BASIC STATISTICS

TYPES OF DATA

1. Categorical Data

- Descriptive data that is present in form of language.
- They usually have fixed number of outcomes like gender, male and female.
- Can be classified as:
  1. **Nominal Data**
     - Data has no meaningful order.
     - Examples are **nationality** [Indian, American etc] and **gender** [male, female].
  2. **Ordinal Data**
     - Data has a meaningful order.
     - Examples are **spicy** [less spicy, medium spicy, very spicy] and **satisfaction** [not satisfied, somewhat satisfied, completely satisfied].

2. Numerical Data

- Data in form of numbers.
- Can be classified as:
  1. **Discrete Data**
     - These are countable and non-divisible form of data.
     - Easily visualised using bar charts, pie charts and histograms.
     - Example, number of students in a class, number of tickets sold.
     - In many cases, discrete data is prefixed with ‘ the number of ’.
  2. **Continuous Data**
     - These are measurable form of data that may or may not have decimal points.
     - They can be any real number in a fixed range.
     - Continuous data changes over time.
     - Example, height and weight of students, weather and temperature, length of customer care calls.

TYPES OF STATISTICS

1. Descriptive Statistics

- These are used to describe the basic features of data that provide a summary of the given dataset.
- It consists of methods for organising, displaying and describing data using tables, graphs and summary measures.
- Categorised as:
  a. Measure of central tendency [mean, median, mode etc]
  b. Measure of dispersion or variability [range, standard deviation, variance etc]
  c. Frequency distribution

2. Inferential Statistics

- It consists of methods that use sample data to make decisions and predict about the population.
- Examples are Linear Regression, Random Forest and Neural Networks.

MEASURE OF CENTRAL TENDENCY

A. Mean

- Equal to sum of all values divided by number of values in dataset.
- Population mean is for complete dataset or population dataset.
- Sample mean is for sample dataset where a sample dataset is a small part of population dataset.
- Population mean is denoted by u and sample mean is denoted by x bar.
- It is not a good measure of central tendency for data with outliers or extreme values.
- For the set {3, 6, 1, 9, 5}, the mean is 4.8. For the set {3, 6, 1, 9, 5, 60}. The mean is 14.

B. Median

- Value at the centre position of an ordered dataset.
- It is a rank based method to measure central tendency of dataset.
- Example: For the set {3, 6, 1, 9, 5}, first sort it to {1, 3, 5, 6, 9}. The median is 5.
- If a dataset has even number of values at the centre, then the average of 2 numbers at the centre of an ordered dataset is median.
- Median is also 50th percentile in a dataset.
- Good measure of central tendency for data with outliers.
- For the set {3, 6, 1, 9, 5}, first sort it to {1, 3, 5, 6, 9}. The median is 5. For the set {3, 6, 1, 9, 5, 60}, first sort it to {1, 3, 5, 6, 9, 60}. The median is 5.5. Median was hardly influenced by the outlier.

C. Mode

- Value with the greatest frequency in the dataset is called mode.
- When 2 values have greatest frequency, dataset is call Bimodal.
- When more than 2 values have greatest frequency, dataset is called Multimodal.
- In case of using 2 modes in a dataset, the one with highest frequency is called major mode and the one with second highest frequency is called minor mode.
- It is used for categorical data as well as numerical data.
- Categorial Example, out of 100 people, 80 people called the dish spicy and remaining 20 chose from the options not spicy, less spicy or very spicy.
- Numerical Example, 10 people rated the dish between 0 to 10, 10 being the most spicy {4, 4, 2, 6, 8, 4, 9, 6, 5, 8}, the mode is 4.

D. Midrange

- Equal to sum of largest and smallest values in dataset divided by 2.
- The midrange is a straightforward measure that considers the range of the data by taking into account both extremes.
- It is particularly useful when you want a basic summary measure that incorporates the highest and lowest values, providing a simple understanding of the central tendency.
- However, it may not be as robust as the median when dealing with skewed distributions or datasets with outliers.

MEASURES OF DISPERSION / VARIABILITY

- Understanding the variability of data, how the data is spread around the mean helps us to understand data better. Variability is everywhere. Your commute time to work varies a bit every day. When you order a favourite dish at a restaurant repeatedly, it isn't exactly the same each time. These are all examples of real-life variability.
- Some degree of variation is unavoidable. However, too much inconsistency can cause problems. If your morning commute takes much longer than the mean travel time, you will be late for work. If the restaurant dish is much different than how it is usually, you might not like it at all.

