

# Measures of Center and Spread

# Measures of Center

- Most commonly used measures of center
  - Mean
  - Median

# Mean

- The mean is also known as
  - *the average*
  - *the arithmetic mean*

# Mean

- The mean is also known as
  - *the average*
  - *the arithmetic mean*
- How do we calculate the mean?

# Mean

- The mean is also known as
  - *the average*
  - *the arithmetic mean*
- To calculate the mean, we take the *sum of all the data values* and *divide this sum by the number of data values.*

# Terminology

- A descriptive measure for a population is called a *parameter*.
- A descriptive measure for a sample is called a *statistic*.

# Terminology

- A descriptive measure for a population is called a parameter.
  - The population mean is a *parameter*.
- A descriptive measure for a sample is called a statistic.
  - The sample mean is a *statistic*.

# Sample Mean

- Suppose that we have  $n$  data values. Let us represent these data values with a subscripted  $x$  to distinguish between the data values, That is, suppose  $x_1, x_2, x_3, \dots, x_n$  are the data.

Then, the sample mean, denoted by  $\bar{x}$ , is given by

$$\bar{x} = (x_1 + x_2 + x_3 + \dots + x_n) / n$$

# Sample Mean

- Suppose  $x_1, x_2, x_3, \dots, x_n$  are the  $n$  data values.

Using summation notation, the sample mean can be expressed as

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

$\sum x_k$  denotes taking the sum of  $x_1, x_2, x_3, \dots,$  and  $x_n$ .

# Population Mean

- Suppose that we have  $n$  data values. Let us represent these data values with a subscripted  $x$  to distinguish between the data values, That is, suppose  $x_1, x_2, x_3, \dots, x_N$  are the data.

Then, the population mean, denoted by  $\mu$ , is given by

$$\mu = (x_1 + x_2 + x_3 + \dots + x_n) / N$$

# Population Mean

- Suppose  $x_1, x_2, x_3, \dots, x_N$  are the  $N$  data values.

Using summation notation, the population mean can be expressed as

$$\mu = \frac{\sum x_k}{N}$$

$\sum x_k$  denotes taking the sum of  $x_1, x_2, x_3, \dots$ , and  $x_N$ .

# Mean

- The mean is the *balance point* (center of gravity) for a distribution.
- Visual estimate: place your finger below the point on the horizontal axis of a dot plot or histogram so that you can *balance* it, half the weight to the left and half the weight to the right

# Mean

- The mean is the *balance point* (center of gravity) for a distribution.
- Visual estimate: for a distribution that is *approximately normal*, the mean will be directly below the highest point on the bell curve, the tallest stack of dots for a dot plot, and the tallest bar in a histogram

# Median

- The median is the middle value for the distribution.

# Median

- The median is the middle value for the distribution.
- The median divides the distribution into two halves.

# Median

- The median is the middle value for the distribution.
- The median divides the distribution into two halves.

**Caution:** The median is the *physical* middle value for the distribution - the middle *data value*

# Median

- The median is the middle value for the *data values in numerical order*.
- The median divides the distribution into two halves.

*Caution*: The median is the *physical* middle value - the actual *value* is not considered other than for order

# Determining the Median

- Arrange the data in numerical order
- Determine the number of data values
- Count off data values from one end to the middle value

# Determining the Median

- If there are an odd number of data values, the median is the middle data value
- If there are an even number of data values, the median is the average of the middle two data values.

# Determining the Median

- If there are an odd number of data values, the median is the *middle data value*
- If there are an even number of data values, the median is the *average of the middle two data values*.

# Units on the Mean

- What units would you expect to have on the mean for the
  - age?
  - height?
  - Speed of the non-predators?

# Units on the Mean

- What units would you expect to have on the mean for the
  - age?
  - height?
  - Speed of the non-predators?
- The units on the mean are the same as the units on the original data.

# Units on the Median

- What units would you expect to have for the median of the
  - Speeds for the animal data?
  - The salaries for the FSC faculty?
  - The speeds of 100 randomly selected drivers on Route 9 at 10:30 am?

# Units on the Median

- What units would you expect to have for the median of the
  - Speeds for the animal data?
  - The salaries for the FSC faculty?
  - The speeds of 100 randomly selected drivers on Route 9 at 10:30 am?
- The units on the median are the same as the units on the original data.

# Range

- The range of the data is the difference between the *largest* and the *smallest* data values.
- The range only involves the largest and smallest values of the data, and tells us *nothing* about the center and spread of the data around the mean or around the median.

# Measuring Spread Around the Median

- When determining the median, we divide the distribution into halves
  - Dividing each of these halves into halves, we determine the first or lower quartile,  $Q_1$ , and the third or upper quartile,  $Q_3$ , for the distribution.

# Measuring Spread Around the Median

- The first or lower quartile,  $Q_1$ , is the median of the lower half of the distribution.
- The third or upper quartile,  $Q_3$ , is the median for the upper half of the distribution.

# Measuring Spread Around the Median

- The median,  $Q_1$ , the first or lower quartile, and  $Q_3$ , the third or upper quartile, divide the distribution into quarters.
- That is, the median,  $Q_1$ , and  $Q_3$  divide the distribution into four pieces (fourths).

