

UNIT - 1

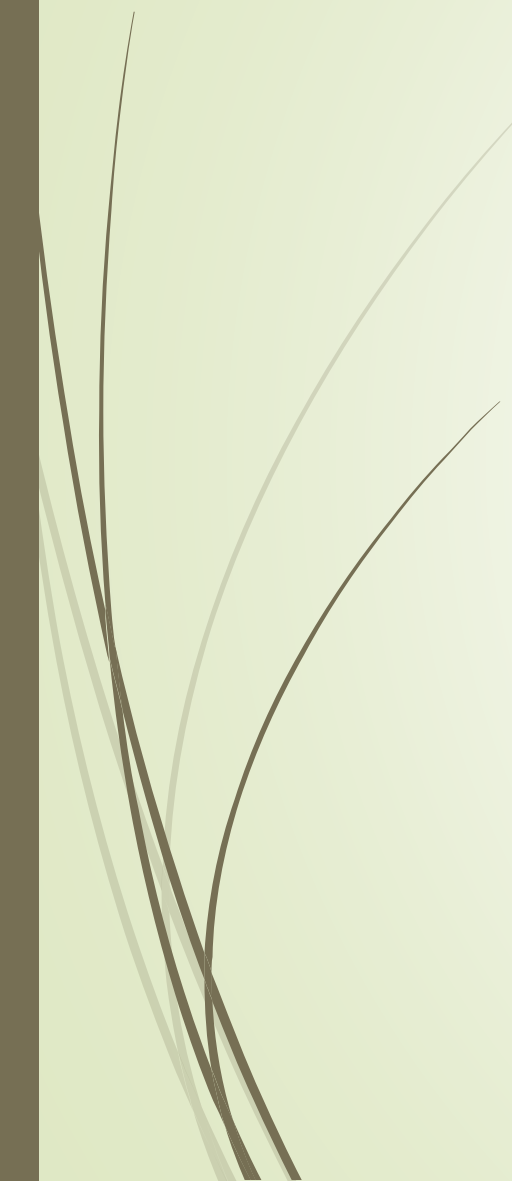
Descriptive statistics, Introduction to Analytics, Business Understanding, Introduction to R- Basics, The R Environment, Inductive and Deductive Logic, Installation of R



