Probability and Statistics

1. Gaussian Distribution and its PDF
2. CDF of Gaussian Normal distribution
3. Central Limit Theorem
4. QQ Plot
5. Correlation Coefficient and its type
6. Power Law
7. Pareto Distribution Function
8. Co-varinace, Pearson coeffient, Spearman rank corelation coeffient
9. Confidence Interval, CI menas for random variable
10. Hypothesis Testing, P-value

1. Random Variable:

- It means set of possible outcomes of an experiment.
- A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.
- Types of Random Variable are -
- Discrete Random Variable
- Continuous Random Variable

1.1 Discrete Random Variable-

- Random Variable which has a finite set of outcomes or values is called a discrete random variable
- Eg - A random experiment of tossing a coin results in either heads or tails, Rolling of a dice

1.2 Continuous Random Variable-

- Random variable which can take any real value is called a continuous random variable.
- Eg- The probability of height of a group of students.

Outlier

An observation point that is distant from other observations is called an outlier.

Note:

Mean and variance gets corrupted by an outlier, hence we use median and median absolute deviation.

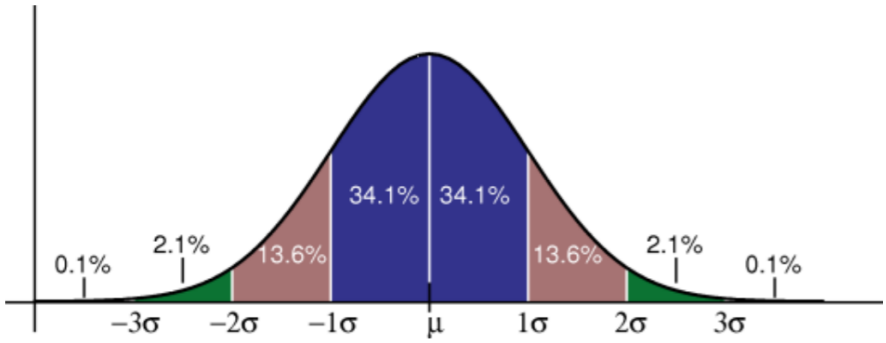
Median Absolute Deviation - MAD

It is the summation of all the absolute differences of the median from the individual points)

Gaussian/Normal Distribution -

If X is a continuous random variable that has a PDF like that of a bell shaped curve, then we say X has a distribution which is a **Gaussian Distribution**.

- Parameters of Gaussian Distribution-
  1. Mean
  2. Variance
- **Variance** is the spread of the curve. So, if it is small, the curve is going to be steeper.
- **Mean(μ)** is generally the peak of the distribution.

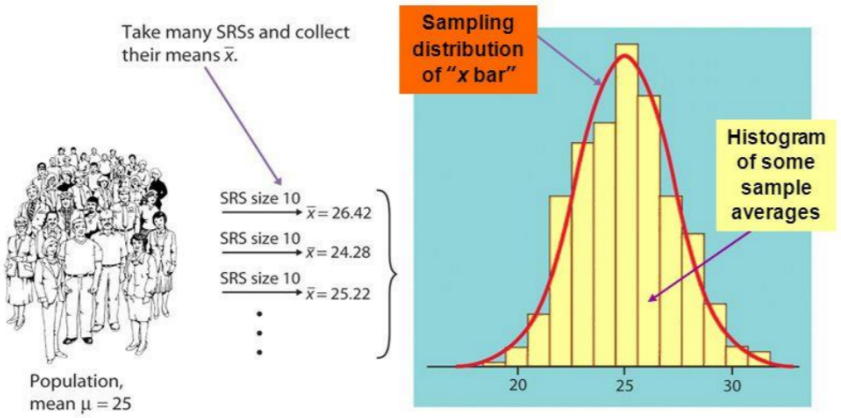


Standard Normal Distribution / Standard Normal Variate (Z):

A standard normal variate is a normal variate with mean μ=0 and standard deviation σ =1.