# Measuring Spread Around the Median

- The interquartile range, denoted IQR, is a measure of spread from the lower quartile to the upper quartile,

$$\text{IQR} = Q_3 - Q_1$$

# Five-Number Summary

- **Minimum** - The smallest value
- **Lower or first quartile,  $Q_1$**  - the median of the lower half of the values
- **Median** - the values that divides the data into halves
- **Upper or third quartile,  $Q_3$**  - the median of the upper half of the values
- **Maximum** - the largest value

# Outlier

- A value is considered to be an outlier if it is more than 1.5 times the interquartile range, IQR, from the nearest quartile.

# Outlier

- *More than 1.5 times the interquartile range, IQR, from the nearest quartile means that*
  - the value is more than  
 $Q_3 + 1.5 \cdot \text{IQR}$
  - the value is less than  
 $Q_1 - 1.5 \cdot \text{IQR}$

# Box Plot

- A box plot is a graphical display of the five-number summary for a data set.
  - Minimum
  - $Q_1$
  - Median
  - $Q_3$
  - Maximum

# Box Plot

(a.k.a. Box and Whiskers Plot)

- To make a box plot,
  - Make a box that starts at  $Q_1$  and ends at  $Q_3$
  - Mark the median in the middle of the box
  - Make “whiskers” that extend from each quartile to the adjacent extreme value

# Box Plot

(a.k.a. Box and Whiskers Plot)

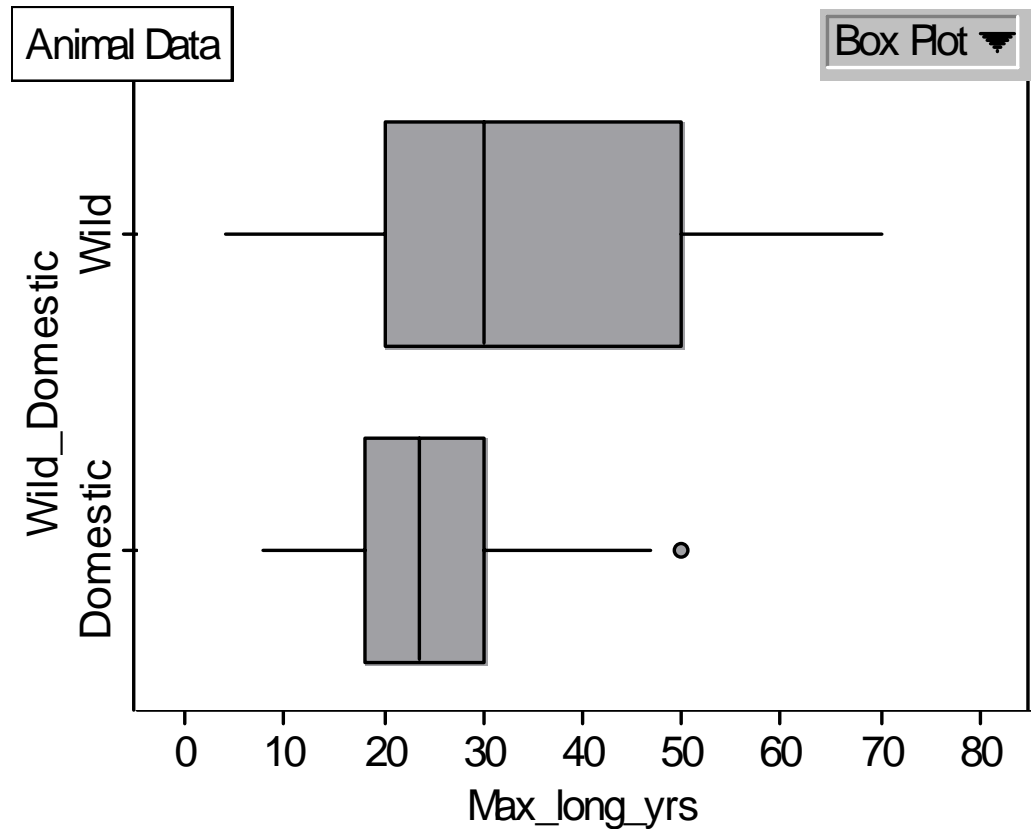
- To make a box plot,
  - Make a *rectangle* that starts at  $Q_1$  and ends at  $Q_3$
  - Mark the median in the middle of the rectangle
  - Make “whiskers” that extend from  $Q_1$  to the minimum and from  $Q_3$  to the maximum

# Modified Box Plot

- Similar to basic box plot except
  - Whiskers extend only as far as the largest and smallest non-outliers for the data
    - Other outliers are marked as individual dots or other symbols
  - Largest and smallest non-outliers are often called the *adjacent values*