A. Range

- Equal to difference between largest and the smallest values in dataset.
- Higher the range, higher the variability in the dataset.
- It is not a good measure for dispersion if outliers are present. Even a single outlier can affect the range drastically.

B. Variance [sigma^2]

- Equal to summation of squared differences from mean divided by N, N being the number of data points.
- To calculate sample variance, the denominator is N-1.
- Standard deviation is more meaningful and easier to interpret than variance.

Population mean is known

$$var(x) = \frac{\sum_i^n (x_i - \mu)^2}{N}$$

Population mean is unknown

$$var(x) = \frac{\sum_i^n (x_i - \overline{x})^2}{N - 1}$$

C. Standard Deviation [sigma]

- Equal to square root of variance.
- Standard deviation has the same unit as data.
- It tells us about the average spread of each data point with respect to mean.
- Standard deviation is a good measure of dispersion. It is more meaningful and easier to interpret than variance because it has the same unit as data.

D. Percentile

- Percentile is basically the percentage of data that falls below a particular value.
- If IQ level of 120 is equal to 90th percentile, it means that 90 percent of the people who took an IQ test have an IQ below 120.
- It helps us to understand how our value stands relative to complete data.
- For example, JEE score of 160 is of no use unless we compare it. JEE percentile of 95 gives us a better understanding of result and performance.
- It is used to identify outliers using a method called capping and flooring.
- Example
  - A list of height measurements of 60 students [162, 158, 173, …, 180]
  - Arrange the list in ascending order
  - We want to find percentile of 176cms in the list
    1. Locate the rank of 176cms in ordered list
    2. Lets assume it is present at 54th rank
    3. Percentile = rank/total * 100
    4. Percentile = 54/60 * 100 = 90%
    5. Hence 90% of students have height below 176cm
  - We want to find the height at 60th percentile
    1. Position = percentile/100 * (total+1)
    2. Position = 60/100 * (60+1) = 36.6 [between 36th and 37th position]
    3. Assume 36th position has 165cms and 37th position has 167cms
    4. Rank = 165 + (0.6 * (167-165)) = 165 + (0.6 * 2) = 165 + 1.2 = 166.2cms
    5. Hence the height at 60th percentile is 166.2cms

E. Quartile

- 3 values that split our dataset into 4 quarters.
  1. 1st quartile Q1 - 25 percent of dataset falls below this value
  2. 2nd quartile Q2 - divides the dataset into 2 halves, also called median
  3. 3rd quartile Q3 - 25 percent of dataset falls above this value
- Interquartile range IQR is equal to Q3 - Q1 where 50 percent of the data is present. High IQR value represents broader spread of the dataset in the middle.
- It can also help us find outliers. Calculate lower fence and upper fence. Values above upper fence and values below lower fence are potential outliers.
  1. Upper fence = Q3 + (1.5*IQR)
  2. Lower fence = Q1 - (1.5*IQR)

Note

- When you have a skewed distribution [outliers in data], median is a better measure of central tendency than mean because mean is highly influenced by outliers. Pairing it with percentile based ranges gives us a good understanding of the data because it divides the dataset into groups with specific proportions.

- For normally distributed data, or even data that aren't terribly skewed, using the tried and true combination of mean and the standard deviation is the way to go. This combination is by far the most common. You can still supplement this approach with percentile-base ranges as you need.

TYPES OF FREQUENCY DISTRIBUTION

1. Absolute frequency

- Number of counts of each category or class.

2. Cumulative frequency

- Frequency of the first class added to the frequency of second class and then added to the frequency of third class and so on.
- Last category will have cumulative frequency equal to total frequency.

