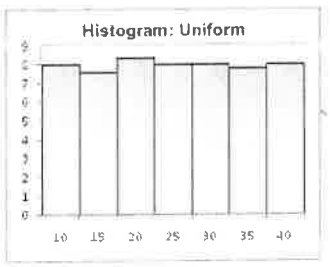
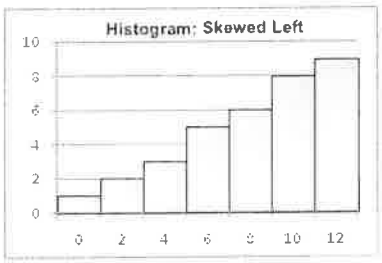
# Chapter 2: Descriptive Statistics

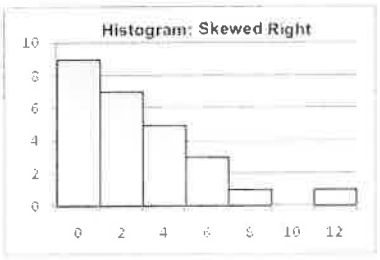
## 2.2: Histograms

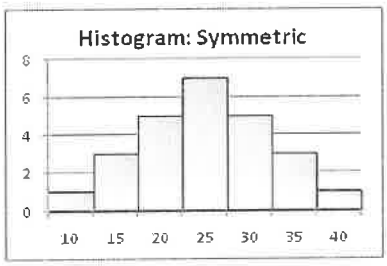
### Graphics for Quantitative Data

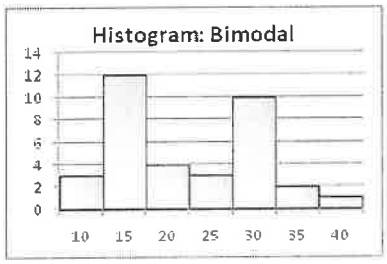
**Histogram** - a graph consisting of bars of equal width drawn adjacent (they touch) to each other. The horizontal scale represents classes of quantitative data values, and the vertical scale represents frequencies. The heights of the bars correspond to the frequency values.



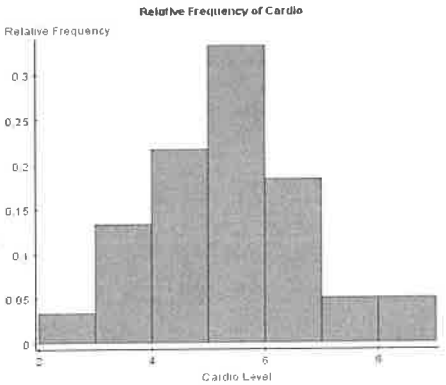



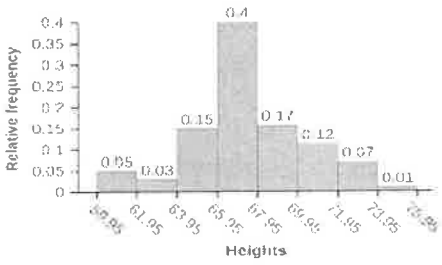




**Relative Frequency Histogram** - a graph with the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies (proportions or percents).



## 2.3, 2.4: Measures of the Location of the Data, Boxplots

Median – a number that measures the "center" of the data. It is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude.

- If the number of data values is odd, the median is the number located in the exact middle of the list.

- If the number of data values is even, the median is found by computing the mean of the two middle numbers.

Example 1: Determine the Median of each of the following set of numbers.