# Modified Box Plot

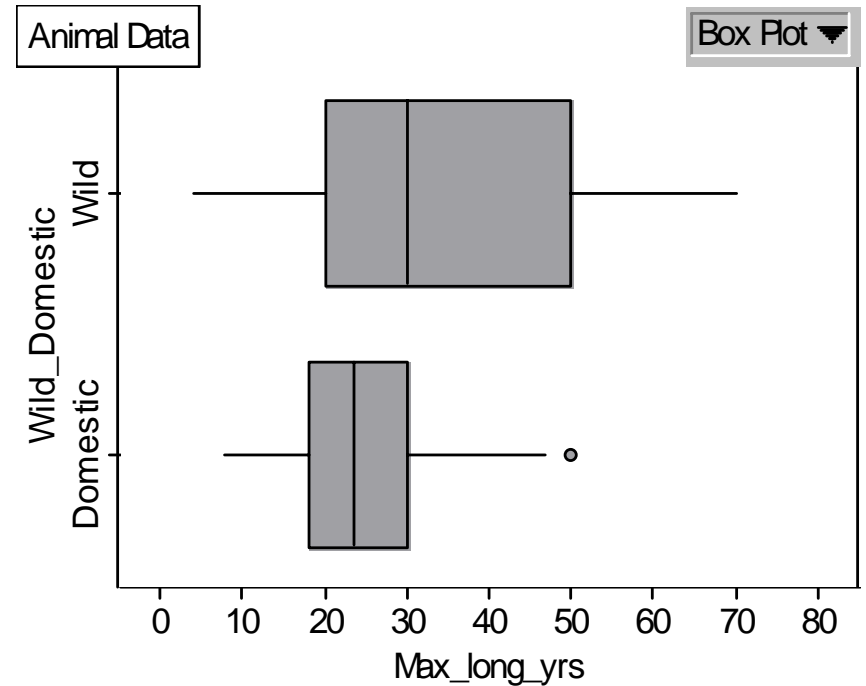
A modified box plot for the animal data. Notice that Fathom indicates the outlier with a dot.



# Modified Box Plot

Animal Data	Summary Table		
↓	Wild_Domestic		Row Summary
	Domestic	Wild	
Max_long_yrs	10	29	39
	26.1	32.413793	30.794872
	23.5	30	27
	18	20	20
	30	50	47
	12	30	27
	8	4	4
	50	70	70
	0	-25	-20.5
	48	95	87.5

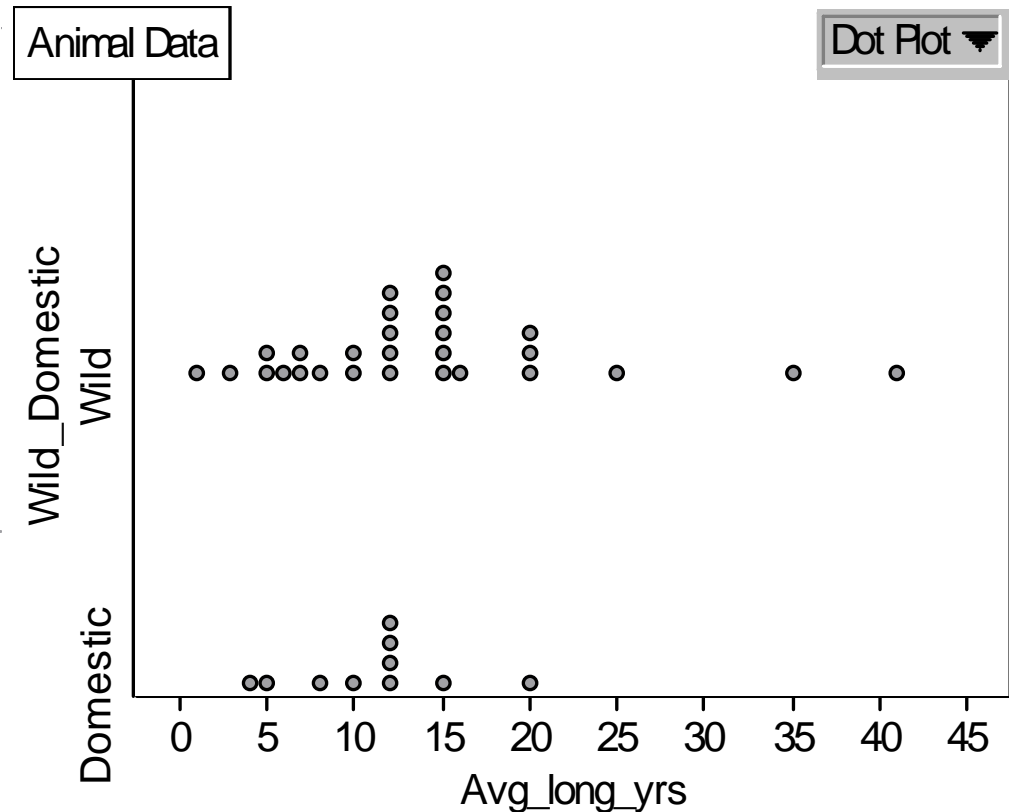
S1 = count ( )  
 S2 = mean ( )  
 S3 = median ( )  
 S4 = Q1 ( )  
 S5 = Q3 ( )  
 S6 = iqr ( )  
 S7 = min ( )  
 S8 = max ( )  
 S9 = Q1 ( ) - 1.5 ( iqr ( ) )  
 S10 = Q3 ( ) + 1.5 ( iqr ( ) )



# Average Longevity for Animal Data

Animal Data	Summary Table		
↓	Wild_Domestic		Row Summary
	Domestic	Wild	
Avg_long_yrs	10	28	38
	11	13.892857	13.131579
	12	12	12
	8	7.5	8
	12	15.5	15
	4	8	7
	4	1	1
	20	41	41
	2	-4.5	-2.5
	18	27.5	25.5

