Date :
Page No.

Probability And Statistics

→ fundamental area

→ histogram, Pdf, cdf

→ random variable

$x = 2, 3 \ldots$

dice = {1, 2, 3, 4, 5, 6} Discrete Random
rolled a fair dice → equally likely

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

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

$P(x \text{ is even}) = \frac{1}{2} = P(x=2) + P(x=4) + P(x=6)$

$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

$P(X \text{ is odd}) = \frac{1}{2}$

Height of a Random Student
→ Continuous Random Variable

Outlier

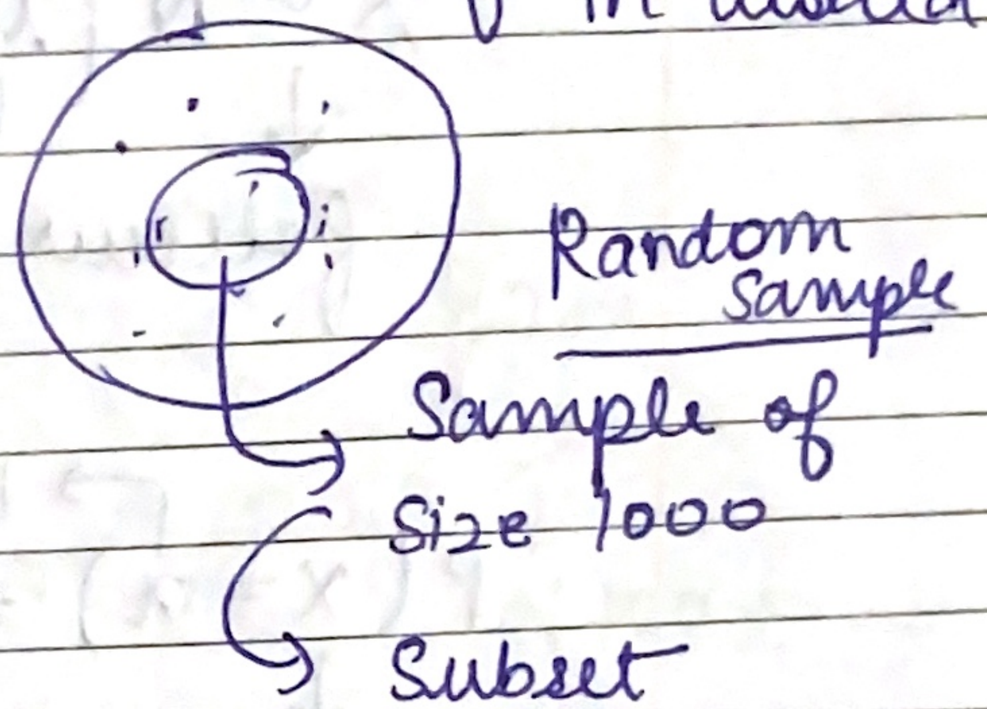
Y: Height of Student

[122.2, 140.4, 132.5 ---- 12.26 156.3]
→ Outlier

Date :
Page No.

Population and Sample

→ estimate average height of human?



Mean $\mu = \frac{1}{7B} + \sum_{i=1}^{7B} h_i$

$$\bar{h} = \frac{1}{1000} + \sum_{i=1}^{1000} h_i$$

↳ height in my sample

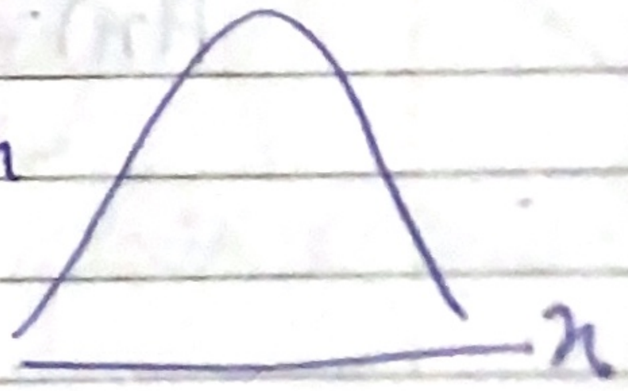
As sample size increases

$\bar{x} = \mu$

↳ sample mean → population Mean

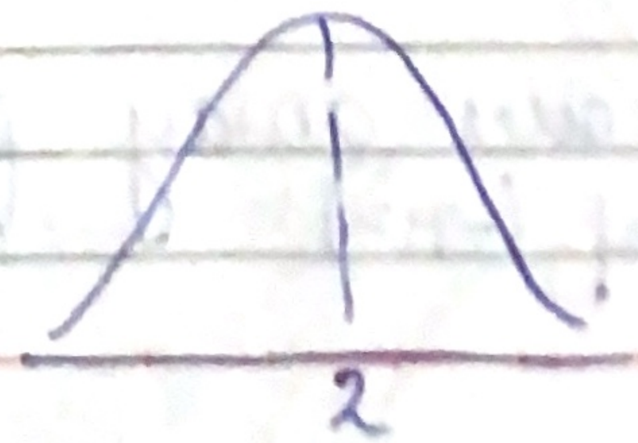
Gaussian Distribution

PDF of a Gaussian distribution random variable



X: Continuous random Variable

Parameters of Gaussian dist = μ, σ^2



$\mu = 2$

Date :
Page No.