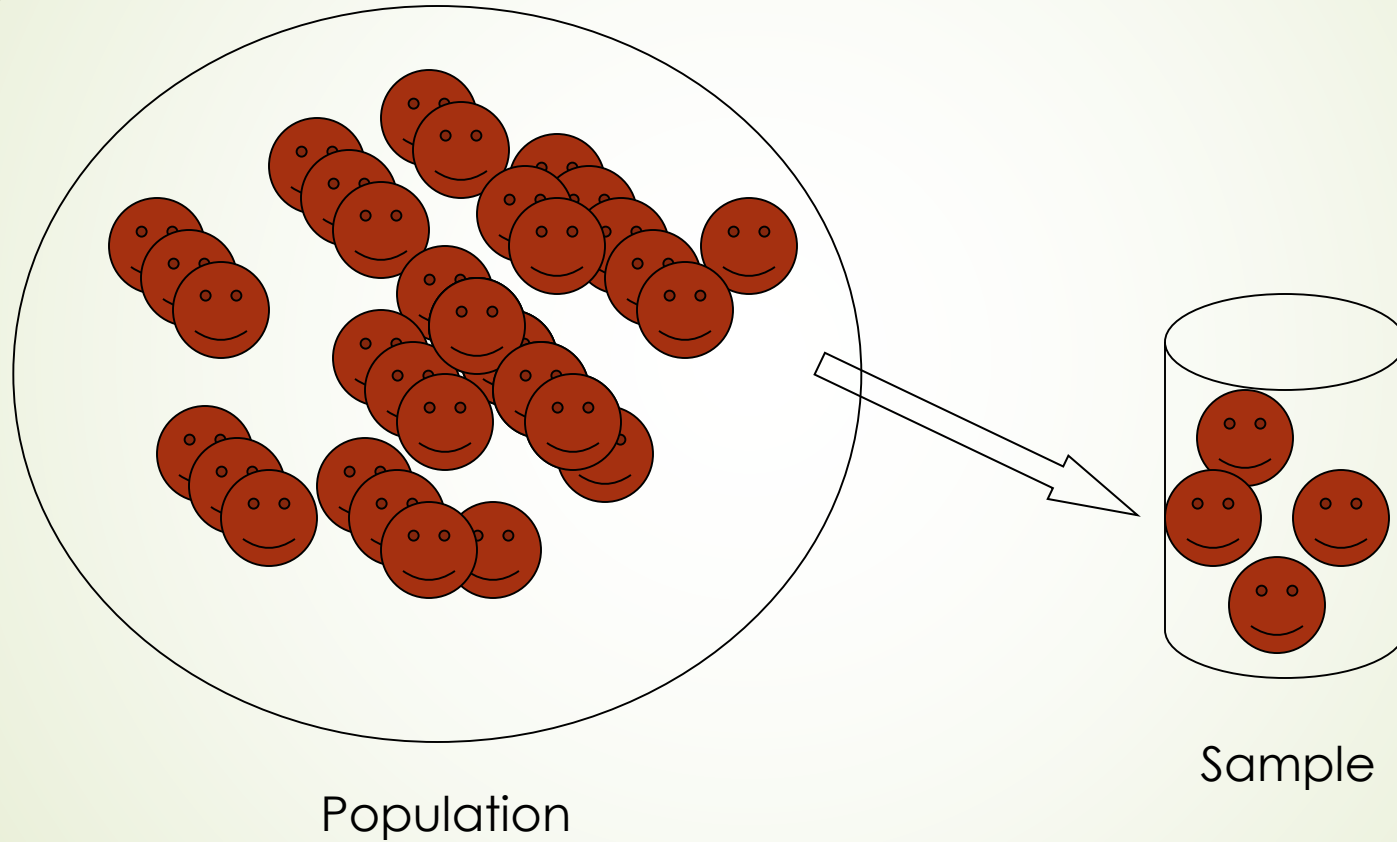
Descriptive Statistics



Descriptive Statistics

- Descriptive Statistics are Used by Researchers to Report on Populations and Samples
 - Summary descriptions of measurements (variables) taken about a group of people
 - By Summarizing Information, Descriptive Statistics Speed Up and Simplify Comprehension of a Group's Characteristics
- 

Sample vs. Population: a naïve description



Descriptive Statistics: an Illustration

Which Group/Class is Smarter?

Class A--IQs of 13 Students

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

Class B--IQs of 13 Students

127	162
131	103
96	111
80	109
93	87
120	105
109	

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



Descriptive Statistics

Which class is smarter now?

Class A: Average IQ

Class B: Average IQ

110.54

110.23

The two are roughly the same.

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



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
 - Frequency Polygon

Example: Frequency Distribution

IQ

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 82.00	1	4.2	4.2	4.2
87.00	1	4.2	4.2	8.3
89.00	1	4.2	4.2	12.5
93.00	2	8.3	8.3	20.8
96.00	1	4.2	4.2	25.0
97.00	1	4.2	4.2	29.2
98.00	1	4.2	4.2	33.3
102.00	1	4.2	4.2	37.5
103.00	1	4.2	4.2	41.7
105.00	1	4.2	4.2	45.8
106.00	1	4.2	4.2	50.0
107.00	1	4.2	4.2	54.2
109.00	1	4.2	4.2	58.3
111.00	1	4.2	4.2	62.5
115.00	1	4.2	4.2	66.7
119.00	1	4.2	4.2	70.8
120.00	1	4.2	4.2	75.0
127.00	1	4.2	4.2	79.2
128.00	1	4.2	4.2	83.3
131.00	2	8.3	8.3	91.7
140.00	1	4.2	4.2	95.8
162.00	1	4.2	4.2	100.0
Total	24	100.0	100.0	

Frequency Distribution

Frequency Distribution of IQ for Two Classes



IQ	Frequency
82.00	1
87.00	1
89.00	1
93.00	2
96.00	1
97.00	1
98.00	1
102.00	1
103.00	1
105.00	1
106.00	1
107.00	1
109.00	1
111.00	1
115.00	1
119.00	1
120.00	1
127.00	1
128.00	1
131.00	2
140.00	1
162.00	1
Total	24