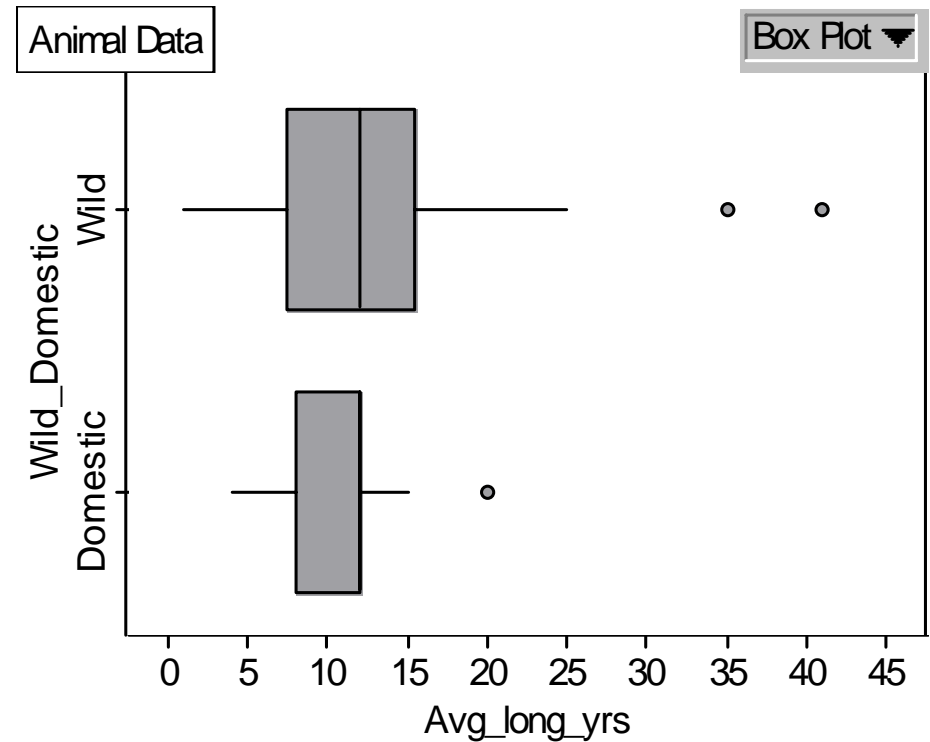
- S1 = count ( )
- S2 = mean ( )
- S3 = median ( )
- S4 = Q1 ( )
- S5 = Q3 ( )
- S6 = iqr ( )
- S7 = min ( )
- S8 = max ( )
- S9 = Q1 ( ) - 1.5 ( iqr ( ) )
- S10 = Q3 ( ) + 1.5 ( iqr ( ) )



# Average Longevity for Animal Data

Animal Data	Summary Table		
↓	Wild_Domestic		Row Summary
	Domestic	Wild	
Avg_long_yrs	10	28	38
	11	13.892857	13.131579
	12	12	12
	8	7.5	8
	12	15.5	15
	4	8	7
	4	1	1
	20	41	41
	2	-4.5	-2.5
	18	27.5	25.5

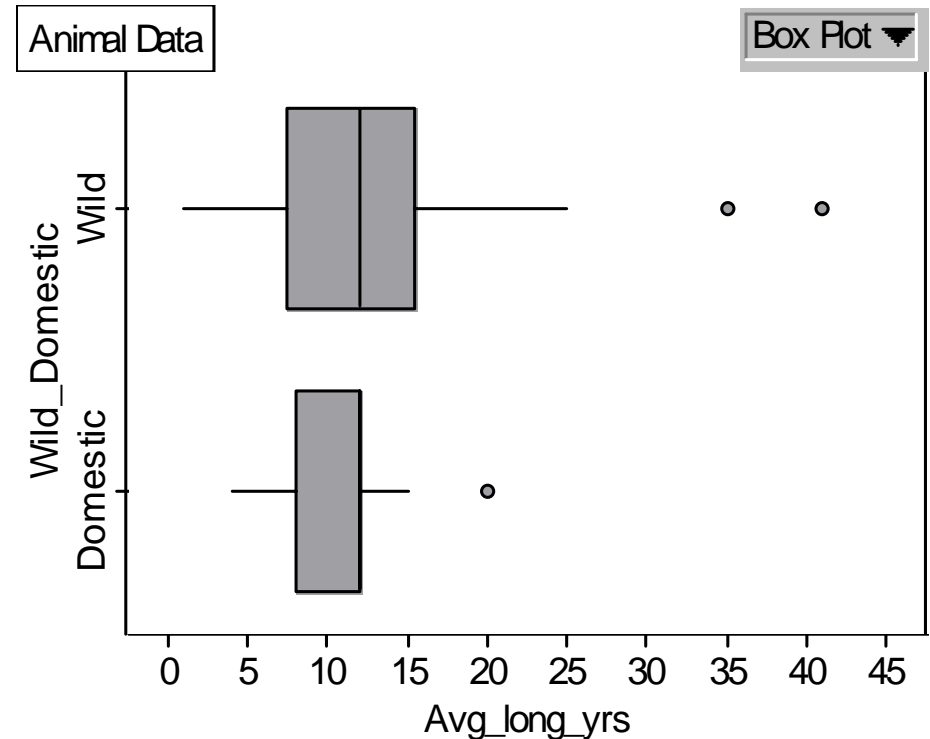
S1 = count ( )  
 S2 = mean ( )  
 S3 = median ( )  
 S4 = Q1 ( )  
 S5 = Q3 ( )  
 S6 = iqr ( )  
 S7 = min ( )  
 S8 = max ( )  
 S9 = Q1 ( ) - 1.5 ( iqr ( ) )  
 S10 = Q3 ( ) + 1.5 ( iqr ( ) )