Normal / Gaussian

$X \sim N(\mu, \sigma^2)$

follows — mean — Variance

$P(X = x) = P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$

$x \to 160$

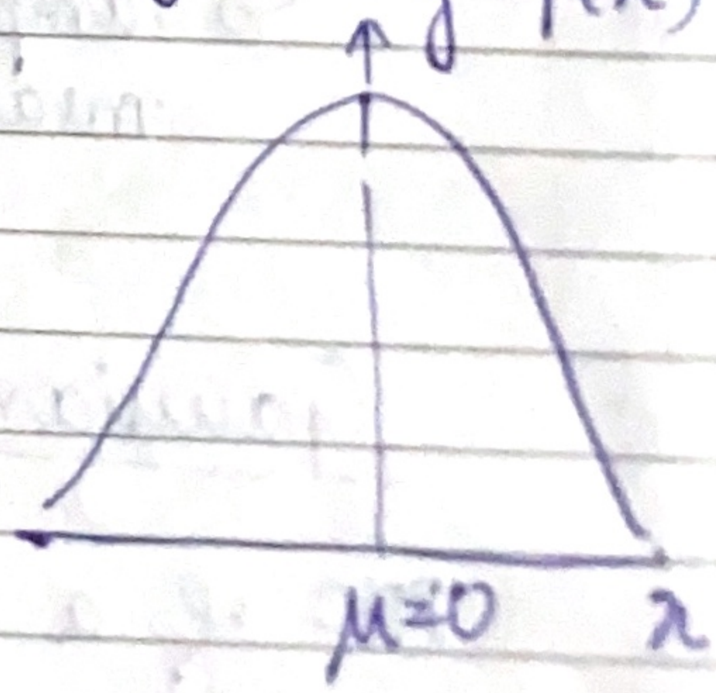
r.v

Let $\mu = 0 \quad \sigma^2 = 1$

$\sigma = 1$

$P(x) = \boxed{\frac{1}{\sqrt{2\pi}}} \exp\left\{\boxed{\frac{-1}{2}} x^2\right\} = y$

Constant



$P(x) = y = \exp(-x^2)$

$x \uparrow \quad (-x^2) \downarrow$

$\exp(-x^2) \downarrow$

① as x moves away from μ, $y \downarrow$

② symmetric

③ x moves away from μ, y reduces $\exp(-x^2)$

Date :
Page No.

$$y = \exp(-x^2)$$

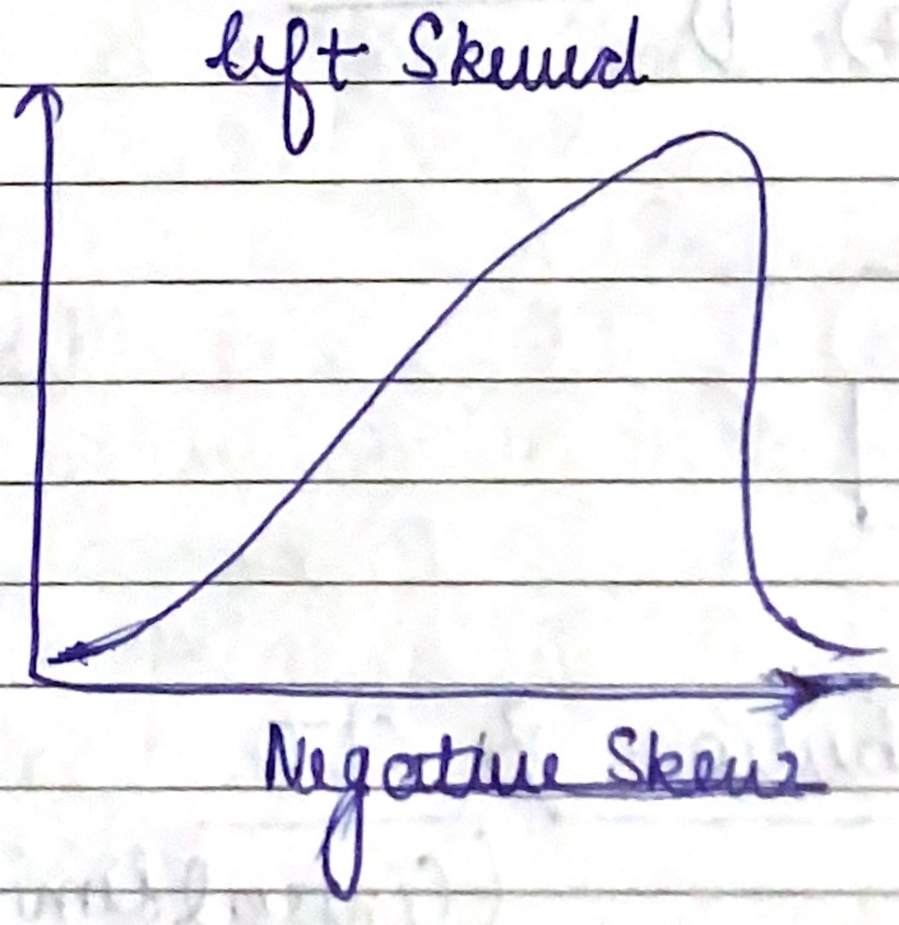
$x=0 \quad y=1$

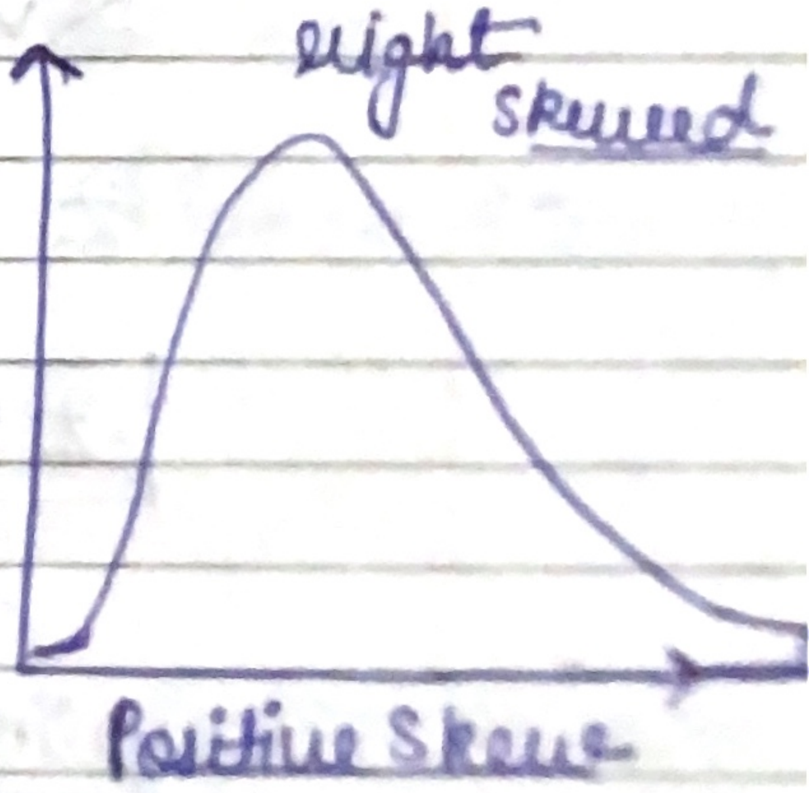
$x=1 \quad y = \exp(-1) = \frac{1}{e^1} = 0.3678$ } 2.0x

