12.02.2022

✓ Univariate.

Four point { Mean, S.d. skewness & Kurtosis.

Five point — Min Q1 Q2 Q3 Max

EDA.

Exploring Data Analysis

Descriptive ✓ → Univariate, Bivariate, Graphical, Explore

Diagnostic ✓

Traditional:

Frequency, Mean, Median

Prescriptive

Data → Primary, Secondary

Random experiment

Coin

(1) Probability: variables, Prob.

(2) Conditional Prob.

(3) Bayes Th.

(4) Random variable → Discrete, Continuous

Discrete. Ex.

(1) ✓ Countable Finite.

(2) Countable infinite. ✓

20 coins are tossed.

How many Heads

✓ No. of mobile calls.

✓ No of deaths.

✓ No of births

✓ Continuous Ex.

Income, Age →

Images

uncountable. 0-1. 0.1, 0.01, 0.001

Discrete r.v. → Prob. → P.m.f

Prob. means summation

Continuous → Prob. → Prob. density function → p.d.f

$\int$

$\sum$

Most Most imp

Pattern of Data

Parametric ✓ ← Non-Parametric ✓

Discrete

Continuous

Theoretical Prob. Distribution

(1) Binomial (2) Poisson → (Discrete)

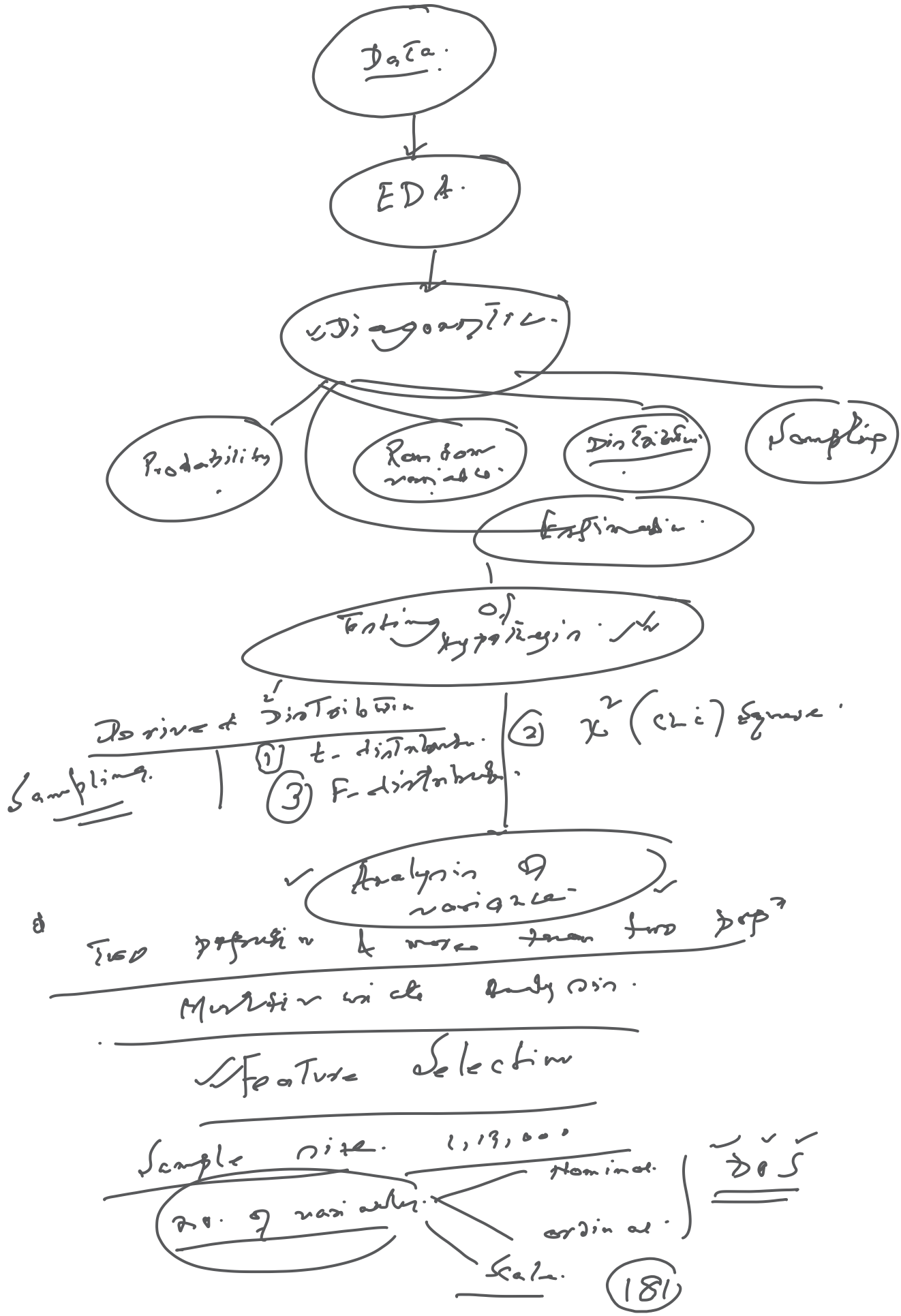
Theoretical Continuous Dist. → Normal

Sampling

Ex: To estimate population parameter with the help of sample statistic.

To minimize error for ⇒ Dimension.

Estimation


Data.
EDA.
Diagnostic.
Probability
Random variable.
Distribution
Sampling
Estimation.
Testing of hypothesis.
Derived distribution
Sampling.
(1) t-distribution.
(2) χ^2 (chi) square.
(3) F-distribution.
Analysis of variance
Two population & more than two pop