(a) 132, 162, 133, 145, 148, 139, 147, 150, 153

132 133 139 145 147 148 150 153 162

(b) 40, 38, 42, 39, 43, 39

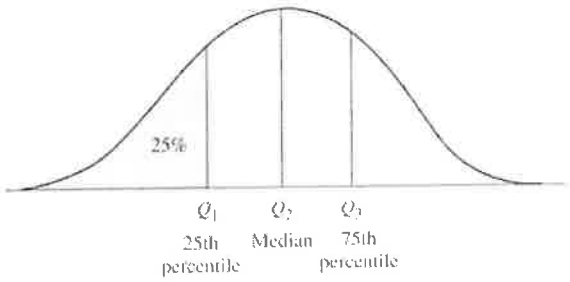
38 39 39 40 42 43 $\frac{39+40}{2} = 39.5$

Common measures of location are **Quartiles** and **Percentiles**. Quartiles are special percentiles.

The first quartile, $Q_1$, is the same as the 25th percentile, and the third quartile, $Q_3$, is the same as the 75th percentile. The median, *M*, is called both the second quartile, $Q_2$, and the 50th percentile.

**Quartiles** are the numbers that separate the data into quarters. (Four parts)

The 25th, 50th, and 75th percentiles denoted by $Q_1$, $Q_2$, $Q_3$, respectively.




**First quartile $Q_1$** = 25th percentile or ($P_{25}$) = number in which 25% of scores are below and 75% of scores are above.

**Second quartile $Q_2$** = 50th percentile or ($P_{50}$) = number in which 50% of scores are below and 50% of scores are above. Known as the median.

**Third quartile $Q_3$** = 75th percentile or ($P_{75}$) = number in which 75% of scores are below and 25% are above.

**Interquartile Range –** a number that indicates the spread of the middle 50% of the data. It is the difference between the third quartile $Q_3$ and the first quartile $Q_1$.

$$IQR = Q_3 - Q_1$$

Example 2: In a park that has several basketball courts, a student counts the number of players playing basketball each day over a two-week period and records the following data.

10, 90, 30, 20, 50, 30, 60, 40, 70, 40, 30, 60, 80, 20

m = 40

10 20 20 30 30 30 40 40 50 60 60 70 80 90

a.) Find and interpret the first, second and third quartiles for the number of basketball players on the courts over the two-week period.

$Q_1 = 30$ — 25% of the data is at 30 or less

$Q_2 = m = 40$ — 50% of the data is at 40 or less

$Q_3 = 60$ — 75% of the data is at 60 or less

b.) Determine and interpret the interquartile range.

$IQR = Q_3 - Q_1$
$= 60 - 30$
$= 30$

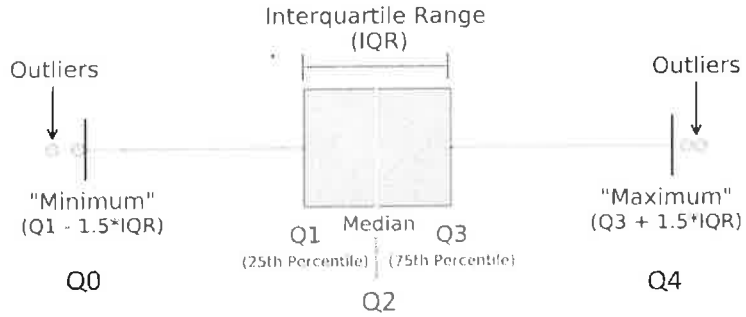
middle 50% of the data

**5 Number Summary:** a subset of the data that consists of the minimum value, the first quartile, the median, the third quartile, and the maximum value.

**Boxplots** - also called box-and-whisker plots; constructed from the 5-number-summary; shows how far extreme values are from the bulk of the data.

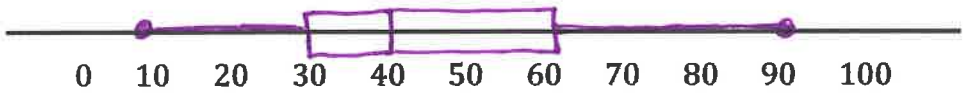
- **Strengths:** give a direct look at location and spread; outliers identified; great for comparing.
- **Weaknesses**: symmetry and skewness can be judged, but not so much shape.

**Outlier** – a data point that is not consistent with the bulk of the data from that group.



Example 3: Find the **5-number summary** and construct a boxplot on the number of players playing basketball each day over a two-week period.

