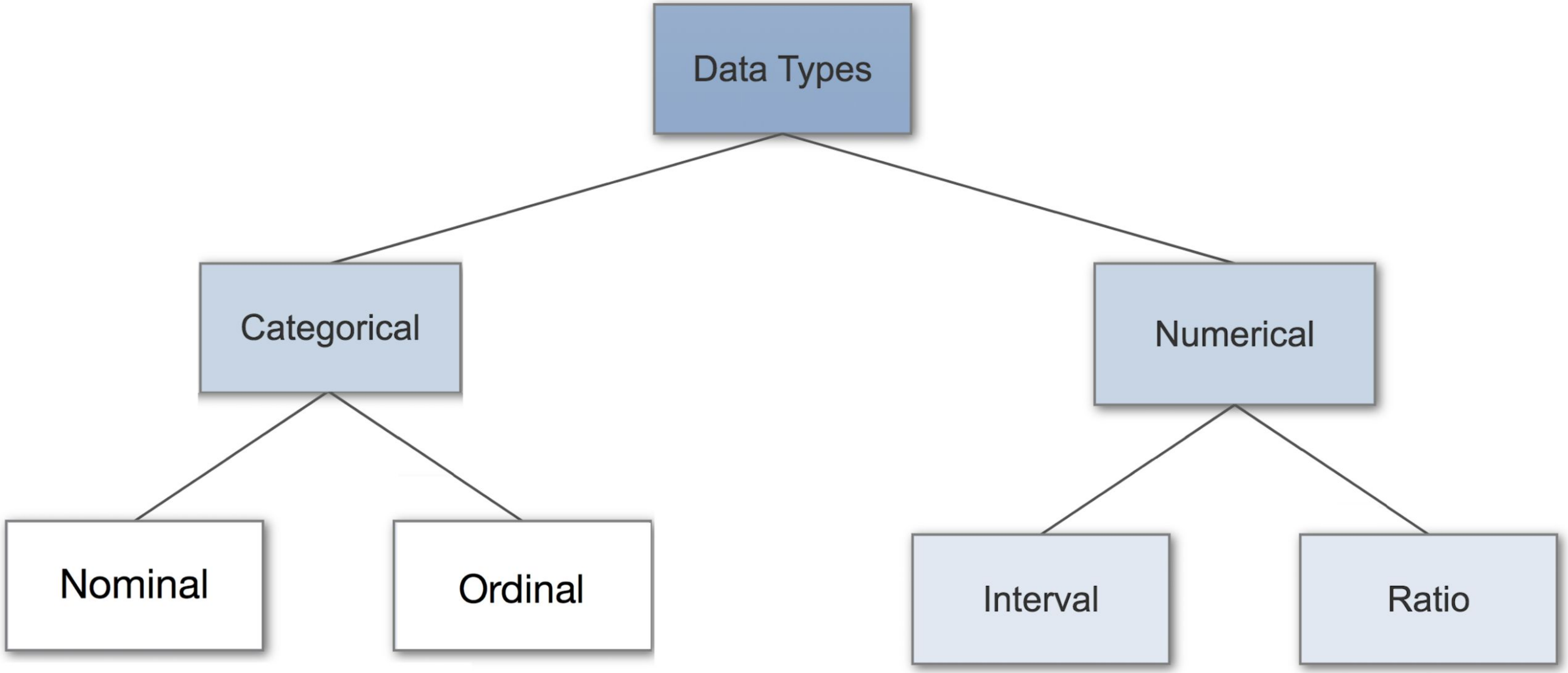
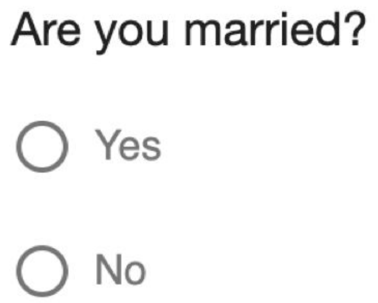
Statistics

Data Types.


Data Types
Categorical
Numerical
Nominal
Ordinal
Interval
Ratio

Nominal Data

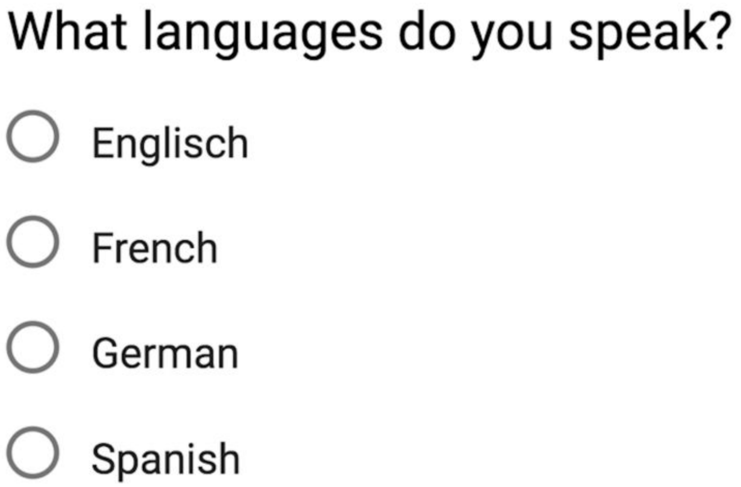
Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as „labels“. Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change.



