Chapter 3

Numerical Summaries of Center and Variation

Describing Numerical Distributions

- We learned in Chapter 2 the things to always describe when considering a distribution
    - Shape - how many peaks, symmetric or skewed
    - Center - the “typical” value
    - Variability (Spread) - how spread out the data is

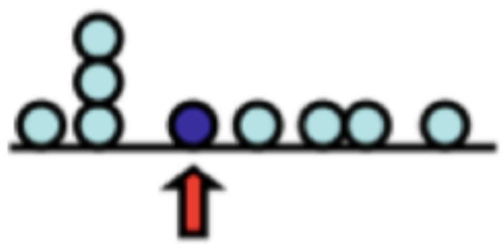
- In Chapter 3 we will learn about what values to use to measure center and spread.

Center - The Typical Value

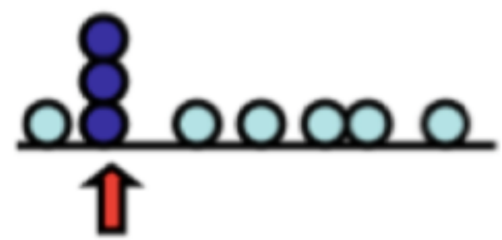
Mean is the arithmetic average

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

Median is the midpoint of ranked values



Mode is the most frequently observed value



Mean

- Mean is the arithmetic average.

Below are the final scores of 5 previous Stat 10 students:

79, 82, 94, 83, 92

The mean of this data is:

$$\frac{79+82+94+83+92}{5}=\frac{430}{5}=86$$

Example

Clicker!

What is the mean?

309, 278, 5, 311, 162, 24

Answer:

A. 160
B. 171
C. 181.5
D. 191.5

Example

What is the mean?

309, 278, 5, 311, 162, 24

Answer: 181.5

$$\frac{309+278+5+311+162+24}{6}=181.5$$

Median

- Median is the midpoint of ranked values.

Below are the final scores of 5 previous Stat 10 students:

79, 82, 94, 83, 92

The order to find the median we must first put the values in increasing order:

79, 82, 83, 92, 94

The median will be the value in the middle of the data set:

79, 82, **83**, 92, 94

Median

- Let's add an outlier to the example data set:

3, 79, 82, 83, 92, 94

Notice we now have an even number of values so finding the median requires an extra step:

3, 79, 82, 83, 92, 94

We take the mean (or average) of 82 and 83:

$$\frac{82+83}{2}=82.5$$

Example

Clicker!

- What is the median?

309, 278, 5, 311, 320, 321, 309, 355, 355

Answer:

A. 309

B. 355

C. 311

D. 320

Example

- What is the median?

309, 278, 5, 311, 320, 321, 309, 355, 355

Answer: 311

5, 278, 309, 309, **311**, 320, 321, 355, 355

Spread

- The standard deviation is described by the square root of the variance. The average distance of a value from the mean.

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

- The interquartile range (IQR) is the third quartile minus the first quartile:

IQR = Q3 - Q1

- The range is the maximum value minus the minimum value:

Range = Max - Min

Standard Deviation

- The standard deviation is described by the square root of the variance. The average distance of a value from the mean.

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

- It represents a typical distance from the mean of observations.
- $(x-\bar{x})^2$ means to take each data point, subtract the mean, and then square the difference, this is called a deviation.
- The $\sum$ (sigma) means to add up all the deviations.

Steps to Finding the Standard Deviation

1. Find the mean of your data.
2. Subtract the mean from each data point and then square those differences.
3. Add (sum) all the squared values from Step 2.
4. Divide the value from Step 3 by the number of your data points minus one. This gives you variance.
5. Take the square root of Step 4 (the variance) to get the standard deviation.

Standard Deviation Example

- Let's again consider the final scores of 5 Stats 10 students:

79, 82, 94, 83, 92

1. We found the mean, $\bar{x}$, to be 86.
2. Now subtract the mean from each value and square the difference:

$$(79-86)^2 = 49$$

$$(82-86)^2 = 16$$

$$(94-86)^2 = 64$$

$$(83-86)^2 = 9$$

$$(92-86)^2 = 36$$

Standard Deviation Example

3. Sum all the squared values from Step 2.

$$49 + 16 + 64 + 9 + 36 = 174$$

4. Divide Step 3 by the number of data points minus one.

$$174/(5-1) = 174/4 = 43.5$$

The variance is 43.5

5. Take the square root of the variance to get the standard deviation.

$$s = \sqrt{43.5} = 6.6$$

Example

Clicker!

What is the standard deviation?

1, 5, 5, 1, 8

Answer:

A. 1
B. 3
C. 5
D. 8

Example

What is the standard deviation?

1, 5, 5, 1, 8

Answer: 3

1) $\text{mean} = \dfrac{1+5+5+1+8}{5} = \dfrac{20}{5} = 4$

2)

$(1-4)^2 = 9$

$(5-4)^2 = 1$

$(5-4)^2 = 1$

$(1-4)^2 = 9$

$(8-4)^2 = 16$

3) $9+1+1+9+16 = 36$

4) $\dfrac{36}{(5-1)} = 9$

5) $\sqrt{9} = 3$

Comparing Standard Deviations

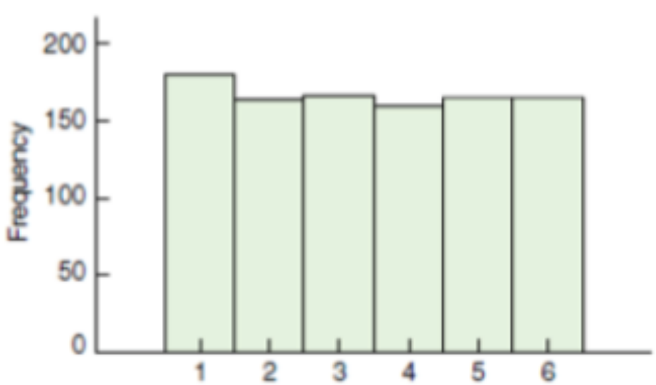
- Each distribution contains the same number of observations. The mean is about 3.5 for all 3 graphs.
- Which distribution has the largest standard deviation?

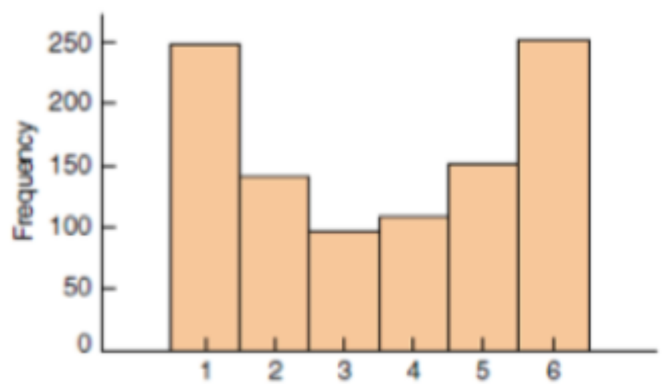
    A. Green

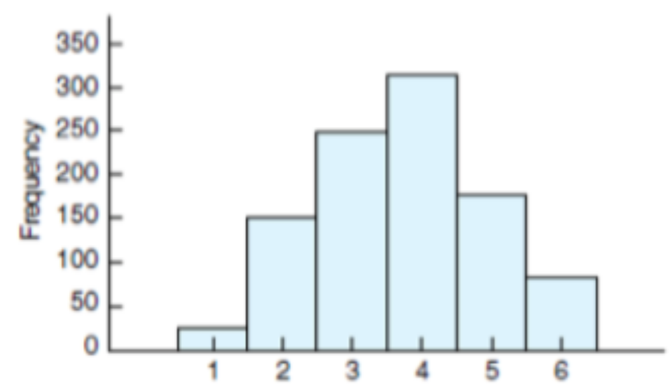
    B. Orange

    C. Blue

Clicker!







Comparing Standard Deviations

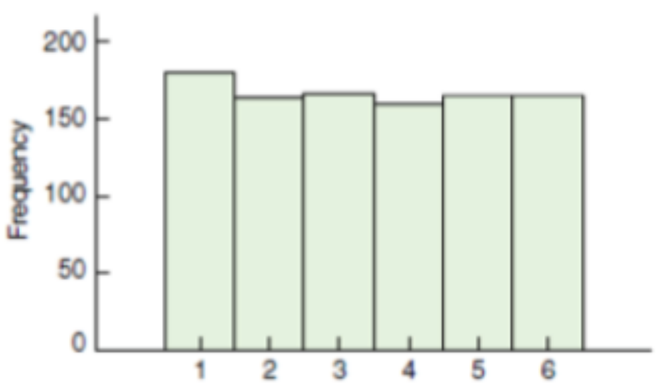
- Each distribution contains the same number of observations. The mean is about 3.5 for all 3 graphs.
- Which distribution has the smallest standard deviation?

