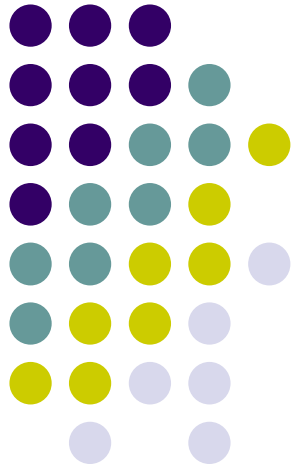
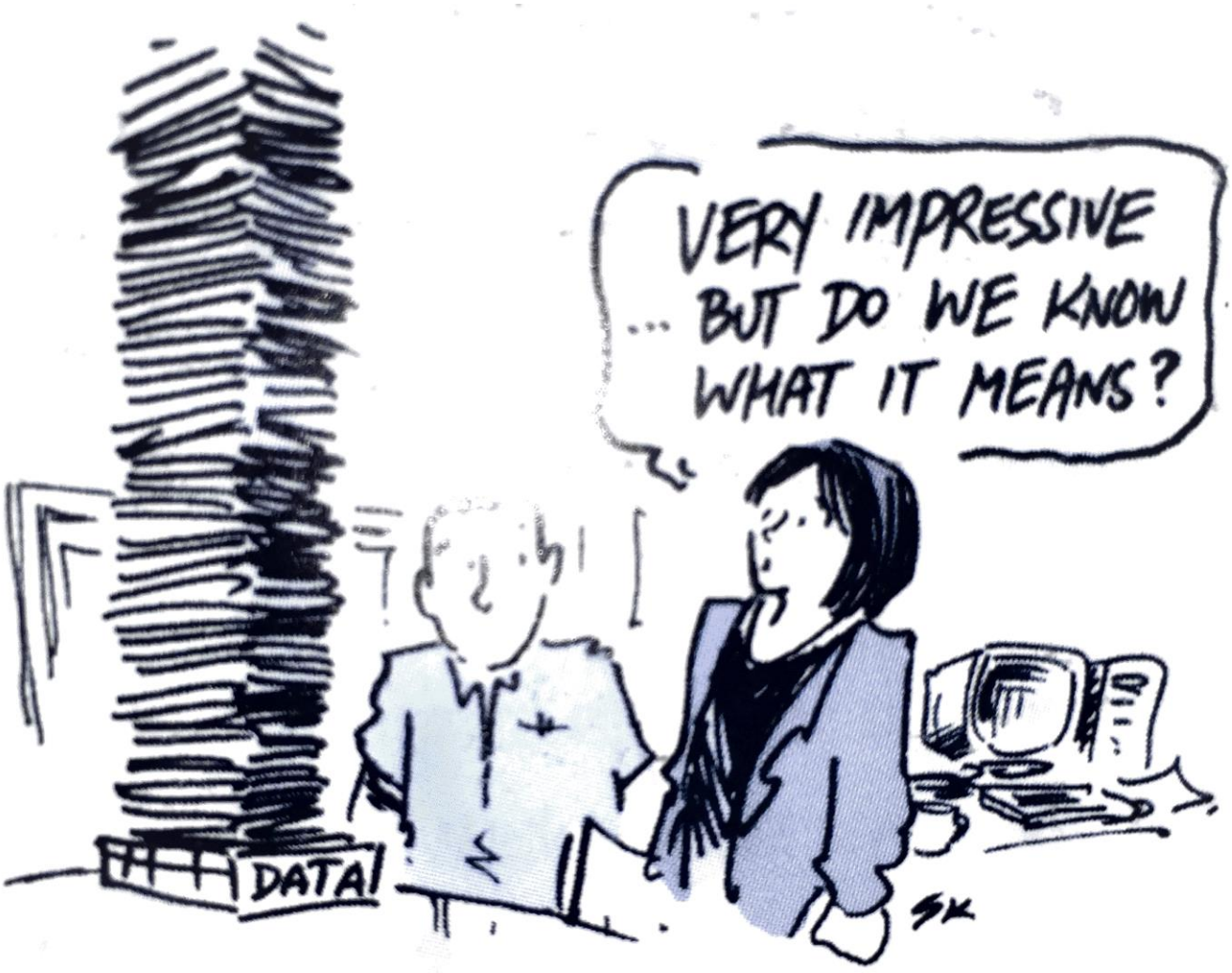
Descriptive Statistics