$x=2 \quad y = \exp(-4) = \frac{1}{e^4} = 0.018$ } 100x

$x=3 \quad y = \exp(-9) = \frac{1}{e^9} = 0.000123$

exponentially Quadratic drop





Standard Normal Variate (Z)

Date :
Page No.

① $Z \sim N(0,1)$

$\mu = 0$

$\sigma^2 = 1$

② let $X \sim N(\mu, \sigma^2)$

↳ $[x_1, x_2 -- x_{50}]$

Standardization

$x_i' = \frac{x_i - \mu}{\sigma}$

$x_i' \sim N(0,1)$

↑ standard normal variant

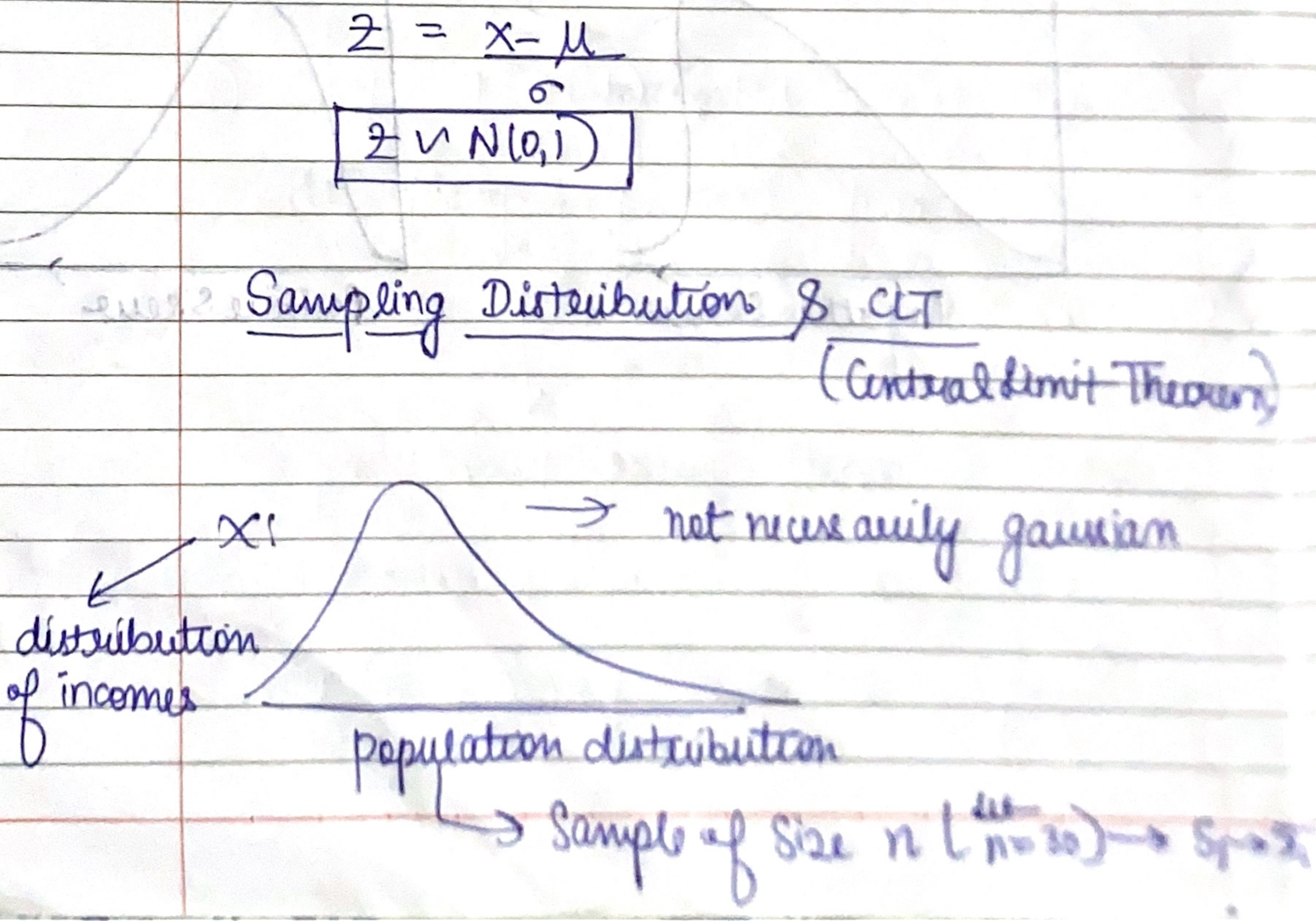
Standardization of Data

$X \sim N(\mu, \sigma 2)$

$$Z = \frac{x - \mu}{\sigma}$$

$$\boxed{Z \sim N(0,1)}$$

Sampling Distribution & CLT (Central limit Theorem)



Date :
Page No.

Sample mean

→ Sample of size (n) → S_2 → $\bar{x}_2$

3

→ Sm → $\bar{x}_m$

$\bar{x}_1, \bar{x}_2, \ldots \bar{x}_m$:- m sample - means

$\bar{x}_i$ ~ distribution

↳ distribution of sample means

dist of $\bar{x}_i$ = Sampling dist of sample - mean

CLT

If → need not gaussian

X: finite mean and variance (μ) (σ^2)

infinite μ & σ^2

(pareto dist)

→ $S_1, S_2 \ldots S_m$ (Sample of size n)

Sample mean ← $\bar{x}_1, \bar{x}_2 \ldots \bar{x}_m$

dist of $\bar{x}_i$ = Sampling dist of sample mean.