Multiple variable analysis.
Feature selection
Sample size. 1,13,000
No. of variables.
Nominal.
ordinal.
Scale.
DOS
(181)
1,13,000 - rows.
(281 - Col.

① Some variables / some attributes are not imp => Discard using some statistical methods.

② Some variables / attributes can behave of a homogeneous nature. They are highly associated each other.

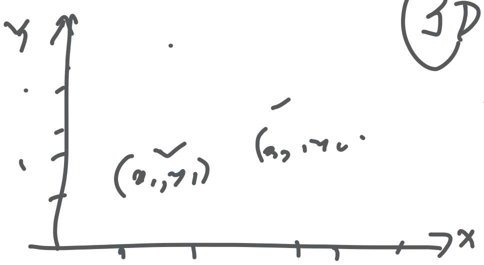
Several variables P C M B E H G.

100 { underlying construct

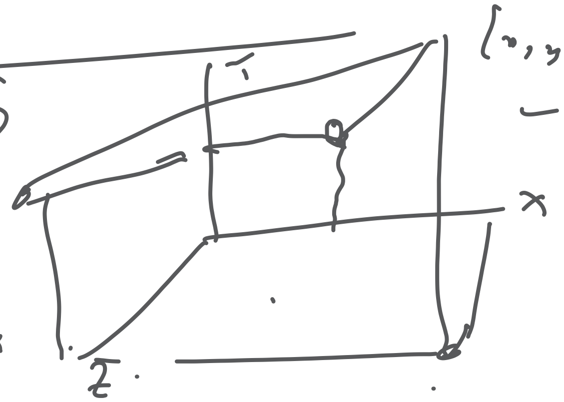
P C M → Group 1. Basic science. B E H G Language.

Dimension Reduction.

2D



3D



4D

Predictive model.

① Cross sectional ② Longitudinal ③ Panel.

Predictive model

Dependence (supervised)
Output = f (inputs)
Dependent variable / independent variable.

inter-dependence (unsupervised)
Inputs.

① Regression / Time Series

② PCA, Factor Analysis

Classification

Cluster Analysis

Prescriptive

Optimization

Local → Global

min

Try to optimize error

DDP

DDP

$\begin{cases} P = \text{population} \\ S = \text{Sample} \end{cases}$ $|P - S| \leq \epsilon$

ϵ = Smallest gap in the universe.

$(10^{-10}, 10^{-9}, \ldots)$

Stock Prices

Year \ Country	USA	Japan	UK	India	China
2001					
⋮					
2021					

Country | Stock Prices → Probability

Crores → ✓

(1) Trial (2) Event (3) Mutually exclusive event (4) Exhaustive Cases (5) Equally Likely Cases

Tossing a coin (Die) — Trial.

Outcomes — Head, Tail.

Possible Cases.

