Lecture-07

→ P value and Significance Value.

→ Distribution.

→ Central limit theorem.

→ Bernoulli's distribution.

→ Binomial distribution.

→ Pareto's distribution. [Power law distribution]

→ Log Normal distribution.

→ F Test (ANNOVA)

P Value and Significance Value:

↳ Derive the p value.

1. The average weight of all residents in Bengalore city is 168 pound with a standard deviation 3.9. We take a sample of 36 individuals and the mean is 169.5 pounds with 95% of Confidence Interval.

Given,

$\mu = 168 \quad \sigma = 3.9 \quad n = 36 \quad \bar{x} = 169.5$

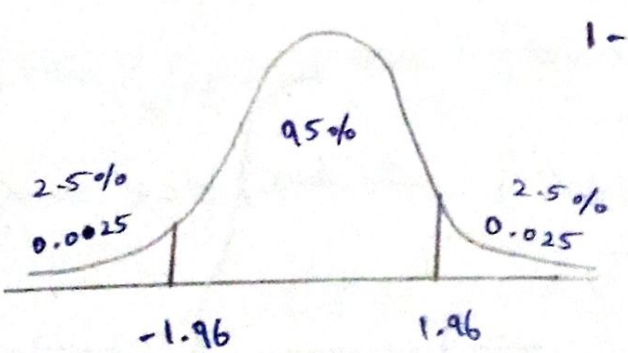
$\alpha = 0.05$

1. Null hypothesis, $(H_0) = \mu = 168$

2. Alternate hypothesis, $(H_1) = \mu \neq 168$

3. Significance Value $= 0.05$

4. Decision boundary, [obiously Two tail]



5. Test statistics,

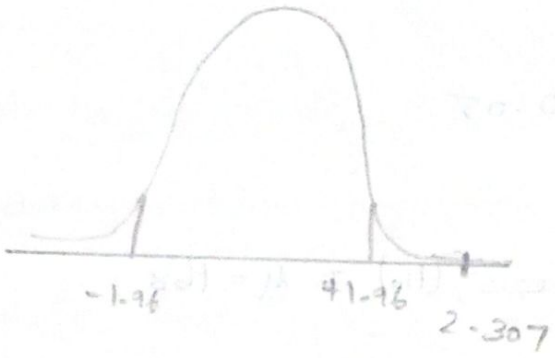
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}} = \frac{1.5}{3.9} \times 6$$

$$= 2.307$$

$Z = 2.307$

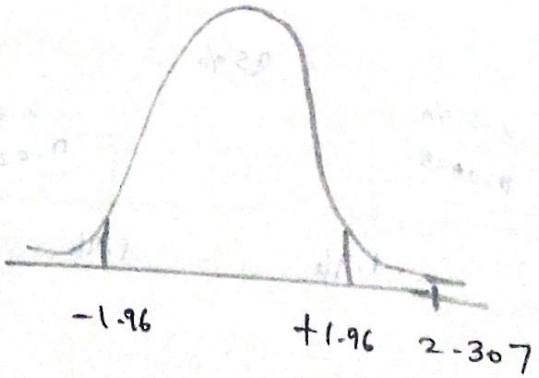
6. State Own decision,

$Z = 2.307 > 1.96$



So, obiviously reject the null hypothesis.

7. State Own decision with the help of P-Value,



Some extend the Curve,

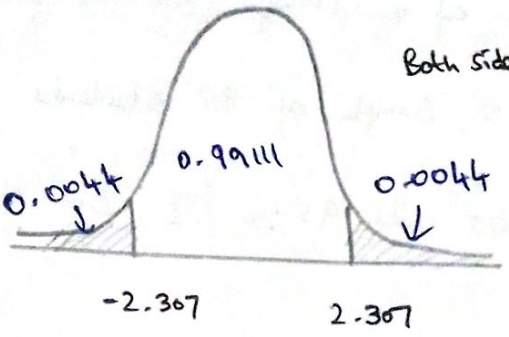
Two Tail,

$1 - 0.99111 = 0.00889$

↳ Area of one side

Both sides are Symmetric, So

$\frac{0.0089}{2} = 0.0044$



Add,

$0.0044 + 0.99111 + 0.0044$

$= 1$ [whole distribution]

⇓ See Z score Value in Z table

0.99111

(Area under the Curve)

So, the P-value is calculated with the help of two tail,

$P\ value = 0.0044 + 0.0044$

$\boxed{Pval = 0.0088}$ [Probability of data fall within the distribution]

We decide,

$P\ value < Significant\ Value$ [Reject H_0]

$P\ value > Significant\ Value$ [Accept H_0]

$P\ value < 0.05$

$\boxed{0.0088 < 0.05}$ [Reject the null hypothesis]

So, The average weight of all residents in Bengaluru City is 168 is statistically accepted.

2. Average age of a College is 24 years with a Standard deviation 1.5. Sample of 35 students Mean is 25 years. with $\alpha = 0.05$ with 95% CI, Does the age vary?

Given,

$\mu = 24 \quad \sigma = 1.5 \quad \bar{x} = 25 \quad n = 35 \quad \alpha = 0.05$

(i) Null hypothesis, $H_0 = \mu = 24$

(ii) Alternate hypothesis, $H_1 = \mu \neq 24$.

