Unit Two
Descriptive Biostatistics

Dr. Hamza Aduraidi

Biostatistics

- What is the biostatistics? A branch of applied math. that deals with collecting, organizing and interpreting data using well-defined procedures.
- Types of Biostatistics:
  - Descriptive Statistics. It involves organizing, summarizing & displaying data to make them more understandable.
  - Inferential Statistics. It reports the degree of confidence of the sample statistic that predicts the value of the population parameter.

Descriptive Biostatistics

- The best way to work with data is to summarize and organize them.
- Numbers that have not been summarized and organized are called raw data.

Definition

- Data is any type of information
- Raw data is a data collected as they receive.
- Organize data is data organized either in ascending, descending or in a grouped data.

Descriptive Measures

- Measures of Location
  - Measures of central tendency: Mean; Median; Mode
  - Measures of non-central tendency - Quantiles
    - Quartiles; Quintiles; Percentiles
- Measures of Dispersion
  - Range
  - Interquartile range
  - Variance
  - Standard Deviation
  - Coefficient of Variation
- Measures of Shape
  - Mean > Median-positive or right Skewness
  - Mean = Median- symmetric or zero Skewness
  - Mean < Median-Negative of left Skewness

Measures of Location

- It is a property of the data that they tend to be clustered about a center point.
- Measures of *central tendency* (i.e., central location) help find the approximate center of the dataset.
- Researchers usually do not use the term average, because there are three alternative types of average.
- These include the *mean*, the *median*, and the *mode*. In a perfect world, the mean, median & mode would be the same.
  - mean (generally not part of the data set)
  - median (may be part of the data set)
  - mode (always part of the data set)

Commonly Used Symbols

For a Sample

$\bar{x}$	sample mean
s^2	sample variance
s	sample standard deviation

For a Population

μ	population mean
σ^2	population variance
σ	population standard deviation

The Mean

- The sample mean is the sum of all the observations (ΣX_i) divided by the number of observations (n):

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{where } \Sigma X_i = X_1 + X_2 + X_3 + X_4 + \ldots + X_n$$

- **Example.** 1, 2, 2, 4, 5, 10. Calculate the mean. Note: n = 6 (six observations)

$$\Sigma X_i = 1 + 2 + 2 + 4 + 5 + 10 = 24$$

$$\bar{X} = 24 / 6 = 4.0$$

General Formula--Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

μ = population mean

Σ = summation sign

x_i = value of element i of the sample

N = population size

Notes on Sample Mean $\overline{X}$

- **Formula**

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Summation Sign**
  - In the formula to find the mean, we use the "summation sign" — $\sum$
  - This is just mathematical shorthand for "add up all of the observations"

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \dots\dots + X_n$$

Notes on Sample Mean

- Also called *sample average* or *arithmetic mean*
- Mean for the sample = $\overline{X}$ *or M, Mean for population = mew (μ)*
- Uniqueness: For a given set of data there is one and only one mean.
- Simplicity: The mean is easy to calculate.
- Sensitive to extreme values

The Mean

Example.

For the data: 1, 1, 1, 1, 51. Calculate the mean.

Note: n = 5 (five observations)

$\Sigma X_i = 1 + 1 + 1 + 1 + 51 = 55$

$\bar{X} = 55 / 5 = 11.0$

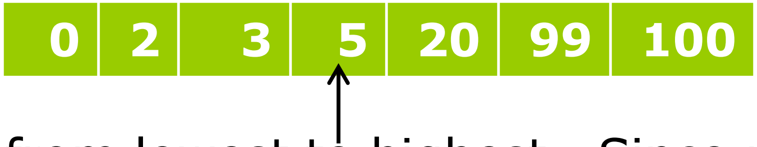
- Here we see that the mean is affected by extreme values.

The Median

- The median is the middle value of the *ordered* data
- To get the median, we must first rearrange the data into an ***ordered array*** (in ascending or descending order). Generally, we order the data from the lowest value to the highest value.
- Therefore, the median is the data value such that half of the observations are larger and half are smaller. It is also the 50th percentile.
- If n is odd, the median is the middle observation of the ordered array. If n is even, it is midway between the *two* central observations.

The Median

Example:



0	2	3	5	20	99	100

Note: Data has been ordered from lowest to highest. Since n is odd (n=7), the median is the (n+1)/2 ordered observation, or the 4th observation.