Dr. Yazdan Asgari



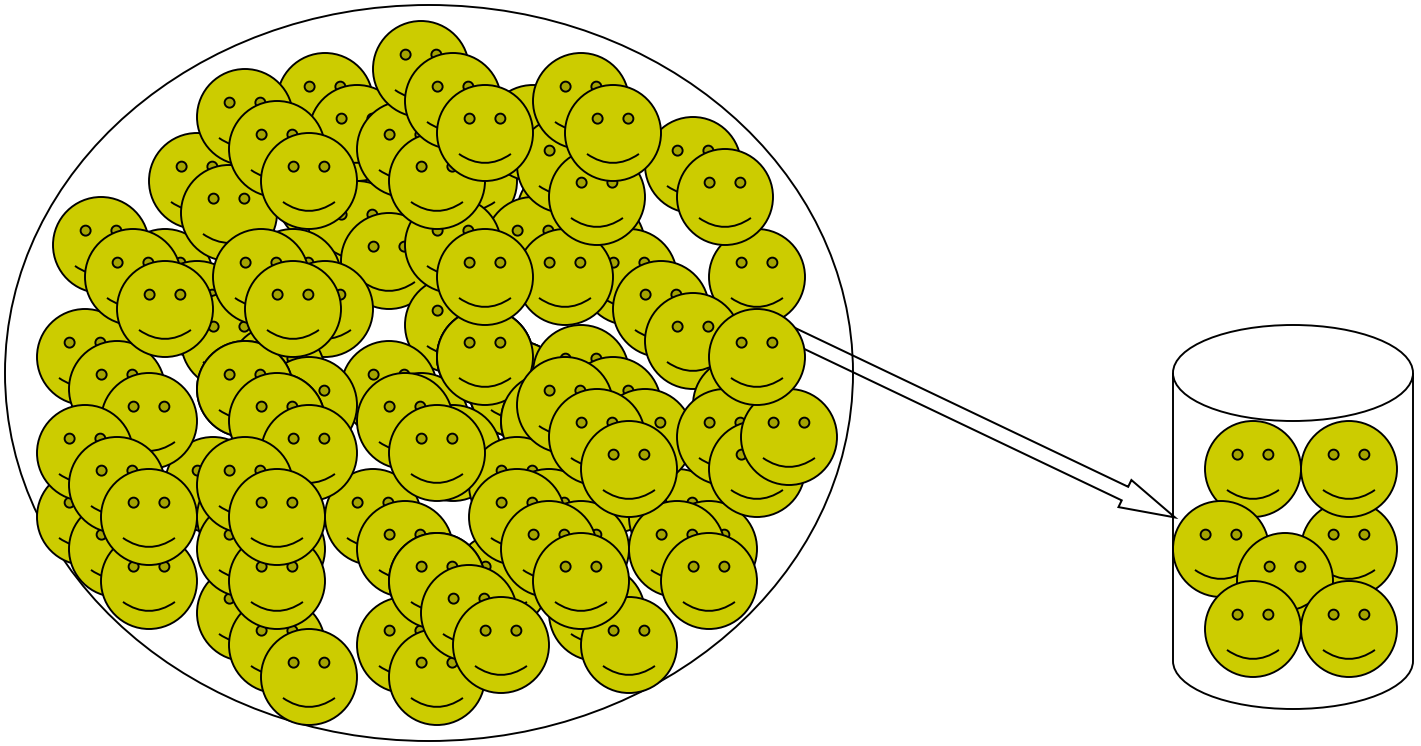
2019

Descriptive Statistics - Goal



https://images.google.com/

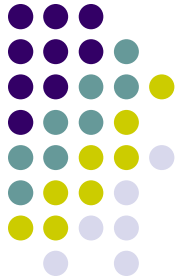
Sample vs. Population



Population

Sample

Frequency distributions



Example:

80 data of emission (in grams) of carbon monoxide from cars at a specified university in one month

26.4	17.3	11.2	23.9	24.8	18.7	13.9	9.0
13.2	22.7	9.8	6.2	14.7	17.5	26.1	12.8
28.6	17.6	23.7	26.8	22.7	18.0	20.5	11.0
20.9	15.5	19.4	16.7	10.7	19.1	15.2	22.9
26.6	20.4	21.4	19.2	21.6	16.9	19.0	18.5
23.0	24.6	20.1	16.2	18.0	7.7	13.5	23.5
14.5	14.4	29.6	19.4	17.0	20.8	24.3	22.5
24.6	18.4	18.1	8.3	21.9	12.3	22.3	13.3
11.8	19.3	20.0	25.7	31.8	25.9	10.5	15.9
27.5	18.1	17.9	9.4	24.1	20.1	28.5	15.8

Frequency distributions

- A **frequency distribution** is a tabular arrangement of data whereby the data is grouped into different intervals, and then the number of observations that belong to each interval is determined.
- Data that is presented in this manner are known as **grouped data**.

Frequency distributions

Class	Frequency
Less than 5	0
Less than 9	3
Less than 13	13
Less than 17	27
Less than 21	52
Less than 25	69
Less than 29	78
Less than 33	80

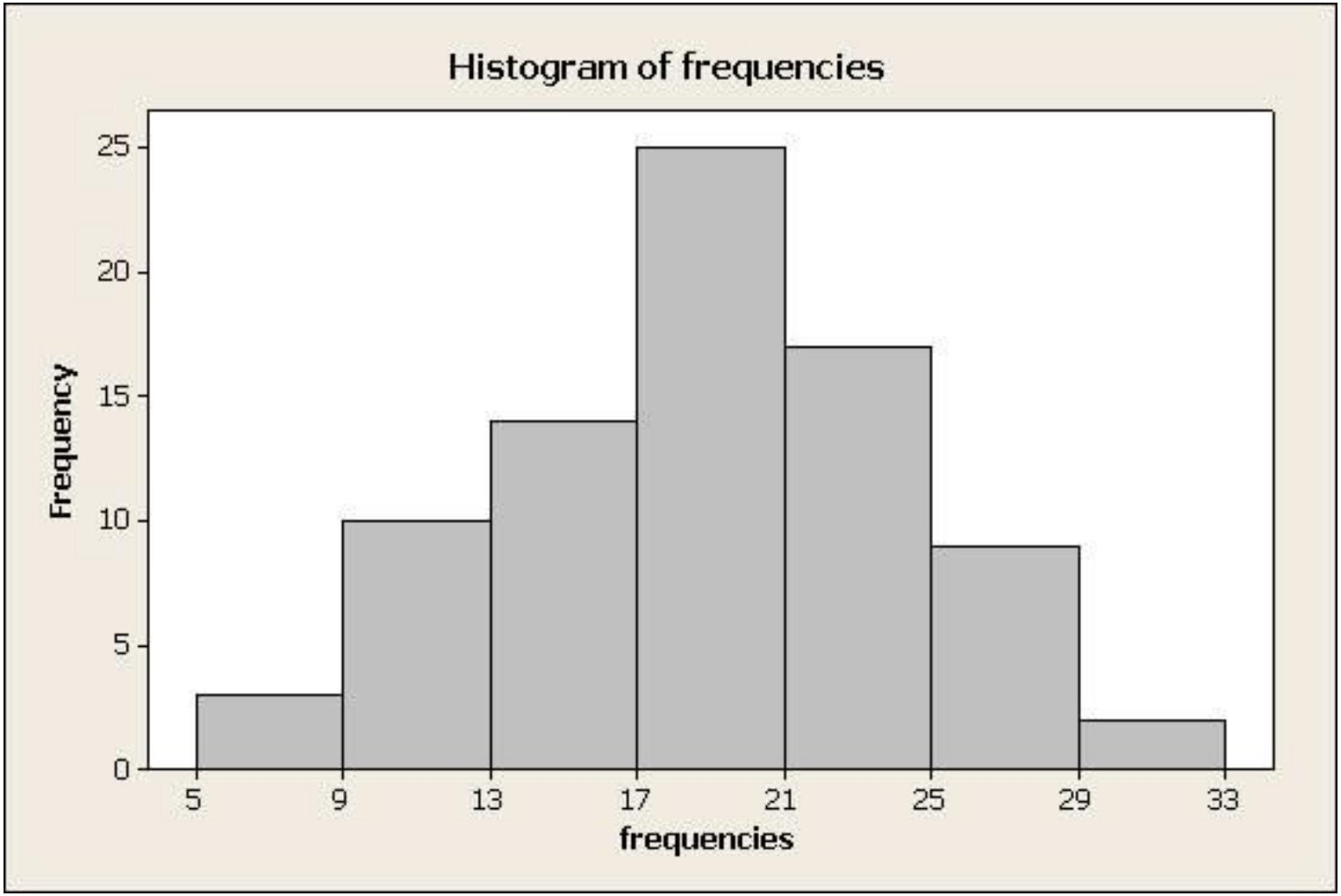
Frequency Table

Class	Frequency
5.0 -- 8.9	3
9.0 – 12.9	10
13.0 – 16.9	14
17.0 – 20.9	25
21.0 – 24.9	17
25.0 – 28.9	9
29.0 – 32.9	2
Total	80

