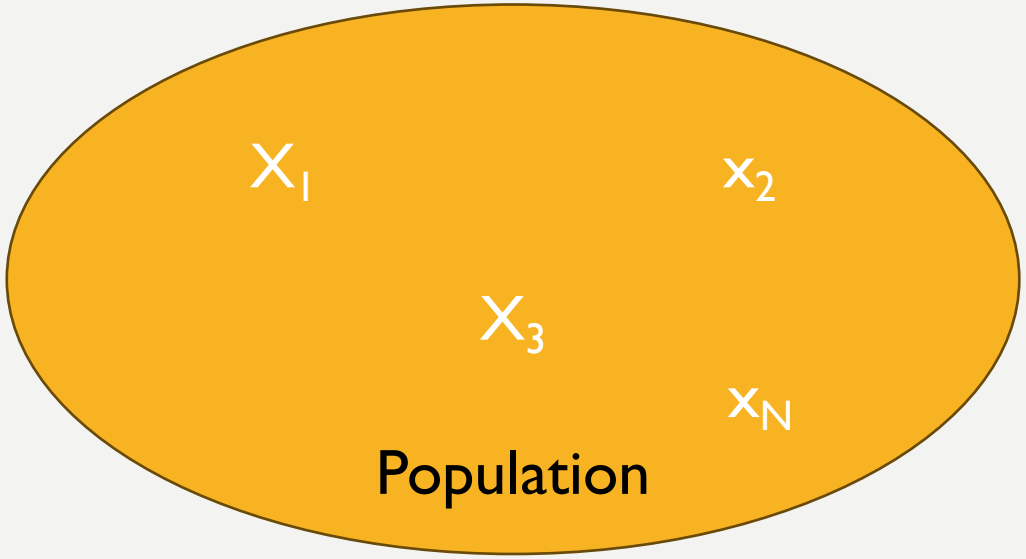
SECTION 1.3

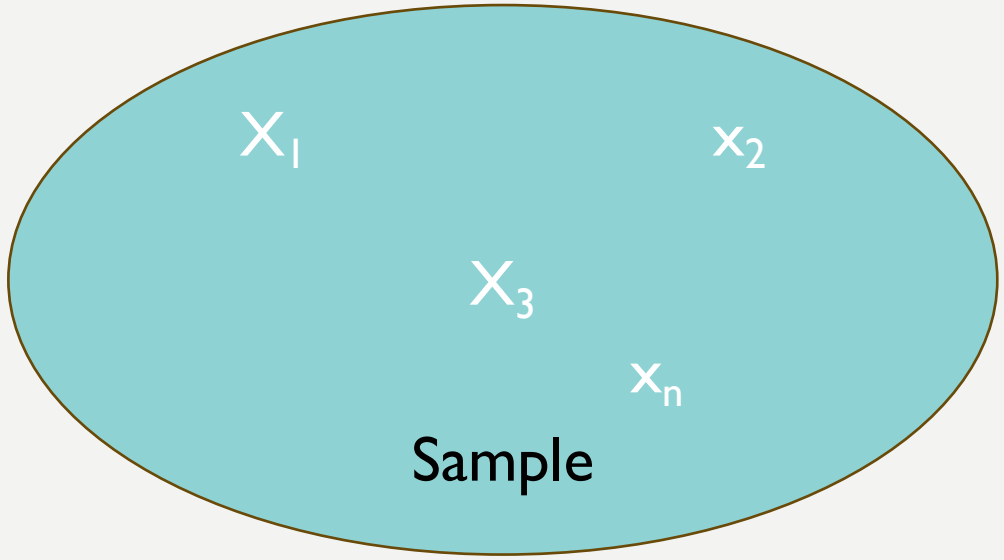
MEASURES OF LOCATION

OBJECTIVES

- Calculate the mean, median, and mode.
- Calculate quartiles, percentiles, and trimmed means.
- Determine the most appropriate measure of location.

NOTATION





Population	Sample
• Population size: N • The i^{th} data value: x_i • Population mean: μ	• Sample size: n • The i^{th} data value: x_i • Sample mean: $\bar{x}$

MEAN

Population Mean	Sample Mean
$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$ $= \frac{\sum x_i}{N}$	$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$ $= \frac{\sum x_i}{n}$

MEDIAN

Finding the Median of a Data Set

1. List the data in ascending (or descending) order, making an ordered array.
2. If the data set contains an ODD number of values, the median is the middle value in the ordered array.
3. If the data set contains an EVEN number of values, the median is the arithmetic mean of the two middle values in the ordered array. Note that this implies that the median may not be a value in the data set.