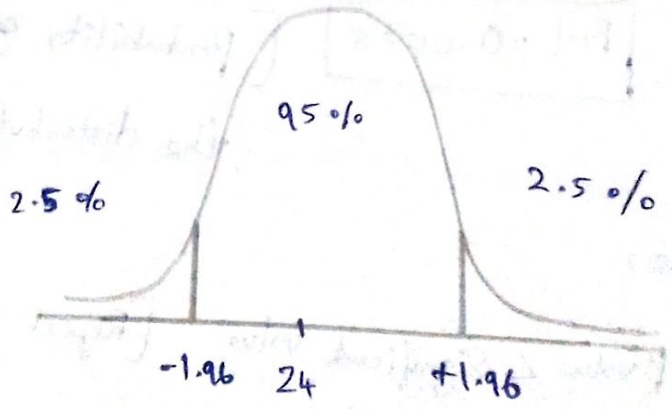
(iii) Significance Value = 0.05

(iv) State Decision boundary,

$1 - 0.025 = 0.975$

↓ Z-table

1.96



(V) Test statistics (z-test),

$$\text{Z-Score} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{25 - 24}{1.5} \times 6$$

$= \frac{1 \times 6}{1.5} = 4$

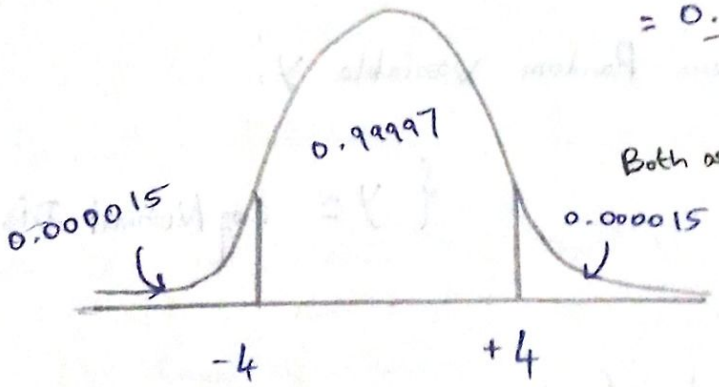
Z-score = 4

(vi) State Own decision,

$Z = 4 > 1.96$

So, reject the null hypothesis.

(vii) State Own decision with P-value,



$1 - 0.99997$

$= 0.00003$

↳ Area of one side

Both are symmetric,

$\frac{0.00003}{2}$

$= 0.000015$

↓ See Zscore value in Z table

0.99997

(Area Under the Curve)

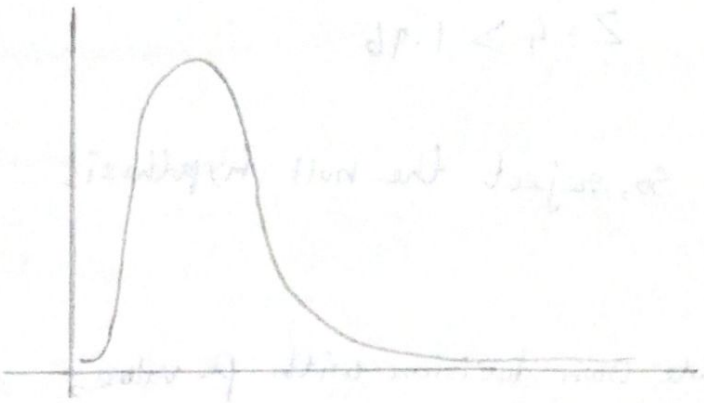
P-value $= 0.000015 + 0.000015$

$= 0.00003$

P value $\leq$ Significance Value

So, reject the null hypothesis.

Log Normal distribution:



Log Normal distribution

Suppose Random Variable y,

{ $y \simeq$ Log Normal Distribution }

Using log function,

Apply: log(y) of the Random Variable y, It Converted into Normal distribution.

Suppose we have log normal distribution, We need

to change the distribution to normal distribution for analysis and research.

$\{y \simeq$ Log Normal Distribution$\}$

$\Downarrow$

$\log(y) \xrightarrow{\text{Convert to}}$ Normal distribution.

Example:

→ Wealth distribution.

→ People writing big Comments, etc.

Bernoulli's Distribution:

Bernoulli's distribution is focus on two Outcomes One is Success and another one is failure in the term of probability.

Tossing a Coin,

$P(H) = 0.5$

So Consider,

$p = 0.5$ [Probability of Success] (or) [One change of outcome]

$q = 1 - p$

$q = 1 - 0.5$

$\boxed{q = 0.5}$ Probability of failure (or) other change of outcome

Suppose, we do not a fair coin, [one-time]

$\boxed{P(H) = 0.3}$

$P(T) = 1 - P(H)$

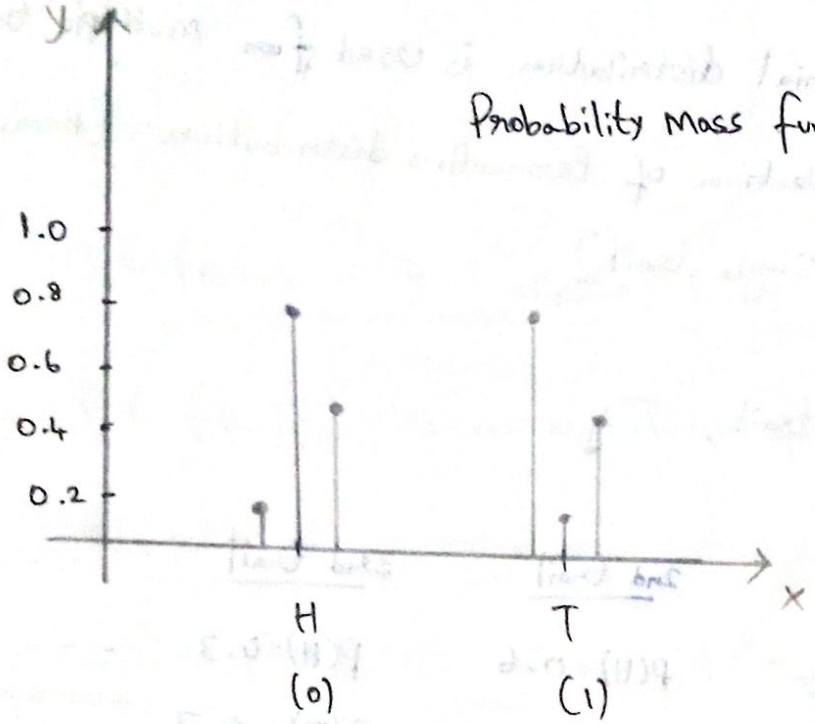
$= 1 - 0.3$

$\boxed{P(T) = 0.7}$, These type of scenario, we use the bernoulli's distribution.

