## Uni+ 6: Describing Da+a

## QUIZ: NEXT WEEK <br> TEST: WEEK AFTER SPRING BREAK

vocab to KNOW

## VOCAB TO KNOW

$\times$ Histogram shows frequencies of data values in intervals of the same size

x Box Plot (Box-and-Whisker): shows the variability of a data set using quartiles
$\times$ Dot Plot: shows the number of times each value occurs in a data set


## MEASURES OF CENTER

A measure of center is a measure that represents the center, or typical value, of a data set.
$\Rightarrow$ The mean, median, and mode are measures of center.
x Mean: The mean of a numerical data set is the sum of the data divided by the number of data values. The symbol $\bar{x}$ represents the mean. It is read as " $x$-bar."
$\times \quad$ Median:
The median of a numerical data set is the middle number when the values are written in numerical order. When a data set has an even number of values, the median is the mean of the two middle values.
$\times$
Mode: The mode of a data set is the value or values that occur most often. There may be one mode, no mode, or more than one mode.

Outlier: An outlier is a data value that is much greater than or much less than the other values in a data set.

## MEASURES OF CENTER

$>$ The mean, median, and mode are measures of center.
a) Find the mean, median, and mode of the hourly wages.

| Students' Hourly Wages |  |
| :---: | :---: |
| $\$ 16.50$ | $\$ 8.25$ |
| $\$ 8.75$ | $\$ 8.45$ |
| $\$ 8.65$ | $\$ 8.25$ |
| $\$ 9.10$ | $\$ 9.25$ |

SOLUTION
a. Mean $\quad \bar{x}=\frac{16.5+8.75+8.65+9.1+8.25+8.45+8.25+9.25}{8}=9.65$

Median $8.25,8.25,8.45,8.65,8.75,9.10,9.25,16.50$ Order the data.

$$
\frac{17.4}{2}=8.7
$$

Mean of two middle values
Mode $8.25,8.25,8.45,8.65,8.75,9.10,9.25,16.50 \quad 8.25$ occurs most often.
The mean is $\$ 9.65$, the median is $\$ 8.70$, and the mode is $\$ 8.25$.
Mean $=\$ 9.65$, Median $=\$ 8.70$, Mode $=\$ 8.25$
b) Which measure of center best represents the data? Explain.
$\mathbf{x}$ The median best represents the data.
X The mode is less than most of the data, and the mean is greater than most of the data.

## MEMSTRES OF CENTER $>$ The mean, median, and mode are measures of center

a) Find the mean, median, and mode of the hourly wages.

| Students' Hourly Wages |  |
| :---: | :---: |
| $\$ 16.50$ | $\$ 8.25$ |
| $\$ 8.75$ | $\$ 8.45$ |
| $\$ 8.65$ | $\$ 8.25$ |
| $\$ 9.10$ | $\$ 9.25$ |

Or... We can work smarter and not harder... and use Desmos https://www.desmos.com/calculator/f Ipios9rem

$$
A=[16.50,8.25,8.75,8.45,8.65,8.25,9.10,9.25]
$$

$$
A=8 \text { element list }
$$

b) Which measure of center best represents the data? Explain.
$\mathbf{x}$ The median best represents the data.
stats $(A)$
$\mathbf{x}$ The mode is less than most of the data, and

## MEMSTRES OF CENTER $>$ The mean, median, and mode are measures of center.

Or... We can work smarter and not harder... and use Desmos https://www.desmos.com/calculator/f Ipios9rem


## Symmetric? Skewed left? Skewed right?



| stats(A) |  | $\times$ |
| :---: | :---: | :---: |
|  | Min | 8.25 |
|  | Q1 | 8.35 |
|  | Median | 8.7 |
|  | Q3 | 9.175 |
|  | Max | 16.5 |
| mean(A) |  | $\times$ |
|  |  | $=9.65$ |
| median(A) |  | $\times$ |
|  |  | $=8.7$ |

## MEASURES OF CENTER

b) Which measure of center best represents the data? Explain. Choosing Appropriate Measures
When a data distribution is symmetric,

- use the mean to describe the center and
- use the standard deviation to describe the variation.

When a data distribution is skewed,

- use the median to describe the center and
- use the five-number summary to describe the variation.



## MEASURES OF CENTER

b) Which measure of center best represents the data? Explain.

- Answer: The median best represents the data.
$>$ When your data is SKEWED... it is best to use the MEDIAN. The median would be the measure which most accurately represents the center of the data.
$>$ The mode is less than most of the data
$>$ The mean is greater than most of the data


## MEMSTRES OF CENTER $>$ The mean, median, and mode are measures of center

OUTLIERS: An outlier is a data value that is much greater than or much less than the other values in a data set.