CLT :-

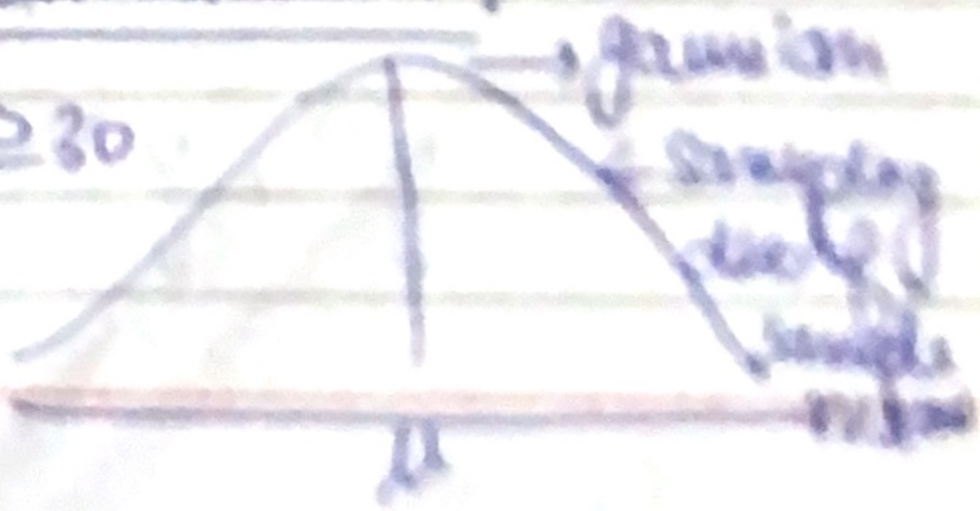
$$\bar{x}_i \to N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \to \infty$$

↑ gaussian distribution

$\mu \cong$ mean of x_i &

$\frac{\sigma^2}{n} \cong$ var of $\bar{x}_i$

$n \geq 30$



Quantile-Quantile plot: (Q-Q plot)

$X: x_1, x_2, x_3 \text{ --- } x_{500}$

(Q) is X gaussian dist? — QQ plot (graphical method)
— statistical testing (KS, AD)

How

① Sort x_i and compute percentiles

$x_1, x_2, \text{ ------ } x_{500}$

↓

Sort (asc)

$x_1', x_2', \text{ ---- } x_{500}' \xrightarrow{\text{Percentile}}$ ① ② ... (100)

$x_5', x_{10}', x_{15}' \text{ --}$ x_{500}'

$(x^{(1)}\ x^{(2)}\ x^{(3)}$

↳ 1st percentile value of x_1

② $y \sim N(0, 1)$ → Std Normal dist

$y_1, y_2, y_3 \text{ --- } y_{1000}$

↳ 1000 obs from N(0,1)

↓

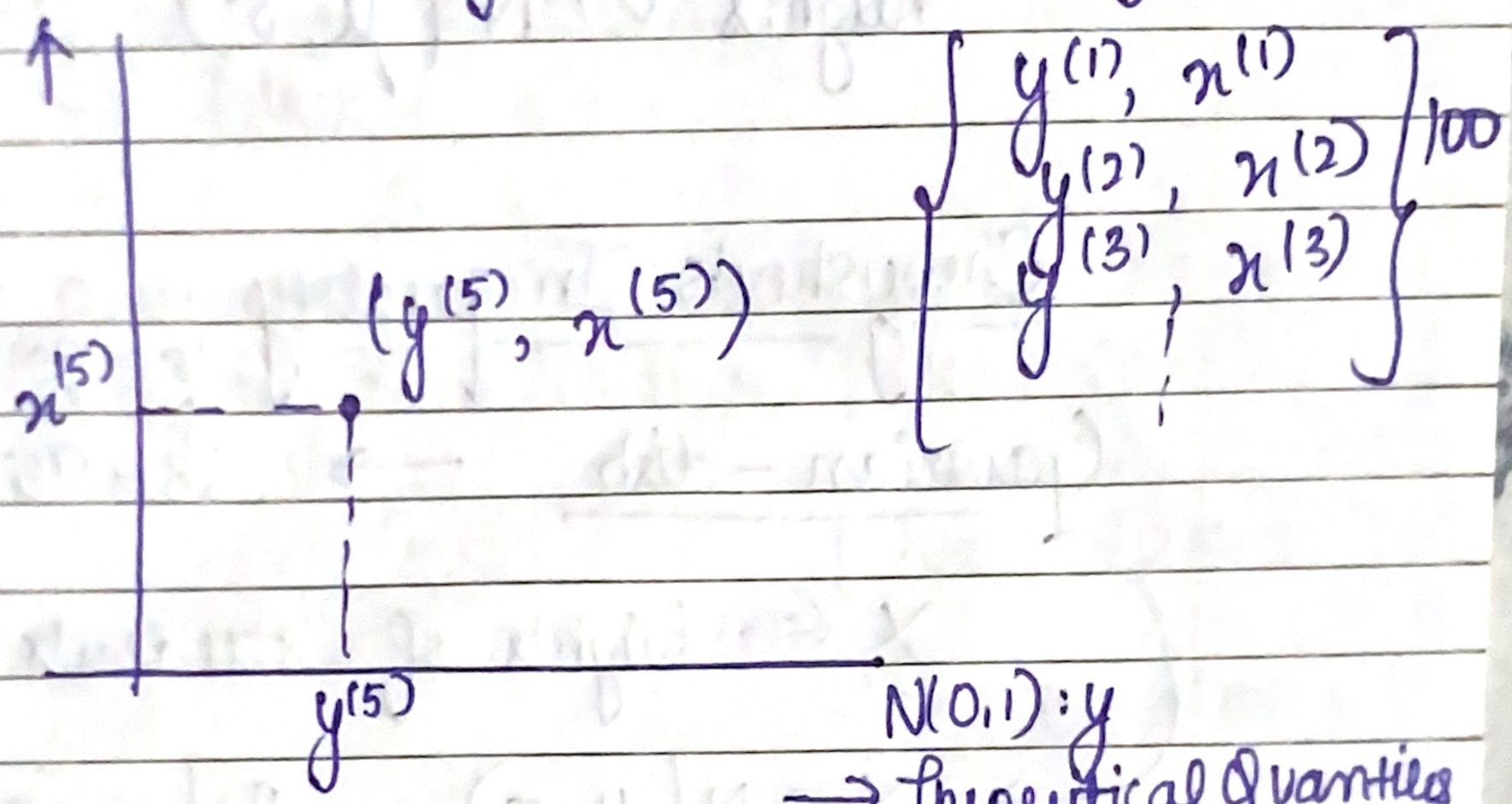
$y_1', y_2', y_3' \text{ --- } y_{1000}'$