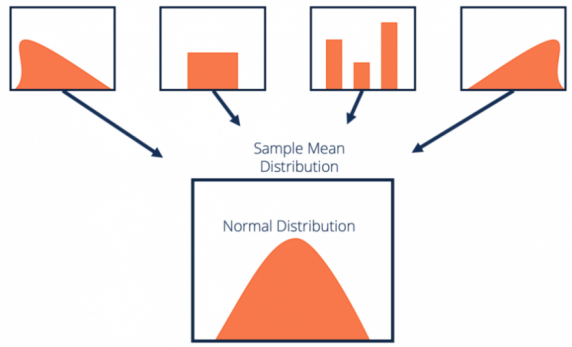
Central Limit Theorem CLT -

- Given a dataset with unknown distribution (it could be uniform, binomial or completely random), the sample means will approximate the normal distribution. Example -



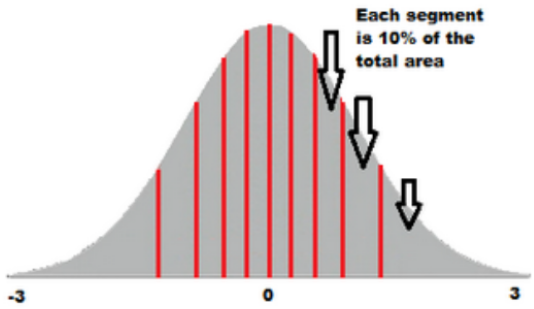
If we take m samples with n data points from a population distribution X which may not be Gaussian, and we calculate the mean of every samples, then, CLT states that-

1. **The distribution of the mean of the samples will be a Normal distribution with the population mean as it's Mean** (roughly).
2. The standard deviation equal to (population variance/the number of data points in each sample) which can be written as, N(m, s^2/n) as n->∞.
3. If n is sufficiently large then sample means has Normal Distribution.



Q-Q Plot -

Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities or dividing the observations in a sample in the same way.



- 2 quantile is known as the Median
- 4 quantile is known as the Quartile
- 10 quantile is known as the Decile
- 100 quantile is known as the Percentile

Ques- How to check random variable(X) is Gaussian(Normal) Distributed ?

1. Sort the data of X in ascending order and then compute percentile of data.
2. Create a Y ~ N(0,1) which is Standard Normal Distribution and sort them then compute Percentile.
3. Plot Q-Q plot of X and Y, if all points lie on same straight line then random variable X and Y have similiar distribution. If X is Gaussian,Normal Distribution or Pereto then Y is also Gaussian,Normal Distribution or Pereto respectively.

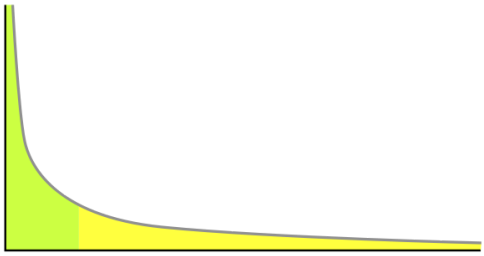
Ques- How to check two distribution is statistically same, Here X and Y have different distribution ?

1. First, sort the given random variable in ascending order.
2. Then, generate a theoretical Quantile of the test distribution and sort it.
3. Then, graph the respective elements matching each of the given random variable and the theoretical distribution.
4. If the graph generates a straight line, then the 2 distributions are statistically same, else not.

Limitation- If no. of sample is small then its hard to interpret QQ-Plot.