The Bernoulli distribution discovered by Jacob Bernoulli and this distribution is for discrete variables. So, it is called as discrete probability distribution of a random variable.

This distribution ensures the PMF (probability mass function) because it is discrete supported One and for Categorical variable. But pdf are used in Continuous variable. (i.e) Normal distribution.

Probability Mass function of bernoulli's



$P(x=0) = 0.2$ and $P(x=1) = 0.8$

$P(x=0) = 0.8$ and $P(x=1) = 0.2$

$P(x=0) = 0.5$ and $P(x=1) = 0.5$

Parameter, $0 \leq P \leq 1$

$q = 1 - P$

formula,

Pmf,	CDF,
$\begin{cases} q = 1-P & \text{if } k=0 \\ P & \text{if } k=1 \end{cases}$	$\begin{cases} 0 & \text{if } k<0 \\ 1-P & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$
$P^k(1-P)^{1-k}$	

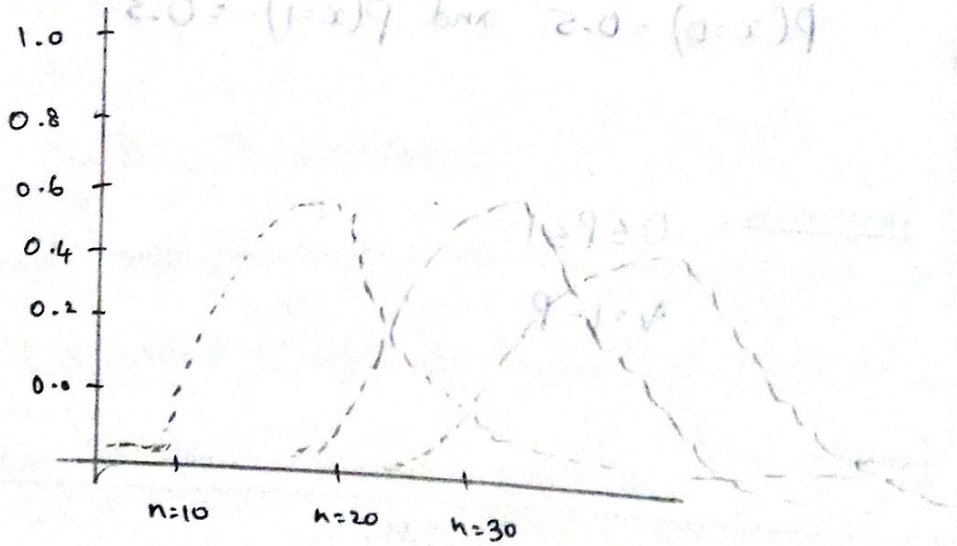
Binomial Distribution:

Binomial distribution is used for multiple trails. It is a collection of Bernaulli's distribution. [Because it use only single trail].

Multiple trail, ↘

1st trail	2nd trail	3rd trail	. . .
P(H) = 0.5	P(H) = 0.6	P(H) = 0.3	. . .
P(T) = 0.5	P(T) = 0.4	P(T) = 0.7	. . .

Probability Mass function:



p = 0.5 and n = 10

p = 0.5 and n = 20

p = 0.4 and n = 30

Cumulative Mass function graph based on the cumulative sum of the each sample.

Parameter,

$n \in \{0, 1, 2, \dots\}$ number of trails

$p \in \{0, 1\}$ → Success probability for each trail

$q = 1 - p$

formula,

PMF

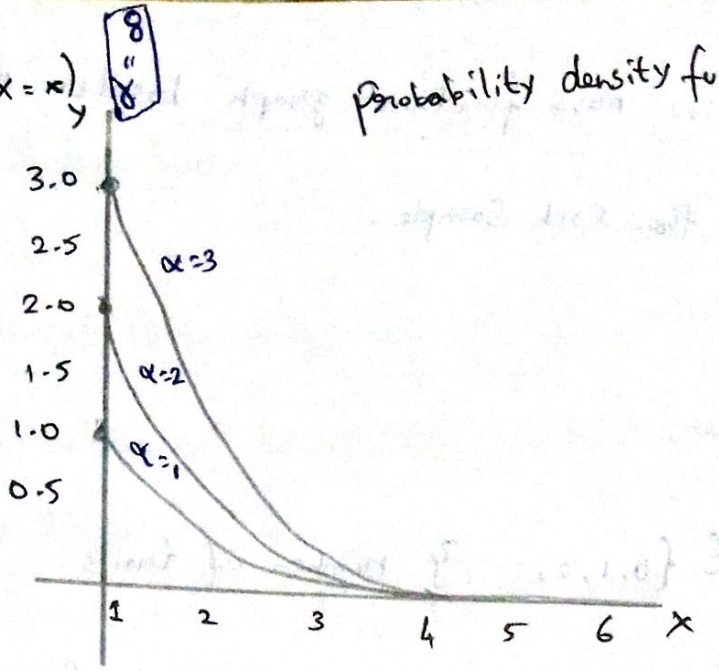
$$\binom{n}{k} p^k q^{n-k}$$

CDF,

Pareto Distribution: [Non-Gaussian distribution]

→ Pareto distribution is a power-law probability distribution that is used in description of social, quality, control, scientific and many type of observable phenomena.

The structure would be an,



probability density function

This structure of distribution is called as the Pareto distribution.

