Probability

- An event is a subset of sample space.
- if $A \cap B = \emptyset$ then A & B are mutually exclusive or disjoint sets.
- The no of permutations of n distinct objects taken r at a time is ${}^nP_r = \frac{n!}{(n-r)!}$
- The no of combinations of n distinct objects taken r at a time is ${}^nC_r = \frac{n!}{r!(n-r)!}$ or denoted by $\binom{n}{r}$.
- $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$ → Additive rule.

  and if mutually exclusive then $P(A \cup B) = P(A) + P(B)$ bcoz $P(A \cap B) = \emptyset$.
- $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
- $P(A) + P(A') = 1$ Complement rule
- $P(B|A) = \frac{P(A \cap B)}{P(A)}$ where $P(A) > 0$
- $P(\bar{A} \cup \bar{B}) = P(\bar{A}) + P(\bar{B}) - P(\bar{A}\bar{B})$ →
- $P(B|A) = P(B)$ or $P(A|B) = P(A)$ → independent event
- $P(A \cap B) = P(A)\,P(B|A)$
- $P(A \cap B) = P(A)\,P(B)$ → independent event

$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|(A \cap B))$

$P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B)$

Bayes theo:- $P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)}$

$= \frac{P(B_r)\,P(A|B_r)}{\sum_{i=1}^{k} P(B_i)\,P(A|B_i)}$

for $r = 1, 2, \ldots, k$

$$\text{Mean } \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

if data is in increased order

$$\bar{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{if } n \text{ is even} \end{cases}$$

Sample variance denoted by s^2,

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Sample standard deviation $\Rightarrow s = \sqrt{s^2} = \sqrt{\text{Variance}}$

Median = average of $\left(\frac{n}{2}\right)$ & $\left(\frac{n}{2}+1\right)^{th}$ value

$$= \begin{cases} \frac{n}{2}\left(\frac{n}{2}+1\right) & \text{for } n \text{ even} \\ \frac{n+1}{2} & \text{for } n \text{ odd} \end{cases}$$

Trimmed mean :- Suppose 50 data is given and you have to calculate 10% trimmed mean $\rightarrow \frac{10}{100} \times 50 \Rightarrow 5$

∴ Remove 5 largest values and remove 5 smallest values

Range = max val - min val

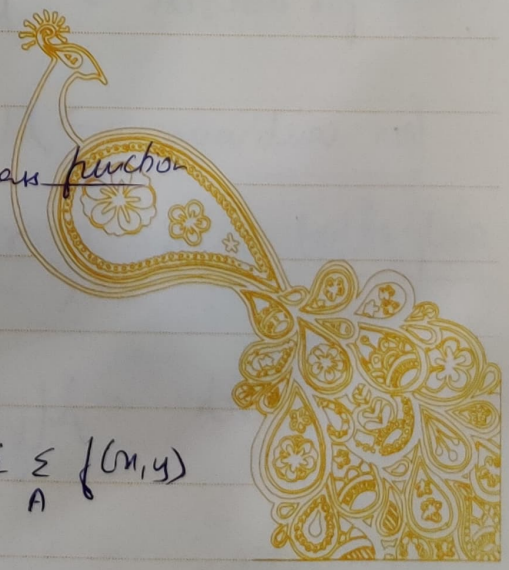
Joint probability distribution or probability mass function

① $f(x,y) \geq 0$ for all (x,y)

② $\sum_x \sum_y f(x,y) = 1$

③ $P(X=x, Y=y) = f(x,y)$

For any region A in x-y plane $P[(x,y) \in A] = \sum_A \sum f(x,y)$



Joint density function

① $f(x,y)$ is jdf of continuous random variables x & y if
i) $f(x,y) \geq 0$ for all (x,y)
ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

② $P[(x,y)\in A] = \iint_A f(x,y)\,dx\,dy$

Let $f(x,y)$

Marginal distributions of x alone & y alone

$g(x) = \sum_y f(x,y)$ and $h(y) = \sum_x f(x,y)$ for discrete case and

$g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$ & $h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$ (for continuous case)

Conditional distribution

$f(y|x) = \frac{f(x,y)}{g(x)}$, $g(x)>0$ & $f(x|y) = \frac{f(x,y)}{h(y)}$, $h(y)>0$

$P(a<x<b|y=y) = \sum_{a<x<b} f(x|y)$ $\quad P(a<x<b|y=y) = \int_a^b f(x|y)\,dx$

~~Statistical independence~~

Mathematical Expectation

Let x → random variable with probability distribution $f(x)$.

mean or expected value :—

for discrete → $\mu = E(x) = \sum_x x f(x)$ or $\sum_x \sum_y x f(x,y)$ then use marginal

for continuous → $\mu = E(x) = \int_{-\infty}^{\infty} x f(x)\,dx$

expected value of random variable—

$\sum_x g(x)$ $(g(x))$

(for discrete) $\mu_{g(x)} = E[g(x)] = \sum_x g(x) f(x)$

(for continuous) $= \int_{-\infty}^{\infty} g(x) f(x)\,dx$

Correlation coeff

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Let X & Y be random variable with joint Prob dist $f(x,y)$.