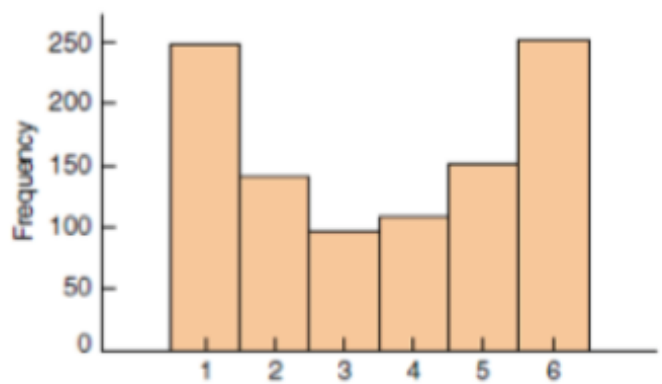
  A. Green

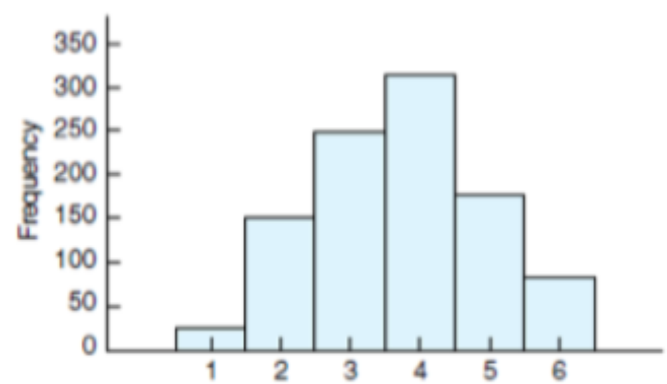
  B. Orange

  C. Blue

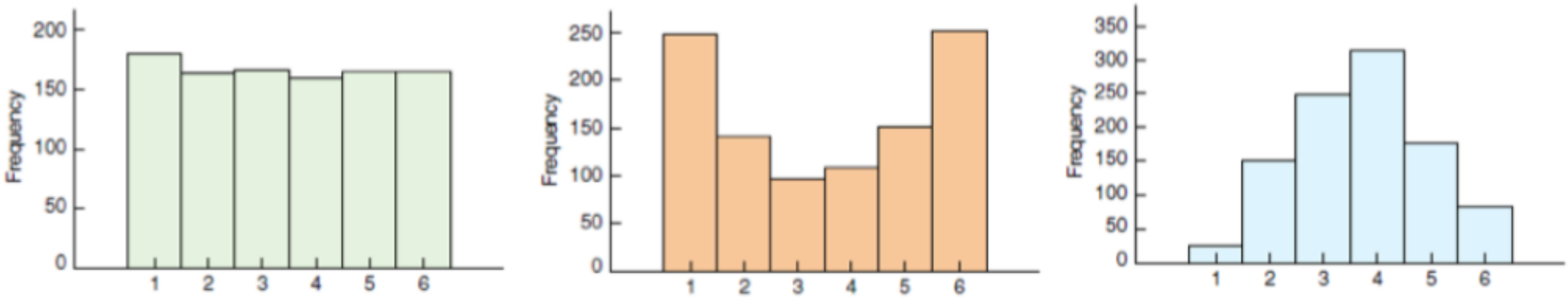
Clicker!







Comparing Standard Deviations



- The distribution shown by the orange histogram has the largest standard deviation.
- The distribution shown by the blue histogram has the smallest standard deviation.
- All these groups have the same mean, min and max values.
- The standard deviation measures how widely spread the points are from the mean.

Q1, Q3, and IQR

- Quartiles:
  - Q1 is the first quartile, 25% of the data are below this point
  - Q2 is the median, 50% of the data are below this point
  - Q3 is the third quartile, 75% of the data are below this point

- IQR is the difference between Q3 and Q1, i.e. the middle 50% of the data, so IQR = Q3 - Q1

Finding Q1, Q3, and IQR

- Let's again consider the final scores of 5 Stats 10 students: 79, 82, 94, 83, 92
- In order to find Q1 and Q3 we must first put the values in increasing order: 79, 82, 83, 92, 94

1. Find Q2. What is the median? Q2 = 83.
2. Find Q1, the median of the numbers less than 83.

$$Q1 = \frac{79 + 82}{2} = 80.5$$

3. Find Q3, the median of the numbers greater than 83.

$$Q3 = \frac{92 + 94}{2} = 93$$

5. Find IQR. IQR = Q3 - Q1 = 93 - 80.5 = 12.5

Example

Clicker!

What is the IQR?

1, 5, 5, 1, 7, 3, 1, 5, 9, 7, 8

Answer:

A. 1
B. 5
C. 6
D. 7

Example

What is the IQR?

1, 5, 5, 1, 7, 3, 1, 5, 9, 7, 8

Answer: 6

1, 1, 1, 3, 5, 5, 5, 7, 7, 8, 9

Q1 median Q3

Q3-Q1
= 7-1
= 6

Range

- Range = Max - Min
- Again, consider our five Stat 10 students:

79, 82, 94, 83, 92

- Range = 94 - 79 = 15
- Range is a poor measure of spread because it is not resistant to outliers and generally doesn't tell us where most of the data is located.

Thinking About Variation

- Since statistics is about variation, spread is an important fundamental concept of statistics.
- Measures of spread help us talk about what we don’t know.
- When the data values are tightly clustered around the center of the distribution, the IQR and standard deviation are small.
- When the data values are scattered far from the center, the IQR and standard deviation are large.

Which Center and Spread are Best?

- Use the mean and standard deviation when the distribution is symmetric and unimodal.
- Use the median and IQR when the distribution is left skewed or right skewed.
- If the distribution is not unimodal, it may be better to split the data.
    - In this case, neither the mean nor the median represent typical values or the center
    - Investigate further into possible separate sub-populations
    - Present graphs and statistics of sub-populations separately

Review: Describing Distributions

The things to always describe when considering a distribution:

- Shape - examples: unimodal, symmetric
- Center - the “typical” value
    - Use the mean as the typical value for symmetric distribution
    - Use the median as the typical value for skewed distribution
- Spread - how spread out the data is
    - Use the standard deviation as the spread for symmetric distribution
    - Use the IQR as the spread for skewed distribution

The Empirical Rule

- A rough guideline, a rule of thumb, that helps us understand how the standard deviation measures variability.

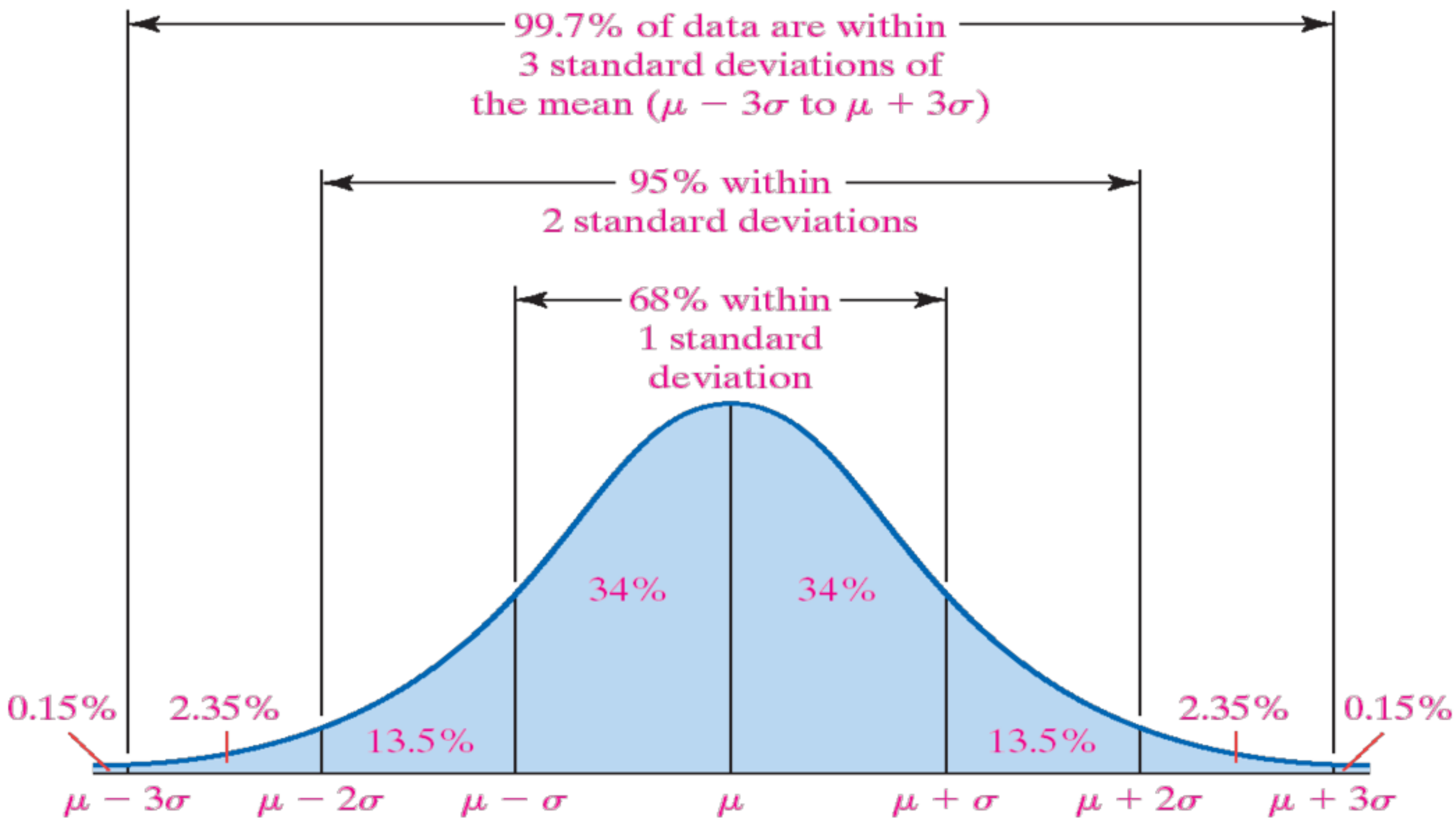
If a distribution is unimodal and symmetric:

- About 68% of the observations fall within one standard deviation of the mean
- About 95% of the observations fall within two standard deviations of the mean
- About 99.7% (nearly all) of the observations fall within three standard deviations of the mean

Empirical Rule for Data with a Bell Shape

If the distribution of a data set is approximately bell shape, than

•68% of the data values fall within 1 standard deviation of the mean.
•95% of the data values fall within 2 standard deviations of the mean.
•99.7% of the data values fall within 3 standard deviations of the mean.