Are you married?

○ Yes

○ No



What languages do you speak?

○ Englisch

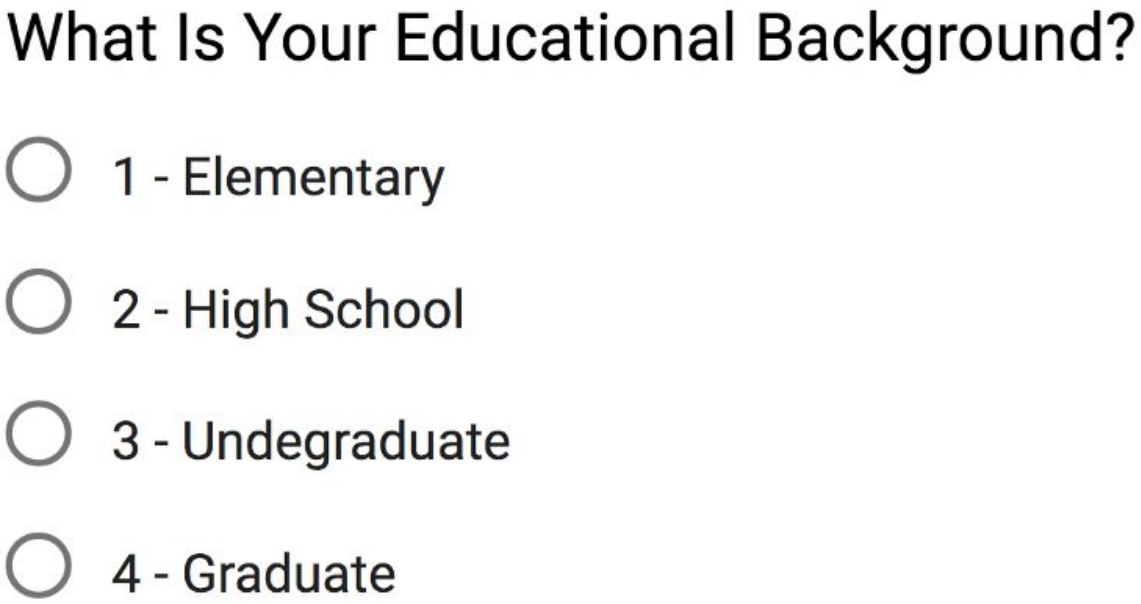
○ French

○ German

○ Spanish

Ordinal Data

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it’s ordering matters. You can see an example below:



What Is Your Educational Background?

○ 1 - Elementary

○ 2 - High School

○ 3 - Undegraduate

○ 4 - Graduate

Numerical Data

1. Discrete Data

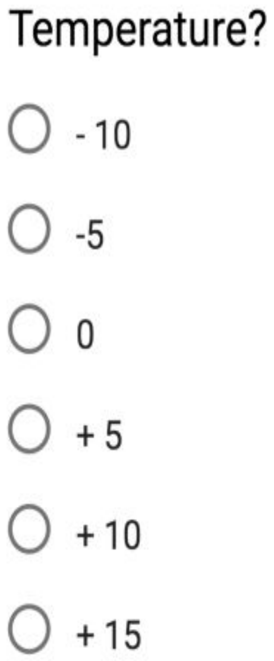
We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data **can’t be measured but it can be counted**.

2. Continuous Data

Continuous Data represents measurements and therefore their values **can’t be counted but they can be measured**. An example would be the height of a person, which you can describe by using intervals on the real number line.

Interval Data

Interval values represent **ordered units that have the same difference**. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values.



Temperature?

○ - 10

○ -5

○ 0

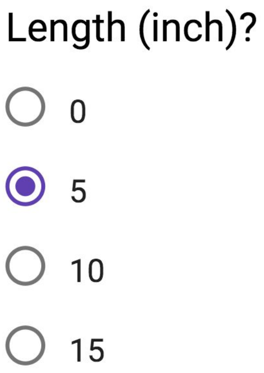
○ + 5

○ + 10

○ + 15

Ratio Data

Ratio values are also ordered units that have the same difference. Ratio values are **the same as interval values, with the difference that they do have an absolute zero**.



Length (inch)?

○ 0

● 5

○ 10

○ 15

Measure of Dispersion

- **Dispersion** is a way of describing how data is spread around an average value.
- data set have large differences between data values, then data set is said as widely scattered data set.
- data values are close to each other the data set is said to be tightly clustered data set.
- E.g.
  1. 55,57,55,58,56
  2. 10,22,31,45,27

Range

- The difference between largest and smallest data value in given data set is known as Range of given data set.
- For example, if we have data set as1,3,4,2,7,8,12,6.