3. **Relative frequency**
   - Fraction or percentage of frequency for a specific category or class with respect to total frequency.
4. **Cumulative relative frequency**
   - The fraction or percentage of cumulative frequency of different classes with respect to the cumulative frequency. The cumulative relative frequency of the last class interval is equal to 1 or 100%

Score	Absolute Frequency	Absolute Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
1	0	0	0%	0%
2	2	2	10%	10%
3	1	3	5%	15%
4	2	5	10%	25%
5	2	7	10%	35%
6	1	8	5%	40%
7	3	11	15%	55%
8	4	15	20%	75%
9	1	16	5%	80%
10	4	20	20%	100%
Total	20		100%	100%

1. **Grouped frequency distribution**
   - We form class intervals and tally the frequency of data that lies in that interval.
   - It is not possible to create a frequency table for data having lots of distinct values. In such a case, creating class intervals can help us overcome the problem and get a clean and short frequency table.
2. **Ungrouped frequency distribution**
   - We do not form class intervals but use the exact category and tally the frequency for that particular category.

FREQUENCY DISTRIBUTION GRAPHS

1. Symmetric

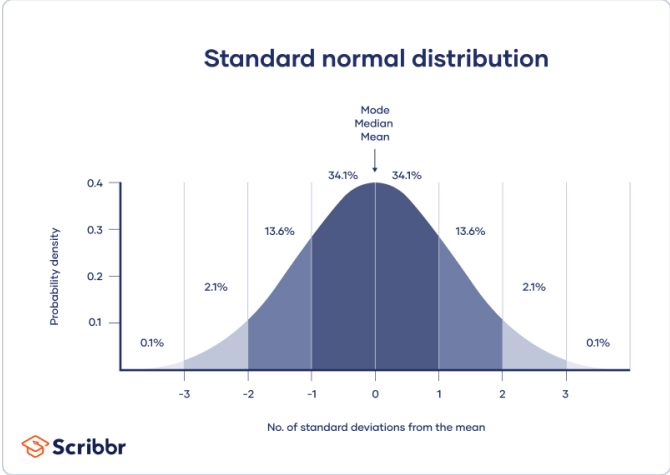
A frequency distribution cut into 2 equal halves vertically when gives us exact mirror images is called symmetric frequency distribution.

1. Normal Distribution

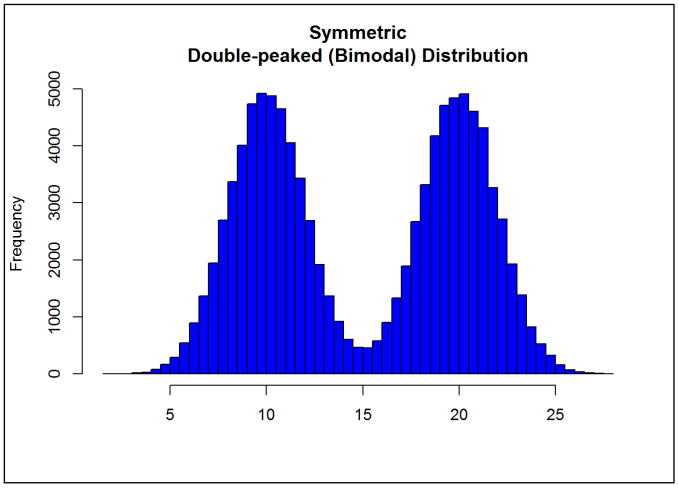
- Most common form of symmetric distribution.
- If a frequency distribution is symmetric and shaped like a bell, then it has a normal distribution. The curve exponentially reduces on both ends indefinitely.
- Mean, median and mode are equal to one another. The average, 50th percentile and value having the highest frequency, all are equal to one another.
- Normal distribution has 2 parameters, mean and standard deviation.
- Empirical rule of normal distribution, mean +/- nSD is equal to X percentage of data contained
  - n = 1, X = 68%
  - n = 2, X = 95%
  - n = 3, X = 99.7%

2. Bimodal Distribution

- Bimodal distribution is much alike to a normal distribution but instead of having a single peak, a bimodal distribution has 2 peaks. This means that the dataset is bimodal.
- For the graph below, the mean and median are equal to one another, around 15. Modes are 10 and 20. Hence, mean and median are still equal for bimodal distribution.



- Mean and median are equal to 15 and to one another in the distribution below because what we have below is a symmetric bimodal distribution. This may or may not be the case every time.