Example

- High temperatures in San Francisco follow a unimodal and symmetric distribution with mean 65 degrees and standard deviation 8 degrees.
    - What is the range of temperatures that includes the middle 95% of high temperature days in SF?

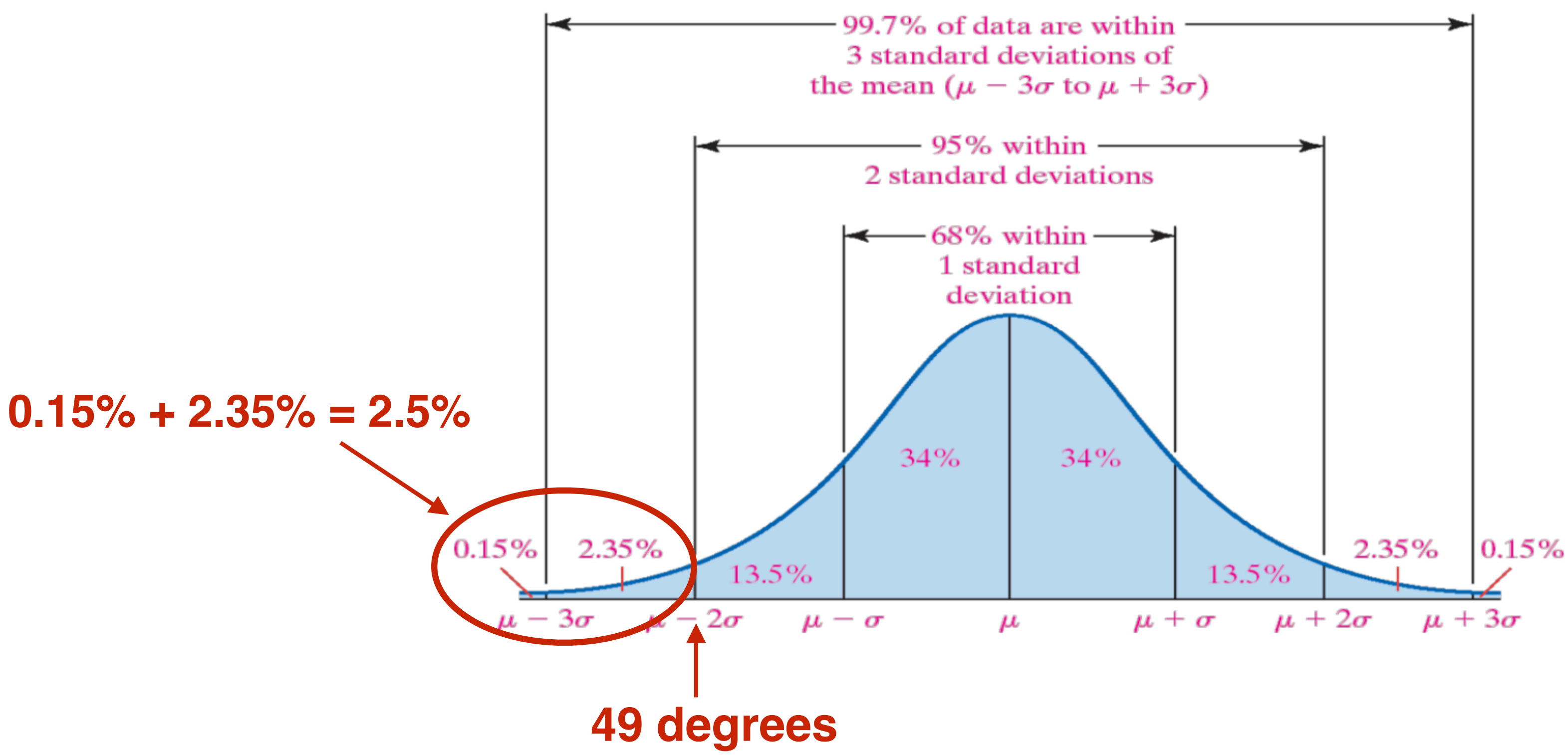
$$65 - 2 \times 8 = 49 \qquad\qquad 65 + 2 \times 8 = 81$$

From the Empirical Rule we know that 95% of all days in SF are between 49 and 81 degrees.

Example

- High temperatures in San Francisco follow a unimodal and symmetric distribution with mean 65 degrees and standard deviation 8 degrees.
  - Is a temperature colder than 49 degrees unusual?

Only 5% of days are warmer than 81 degrees or cooler than 49 degrees. More specifically, only 2.5% of days are colder than 49 degrees. Yes a temperature colder than 49 degrees is unusual.



Example

Clicker!

- SAT scores are unimodal and symmetric with a mean of 1500 and standard deviation of 300. What percentage of students scored between 1200 and 1800?

    a) 95%

    b) 99.7%

    c) 68%

    d) 100%

Example

- SAT scores are unimodal and symmetric with a mean of 1500 and standard deviation of 300. What percentage of students scored between 1200 and 1800?

    a) 95%
    b) 99.7%
    c) 68%
    d) 100%

$$68\% \rightarrow 1500 \pm 1SD = 1500 \pm 1(300) = (1200, 1800)$$

$$95\% \rightarrow 1500 \pm 2SD = 1500 \pm 2(300) = (900, 2100)$$

$$99.7\% \rightarrow 1500 \pm 3SD = 1500 \pm 3(300) = (600, 2400)$$

Comparing Apples to Oranges

- Comparing test scores of students from different tests (ex, SAT and ACT)
- Comparing weight loss of subjects from two groups that are vey different in average weight
- Growth (height and weight) comparisons with children or babies from different age groups

SAT vs ACT Example

A college admissions officer is looking at the files of two candidates, Sophie and Adam. Sophie has a total SAT score of 1800 (out of 2400) and Adam has a total ACT score of 27 (out of 36). Which candidate scored better?

- Since these scores are on different scales we can't just compare 1800 to 27.
- Instead we should try to figure out how well Sophie did in comparison to all the other students who took the SAT, and how well Adam did in comparison to all the other students who took the ACT

SAT vs ACT Example

- SAT and ACT scores are assumed to follow a unimodal, symmetric distribution. The average SAT score was 1500 with a standard deviation of 300 and the average ACT score was 21 with a standard deviation of 5. Remember that Sophie scored 1800 on the SAT and Adam scored 27 on the ACT.

Sophie:	**Adam:**
1800 - 1500 = 300	**27 - 21 =6**

- So Sophie scored 300 points above the mean and Adam scored 6 points above the mean. But these values are still not comparable. We need to convert them into the same units in order to be able to compare them.

SAT vs ACT Example

- Remember the standard deviation for the SAT is 300 and the standard deviation for the ACT is 5. To compare the two we use the standard deviation. We calculate how many standard deviations above the mean they scored.

Sophie:

$$\frac{1800-1500}{300}=\frac{300}{300}=1.0$$

Adam:

$$\frac{27-21}{5}=\frac{6}{5}=1.2$$

- So Sophie scored 1 standard deviation above the mean and Adam scored 1.2 standard deviations above the mean. Now the units are comparable. So we can conclude that Adam scored better.

Standardizing with Z-scores

- The scores we calculated are called standardized scores denoted as z.
- They are also called z-scores.
- Z-scores are used to compare individual data values to their mean relative to their standard deviation.
- The formula for calculating the z-score of a data value is:

$$z = \frac{x - \bar{x}}{s}$$

Properties of Z-scores

- Z-scores have no units.
- Z-scores measure the distance of each data value from the mean in standard deviations.
- A negative z-score tells us that the data value is below the mean, while a positive z-score tells us that the data value is above the mean.

Benefits of Standardizing

- Standardized values have been converted from their original units to the standard statistical unit of standard deviations from the mean.
- We can compare values that are measured on different scales, with different units, or from different populations.

Example

- In the 2012 Olympics, Michael Phelps won Olympic gold in the 200m IM with a time of 114.27 seconds. The mean finals time in the event was 119.83 seconds with a standard deviation of 1.91 seconds.
- Missy Franklin won gold in the 200m backstroke in a time of 124.06 seconds. The mean finals time in the backstroke was 130.33 seconds with a standard deviation of 3.47 seconds.
- Who performed better in relation to the rest of the Olympic finals field in their specific events?

