UNIT 1

Basic Statistics and Introduction to Probability

WHY STUDY STATISTICS?

1. Data are everywhere

2. Statistical techniques are used to make many decisions that affect our lives

3. No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions effectively

STATISTICS

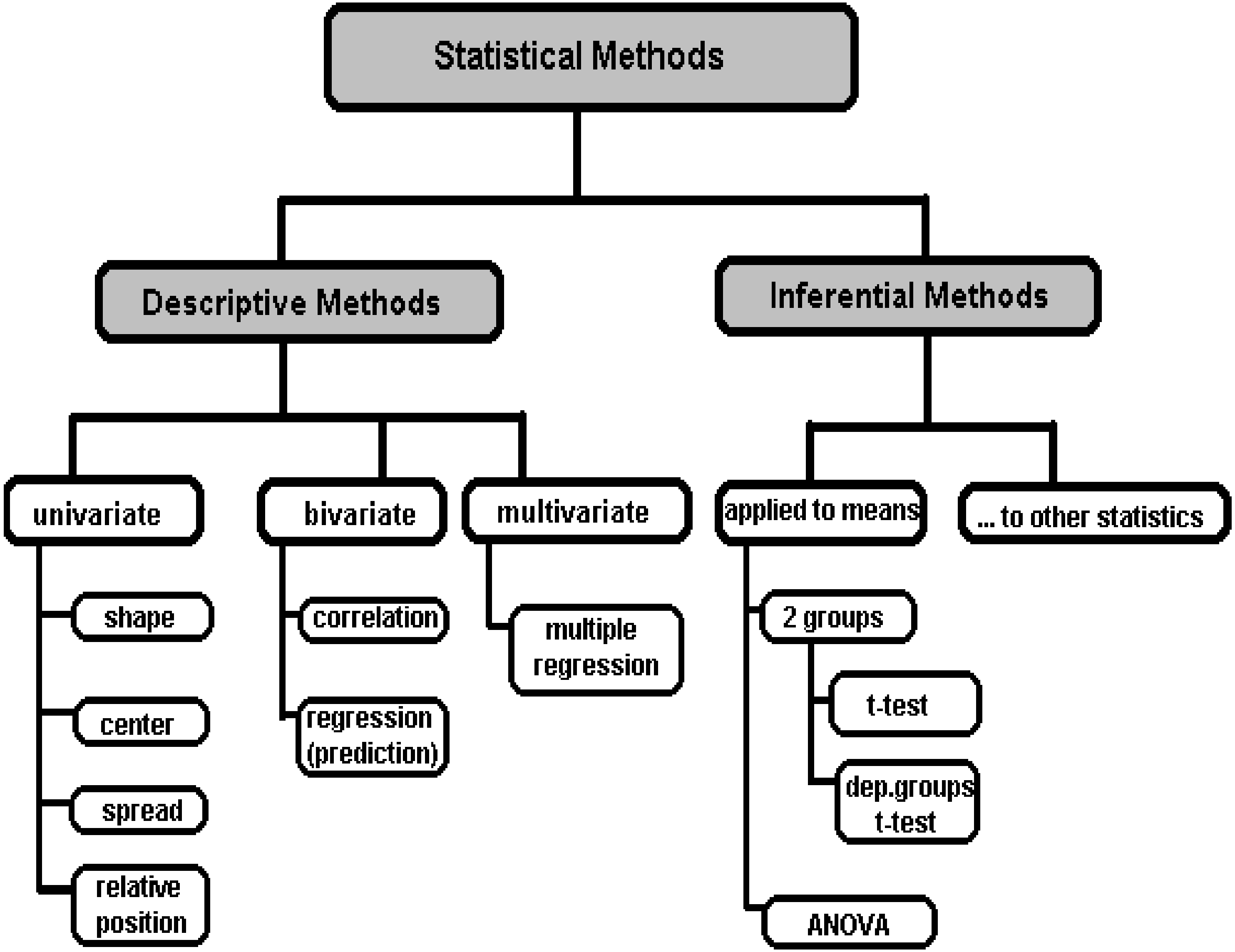
- The science of collectiong, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions

- Statistical analysis – used to manipulate summarize, and investigate data, so that useful decision-making information results.

