* Variance & Standard Deviation ↴

Variance (Population Data)	Variance (Sample Data)
↳ $\sigma^2 = \sum_{i=1}^{N} \frac{(x-\mu)^2}{N}$	↳ $\sigma^2 = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1}$
$\boxed{std = \sqrt{\sigma^2}}$	$std = \sqrt{\sigma^2}$
	Bessel error
	d.o.f = n-1

* Covariance ↴

$$\mathrm{Cov}(x,y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

↳ $\mathrm{Cov}(x,x) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(x_i-\bar{x})}{n-1} = \boxed{\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}}$

↳ $\mathrm{Cov}(x,x) = \mathrm{Var}(x)$

Adv. of Covariance → Quantify the relationship b/w X & Y

Disadv. of Covariance ⇒ Covariance does not have a specific limit value
↳ Cov(X, Y) ⇒ $-\infty$ to $+\infty$

* Correlation
→ Pearson Corr. Coeff.
→ Spearman " "

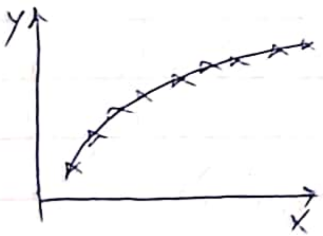
① Pearson corr. coeff. ⇒ [-1 to +1] ← limit

$$\boxed{\rho_{x,y} = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y}}$$

↳ More the value to +1 → more +ve correlated X & Y
" " " " -1 → " -ve " " "

Note:- Pearson corr. not suitable for non-linear data

② Spearman rank corr. ⇒

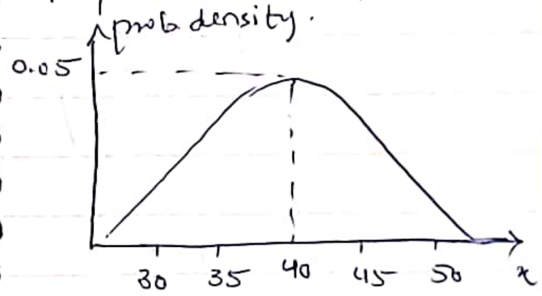


$$\boxed{r_s = \frac{Cov(R_{(x)}, R_{(y)})}{\sigma(R_{(x)}) * \sigma(R_{(y)})}}$$

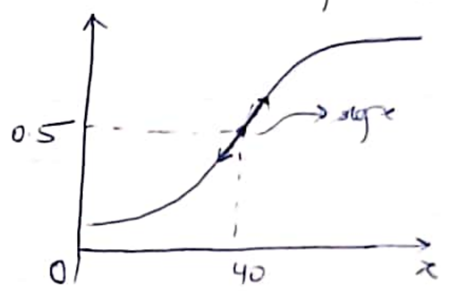
Ex:-

x	y	R(x)	R(y)
1	2	2	1
3	4	3	2
5	6	4	3
7	8	5	5
0	7	1	4

* Probab. Density Func



Cummulative probab.



① probab. Mass func" (PMF) → used for discrete random variab.

② probab. Density func" (PDF) → used for contin. random varia

③ Cummu. Distributive func"

Imp. * probab. Density ⇒ Gradient of Cummulative density func"

* p.d.f. properties →

i) Non-negative ; $f(x) \geq 0 \quad \forall x$

ii) The total area under the p.d.f curve is equal to 1

$$\boxed{\int_{-\infty}^{\infty} f(x)\, dx = 1}$$

Types of Probab. Distribution → [p.d.f, p.m.f, c.d.f]

① Bernoulli Distribution

→ outcomes are binary (p.m.f) like tossing of coins

② Binomial Distribution → p.m.f.