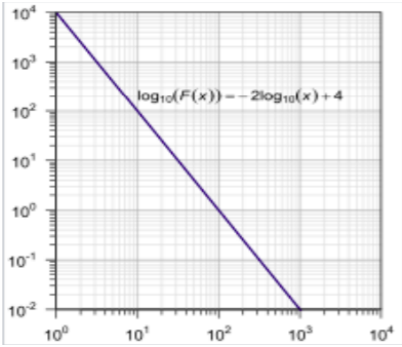
Power Law Distribution-

- The Power Law (Pereto Distribution/Scaling Law) states that a relative change in one quantity results in a proportional relative change in another.
- Eg- The simplest example of the law in action is a square; if you double the length of a side (say, from 2 to 4 inches) then the area will quadruple (from 4 to 16 inches squared).
- A power law distribution has the form Y = k (X^α),
  where: X and Y are variables of interest, α is the law's exponent, k is a constant.
- It roughly follows the **80-20 rule**, i.e., 80% of the points lie in the 20% of the region of the distribution.



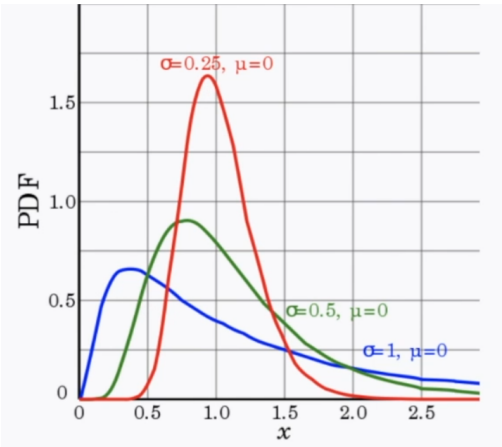
Eg- Distribution of income, Magnitude of earthquakes, Size of corporations.

Ques- How to check Distribution is Power Law or not?
Ans- Take log of x and log of y and if they have slat-straight from upper to lower in increasing order then its follow Power law.



Log Normal Distribution

- Random Variable X is considered to be Log-Normal if log(X) is Normally Distributed.
- It have long tail like graph.
- Eg - Length of comments posted in social media.

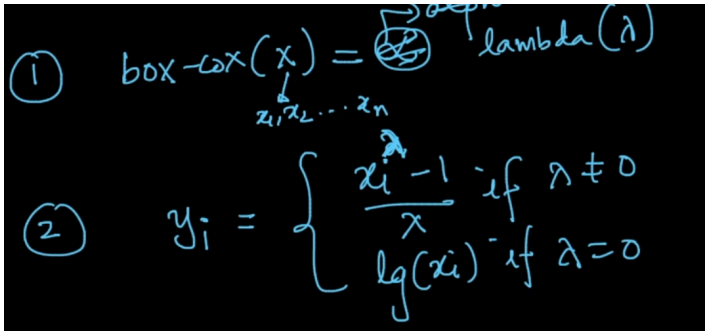


Ques- How to test random variable X(x1,x2,x3...xn) have Log-Normal Distribution ??

1. Take log(x1),log(x2)...log(xn).
2. Sort the data of X in ascending order and then compute percentile of data.
3. Create a Y ~ N(0,1) which is Standard Normal Distribution and sort them then compute Percentile.
4. Plot Q-Q plot of X and Y, if all points lie on same straight line then X have Log-Normal Distribution.

Power transform(Box-cos tranform)

- It is used to tranform Pareto distribution to Guassian distribution.
- By using Box-Cox method in X we get lambda, if lambda is equal to 0 then we get log(x) else we get in below eqn.

