**Probability Notes**

Probability:

If the trail is repeated number of times under homogenous and identical condition then the value of ratio of number of favourable cases to the total number of possible cases is called probability.

Probability = (No. of favourable outcome/ Total no. of possible outcome)

Trial 🡪 Each performance in a random event

Event🡪 Outcome of trail is an event

Random Experiment🡪 To perform more than one

Sample Space 🡪 Set of all possible outcomes.

Question 1: What is the Probability of getting an even when throwing a single die.

Solution – Sample Space (S) = getting an even number

Event (A) = getting an even number

P(A) =?

P(A) = No. of favourable cases / total no of possible outcome

P(A) = 3/6

P(A) =1/2

Question 2: What is the Probability of getting an head when toss a fair coin.

Solution – Sample Space (S)= {H.T}

Event (A) = getting a head

P(A) =?

A={H}

P(A) = No. of favourable cases/ total no of possible outcome

P(A) =1/2

So probability of getting a head on tossing a fair coin is 1/2.

Skewness: measures the distribution of the data. It indicates weather the data is distributed symmetric or not. If the data is distributed symmetric means the data is normally distributed.

Type of skewness:

1. Symmetric
2. Positive skewness
3. Negative skewness

Pearson coefficient of skewness:

Skp = (Mean-Mode)/standard deviation

Three scenarios:

1. Normally distributed (Symmetric) : Mean = Median= Mode
2. Positive skewed (right skewed): Mean=Median=Mode
3. Negative skewed (left skewed): Mode=Mean=Mode

How to handle right skew and left skew?

Right Skew:

Log transformation, square root transformation, cube root transformation and reciprocal.

By using the log transformation in numpy:

Here is an example:

Log\_mileage = np.log(data[‘mileage’])

Log\_mileage.head(5)

#Checking the skewness after the log-transformmation

Log\_mileage.skew()

If you are getting zeros inside the data, refer to root transformation.

Root transformation is square root of original data.

Example:

Sqrt\_mileage = np.sqrt(data[‘mile\_age’])

#Checking the skewness after the root transformation:

Sqrt\_mileage.head(5)

Cube root transformation: The cube root means x to x^(1/3). It is weaker than the logarithm but stronger than the square root transformation. It is also used for reducing right skewness and has the advantage that it can be applied a volume has the units of a length. It is commonly applied to rainfall data.

Cube\_root\_mileage = np.cbrt(data[‘mileage’])

Cube\_root\_mileage.head((5)

Reciprocals transformation: if the value is x=20, reciprocal value is =1/20.

Outlier:

1. A data point that is significantly greater or smaller than other data points in a dataset.
2. It is useful when analyzing data to identify outliers
3. They may affect the calculations of descriptive statistics
4. Outliers can occur in any given dataset and in any distribution

The easiest way to detect them is by graphing the data or using graphical method such as:

1. Histograms
2. Box plots
3. Normal distribution plots
4. Outlier may indicate an experimental error or incorrect recording of data
5. They may also occur by chance

* It may be normal to have high or low data points

1. You need to decide whether to exclude them before carrying out your analysis

* An outlier should be excluded if it is due to human error.

Steps in a data-driven decision making

1. Formulate a hypothesis
2. Find the right test
3. Execute the test
4. Making a decision

Hypothesis:

A hypothesis is an idea that can be tested.

First 🡪 Identify the problem (question)

* Gather data (research)
  + Hypothesis
    - Test Hypothesis(experiment)
  + Does the new data agree?

Measure of central tendency: central tendency measures the center value of the dataset. It gives us idea about the concentration of the value in the central part of the distribution.