- INTRODUCTION
- MULTIVARIATE OBSERVATIONS (Data)

UNIVARIATE STATISTICS

MULTI-VARIATE DOMAIN (corresponding quantities)

① Central Tendencies
- Mean — Mean Vector $\begin{bmatrix} - \\ - \\ - \end{bmatrix}_{p\times 1}$
- Mode
- Median

Dispersion
- Range
- Inter Quartile Range
- STd. deviatn (Variance) — Covariance Matrix ⇒ square $\begin{bmatrix} \diagdown \end{bmatrix}_{p\times p}$ Var-Covariance

& additionally CORRELATION matrix. (as there are more than 1 variables).

② Multivariate (Population) observations (Before data collection)

$$\bar{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}_{p\times 1} \qquad \text{obsr} \quad \begin{array}{c|cccccc} {}_{i}\diagdown^{j} & X_1 & X_2 & X_3 & X_j & \cdots & X_p \\ \hline 1 & X_{11} & X_{12} & X_{13} & X_{1j} & \cdots & X_{1p} \\ 2 & X_{21} & X_{22} & X_{23} & X_{2j} & \cdots & X_{2p} \\ \vdots & & \vdots & & & & \vdots \\ i & X_{i1} & X_{2i} & & X_{ij} & & X_{ip} \\ \vdots & & \vdots & & & & \vdots \\ n & X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{np} \end{array}$$

variables

Data Matrix $\mathbf{X}_{n\times p}$

$i = 1, 2, --- n$ = No. of observations

$j = 1, 2, --- p$ = No. of variables (co-variates)

when data is collected : X_{ij} = i^{th} observatn on the j^{th} variable. to be collected. (∴ it is a random variable).

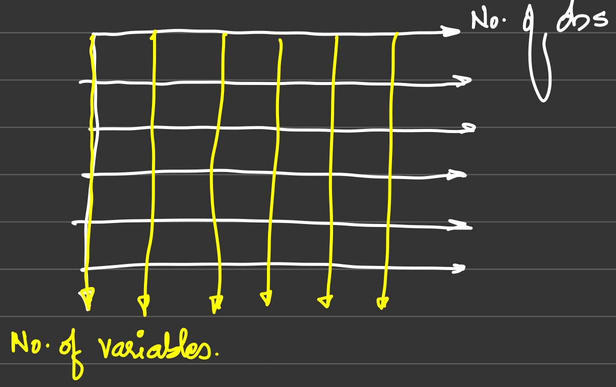
∵ being a r.v, we can EXPECT the value of the r.v.

As in UNIVARIATE case, if 'x' is a r.v then $E[x] = \mu$

Similarly, here in MULTI-VARIATE, every variable $X_1, X_2 --- X_p$ will have an Expected Value

$$X_i = i^{th} \text{ Multivariate OBSERVATION } = \begin{bmatrix} X_{i1} \\ X_{i2} \\ X_{i3} \\ \vdots \\ X_{ip} \end{bmatrix}_{p\times 1} \text{ vector}$$

Instead if we write $X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix}_{n\times 1}$ ie j^{th} variable of ALL observations.



④ when Data is actually collected, we get the similar shaped Data Matrix. (fixed values)

$$\underset{n\times p}{X} = \begin{array}{c|cccccc} & & & & x_j & & \\ & x_{11} & x_{12} & x_{13} & x_{1j} & \cdots & x_{1p} \\ & x_{21} & x_{22} & x_{23} & x_{2j} & \cdots & x_{2p} \\ & \vdots & & & \vdots & & \\ x_i \Rightarrow & x_{i1} & x_{i2} & x_{i3} & x_{ij} & \cdots & x_{ip} \\ & \vdots & & & \vdots & & \\ & x_{n1} & x_{n2} & x_{n3} & x_{nj} & & x_{np} \end{array}$$

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}_{p\times 1} \qquad x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix}_{p\times 1}$$

⑤ Mean Vector (Populatⁿ Parameter)

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_j \\ \vdots \\ \mu_p \end{bmatrix}_{p\times 1} = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_j] \\ \vdots \\ \mathbb{E}[X_p] \end{bmatrix}$$

$$\mu_j = \mathbb{E}[X_j] = \begin{cases} \sum_j x_j \cdot p(x_j) \\ = \int_{-\infty}^{\infty} x_j \cdot f(x_j)\, dx_j \end{cases}$$

Mean Vector (Sample Parameter).