Types of statistics

- **Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way

- **Inferential statistics** – The methods used to determine something about a population on the basis of a sample
  - Population –The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
  - Sample – A portion, or part, of the population of interest

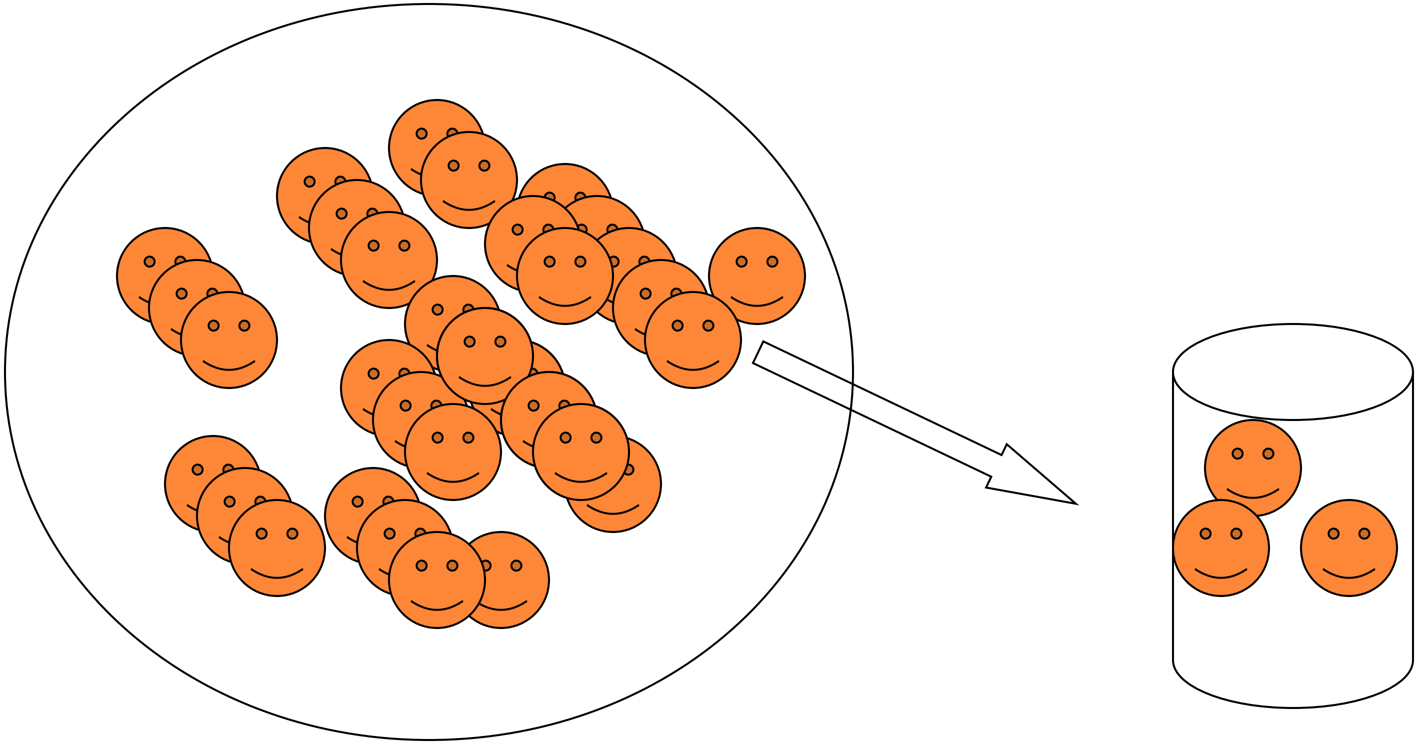

Statistical Methods
Descriptive Methods
Inferential Methods
univariate
bivariate
multivariate
applied to means
... to other statistics
shape
center
spread
relative position
correlation
regression (prediction)
multiple regression
2 groups
t-test
dep.groups t-test
ANOVA

Descriptive Statistics

- Descriptive Statistics are Used by Researchers to Report on Populations <u>and</u> Samples