  Range = 12–1 = 11.

Mean Deviation

- Average of absolute differences from the mean is known as mean deviation.
  1. Calculate the mean for given data set.
  2. Find out the absolute difference of each data value from mean
  3. Add all the values you calculated in second step and divide by total number of data values

$$M.D. = \frac{\sum_{i=1}^{N} |x_i - \bar{x} |}{N}$$

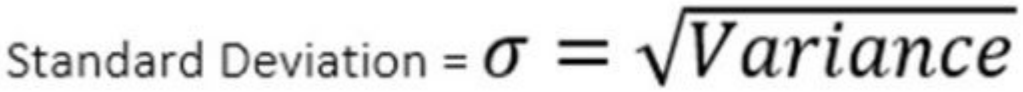
Variance

- The Variance is the average of squared differences from the mean.
  1. Calculate the mean of data values
  2. Find out the absolute difference of each data value from mean
  3. Square each absolute difference you find
  4. Take the mean of all squared values from step 3.

$$Variance = \sigma^2 = \frac{\sum_{i=1}^{N} |x_i - \bar{x}\ |^2}{N}$$

Standard Deviation

- The Standard Deviation is the square root of variance.



Standard Deviation = $\sigma = \sqrt{Variance}$

Measures of Central Tendency

- measures of central tendency are a set of “middle” values representative of the data points.
- Central tendency describes the distribution of data focusing on the central location around which all other data are clustered.
- It is the opposite of **dispersion** that measures how far the observations are scattered with respect to the central value.

Mean

- Mean is the average of some data points.

$$\boxed{\overline{X} = \frac{\sum X}{N}}$$

Median

- Median is the number at the center of a series *after* they are ordered (ascending or descending).

- 2,3,4,5,6

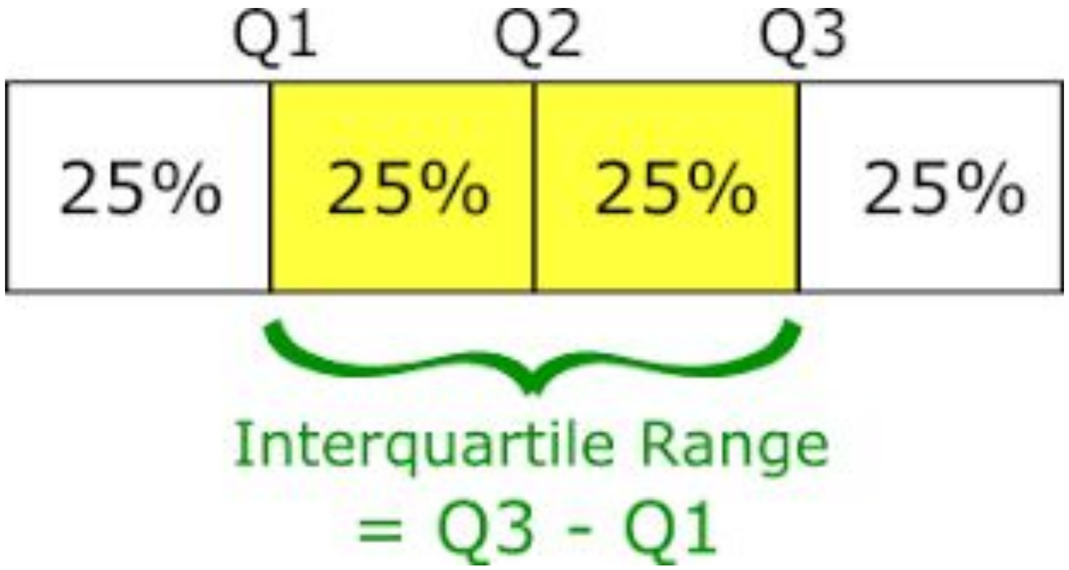
- 4, 7, 6, 2, 10, 8

Mode

- [1,2, 3, 4, 5,5] — the most frequently occurring one is 5; that’s **mode**.

1. A distribution can have more than one mode as in the list [2, 2, 3, 4, 4]; it’s called **bimodal** distribution of a discrete variable.
2. Along this logic, a distribution with more than two modes are called **multimodal** distribution.

- **Quartiles** — Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.
- 25% of the data points lie below Q1 and 75% lie above it.
- 50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.
- 75% of the data points lie below Q3 and 25% lie above it.