The 80 - 20 rule is applicable for Pareto distribution

Parameter,

$x_m > 0$ Scale (real)

$\alpha > 0$ Shape (real)

formula,

PDF

$$\frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$$

CDF

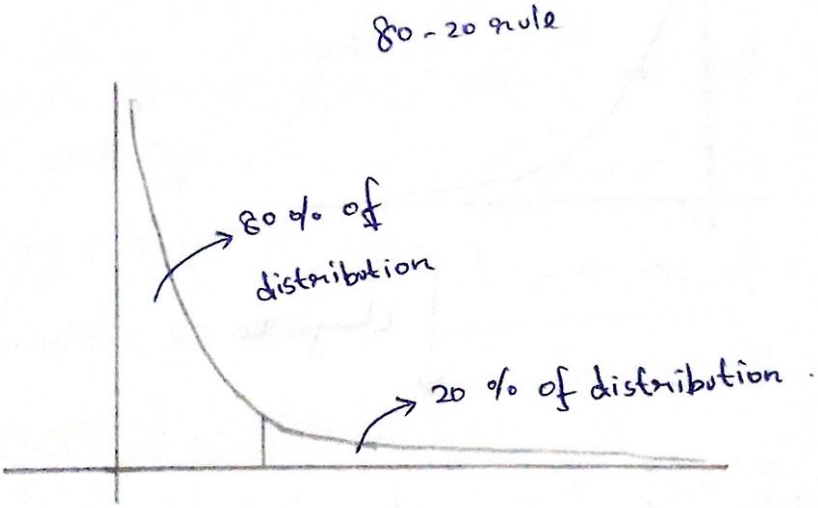
$$1 - \left(\frac{x_m}{x}\right)^{\alpha}$$

Power-law Distribution: [Application of Pareto distribution]

→ Power law is a functional relationship between two quantities.

→ It is the application of Pareto distribution.

The structure would be,



Example,

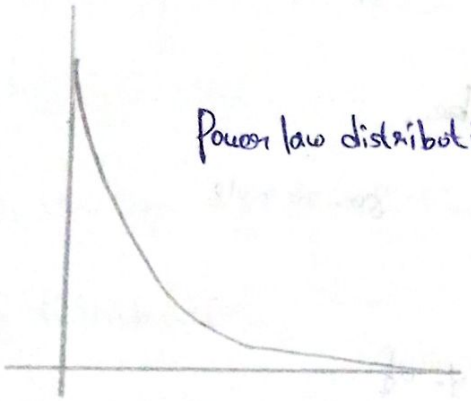
→ 80 % of the wealth is distributed with 20 % of the people.

→ 80% of the company project are done by 20% of the people in a team.

→ 80% of sales is done by the 20 % of the famous product.

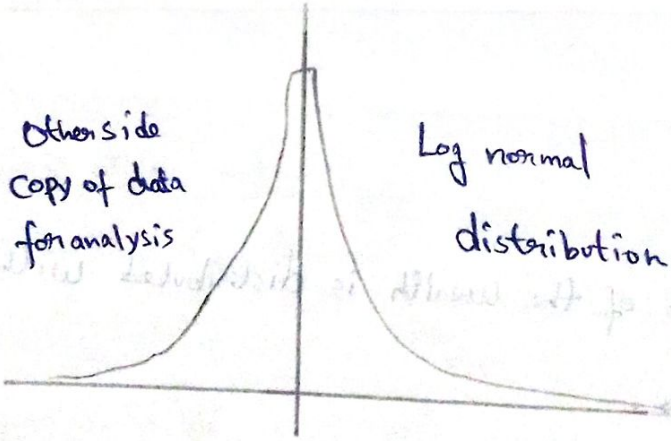
This is also called as Power law (or) Pareto distribution.

One Amazing thing is that we can change this distribution to Normal distribution for analysis.



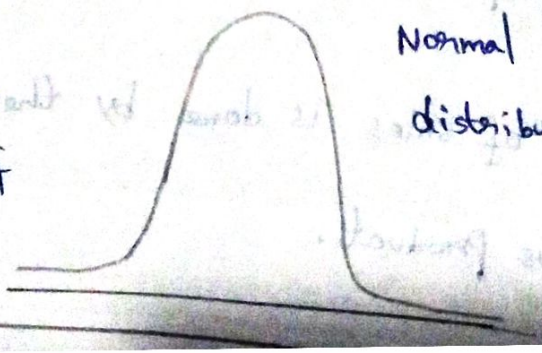
Change to

Other side copy of data for analysis



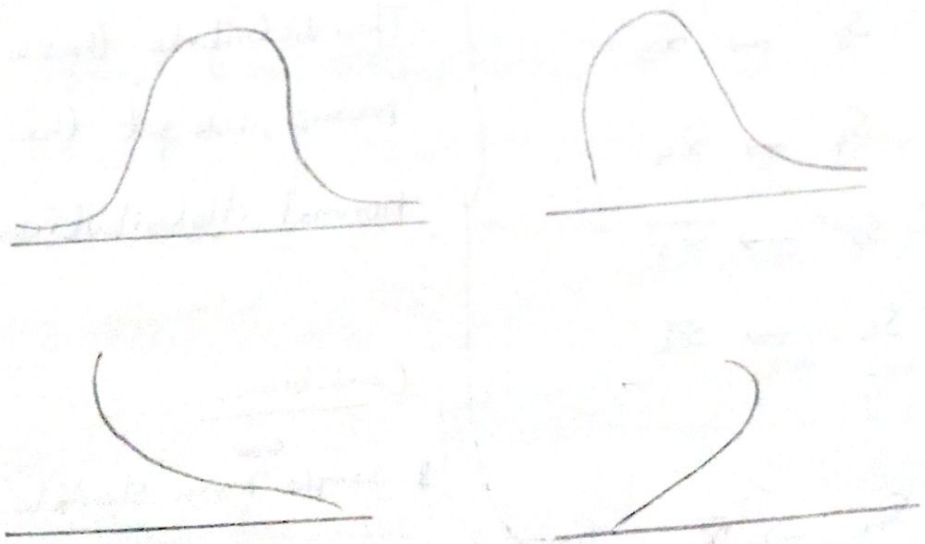
Using log function ↓ Change to

This is called as Transformation of data.



Central limit theorem:

Central limit theorem Says that, In any distribution of data take an multiple Samples ^Mean with Sample Size is greater than 30. Then Say that the and Convert to normal distribution for analysis.