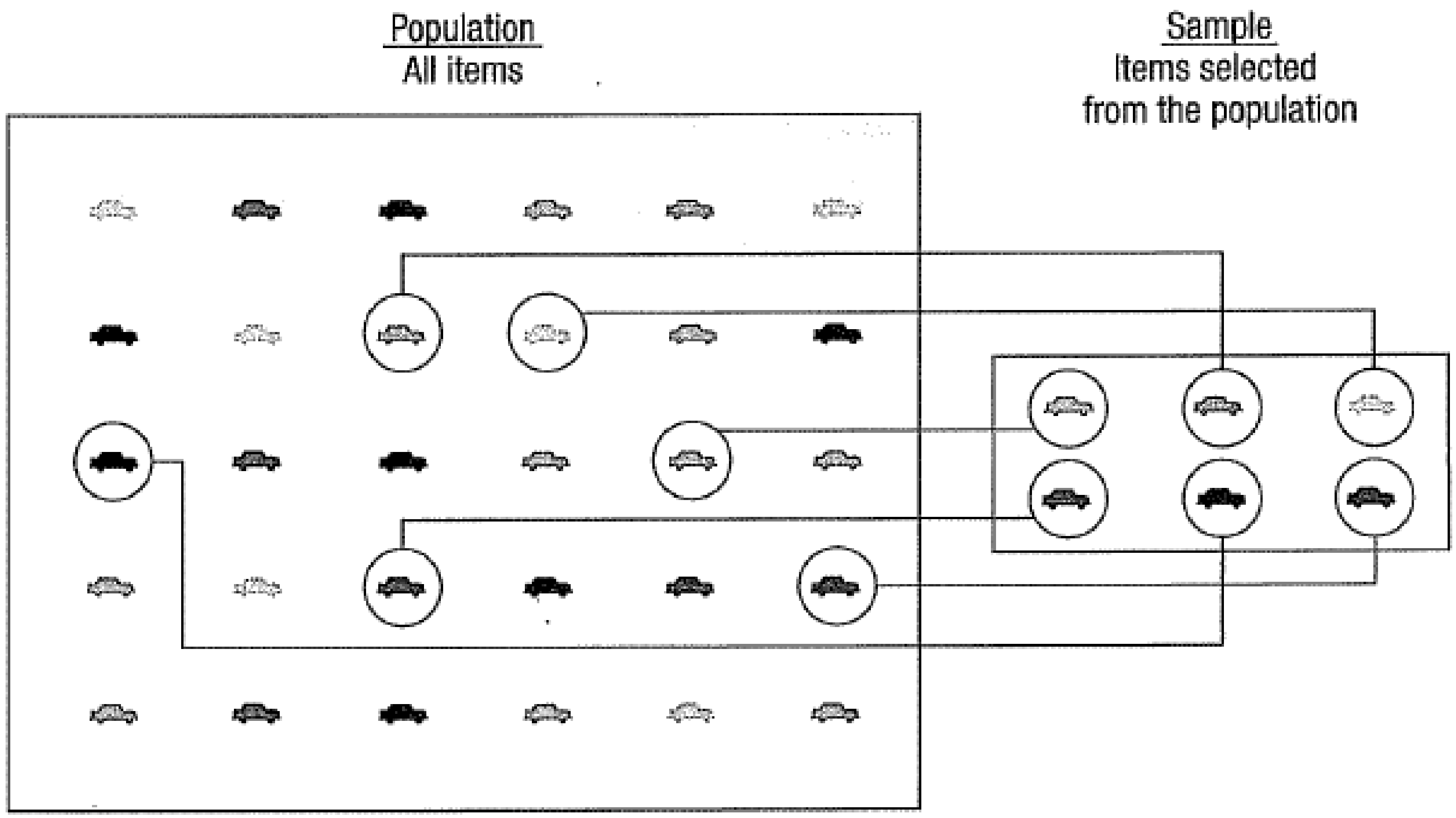
- By Summarizing Information, Descriptive Statistics Speed Up and Simplify Comprehension of a Group's Characteristics

Sample vs. Population



Population

Sample



Population
All items
Sample
Items selected
from the population

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kg

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Descriptive Statistics

An Illustration:

Which Group is Smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.

Descriptive Statistics

Which group is smarter now?

Class A--Average IQ	Class B--Average IQ
110.54	110.23

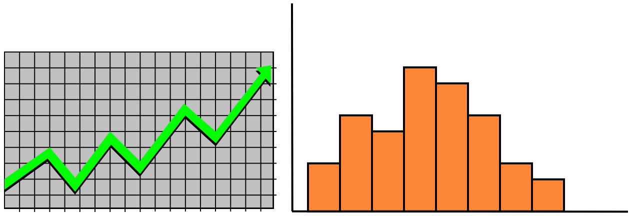
They're roughly the same!

With a summary descriptive statistic, it is much easier to answer our question.