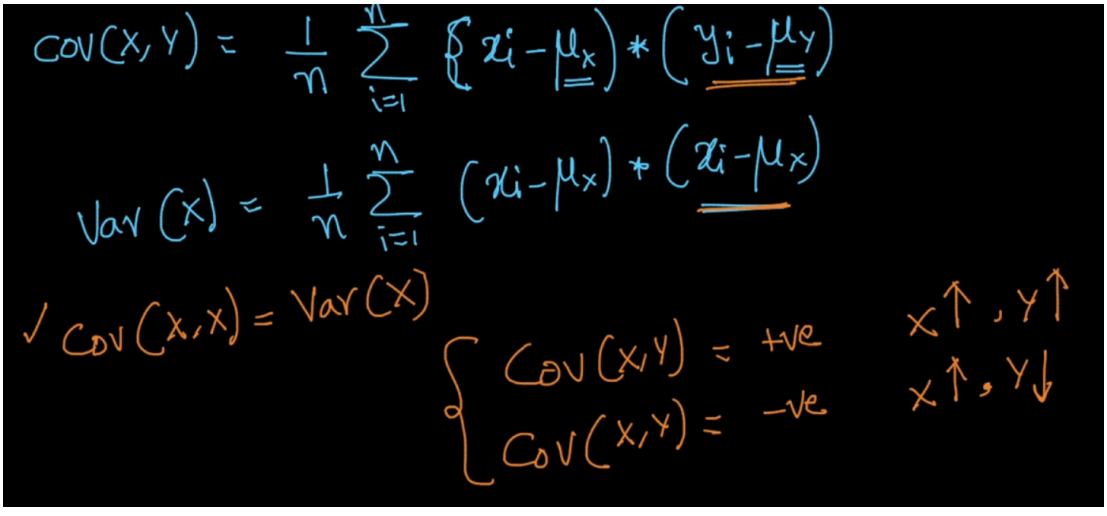


To find relationship betw two variable x,y belongs we use following methods are-

1. Co-variance
2. Pearson correlation coefficient(PCC)
3. Spearman rank corr coe.

1. Covariance -

- Covariance is the find of a relationship between two or more random variable.
- Covariance(X,Y) is +ve, if X and Y increases.
- Covariance(X,Y) is -ve, if X increases and Y decreases.

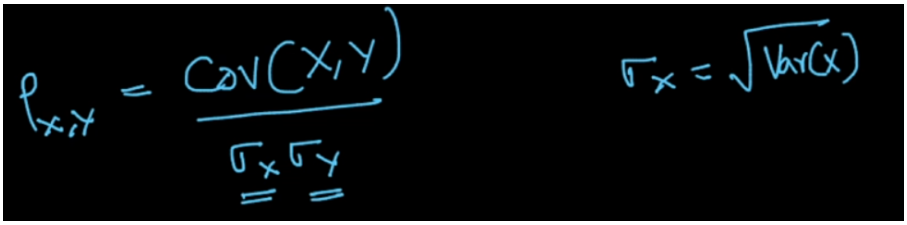


Drawback- By changing the units of measure, co-variance may differ.

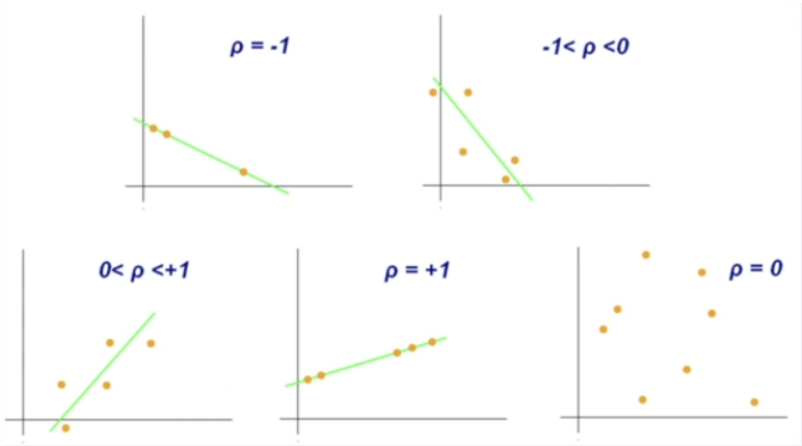
- It means for same dataset of X and Y(cm,kg) have different Covariance if we change the unit of X,Y(feets,pounds).