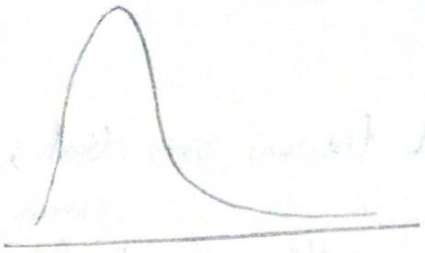
ANY DISTRIBUTION

↓ CONVERT TO [with the help of taking samples Mean]

Normal distribution

Example:

Consider the any type of distribution,



⇓

Take some samples

$S_1 \xrightarrow{\text{Find mean}} \bar{x}_1$

$S_2 \xrightarrow{\text{Find mean}} \bar{x}_2$

$S_3 \rightarrow \bar{x}_3$

$S_4 \rightarrow \bar{x}_4$

$S_5 \rightarrow \bar{x}_5$

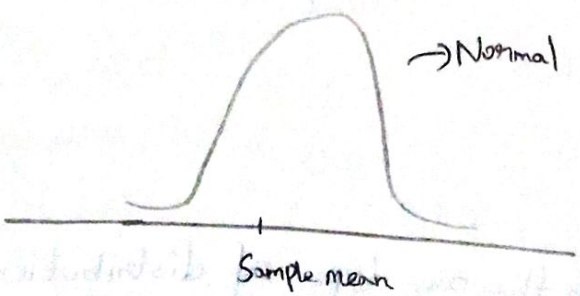
$S_6 \rightarrow \bar{x}_6$

$S_m \rightarrow \bar{x}_m$

Then distribute these sample means, we get the normal distribution.

Conditions:

1. Sample size (n) mean should be $n \geq 30$
2. Samples are allowed any multiple samples. [Higher samples → Good result].



→ Normal distribution.

Poisson distribution:

→ Poisson distribution is a discrete probability distribution for the counts of events that occur randomly over a given period.

→ Many experimental situations occur in which we observe the counts of events within the set units of time, area, volume etc.

→ Discovered by French mathematician Simeon Denis Poisson in 1837.

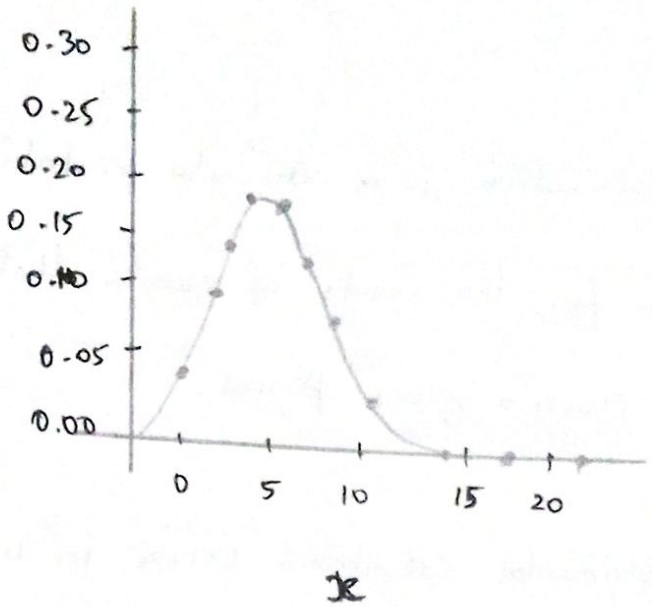
Note,

Poisson distribution is used in cases where the chances of any individual event being a success is very small.

Eg:

The number of plane crash in india in one year.

P(X=x) Probability Mass function $\boxed{\lambda = 4}$



Parameter, $\lambda \in (0, \infty)$ (rate)

Formula,

The probability of observing x events in a given interval is given by,

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad (x = 0, 1, 2 \ldots)$$

e value = 2.7182

It is also a kind of Pareto distribution.

Example Problem:

1. On average Cancer kills five people each year in india, $\lambda = 5$. What is the probability that only one person is

killed this year?

Assuming these are independent random events, the number of people killed in a given year therefore has a Poisson distribution.

$\boxed{\lambda = 5}$

Let x be the number of people killed in a year,

$$P(x=k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$P(x=1) = \frac{e^{-5}5^1}{1!} \approx 0.033$$

↑ One person Probability

2. Suppose the trains arrive at the railway station with an average arrival rate of 4 trains per hour. What is the probability exactly 6 trains will arrive in a two hour period?

$\boxed{\lambda = 4}$ $\boxed{k = 6}$

$$P(x=6) = \frac{2.7182^{(-4)}\,4^6}{6!} = 0.104067$$

$P(x=6) \simeq 0.104067$ (or) 10.41% .

Uniform distribution:

A uniform distribution refers to the probability distribution of discrete and continuous variables. It is applicable for Pdf and Pmf. It describes in which all values within a given range are equally likely to occur.

The same probability of being observed.

Example:

Rolling a fair Six-Sidded Die

The die is unbiased so all six sides have the same chance of landing face up. Each outcome (1,2,3,4,5,6) is equally likely.

In this case,

$P(x=1) = 1/6$

$P(x=2) = 1/6$

$P(x=3) = 1/6$

$P(x=4) = 1/6$

$P(x=5) = 1/6$

$P(x=6) = 1/6$

Another Example:

Random Number Generation.

Random number Generation has equally chance to create (0 to 1) - i.e $P(0 \leq X \leq 1) = 1$.

Formula,

PMF, $Pmf = \frac{1}{n}$

PDF, $Pdf = \begin{cases} \frac{1}{b-a} & ; \text{for } x \in [a,b] \\ 0 & ; \text{otherwise} \end{cases}$

Exponential Distribution:

→ The exponential distribution is a Continuous probability distribution that models the time between events in a poisson process.