Skewness

- Skewness is the measure of symmetry or asymmetry of data distribution.
- A distribution or data set is said to be symmetric if it looks the same to the left and right points of the center.
- Type
  1. **Right skewness or Positive skewness**
  2. **Left skewness or Negative skewness**

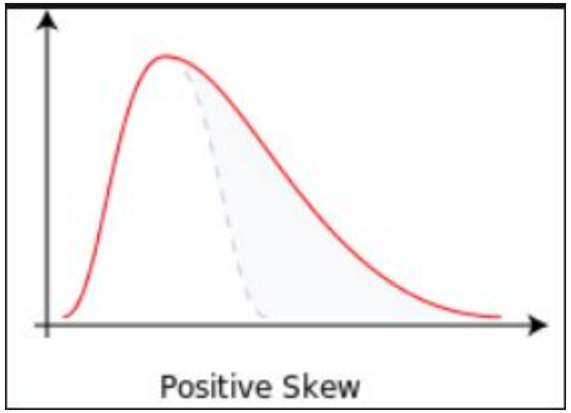
- Positive Skew — This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, mean is greater than the mode.
- Negative Skew — This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, mean is smaller than the mode.

- The most commonly used method of calculating Skewness is

$$Skewness = \frac{3\,(Mean - Median)}{Std \quad Deviation}$$

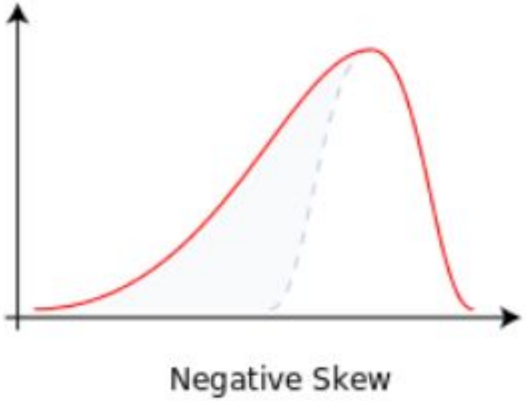
Right skewness

- A right-skewed distribution will have a long tail in the right direction on the number line such that the mean of the total value of all data points will eventually go up.

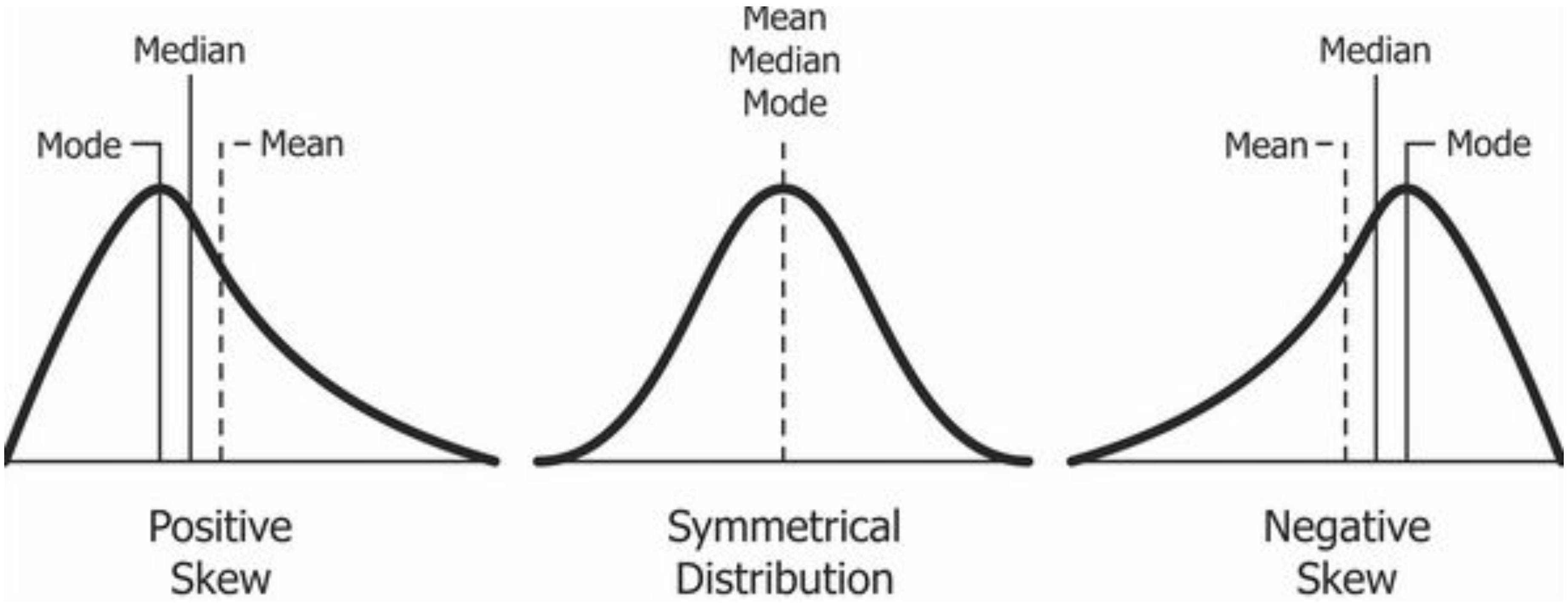


Left skewness

- A left-skewed distribution will have a long tail in the left direction on the number line such that the mean of the total intrinsic value of all data points will eventually go down.



- If the skewness is zero, the distribution is symmetrical. If it is negative, the distribution is Negatively Skewed and if it is positive, it is Positively Skewed.