- Outliers usually have the greatest effect on the mean.

```
\(A=[16.50,8.25,8.75,8.45,8.65,8.25,9.10,9.25]\)
Mean \(=\$ 9.65\), Median \(=\$ 8.70\), Mode \(=\$ 8.25\)
```


## Consider the data in Example 1

(a) Identify the outlier. How does the outlier affect the mean, median, and mode?

| Students' Hourly Wages |  |
| :---: | :---: |
| $\$ 16.50$ | $\$ 8.25$ |
| $\$ 8.75$ | $\$ 8.45$ |
| $\$ 8.65$ | $\$ 8.25$ |
| $\$ 9.10$ | $\$ 9.25$ |

## MEMSTRES OF CENTER $>$ The mean, median, and mode are measures of center.

OUTLIERS: An outlier is a data value that is much greater than or much less than the other values in a data set.

- Outliers usually have the greatest effect on the mean.


Consider the data in Example 1 .
(a) Identify the outlier. The value $\$ 16.50$ is the outlier because it is much greater than the other wages
(b) How does the outlier affect the mean, median, and mode?

$$
\text { Mean }=\$ 8.67, \text { Median }=\$ 8.65, \text { Mode }=\$ 8.25
$$

When you remove the outlier, the mean decreases $\$ 9.65-\$ 8.67=\$ 0.98$
The median decreases $\$ 8.70-\$ 8.65=\$ 0.05$
The mode is the same.

## MEASURES OF UARIATION

$\Rightarrow$ A measure of variation is a measure that describes the spread, or distribution, of a data set
$>$ Range and standard deviation
$\times$ Range: difference of the greatest value and the least value.

| Show A |  |
| :---: | :--- |
| Ages |  |
| 20 | 29 |
| 19 | 22 |
| 25 | 27 |
| 27 | 29 |
| 30 | 20 |
| 21 | 32 |

Two reality cooking shows select 12 contestants each. The ages of the contestants are shown in the tables. Find the range of the ages for each show. Compare your results.

## SOLUTION

Show A 19, 20, 20, 21, 22, 25, 27, 27, 29, 29, 30, $31 \quad$ Order the data.
So, the range is $31-19$, or 12 years.
Show B

$$
\begin{aligned}
& 19,20,21,22,22,24,25,25,27,27,32,48 \\
& \text { So, the range is } 48-19 \text {, or } 29 \text { years. }
\end{aligned}
$$

The range of the ages for Show A is 12 years, and the range of the ages for Show B is 29 years. So, the ages for Show B are more spread out.

## MEASURES OF UARIATION

$>$ A measure of variation is a measure that describes the spread, or distribution, of a data set
$>$ Range and standard deviation

## Standard Deviation

The standard deviation of a numerical data set is a measure of how much a typical value in the data set differs from the mean. The symbol $\sigma$ represents the standard deviation. It is read as "sigma." It is given by

$$
\sigma=\sqrt{\frac{\left(x_{1}-\bar{x}\right)^{2}+\left(x_{2}-\bar{x}\right)^{2}+\cdots+\left(x_{n}-\bar{x}\right)^{2}}{n}}
$$

where $n$ is the number of values in the data set. The deviation of a data value $x$ is the difference of the data value and the mean of the data set, $x-\bar{x}$.

Step 1 Find the mean, $\bar{x}$.
Step 2 Find the deviation of each data value, $x-\bar{x}$.
Step 3 Square each deviation, $(x-\bar{x})^{2}$.
Step 4 Find the mean of the squared deviations. This is called the variance.
Step 5 Take the square root of the variance.

## MEASURES OF UARIATION

$\Rightarrow$ A measure of variation is a measure that describes the spread, or distribution, of a data set

- Range and standard deviation

The meaning of standard deviation is what is important... What is its significance??

OR the short way... (Desmos)

The long way...