Relative Frequency Distribution of IQ for Two Classes

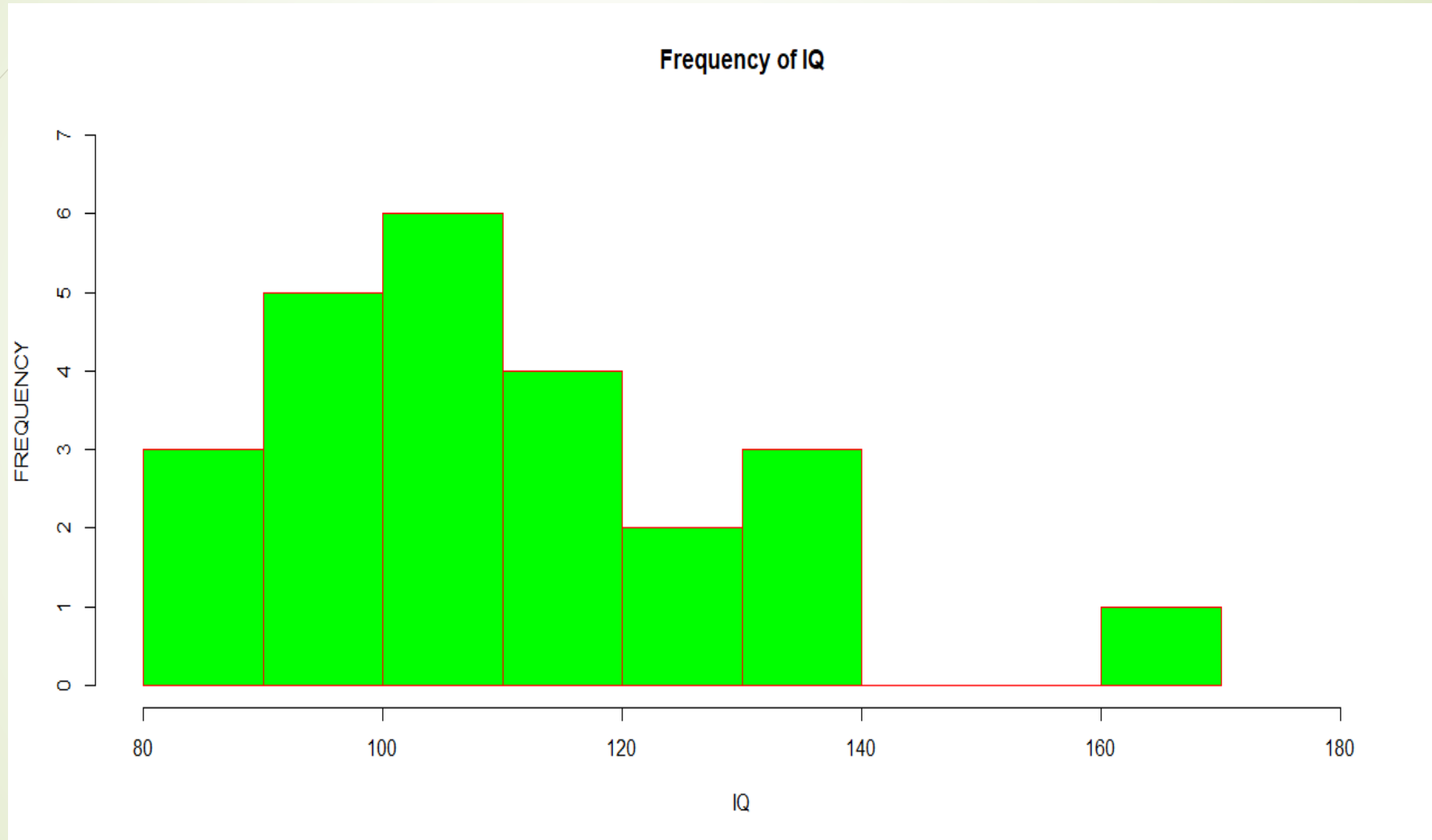
IQ	Frequency	Percent	Valid Percent	Cumulative Percent
82.00	1	4.2	4.2	4.2
87.00	1	4.2	4.2	8.3
89.00	1	4.2	4.2	12.5
93.00	2	8.3	8.3	20.8
96.00	1	4.2	4.2	25.0
97.00	1	4.2	4.2	29.2
98.00	1	4.2	4.2	33.3
102.00	1	4.2	4.2	37.5
103.00	1	4.2	4.2	41.7
105.00	1	4.2	4.2	45.8
106.00	1	4.2	4.2	50.0
107.00	1	4.2	4.2	54.2
109.00	1	4.2	4.2	58.3
111.00	1	4.2	4.2	62.5
115.00	1	4.2	4.2	66.7
119.00	1	4.2	4.2	70.8
120.00	1	4.2	4.2	75.0
127.00	1	4.2	4.2	79.2
128.00	1	4.2	4.2	83.3
131.00	2	8.3	8.3	91.7
140.00	1	4.2	4.2	95.8
162.00	1	4.2	4.2	100.0
Total	24	100.0	100.0	