Kurtosis

- Kurtosis is the characteristic of being flat or peaked. It is a measure of whether data is heavy-tailed or light-tailed in a normal distribution
- Percentile coefficient of Kurtosis
- **Ku=Q / (P90 — P10)**
- Where,
- **Q= Quartile deviation**
- **P90=90th percentile**
- **P10=10th percentile**

- A large value of kurtosis is often considered as riskier because data may tend to give an outlier value as an outcome with greater distance from the mean if applied to any machine learning algorithm.

Types of Kurtosis

1. Mesokurtic
2. Leptokurtic
3. Platykurtic

Mesokurtic

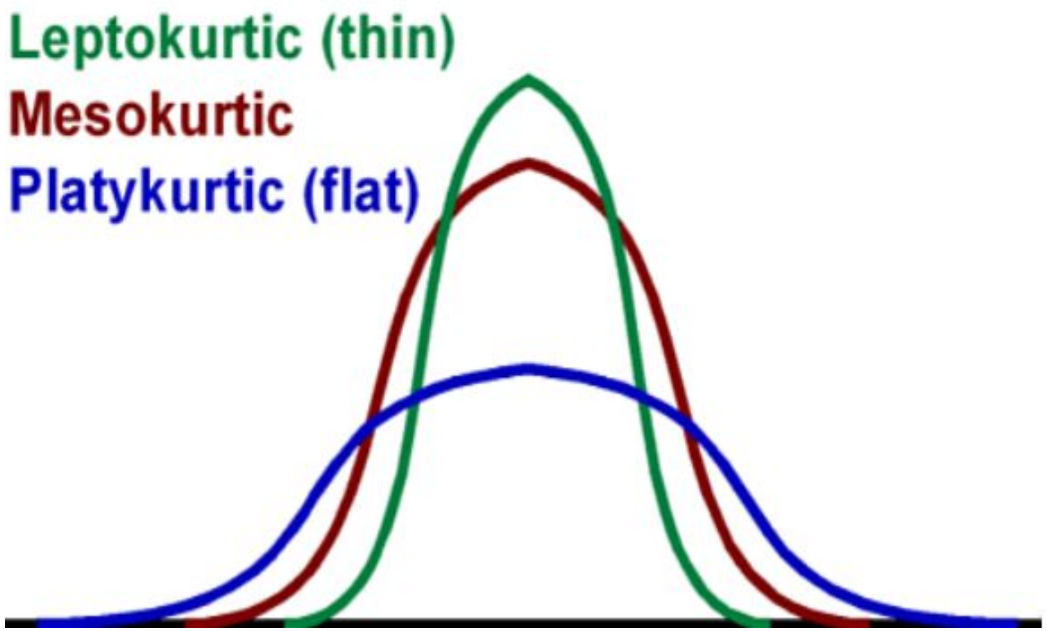
This distribution has the tails often similar to normal distribution.

Leptokurtic

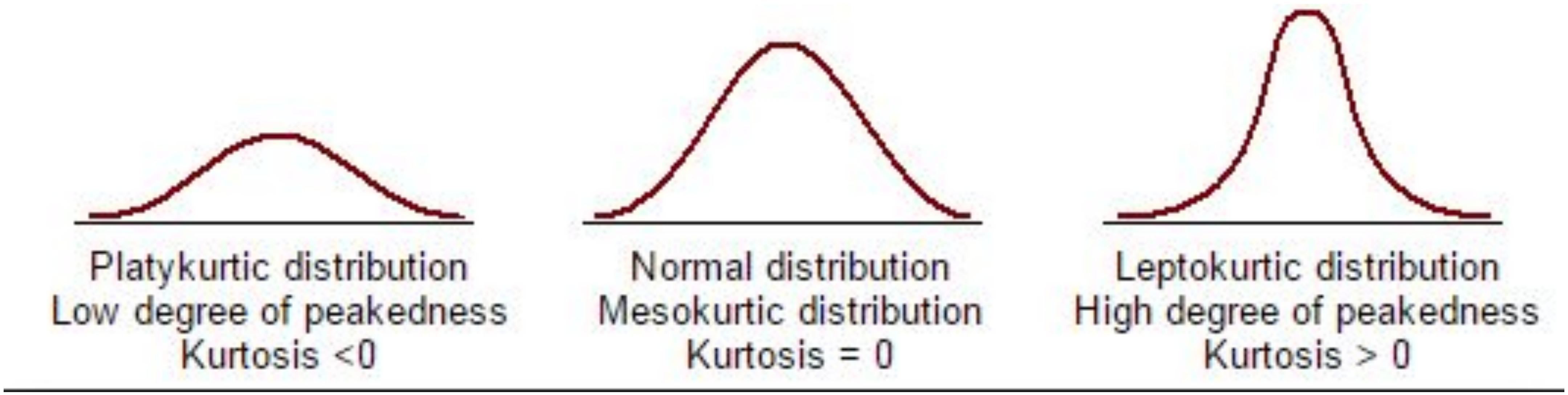
This distribution will be having very long and skinny tails. This means there are more chances of the presence of outliers.