Example

Clicker!

- In the 2012 Olympics, Michael Phelps won Olympic gold in the 200m IM with a time of 114.27 seconds. The mean finals time in the event was 119.83 seconds with a standard deviation of 1.91 seconds.
- What is the z-score?

$$z = \frac{x - \bar{x}}{s}$$

a) 2.91

b) 1.81

c) -2.91

d) -1.81

Example

- In the 2012 Olympics, Michael Phelps won Olympic gold in the 200m IM with a time of 114.27 seconds. The mean finals time in the event was 119.83 seconds with a standard deviation of 1.91 seconds.

- What is the z-score?

  a) 2.91

  b) 1.81

  c) -2.91

  d) -1.81

$$z = \frac{x - \bar{x}}{s}$$

$$z = \frac{114.27 - 119.83}{1.91} = -2.91$$

Example

Clicker!

- Missy Franklin won gold in the 200m backstroke in a time of 124.06 seconds. The mean finals time in the backstroke was 130.33 seconds with a standard deviation of 3.47 seconds.
- What is the z-score?

a) 2.91

b) 1.81

c) -2.91

d) -1.81

$$z = \frac{x - \bar{x}}{s}$$

Example

- Missy Franklin won gold in the 200m backstroke in a time of 124.06 seconds. The mean finals time in the backstroke was 130.33 seconds with a standard deviation of 3.47 seconds.
- What is the z-score?

  a) 2.91

  b) 1.81

  c) -2.91

  d) -1.81

$$z = \frac{x - \overline{x}}{s}$$

$$z = \frac{124.27 - 130.33}{3.47} = -1.81$$

Example

Clicker!

- The z-score for Michael Phelps is -2.91 and the z-score for Missy Franklin is -1.81.
- So who performed better?
  a) Michael Phelps
  b) Missy Franklin

Example

- The z-score for Michael Phelps is -2.91 and the z-score for Missy Franklin is -1.81.
- So who performed better?

a) Michael Phelps

b) Missy Franklin

The Five-Number Summary

- When the data is partitioned into 4 equal segments, five important numbers arise
- The five-number summary consists of:

Min, Q1, Median, Q3, Max

The Five-Number Summary

- What is the five-number summary for the five Stat 10 final grades?

79, 82, 83, 92, 94

Clicker!

- Answer:

a) min=79, Q1=80.5, median=83, Q3=93, max=94

b) min=79, Q1=80.5, median=83, Q3=93, max=92

c) min=79, Q1=82, median=83, Q3=92, max=94

d) min=79, Q1=82, median=83, Q3=92, max=92

The Five-Number Summary

- What is the five-number summary for the five Stat 10 final grades?

79, 82, 83, 92, 94

- Answer:

a) min=79, Q1=80.5, median=83, Q3=93, max=94

b) min=79, Q1=80.5, median=83, Q3=93, max=92

c) min=79, Q1=82, median=83, Q3=92, max=94

d) min=79, Q1=82, median=83, Q3=92, max=92

Outliers

- An outlier is a data value that is a distance of more than 1.5 interquartile ranges below the first quartile or above the third quartile.
  - Calculate IQR
  - Find Left limit = Q1 - 1.5 x IQR
  - Find Right limit = Q3 + 1.5 x IQR
  - Any observation less than the left limit or greater than the right limit is an outlier
- Not all data sets will have outliers.

Example

Consider the following test scores:

3, 79, 82, 83, 92, 94

- What is the median?
- Q1? Q3?
- What is the IQR?

Example

Consider the following test scores:

3, 79, 82, 83, 92, 94

- What is the median? 82.5
- Q1? Q3? Q1=79, Q3=92
- What is the IQR? Q3-Q1=92-79=13

Example

Consider the following test scores:

3, 79, 82, 83, 92, 94

- Left limit?
- Right limit?
- Are there any outliers? If so, what?

Example

Consider the following test scores:

3, 79, 82, 83, 92, 94

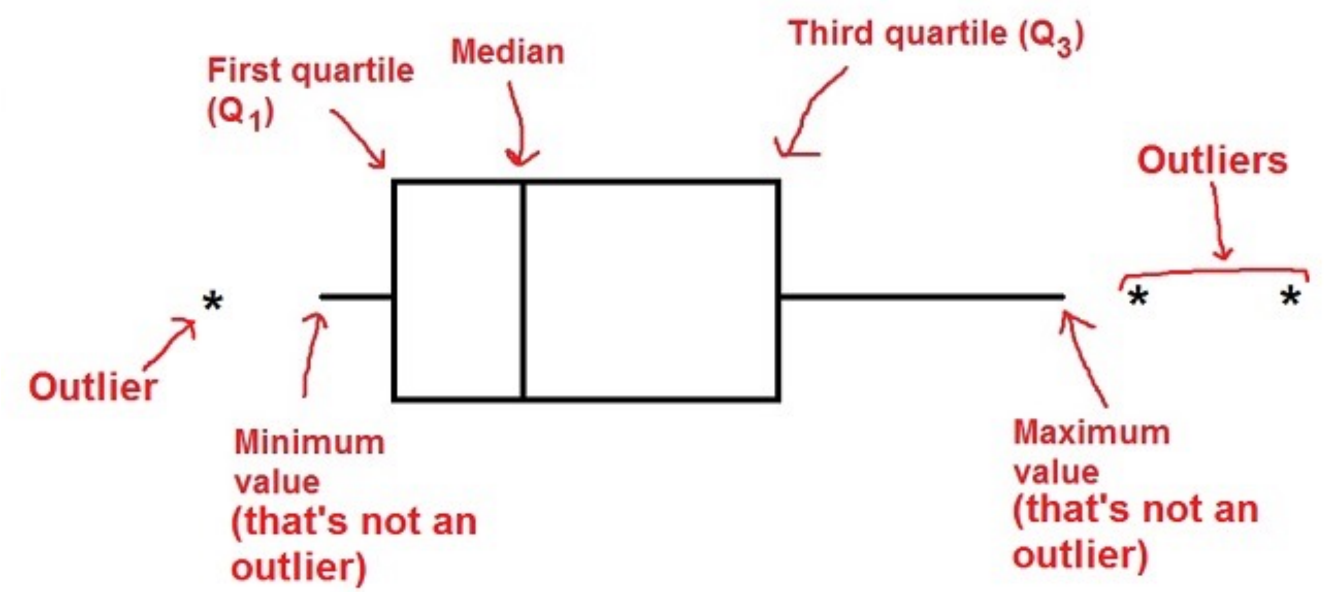
- Left limit? Q1-1.5xIQR= 79-1.5(13) = 59.5
- Right limit? Q3+1.5xIQR= 92+1.5(13) = 111.5
- Are there any outliers? Yes, “3” is an outlier.

Boxplots

- A box plot is a chart that visually displays the five-number summary in a way that allows us to view the distribution (shape, center, spread) of a numerical variable.

Constructing a Boxplot

- Calculate the five-number summary, right and left limits and outliers (if any)
- Draw a horizontal rectangle with a line segment in the middle. The short sides of the rectangle and the line segment in the middle should correspond to Q1, the median and Q3.
- Sketch horizontal line segments (whiskers) on each side that extend to the most extreme values that are not potential outliers.



Boxplot Example

- Consider the six Stat 10 final scores:

  3, 79, 82, 83, 92, 94

Draw the boxplot.

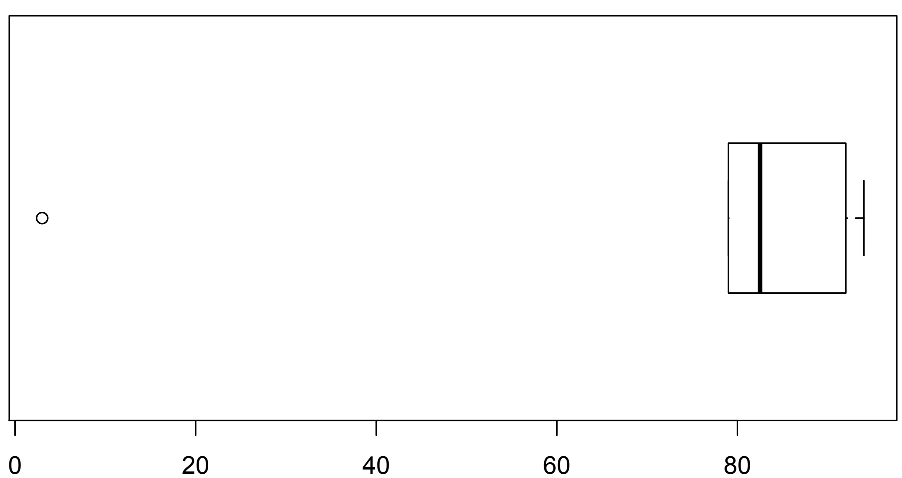
- Median is 82.5, Q1 is 79, Q3 is 92, the IQR is 13.
- Left limit is 59.5, right limit is 111.5 hence 3 is an outlier.