Percentage Distribution

Class	Frequency	Perc. Dist.
[5.0, 9.0)	3	3.75%
[9.0, 13.0)	10	12.5%
[13.0, 17.0)	14	17.5%
[17.0, 21.0)	25	31.25%
[21.0, 25.0)	17	21.25%
[25.0, 29.0)	9	11.25%
[29.0, 33.0)	2	2.5%
Total	80	100%

Histogram


Histogram of frequencies
Frequency
25
20
15
10
5
0
5
9
13
17
21
25
29
33
frequencies

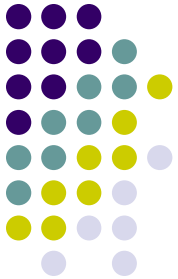
Descriptive Statistics

Example

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	

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

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs
- Summarize Data
  - Central Tendency
  - Variation

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

Class A--IQs of 13 Students

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

$\Sigma Y_i = 1437$

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

Class B--IQs of 13 Students

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

$\Sigma Y_i = 1433$

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

Median

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

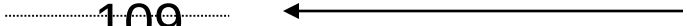
When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it.

The 50^{th} 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

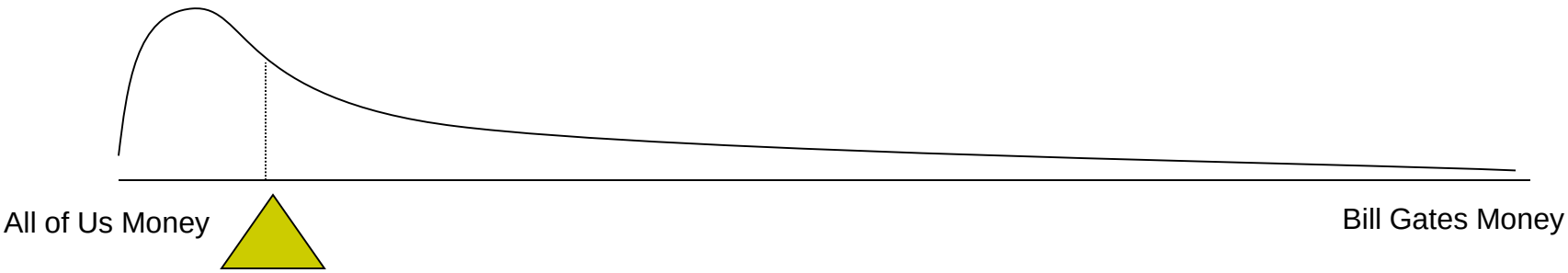
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

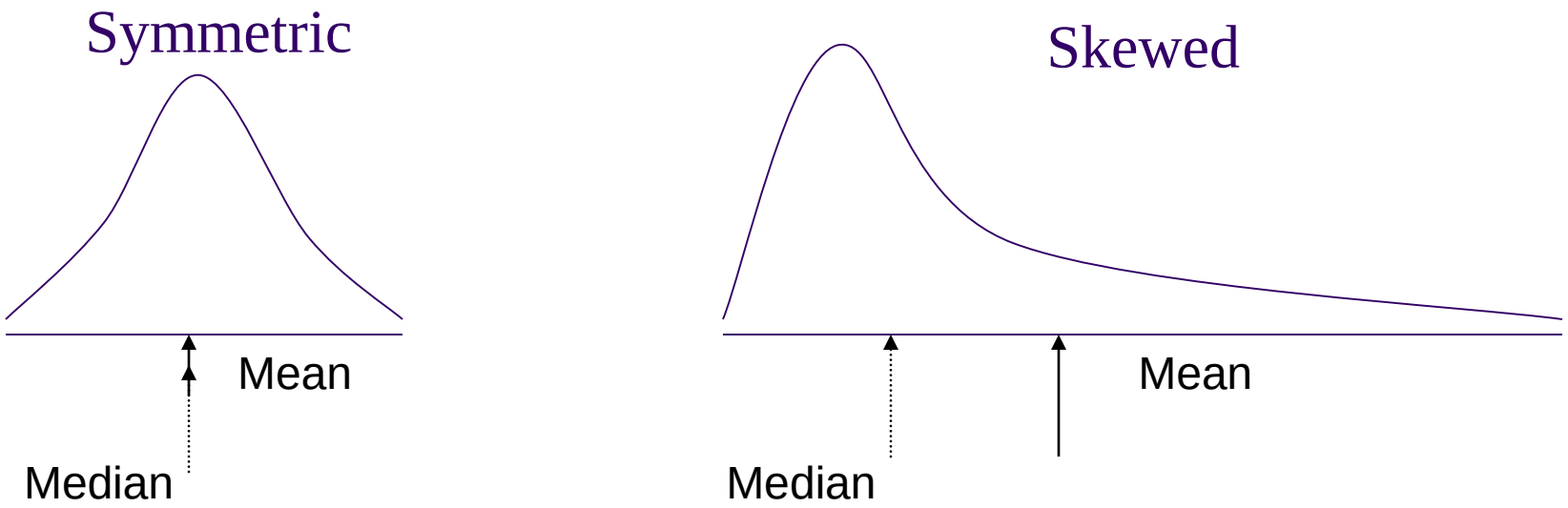
Median Properties

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 Properties

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.