(Average Vector)

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{bmatrix}_{p\times 1} = \begin{bmatrix} 1/n \sum_{i=1}^{n} x_{i1} \\ 1/n \sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{ip} \end{bmatrix}_{p\times 1}$$

↓ estimate of 'μ' in this case

where

$$\bar{x} = \frac{1}{n} X^T \mathbb{1}$$

In case of sample (after data is collected)

$\bar{x}$ = Average is an ESTIMATE of the mean = $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\underset{\text{avg. vector for sample}}{\bar{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_j \\ \vdots \\ \bar{X}_p \end{bmatrix}_{p\times 1} = \begin{bmatrix} 1/n \sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{bmatrix}_{p\times 1}$$

NOTE: Here we will not got for individual Average Calculatn $\frac{1}{n}\sum_{i=1}^{n} x_{ij}$

Instead we will use a cool Matrix Multiplication TRICK

Consider the data Matrix

$$\underset{n\times p}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p} \Rightarrow \text{we want } \bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}_{p\times 1}$$

where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$

To go from $(n\times p)$ matrix to $(p\times 1)$ matrix.

We create a matrix (all ones), & use it the data Matrix to directly get Avg. vector simultaneously

$$\mathbb{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}$$

(*) Think of $(p\times \underbrace{n)\times(n}\times 1) = (p\times 1)$

To do this, we will have a TRANSPOSE of the data Matrix

$\therefore \underset{p\times n}{X^T} \cdot \underset{n\times 1}{\mathbb{1}} = \underset{p\times 1}{}$

To make the computation less tedious, let $n=3$ and $p=2$ (ie consider the Bivariate case)

$$\underset{3\times 2}{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}_{3\times 2}$$

obsv ↑ variates ↑

since there are $(n=3)$ observations, ∴ Create a Unit vector

$$\mathbb{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Now consider

$$X^T\cdot\mathbb{1} = \begin{bmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{bmatrix}_{2\times 3} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}_{3\times 1} = \begin{bmatrix} (x_{11}+x_{21}+x_{31}) \\ (x_{12}+x_{22}+x_{32}) \end{bmatrix}_{2\times 1} = \begin{bmatrix} \Sigma x_{i1} \\ \Sigma x_{i2} \end{bmatrix}$$

Now, we can simply divide each term by 'n' to get the average vector $\boxed{\bar{X} = \frac{1}{n}\, X^T\cdot\mathbb{1}}$

NOTE: give an EXERCISE here for a sample.csv data sheet for practice. let the students use either NUMPY or EXCEL.

⑥ Population Covariance Matrix.

6.1
$$\sigma_j^2 = \sigma_{jj} = E\left[(x_j-\mu_j)^2\right] = \begin{cases} \sum\limits_{\forall x_j} (x_j-\mu_j)^2 \cdot f(x_j) & \text{— for discrete } x_j \\ \int\limits_{-\infty}^{\infty} (x_j-\mu_j)^2 \cdot f(x_j)\, dx_j & \text{— for continuous } x_j \end{cases}$$

6.2
$$\therefore \operatorname{cov}(x_j, x_k) = \sigma_{jk} = E\left[(x_j-\mu_j)(x_k-\mu_k)\right]$$
$$= \sum_{\text{all } x_j} \sum_{\text{all } x_k} (x_j-\mu_j)(x_k-\mu_k) \cdot f(x_j, x_k).$$

Since there are a total of 'p' variables, we can now calculate a COVARIANCE MATRIX with 'p' rows & 'p' columns.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1j} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2j} & \cdots & \sigma_{2p} \\ \vdots & & & \vdots & & \vdots \\ \sigma_{j1} & \sigma_{2j} & \cdots & \sigma_{ij} & \cdots & \sigma_{ip} \\ \vdots & & & & & \\ \vdots & & & & & \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pj} & \cdots & \sigma_{pp} \end{bmatrix}_{P \times P}$$

NOTE: This is a SYMMETRIC MATRIX

Discuss Covariance Matrix, firstly from the population point of view & then sample pov

If $x \rightarrow$ r.variable

then Variance $V(x) = E\left[(x-\mu)^2\right]$ $= \sum\limits_{\forall x} (x-\mu)^2 \cdot f(x)$
$= \int\limits_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x) \cdot dx$ $\Bigg\} = \sigma^2$ (for population)

Imp

if X_1 & X_2 are Dependent then there will be correlatn Simultaneously vary. ie there is COVARIANCE

$$\therefore \text{Variance } (x_j) = E\left[(x_j-\mu_j)^2\right]$$
$$= \sum_{\forall x_j} (x_j-\mu_j)^2 \cdot f(x_j) \longleftarrow \text{Discrete.}$$
$$= \int_{-\infty}^{\infty} (x_j-\mu_j)^2 \cdot f(x_j)\, dx_j \longleftarrow \text{Continuous.}$$

If we sub. $j = 1, 2, \ldots p$. if $j=1 \Rightarrow$ we will get σ_1^2
if $j=2 \Rightarrow$ we will get σ_2^2

Consider Variance $V(x_j) = E[(x_j - \mu_j)^2]$

Using this we write $cov(x_j, x_k) = E[(x_j-\mu_j)(x_k-\mu_k)] = \sigma_{jk} = \sigma_{kj}$

$$\sigma_{kj} = \sigma_{jk} = \sum_{\text{all } x_j, x_k} (x_j-\mu_j)(x_k-\mu_k)\cdot \underbrace{p(x_j, x_k)}_{\text{joint probability}}$$