Grouped Relative Frequency Distribution

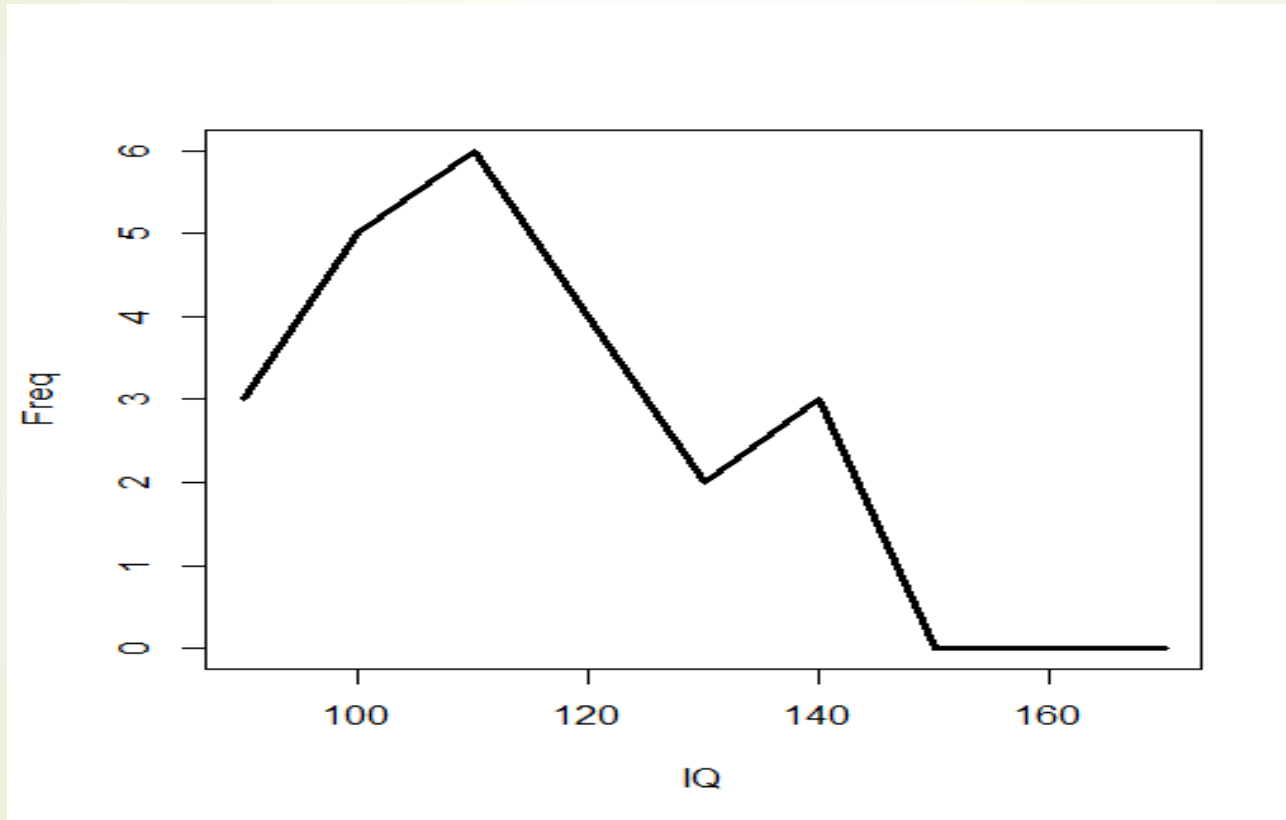
Relative Frequency Distribution of IQ for Two Classes

IQ	Frequency	Percent	Cumulative Percent
80 – 89	3	12.5	12.5
90 – 99	5	20.8	33.3
100 – 109	6	25.0	58.3
110 – 119	3	12.5	70.8
120 – 129	3	12.5	83.3
130 – 139	2	8.3	91.6
140 – 149	1	4.2	95.8
>= 150	1	4.2	100.0
Total	24	100.0	100.0

IQ Histogram



IQ Frequency Polygon





Descriptive Statistics

Summarizing Data:

- Measures of Central Tendency (or Groups' "Middle Values")
 - Mean
 - Median
 - Mode
- Measures of Dispersion / Variation (or Summary of Differences Within Groups)
 - Range
 - Interquartile Range
 - Variance
 - Standard Deviation