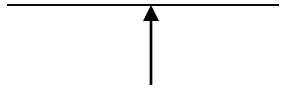


Mode

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

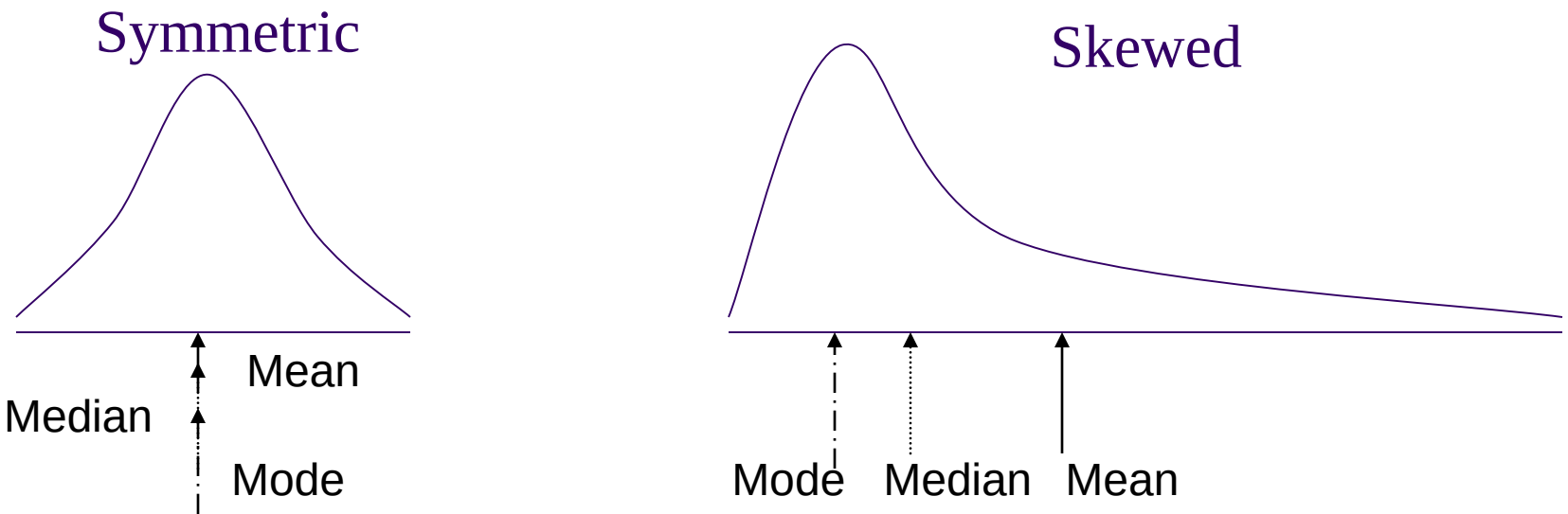


A la mode!!

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

Mode Properties

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.



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

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

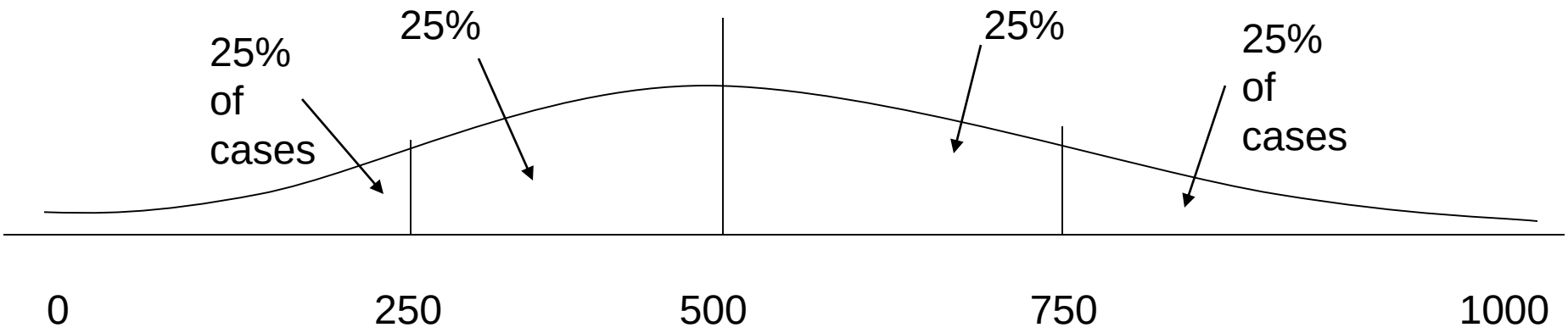
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

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.

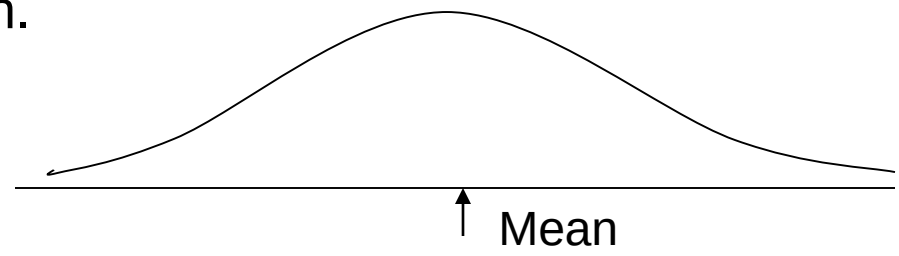
The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



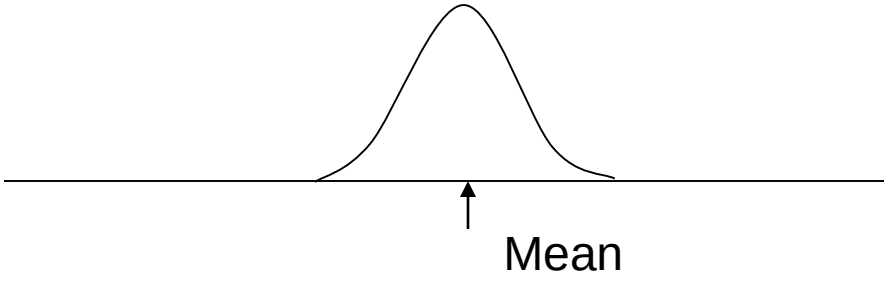
Variance

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.



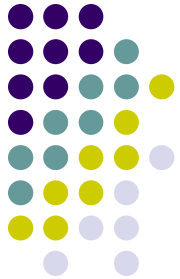
Variance

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	

$\bar{Y} = 110.54$

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	

$\bar{Y} = 110.54$

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: $(Y_i - \bar{Y})^2$

Back to the IQ example,

A deviation squared for 102 is: $(102 - 110.54)^2 = (-8.54)^2 = 72.93$

of 115: $(115 - 110.54)^2 = (4.46)^2 = 19.89$

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) = $\Sigma (Y_i - \bar{Y})^2$

$SS = (Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \ldots + (Y_n - \bar{Y})^2$

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	

$\bar{Y} = 110.54$

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

Variance

For Class A, Variance = 2825.39 / n - 1

= 2825.39 / 12 = 235.45

How helpful is that???

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.

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

Standard Deviation

For Class A, the standard deviation is:

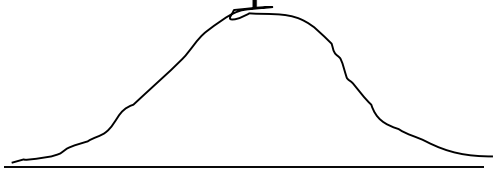
$$\sqrt{235.45} = 15.34$$

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

Standard Deviation

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

   For example:



19 25 31

$\bar{Y} = 25$

s.d. = 3

13 25 37

$\bar{Y} = 25$

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.