Answer: The median is 5.

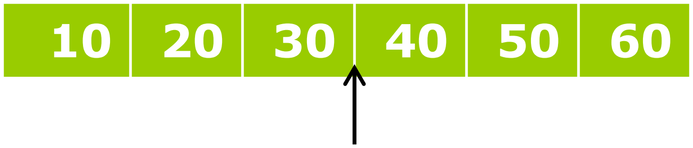
The mean and the median are unique for a given set of data. There will be exactly one mean and one median.

Unlike the mean, the median is not affected by extreme values.

Q: What happens to the median if we change the 100 to 5,000? Not a thing, the median will still be 5. Five is still the middle value of the data set.

The Median

Example:

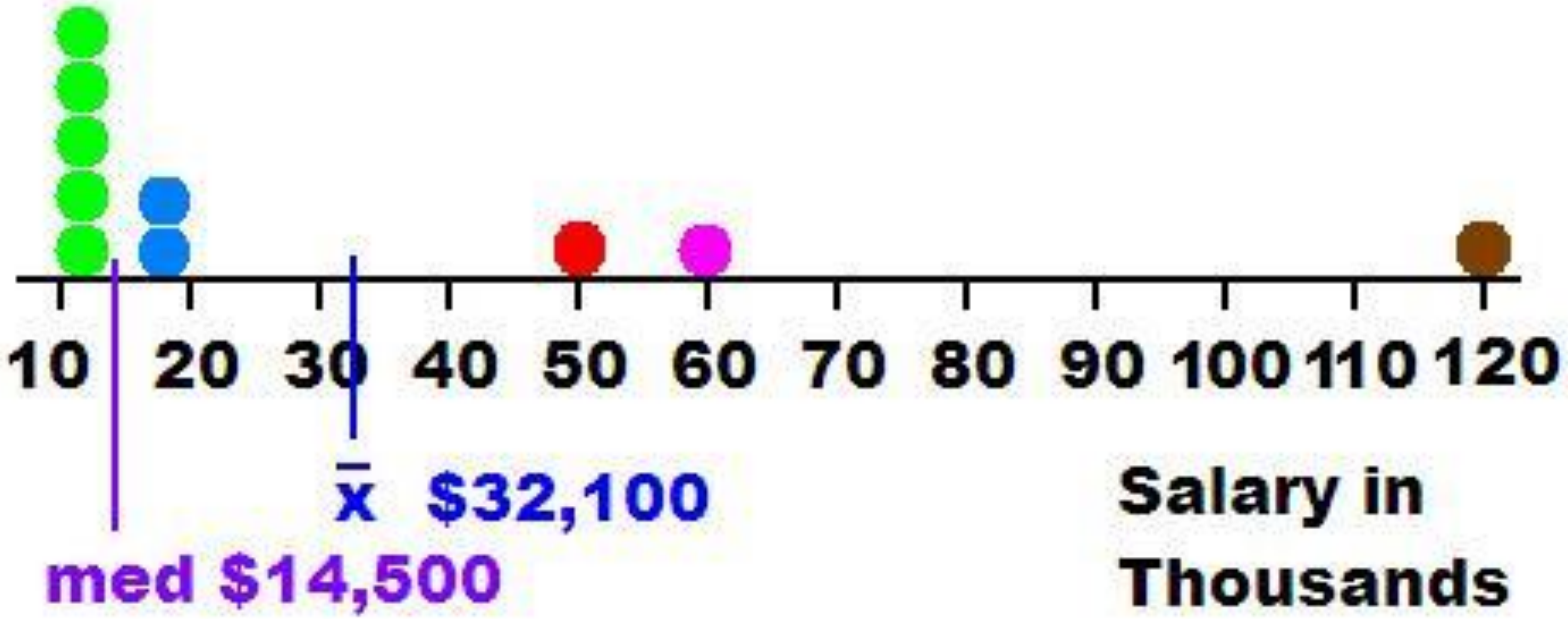


10 20 30 40 50 60

Note: Data has been ordered from lowest to highest. Since n is even (n=6), the median is the (n+1)/2 ordered observation, or the 3.5th observation, *i.e.,* the average of observation 3 and observation 4.

Answer: The median is 35.