The mean/expected of random variable $g(x,y)$ is

for discrete → $\mu_{g(x,y)} = E[g(x,y)] = \sum_x \sum_y g(x,y) f(x,y)$

for continuous ⇒ " " " = $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\, dx\, dy$

Variance:- for discrete → $\sigma^2 = E[(x-\mu)^2] = \sum_x (x-\mu)^2 f(x)$

for continuous ⇒ " " = $\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$

▷ for joint prob $E[(x-\mu_x)(x-\mu_y)]$

▷ variance of random variable X is $\sigma^2 = E(X^2) - \mu^2$ → $\sigma_x^2 = E(x^2) - \mu_x^2$

▷ covariance → $\sigma_{xy} = E(xy) - \mu_x \mu_y$

▷ $\int_{-\infty}^{\infty} (g(x) - \mu_{g(x)})^2 f(x)\, dx$

Discrete Probability Distributions

Binomial distribution :- ${}^nC_r\, p^r q^{n-r}$ (r or x)

Bernoulli trial

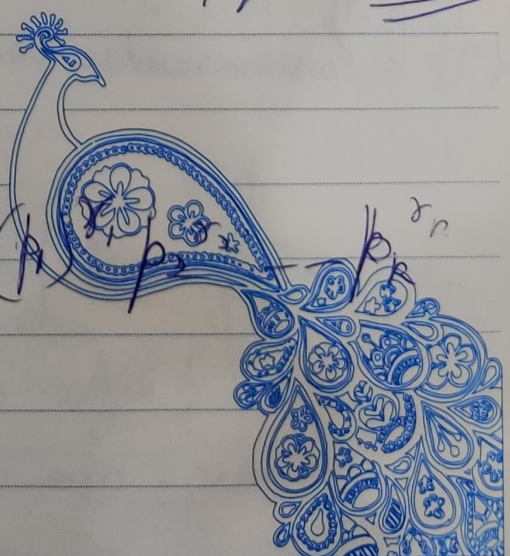
or $b(r; n, p)$

The mean and the variance of binomial distribution $b(r; n, p)$ are $\mu = np$ and $\sigma^2 = npq$ (or $np(1-p)$)

Multinomial distribution :- ✗

$$f(r_1, r_2 \dots r_k; p_1, p_2 \dots p_k, n) = \binom{n}{r_1 r_2 \dots r_k} p_1^{r_1} p_2^{r_2} \dots p_k^{r_k}$$

and $\sum_{i=1}^{k} r_i = n$ & $\sum_{i=1}^{k} p_i = 1$



Hypergeometric Distribution :-

Hypergeometric random variable x, the number of successes in a random sample of size n selected from N items of which k are labelled success and $N-k$ labeled failure, is

$$h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad \max\{0, n-(N-k)\} \le x \le \min\{n, k\}$$

Both the

The mean and the variance of the hypergeometric distribution $h(x; N, n, k)$ are

$$\mu = \frac{nk}{N} \quad \text{and} \quad \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right)$$

$$\mu = np = \frac{nk}{N} \quad \text{and} \quad \sigma^2 = npq = n \cdot \frac{k}{N}\left(1 - \frac{k}{N}\right)$$

$\frac{N-n}{N-1}$ is negligible when n is small relative to N.

Multivariate hypergeometric distribution :-

Poisson ratio Distribution :-

$\mu = \left(\frac{q}{p}\right)$ of $\left(\frac{1}{p} - 1\right)$ Poisson random variable $x \to$ no. of outcomes occuring in a given time interval or specific region denoted by t is

or

$$\frac{e^{\mu}(\mu)^n}{x!} \Rightarrow \mu = np \quad p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots$$

where λ is the average no. of outcomes per unit time, distance, area or volume and $e = 2.71828\ldots$

Both the mean and variance of poisson distribution $p(x; \lambda t)$ is λt.