MEDIAN

Example:

Given the numbers of absences for samples of students in two different classes, find the median for each sample.

a. 3, 4, 6, 7, 2, 8, 9

b. 5, 7, 8, 1, 4, 9, 8, 9

MODE

- The **mode** is the value in the data set that occurs most frequently.
- If the data values only occur once or an equal number of times, we say there is **no mode**.
- If one value occurs most often, then the data set is said to be **unimodal**.
- If exactly two values occur equally often, then the data set is said to be **bimodal**.
- If more than two values occur equally often, the data set is said to be **multimodal**.

MODE

Example:

Given the number of phone calls received each hour during business hours for three different companies, find the mode of each, and state if the data set is unimodal, bimodal, or neither.

a. 6, 4, 6, 1, 7, 8, 7, 2, 5, 7

b. 3, 4, 7, 8, 1, 6, 9

c. 2, 5, 7, 2, 8, 7, 9, 3

d. 2, 2, 3, 3, 4, 4, 5, 5

CALCULATING MEASURES OF LOCATION

Given the following two data sets of Hank Aaron's number of home run counts for his 23 years in baseball and Barry Bonds's home run count for his 22 years in baseball. Calculate the Mean, Median and Mode.

Aaron's: 10, 12, 13, 20, 24, 26, 27, 29, 30, 32, 34, 38, 39, 39, 40, 40, 44, 44, 44, 44, 45, 47

Bonds's: 5, 16, 19, 24, 25, 25, 26, 28, 33, 33, 34, 34, 37, 37, 40, 42, 45, 45, 46, 46, 49, 73

a. From the data, can we conclude that one player performed better than the other?
b. What would happen to the mean for Bonds's homerun count if we did not include the 73 homeruns?

Solution:

a)

Hank Aaron	Barry Bonds
Mean = $\frac{721}{22} = 32.77$ homeruns	Mean= $\frac{762}{22} = 34.64$ homeruns
Median = 36 homeruns	Median = 34 homeruns
Mode= 44 homeruns	Mode = 25, 33, 34, 37, 45, 46

b) Barry Bonds: $\frac{689}{21} = \mathbf{32.81}$ homeruns

DETERMINING THE MOST APPROPRIATE MEASURE OF CENTER

1. For qualitative data, the mode should be used.
2. For quantitative data, the mean should be used, unless the data set contains outliers or is skewed.
3. For quantitative data sets that are skewed or contain outliers, the median should be used.

Example:

Choose the best measure of center for the following data sets.

a. T-shirt sizes (S, M, L, XL) of American women

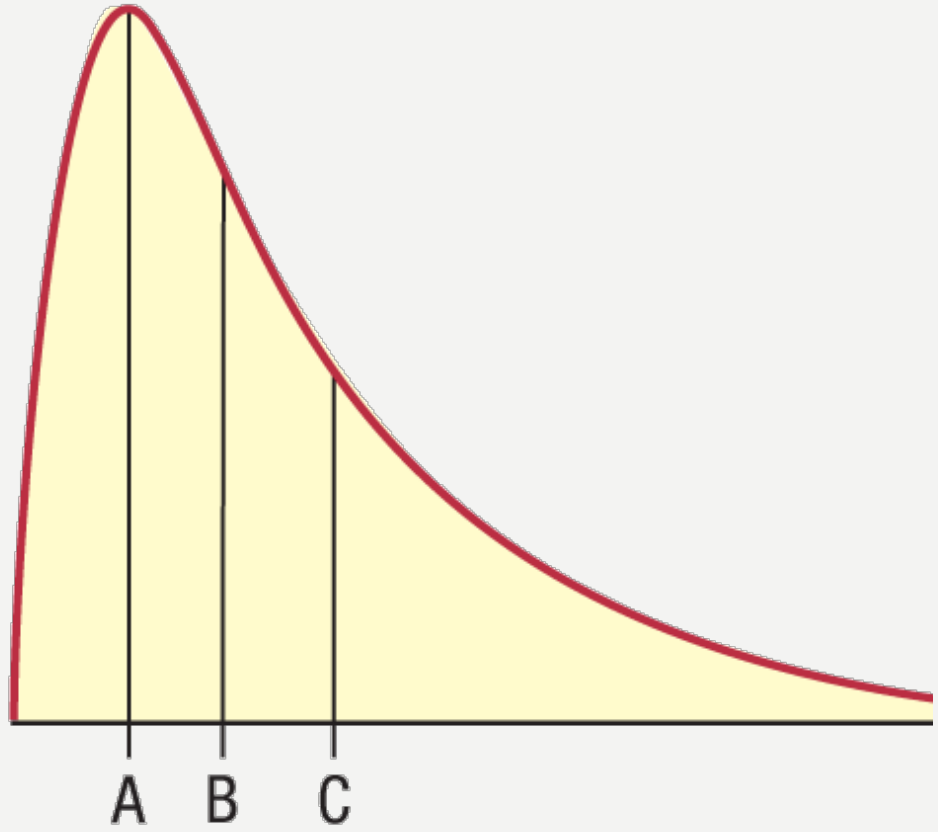
b. Salaries for a professional team of baseball players