Boxplot Example

- Consider the six Stat 10 final scores:

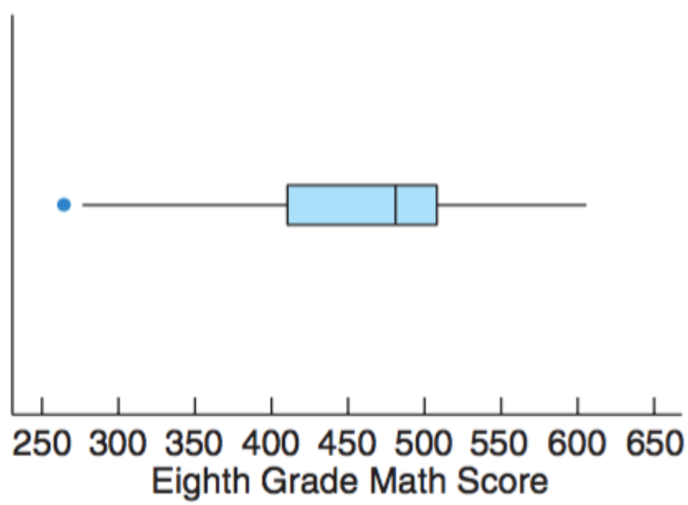
  3, 79, 82, 83, 92, 94

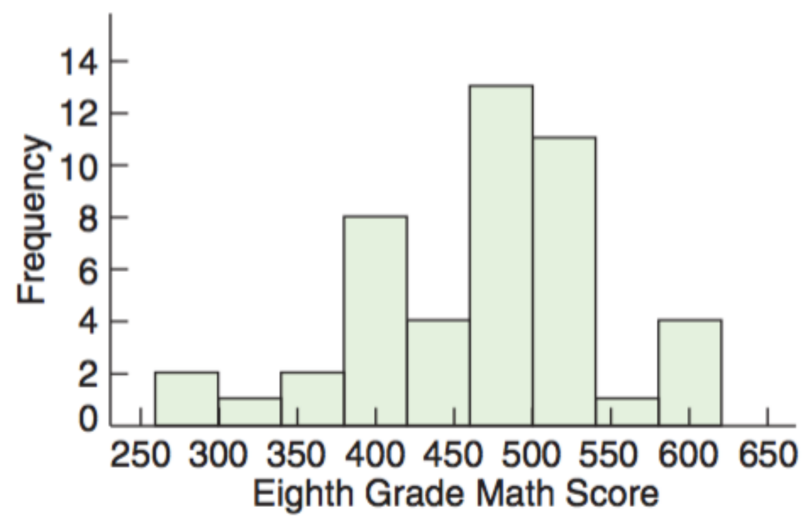
- Median is 82.5, Q1 is 79, Q3 is 92, the IQR is 13.
- Left limit is 59.5, right limit is 111.5 hence 3 is an outlier.
- Whiskers: Max value is 94 and min value is 79.



Investigating Potential Outliers

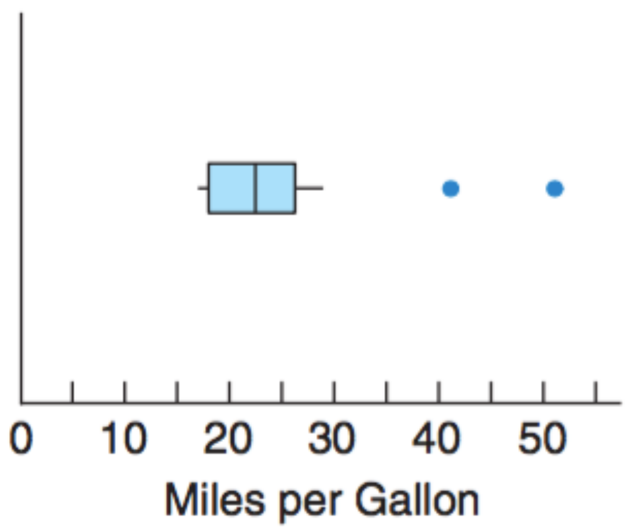
- We always investigate potential outliers. A potential outlier might not be an outlier at all.
- Below we have a boxplot of the NAEP international math scores for 42 countries. One country is flagged as a potential outlier.
- However, when we examine the histogram, we see that this outlier is really not that extreme. We probably wouldn't consider this to be an outlier because it is not separated from the bulk of the distribution in the histogram.

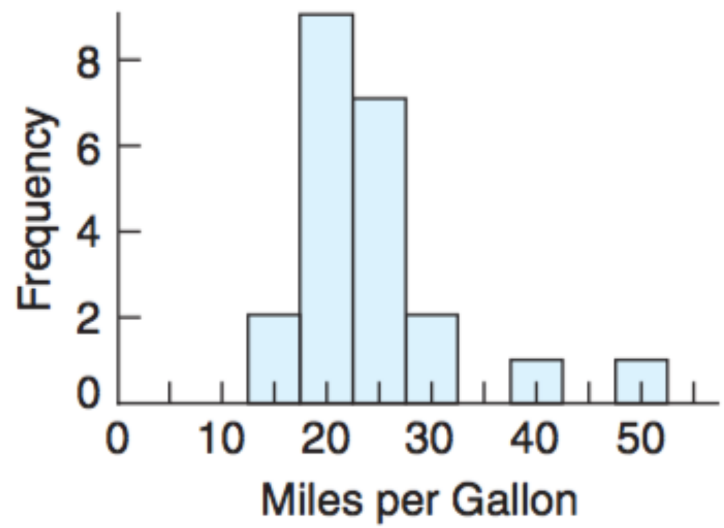




Investigating Potential Outliers

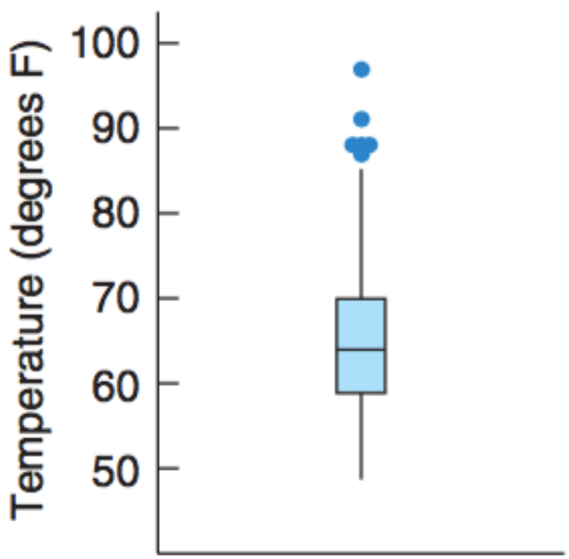
- Here we have a boxplot and histogram for the fuel economy (in city driving) of the 2010 model sedans from Ford, Toyota and GM, in miles per gallon.
- Two potential outliers are shown in the boxplot. When examining the histogram, we can confirm these two outliers because they are far enough from the bulk of the distribution. These outliers happen to be hybrid cars (Ford Fusion and Toyota Prius) which deliver much better fuel economy.

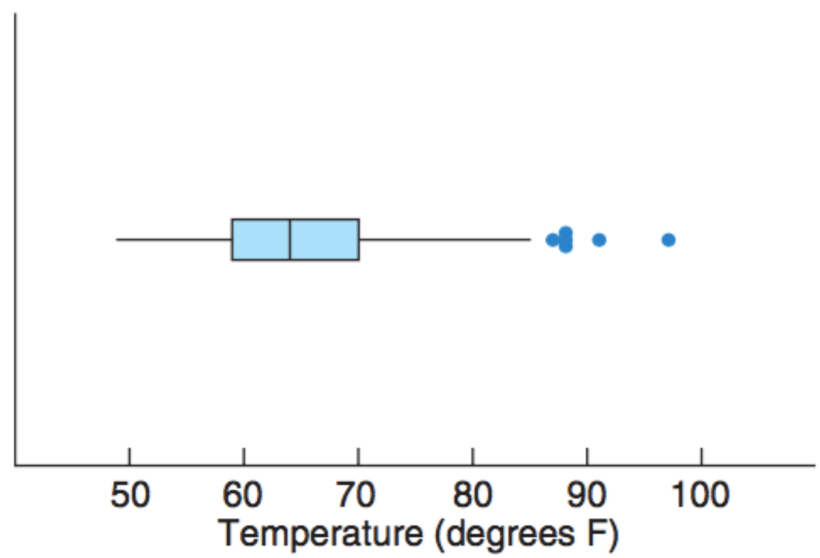




Horizontal or Vertical?