Mean and Median in a set of Salary Data
XYZ Corp



The Mode

- The mode is the value of the data that occurs with the greatest frequency.
- Unstable index: values of modes tend to fluctuate from one sample to another drawn from the same population

Example. 1, 1, 1, 2, 3, 4, 5

Answer. The mode is 1 since it occurs three times. The other values each appear only once in the data set.

Example. 5, 5, 5, 6, 8, 10, 10, 10.

Answer. The mode is: 5, 10.

There are two modes. This is a ***bi-modal*** dataset.

The Mode

- The mode is different from the mean and the median in that those measures always exist and are always unique. For any numeric data set there will be one mean and one median.
- *The mode may not exist*.
  - Data: 1, 2, 3, 4, 5, 6, 7, 8, 9, 0
  - Here you have 10 observations and they are all different.
- *The mode may not be unique*.
  - Data: 0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7
  - Mode = 1, 2, 3, 4, 5, and 6. There are ***six*** modes.

Comparison of the Mode, the Median, and the Mean

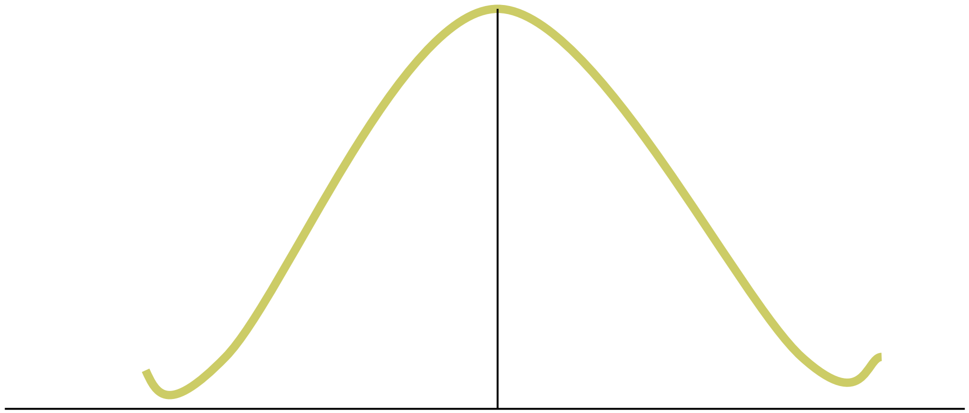
- In a normal distribution, the mode , the median, and the mean have the same value.
- The mean is the widely reported index of central tendency for variables measured on an interval and ratio scale.
- The mean takes each and every score into account.
- It also the most stable index of central tendency and thus yields the most reliable estimate of the central tendency of the population.

Comparison of the Mode, the Median, and the Mean

- The mean is always pulled in the direction of the long tail, that is, in the direction of the extreme scores.
- For the variables that positively skewed (like income), the mean is higher than the mode or the median. For negatively skewed variables (like age at death) the mean is lower.
- When there are extreme values in the distribution (even if it is approximately normal), researchers sometimes report means that have been adjusted for outliers.
- To adjust means one must discard a fixed percentage (5%) of the extreme values from either end of the distribution.