↓ percentile

$y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)} \text{ --- } y^{(100)}$

Date :
Page No.

(3) Plot Q-Q plot using $x^{(1)}, x^{(2)} \ldots x^{(100)}$
$y^{(1)}, y^{(2)} \ldots y^{(100)}$



(100) x_{sod}

If $(y^{(i)}, x^{(i)})$ i→100 lie on straight line then x & y have similar distribution.

How / Where to use distributions?

probability → data analysis → answering Ques about data

(Q1) Company – XYZ

Task: Order t-shirts for all employees (100K)
S, M, L, XL

(a) How many XL t-shirts to order?

① collect data for all (100K)

height → 180cm → XL
160 → L

collect heights of 500 random employees

↓

mean, std dev

heights ~ N(μ, σ)

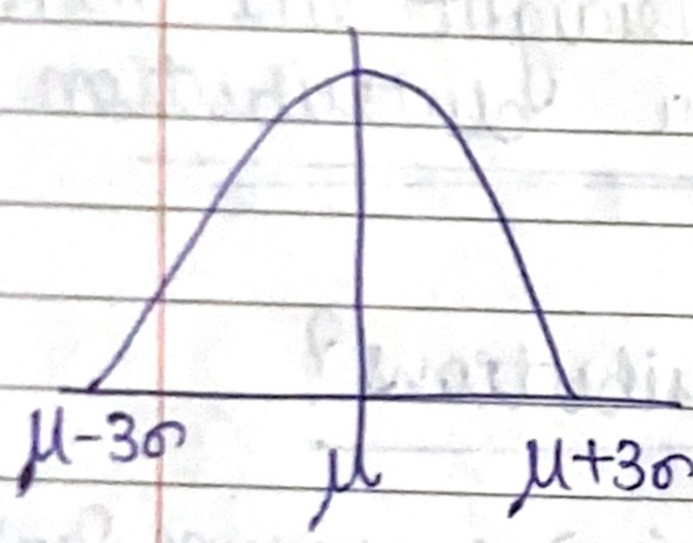
Chebyshev's Inequality

Gaussian - dist → 1σ 2σ 3σ → 68 - 95 - 99.7 rule

X ← heights of students

→ X ~ N(μ, σ) Pr(μ-2σ ≤ X μ+2σ) = 95%

95% of students heights lie in the [130, 170]



Q) what if I don't know dist

know μ, σ

↑ ↑

finite └ non zero & finite

finite

x% of data lies μ-2σ & μ+2σ

y% of data " μ-1-5σ & μ+1-5σ

Why?

X: r.v

Salaries of people ← don't know dist

μ, σ ← know

→ $10k

→ $40k

26 (Q) What percentage of individuals have Salary in range [20k, 60k]

36 (Q) " " " [10k, 70k]

→ 40 − 3(10k)

5%

Chebyshev's Inequality

X: r.v μ = finite mean

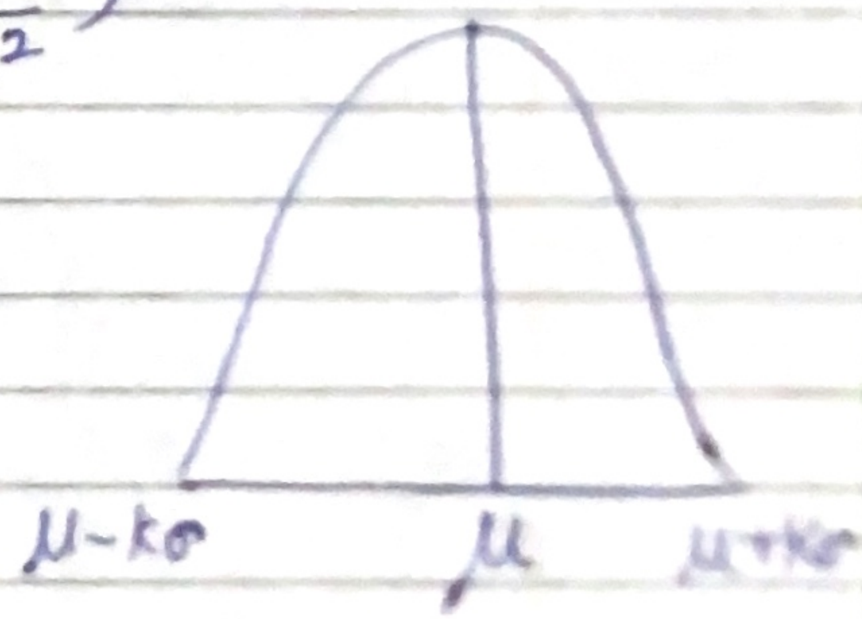
σ = non zero & finite

P

$$P(|x-\mu| \geq k\sigma) \leq \frac{1}{k^2})$$

$$x - \mu \geq k\sigma$$

& $$x - \mu \leq k\sigma$$



$$P\left(\begin{matrix} x \geq \mu + k\sigma \\ x \leq \mu - k\sigma \end{matrix}\right) \leq \frac{1}{k^2}$$

