Lec I

Friday, 25 July 2025 10:10 AM

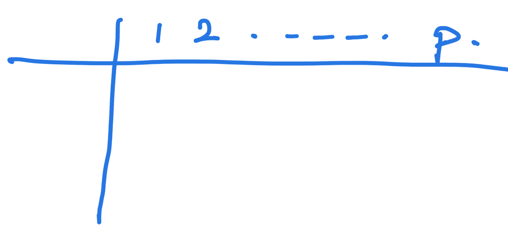
Books – ① Applied multivariate Stat. Analysis
Richard Johnson & D Wichern.
② multivariate analysis
K.V Mardia & 2 authors.

---

$p \to$ features / attributes $\to$ columns. $p > 1$. $p = 1$ r.v.

$\underset{\sim}{X} \to$ random vector of p-dimension.

$$\underset{\sim}{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

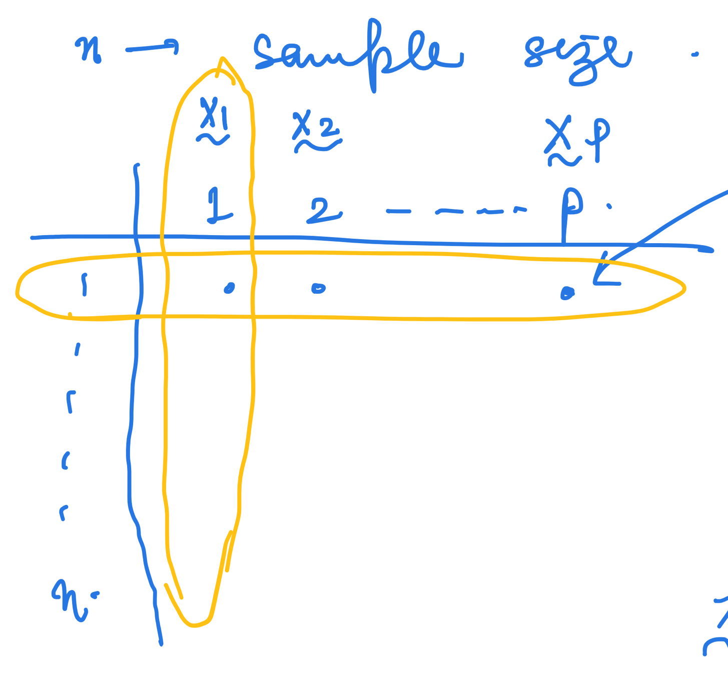



Population – collection of all objects. entire universal set.
uncountably infinite.

Sample – Finite subset of population.
There are methodologies to collect samples.

we assume sampling is random & iid.
if $n$ is big, we can use CLT.

$n \to$ sample size.



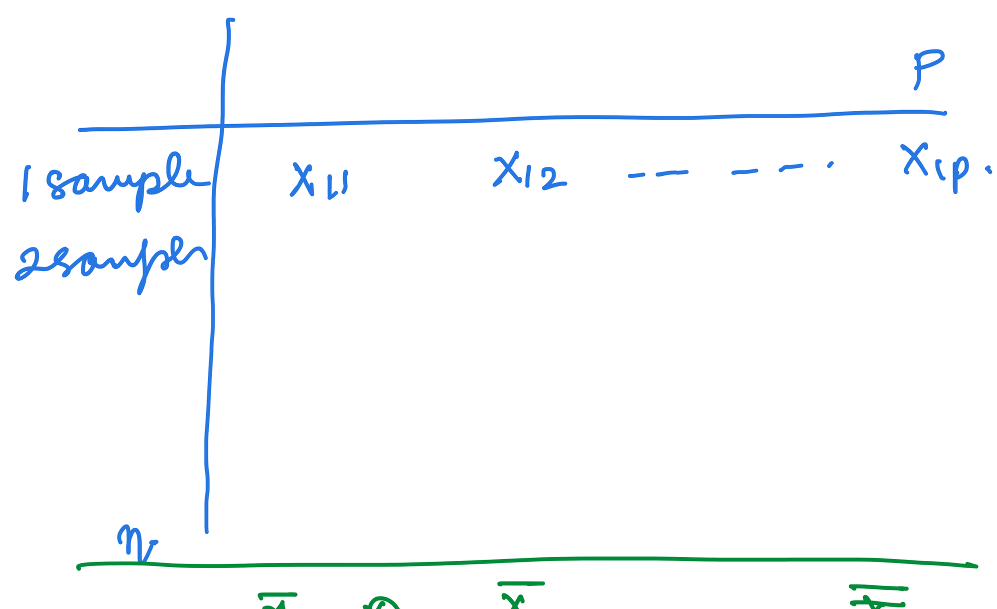

random variables.
sample.

$\underset{\sim}{x}_1^T$ (as a vector means a col. vector).

$\underset{\sim}{x}_2^T$

$\underset{\sim}{x}_n^T$.

← denotes vector.




$\overline{x}_1$ $\overline{x}_2$ $\overline{x}_p$.

$$\underset{\sim}{X}_1^T = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{bmatrix}$$

→ random var of 1-dim.

random variable (pre-conducting) → experiment & question set up.
Not yet asked the question. There is a variable which can take possible values.
whereas sample variable is the realisation ~~of~~ value of the r.v.

$x_{11}$ r.v.

Sample r.v. i.i.d. independent & identically distribution

$f(\underset{\sim}{x}_1, \underset{\sim}{x}_2 \cdots \underset{\sim}{x}_p, \theta)$.
↳ parameter of 1-dim or higher dim.

sample space $\mathbb{R}^p$.
rows are independent, not columns.
so joint pdf in terms of columns.
pdf doesn't depend on $n$, no. of samples but the attributes.

Sample mean. $\overline{\underset{\sim}{X}}$

$$\overline{\underset{\sim}{X}} = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{bmatrix}, \qquad \overline{x}_j = \frac{\sum_{r=1}^{n} x_{rj}}{n}$$

$1 \le j \le p$.

Population mean: $\underset{\sim}{\mu}$ : $p \times 1$.

Covariance : $i, j$ two features.

$$s_{ij} = \frac{1}{n} \sum_{r=1}^{n} (x_{ri} - \overline{x}_i)(x_{rj} - \overline{x}_j).$$

Covariance of $\underline{x_i}$ with $\underline{x_j}$.
$Cov(\underset{\sim}{x}_i, \underset{\sim}{x}_j)$

$1 \le i \le p$
$1 \le j \le p$.

if $i = j$ $s_{ii}$ = variance in $x_i$.
$Var(\underset{\sim}{x}_i) = Cov(\underset{\sim}{x}_i, \underset{\sim}{x}_i)$.

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & & & \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} p \times p.$$

Variance co-variance matrix of sample.

$$s_{ij} = \underbrace{\frac{1}{n-1}}_{\text{for pop sometimes we take } n-1} \sum_{r=1}^{n} (x_{ri} - \overline{x}_i)(x_{rj} - \overline{x}_j).$$

$p=1$
$\sigma^2$: population variance
$s$ : sample variance
$E(s) = \sigma^2$ unbiased estimator of $\sigma^2$
$s = \frac{1}{n-1} \sum_{r=1}^{n} (x_r - \overline{x})^2$

if $s = \frac{1}{n}$ ( " )

$E(s) = \sigma^2$ will get violated.
but if $n$ is large enough, it does not matter.
if $n$ is small matters.

Population variance - covariance matrix $p \times p$.

$\Sigma$

$\underset{\sim}{\mu}$, $\Sigma \to$ Pop.
$\underset{\sim}{X}$, $S \to$ sample.