|  |  |  |  | Find the standard deviation of the ages for Show A in Example 3. Use a table to organize your work. Interpret your result. <br> SOLUTION |
| :---: | :---: | :---: | :---: | :---: |
| $x$ | $\bar{x}$ | $x-\bar{x}$ | $(x-\bar{x})^{2}$ | Step 1 Find the mean, $\bar{x}$. |
| 20 | 25 | -5 | 25 | $\bar{x}=\frac{300}{12}=25$ |
| 29 | 25 | 4 | 16 | Step 2 Find the deviation of each data value, $x-\bar{x}$, as shown. |
| 19 | 25 | -6 | 36 | Step 3 Square each deviation, $(x-\bar{x})^{2}$, as shown. |
| 22 | 25 | -3 | 9 |  |
| 25 | 25 | 0 | 0 | Step 4 Find the mean of the squared deviations, or variance. |
| 27 | 25 | 2 | 4 | $\frac{\left(x_{1}-\bar{x}\right)^{2}+\left(x_{2}-\bar{x}\right)^{2}+\cdots+\left(x_{n}-\bar{x}\right)^{2}}{n}=\frac{25+16+\cdots+36}{12}=\frac{212}{12} \approx 17.7$ |
| 27 | 25 | 2 | 4 | Step 5 Use a calculator to take the square root of the variance. |
| 29 | 25 | 4 | 16 | $\left(x_{1}-\bar{x}\right)^{2}+\left(x_{2}-\bar{x}\right)^{2}+\cdots+\left(x_{n}-\bar{x}\right)^{2}{ }^{2}=\sqrt{212}$ |
| 30 | 25 | 5 | 25 | $1 \frac{\left(x_{1}-\bar{x}\right)^{2}+\left(x_{2}-x^{2}\right.}{n}=\sqrt{12}$ |
| 20 | 25 | -5 | 25 | - The standard deviation is about 4.2. This means that the typical age of a |
| 21 | 25 | -4 | 16 | contestant on Show A differs from the mean by about 4.2 years. |
| 31 | 25 | 6 | 36 |  |

DOT PLOTS

## DOT PLOTS: DISTRIBUTIONS

The dot plot uses a number line to show the number of times each value in a data set occurs.

- show clusters, peaks, and gaps in a data set
- can also use a dot plot to identify the shape of a distribution



## DOT PLOTS: EXAMPIES

1. Televisions in households:
$2,4,3,6,2,1,3,4,0,1,3,3,2,2,5,2,1,7,5,4$
A peak occurs at 2 . The distribution is skewed right.

2. Ages of new drivers:
$15,16,16,17,16,15,15,18,16,17,18,18,15,17,17,18$
All the dots are the same height. The distribution is uniform.
3. Heights of basketball players (in inches):
$74,79,80,81,71,73,73,72,78,79,80,79,72,73$
There are two clusters of dots. Peaks occur at 73 inches and 79 inches. There is a gap between 74 inches and 78 inches.


The distribution is symmetric.


## HISTOGRAMS: DISTRIBUTIONS

s.


Skewed left

- The "tail" of the graph extends to the left.
- Most of the data are on the right.


Symmetric

- The data on the right of the distribution are approximately a mirror image of the data on the left of the distribution.


Skewed right

- The "tail" of the graph extends to the right.
- Most of the data are on the left.


## HISTOGRAMS: DISTRIBUTIONS

## Types of Histograms

Symmetrical


## Uniform

(all frequencies the same) (no visible trend)


Skewed Left
(most data to left of high point)


Skewed Righ
(most data to right of high point)


## HISTOGRAMS: DISTRIBUTIONS

## Choosing Appropriate Measures

When a data distribution is symmetric,

- use the mean to describe the center and
- use the standard deviation to describe the variation.


When a data distribution is skewed,

- use the median to describe the center and
- use the five-number summary to describe the variation.



## HISTOGRAMS: COMPARING

The table shows the results of a survey that asked men and women how many pairs of shoes they own.
a. Make a double box-and-whisker plot that represents the data. Describe the shape of each distribution.
b. Compare the number of pairs of shoes owned by men to the number of pairs of shoes owned by women.
c. About how many of the women surveyed would you expect to own between 10 and 18 pairs of shoes?

|  | Men | Women |
| :--- | :---: | :---: |
| Survey size | 35 | 40 |
| Minimum | 2 | 5 |
| Maximum | 17 | 24 |
| 1st Quartile | 5 | 12 |
| Median | 7 | 14 |
| 3rd Quartile | 10 | 17 |
| Mean | 8 | 14 |
| Standard deviation | 3 | 4 |



## BOX PLOTS



## BOX PLOTS: DEFINITIONS

## Box-and-Whisker Plot

A box-and-whisker plot shows the variability of a data set along a number line using the least value, the greatest value, and the quartiles of the data. Quartiles divide the data set into four equal parts. The median (second quartile, Q2) divides the data set into two halves. The median of the lower half is the first quartile, Q1. The median of the upper half is the third quartile, Q3.


The five numbers that make up a box-and-whisker plot are called the five-number summary of the data set.

## BOX PLOTS: DEFINITIONS



Another measure of variation for a data set is the interquartile range (IQR), which is the difference of the third quartile, Q3, and the first quartile, Q1. It represents the range of the middle half of the data.

## BOX PLOTS: MAKING A BOX PLOT

Make a box-and-whisker plot that represents the ages of the members of a backpacking expedition in the mountains.
$24,30,30,22,25,22,18,25,28,30,25,27$

## SOLUTION

Step 1 Order the data. Find the median and the quartiles.