# Average Longevity for Animal Data

Animal Data	Summary Table		
↓	Wild_Domestic		Row Summary
	Domestic	Wild	
Avg_long_yrs	10	28	38
	11	13.892857	13.131579
	12	12	12
	8	7.5	8
	12	15.5	15
	4	8	7
	4	1	1
	20	41	41
	2	-4.5	-2.5
	18	27.5	25.5

S1 = count ( )  
 S2 = mean ( )  
 S3 = median ( )  
 S4 = Q1 ( )  
 S5 = Q3 ( )  
 S6 = iqr ( )  
 S7 = min ( )  
 S8 = max ( )  
 S9 = Q1 ( ) - 1.5 ( iqr ( ) )  
 S10 = Q3 ( ) + 1.5 ( iqr ( ) )



Where is the median line for the domestic animals?

# Modified Box Plots

- Used for quantitative variable
- Does not record individual data values
- Records five-number summary of data with outliers
  
- **NOTE:** We will make a modified box plot for data whenever a box plot is requested.

# Box Plots

- Useful when plotting a single quantitative variable
  - Compare shape, center, spread for two or more distributions
  - When distribution has too many values or would require too much space to make a stemplot
  - Do not need to see individual values
  - Do not need more than five-number summary with outliers marked

# Exploring Spread Around The Mean

- Consider the following numbers:  
3, 5, 1, 8, 0, 7, 3, 6, 5, 2
- Determine the mean.
- Sketch a dot plot.
- Explore the spread around the mean.

# Exploring Spread Around The Mean

$x_k$

---

3

5

1

8

0

7

3

6

5

2

# Exploring Spread Around The Mean

$x_k$

$x_k - \bar{x}$

---

3

5

1

8

0

7

3

6

5

2

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$
-------	-----------------

---

3	-1
5	1
1	-3
8	4
0	-4
7	3
3	-1
6	2
5	1
2	-2

# Exploring Spread Around The Mean

$x_k$

$x_k - \bar{x}$

---

3

-1

5

1

1

-3

8

4

0

-4

7

3

3

-1

6

2

5

1

2

-2

deviations  
or  
deviations from  
the mean

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$
-------	-----------------

---

3	-1
5	1
1	-3
8	4
0	-4
7	3
3	-1
6	2
5	1
2	+ -2

---

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$
-------	-----------------

---

3	-1
5	1
1	-3
8	4
0	-4
7	3
3	-1
6	2
5	1
2	+ -2

---

0

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$
-------	-----------------

---

3	-1
5	1
1	-3
8	4
0	-4
7	3
3	-1
6	2
5	1
2	+ -2
	<hr/>
	0

The sum of the deviations from the mean is zero.

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$	$(x_k - \bar{x})^2$
-------	-----------------	---------------------

---

3	-1	1
5	1	1
1	-3	9
8	4	16
0	-4	16
7	3	9
3	-1	1
6	2	4
5	1	1
2	-2	4

# Exploring Spread Around The Mean

$x_k$        $x_k - \bar{x}$        $(x_k - \bar{x})^2$

---

3	-1	1
5	1	1
1	-3	9
8	4	16
0	-4	16
7	3	9
3	-1	1
6	2	4
5	1	1
2	-2	4

The squared deviations

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$	$(x_k - \bar{x})^2$
-------	-----------------	---------------------

---

3	-1	1
5	1	1
1	-3	9
8	4	16
0	-4	16
7	3	9
3	-1	1
6	2	4
5	1	1
2	-2	+ 4

---

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$	$(x_k - \bar{x})^2$
-------	-----------------	---------------------

---

3	-1	1
5	1	1
1	-3	9
8	4	16
0	-4	16
7	3	9
3	-1	1
6	2	4
5	1	1
2	-2	+ 4

---

62

# Exploring Spread Around The Mean

$x_k$	$x_k - \bar{x}$	$(x_k - \bar{x})^2$
-------	-----------------	---------------------

---

3	-1	1
5	1	1
1	-3	9
8	4	16
0	-4	16
7	3	9
3	-1	1
6	2	4
5	1	1
2	-2	4

---

+ 4

---

62

The sum of the squared deviations is not zero.

# Standard Deviation

- **Sample Standard Deviation**

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

- **Population Standard Deviation**

$$\sigma = \sqrt{\frac{\sum (x_k - \mu)^2}{N}}$$

# Standard Deviation

- **Sample Standard Deviation**

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

**Used for  
statistical  
inference**

- **Population Standard Deviation**

$$\sigma = \sqrt{\frac{\sum (x_k - \bar{x})^2}{N}}$$

# Standard Deviation

- **Sample Standard Deviation**

$$s = \sqrt{\frac{n \left( \sum x_k^2 \right) - \left( \sum x_k \right)^2}{n(n-1)}}$$

- **Population Standard Deviation**

$$\sigma = \sqrt{\frac{N \left( \sum x_k^2 \right) - \left( \sum x_k \right)^2}{N^2}}$$

# Standard Deviation

- Use  $\sigma$ , the *population standard deviation*, when you *know* all the values in a population
- Use  $s$ , the *sample standard deviation*, when you *have* a random sample *chosen* from the population

# Standard Deviation

- Use  $\sigma$ , the *population standard deviation*, when you *know all the values in a population*
- Use  $s$ , the *sample standard deviation*, when you *have a random sample chosen from the population*

# Standard Deviation

- Used to measure the spread of the data from the mean
- A measure of the dispersion of a distribution

# Standard Deviation

- Used to measure the spread of the data from the mean
- A measure of the dispersion of a distribution

What is "dispersion"?