Or

$$P\left(\mu - k\sigma < x < \mu + k\sigma\right) > 1 - \frac{1}{k^2}$$

Date :
Page No.

all values equally probable

Uniform distribution

- discrete (only take discrete value)
- continuous

PMF → discrete r.v

PDF → continuous r.v

eg → Throwing dice

1, 2, 3, 4, 5, 6

↳ equally likely [1/6]

$a = 1$

$b = 6$

$n = b - a + 1 = 6 - 1 + 1 = 6$

$U(2,6)$

$a = 2$

$b = 6$

$U\{a, b\}$

$a \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$

$b \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}, b \geq a$

$n = b - a + 1$

$k \in \{a, a+1, \ldots b-1, b\}$

pmf $\frac{1}{n}$

Mean $a+b/2$

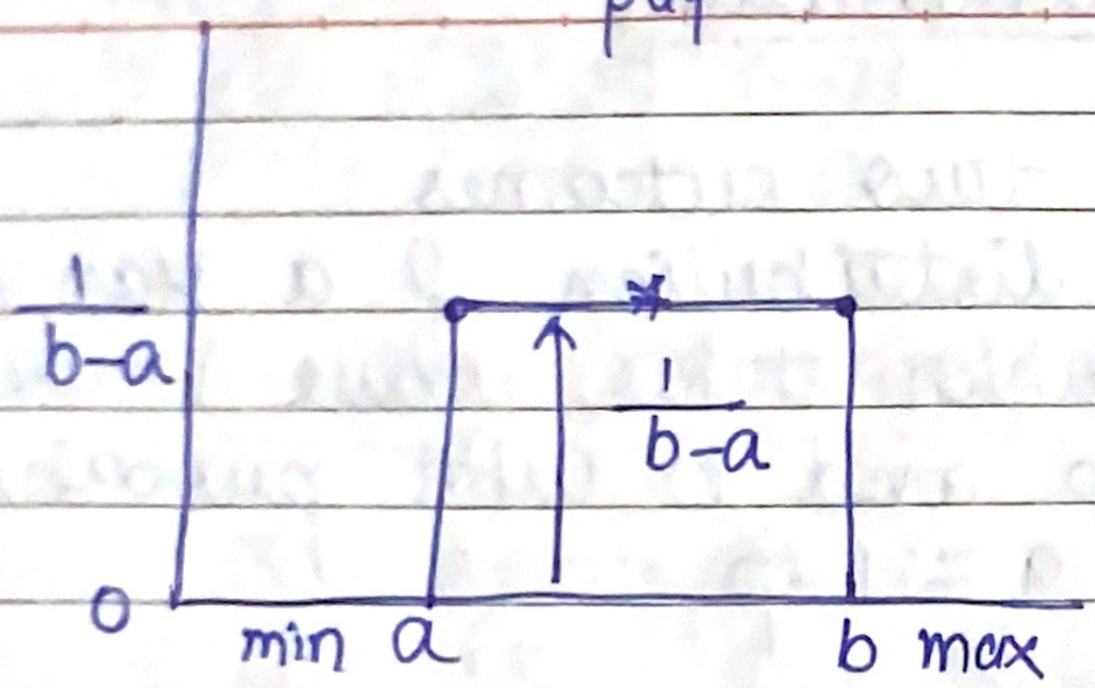
Median $a+b/2$

Variance $\frac{(b-a+1)^2 - 1}{12}$

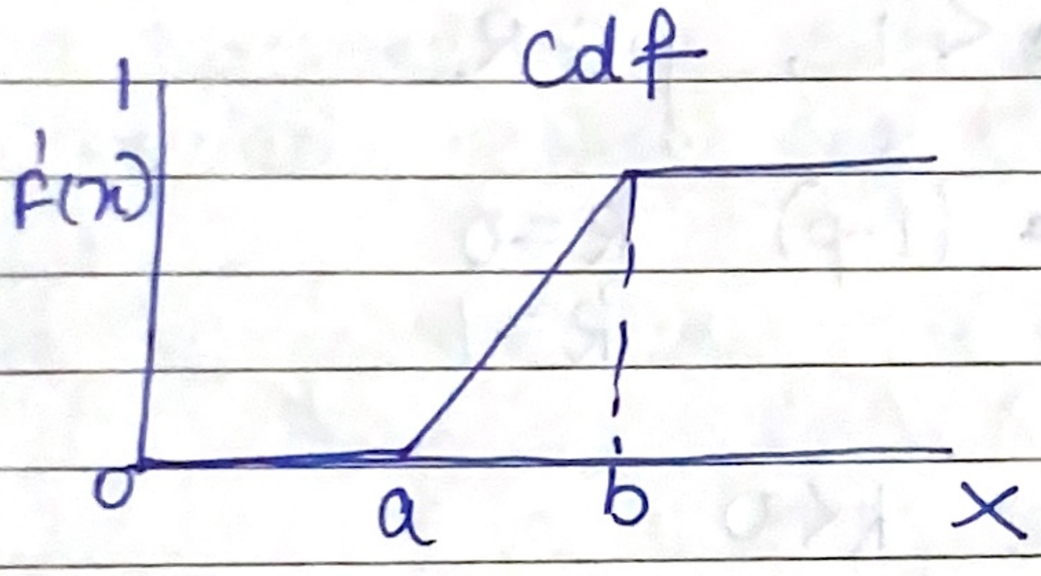
Skewness = 0

Date :
Page No.

pdf



cdf



$U(a,b)$

$-\infty < a < b < \infty$

$x \in [a,b]$

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

cdf $\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b) \\ 1 & x \geq b \end{cases}$

Mean $\frac{1}{2}(a+b)$

Median $\frac{1}{2}(a+b)$

Date :
Page No.