c. Prices of homes in a subdivision of similar homes

d. Rankings of services on a scale of *best*, *average*, and *worst*

GRAPHS AND MEASURES OF CENTER

1. The mode is the data value at which a distribution has its highest peak.
2. The median is the number that divides the area of the distribution in half.
3. The mean of a distribution will be pulled toward any outliers.



PROPERTIES OF MEAN, MEDIAN, AND MODE

Mean	Median	Mode
"Average"	"Middle value"	"Most frequent value"
May not be a data value	May not be a data value	Must be a data value
Single Value	Single Value	Could be one value, multiple values, or not exist
Affected by outliers	Not affected by outliers	Not affected by outliers
Use for quantitative data with *no* outliers	Use for quantitative data with outliers	Use for qualitative data

TRIMMED MEAN

- A trimmed mean (truncated mean) is a method of **averaging that removes a small percentage of the largest and smallest values** before calculating the mean.
- Using a trimmed mean helps eliminate the influence of outliers or data points on the tails that may unfairly affect the traditional mean.
- Three commonly applied trim percentages: 5%, 10%, and 20%.

To calculate a X% trimmed mean, you can use the following steps:

- **Step 1**: Order each value in a dataset from smallest to largest.
- **Step 2**: Remove the values in the bottom X% and top X% of the dataset.
- **Step 3**: Calculate the mean of the remaining values.

Example:

Given the data set: 4, 8, 12, 15, 9, 6, 14, 18, 12, 9. Calculate the 10% trimmed mean.

Solution:

Ordered Dataset: 4, 6, 8, 9, 9, 12, 12, 14, 15, 18

Trimmed Dataset: 6, 8, 9, 9, 12, 12, 14, 15

10% trimmed mean = (6+8+9+9+12+12+14+15) / 8 = 10.625

The 10% trimmed mean is **10.625**.

MEASURES OF RELATIVE POSITION

P^{th} Percentile of a Data Value

The P^{th} percentile of a particular value in a data set is given by

$$P = \frac{l}{n} \cdot 100$$

where P is the percentile rounded to the nearest whole number,

l is the number of values in the data set *less than or equal to* the given value, and

n is the number of data values in the sample.

Example:

a) If the scores of a set of students in a math test 20, 30, 15 and 75, what is the percentile rank of the score 30?

b) The scores obtained by 10 students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92. Using the percentile formula, calculate the percentile for score 70?

Location of Data Value for the P^{th} Percentile

To find the *data value* for the P^{th} percentile, the location of the data value in the data set is given by

$$l = n \cdot \frac{P}{100}$$

where l is the location of the P^{th} percentile in the *ordered array* of data values.

n is the number of data values in the sample, and P stands for the P^{th} percentile.

- If the formula results in a decimal value for l, the location is the next *larger* whole number. (*Note:* Other methods may be used here.)

Example:

Consider the data set {50, 45, 60, 25, 30}. Find the 5th, 30th, 40th, 50th and 100th percentiles of the list given.