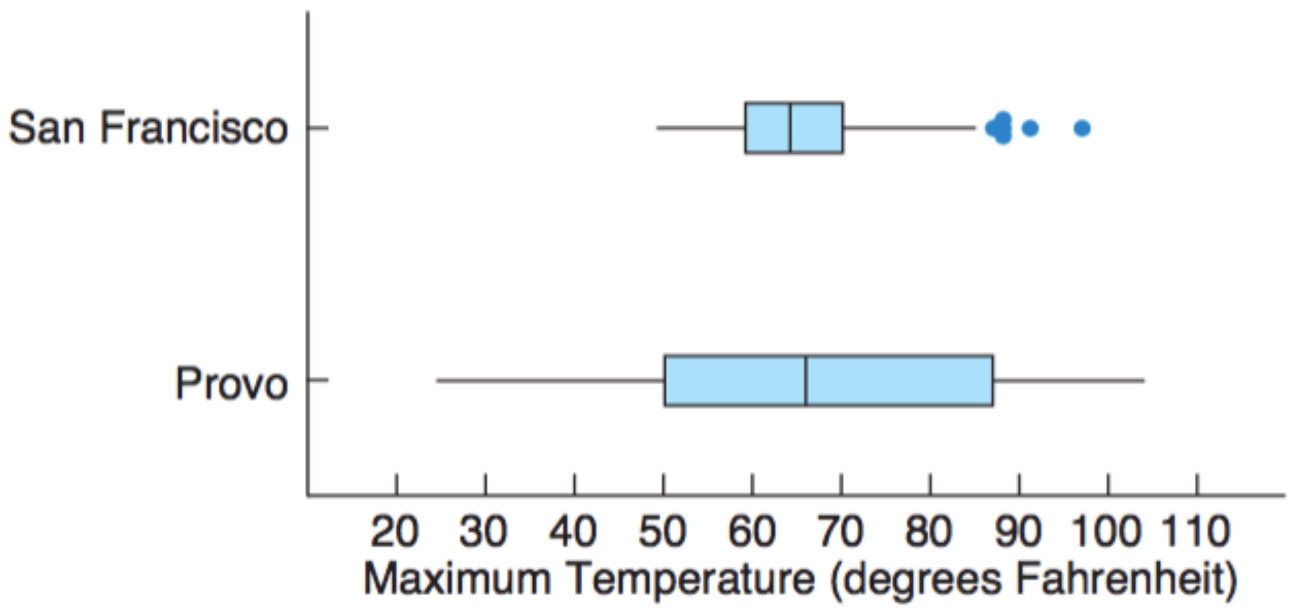
- Boxplots can be either horizontal or vertical.
- Direction is not important.
- It is up to the researcher to decide which one is more readable.





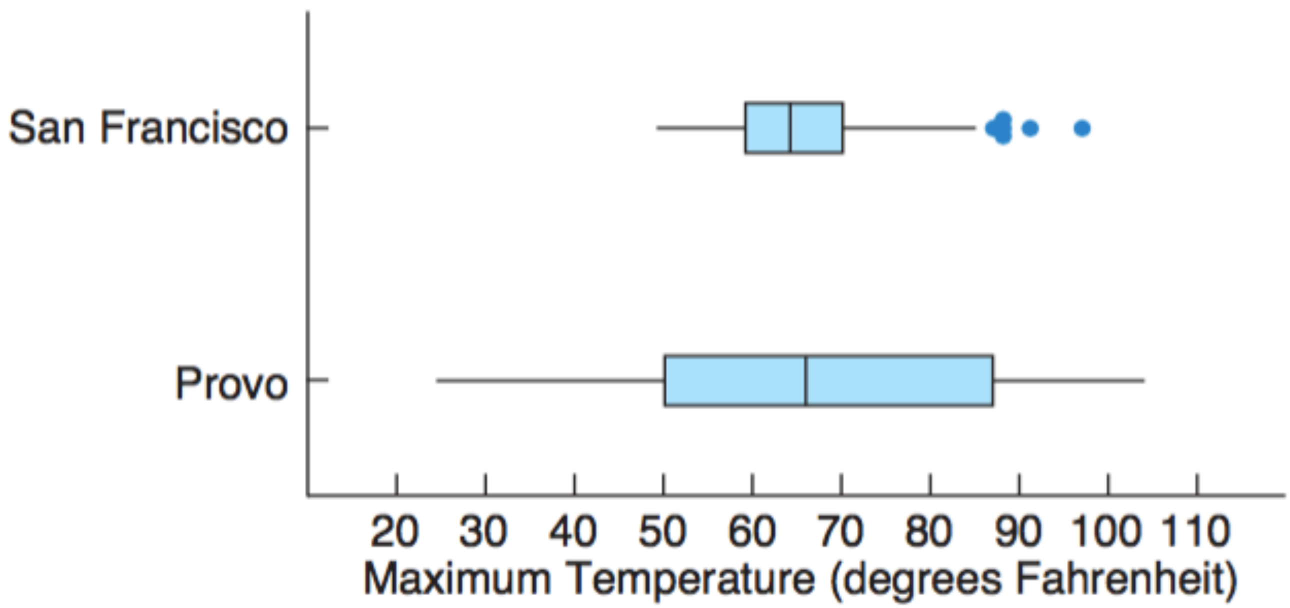
Comparing Distributions with Boxplots

- Boxplots are useful for comparing two or more distributions.
- These boxplots show the daily maximum temperatures for San Francisco and Provo.
- How do the temperatures in Provo compare with those in San Francisco?



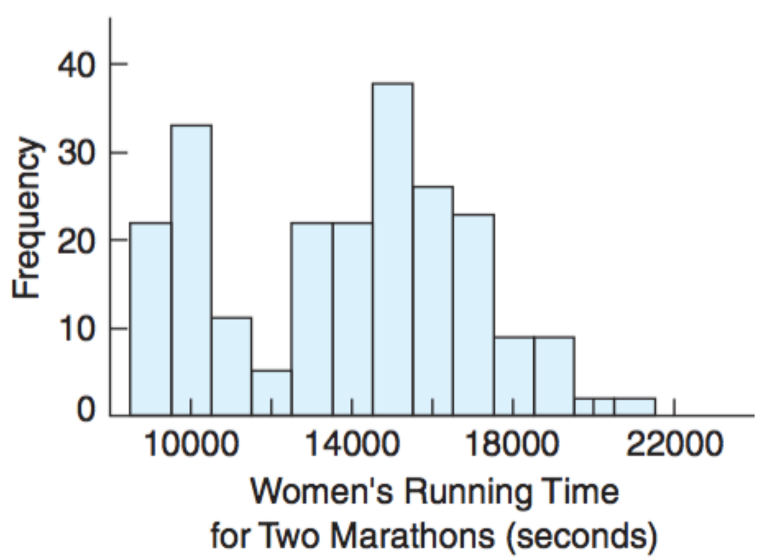
Comparing Distributions with Boxplots

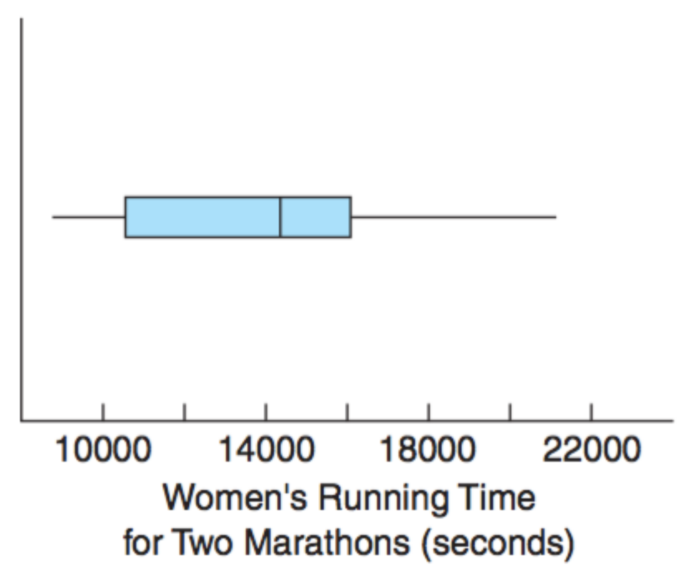
- The median temperatures are about the same.
- The medians are in the center of the boxes and boxplots are fairly symmetric hence the distributions are fairly symmetric.
- The variation in daily temperatures in Provo is greater than in SF.
- The boxplot for SF shows some potential outliers.



Things to Watch for with Boxplots

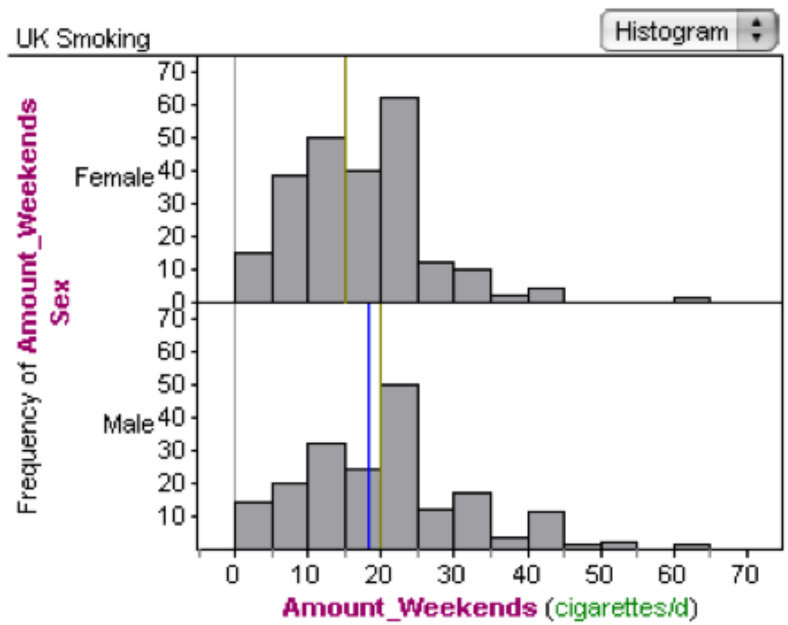
- Boxplots are best used only for unimodal distributions.
- Boxplots don't show bimodality or multimodality. Can be misleading.
- Below we have a histogram of marathon running times for two groups of women runners: amateurs and Olympians. The distribution is bimodal. However, when we look at the boxplot it doesn't show bimodality.
- Boxplots shouldn't be used for very small data sets. The data set should have at least five numbers to make a boxplot.