Bernoulli Distribution

→ having only two outcomes
→ probability distribution of a random variable which takes value 1 with probability p and 0 with probability $q = 1-p$

$$0 < p < 1, \quad p \in R$$

$$\text{pmf} \begin{cases} q = (1-p) & k=0 \\ p & k=1 \end{cases}$$

$$\text{cdf} \begin{cases} 0 & k<0 \\ 1-p & 0 \le k < 1 \\ 1 & k \ge 1 \end{cases}$$

Mean p

Variance $p(1-p) = pq$

Binomial Distribution

$Y \sim Bin(n, p)$ → prob of getting head
↓
no of trials

$B(n, p)$

$n \in N_0$ — number of trials
$p \in [0,1]$ — success probability in each trial

Date :
Page No.

pmf $\binom{n}{k} p^k (1-p)^{n-k}$

Log Normal Distribution

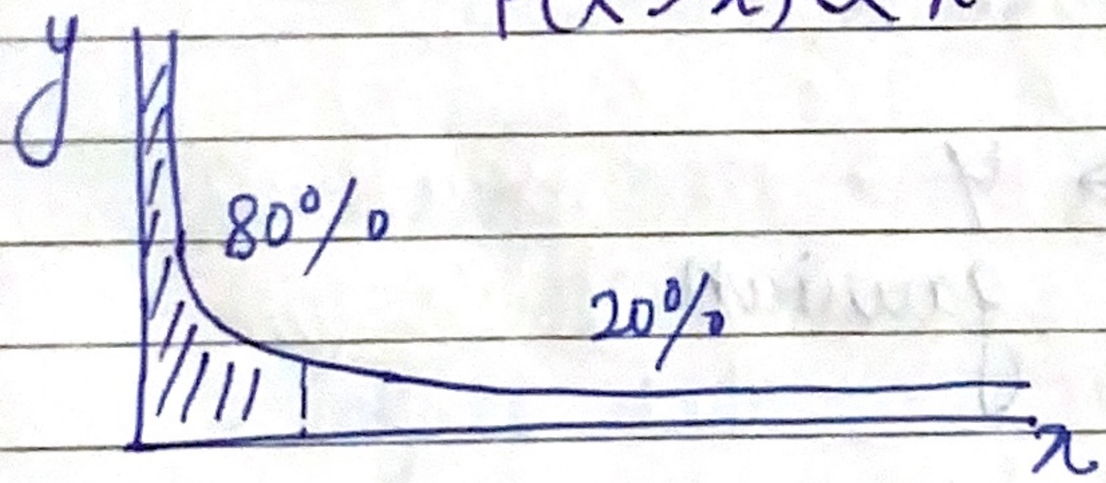
$X \sim$ log normal
if $\log(x)$ is normally distributed

log normal is a cont prob dist of random variable whose log is normally dist.

Thus if rand vars X is log normally dist then $y = \ln(x)$ has normal distribution

Power Law [called Pareto distribution]

$P(X > x) \propto x^{-\alpha}$



[80-20 rule]

Power law is a relationship b/w two quantities where a relative change in one quantity results in proportional relative change in other quantity

∴ One quantity varies as power of another

log-log plot → To determine whether dist follows power law

→ Plot $\log(x)$ vs $\log(P(x))$
→ Straigh line suggests power law

$$P(X>x) \propto x^{-\alpha}$$

- $P(X>x)$ is prob that variable X is greater than some value x.
- $\alpha > 1$ is power law exponent.
- implies that small events are common, large events are rare but not negligible.

Power transform (Box-Cox transform)

$$X \xrightarrow{\ln} Y$$

$\downarrow$ $\downarrow$

log normal gaussian

$$X \xrightarrow{?} Y$$

Power Law / pareto gaussian

Conversion → $X: x_1, x_2 \sim$ [Pareto]

$Y: y_1, y_2 \sim$ (gaussian)

① Box-Cox (x) = $\lceil$alpha lambda (λ)

$(x_1, x_2 \ldots x_n)$

Date :
Page No.

② $$y_i = \begin{cases} \frac{x_i^{\alpha} - 1}{\alpha} & \text{if } \alpha \neq 0 \\ \log(x_i) & \text{if } \alpha = 0 \end{cases}$$

$\forall_i : 1 \to n$

Applications of non-gaussian dish

↓

100 of dish
↳ Why?

uniform → random nos
Bernoulli, Binomial
log normal, Pareto

Weibull Distribution

↳ dams → how high dam should be
↳ water storage (max rainfall)

Civil engg:
Max one day rainfall
rainfall data → 30 yrs, 50 yrs
↳ few hundreds of data pts

$P(X > 20cm)$ is v small ⟶ bfore computers
$\leq 0.1\%$

$P(X \geq 20cm) = ?$

percentiles
$\frac{1}{200} = 0.5\%$

dist

Date :
Page No.

100s of dist → gaussian — ✗
→ log normal — ✗
→ weibull — ✓

X = 1, 2, 3 → fit a known & well studied dist / theoretical model

① X = max daily rainfall
→ Weibull (parameters) ②

② CDF & PDF $P(X \geq 20cm) = 1 - P(X < 20cm)$
↓
cdf of X

Box-Cox transform

X → X' → all analysis → gaussian dist
↓
non gaussian
→ max dainly rainfall

Date :
Page No.