Mean

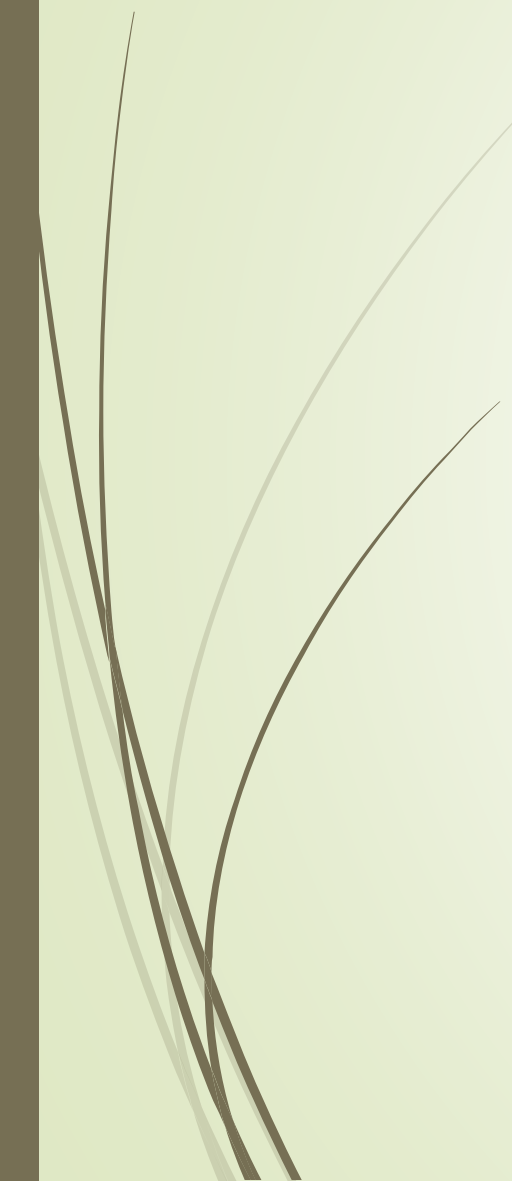
Let us consider that we have n observations, x_1, x_2, \dots, x_n as the outcome of an experiment. Then the mean of these observations is defined as $\bar{x} = (\sum_{i=1}^n x_i)/n$. Mean is also called as arithmetic mean associated with these observations.

Properties of Mean

- ▶ Sum of deviations of the observations from their mean is always zero.
- ▶ Sum of squares of deviations of observations is minimum when taken from their mean.
- ▶ If a constant is added to the observations, the mean of these observations also gets added by this constant.
- ▶ If every observation is multiplied by a constant, the mean of these observations also gets multiplied by this constant.



Mean



Mean is very widely used central tendency. However, it is very much affected by extreme observations and as such, this is not a very good representation for the data consisting of extreme values.



Median

Median of a distribution is the value that divides it into two equal parts. In terms of frequency curve, the ordinate drawn at median divides the area under the curve into two equal parts.

Let us again take n observations, x_1, x_2, \dots, x_n as the outcome of an experiment. In order to calculate the median, we first arrange these observations in either ascending or descending order. The median is given by the size of $[(n + 1)/2]^{\text{th}}$ observation if n is odd. The median is given by the mean of the sizes of $[n/2]^{\text{th}}$ and $[n/2 + 1]^{\text{th}}$ observations if n is even.



Mode

Mode is the observation that occurs maximum number of times in a given distribution and other observations are densely distributed around this number. One definition of mode tells us that “mode is the value which has the greatest frequency density in its immediate neighborhood”.

Measures of Dispersion

- Dispersion is the term associated with the variability in the data. This variability is measured in terms of the deviations from a central tendency. And a suitable average of these deviations is called the measure of dispersion.
- According to A.L. Bowley, “Dispersion is the measure of variation of the items” and according to Spiegel, “The degree to which numerical data tend to spread about an average value is called variation or dispersion of data”.
- Some important measures of dispersion are range, inter-quartile range, mean deviation and standard deviation. Out of these standard deviation is an important measure.

Range

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

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 = $162 - 80 = 82$

Inter-quartile range

- In descriptive statistics, the interquartile range (IQR), also called the midspread, middle 50%, or H-spread, is a measure of dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.
- The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. $IQR = Q3 - Q1$.