10, 90, 30, 20, 50, 30, 60, 40, 70, 40, 30, 60, 80, 20




## 2.5: Measure of the Center of the Data

A **measure of center** is a value at the center (or middle) of a data set of numbers.

The three forms of center are mean, median, and mode.

- The **mean (average)** of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set.
- The **median** is the middle value when a data set is ordered from least to greatest.
- The **mode** is the number that occurs most often in a data set.

### Statistical Notation

| Notation | Definition |
|---|---|
| $\Sigma$ | sum of a set of data values |
| $x$ | variable used to represent the individual data values |
| $n$ | # of data values (x) in a sample |
| $N$ | **# of data values (x) in a population** |
| $\bar{x} = \frac{\Sigma x}{n}$ | mean of a set of sample values |
| $\mu = \frac{\Sigma x}{N}$ | mean of a set of population values |

Sigma

"x̄ - bar"

"mu"

- Is $\bar{x}$ a parameter or a statistic?
- Is $\mu$ a parameter or a statistic?

Example 4: Use Table 3

**Table 3: Student Test Scores**

| Student | Score | |
|---|---|---|
| Michelle | 82 | |
| Ryanne | 77 | * |
| Bilal | 90 | |
| Pam | 71 | |
| Jennifer | 62 | |
| Dave | 68 | * |
| Joel | 74 | * |
| Sam | 84 | * |
| Justine | 94 | |
| Juan | 88 | |

a.) Compute the **population mean and median** of student test scores. $N=10$

$\mu = \frac{\sum x}{N} = \frac{82+77+\dots+88}{10} = 79$

$m = \frac{77+82}{2} = 79.5$

b.) Take a sample from this class by randomly selecting four students, then compute the sample mean. $n=4$

$\bar{x} = \frac{\sum x}{n} = \frac{77+68+74+84}{4} = 75.75$

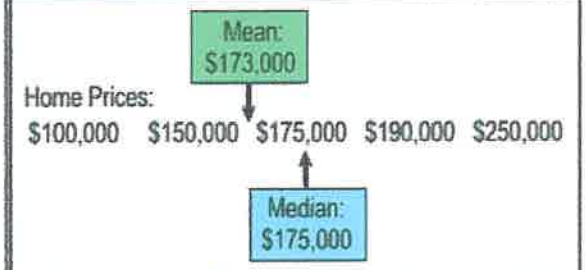
c.) Now, suppose we replace the lowest test score to 0. Find the new population mean and median.

$m = \frac{77+82}{2} = 79.5$

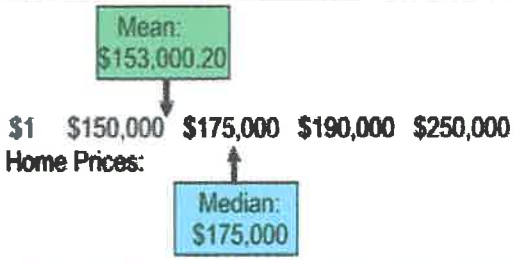
$\mu = \frac{0+82+\dots+88}{10} = 72.8$

0 62 68 71 74 77 | 82 84 88 90 94

- Notice what happens to the Mean and Median when an outlier appears:




This graphic shows a listing of home prices arranged in ascending order. The mean and median are shown.




This graphic is the same as the one on the left, except the lowest price has been replaced with an outlier. The $100,000 home has been "gifted" to a relative for $1.

**Conclusion:** With the outlier, the mean changed.
With the outlier the median did **not** change.

d) Which measure of center (mean or median) is more appropriate to use from (c) above?

median

- The median is not affected by an outlier, whereas the mean will be pulled in the direction of the outlier

**Resistant** - if extreme values (very large or very small) relative to the data do not affect its value substantially, then, the data is said to be resistant.