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
- …Wait! There's more

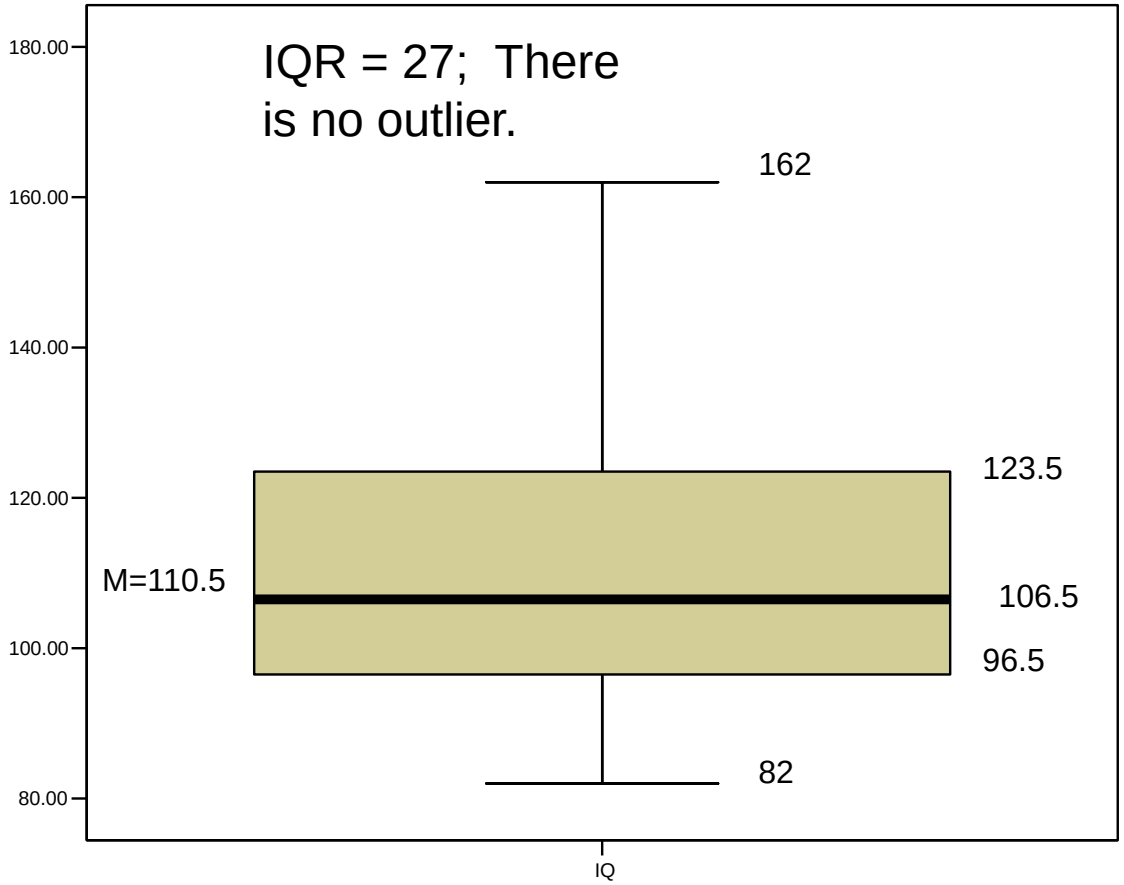
Box-Plots

A way to graphically portray almost all the descriptive statistics at once is the box-plot.

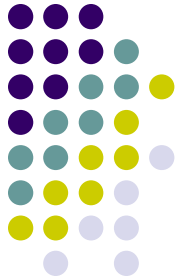
A box-plot shows:

- Upper and lower quartiles
- Mean
- Median
- Range
- Outliers (1.5 IQR)

Box-Plots


IQR = 27; There is no outlier.
180.00
160.00
140.00
120.00
100.00
80.00
162
123.5
M=110.5
106.5
96.5
82
IQ

IQV—Index of Qualitative Variation



- For nominal variables
- Statistic for determining the dispersion of cases across categories of a variable.
- Ranges from 0 (no dispersion or variety) to 1 (maximum dispersion or variety)
- 1 refers to even numbers of cases in all categories, NOT that cases are distributed like population proportions
- IQV is affected by the number of categories

IQV—Index of Qualitative Variation

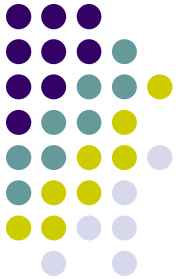
To calculate:

$$IQV = \frac{K(100^2 - \Sigma\ cat.\%^2)}{100^2(K - 1)}$$

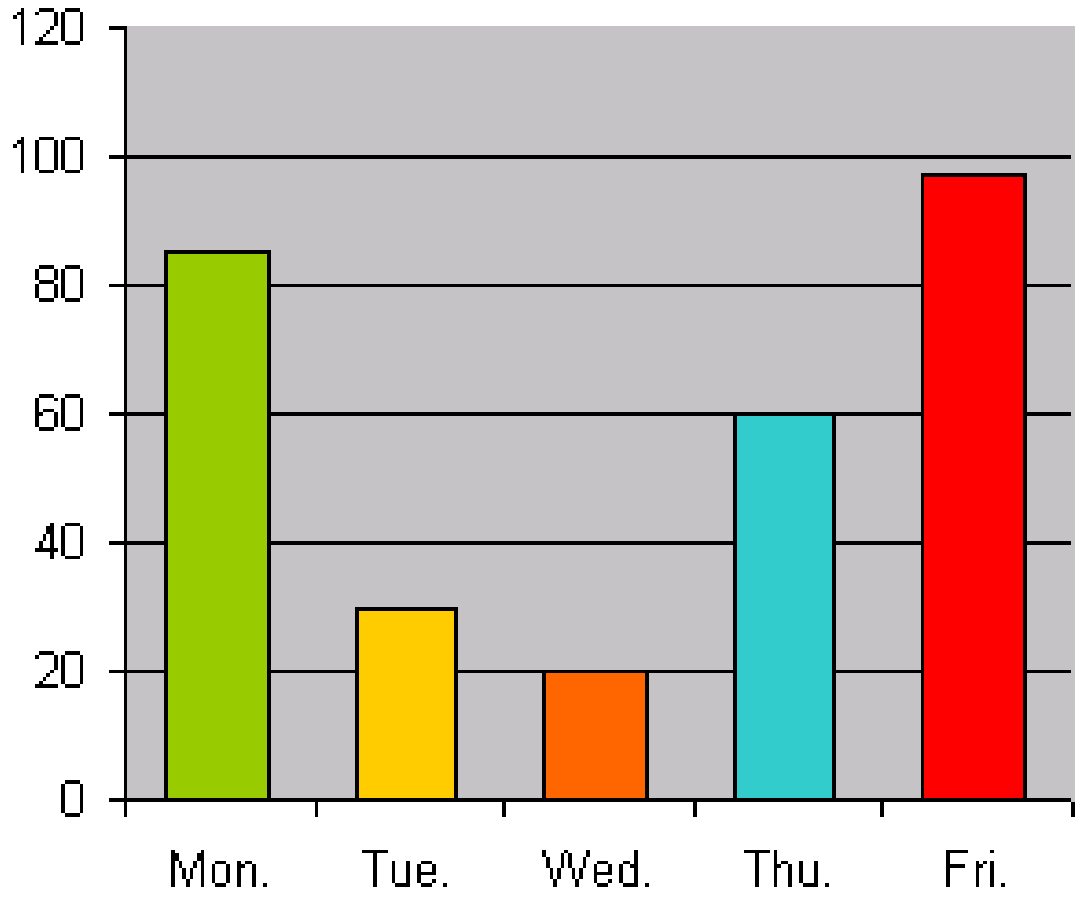
K=# of categories

Cat.% = percentage in each category

Some useful Graphs

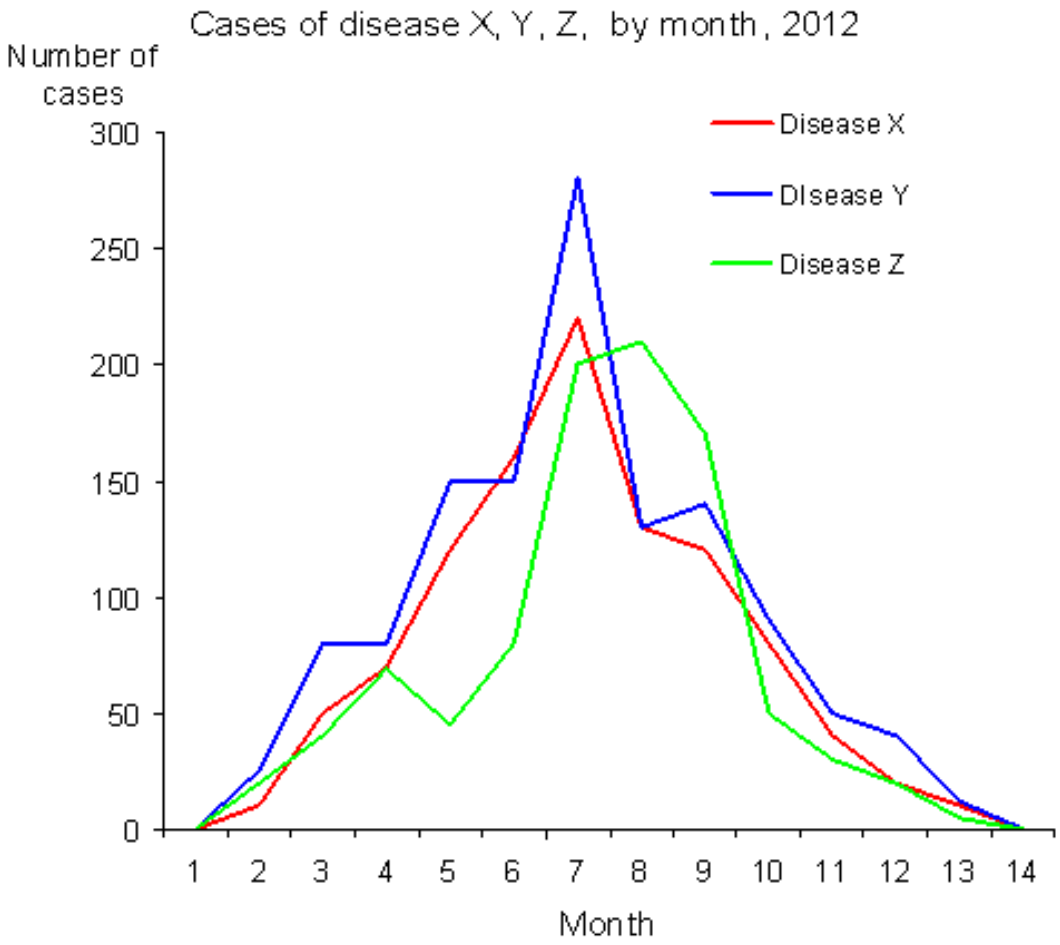