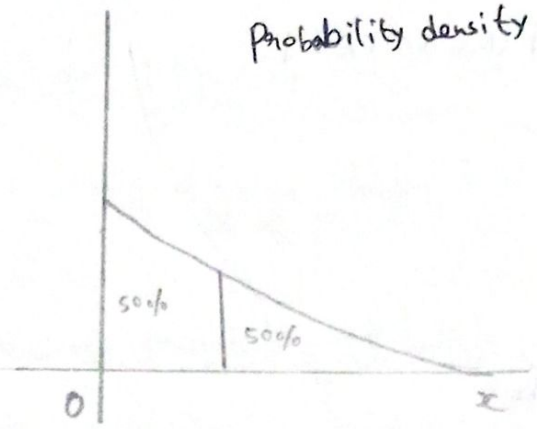
→ It is Commonly used to describe the time between Occurence of Continuous and memory less event.

Pdf,

$$f(x|\lambda) = \lambda e^{(-\lambda x)}$$

Cdf,

$$F(x/\lambda) = 1 - e^{(-\lambda x)}$$



Interarrival time:

It refers to the time elapsed between Consecutive Occurence of events. It is the time gap or duration between one event happening and next event Occuring.

Example: One Customer arrive at 10 A.M, Another Customer from 10 to 11 Am like that.

Memoryless Events:

The memoryless event means that the past does not affect the future events.

Example: Bus waiting, Probability of head in fair coin.

Two types of problem in Exponential distribution:

Method 1: [With Interarrival time]

Time between Website pages requests.

Suppose you are an IT analyst monitoring the website traffic for a ~~popular~~ popular Online store. You record the time stamps of user page requests and find the following time Interval (Interarrival time) between Consecutive Page requests. [0.2, 0.3, 0.1, 0.4, 0.2, 0.3, 0.5, 0.2, 0.3, 0.4]. So, what is the probability of next page request will occur within 5 seconds?

Soln:

Step1: Calculate the average interarrival time.

$$\text{Average Interarrival time} = \frac{0.2+0.3+0.1+0.4+0.2+0.3+0.5+0.2+0.3+0.4}{10}$$

$$= 0.3 \text{ Seconds}$$

Step 2: Fit the exponential distribution.

Fit the exp dist, 0.3 seconds as the rate Parameter.

Step 3: Probability Calculation

Why use cdf formula?
because the value of time ($x_{(0.3)}$) is less than $x_{(0.5)}$.
So, use cdf formula, otherwise use pdf formula.

Find $P(x \leq 0.5) = 1 - e^{(-0.3 * 0.5)}$

$= 1 - e^{(-0.15)}$

$= 1 - 0.8607$

≈ 0.1393

$P(x \leq 0.5) \approx 0.1393$ (or) 13.93%

Method 2: [without interarrival time]

Suppose you are IT analyst monitoring traffic for popular online store. You observe the time between Consecutive Page requests made by users on the website, and you find that the time between page requests follows an exponential distribution with an average rate of 0.1 requests per second. ($\lambda = 0.1$ request per second)

What is the probability that the next page request will occur within 5 seconds?

Directly give $\lambda = 0.1$

Step 1: Convert the time to same unit of rate parameter.

$\lambda = 0.1$ (Rate parameter)

$x = 0.5$ seconds

Step 2: Calculate probability,

$$P(x \leq 5) = 1 - e^{(-0.1 * 0.5)}$$

$$= 1 - e^{(-0.05)}$$

$$= 1 - (0.951229424 5)$$

$$= 0.0488$$

$P(x \leq 5) = 0.0488$ (or) 4.8%

Method 1 Question Suppose the interarrival time is 0.6 and Say probability of next page request within 0.5 seconds,

[Use pdf]

So, $\lambda = 0.6$ [Rate parameter]

$x = 0.5$ Sec.

Calculate,

$$\text{Pdf}, f(x/\lambda) = \lambda e^{(-\lambda x)}$$

$$f(x|0.6) = 0.6 * e^{(-0.6x)}$$

$$P(x \leq 0.5) = \int_{[0,0.5]} f(x|0.6)\, dx$$

$$P(x \leq 0.5) = \int_{0}^{0.5} [0,0.5]\ (0.6 * e^{(-0.6x)})\, dx$$

$$P(x \leq 0.5) = \left[-e^{(-0.6x)}\right] \text{ from } 0 \text{ to } 0.5$$

$$P(x \leq 0.5) = \left[-e^{(-0.6 * 0.5)}\right] - \left[-e^{(0)}\right]$$

$$P(x \leq 0.5) = [-0.30119] - [-1]$$

$$P(x \leq 0.5) \approx 0.69881$$

The probability of the next page request will occur within 0.5 seconds is 0.69881 (or) 69.88%.

- (or) this problem also can be solved with the help of complementary of cdf. [CCDF].

ANNOVA TEST: [F-Test]

Analysis of Variance (ANOVA) is a statistical test for detecting difference in the group means when there is One parametric dependent variable and One or more independent variable.

Types:

One-way Anova:

One way Anova is used when you have one independent variable (factor) that divides your data into two or more groups.

Two way Anova:

Two way Anova is used when you have two independent variables (factors) that divides your data into different groups.

In Annova, F-statistics will be used.

Formula, [ONE WAY ANOVA]

$$F = MSB / MSW$$

MSB → Mean Square Between, Calculated as the Sum of Squares between groups (SSB) divided by the degree of freedom.

$$MSB = SSB / dfB$$

SSB → Sum of squares between, the variability between the means of the group.

dfB → Degree of freedom between, the degrees of freedom associated with between-group variability.

$$MSW = SSW / dfW$$

SSW → Sum of squares within, the variability within each group.

dfW → Degrees of freedom within, the degrees of freedom associated with group variability.

$SSB = \sum (\text{Group Mean} - \text{Grand Mean})^2 * (\text{Number of Observation in Group})$

$dfB = \text{Number of Groups} - 1$

$SSW = \sum (\text{observation} - \text{Group Mean})^2$

$dfw = \text{Total number of Observation} - \text{Number of Groups}.$

Formula, [Two way ANOVA]

$$F = (SSAB / dfAB) / (ssw / dfw)$$

↳ Both Combine factor formula.

and

$$F = (SSA / dfA) / (ssw / dfw)$$

↳ one factor

Example for One way and Two way Anova:

One-way Anova:

1. Scenario:

Imagine you work as a quality Control Manager in a factory that produces three different types of Cookies: A, B and C. You want to determine if there are any Significant differences in the average diameter of the Cookies produced by these three types.