Solution:

Ordered list – 25, 30, 45, 50, 60

N = 5

Percentile (P)	Ordinal rank	Percentile value
5th	(5/100) × 5 = [0.25] = 1	1st number in the ordered list = 25
30th	(30/100) × 5 = [1.5] = 2	2nd number in the ordered list = 30
40th	(40/100) × 5 = 2	2nd number in the ordered list = 30
50th	(50/100) × 5 = [2.5] = 3	3rd number in the ordered list = 45
100th	(100/100) × 5 = 5	5th number in the ordered list = 60

Quartiles

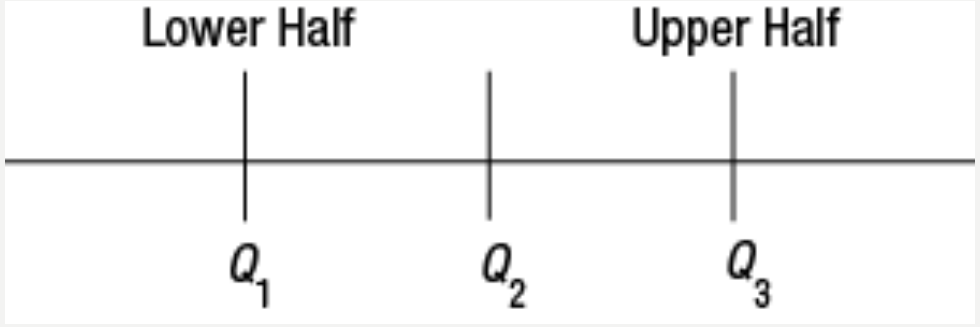
Q_1 = First Quartile: 25% of the data are less than or equal to this value.

Q_2 = Second Quartile: 50% of the data are less than or equal to this value.

Q_3 = Third Quartile: 75% of the data are less than or equal to this value.

Five-number summary

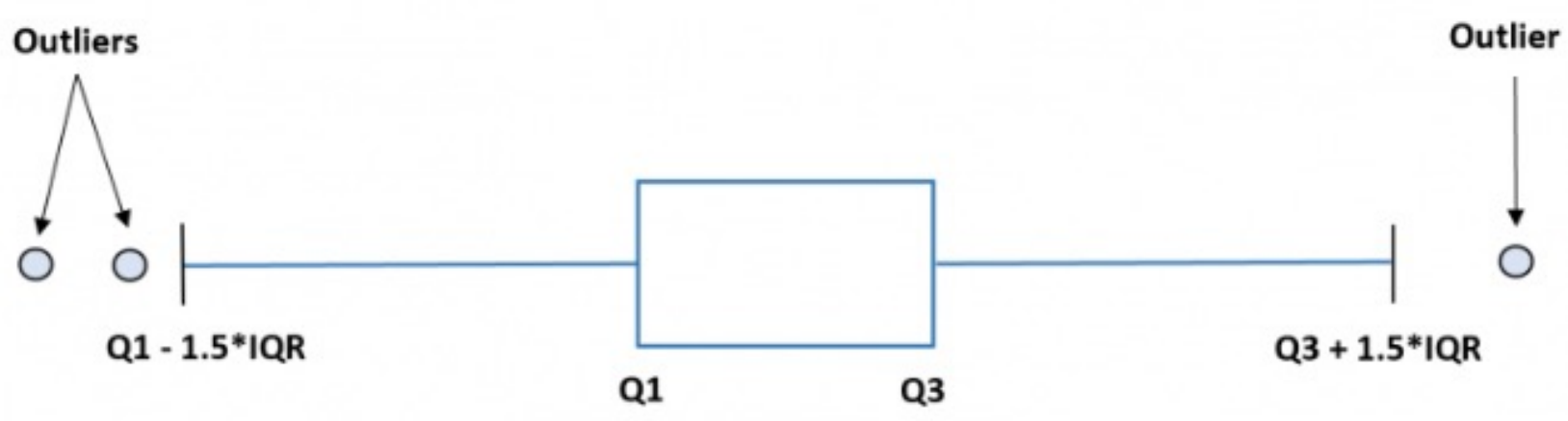
- The minimum value
- The first quartile, Q_1
- The median, or second quartile, Q_2
- The third quartiles, Q_3
- The maximum value



Interquartile Range: $IQR = Q_3 - Q_1$

Outliers

- Any value that is 1.5 x IQR greater than the Q_3 is designated as an outlier and any value that is 1.5 x IQR less than the Q_1 is also designated as an outlier.



- An outlier is extreme if it is more than 3 x IQR from the nearest quartile. Otherwise, it is mild.