3. Rectangular or Uniform Distribution

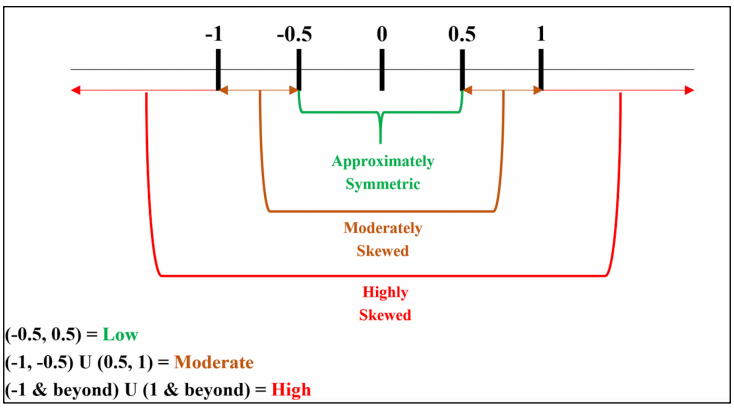
- This happens when frequency of all categories or classes in a frequency distribution are equal to on another.

2. Asymmetric

- This frequency distribution has a varying density in different parts of graphical representation.

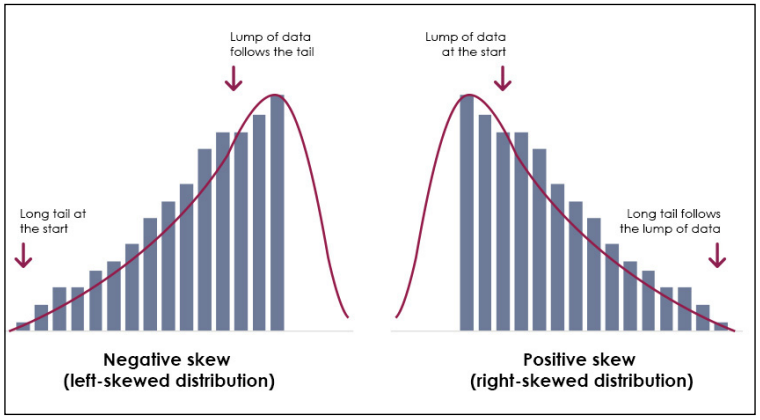
1. Skewed Distribution

- Skewness is the degree to which a frequency distribution graph shifts horizontally with respect to normal distribution.
- Skewness = (mean - mode)/SD
- Division by standard deviation keeps the skewness in same standard scale.



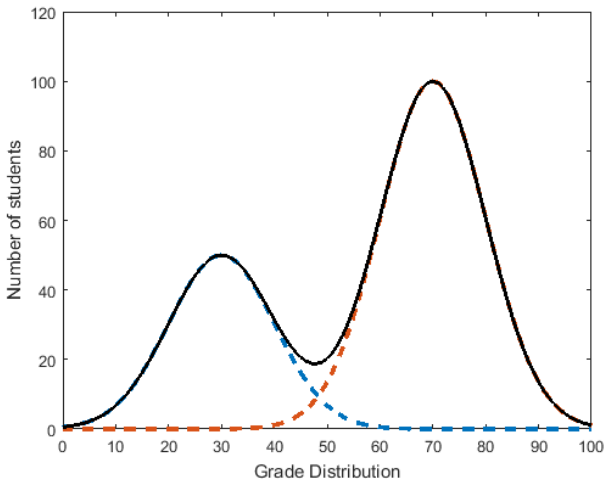
- Right/Positively Skewed
  - Most of the values will fall towards left of the distribution.
  - The tail will exponentially reduce towards right indefinitely.
  - The mode will lie somewhere near the bucket with highest frequency which is towards the left.
  - Mean has the highest value.
  - Median lies somewhere between mean and mode.

- Left/Negatively Skewed
  - Most of the data will fall towards right of the distribution.
  - The tail will exponentially reduce towards left indefinitely.
  - The mode will lie somewhere near the bucket with highest frequency which is towards the right.
  - The mean will have the least value.
  - Median will lie between mean and mode.