Platykurtic

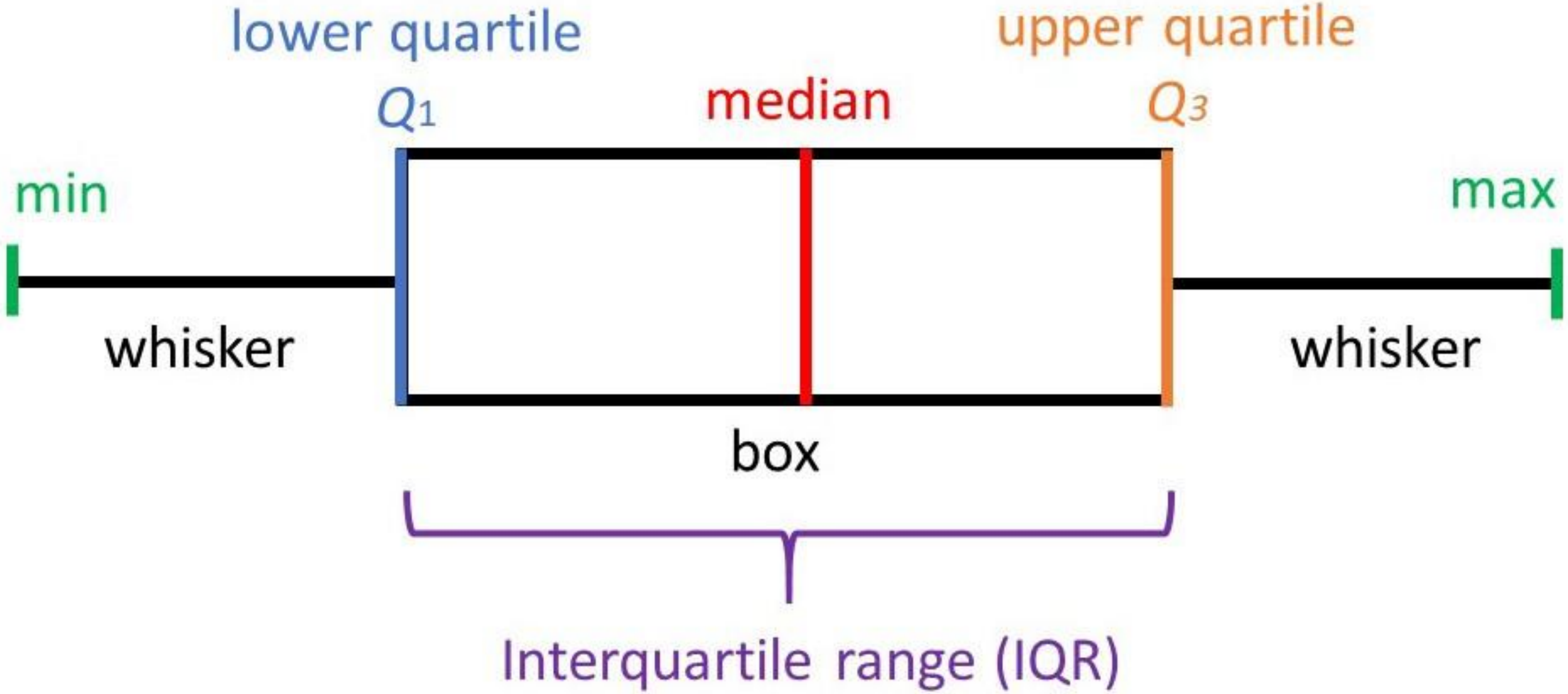
This distribution will be having very low and stretched around center tails which means most of the data points are present in high proximity with mean.



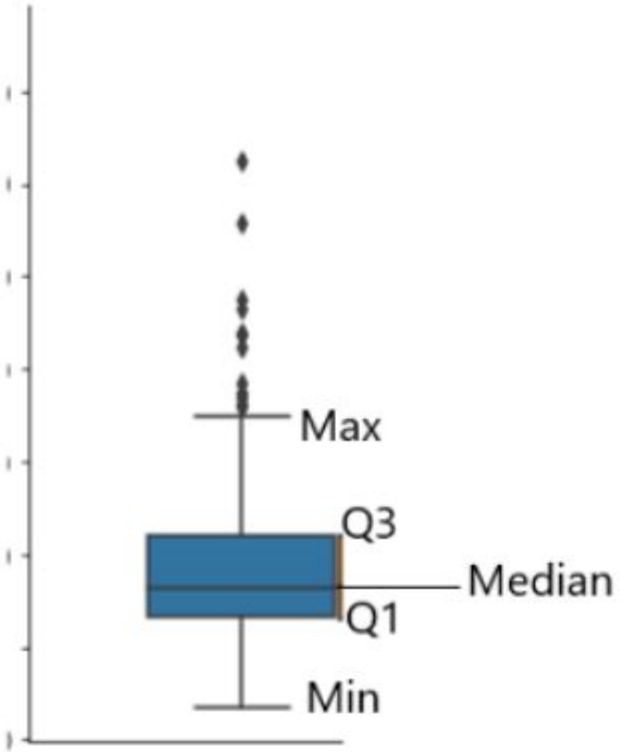
- Mesokurtic — This is the case when the kurtosis is zero, similar to the normal distributions.
- Leptokurtic — This is when the tail of the distribution is heavy (outlier present) and kurtosis is higher than that of the normal distribution.
- Platykurtic — This is when the tail of the distribution is light(no outlier) and kurtosis is lesser than that of the normal distribution.