③ Normal / Gaussian distrib. → (p.d.f)

→ assumption

④ Poisson Distribution ⇒ (p.m.f)

⑤ Log Normal Distrib. ⇒ (p.d.f)

⑥ Uniform Distrib. (p.m.f)

Example —

Dataset → House price prediction Dataset

Size of House	No. of Rooms	Location	Floor	Sea side	Price
Continuous random variable	Discrete		Discrete	0 & 1	Continuous
pdf	pmf.		p.m.f	p.m.f	pdf

① Bernoulli Distribution → success (1), fail (0)

↳ i) Discrete Random variable (p.m.f)

ii) Outcomes are Binary

↳ P(success) → (P) so, P(lose) = (1 − P)

NOTE :- $\boxed{P.m.f = p^{k} * (1-p)^{1-k}}$

where {k = 0, 1}

0 → failure, 1 → success

Scanned with CamScanner

Also,

$$\text{p.m.f} = \begin{cases} q = 1-p & ; \text{ if } k = 0 \\ p & ; \text{ if } k = 1 \end{cases}$$

→ Mean of Bernoulli distrib.

$$\boxed{E(x) = \sum_{k=0}^{1} k \cdot p(k)}$$

→ Median of Bernoulli Distrib

$$\text{Median} = \begin{cases} 0 & \text{if } p < 1/2 \text{ (or) } q > p \\ [0, 1] & \text{if } p = 1/2 \text{ (or) } q = p \\ 1 & \text{if } p > 1/2 \text{ (or) } q < p \end{cases}$$

→ Mode of Bernoulli

$p > q$ ⟶ p will be mode
Else q will be mode

→ Variance ⟶

$$\sigma^2 = pq = p(1-p)$$
$$\sigma = \sqrt{pq} = \sqrt{p(1-p)}$$

(2) Binomial Distribution

- → Multiple sequence/cases of Bernoulli distribution
- → success / → failure [Binary]
- → discrete r.v.
- → perform for n-trials

Ex:- Tossing a coin 10 times ⟶ $\boxed{n = 10}$

↳ Notation ⟶ $B(n, p)$

parameters ⟶ $n \in \{0, 1, 2, \ldots\}$ → No of trails
$p \in [0, 1]$ → success probab. for each trail
$q = 1-p$

support ⟶ $k \in \{0, 1, 2, \ldots n\}$ → No. of success

* p.m.f for Binomial

$$\longrightarrow \quad P_r(k, n, p) = {}^nC_k \, p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots n$ where

Binomial coeff. $\boxed{{}^nC_k = \frac{n!}{k!\,(n-k)!}}$

Scanned with CamScanner

→ Mean ⇒ $n.p$

→ Variance ⇒ $n.p.q$

→ σ (std) ⇒ $\sqrt{n.p.q}$

③ Poisson Distribution ↴

↳ Discrete r.v. (p.m.f)

↳ no. of events in fixed interval of time

Ex:- No. of people visiting hospital every hr
" " " " banks " "

⇒ { λ → Expected no. of events occuring at every time interval

→ p.m.f for poisson ⇒ $\boxed{p(x) = \frac{e^{-\lambda}\lambda^x}{x!}}$

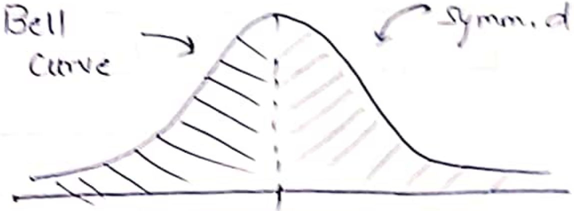
→ mean = $E(x) = \mu = \lambda * t$
where t → time interval

④ Normal / Gaussian Distribution ↴

↳ continuous r.v. (p.d.f)