2. Bimodal Distribution

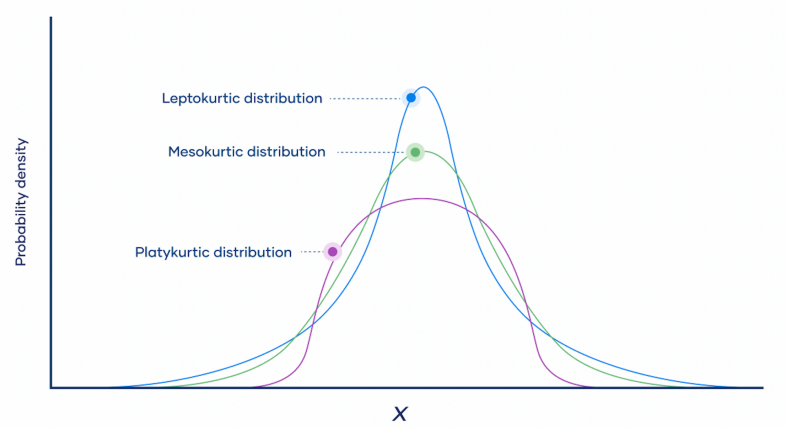
- Bimodal distribution is form of frequency distribution having two modes, a major mode and a minor mode. A graphical representation of bimodal distribution displays two peaks.
- It mostly occurs when we are dealing with combined population.
- Identifying a bimodal distribution is a valuable finding. This can change the understanding of the data. It is better to assess the subpopulations separately to understand individual distribution. Both of them have different central tendency and variability that is not possible to understand when put together
- Example - a frequency distribution of weights of 100 people, 50 are men and 50 are women. Men and women have different average weight and different variability. Average weight for men is way above average weight for women. A graphical representation for a such a frequency distribution will be bimodal



3. Kurtosis
   - Just like skewness is the horizontal distortion from normal distribution, kurtosis is the vertical distortion from a normal distribution.
   - Kurtosis is the measure of tailedness of a distribution where the tail is the reducing curve that goes indefinitely. The tail tells us about the frequency of values that are extremely high or low relative to the mean.
   - The range of kurtosis is from 1 to infinity
   - It is the outliers that affect the kurtosis majorly rather than the values near the mean
   - Excess kurtosis is the measure of tailedness of a distribution relative to normal distribution. It is equal to kurtosis minus 3 where 3 is the kurtosis for normal distribution. The range of excess kurtosis is -2 to infinity

      Mesokurtic [close to normal distribution]
      - Kurtosis is approximately equal to 3, so excess kurtosis is approximately equal to 0
      - Medium tailed, so the frequency of outliers is neither high nor low

   2. Platykurtic [also called negative kurtosis]



      - Kurtosis is less than 3, so excess kurtosis is less than 0
      - Thin tailed, so the frequency of outliers is low
      - A good example is uniform distribution. They rarely have any outliers

   3. Leptokurtic [also called positive kurtosis]
      - Kurtosis is greater than 3, so excess kurtosis is greater than 0
      - Fat tailed, so the frequency of outliers is high
      - A good example is Laplace distribution

MEASURES OF RELATIONSHIP

1. Covariance

- Covariance is a statistical measure of the variability of 2 variables.
- It tells us about the relationship of 2 variables. It answers the question of what happens to a one variable when another one changes.

Population mean is known

$$cov(x,y) = \frac{\sum_i^n (x_i - \mu) \cdot (y_i - \mu)}{N}$$

Population mean is unknown

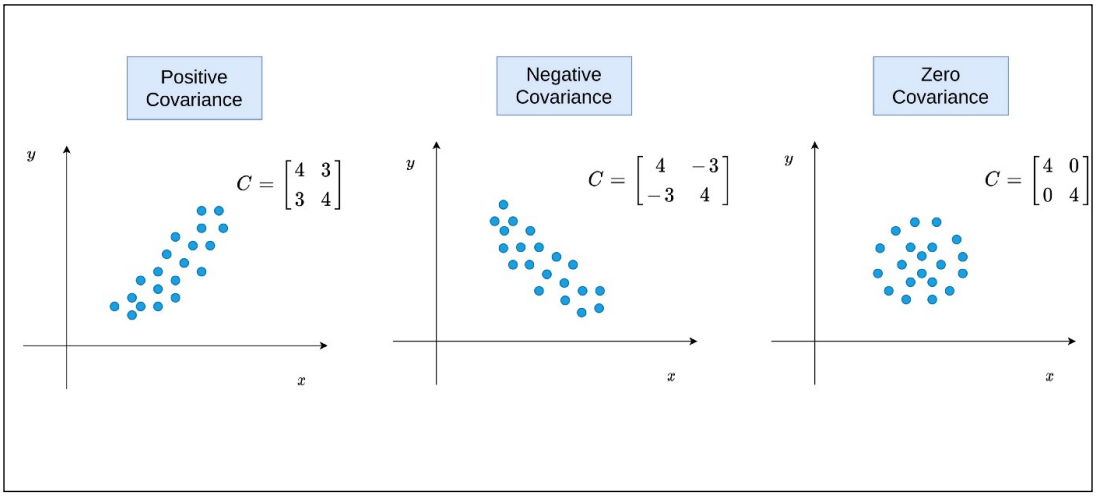
$$cov(x) = \frac{\sum_i^n (x_i - \overline{x}) \cdot (y_i - \overline{y})}{N - 1}$$