Interquartile Range

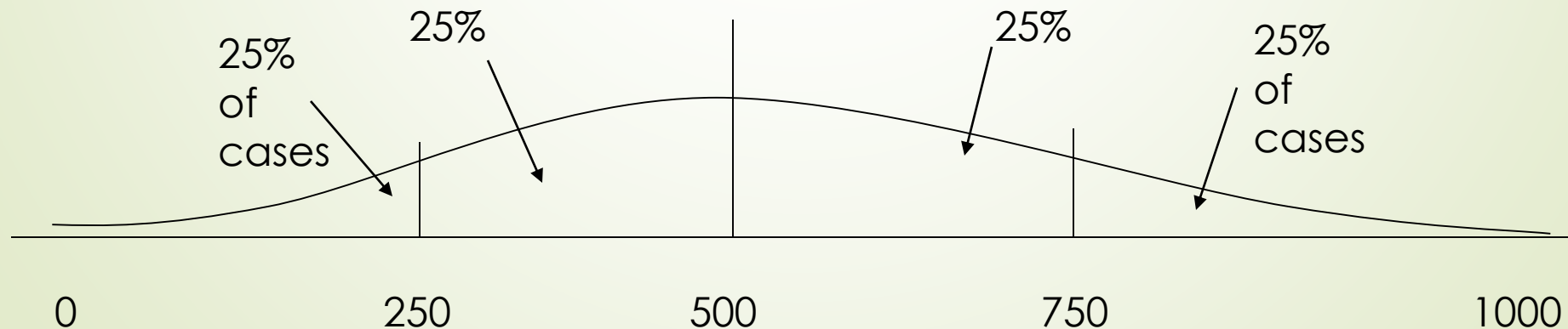
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

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

25th percentile is a quartile that divides the first $\frac{1}{4}$ of cases from the latter $\frac{3}{4}$.

75th percentile is a quartile that divides the first $\frac{3}{4}$ of cases from the latter $\frac{1}{4}$.

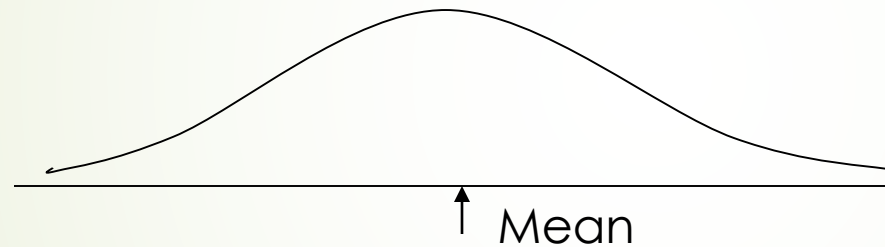
The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



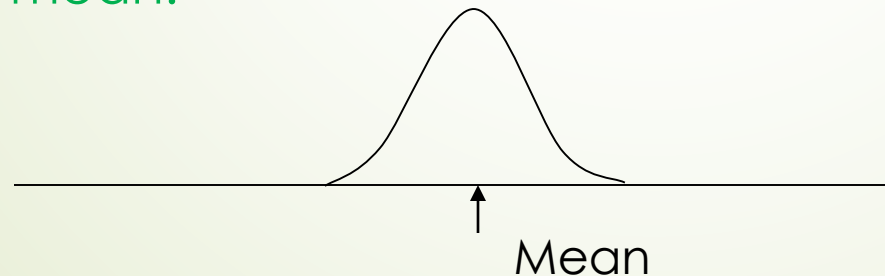
Variance

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

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



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



Variance

- $S^2 = \sum_{i=1}^n (x(i) - \bar{x})^2 / n$ #Variance in the data
- $S^2 = \sum_{i=1}^n (x(i) - \bar{x})^2 / (n - 1)$ #Sample Variance

Positive square root s of this quantity is called the standard deviation of the observations.

**Coefficient of Variation = Standard Deviation of data / mean of data
(sigma / mu)**

Descriptive Statistics: Example

Class A--IQs of 13 Students

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

Class A

Mean = 110.5385

Median = 109

Range = 140 – 89 = 51

Q1 = 98, Q2 = 109, Q3 = 119

IQR = 21

Variance = 240.9359

Standard Deviation = 15.52211

Coefficient of Variation = 0.14042

Class B--IQs of 13 Students

127	162	109
131	103	
96	111	
80	109	
93	87	
120	105	

Class B

Mean = 110.2308

Median = 109

Range = 162 – 80 = 82

Q1 = 96, Q2 = 109, Q3 = 120

IQR = 24

Variance = 460.359

Standard Deviation = 21.45598

Coefficient of Variation = 0.194646

Which class is smarter now?



Descriptive Statistics Ends

Thank You !