Continous probability distribution

Uniform distribution:-

The density function of the continous uniform random variable x on interval $[A, B]$ is

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \le x \le B \\ 0, & \text{elsewhere} \end{cases}$$

Mean & Variance of uniform distribution is $\mu = \frac{A+B}{2}$ and $\sigma^2 = \frac{(B-A)^2}{12}$

Normal distribution :-

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty \le x \le \infty$$

$\pi = 3.14159$ and $e = 2.71828$

Mean = np

Standard deviation = $\sigma = \sqrt{npq}$

Area under curve

$\therefore Z = \frac{X-\mu}{\sigma}$ $\therefore$ Z is normal random variable with mean 0 & var ± 1

Normal approximation :-

If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$ then the limiting form of distribution of

$$Z = \frac{X - np}{\sqrt{npq}}$$

as $n \to \infty$, is the standard normal distribution

Central limit theorem

If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$

$$\frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ → SD of observed pop

↓

SD of mean for sampling

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

Standard error of Means :- SEM :- $\frac{\sigma}{\sqrt{n}}$

Interquartile

quartiles

Level of confidence =

$100(1-\alpha)$ = Say 75

$\alpha = 0.05$

Range of $|\bar{x} - \mu|$

$|\bar{x} - \mu| \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Max error $= z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Sampling Distribution

Sample mean :- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Sample Median :- $\begin{cases} x_{(n+1)/2} & \rightarrow n \text{ odd} \\ \frac{1}{2}\left(x_{n/2} + x_{(n/2+1)}\right) & \rightarrow n \text{ is even} \end{cases}$

Sample variance :- $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

$$s^2 = \frac{1}{n(n-1)} \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right]$$

Range $= x_{max} - x_{min}$

mean $= \mu$ var $= \frac{\sigma^2}{n}$

(1) When σ known $\rightarrow \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

(2) when σ unknown $\rightarrow$ pop follows normal dis

t-distribution is ~~after~~ at last

Hypothesis

H_0 → null hypothesis H_1 → alt. hypothesis → claim

Type I error :- Rejection of H_0 when its true → Prob (level of signific.) α

Type II error :- Non rejection of H_0 when its false → β

- Prob of Type I error $\propto \frac{1}{\text{Prob of Type II error}}$
- greater n, will reduce α, β

Power :- prob. of rejecting H_0 given that H_1 is true

↳ $1-\beta$

Level of confidence + level of sign = 1

Test statistics :-

$$Z = \left| \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \right| \qquad \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

$$s_p^2 = \frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1 + n_2 - 2}$$

↓ ↓

diff of two means — SD is known

↓

but SD is unknown

Conclusion

- know → z dist
- unknown → t

Conclusion of hypothesis:-

Reject H_0 or fail to reject H_0

⇒ Level of significance (α)

* standard value of α = 5%

one sample	2 samples	More than 2 samp
1) Mean (μ)	1) Diff in means	Multiple means
2) Variance (σ^2)	2) Ratio of var	ANOVA

for mean (μ):- ① critical Region approach

if $n < 30$ → t- dist^n

if $n \geq 30$ → z- dist^n

for 2 tailed hyp $\alpha = \alpha/2$

" 1 " " $= \alpha = \alpha$

$(t_{\alpha/2}$ or $z_{\alpha/2})$ or $(t_\alpha$ or $z_\alpha)$

where null hypothesis is rejected ← critical region

→ for two tail hyp

CR:- $(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$; $n < 30$

$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$; $n \geq 30$

→ for 1 tailed hyp

CR:- $(-\infty, -t_\alpha)$ OR $(-\infty, -z_\alpha)$ { for less than sign }

(t_α, ∞) or (z_α, ∞) { for greater than sign }

H_1	CR
$\mu > \mu_0$	$z > z_\alpha$
$\mu < \mu_0$	$z < -z_\alpha$
$\mu \neq \mu_0$	$z < -z_{\alpha/2}$, $z > z_{\alpha/2}$

P-value approach

if P value $< \alpha \Rightarrow$ Reject H_0
P value $\geq \alpha \Rightarrow$ fail to Reject H_0

P-value = $P(z < z^*)$ $\Rightarrow$ for left Tail test
(z^* ↓ test stats)

$P(z > z^*)$ $\Rightarrow$ for Right tail test

$= 2[P(z \leq z^*)]$ $\Rightarrow$ for 2 tails

Starting of ques: H_0, H_1 → α → CR → z^* or t^* → P value → Ans

CI → 1 sample α = significance
$1-\alpha$ = confidence

Marginal error $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$(1-\alpha)\%$ CI of $\mu = \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{ME}\right)^2$