Step 2 Draw a number line that includes the least and greatest values. Graph points above the number line for the five-number summary.
Step 3 Draw a box using Q1 and Q3. Draw a line through the median. Draw whiskers from the box to the least and greatest values.


## BOX PLOTS: EXAMPLE

The box-and-whisker plot represents the lengths (in seconds) of the songs played by a rock band at a concert.

a. Find and interpret the range of the data.
b. Describe the distribution of the data.
c. Find and interpret the interquartile range of the data.
d. Are the data more spread out below Q1 or above Q3? Explain.

## BOX PLOTS: COMPARING

## SOLUTION

a.

$>$ The distribution for men is skewed right, and the distribution for women is symmetric.
b. The centers and spreads of the two data sets are quite different from each other. The mean for women is twice the median for men, and there is more variability in the number of pairs of shoes owned by women.
c. Assuming the symmetric distribution is bell-shaped, you know about $68 \%$ of the data lie within 1 standard deviation of the mean. Because the mean is 14 and the standard deviation is 4 , the interval from 10 to 18 represents about $68 \%$ of the data. So, you would expect about $0.68 \cdot 40 \approx 27$ of the women surveyed to own between 10 and 18 pairs of shoes.

## BOX PLOTS: EXAMPLE

## SOLUTION

a. The least value is 160 . The greatest value is 300 .

So, the range is $300-160=140$ seconds. This means that the song lengths vary by no more than 140 seconds.
b. Each whisker represents $25 \%$ of the data. The box represents $50 \%$ of the data. So,

- $25 \%$ of the song lengths are between 160 and 220 seconds.
- $50 \%$ of the song lengths are between 220 and 280 seconds.
- $25 \%$ of the song lengths are between 280 and 300 seconds.
c. $\mathrm{IQR}=\mathrm{Q} 3-\mathrm{Q} 1=280-220=60$
$>$ So, the interquartile range is 60 seconds. This means that the middle half of the song lengths vary by no more than 60 seconds.
d. The left whisker is longer than the right whisker.

So, the data below Q1 are more spread out than data above Q3.

## BOX PLOTS: DISTRIBUTION

## Shapes of Box-and-Whisker Plots



Skewed left

- The left whisker is longer than the right whisker.
- Most of the data are on the right side of the plot.


Symmetric

- The whiskers are about the same length.
- The median is in the middle of the plot.


Skewed right

- The right whisker is longer than the left whisker.
- Most of the data are on the left side of the plot.


## BOX PLOTS: COMPARING

The double box-and-whisker plot represents the test scores for your class and your friend's class.

a. Identify the shape of each distribution.
b. Which test scores are more spread out? Explain.

SOLUTION
a. For your class, the left whisker is longer than the right whisker, and most of the data are on the right side of the plot. For your friend's class, the whisker lengths are equal, and the median is in the middle of the plot.
$>$ So, the distribution for your class is skewed left, and the distribution for your friend's class is symmetric.
b. The range and interquartile range of the test scores in your friend's class are greater than the range and interquartile range in your class.
$>$ So, the test scores in your friend's class are more spread out.

## PRACTICE TOGETHER

Find the mean, median, and mode of the data. (Section 7.1)
1.

3.

Hours Spent on Project

| $3 \frac{1}{2}$ | 5 | $2 \frac{1}{2}$ |
| :---: | :---: | :---: |
| 3 | $3 \frac{1}{2}$ | $\frac{1}{2}$ |

2. 

| Checkbook Balances (dollars) |  |  |
| :---: | ---: | ---: |
| 40 | 10 | -20 |
| 0 | -10 | 40 |
| 30 | 40 | 50 |

4. Students in a Grade

| Stem | Leaf |  |  |
| ---: | :--- | :--- | :--- |
| 9 | 4 | 9 |  |
| 10 | 1 | 2 | 6 |
| 11 | 3 | 3 |  |
| 12 | 0 |  |  |

Key: $10 \mid 6=106$ students

## PRACTICE TOGETHER

Make a box-and-whisker plot for the data. (Section 7.2)
5. Hours spent on each babysitting job: 2, 4, 7, 5, 4, 1, 7, 4
6. Minutes of violin practice: $20,50,60,40$, 40, 30, 60, 40, 50, 20, 20, 35
7. Players' scores at end of first round: 200, $-100,100,350,-50$, $0,-50,300$

## PRACTICE TOGETHER


10. QUIZ SCORES The graph shows the quiz scores of students in a class. (Section 7.1)
a. Identify the outlier.
b. Which measure of central tendency will be most affected by removing the outlier? Explain.