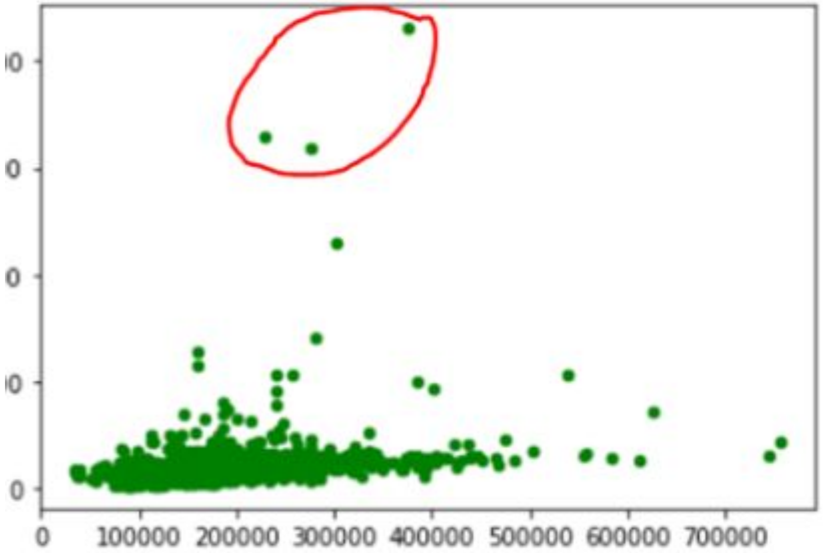


Detecting outliers using Box-And-Whisker Diagrams and IQR



- Minimum Value: It is the least numerical value of a dataset. This defines the left boundary.
- Median (Q2): It is the number that divides the data into two halves. If the number of numbers in the data is ODD, the median number is the middle data value. If the number of numbers in the data is EVEN, the median number is the mean of the two middle data values. This is only applicable if the numbers in the dataset are arranged in increasing/ascending order.
- Lower Half: It is the set of values that lie to the left of the Median (Q2)
- Upper Half: It is the set of values that lie to the right of the Median (Q2)
- Q1 (25th Percentile): It is the median of the lower half of the dataset.
- Q3 (75th Percentile): It is the median of the upper half of the dataset.
- Maximum Value: It is the greatest numerical value of a dataset. This defines the right boundary.
- Range: It is the difference between the maximum and the minimum value.
- Interquartile Range (IQR): It is the difference between Q3 and Q1.


Max
Q3
Median
Q1
Min


0
100000
200000
300000
400000
500000
600000
700000

Probability :*How **likely** something is to happen.*

Tossing a Coin

When a coin is tossed, there are two possible outcomes:

- heads (H) or
- tails (T)

We say that the probability of the coin landing **H** is ½

And the probability of the coin landing **T** is ½

When a single die is thrown, there are six possible outcomes: **1, 2, 3, 4, 5, 6**.

The probability of any one of them is ***1/6***

Probability of an event happening = *Number of ways it can happen*

Total number of outcomes

Experiment: a repeatable procedure with a set of possible results.

Example: Throwing dice

We can throw the dice again and again, so it is repeatable.

The set of possible results from any single throw is {1, 2, 3, 4, 5, 6}

Sample Space: all the possible outcomes of an experiment.

Example: choosing a card from a deck

There are 52 cards in a deck (not including Jokers)

So the **Sample Space is all 52 possible cards**: {Ace of Hearts, 2 of Hearts, etc... }

Outcome: A possible result of an experiment.

Example: Getting a "6"

Sample Point: just one of the possible outcomes

Example: Deck of Cards

the 5 of Clubs is a sample point

- the King of Hearts is a sample point

"King" is not a sample point. There are 4 Kings, so that is 4 *different* sample points.

Event: one **or more** outcomes of an experiment

Example Events:

An event can be just one outcome:

- Getting a Tail when tossing a coin
- Rolling a "5"

An event can include more than one outcome:

- Choosing a "King" from a deck of cards (any of the 4 Kings)
- Rolling an "even number" (2, 4 or 6)

- **Dependent Events** where what happens **depends on** what happened before, such as taking cards from a deck makes less cards each time.

- **Independent Events** which we learn about here.