Coin {H, T}

Dice {1, ..., 6}

{1, ..., 52}

Defⁿ. Prob. = Favourable no of Cases of an event / Total no of " " " "

$= \frac{m}{n}$

Voting. H = 100

60 favours A. $P(A) = \frac{60}{100}$

40 " B. $P(B) = \frac{40}{100}$

A B.

$p + q = 1$. Total Prob = 1.

Prob = $\frac{m}{n}$. $0 < p < 1$

$p = 0 \Rightarrow$ Impossible event

$p = 1 \Rightarrow$ Certain event

$0 <$ Uncertain event < 1.

60 — A $P(A) = \frac{60}{100}$

$100 <$

40 — B $P(B) = \frac{40}{100}$

$P(A) + P(B) = 1$.

$N = \{1, 2, \ldots, 10\}$ $A = \{1, 2, 3, 4, 5\}$ $n(A)$ = cardinality of A = number of elements in it = 5 ✓

$= 10$

$B = \{3, 4, 5, 6\}$.

$n(B) = 4$ ✓

$A \cup B = \{1, 2, 3, 4, 5, 6\}$

$n(A \cup B) = 6$.

$A \cap B = \{3, 4, 5\}$ $n(A \cap B) = 3$ ✓

$n(A \cup B) = n(A) + n(B) - n(A \cap B)$

$6 = 5 + 4 - 3$

$\frac{n(A \cup B)}{N} = \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N}$

Favourable no of cases

$\Rightarrow \boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$ GOC

when A & B are not mutually exclusive events

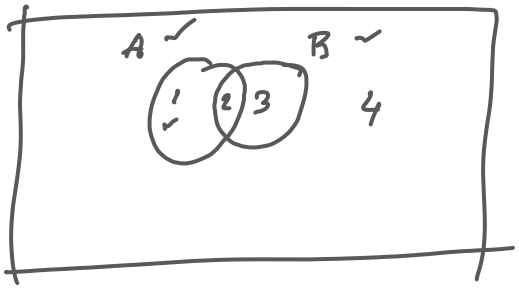
A B.

Only A : 1 = $n(A \cap B^c)$

A & B Both : 2 = $n(A \cap B)$

Only B : 3 = $n(A^c \cap B)$

No A & no B : 4 = $n(A^c B^c)$



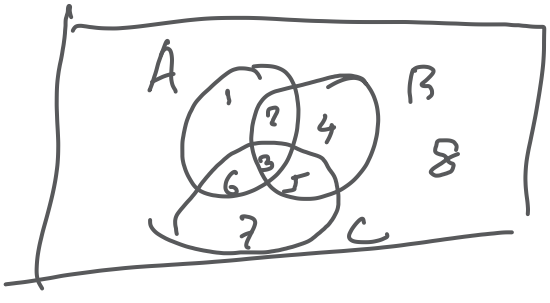
~~$P(AB^c) + P$~~

1 = $n(AB^c C^c)$

2 = $n(AB^c C)$

3 = $n(ABC)$

4 = $n(A^c B C^c)$



$5 = n(A^c B C)$ $6 = n(A B^c C)$ $7 = n(A^c B^c C)$

$8 = n(A^c B^c C^c)$

when there are two sets,

How many points $= 2^2 = 4$.

When there are three sets $= 2^3 = 8$.

p sets / p variables $= 2^p$

$p = 20$ $2^{20} = 10,48,576$

Dimension reduction

Current Salary depending on

Beginning sal, Prev. Exp, Qualification,

opening since Stines

(4) $2^4 = 16$

Stock - mkt

Stock Price

macro-economic Parameters

Sensex $= f($Gold Price, Dollar, RR, foreign FDI, forex, foreign$)$

↓

Dr

$2^6 = 64$

covid. RTPCR.

	Age group	Total intake	No of death
G1	< 30	5000	5
G2	30 - 50	3000	15
G3	≥ 50	2000	40
		N = 10000	60

$p(D) = \frac{60}{10000}$

$= 0.6\%$

$P(D/G_1) = 5/5000 = 0.1\%$

$P(D/G_2) = 15/3000 = 0.2\%$

$P(D/G_3) = 40/2000 = 2\%$

Conditional

Prob.

Unconditional

Conditional

Stock A.