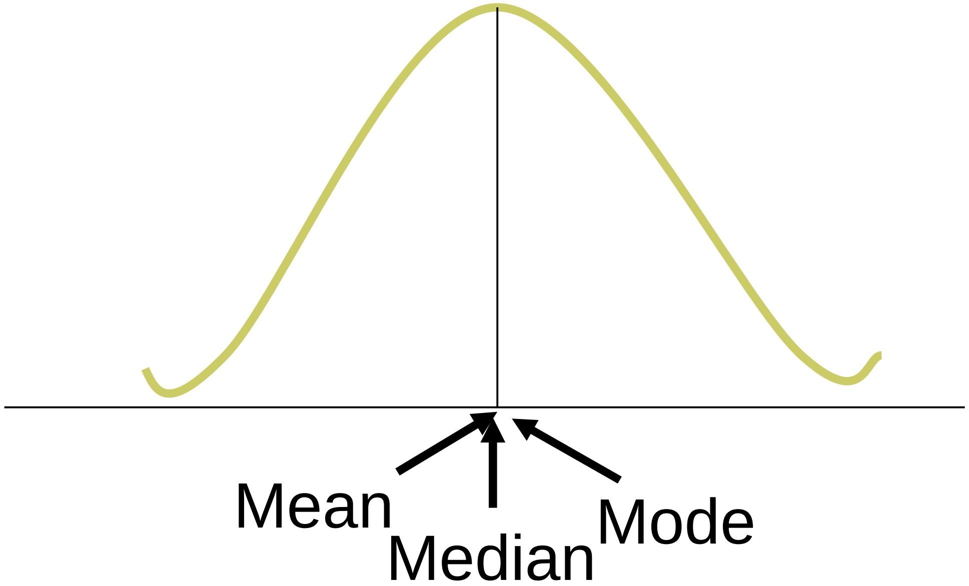
Distribution Characteristics

- Mode: Peak(s)
- Median: Equal areas point
- Mean: Balancing point



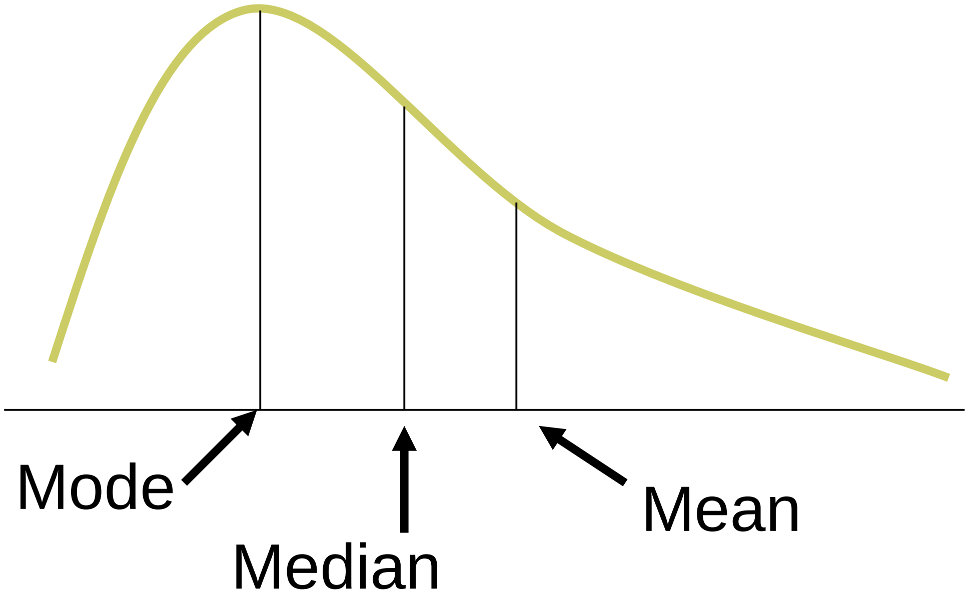
Shapes of Distributions

- **Symmetric** (Right and left sides are mirror images)
  - Left tail looks like right tail
  - Mean = Median = Mode



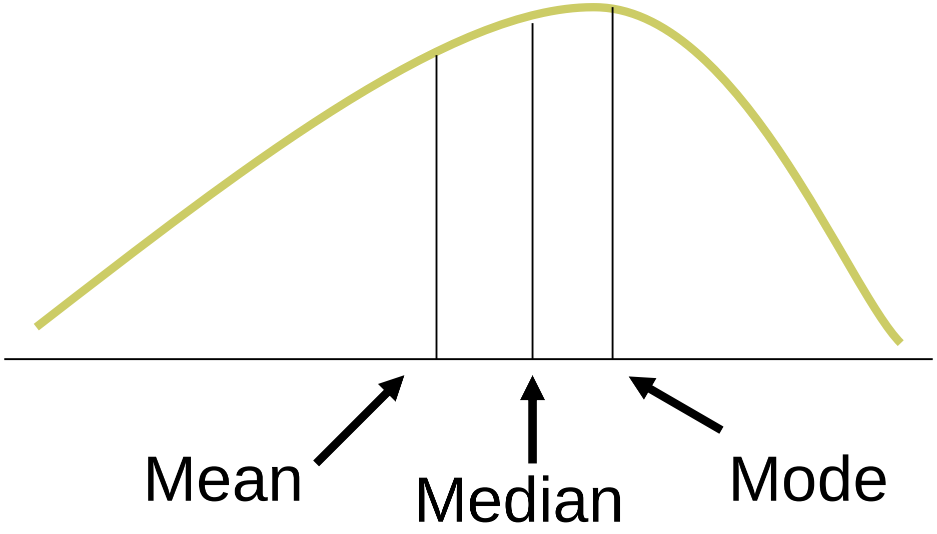
Shapes of Distributions

- **Right skewed** (positively skewed)
  - Long right tail
  - Mean > Median



Shapes of Distributions

- **Left skewed** (negatively skewed)
  - Long left tail
  - Mean < Median



Quantiles

- Measures of non-central location used to summarize a set of data
- Examples of commonly used quantiles:
  - Quartiles
  - Quintiles
  - Deciles
  - Percentiles