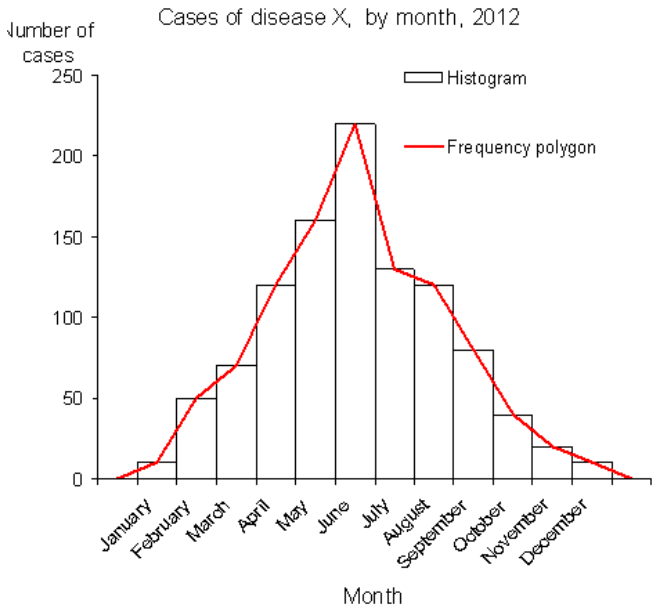
Bar Graphs



https://images.google.com/

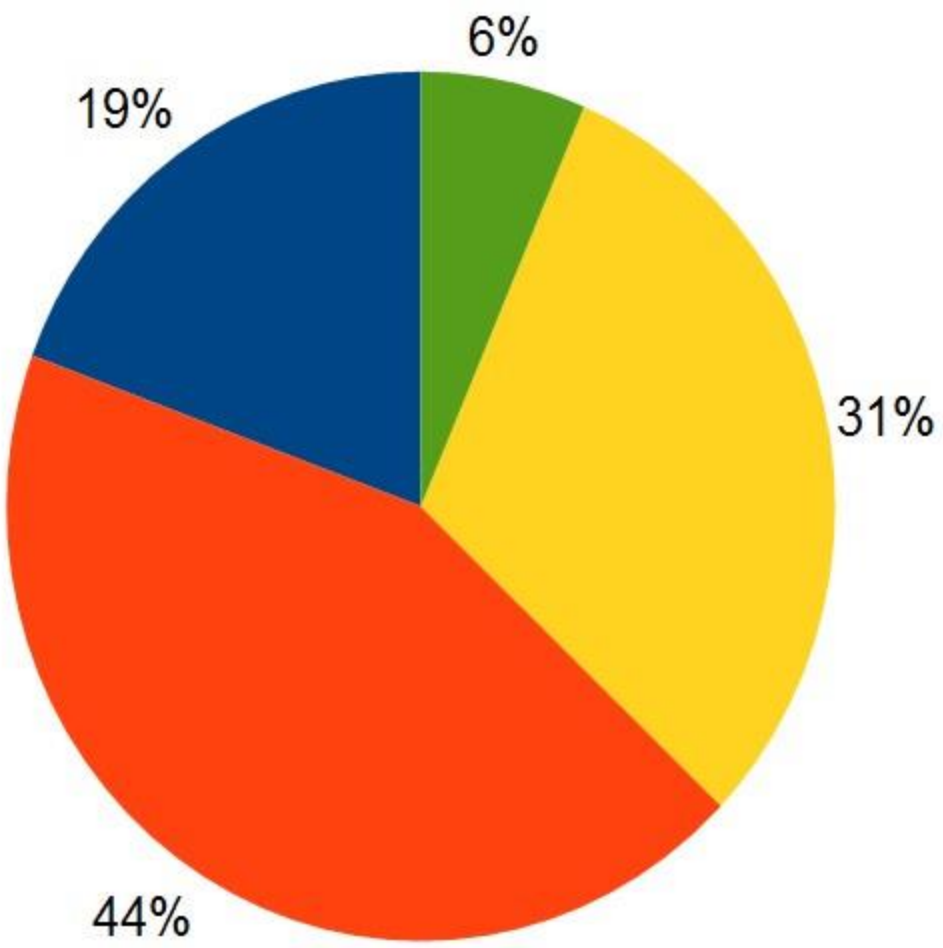
Polygons





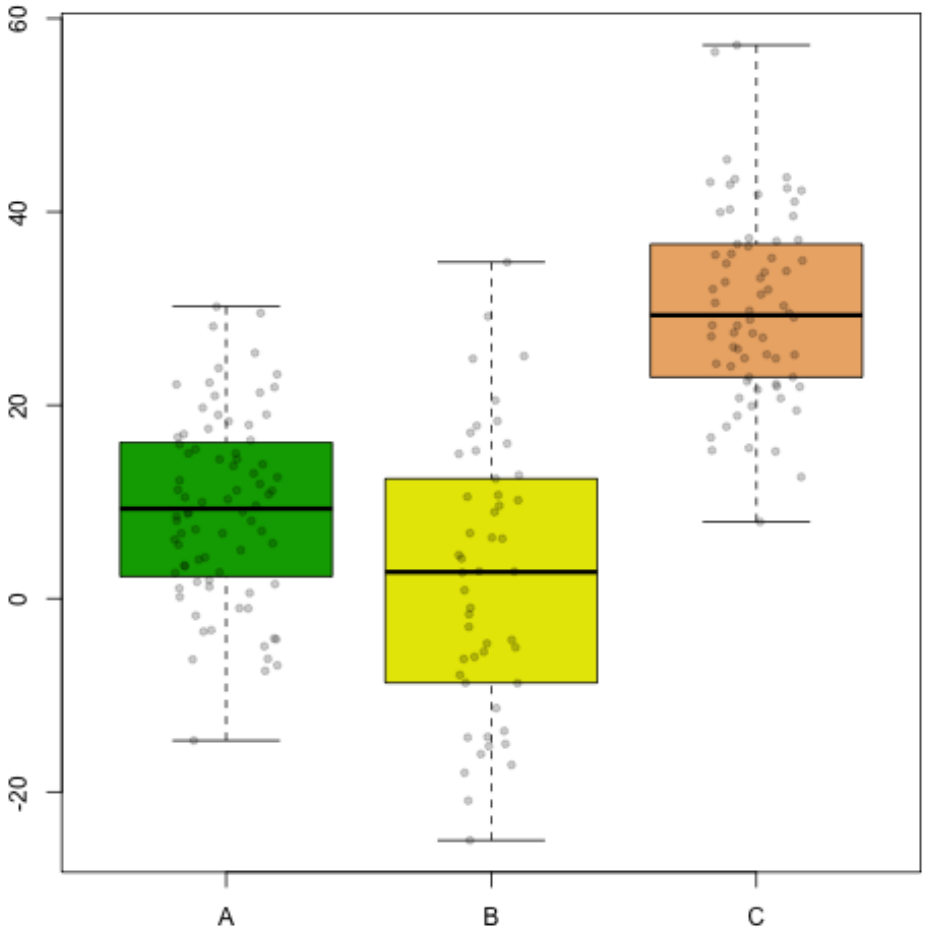
https://images.google.com/

Pie chart