Which measure of center is more resistant to outliers – mean or median?

median

Which measure of center is more appropriate to use when outliers are present?

median

Example 5:

1. Sam has 20 rose bushes, but only counted the flowers on 6 of them.

Sam's flower counts are: 9, 2, 5, 4, 12, 7

a.) Find the mean and include the unit of measure. $n=6$

$\bar{x} = \frac{\Sigma x}{n} = \frac{9+2+5+4+12+7}{6} = 6.5$ flowers

b.) Find the median and include the unit of measure.

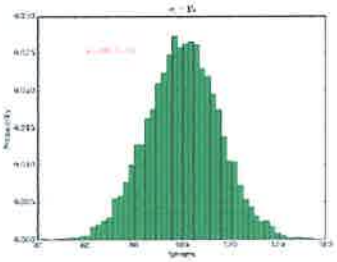
2 4 5 7 9 12

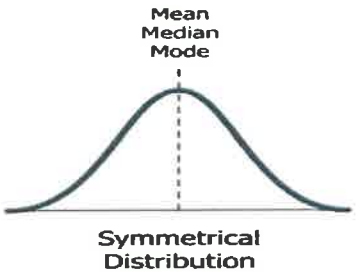
$\frac{5+7}{2} = 6$ flowers

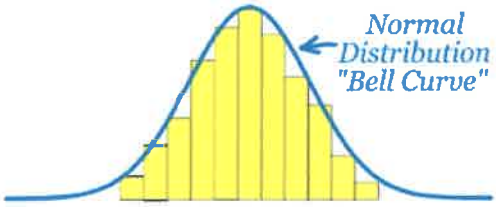
## 2.6: Skewness and the Mean, Median, and Mode

### Shape of a Distribution

**Bell-shaped Distribution** – the highest frequency occurs in the middle and frequencies tail off to the left and the right of the middle. This shape will be used in our study of continuous probability distributions and will be known as a **normal curve**.

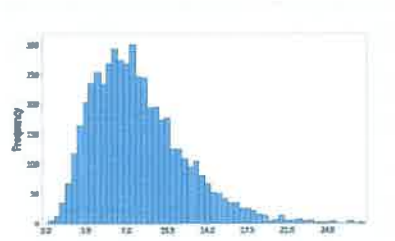


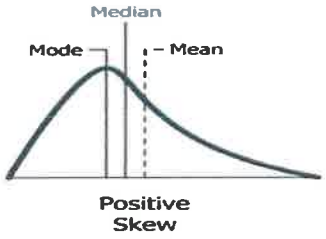



**Right-Skewed (positively skewed)** – The tail extends to the right of the peak longer than to the left.

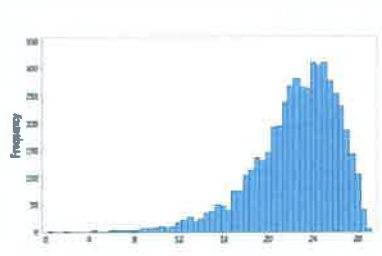
There are extreme values (outliers) to the right.

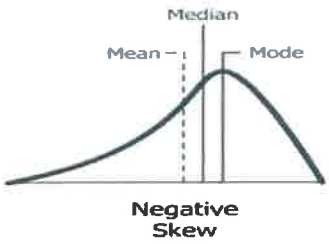





**Left-Skewed (negatively skewed)** – The tail extends to the left of the peak longer than to the right.

There are extreme values (outliers) to the left.





## 2.7: Measures of the Spread of the Data

**Variation** - the degree to which the data is spread out.

Are the data values in our set of data concentrated closely near the mean (the center) OR are the data values more widely spread out from the mean (the center)? To answer this question, we need to learn about the most common measure of variation, or spread, which is the **standard deviation**.

**Standard deviation** – A number that measures how far data values are from their mean. (center)
The standard deviation provides a measure of the overall variation in a data set.