# Standard Deviation

- Used to measure the spread of the data from the mean
- A measure of the **dispersion** of a distribution

**Dispersion is the degree of scatter of data around the mean. ...**

# Standard Deviation

- Used to measure the spread of the data from the mean
- A measure of the **dispersion** of a distribution

**Dispersion is the scattering of the values of a frequency distribution from the mean.**

...

# Standard Deviation

- Used to measure the spread of the data from the mean
- A measure of the **dispersion** of a distribution

**Dispersion is the spread of a distribution around the central value.**

# Standard Deviation

- Used to measure the spread of the data from the mean
- A measure of the **dispersion** of a distribution

Other measures of dispersion include the *semi-interquartile range* and the mean absolute deviation.

# Variance

- The variance of a set of values is a measure of variation equal to the square of the standard deviation.
- Variation is a general description of the amount that values vary among themselves. (The terms *dispersion* and *spread* are often used instead of *variation*.)

**Which summary statistics should I use to describe a distribution?**

# Which summary statistics should I use to describe a distribution?

- **Normal Distribution:** use mean and standard deviation
- **Skewed Distribution:** use median and quartiles

# So, what should I do first?

- Since we need to know the shape of the distribution in order to determine the distribution type, we should always start by graphing the distribution.

**What graph(s) should I use?**

# What graph(s) should I use?

- What graphs display the shape of the distribution?

# What graph(s) should I use?

- What graphs display the shape of the distribution?
  - Dot plots
  - Histograms

# Uses for mean and standard deviation for distributions that are not normal

- Suppose you have a representative sample for prices of cars of a particular make or class and you want to use the sample to represent the price of all the cars of this make or class
  - Use the mean

# Uses for mean and standard deviation for distributions that are not normal

- Suppose you have a representative sample for prices of cars of a particular make or class and you want to use the sample to represent the price of all the cars of this make or class
  - Use the mean **Why?**

# Uses for mean and standard deviation for distributions that are not normal

- Suppose you have a representative sample for prices of cars of a particular make or class and you want to use the sample to represent the price of all the cars of this make or class
  - Use the mean **Why?**
  - Sample means are approximately normal

# Recentering and Rescaling Data

- **Recentering** - adding the same number to all data values
  - Does not change the shape
  - Does not change the spread
  - "slides" (along horizontal axis for graph) distribution by amount  $c$ 
    - Changes mean and median by amount  $c$

# Recentering and Rescaling Data

- Rescaling - multiplying all data values by same *nonzero* number
  - Does not affect the *basic* shape
  - Stretches or shrinks the distribution
  - IQR multiplied by  $|d|$
  - Mean and median multiplied by  $d$

# Influence of Outliers

A summary statistic is

- *Resistant to outliers* if the summary statistic does not change very much if an outlier is removed from a data set

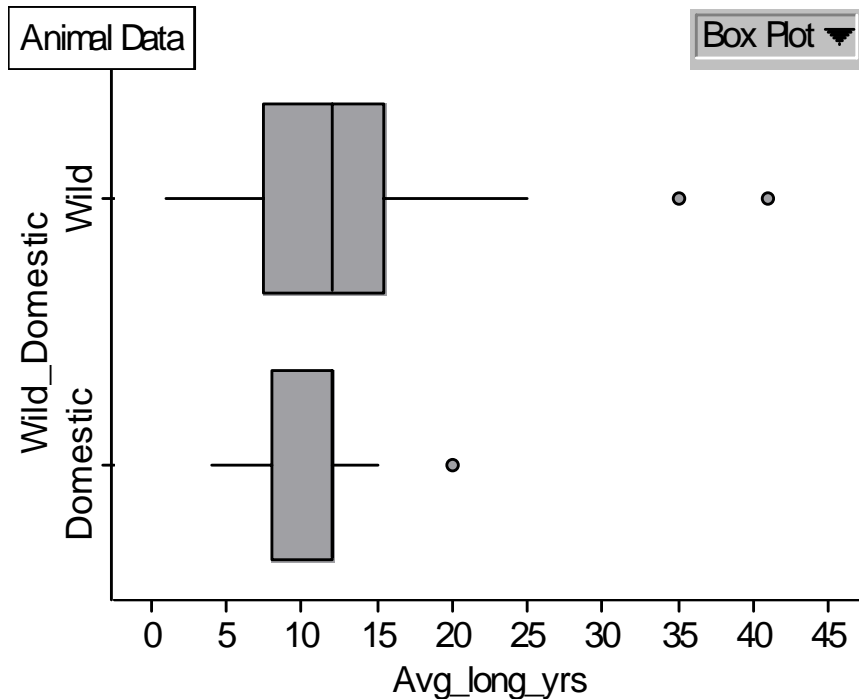
# Influence of Outliers

A summary statistic is

- *Sensitive to outliers* if the summary statistic changes when an outlier is removed from a data set

