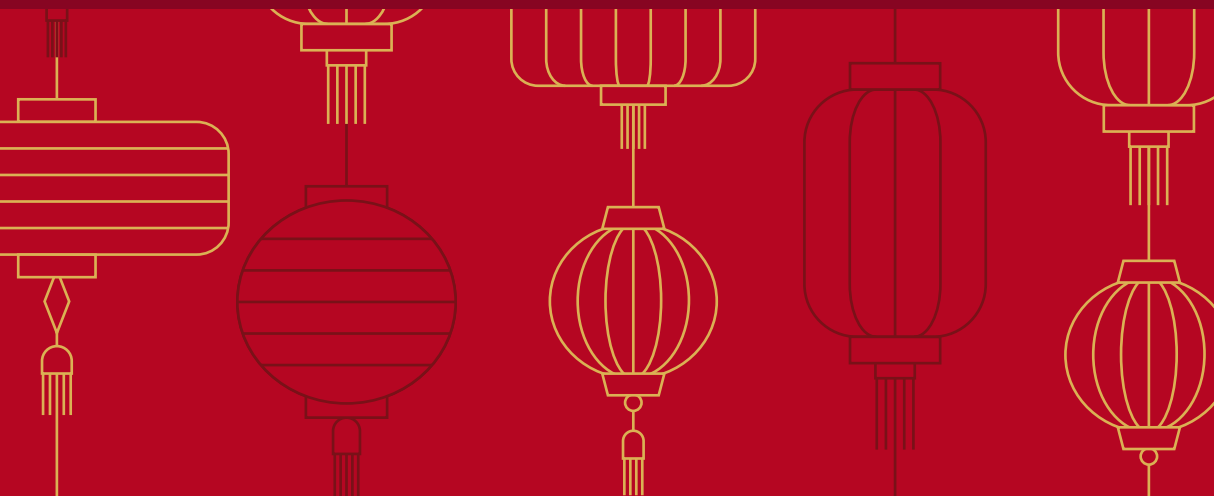


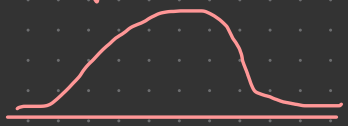
MULTIVARIATE STATISTICAL MODELING

Lecture-1

Applied → Refers to using base to make it useful

Normal Distribution :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty$$



Variable :

→ Takes different values

→ x = ID

i	ID
1	x_1
2	x_2
⋮	⋮
n	x_n

① Fixed / Deterministic

② Random / Probabilistic

→ Different features

	x_1 = ID	x_2 = DD	x_3 = TR	⋯	x_p
1	x_{11}	x_{12}	x_{13}	⋯	x_{1p} → observations on multiple variables
2	x_{21}	x_{22}	x_{23}	⋯	
3	⋮	⋮	⋮	⋯	
n	x_{n1}	x_{n2}	x_{n3}	⋯	x_{np}

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \\ \\ x_{ip} \end{pmatrix}$$ → Each row can be taken out separately

i^{th} Observation on 'p' variables

When each observation has multiple values, we can it a **multivariate**

Linear combination of variables x_1, x_2, x_3 like $\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ with empirically determined weights $\beta_1, \beta_2, \beta_3$ is known as **variate**

$p = 1$, Univariate
$p = 2$, Bivariate
$p \geq 2$, Multivariate

- Field Based Observation
- Expt
- Naturalistic

Lecture 2

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \\ \\ x_{ip} \end{pmatrix}$$ → Check if features are correlated or Covariance

Correlation Matrix: $\begin{bmatrix} & \\ & \end{bmatrix}_{p \times p}$

$p \times p$

Mean Matrix μ: $\begin{bmatrix} \mu_1 \\ \mu_2 \\ \\ \mu_p \end{bmatrix}_{p \times 1}$

Behaviour should be analysed for design, improvement

→ Will lose substantial information if covariance/correlation is not considered

Random Variable:

1. Discrete
2. Continous

Data Types:

1) Nominal:

→ Provide identity to some things or items

→ Examples: Months, dept name, brand name

→ Computational limitation: Mathematical operations can't be done

2) Ordinal:

→ Provide order or rank to some item or things

→ Examples: Low, Medium, Good

→ Computational limitations: Mathematical operations can't be done

→ Better than nominal cuz we get a rank for organisation and comparison

3) Interval:

→ Provides continous data in a range

→ Example: Temperature

→ Computational limitation:
- Division can't possible

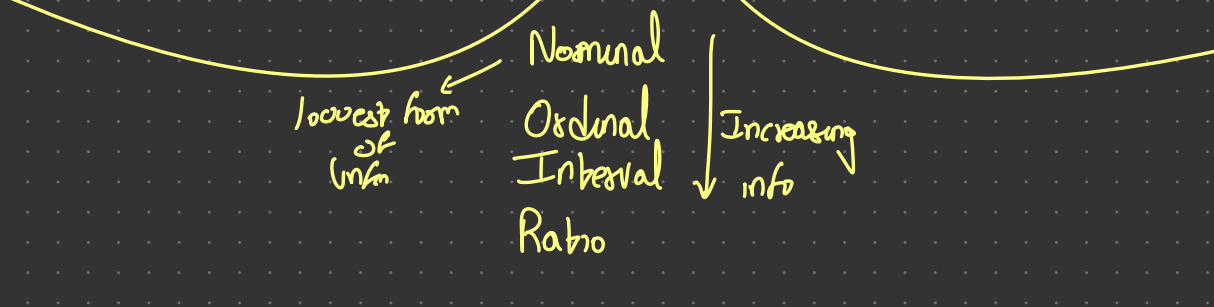
4) Ratio:

→ Contains absolute zero

→ Highest form of data

→ Example: Hours, % etc

→ No computational limitations



Data Sources:

Primary: Collected from source

Secondary: Stored in repo or collected by someone else

Tertiary: Common Knowledge through Wikipedia etc

Model:

- → Mimic reality
- → Explain regularity of phenomenon
- → Can be extended to law or theory

Modelling:

- → Process of building a model
- → Physical, Mathematical, Statistical

Data = Pattern + Errors

↑

Regular/Systematic Component

Principles of modelling

- → Build simple if possible
- → Modelling prob to fit technique (Because)
- → Design phase must be rigorous
- → Verify model before validation
- → Don't take model literally
- → Prob specific
- → Don't oversell model
- → Primary benefits associated with process of developing model
- → Can't replace decision makers
- → Limited to info provided

Lecture-3

Data = Pattern + Errors

↑

Extract this using statistical approaches to problem solving

Practical Problem → Statistical Problem → Statistical Soln

↓

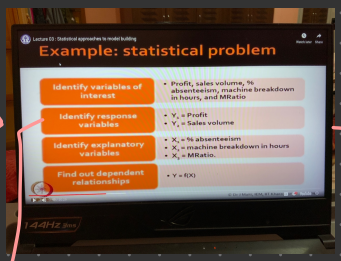
Practical Soln

Example:

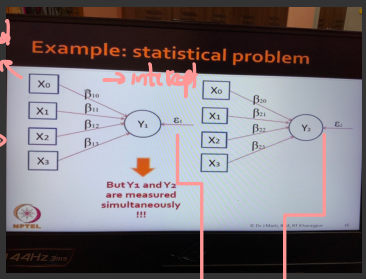
Threat to profit

&

'Sales volume'



Response = Dependent Variables



Error components

Purpose of multivariate modelling:

Description → Explanation → Prediction

System details
Type of data
Problem faced

Dependence of relationships of features

Predict future outcomes

2 types of Multivariate Models

① Dependence model

② Interdependence model

① Response Variables ← Explanatory Variable
(Explains)

② [All variables under one bracket]

Dependence Models:

→ MANOVA
→ Regressions
→ Path Models
→ Structural Equation Modeling
→ Discriminant Analysis

Interdependence Models:

→ Principal Component Analysis
→ FACTOR ANALYSIS
→ Cluster Analysis

MANOVA : Multivariate Analysis of Variance

Lecture 5

Univariate Descriptive Statistics
- Central tendency
- Dispersion

Univariate Inferential Statistics
- Estimation
  - Point Estimation
  - Interval Est
- Hypothesis testing

Population:
- Entirety / Totality of the whole
- Example: Population etc
- Generally a large no of items

Boundary:
- Time Boundary
- Space Boundary

① Finite:
- Size is known
- Sampling with replacement

② Infinite:
- Size is infinite
- Sampling with replacement

Population is characterised by different variables applicable to the population

Discrete Population parameters: $[\mu, \sigma]$

$\text{Mean} = \mu = E(x) = \sum_{\text{all } x} x f(x)$ — x: Discrete RV; $f(x)$: Probability Mass Function

$\text{Variance} = \sigma^2 = E(x-\mu)^2 = \sum_{\text{all } x} (x-\mu)^2 f(x)$

μ, σ^2 are constants for a population (generally)