**Range** – the difference between the maximum data value and the minimum data value.
Range = max value – min value

**Population Standard Deviation**: $\sigma = \sqrt{\frac{\sum(X_i-\mu)^2}{N}}$

**Sample Standard Deviation:** $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$

**Standard Deviation (Properties)**

- The standard deviation is a measure of how much data values deviate from the mean.
- The value of the standard deviation can never be negative. It is positive or zero.
  It is zero if all the values are exactly the same.
- Larger values of standard deviation indicate greater amounts of variation.
- Outlier(s) can drastically change the value of the standard deviation since it is a **"not resistant"** measure.

**Variance** – deviation about the mean – it is the square of the standard deviation.

**Population Variance:** $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$ **Sample Variance:** $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$

Relationship between standard deviation and variance

**variance = $(\text{standard deviation})^2$**

**standard deviation = $\sqrt{\text{variance}}$**

Example 6:

(a) If $\sigma = 9$, find $\sigma^2$.

81

(c) If $s = 4$, find $s^2$.

16

(b) If $s^2 = 36$, find s.

6

(d) If $\sigma^2 = 49$, find $\sigma$.

7

**How to calculate the standard deviation for a sample:**

1. **Calculate the mean of the numbers in the data set.**
2. **Subtract the mean from each data point (x), then square the result.**
3. **Sum the squared differences.**
4. **Take the square root of (sum of sq differences/n-1).**

## Calculate the standard deviation:

The following are the home run totals of a sample of five players who played at least half of the games for the 2017 World Champions Houston Astros:

24, 16, 24, 18, 13

| Sample Standard Deviation | $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}}$ |
|---|---|

Find the standard deviation. Round to two decimal places.

$\bar{x} = 19$ $n = 5$

| $x$ | $(x - \bar{x})$ | $(x - \bar{x})^2$ |
|---|---|---|
| 24 | 5 | 25 |
| 16 | -3 | 9 |
| 24 | 5 | 25 |
| 18 | -1 | 1 |
| 13 | -6 | + 36 |
| | | 96 |

$s^2 = \frac{96}{(5-1)} = \frac{96}{4} = 24$

$s = \sqrt{24} = 4.90$

Example 7:

You try: Given the following sample data, answer the following questions:

2 5 4 9 4 8 10

Find the variance. $n = 7$ $\bar{x} = 6$

| $x$ | $x-\bar{x}$ | $(x-\bar{x})^2$ |
|---|---|---|
| 2 | -4 | 16 |
| 5 | -1 | 1 |
| 4 | -2 | 4 |
| 9 | 3 | 9 |
| 4 | -2 | 4 |
| 8 | 2 | 4 |
| 10 | 4 | 16 |
| | | 54 |

$$s^2 = \frac{54}{(7-1)} = \frac{54}{6} = 9$$

Find the standard deviation.

$$s = \sqrt{s^2} = \sqrt{9} = 3$$

Example 8: Identify the symbol for each of the following.

a) the sample mean $\bar{x}$

b) the population mean $\mu$

c) the sample standard deviation $s$

d) the population standard deviation $\sigma$

e) the sample variance $s^2$

f) the population variance $\sigma^2$

g) the sample size $n$

**Z-score** – the number of standard deviations that a given data value $x$ is above or below the mean. The z-score is unitless. It has a mean of 0 and a standard deviation of 1.

## Z-score (also called Standardized Score)

- It represents the number of standard deviations a data point is above or below the mean.
- If it is positive, then it is above the mean.
- If it is negative, then it is below the mean.
- It is a standardized measurement since it is in terms of standard deviation.
- The z-score has no unit of measurement.