∴ For a total no. of 'p' variables, we can write down the population Cov. Matrix

$$\underset{\text{Population Covariance Matrix}}{\Sigma} = \begin{array}{c|cccccc} & x_1 & x_2 & x_3 & \cdots & x_p \\ \hline x_1 & \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ x_2 & \sigma_{12} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ x_3 & \sigma_{13} & \sigma_{23} & \sigma_{33} & \cdots & \sigma_{3p} \\ \vdots & \vdots & & & \ddots & \vdots \\ x_p & \sigma_{1p} & \sigma_{2p} & \sigma_{3p} & \cdots & \sigma_{pp} \end{array}$$

$p \rightarrow$

$p \downarrow$

off diagonal elements are covariances.

Variance elements along the diagonal.

∴ This matrix is also called VARIANCE – COVARIANCE MATRIX

Now, repeat the same for SAMPLE COVARIANCE calculation.

Data matrix

$$\bar{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{2j} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & & & & & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \Rightarrow$$

$$\frac{\text{mean}}{\text{avg}} = \bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_j \quad \bar{x}_p$$

step#1: subtract avg. value

$$\begin{bmatrix} x_{11}-\bar{x}_1 & x_{21}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & & x_{2p}-\bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & & x_{np}-\bar{x}_p \end{bmatrix}$$

same

Instead we define

$$X^* = \begin{bmatrix} x^*_{11} & \cdots & x^*_{1p} \\ x^*_{21} & \cdots & x^*_{2p} \\ \vdots & \cdots\ x^*_{ij}\ \cdots & \\ \vdots & \cdots & \\ x^*_{n1} & \cdots & x^*_{np} \end{bmatrix}$$

where $x^*_{ij} = x_{ij} - \bar{x}_j$

∴ Covariance Matrix S_{jk} betn x_j, x_k

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} x^*_{ij}\cdot x^*_{ik}$$

Recap:

$$\text{Data Matrix} \quad \bar{X} = \begin{array}{c|cccc} & \multicolumn{4}{c}{\overbrace{\qquad\qquad\qquad}^{\text{Variables}}} \\ & x_1 & x_2 & & x_p \\ \hline & x_{11} & x_{12} & \cdots & x_{1p} \\ & x_{21} & x_{22} & \cdots & x_{2p} \\ & \vdots & & & \\ & x_{i1} & x_{i2} & \boxed{x_{ij}} & x_{ip} \\ & \vdots & & & \\ & x_{n1} & x_{n2} & \cdots & x_{np} \\ \hline \text{mean} & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{array}$$

let in general $x^*_{ij} = x_{ij} - \bar{x}_j$

$$\therefore X^* \begin{bmatrix} x^*_{11} & x^*_{12} & \cdots & x^*_{1p} \\ x^*_{21} & x^*_{22} & & x^*_{2p} \\ | & | & & \vdots \\ | & \vdots & & \vdots \\ x^*_{n1} & x^*_{n2} & & x^*_{np} \end{bmatrix}_{n\times p}$$

Now consider the Transpose x^{*T} then its order will be "p×n" and take a dot product with itself: $X^{*T} \cdot X^* = \left[\quad\right]_{p\times p}$

$$= (n-1)\, S_{p\times p}$$

$$= (n-1)\, \text{Cov}(\bar{x})$$

$$\therefore \boxed{\text{Cov}(\bar{x}) = \frac{1}{n-1} X^{*T} \cdot X^*}$$

X

Population CORRELATION MATRIX : denoted by 'ρ' Rho

Σ = Populat^{n} Cov. Matrix

$$\rho_{p\times 1} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ | & & 1 & \\ | & & & 1 \\ \rho_{np} & - - & & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \sigma_{31} & & \ddots & \\ \vdots & & & \ddots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$\boxed{\text{Correlation}(x_j, x_k) = \frac{\text{Covariance}(x_j, x_k)}{S.D(x_j) \cdot SD(x_k)}}$$

$$\rho_{jk} = \frac{\text{cov}(x_j, x_k)}{\sigma_j \cdot \sigma_k} = \frac{\sigma_{jk}}{\sigma_j \cdot \sigma_k}$$

$$\therefore \boxed{\sigma_{jk} = \rho_{jk} \cdot \sigma_j \sigma_k}$$

if $j = k$

$$\sigma_{jj} = \rho_{jj} \sigma_j \sigma_j$$

$$\downarrow$$

$$\cancel{\sigma_j}\cancel{\sigma_j} = \rho_{jj} \cdot \cancel{\sigma_j}\cancel{\sigma_k}$$

$$\boxed{\therefore \rho_{jj} = 1}$$