Data: you randomly Sample 15 cookies from each type (A, B and C) and measure their diameters. The diameter measurements are within 8.0 to 9.0.

Soln:

Collect Samples,

Type A: 8.3, 8.4, 8.7, 8.9, 8.2, 8.8, 8.6, 8.4, 8.5, 8.7, 8.1, 8.3, 8.6, 8.4.

Type B: 8.5, 8.7, 8.3, 8.6, 8.4, 8.2, 8.8, 8.4, 8.9, 8.6, 8.7, 8.5, 8.3, 8.4, 8.6.

Type C: 8.8, 8.4, 8.9, 8.5, 8.7, 8.3, 8.6, 8.5, 8.7, 8.1, 8.2, 8.4, 8.6, 8.4, 8.5

Null Hypothesis (H_0):

There is no significant difference in the average diameter of cookies produced by type A, B and C.

Alternate Hypothesis (H_1):

There is a significant difference in the average diameter of cookies produced by at least one of the types A, B and C.

Calculations:

Step 1: Calculate mean for each type.

Mean of Type A $(\bar{x}_A) = \frac{(8.3 + 8.4 + 8.7 \ldots + 8.6 + 8.4)}{15}$

$$\boxed{\bar{x}_A = 8.48}$$

Mean of Type B $(\bar{x}_B) = \frac{(8.5 + 8.7 + 8.3 + \ldots + 8.4 + 8.6)}{15}$

$$\boxed{\bar{x}_B = 8.48}$$

Mean of Type C $(\bar{x}_C) = \frac{(8.8 + 8.4 + 8.9 + \ldots + 8.4 + 8.5)}{15}$

$$\boxed{\bar{x}_C = 8.52}$$

Step2:- Calculate Overall mean ($\bar{x}$ - overall) and total Sum of Squares (SS_total).

$$\bar{x}\ \text{overall} = \frac{(\bar{x}\text{-}A + \bar{x}\text{-}B + \bar{x}\text{-}C)}{3} = \frac{8\text{-}48 + 8\text{-}48 + 8\text{-}52}{3}$$

$$\boxed{\bar{x}\ \text{overall} = 8\text{-}4933}$$

$$SS\_total = \sum (xi - \bar{x}\_overall)^2$$

↳ Each data in the three groups.

$$SS\_total = (8\text{-}3 - 8\text{-}4933)^2 + (8\text{-}4 - 8\text{-}4933)^2 + \ldots$$

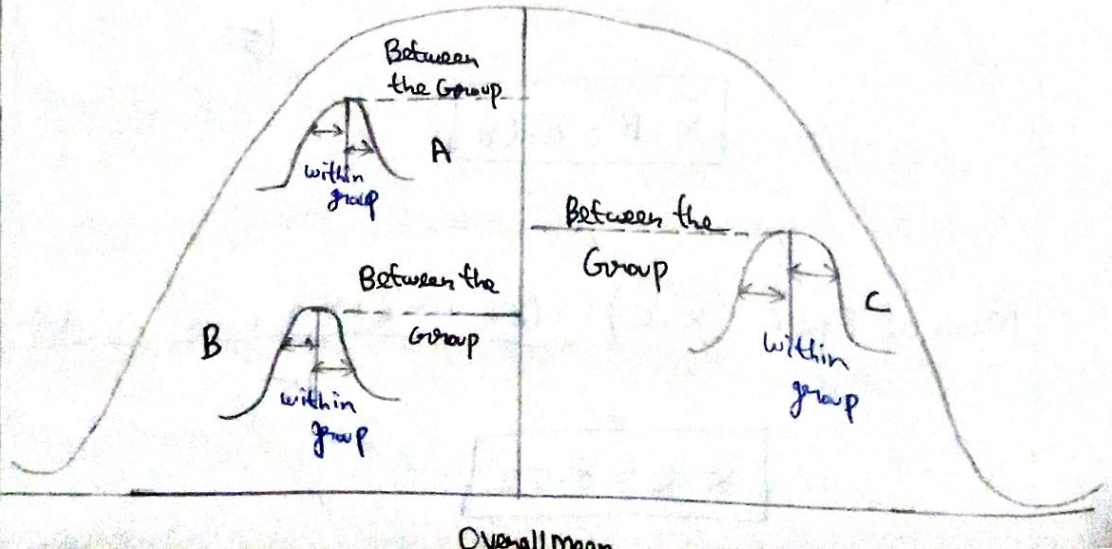
$$+ (8\text{-}4 - 8\text{-}4933)^2 + (8\text{-}5 - 8\text{-}4933)^2$$

Step3:-

Calculate the Sum of Squares between groups (SS_between).

↓ Nothing but

Variability between the group.



$$SS\_between = \sum (\text{no of observation}) * (\text{Group mean} - \text{Grand mean})$$

$$SS\_between = nA * (\bar{x}\_A - \bar{x}\_overall)^2 + nB * (\bar{x}\_B - \bar{x}\ overall)^2$$

$$+ nC * (\bar{x}\_C - \bar{x}\ overall)^2$$

nA, nB, nC → no of observation

Sample (n=15)

$$= 15 * (8.48 - 8.4933)^2 + 15 * (8.48 - 8.4933)^2 +$$

$$15 * (8.52 - 8.4933)^2$$

$$= 0.0544$$

SS_between = 0.0544

Step 4:

Calculate the sum of Squares within groups (SS_within)

↳ variability within the group.

$$SS\_within = \sum (x_i - \bar{x}\_i)^2$$

$$SS\_within = \sum (\text{For each group data points} - \text{For each respective Group mean})^2$$

For Type A,

$$\bar{x}\_A = 8.48$$

$$SS\_within\_A = (8.3 - 8.48)^2 + (8.4 - 8.48)^2 + \ldots + (8.6 - 8.48)^2$$

$$+ (8.4 - 8.48)^2$$

SS_within_A = 0.0328

Same like,

SS_within_B = 0.1064

SS_within_C = 0.2144

Steps: Calculate degree of freedom.

df_between = k-1, where k is the no of groups.

df_within = N-k, where N is the total no of measurements.