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs



- Summarize data
  - e.g., Sample mean $= \dfrac{\sum X_i}{n}$

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs

- Summarize Data
  - Central Tendency
  - Variation

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
  - Tables
    - Frequency Distributions
    - Relative Frequency Distributions
  - Graphs
    - Bar Chart or Histogram
    - Stem and Leaf Plot
    - Frequency Polygon

Descriptive Statistics

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

MEAN

Most commonly called the “average.”

Add up the values for each case and divide by the total number of cases.

$$\text{Y-bar} = \frac{(Y1 + Y2 + \ldots + Yn)}{n}$$

$$\text{Y-bar} = \frac{\Sigma Y i}{n}$$

Mean

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	9	87
93	97	120	105
110		109	

Class A: $\Sigma Y_i = 1437$

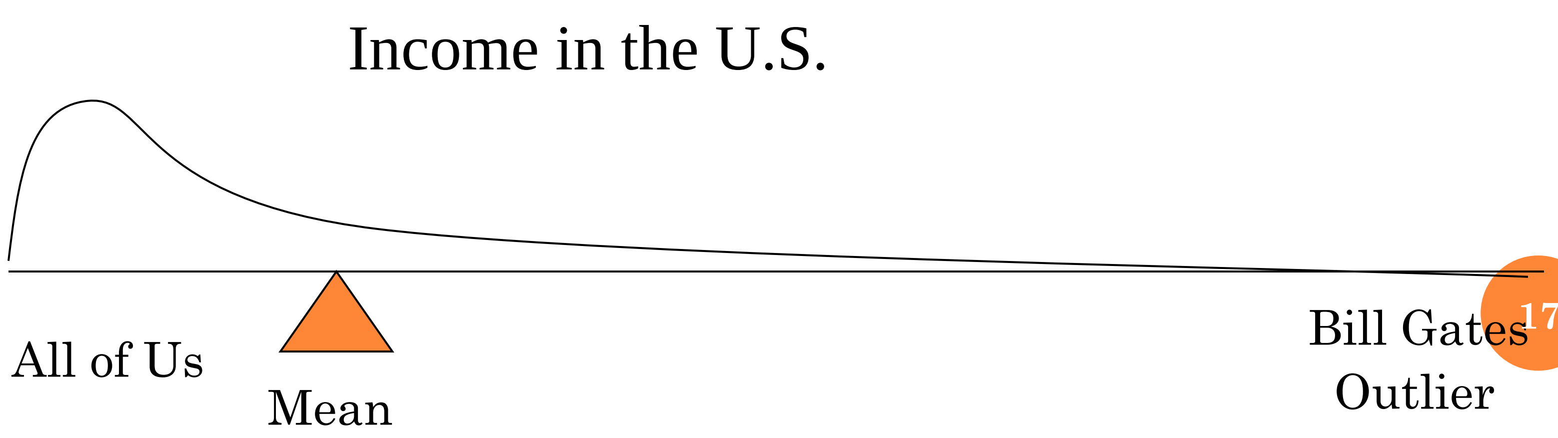
$$\text{Y-bar}_A = \frac{\Sigma Y_i}{n} = \frac{1437}{13} = 110.54$$

Class B: $\Sigma Y_i = 1433$

$$\text{Y-bar}_B = \frac{\Sigma Y_i}{n} = \frac{1433}{13} = 110.23$$

Mean

1. An individual value that falls outside the overall pattern is called an *outlier*.
2. Means can be badly affected by outliers (data points with extreme values unlike the rest)
3. Outliers can make the mean a bad measure of central tendency or common experience



Median

The middle value when a variable's values are ranked in order; the point that divides a distribution into two equal halves.

When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it.

The 50th percentile.

Median

Class A--IQs of 13 Students

89

93

97

98

102

106

109

110

115

119

128

131

140

Median = 109

(six cases above, six below)

MEDIAN

If the first student were to drop out of Class A, there would be a new median:

~~89~~
93
97
98
102
106
109
110
115
119
128
131
140

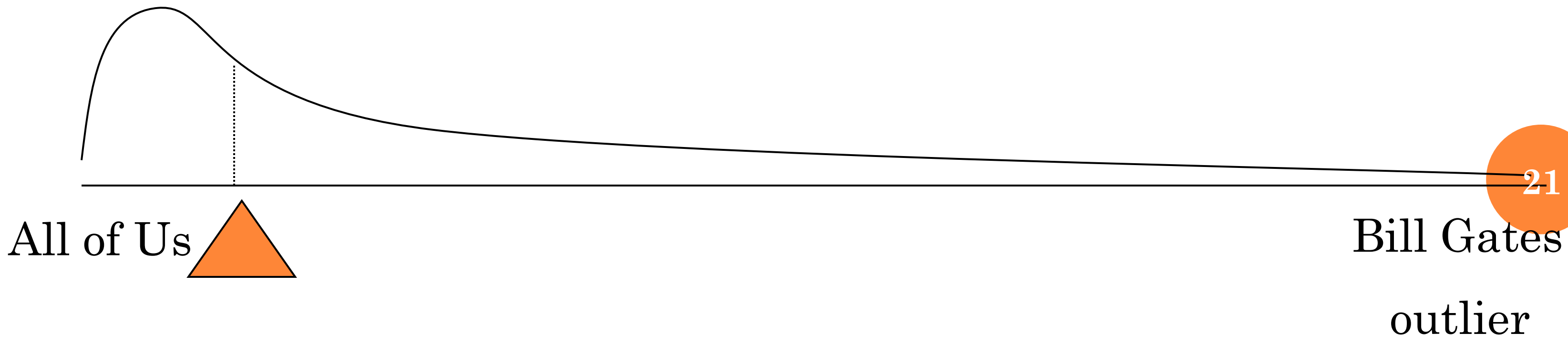
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

MEDIAN

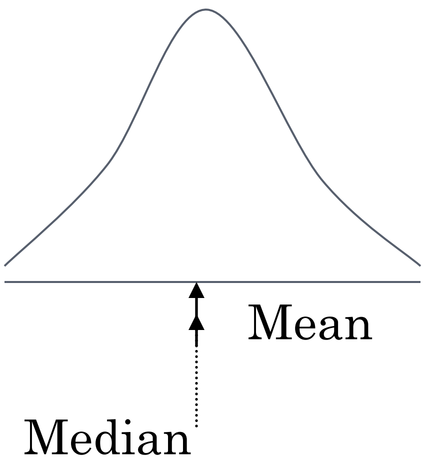
1. The median is unaffected by outliers, making it a better measure of central tendency, better describing the "typical person" than the mean when data are skewed.



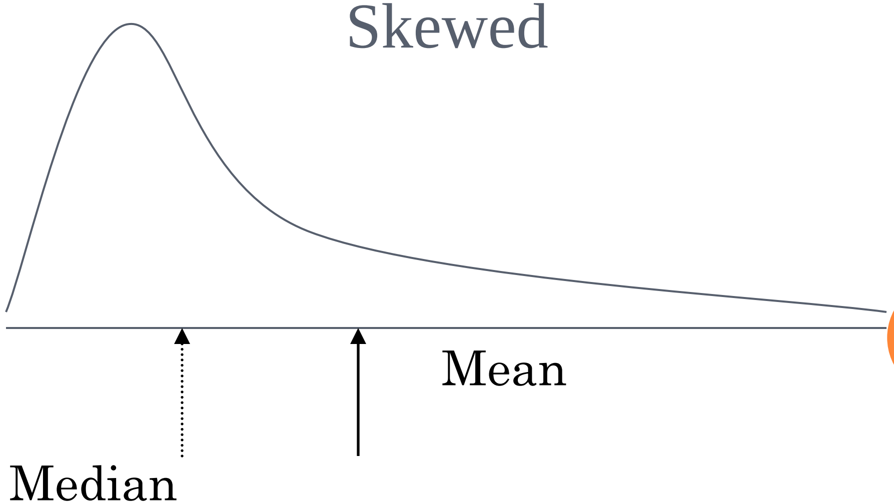
Median

2. If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
3. In skewed data, the mean lies further toward the skew than the median.

Symmetric



Skewed



Mode

The most common data point is called the mode.

The combined IQ scores for Classes A & B:

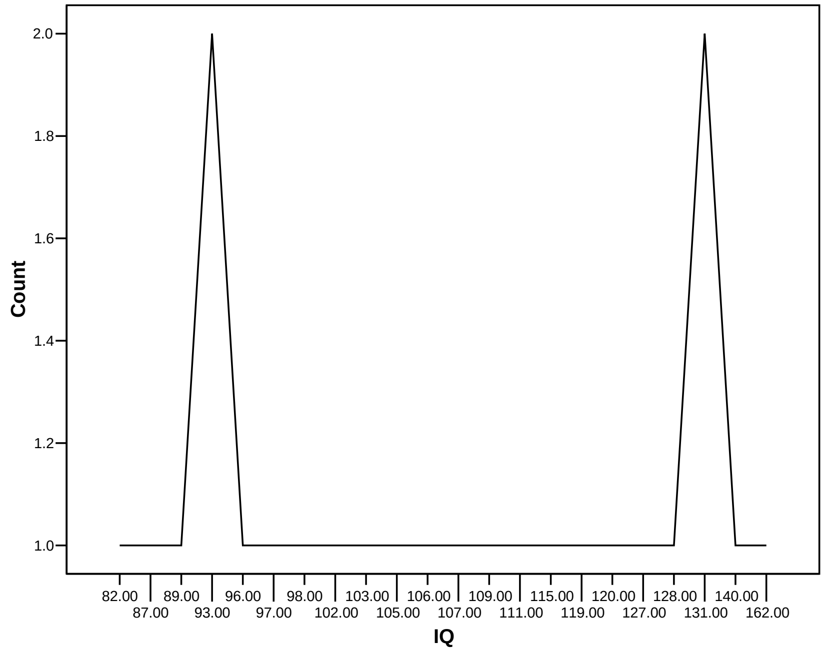
80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115 119 120 127 128 131 131 140 162

BTW, It is possible to have more than one mode!

Mode

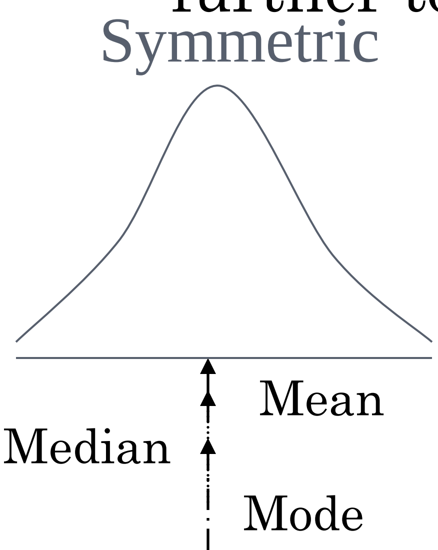
It may not be at the center of a distribution.

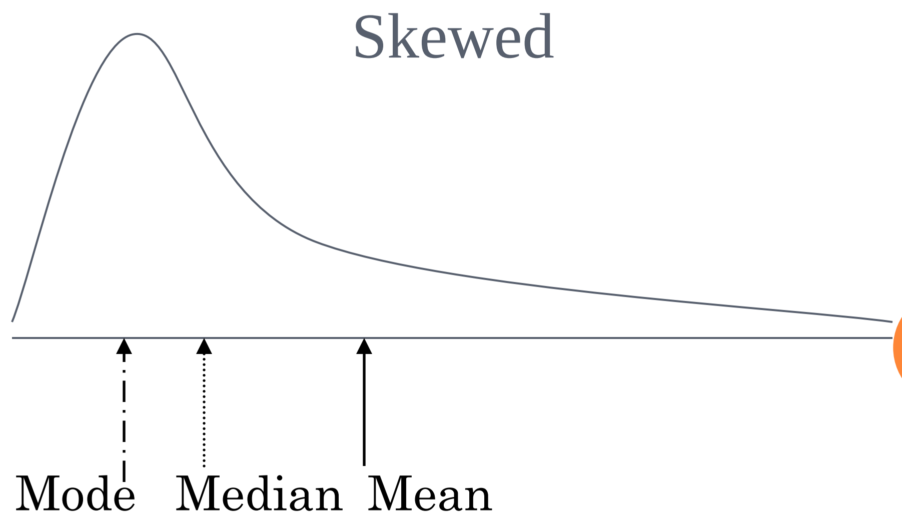
Data distribution on the right is “bimodal” (even statistics can be open-minded)



Mode

1. It may give you the most likely experience rather than the "typical" or "central" experience.
2. In symmetric distributions, the mean, median, and mode are the same.
3. In skewed data, the mean and median lie further toward the skew than the mode.





Range

The spread, or the distance, between the lowest and highest values of a variable.