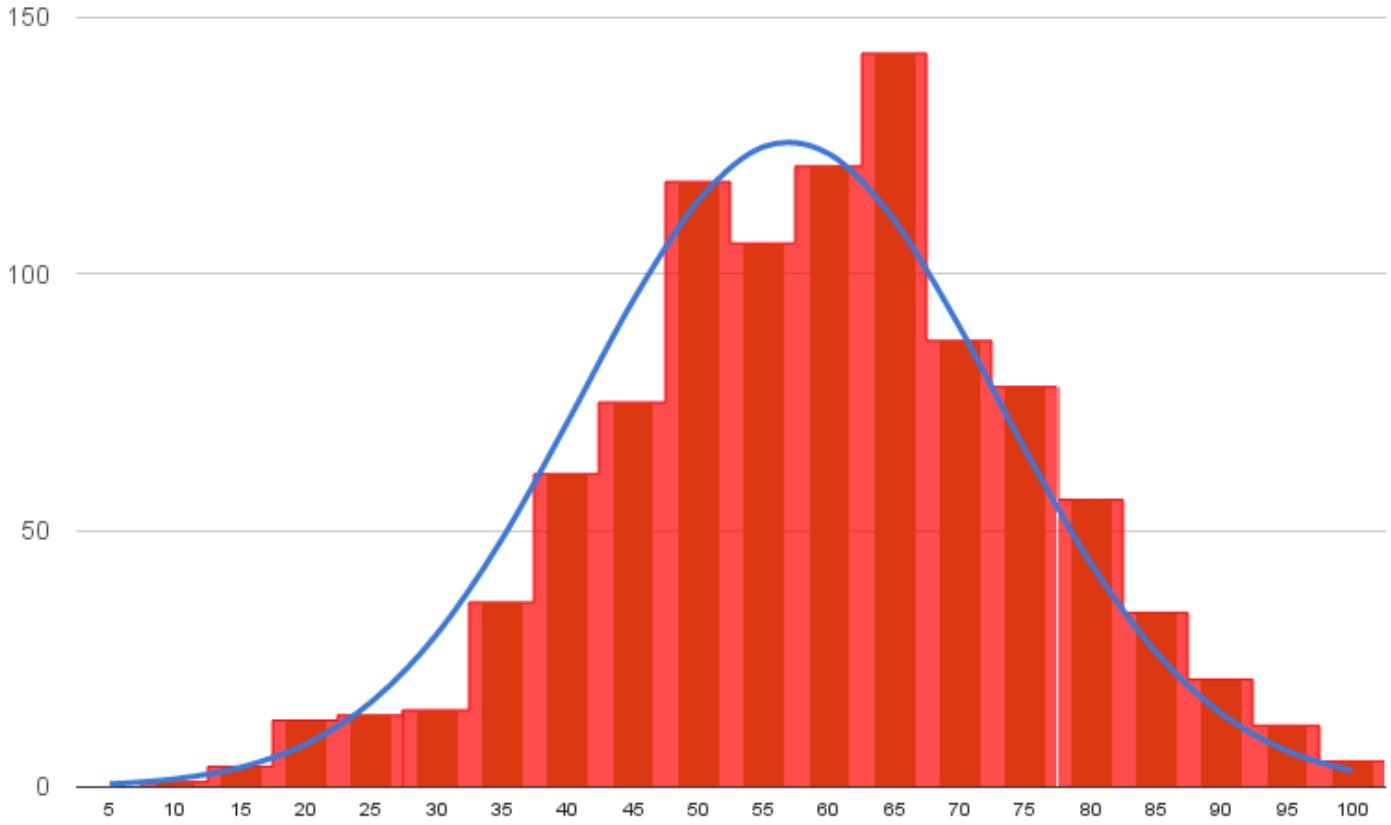
https://images.google.com/

Boxplots

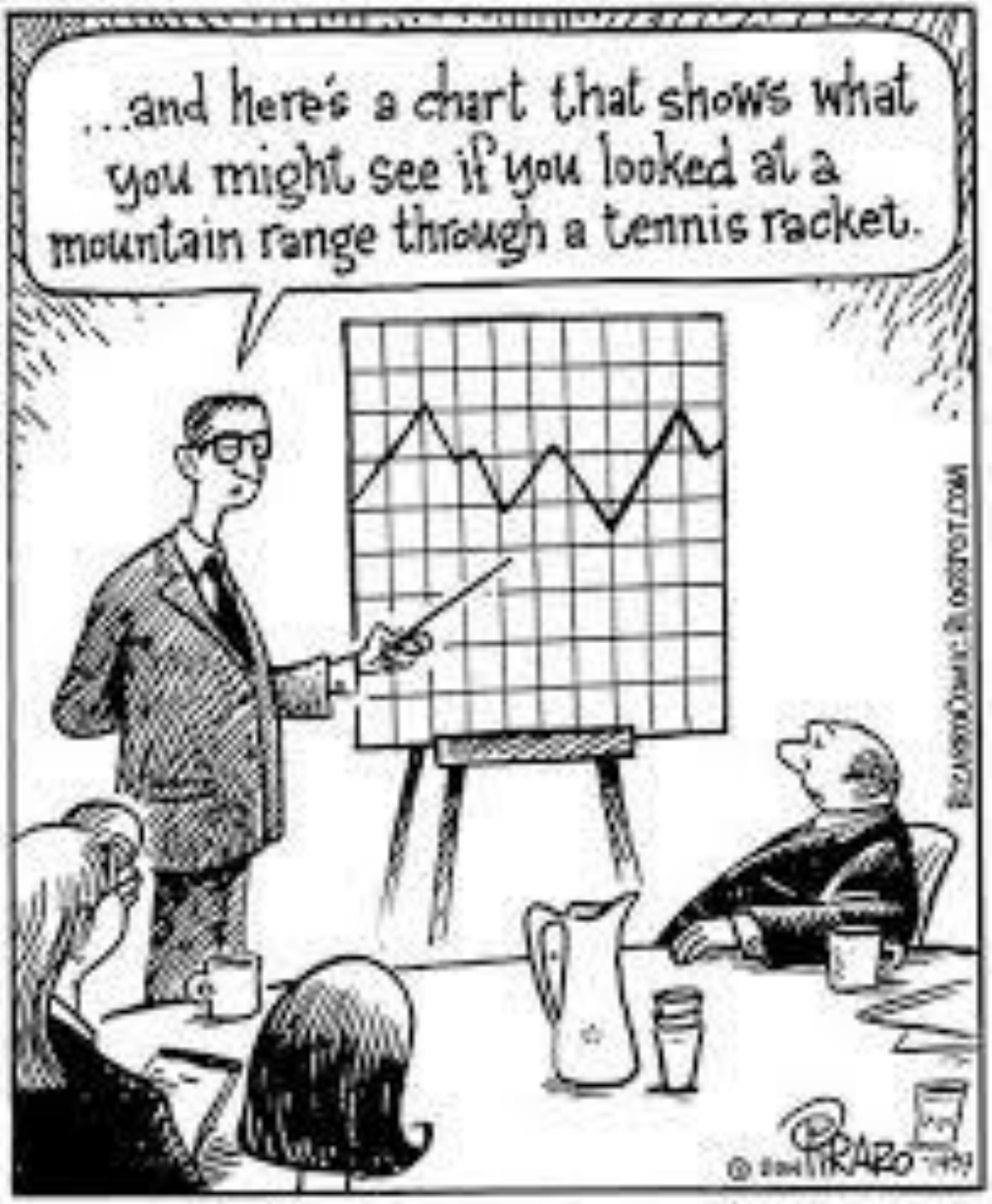


https://images.google.com/

Frequency Distribution curve



https://images.google.com/



Was it useful?