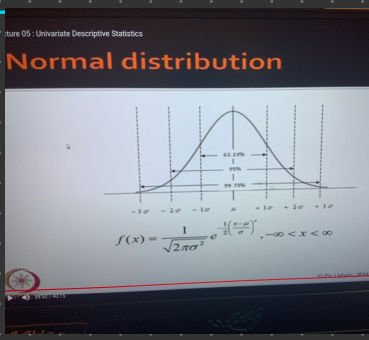
With population parameters, we get a mathematical interpretation

Continuous Population Parameters: $[\mu, \sigma]$

$\text{Mean} = E(x) = \mu = \int_{-\infty}^{\infty} x f(x)\, dx$ — $f(x)$: Probability Density Function

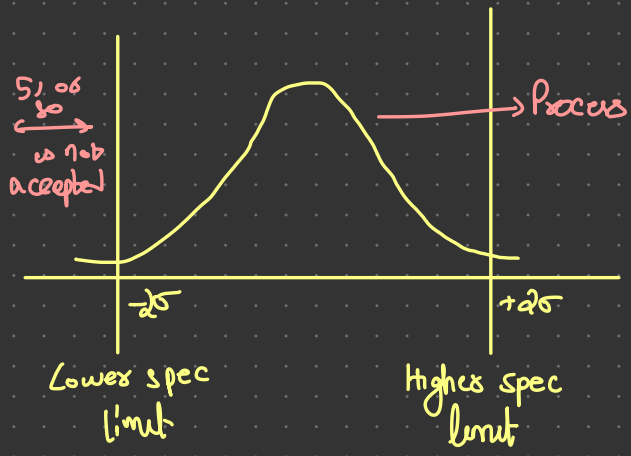
$\text{Variance} = \sigma^2 = E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$

Discrete	Continuous
→ Binomial	→ Normal
→ Poisson	→ Lognormal
→ Negative Binomial	→ Exponential
→ Geometric	→ Weibull
→ Hypergeometric	→ Gamma

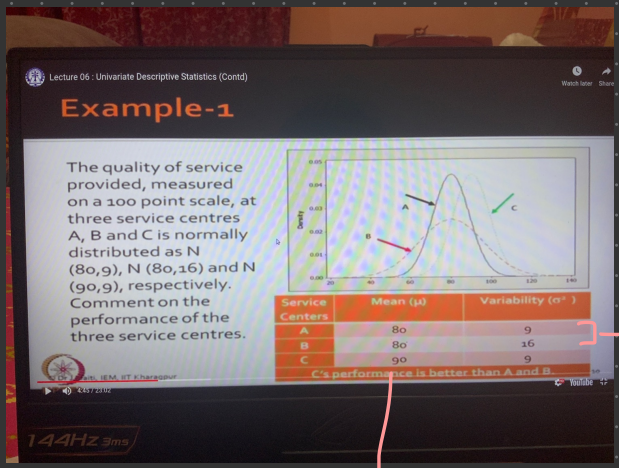


μ

Lecture 6



Example :



→ A is better than B

Higher mean and lesser variability

Variability is very important

Easier to learn as it can be shifted

Population parameters → Prob distribution of variable of interest (x)

Before data collection

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \\ X_N \end{bmatrix} \rightarrow \text{Random \& Unknown}$$

After data collection

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \\ x_n \end{pmatrix} \rightarrow \text{Sample of size } n \text{ which is known and fixed}$$

Mean, Median, Mode

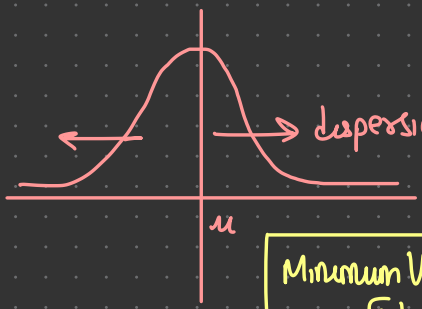
↗ Most common

- Mean ↓ Average
- Median ↓ $\left(\frac{N+1}{2}\right)^{th}$ position
  - If decimal, take average of floor ceiling
- Mode

Dot Plot

No of values helps to find mode

Measure of dispersion



- → Variance = $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$
- → Range = Max − Min
- → Interquartile Range = 3^{rd} quar − 1^{st} quar

$3\left[\frac{N+1}{4}\right]^{th}$ $\left[\frac{N+1}{4}\right]^{th}$

Minimum Variance Unbiased Estimator