60 months

Min ϕ_1, ϕ_2 ϕ_3 max

$f_1 + f_2 + f_3 + f_4 = N = 60$

G_1 : min ϕ_1 — $f_1 = 5$

G_2 $\phi_1 - \phi_2$ $f_2 = 15$

G_3 $\phi_2 - \phi_3$ $f_3 = 30$

G_4 ϕ_3 max $f_4 = 10$

$P_1 + P_2 + P_3 + P_4 = 1$

$P(G_1) = 5/60$

$P(G_2) = 15/60 = 1/4$

$P(G_3) = 30/60 = f_2$

$P(G_4) = \frac{10}{60} = 1/6$

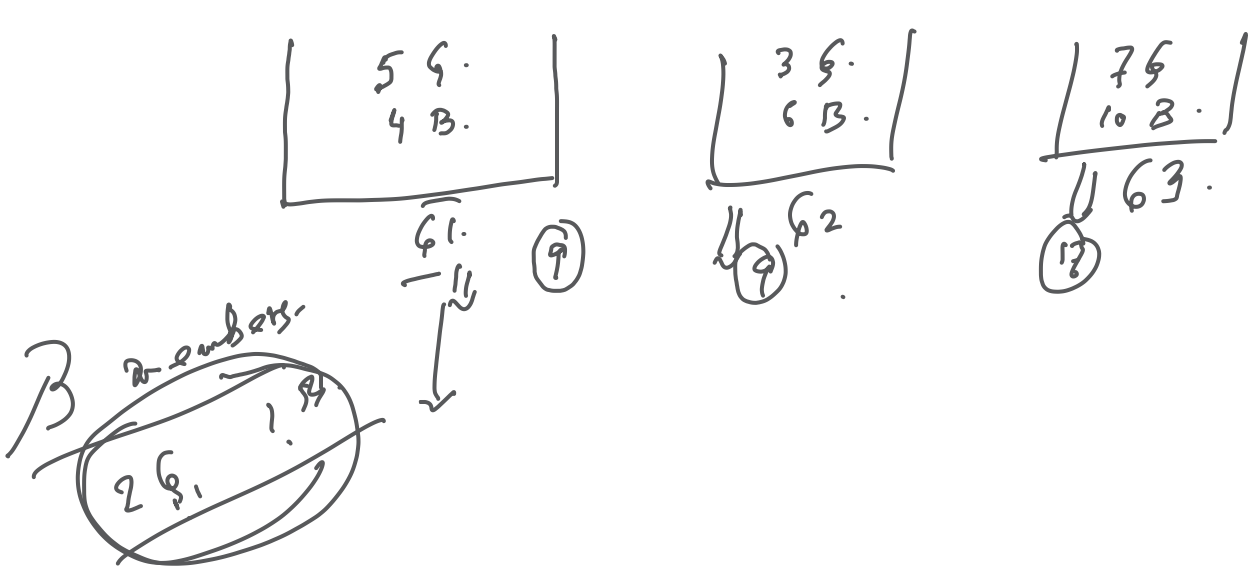
Prob.

Prior (which is known given from the information)

Posterior

BAYE'S

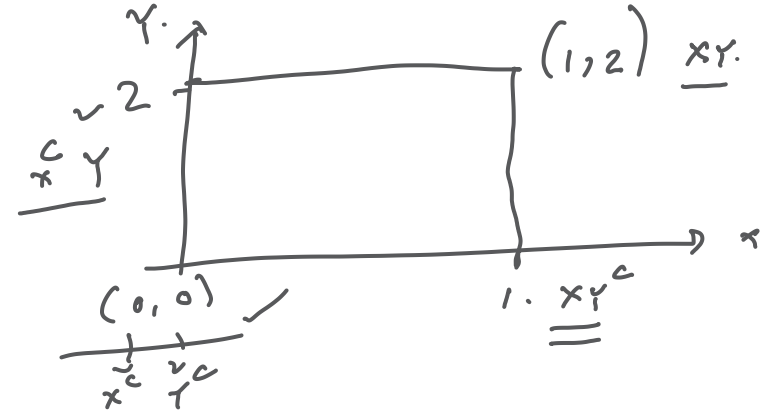
which are new forecast
Diff the help of prior



$A = \{1, 2\}$

no of members = 2.

∴ Total no of subsets $= 2^2 = 4$



$X^c Y$

XY = Power set