Relationship b/w Random Variables

X: Heights
Y: weights

	X=h	Y=W
S_1	160	62
S_2	150	54
	1	1
	/	/
S_n	140	48

Q) relation b/w X & Y

X↑, Y↑

X↑, Y↓

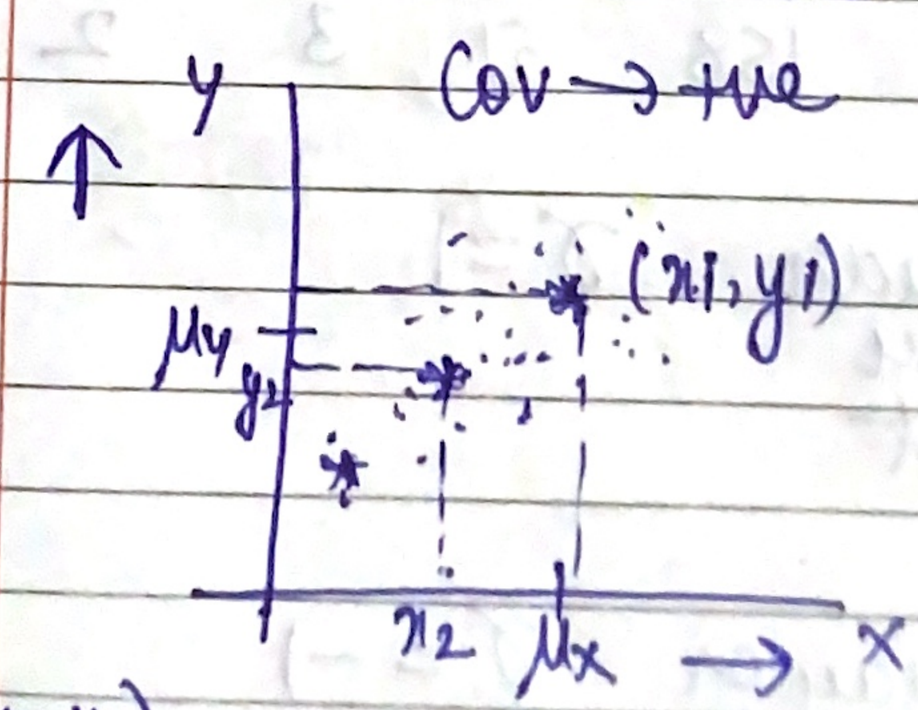
Co-variance, pearson correlation coff, Spearman rank corre. coff

$$\text{Covariance}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)*(y_i-\mu_y)$$

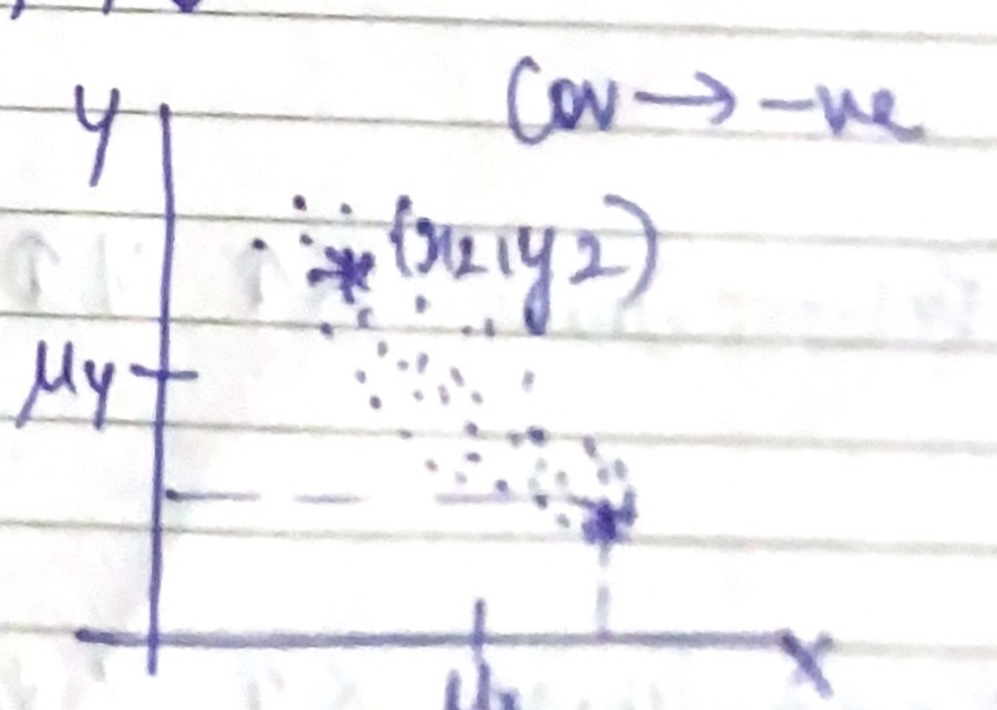
$$\text{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)*(x_i-\mu_x)$$

$$\text{cov}(X,X) = \text{Var}(X)$$

* cov(X,Y) = +ve X↑, Y↑
* cov(X,Y) = -ve X↑, Y↓



(x_1,y_1)	+	+	+
(x_2,y_2)	−	−	+
(x_3,y_3)	−	−	+



(x_1,y_1)	+	−	−
(x_2,y_2)	−	+	−

Date :
Page No.

$$Cov\left(\overset{cm}{x}, \overset{kg}{y}\right) \neq Cov\left(\overset{feet}{x'}, \overset{lbs}{y'}\right)$$

Pearson correlation coeff (pcc)

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

$$\sigma_x = \sqrt{Var(x)}$$

$$-1 \leq \rho < 1$$

→ Only studies linear relations

Spearman rank-corer coeff (γ)

$$\gamma = \rho_{r_x, r_y}$$

			r_x	r_y
S_1	160	52	4	3
S_2	150	66	2	4
S_3	170	68	5	5
S_4	140	46	1	1
S_5	158	51	3	2

$\rho = 1 \leftarrow$ linear $X \uparrow$ $Y \uparrow$ linear or not : $\gamma = 1$

$\rho = -1 \leftarrow$ linear $X \uparrow$ $Y \downarrow$ linear or not $\gamma = -1$

Date :
Page No.

Correlation does not imply causation

X causes Y } X
Y causes X } X

Causal Models → what causes what

Application of Correlation → not causation

SRCC
PCC

(1) Is salary correlated with sq. feet of home?

(2) Is no of years of education correlated with income?

(3) ecommerce: amazon
Is time spent in 24 hrs → money spent in 24 hrs

unique visitors in day Vs $ sale in day

100K → $1M
120K → $1.6M

(4) Medicine
(dosage of drug) (reduction in sugar)

1mg x
2mg y
3mg z