To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students	
102	115
128	109
131	89
98	106
140	119
93	97
110	

Class A Range = 140 - 89 = 51

Class B--IQs of 13 Students	
127	162
131	103
96	111
80	109
93	87
120	105
109	

Class B Range = 162 - 80 = 82

Interquartile Range

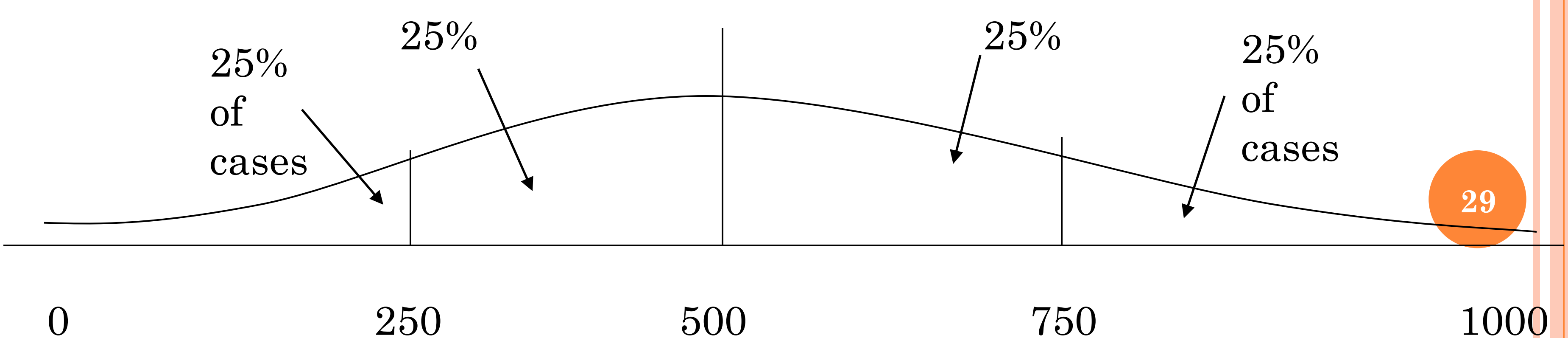
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25th percentile is a quartile that divides the first ¼ of cases from the latter ¾.

75th percentile is a quartile that divides the first ¾ of cases from the latter ¼.

The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



In the following example Q1= ((15+1)/4)1 =4^{th} observation of the data. The 4^{th} observation is 11. So Q1 is of this data is 11.

An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

Q1 Q2 Q3

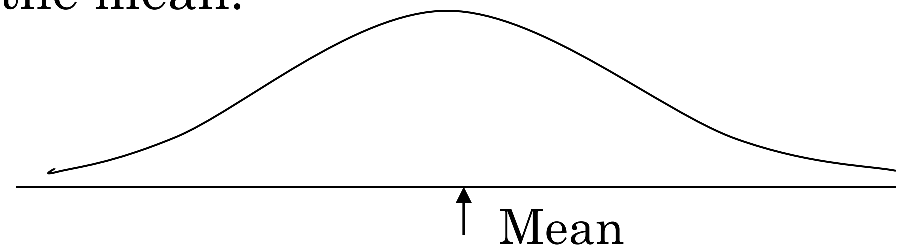
The first quartile is Q1=11. The second quartile is Q2=40 (This is also the Median.) The third quartile is Q3=61.

Inter-quartile Range: Difference between Q3 and Q1. Inter-quartile range of the previous example is 61- 40=21. The middle half of the ordered data lie between 40 and 61.

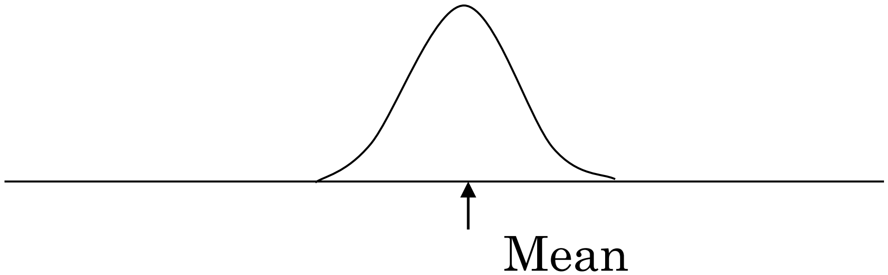
Variance

A measure of the spread of the recorded values on a variable. A measure of dispersion.

The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



Variance

Variance is a number that at first seems complex to calculate.

Calculating variance starts with a "deviation."

A deviation is the distance away from the mean of a case's score.

Y*i* – Y-bar

Variance

The deviation of 102 from 110.54 is?   Deviation of 115?

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar$_A$ ***= 110.54***

Variance

The deviation of 102 from 110.54 is?

