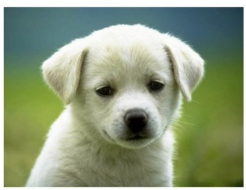
Day 3 - Miscellaneous -2

Session-2

Jan 14, 2024



Agenda

1. 2 - Way anova
2. KS - Test
3. AB - Test
4. Parametric Vs Non - parametric tests
5. Feature Engineering
   - → Heatmap
   - → Skewness & Kurtosis
   - → Outlier treatment
   - → Normalizing Vs Standardizing

2- Way Anova

- One - way anova ??

Cat Vs Numerical

↓

→ 2 Cat - variable

Aerofit Example → 3 Products Vs (num)

10 Products vs Num. Value

① T-test → 2 Products → T-val

↓

P-value

how many times ??

$^{10}C_2$ ⇒ 45 tests

$^{1000}C_2$ → we don't do this

→ too troublesome

(2) Type-II / I error

0.05 → alpha (Level of significance)

P-value < alpha → Ho actually is still true

Incorrectly reject Ho

0.05 - chance of error in T test

45 time

45 x 0.05

→ We don't use T-test for more than 2-cat value

→ anova → more than 2 cat values

1 cat - variable

→ 2-way anova → 2 seperate independent cat variables at once

(1) Flavour has an impact → sales
(2) location where it's sold → sales
(3) Impact of flavour on sale

depends on location or not

Interaction effect — 2-way anova

II Diet / Exercise — Weight loss

→ Does diet impact weight loss or not
→ " Exercise " " " " "

→ Does impact of diet on weight loss

impacted by exercise or not

III Impact of Fertilizer type
Watering Frequency,
growth of plant

1. Impact Fert. → growth
2. " watering → "
3. Impact of Fert. on growth

impacted by watering freq or not.

ttest_1samp(f ?, value)

→ coming from &

t-test → 1 Cat Varg:

→ multiple → cat Vals

drug/2 ← 2 -cat Valu

↳ num counterpart

F- one way anova

↳

Null hyPo.

→ Flavour: No impact on Sale

location: " " " "

Interaction: There's no interaction b/w Flavour & location

IRS - Test

$M_1 \rightarrow [5, 6, 8, \ldots]$ → mean std

$M_2 \rightarrow [9, 10 \ldots]$ → mean std

2 talk samples

Use both ttest & z-test

ttest approximates z-test

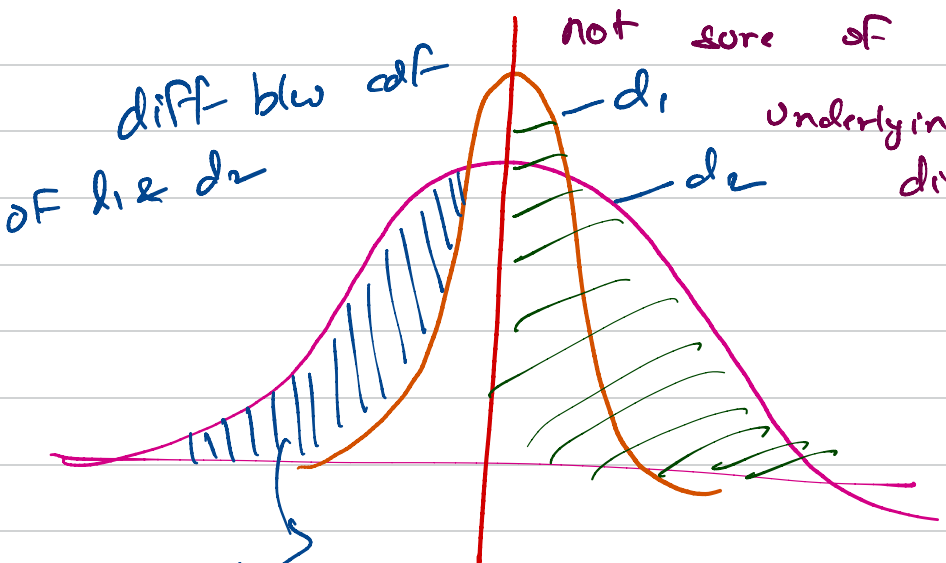
For large samples

→ Here you make assumptions

(1) Normality

K S - Test

→ distribution - Free



① Cat Vs Num → < 2 cat val → ttest
→ > 2 cat val → one-way anova
↳ 2 cat columns vs num → 2-way anova

② Cat Vs Cat → Chi-square test

③ Num Vs Num → Pearson corr
→ Spearman ..

KS – test

H_0: The two dist are same

$\mu_1 = \mu_2$ ← Reject this???