Independent Events

Independent Events are **not affected** by previous events.

A coin does not "know" it came up heads before.
And each toss of a coin is a perfect isolated thing.

"P" to mean "Probability Of", So, for Independent Events:

P(A and B) = P(A) × P(B)

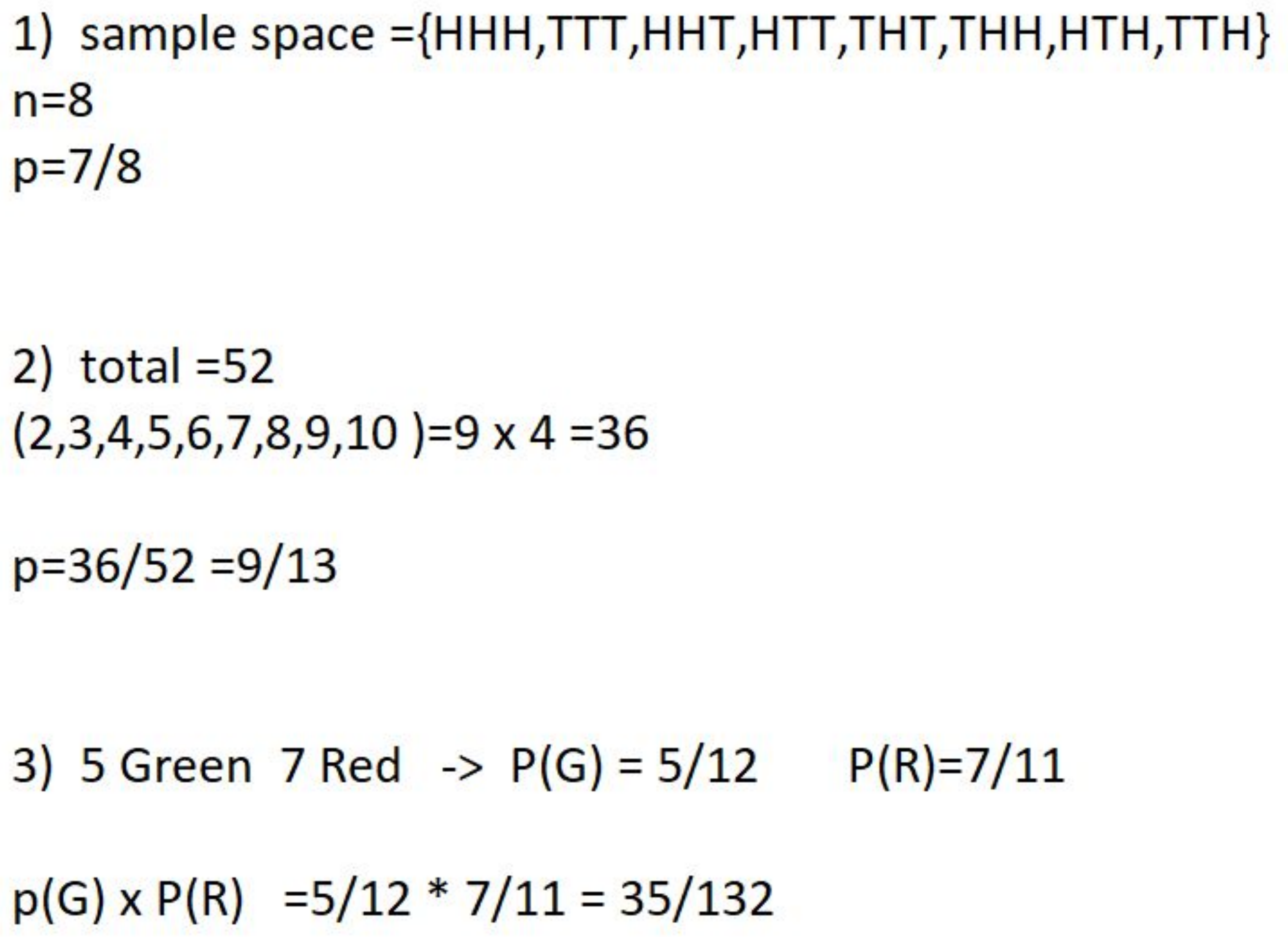
Probability of A and B equals the probability of A times the probability of B

- **Example 1:** A coin is thrown 3 times .what is the probability that atleast one head is obtained?

- **Example 2:** Find the probability of getting a numbered card when a card is drawn from the pack of 52 cards.

- **Example 3:** There are 5 green 7 red balls. Two balls are selected one by one without replacement. Find the probability that first is green and second is red.

Answers:



1) sample space ={HHH,TTT,HHT,HTT,THT,THH,HTH,TTH}

n=8

p=7/8

2) total =52

(2,3,4,5,6,7,8,9,10)=9 x 4 =36

p=36/52 =9/13

3) 5 Green 7 Red -> P(G) = 5/12 P(R)=7/11

p(G) x P(R) =5/12 * 7/11 = 35/132