BOX PLOT

Creating a Box Plot

1. Begin with a horizontal (or vertical) number line that contains the five-number summary.

2. Draw a small line segment above (or next to) the number line to represent each of the numbers in the five-number summary.

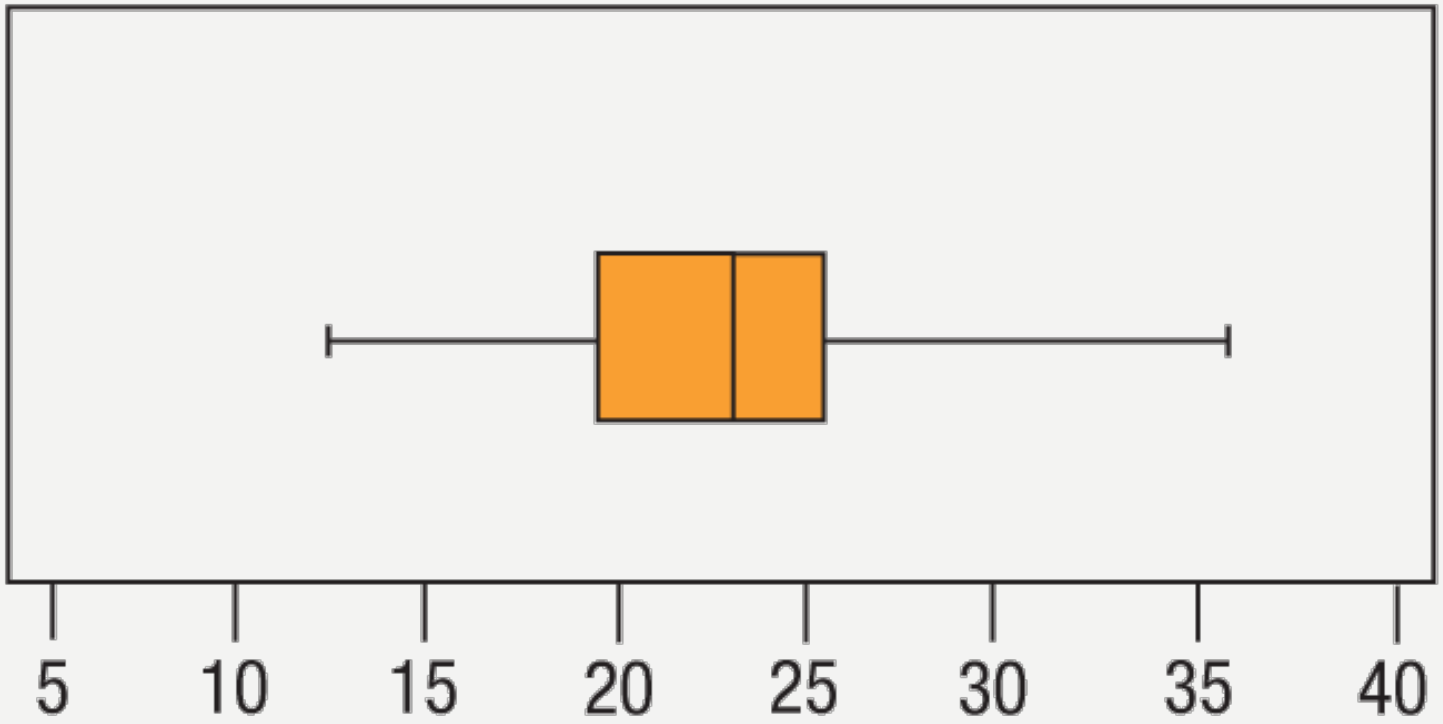
3. Connect the line segment that represents the first quartile to the line segment representing the third quartile, forming a box with the median's line segment in the middle.

4. Connect the "box" to the line segments representing the minimum and maximum values to form the "whiskers."

Example:

Draw a box plot to represent the five-number summary:

12.1, 19.8, 23.6, 25.3, 35.9



Example:

The box plots below are from the US Geological Survey website. Use them to answer the following questions.

a. What do the top and bottom bars represent in these box plots according to the key?

b. Which subbasin had the highest median average spring total phosphorus concentration?

c. Which subbasin had the lowest average spring total phosphorus concentration? (**Note:** Each data value is an average of April's and May's totals, and the lowest average shown for each subbasin is the *10th percentile*.)

d. Which subbasin had the largest interquartile range?