Quartiles

- *Quartiles* split a set of ordered data into four parts.
  - Imagine cutting a chocolate bar into four equal pieces… How many cuts would you make? (yes, 3!)
- Q_1 is the First Quartile
  - 25% of the observations are smaller than Q_1 and 75% of the observations are larger
- Q_2 is the Second Quartile
  - 50% of the observations are smaller than Q_2 and 50% of the observations are larger. Same as the Median. It is also the 50th percentile.
- Q_3 is the Third Quartile
  - 75% of the observations are smaller than $Q_3$and 25% of the observations are larger

Other Quantiles

- Similar to what we just learned about quartiles, where 3 quartiles split the data into 4 equal parts,
  - There are 9 ***deciles*** dividing the distribution into 10 equal portions (tenths).
  - There are four ***quintiles*** dividing the population into 5 equal portions.
  - … and 99 ***percentiles*** (next slide)
- In all these cases, the convention is the same. The point, be it a quartile, decile, or *percentile*, takes the value of one of the observations or it has a value halfway between two adjacent observations. It is never necessary to split the difference between two observations more finely.

Measures of Dispersion

- It refers to how spread out the scores are.

- In other words, how similar or different participants are from one another on the variable. It is either homogeneous or heterogeneous sample.

- Why do we need to look at measures of dispersion?

Measures of Dispersion

- We will study these five measures of dispersion
  - Range
  - Interquartile Range
  - Standard Deviation
  - Variance
  - Coefficient of Variation
  - Relative Standing.

The Range

- Is the simplest measure of variability, is the difference between the highest score and the lowest score in the distribution.
- In research, the range is often shown as the minimum and maximum value, without the abstracted difference score.
- It provides a quick summary of a distribution's variability.
- It also provides useful information about a distribution when there are extreme values.
- The range has two values, it is highly unstable.

The Range

- Range = Largest Value – Smallest Value

  Example: 1, 2, 3, 4, 5, 8, 9, 21, 25, 30

  Answer: Range = 30 – 1 = 29.

- Pros:
  - Easy to calculate
- Cons:
  - Value of range is only determined by two values
  - The interpretation of the range is difficult.
  - One problem with the range is that it is influenced by extreme values at either end.

Standard Deviation

- The standard deviation, *s*, measures a kind of "average" deviation about the mean. It is not really the "average" deviation, even though we may think of it that way.

- Why can't we simply compute the average deviation about the mean, if that's what we want?

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n}$$

- If you take a simple mean, and then add up the deviations about the mean, as above, this sum will be equal to 0. Therefore, a measure of "average deviation" will not work.

Standard Deviation

- Instead, we use:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- This is the "definitional formula" for standard deviation.
- The standard deviation has lots of nice properties, including:
  - By squaring the deviation, we eliminate the problem of the deviations summing to zero.
  - In addition, this sum is a minimum. No other value subtracted from X and squared will result in a smaller sum of the deviation squared. This is called the "least squares property."
- Note we divide by (n–1), not n. This will be referred to as a loss of one degree of freedom.

Standard Deviation

- The **smaller** the standard deviation, the **better** is the mean as the summary of a typical score. E.g. 10 people weighted 150 pounds, the SD would be zero, and the mean of 150 would communicate perfectly accurate information about all the participants wt. Another example would be a heterogeneous sample 5 people 100 pounds and another five people 200 pounds. The mean still 150, but the SD would be 52.7.

Standard Deviation

Example. Two data sets, X and Y. Which of the two data sets has greater variability? Calculate the standard deviation for each.