2. Pearson Correlation Coefficient PCC-

- PCC lie between -1<=PCC<1, PCC =0, when there is no relation betwn x and y, refer graph in each case.



mui(x) is Standard Deviation of x



- If Rho=0, then there is no relation between X and Y
- The slope of the line doesn't matter
- **Drawback-** PCC cannot capture complex non-linear relationships. Like sinusodial wave.

3. Spearman's Rank-Correlation Coefficient-

- Here we find rank(x) and rank(y), and calculate **PCC on rank(x,y)**
- To find rank just range in increasing order of x and y.
- As X and Y increases for linear or non-linear distri. rank =1
- As X increases and Y decreases linear or non-linear distri. rank =-1.
- It works fine in monotonous increasing curve.

Correlation VS Causation

If two variable are co-related with each other that does not means one is the cause of other.

Hypothesis Testing

Hypothesis Testing a concept by which a statement is validated through a Proof by Contradiction.

Hypothesis testing a concept by which a statement is validated through a Proof by Contradiction.

1. We first start by choosing a Test-Statistics(μ2-μ1)
2. We provide a Null hypothesis rejecting the test statistics. We choose an Alternate hypothesis which is the complement of Null hypothesis.
3. Set the Significance Level or Critical Value.
4. We check the p-value or the Probability of the Test Statistics, and assume the Null hypothesis is True.
    - If, p-value is closer to 1, we accept the Null Hypothesis.
    - Else if p-value is closer to 0, we accept the Alternate Hypothesis.

Critical Value-

```
* Normally Critical Value is 5%.
* If our test score lies in the critical zone, we reject the Null Hypothe
sis and accept the Alternate Hypothesis
```

p-Value-

- p-Value is the probability of observation that we already made given Null-Hypothesis is True.
  p-value = P(Obs/H0)

Example -

- Given a coin, determine if the coin is biased towards Head or Not.
  Basic probability we know-
  Biased towards Head- P(H) > 0.5
  Not-Biased towards Head - P(H) = 0.5

  Experiment - Flip a coin 5 times and count the no. of heads. To check Coin is Biased or not.
  Count no. of Head(X) is our Test-Statistic.
  After performing experiment we get no. of Head, X = 5
  Our Null-Hypothesis(H0) is Coin is not biased towards Heads.
  Our Alternate Hypothesis(H1) is Coin is biased towards Heads.<

  P(X=5|coin is not Biased) = P(Obs|H0)

  Probability of coin is not biased = 1/2
  Probability of getting 5 Heads in 5 toss assuming to the coin is not biased is 1/2^5
  P(Obs|H0) = 1/2^5

  ```
  = 1/32 = 0.03 = 3%
  So here p-value = 3%
  ```

- It means that there is 3% chance of getting 5 Heads in 5 flips if the coin is not biased towards Head.
- If P(Obs|H0) < 5%, then H0 may be incorrect.
  It means our assumption or H0 is not True => Reject H0 or Accept H1 => Coin is Biased Towards Head.

Permutation Test(Resampling) -

- It is used to find p-value from sample.
- In Real World there is no exact p-value exist, its depend on where we are using p-value, Eg in medicine we take p-value is 1% it means here are very-very sure about certain things.

Example-
We are here to find which class students have heights greater than other class.

- Let there are two class A and class B having 50-50 students in each class with μ1 and μ2 as mean of Class A,B respectively. Mean Difference is μ1 - μ2 = Δ
- We jumble both class students and then randomly separate with 50-50 students in let say class X and Y, with μ1 and μ2 as mean value.
- Difference of μ1 and μ2 is δ1.
- Repeat step 2 and 3 to 10k times and we get δ2, δ3...δ10k.
- After getting all δ1, δ2, δ3...δ10k, Sort them to get - δ1', δ2', δ3'...δ10k'.
- Then we fit our Actual Δ to above sorted list of sorted means.
- After placing our Δ in above sorted list of sorted means and Lets assume there are x% points greater than sorted means then we can say our p-value is x%. If x is 5 then p-value = 5%