2 sample mean

$\overbrace{\mu_1 - \mu_2}^{CI} = \bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right), \; // \; // \; + z_{\alpha/2} \; // \; //$

Inference of diff of two means

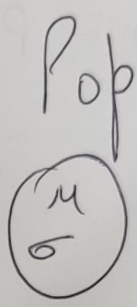
Case I :- when pop SD is known

μ_1, μ_2 , $\bar{x}_1, \bar{x}_2$, σ_1, σ_2 n_1, n_2

let $\mu_d = \mu_1 - \mu_2$

* $H_0 = \mu_0 = \mu_0$

$H_1 = \mu_d \neq \mu_0$, $< \mu_0$, $> \mu_0$



Sample
$\bar{x}$
SD

$$Z^* = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\mu_1 - \mu_0 < \mu_0$
$\mu_1 - \mu_2 > \mu_0$
$\mu_1 - \mu_2$

$$P\text{-value} = 2P(Z < -Z^*) = 0 < \alpha$$

↓

Reject H_0 in H_1

Ans is statem → Data provide sufficient evidence to reject H_0 in favor of H_1

Case II when SD in unknown

$\therefore \rightarrow \sigma_1 = \sigma_2$

Pooled Variance $= S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

$S_1^2 = \frac{1}{n_1 - 1} \Sigma (\bar{x}_i - \bar{x}_1)^2$

$\mu_1 - \mu_2 < \mu_0 / t < -t_\alpha$

$\mu_1 - \mu_2 > \mu_0 / t > t_\alpha$

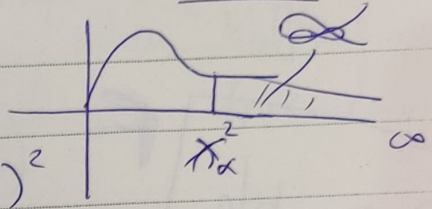
$\mu_1 - \mu_2 \neq \mu_0 / t < -t_{\alpha/2}$
$t > t_{\alpha/2}$

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_d}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Inference on Pop Variance (or SD)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$\nu = n-1$



Chi-Squared dist^n

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2}$$

Single sample

$H_0 : \sigma^2 = \sigma_0^2$

$H_1 : \sigma^2 \neq \sigma_0^2, < \sigma_0^2, > \sigma_0^2$

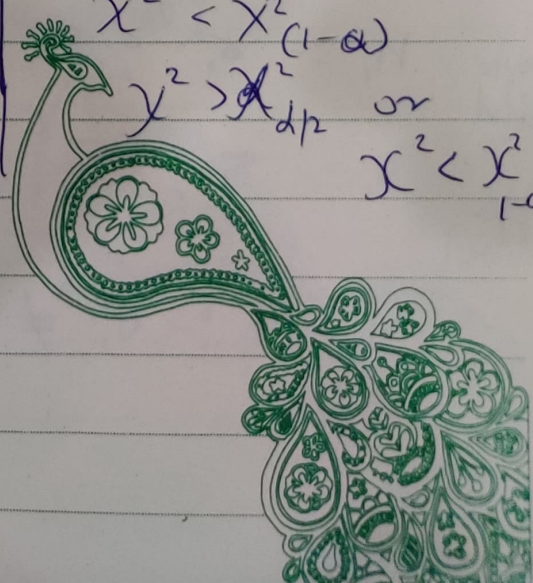
$\sigma_0 \rightarrow \chi^2_\alpha$ or $\chi^2_{\alpha/2}$

OR

Sample Summary

Test stats $\chi_0^2 = \frac{(n-1)s^2}{\sigma^2}$

	$H_0 : \sigma^2 = \sigma_0^2$	
upper tailed	$H_1 : \sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_\alpha$
lower tailed	$H_1 : \sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{(1-\alpha)}$
two tailed	$H : \sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{\alpha/2}$ or $\chi^2 < \chi^2_{1-}$



Type 1 - Type 2

Type 1 error

α = (Reject H_0 / when H_0 is true)

↓ ↓

CR mai lie krega

eg:- ($z <$ ~~ / α or prob)

Type 2 error

β = (fail to reject H_0 or reject H_1 / when H_0 is false)

↓

CR mai lie nhi karne wala area

eg ($z <$ ~~ / μ or prob)

Power = $1 - \beta$

→ 1.