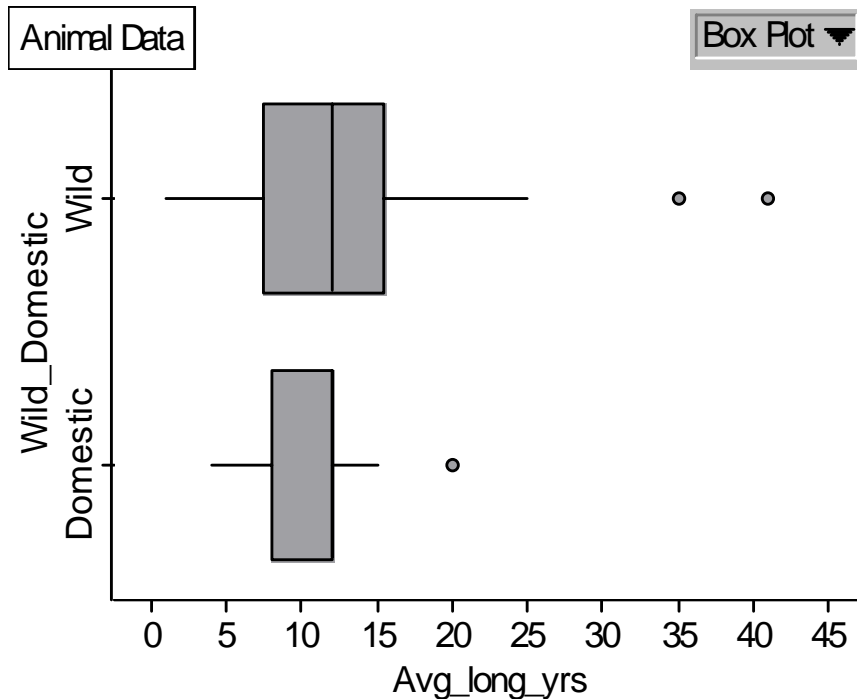
# Influence of Outliers

Let us explore the Average Longevity for the Animal data.



# Influence of Outliers

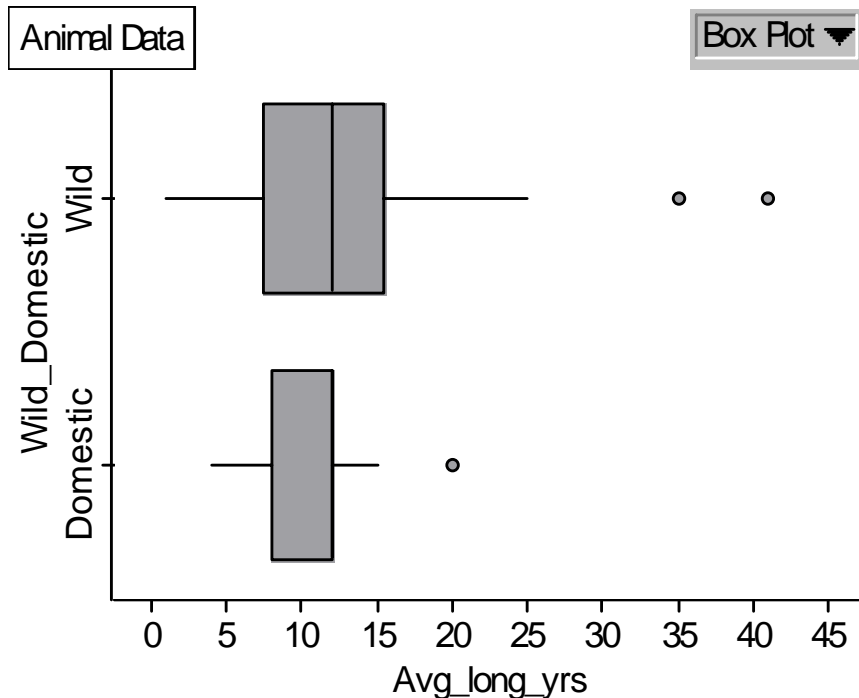
Let us explore the Average Longevity for the Animal data.



Are the measures of center affected by outliers?

# Influence of Outliers

Let us explore the Average Longevity for the Animal data.

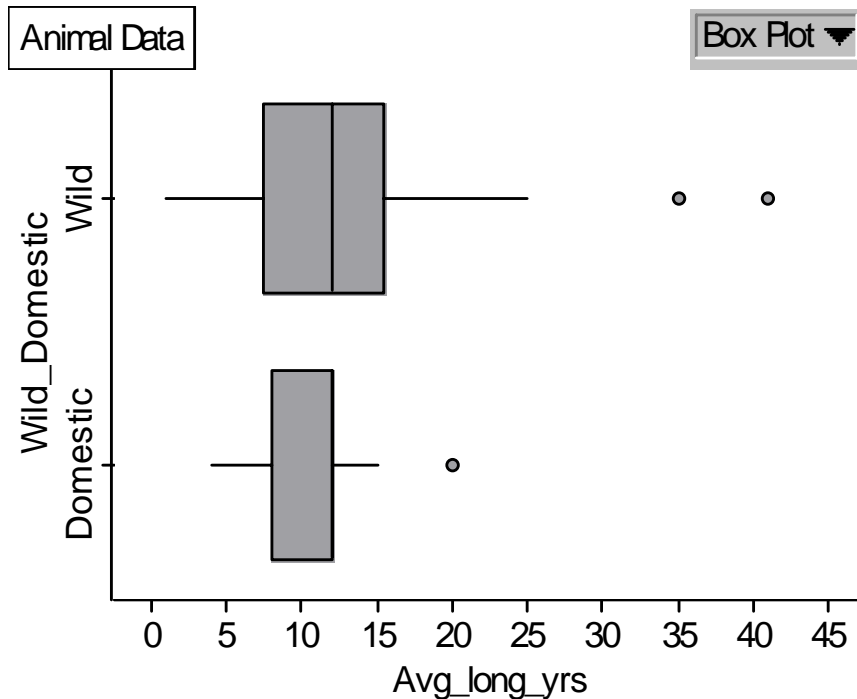


Are the measures of center affected by outliers?