Comparing Groups with Histograms

The histograms below show the distributions of the amount of cigarettes smoked during the weekends by males and females



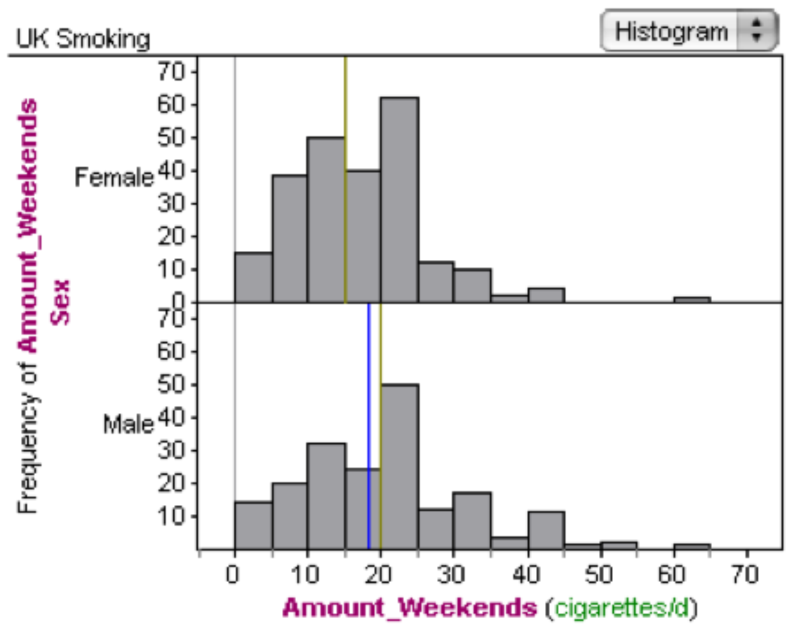
Describe the shape.

Shape

Shape: The distribution of the amount of cigarettes smoked on weekends for females appears to be fairly symmetric with a possible outlier around 60. The distribution for males is right-skewed.

Comparing Groups with Histograms

The histograms below show the distributions of the amount of cigarettes smoked during the weekends by males and females



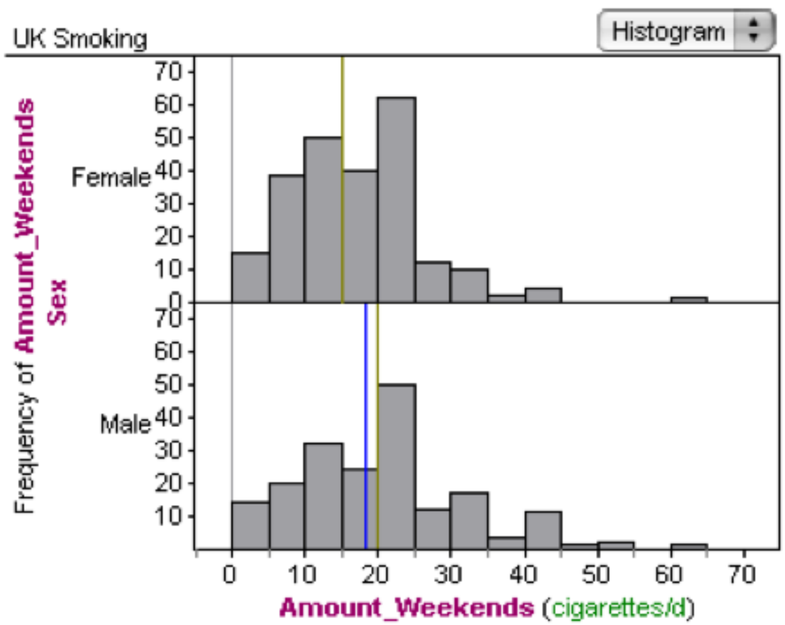
Describe the center.

Center

Center: The mean and the median for females is approximately 15. The mean for males is approximately 20 with a median at 18.

Comparing Groups with Histograms

The histograms below show the distributions of the amount of cigarettes smoked during the weekends by males and females

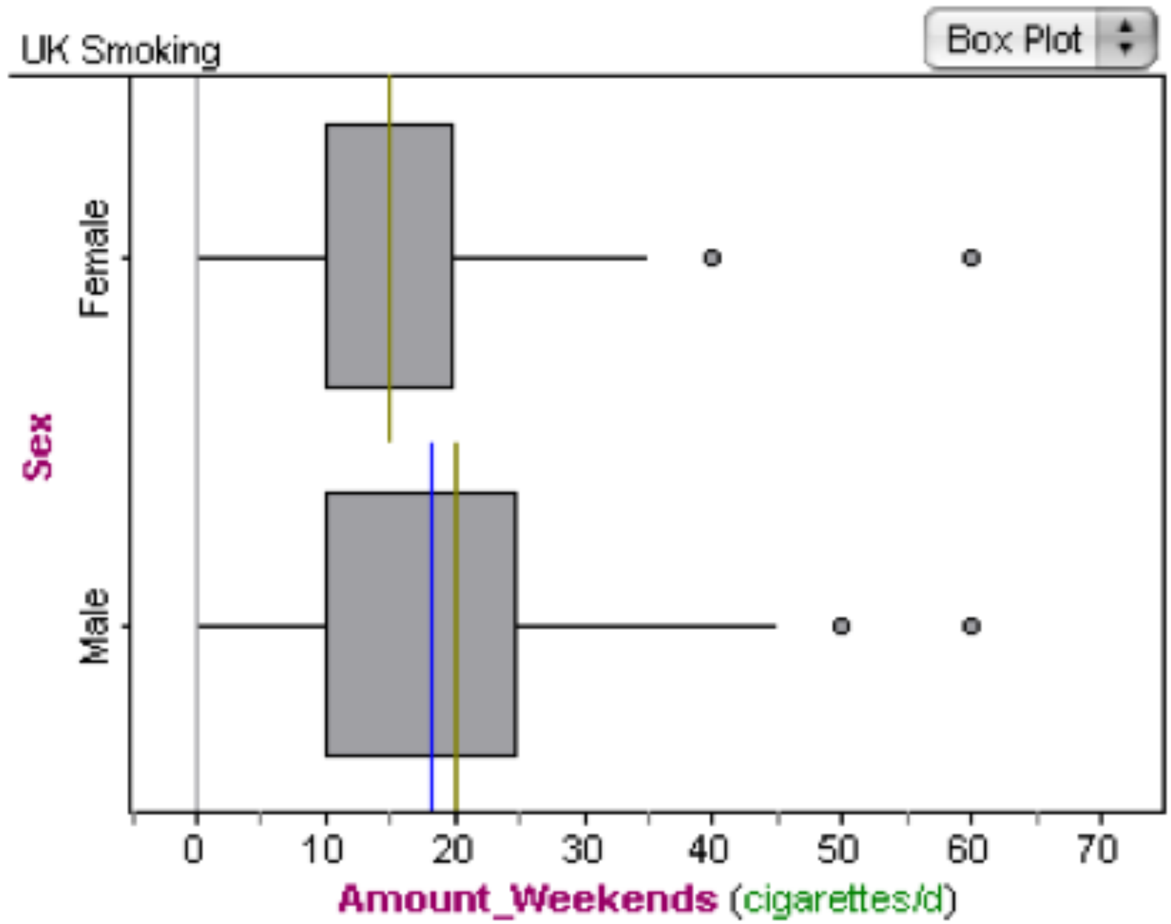


Describe the spread.

Spread

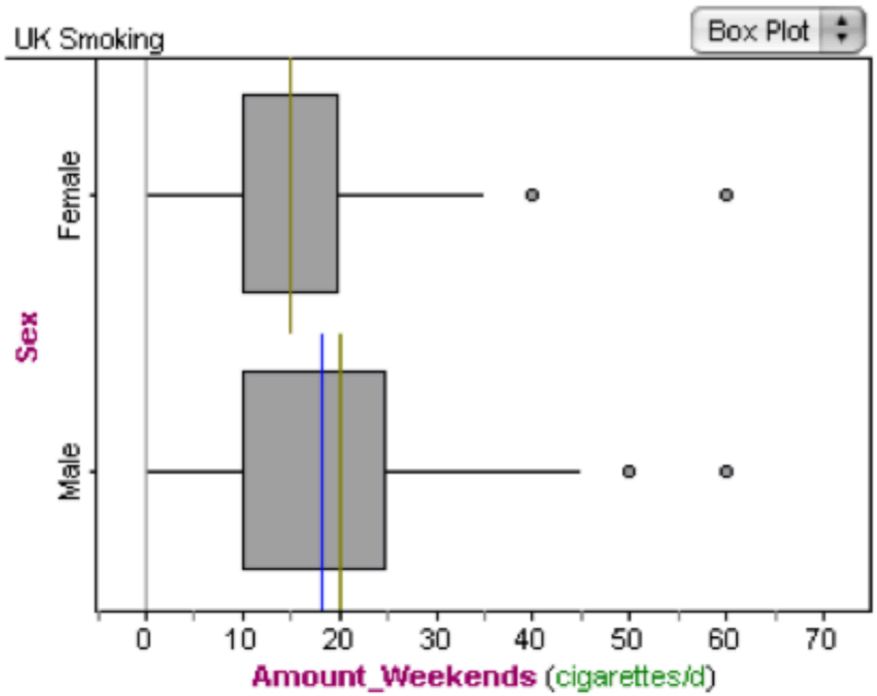
Spread: There is more variability in the amount of cigarettes smoked on weekends by males than by females.