Bell curve →

symm. distribution



$\boxed{\mu = \text{median} = \text{mode}}$

$\sigma^2 \uparrow\uparrow$ → spread $\uparrow\uparrow$

Notation → $N(\mu, \sigma^2)$

Parameters ⇒
$\mu \in R$ → mean
$\sigma^2 \in R > 0$ → variance
$x \in R$

→ p.d.f for Gaussian
$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

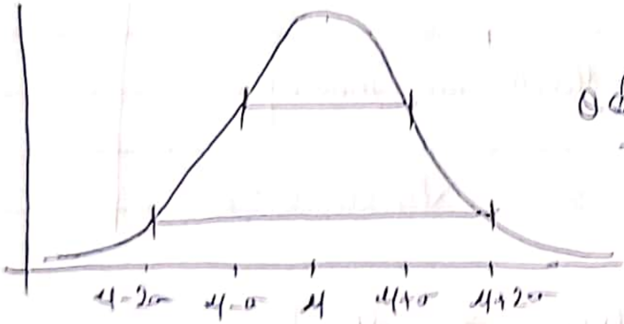
→ Mean ⇒ $\mu = \sum_{i=1}^{n} \frac{x_i}{n}$

Scanned with CamScanner

→ Variance ⇒

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

* Emperical rule of Gaussian distrib^n ⌉



QQ Plot

$\boxed{68 - 95 - 99.7}$ Rule

↳ $P(\mu - \sigma \le x \le \sigma + \mu) \simeq 68\%$

$P(\mu - 2\sigma \le x \le \mu + 2\sigma) \simeq 95\%$

$P(\mu - 3\sigma \le x \le \mu + 3\sigma) \simeq 99.7\%$

NOTE :- Std. Normal Distrib. is the one when we convert p.d.f to diffe. p.d.f with $\boxed{\mu = 0 \ \& \ \sigma = 1}$

↓ Can be converted through →

$\boxed{\text{Z-score} = \frac{x_i - \mu}{\sigma}}$

*** We do standardization just to bring every columns to same unit of measurement

★ Log Normal Distribution ⌉

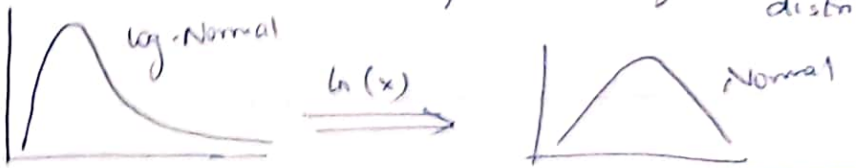
↳ p.d.f → random variable

If r.v. "x" is log normally distributed, the $\boxed{Y = \ln(x)}$ → Normal Distribution

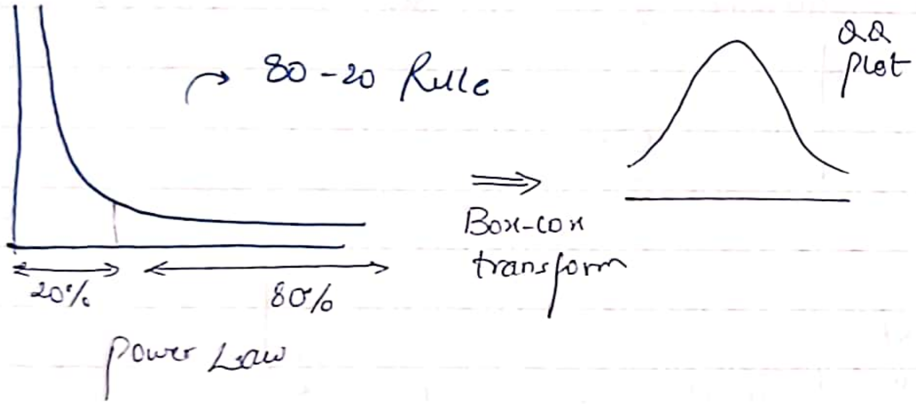
Simi,

Y → normally distributed, then

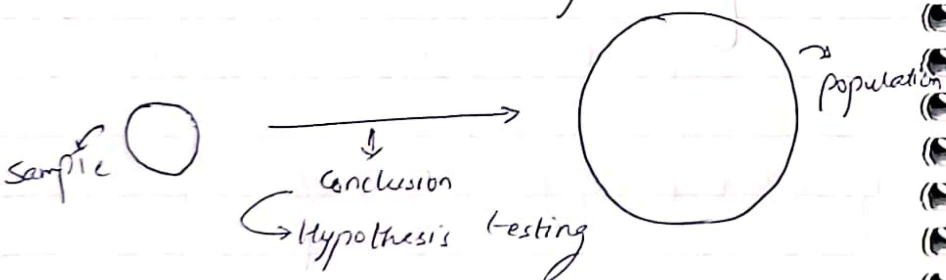
X = exp(Y) → log-normally distribute



* Power-law Distrib. ↴



* Inferential Statistics ↴ { Hypothesis testing }
↳ conclusion or Inference



Hypothesis Testing Mechanism

① Null Hypothesis (H_0)
↳ assumption to begin with.

