yorkers:

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85

Travel times for 20 New Yorkers:

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85

$Q_1 = 15$ *Median* = 22.5 $Q_3 = 42.5$

Travel times for 20 New Yorkers:

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85

$Q_1 = 15$ *Median* = 22.5 $Q_3 = 42.5$

$$
\begin{aligned}
IQR &= Q_3 - Q_1 \\
&= 42.5 - 15 \\
&= 27.5 \text{ minutes}
\end{aligned}
$$

Travel times for 20 New Yorkers:

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85

$Q_1 = 15$ *Median* = 22.5 $Q_3 = 42.5$

$$IQR = Q_3 - Q_1$$
$$= 42.5 - 15$$
$$= 27.5 \text{ minutes}$$

Interpretation: The range of the middle half of travel times for the New Yorkers in the sample is 27.5 minutes.

8. **Find the Interquartile Range**

2	4	5	6	7
8	8	9	15	16