[i.e group A=15, group B=15, group C=15]

N=45

df between = 3-1=2

df within = 45-3=42

Step 6: Calculate MSB and MSW

MSB → Mean Squares between

MS_between = SSB / dfB = SS_between / degree of freedom between

= 0.0544 / 2 = 0.0272

Ms_between = 0.0272

$$MS\_within = \frac{SSW}{dfw} = \frac{SS\text{-}within}{degree\ of\ freedom\ within}$$

$$MS\_within\_A = \frac{0.0328}{14}$$

$$\boxed{MS\_within\_A \approx 0.002343}$$

Same like,

$$\boxed{MS\_within\_B \approx 0.007600}$$

$$\boxed{MS\_within\_C \approx 0.015314}$$

Step7: Calculate the F-statistic

$$F = \frac{MSB}{MSW}$$

Type A,

$$MSB = 0.0272$$

$$MSW = 0.002343$$

$$f\text{-}stat\_A = \frac{0.0272}{0.002343} \approx 11.588$$

$$\boxed{f\text{-}stat\_A \approx 11.588}$$

Type B,

MSB = 0.0272

MSW = 0.007600

F-stat_B ≈ 3.579

Type C,

MSB = 0.0272

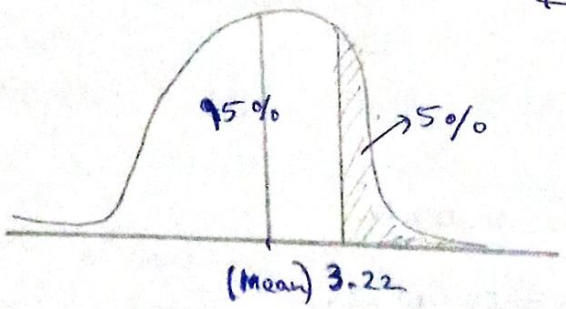
MSW = 0.015314

F-stat_C ≈ 1.776

Step-8:

One-tail because [Is any significant difference or not]

$\alpha = 0.05$



1 - 0.05
= 0.95

See F-Test table with first and second degree of freedom,

df between = 2 (See Second column in table)

df within = 42 (See 42nd row in table)

$(2,42) = 3.22$

Step 9: [Decision]

~~Type B:~~

F-stat_A $\approx$ 11.588 $>$ 3.22 [Reject]

F-stat_B $\approx$ 3.579 $>$ 3.22 [Reject]

F-stat_C $\approx$ 1.776 $<$ 3.22 [Accept]

The Two groups are rejected and only one group is accepted. Therefore Null hypothesis are rejected and Alternate hypothesis are accepted.

Two-way Anova:

It is Quite large calculation. The calculation is much Similar to one-way Anova but the factors difference.

Calculation step:

Scenario:

Imagine you are Conducting experiment to investigate two factors, "Type of fertilizer" and "Amount of water" on growth of Plants. you have three types of fertilizers (A, B and C) and three levels of water (low, medium and

High). you measure the height of the plants after a certain period to determine if there are any significant difference due to two factors.

fertilizer/water	Low	medium	High
A	15	20	25
B	18	22	26
C	12	19	24

Step 1:-

Mean of each type of Sample.

TypeA_fert mean, TypeB_fert mean, TypeC_fert mean

TypeA_water mean, TypeB_water mean, TypeC_water mean

Step2:- Overall Mean

$\bar{x}_{grand} = \frac{15+20+25+18+22+26+12+19+24}{9} = 20.4444$

Step3:- SS_total = 107.1111

Step4:- SS_between Using formula.

Step5:- SS_within Using formula

Step6:- degree of freedom Calculate

Step7:- MSB and MSW

Step8:- F-statistic

Note: Same like apply water factor every step.

- SS_between water
- SS_between fertilizer like this for all.

[df_interaction] is some very.

[Refer chatgpt]

CHEBYSHEV'S INEQUALITY:

Let Consider,

$X \approx$ follows Gaussian distribution (μ, σ).

(Random Variable)

Let Say, Empirical formula for data distribution,

$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 68\%$

$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 95\%$

$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 99.7\%$.

Suppose,

$Y \approx$ Does not follow Gaussian distribution.

(Random variable)

How to find distribution respect first, second and third Sd? that why Chebyshev's inequality Come.

Formula, $P(\mu - k\sigma < \text{Random variable} < \mu + k\sigma) \geq \boxed{1 - \frac{1}{k^2}}$

$k \rightarrow$ Range of standard deviation.

We want to find range = 2,

$\boxed{k=2}$

$$P(\mu - 2\sigma < y < \mu + 2\sigma) > 1 - \frac{1}{2^2}$$

$$\therefore 1 - \frac{1}{4} = \frac{3}{4}$$

$$= 75\%$$

The Second Standard deviation of random variable y Contains the 75% of total distribution of data.

Some Assumption for Each test: [JUST NOTES ALREADY DISCUSSED]

Z-TEST	T-TEST	CHI-SQUARE TEST	ANNOVA TEST
→ Random Sampling. → Normality (data follow GD). → Population SD is known. → Sample Size $n > 30$ is Produce Super result.	ONE Sample, → Follow GD. → Random Sample from population. → Data points are independent. Two Sample, → follow GD. → Random Sample. → Homoscedasticity.	→ Independence [data is independence to each other] → Sample Size at least 5. → Random Sampling. → Data should be Categorical.	→ Independence between each group. → Normality [data follow GD] → Homogenity of variance (Homoscedasticity) If means two group variance is approx similar.