Comparing Groups With Box Plots



What is the five-number summary?

Comparing Groups with Boxplots

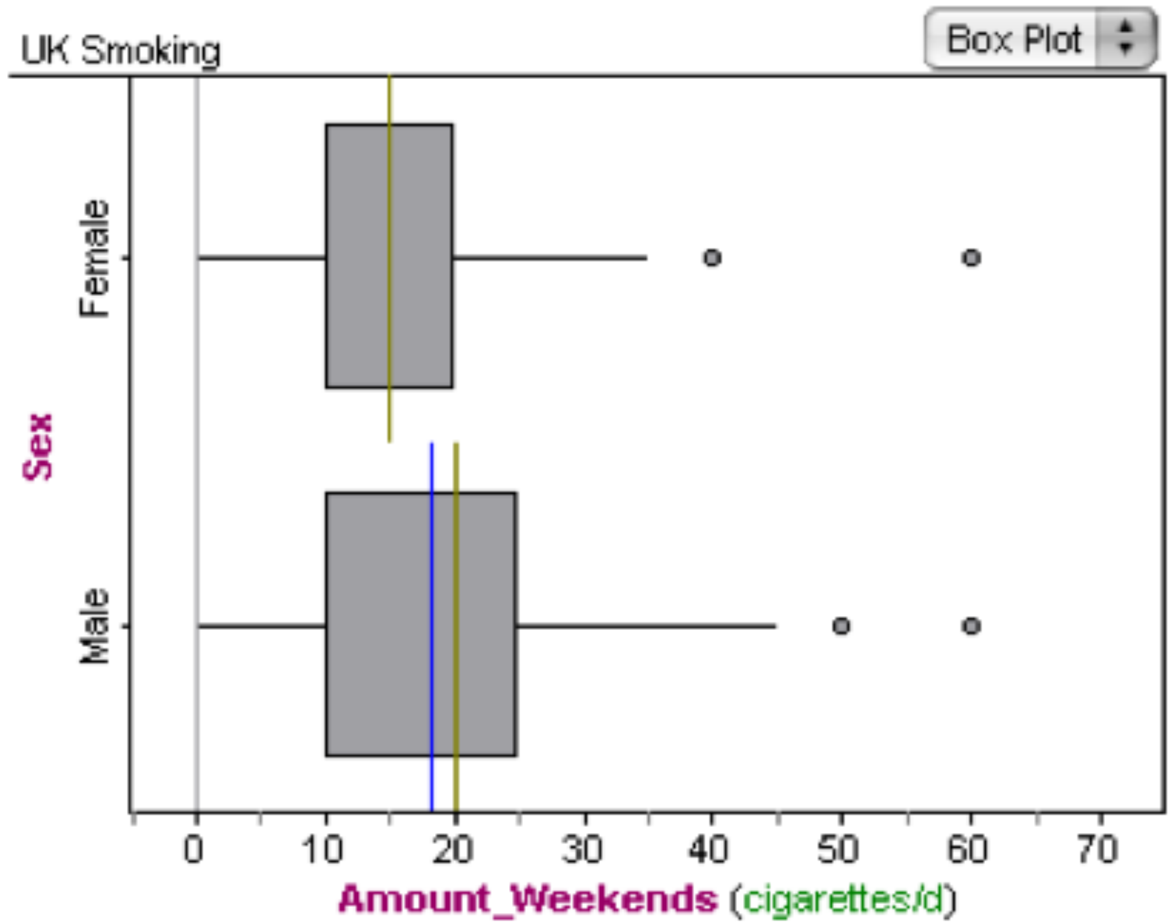


Min/Max: Both distributions have a minimum at 0 and a maximum at 60.

Q1/Q3: Both distributions have a Q1 at 10, meaning that 25% of the males and females smoke less than 10 cigarettes per day on weekends. The female distribution has a Q3 at 20 cigarettes, meaning that 75% of females smoke less than 20 cigarettes per day on the weekend. Q3 is 25 cigarettes for males.

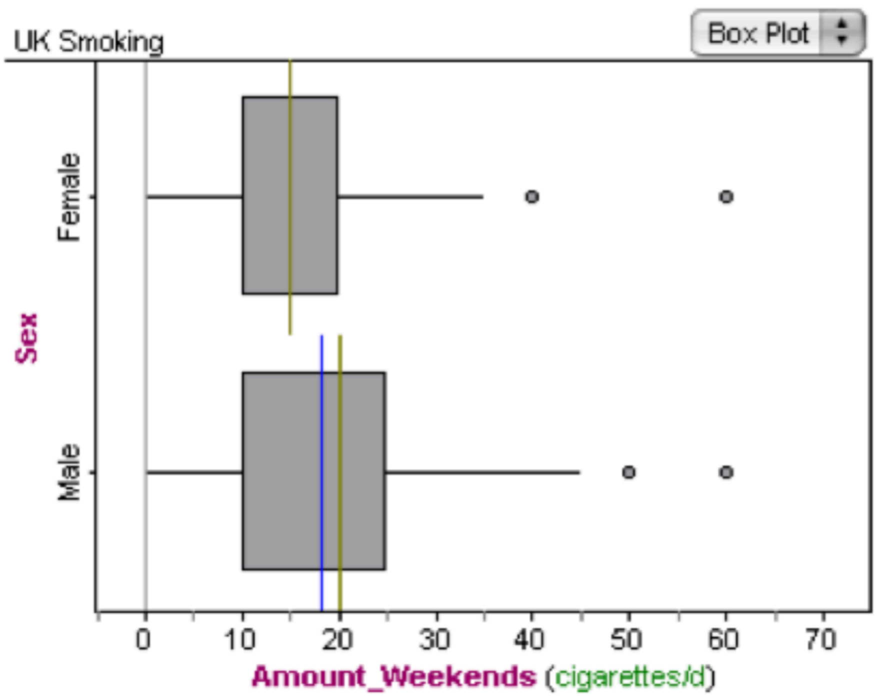
Median: The median amount of cigarettes smoked per day on a weekend for females is 15 while it's 18 for males.

Comparing Groups With Box Plots



Are there outliers? If so, what are the outliers?

Comparing Groups with Boxplots



Outliers: Both distributions have two outliers on the positive end.

For males, the outliers are 50 and 60.

For females, the outliers are 40 and 60.

Example

Below is a sample from the results of a survey that asked how many girlfriends/boyfriends UCLA students had during their time at UCLA:

0, 0, 1, 3, 4, 4, 5, 6, 15

Example

Clicker!

What is the median?

0, 0, 1, 3, 4, 4, 5, 6, 15

Answer:

A. 0

B. 3

C. 4

D. 4.5

Example

Clicker!

What is Q1?

0, 0, 1, 3, 4, 4, 5, 6, 15

Answer:

A. 0
B. 0.5
C. 1
D. 5

Example

Clicker!

What is Q3?

0, 0, 1, 3, 4, 4, 5, 6, 15

Answer:

A. 4
B. 15
C. 5
D. 5.5

Example

Clicker!

What is IQR?

0, 0, 1, 3, 4, 4, 5, 6, 15

Answer:

A. 4

B. 15

C. 5

D. 5.5

Example

0, 0, 1, 3, 4, 4, 5, 6, 15

Q1 median Q3

Median (the center) = 4
Q1 (the median of the numbers less than the center)= 0.5
Q3 (the median of the numbers greater than the center)= 5.5
IQR = Q3 – Q1 = 5.5 – 0.5 = 5

Example

Clicker!

What is left limit?

0, 0, 1, 3, 4, 4, 5, 6, 15

Answer:

A. -1
B. -7
C. 13
D. 15

Example

Clicker!

What is right limit?

0, 0, 1, 3, 4, 4, 5, 6, 15

Answer:

A. -1
B. -7
C. 13
D. 15

Example

Clicker!

Are there any outliers?

0, 0, 1, 3, 4, 4, 5, 6, 15

Answer:

A. Yes
B. No

Example

Left limit = Q1 - 1.5*IQR = 0.5 – 1.5(5) = -7
Right limit = Q3 + 1.5*IQR = 5.5 +1.5 (5) = 13

Are there any outliers? By looking at the range -7 to 13 we can see that one value (15) is an outlier.

Example

Draw your boxplot!

Example

Draw your boxplot!