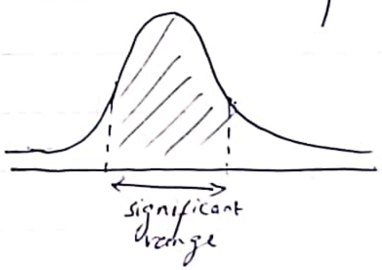
② Alternate Hypothesis (H_1)
↳ opp. of (H_0)

③ Experiments → statistical Analysis
→ Correct Proof

④ Accept the (H_0) or Reject the (H_0)

P-Value ⇒

use to accept or reject the null Hypothesis
↳ lies in significant value range → Accept H_0
if not → Reject H_0



* Hypothesis Testing & Statistical Analysis ↴

① Z-Test } → Average ⇒ Z table → Z score & P value
② t-Test } t table
③ Chi-Square ⇒ Categorical data
④ Annova ⇒ Variance

NOTE :- For Z test → we need pop. std. deviation & $n > 30$

so,

$$\boxed{Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}}$$ for pop. data.

↳ Now, for case when we don't know pop. std. dev. then we use t-test

$$\boxed{t = \frac{\bar{x} - \mu}{s / \sqrt{n}}}$$ S = sample std. deviation

→ we use dof $= n-1$

* Type 1 & type 2 Errors ↴

Outcome 1 :- We reject the null hypo. when in reality it is false → Good

Outcome 2 :- We reject " " " " " it is true → Type 1 Error

Outcome 3 :- We retain the null hypo. when in reality it is false → Type 2 Error

Outcome 4 :- " " " " " " it is true → Good

* Bayes Statistics (Bayes Theorem) ↴

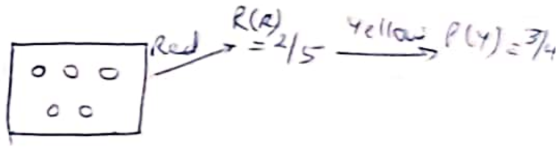
probab. → Independent Events
→ Dependent Events

① Indepen. Events

Ex:- Rolling a dice

$P(x) = 1/6$

② Dependen. Events



↳ $\boxed{P(R \& Y) = P(R) * P(Y/R)}$ conditional probab.

$= 2/5 * 3/4 = 6/20$

↳ $P(A \text{ and } B) = P(B \text{ and } A)$

$P(A) * P(B/A) = P(B) * P(A/B)$

$$P(B/A) = \frac{P(B) * P(A/B)}{P(A)}$$

$$P(A/B) = \frac{P(A) * P(B/A)}{P(B)}$$

Scanned with CamScanner

A, B ⟶ events

$P(A/B)$ ⇒ prob. of A given B is true
$P(B/A)$ ⇒ " " B " A is true

$P(A), P(B)$ ⇒ Independ. probab of A & B

* Confidence Interval

Z-test ⇒ point Estimate ± Margin Error

High & Low C.I. ⟶ $\bar{x} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$

t-test ⇒ $\bar{x} \pm t_{\alpha/2}\, \sigma/\sqrt{n}$

* Chi-Square Test

→ The Chi sq. test for goodness of fit test claims about population proportions.

→ It is a non-parametric test i.e. performed on categorical [ordinal & nominal] data.

$$\boxed{\chi^2 = \sum \frac{(O-E)^2}{E}}$$

O → Observed
E → Expected

Scanned with CamScanner