- Covariance matrices

- They are plotted to see covariance between 2 variables.
- The diagonal plots variance of variables because the covariance between same variables will always be equal to variance.

$$\begin{array}{c} \\ x \\ y \end{array}\begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ x \\ y \\ z \end{array}\begin{array}{c} \begin{array}{ccc} x & y & z \end{array} \\ \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix} \end{array}$$

- Range
  - It can range from negative to positive infinity which is a drawback.
  - Negative means inverse covariance and positive means direct covariance.
  - Looking at the formula, if both the terms are negative or positive, only then can we have a positive covariance, otherwise negative covariance.
  - If in any of the terms, the difference is 0, then there is no covariance between the variables.
  - Like below, if the graphical representation of data is horizontal or vertical, then also we can assume that covariance is almost equal to zero
  - Covariance is not normalised. So this is not a good measure to understand relationship between variables. If 2 variables have values between 1000 and 2000, covariance will be high. If 2 variables have values between 1 and 2, covariance will be low. So, it is a good measure to understand the direction of relation between variables but not the strength.



2. Correlation

- Correlation is also a measure of the variability of 2 variables.
- But it gives a better understanding than covariance because it is actually a normalised form of covariance.
- The range for correlation is from -1 to +1 rather that from negative to positive infinity for covariance.

$$Correlation = \rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

3. Correlation and Causation Difference

- Correlation is a statistical indicator that says that there is a relationship between 2 variables.
- Causation mean that a change in one variable will bring change in another variable. There is a cause-and-effect relationship between 2 variables.
- Causation implies that there is correlation but correlation does not necessarily imply that there is causation.
- Assume a house price prediction. We know that when the number of rooms increase, the price of the house increases, so there is a causal effect, causation and hence correlation also. Now if we consider a third variable like airports, the number of rooms and airport might have a correlation but there is definitely no causation between them.
- Why correlation does no imply causation
    1. Third variable problem
        - Sometimes there is a correlation between 2 variables but it cannot be explained. So a confounding variable, a third variable is used to understand such a correlation.
        - For example, ice cream sales and crime rates are closely correlated, but they are not causally linked with each other. Instead, hot temperatures, a third variable, affects both separately.
    2. Dimensionality problem
        - Sometimes there is causation between 2 variables but it is not clear which of them causes a change in another.
        - For example, vitamin D and depression are correlated to each other. But is not clear whether a change in vitamin D causes change in depression or a change in depression causes change in vitamin D.

ODDS & LOG(ODDS)

1. Odds is the ratio of probability of an event happening (p) to probability of an event not happening (1-p)
2. Odds = p / (1-p)
3. Example, out of 5 matches, a team won 1 match.

   probability of winning = 0.2
   probability of losing = 0.8
   Odds of winning = probability of winning / probability of losing = 0.2 / 0.8 = 0.25
   Odds of losing = probability of losing / probability of winning = 0.8 / 0.2 = 4
4. But simply using odds is not a good measure because of lack of symmetry.
5. Example, a team won 1 match out of 20 matches.

   probability of winning = 0.05
   probability of losing = 0.95
   Odds of winning = probability of winning / probability of losing = 0.05 / 0.95 = 0.052
   Odds of losing = probability of losing / probability of winning = 0.95 / 0.05 = 19
6. Above example shows that odds of winning has a range of (0,1) and odds of losing has a range (1, infinity)
7. The log of odds solves this problem.

   Log(0.25) = -0.602
   Log(4) = +0.602