102 - 110.54 = -8.54

Deviation of 115?

115 - 110.54 = 4.46

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar$_A$ ***= 110.54***

Variance

- We want to add these to get total deviations, but if we were to do that, we would get zero every time. Why?
- We need a way to eliminate negative signs.

Squaring the deviations will eliminate negative signs...

A Deviation Squared: $(\mathrm{Y}i - \text{Y-bar})^2$

Back to the IQ example,

A deviation squared for 102 is: of 115:

$(102 - 110.54)^2 = (-8.54)^2 = 72.93$ $(115 - 110.54)^2 = (4.46)^2 = 19.89$

Variance

If you were to add all the squared deviations together, you'd get what we call the "Sum of Squares."

Sum of Squares (SS) = $\Sigma\ (Y\textit{i} - \text{Y-bar})^2$

SS = $(Y1 - \text{Y-bar})^2 + (Y2 - \text{Y-bar})^2 + \ldots + (Yn - \text{Y-bar})^2$

Variance

Class A, sum of squares:

$(102 – 110.54)^2 + (115 – 110.54)^2 +$
$(126 – 110.54)^2 + (109 – 110.54)^2 +$
$(131 – 110.54)^2 + (89 – 110.54)^2 +$
$(98 – 110.54)^2 + (106 – 110.54)^2 +$
$(140 – 110.54)^2 + (119 – 110.54)^2 +$
$(93 – 110.54)^2 + (97 – 110.54)^2 +$
$(110 – 110.54) = SS = 2825.39$

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar = 110.54

VARIANCE

The last step…

The approximate average sum of squares is the variance.

SS/N = Variance for a population.

SS/n-1 = Variance for a sample.

Variance = $\Sigma(Y_i - \text{Y-bar})^2 / n - 1$

Variance

For Class A, Variance = 2825.39 / n - 1

= 2825.39 / 12 = 235.45

Standard Deviation

To convert variance into something of meaning, let's create standard deviation.

The square root of the variance reveals the average deviation of the observations from the mean.

$$\text{s.d.} = \sqrt{\frac{\Sigma(Y_i - \text{Y-bar})^2}{n - 1}}$$

Standard Deviation

For Class A, the standard deviation is:

$$\sqrt{235.45} = 15.34$$

The average of persons' deviation from the mean IQ of 110.54 is 15.34 IQ points.

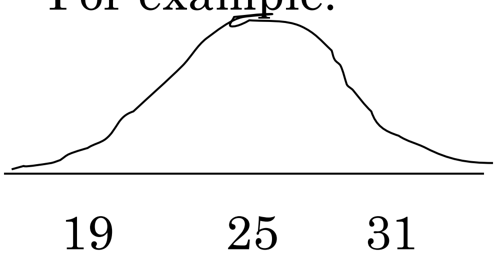
Review:

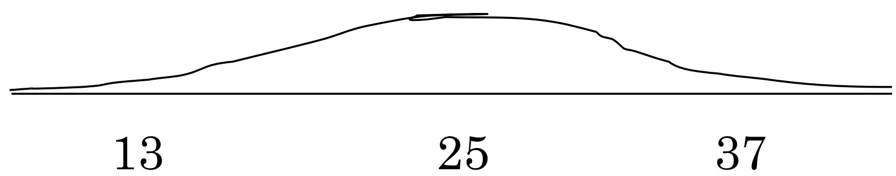
1. Deviation
2. Deviation squared
3. Sum of squares
4. Variance
5. Standard deviation

STANDARD DEVIATION

1. Larger s.d. = greater amounts of variation around the mean.

   For example:





$\overline{Y} = 25$ s.d. = 3

$\overline{Y} = 25$ s.d. = 6

2. s.d. = 0 only when all values are the same (only when you have a constant and not a “variable”)
3. If you were to “rescale” a variable, the s.d. would change by the same magnitude—if we changed units above so the mean equaled 250, the s.d. on the left would be 30, and on the right, 60
4. Like the mean, the s.d. will be inflated by an outlier case value.

Deciles and Percentiles

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25^{th} percentile is the Q1, 50^{th} percentile is the Median (Q2) and the 75^{th} percentile of the data is Q3.

In notations, percentiles of a data is the ((n+1)/100)p th observation of the data, where p is the desired percentile and n is the number of observations of data.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

$$\text{Coefficient of Variation} = \frac{\sigma}{\bar{x}} \times 100$$