We note that both sets of data have the same mean:

$$\overline{X} = 3$$

$$\overline{Y} = 3$$

X_i	Y_i
1	0
2	0
3	0
4	5
5	10

(continued…)

Standard Deviation

$$S_X = \sqrt{\frac{10}{4}} = 1.58$$

X	$\overline{X}$	$(X-\overline{X})$	$(X-\overline{X})^2$
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
		$\sum=0$	10

$$S_Y = \sqrt{\frac{80}{4}} = \, = 4.47$$

Y	$\overline{Y}$	$(Y-\overline{Y})$	$(Y-\ \overline{Y})^2$
0	3	-3	9
0	3	-3	9
0	3	-3	9
5	3	2	4
10	3	7	49
		$\sum=0$	80

[Check these results with your calculator.]

Variance

- The variance, s^2, is the standard deviation (s) squared. Conversely, $s = \sqrt{variance}$.

Definitional formula: $s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$

Computational formula: $s^2 = \frac{\sum_{i=1}^{n}(X_i)^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}$

This is what computer software
(e.g., MS Excel or your calculator key) uses.

Coefficient of Variation (CV)

- The problem with s^2 and s is that they are both, like the mean, in the “original” units.
- This makes it difficult to compare the variability of two data sets that are in different units or where the magnitude of the numbers is very different in the two sets. For example,
  - Suppose you wish to compare two stocks and one is in dollars and the other is in yen; if you want to know which one is more volatile, you should use the coefficient of variation.
  - It is also not appropriate to compare two stocks of vastly different prices even if both are in the same units.
  - The standard deviation for a stock that sells for around $300 is going to be very different from one with a price of around $0.25.
- The *coefficient of variation* will be a better measure of dispersion in these cases than the standard deviation (see example on the next slide).

$$CV = \frac{s}{\bar{X}}(100\%)$$

Coefficient of Variation (CV)

$CV = \frac{s}{\bar{X}}(100\%)$

CV is in terms of a percent. What we are in effect calculating is what percent of the sample mean is the standard deviation. If CV is 100%, this indicates that the sample mean is equal to the sample standard deviation. This would demonstrate that there is a great deal of variability in the data set. 200% would obviously be even worse.

IQR

- The Interquartile range (IQR) is the score at the 75th percentile or 3rd quartile (Q3) minus the score at the 25th percentile or first quartile (Q1). Are the most used to define outliers.
- It is not sensitive to extreme values.

Inter-Quartile Range (IQR)

- IQR = $Q_3 - Q_1$
- Example (n = 15):

  0, 0, 2, 3, 4, 7, 9, 12, 17, 18, 20, 22, 45, 56, 98

  $Q_1 = 3$, $Q_3 = 22$

  IQR = 22 – 3 = 19 (Range = 98)
- This is basically the range of the central 50% of the observations in the distribution.
- Problem: The Interquartile range does not take into account the variability of the *total* data (only the central 50%). We are "throwing out" half of the data.

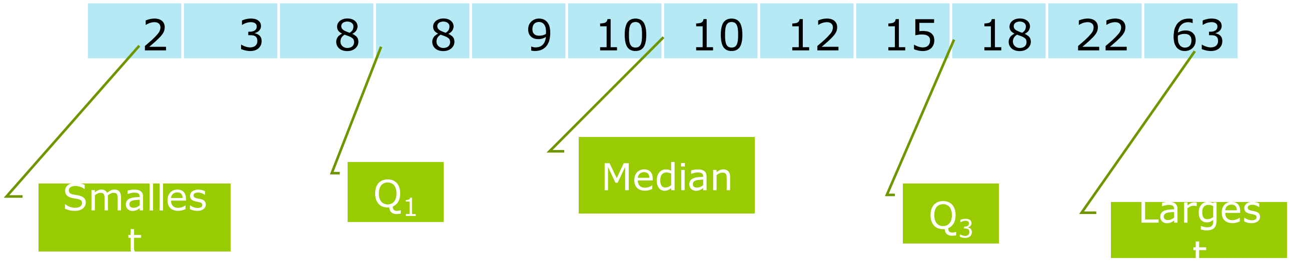
Five Number Summary

- When examining a distribution for shape, sometime the five number summary is useful:

  Smallest| Q1 | Median | Q3 | Largest

- Example:

$\overline{X} = 15$



2	3	8	8	9	10	10	12	15	18	22	63

Smallest

Q_1

Median

Q_3

Largest

5-number summary: 2 | 8 | 10 | 16.5 | 63

This data is right-skewed.

In right-skewed distributions, the distance from Q_3 to $X_{largest}$ (16.5 to 63) is significantly greater than the distance from $X_{smallest}$ to Q_1(2 to 8).