**Mean?** **Median?**

# Influence of Outliers

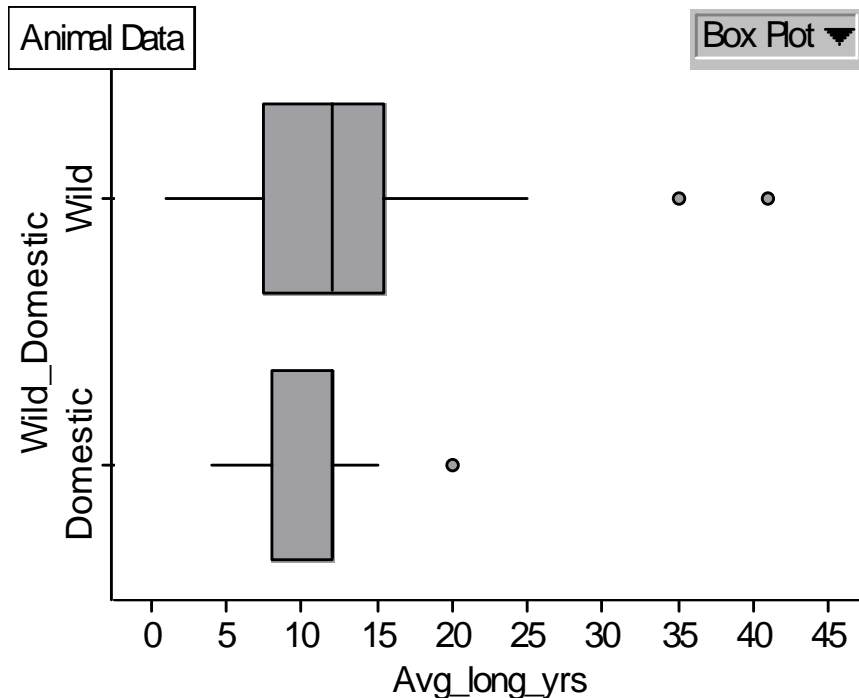
Let us explore the Average Longevity for the Animal data.



Are the measures of spread affected by outliers?

# Influence of Outliers

Let us explore the Average Longevity for the Animal data.



Are the measures of spread affected by outliers?

Range? IQR?  
Standard Deviation?

# Mean and Standard Deviation from a Frequency Table

- Consider the following data:

<u>Outcome</u>	<u>Frequency</u>
1	27
2	31
3	42
4	40
5	28
6	32

# Mean and Standard Deviation from a Frequency Table

- Suppose each data value  $x_k$  occurs with frequency  $f_k$ .
- The sample mean of a frequency table is given by

$$\bar{x} = \frac{\sum x_k \cdot f_k}{n}$$

where

$$\sum f_k = n$$

# Mean and Standard Deviation from a Frequency Table

- Suppose each data value  $x_k$  occurs with frequency  $f_k$ .
- The population mean of a frequency table is given by

$$\mu = \frac{\sum x_k \cdot f_k}{N}$$

where

$$\sum f_k = N$$

# Mean and Standard Deviation from a Frequency Table

- Suppose each data value  $x_i$  occurs with frequency  $f_i$ .
- The standard deviation of a frequency table is given by

$$s = \sqrt{\frac{\sum (x_k - \bar{x})^2 \cdot f_k}{n-1}}$$

# Mean and Standard Deviation from a Frequency Table

- Suppose each data value  $x_i$  occurs with frequency  $f_i$ .
- The standard deviation of a frequency table is given by

$$s = \sqrt{\frac{\sum \left\{ \left( x_k - \bar{x} \right)^2 \cdot f_k \right\}}{n-1}}$$

# Mean and Standard Deviation from a Frequency Table

- Suppose each data value  $x_k$  occurs with frequency  $f_k$ .
- The standard deviation of a frequency table is given by

$$s = \sqrt{\frac{\sum (x_k - \bar{x})^2 \cdot f_k}{n-1}}$$

**Sample Standard Deviation**

# Mean and Standard Deviation from a Frequency Table

- Suppose each data value  $x_k$  occurs with frequency  $f_k$ .
- The standard deviation of a frequency table is given by

$$\sigma = \sqrt{\frac{\sum (x_k - \bar{x})^2 \cdot f_k}{N}}$$

**Population Standard Deviation**

# Mean and Standard Deviation from a Frequency Table

- Consider the following data:

<u>Outcome</u>	<u>Frequency</u>
1	27
2	31
3	42
4	40
5	28
6	32

# Mean and Standard Deviation from a Frequency Table

- Consider the following data:

<u>Speed</u>	<u>Frequency</u>
42-45	25
46-49	14
50-53	7
54-57	3
58-61	1