### Z-Score Formula

$$Population\ z\text{~}score: z = \frac{x - \mu}{\sigma} \qquad or \qquad z = \frac{x - mean}{standard\ deviation}$$

**Discovery:**

Algebra quiz: Mean = 90 St. dev = 10

Find the z scores for the following students' grades:

June: 80 $\frac{80-90}{10} = \frac{-10}{10} = -1$

Dahn: 95 $\frac{95-90}{10} = \frac{5}{10} = 0.5$

Linh: 73 $\frac{73-90}{10} = \frac{-17}{10} = -1.7$

**Compare…using z-scores. On which test did I do "relatively" better?**

**History Test**
Mean = 92
St. Dev = 3
My Score = 95

$z = \frac{95-92}{3} = \frac{3}{3} = 1$

**Math Test**
Mean =80
St. Dev. 5
My score = 90

$z = \frac{90-80}{5} = \frac{10}{5} = 2$

math

- a higher z-score will produce a "relatively better" outcome.

**You try:**
Alex got an 82 on the first exam and an 84 on the second exam. The class averaged a 79 on the first exam with a standard deviation of 3, while they averaged a 79 on the second exam with a standard deviation of 5.
On which exam did Alex do better?

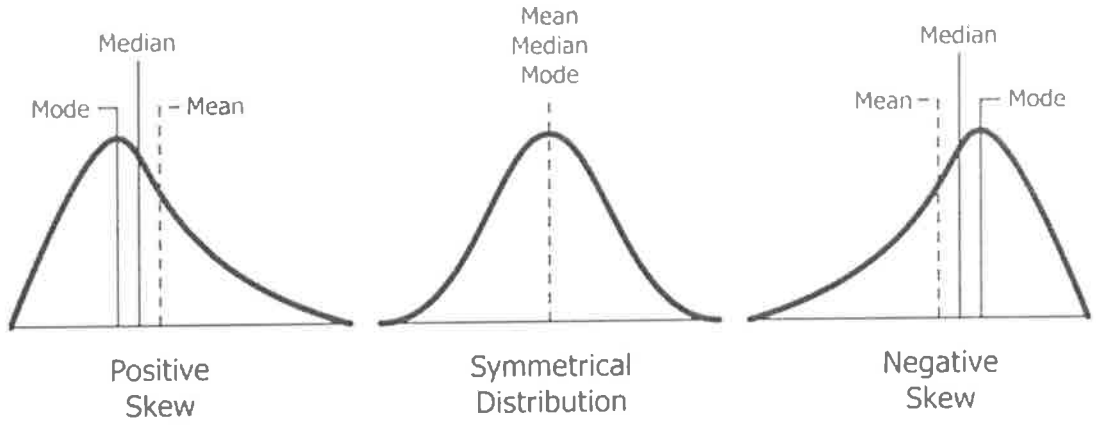
$z = \frac{82-79}{3} = \frac{3}{3} = 1$

$z = \frac{84-79}{5} = \frac{5}{5} = 1$

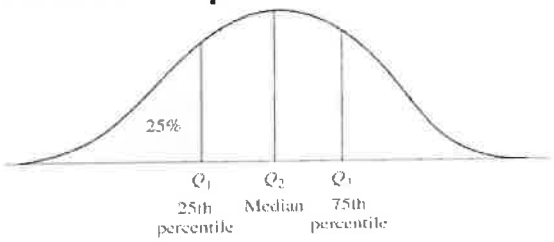
same

**Key Takeaways:**
The median, quartiles, IQR are **RESISTANT (Robust)** measures. They can resist going toward an extreme value. (or outlier)

The mean, range, standard deviation, and variance are **NOT RESISTANT (Not Robust)** measures. These measures are affected by extreme values. These measures will be pulled toward those outliers in the right or left tail.




## Relationship between Quartiles and Percentiles




## Units of measure

- The z-score has no unit of measurement.
- Variance has squared units. (points squared, inches squared, homeruns squared)
- All other measures have single units (points, inches, homeruns)

**Z-score –** the number of standard deviations that a data value $x$ is above or below the mean. Also called standardized score or standardized value

Question: Who did **relatively better?**

Answer: The **higher** z-score

**Measures of center: μ, $\bar{x}$, mean, median, mode**

**Measures of spread: variance, standard deviation, range, IQR, σ, $\sigma^2$, $s^2$, s**