H_a: Not same

A B – Test

→ Amazon → M_1 model → 80% / stable model

M_2 model → 90%

↓

offline data

Concerns

① Acc may not hold in real data

② May not be stable → Low throughput

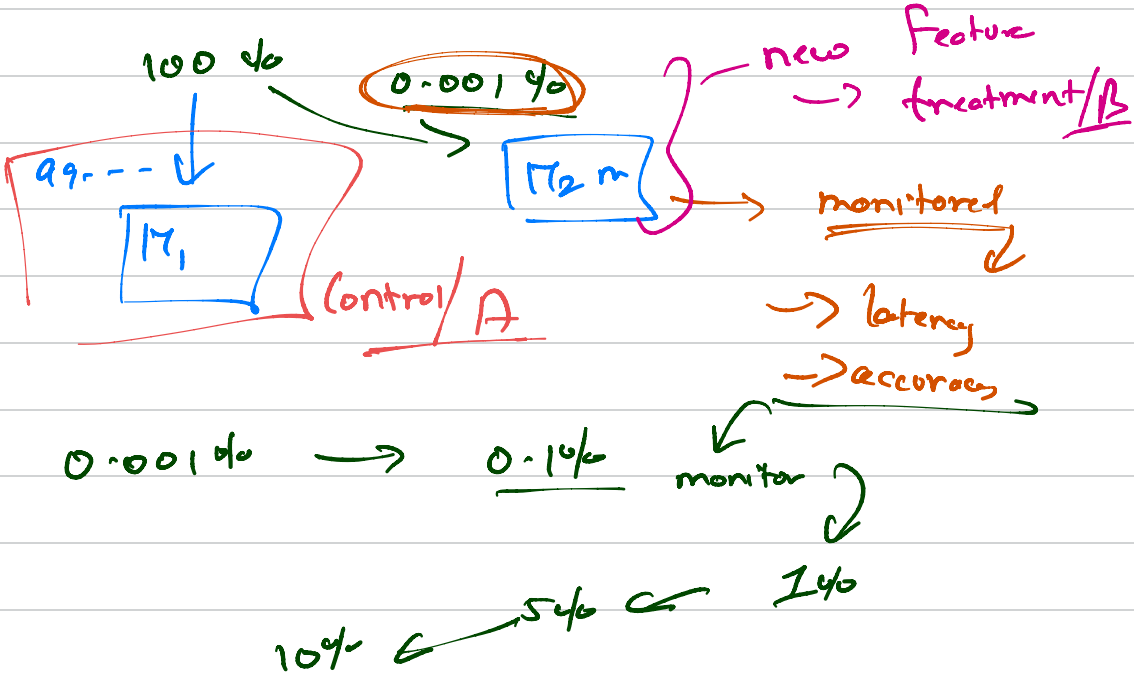
→ high latency.

U.U.U. accurate → entire day??

Compromise b/w

Speed & Accuracy.

M_2 model → live traffic



A/B Test

Drug / Medicine

(200)

(Random)

(100) Control or A (Paracetomol) or Placebo

(100) Treatment or B (New medicine)

days to recover

A/B Test

randomized control-Treatment test

old model

recommendation algo

New model

V2 - recommendation algo

Engagement metric

$$\frac{\#\text{ of engaged user}}{\#\text{ of user}}$$

Z-Proportion test

throughput → no of requests you can handle in one minute

latency → amount of time → respond to one req

10 ms

→ $\frac{1000}{60}$ → $\frac{60}{1000}$

throughput → 1000 → 1 min

→ 1 → $\frac{1 \text{ min}}{1000}$

A D Sense → one of the highest throughput

4 month → women → birth

2 women → 4.5 month

[— — — —] num data

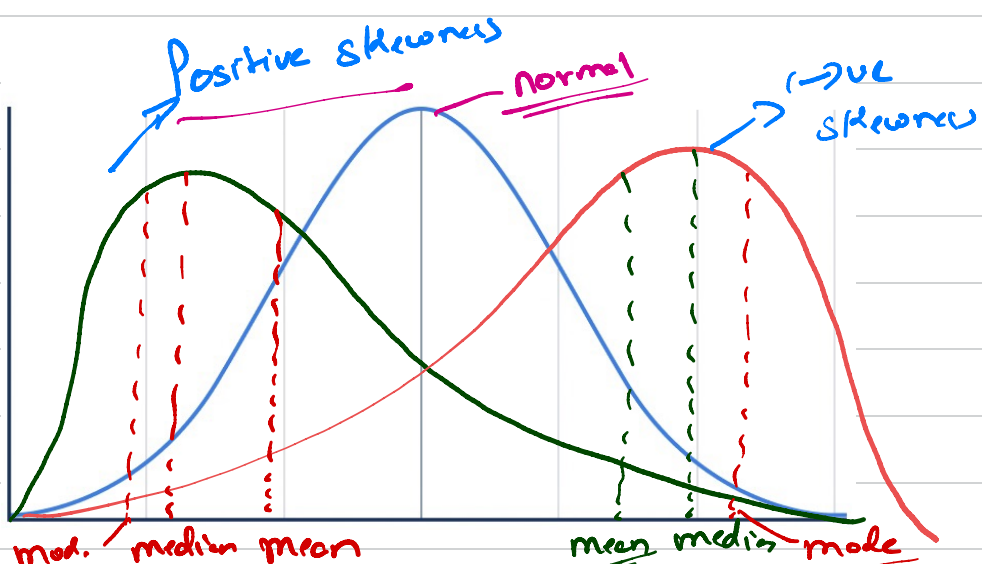
[— — —] num data

Parametric Vs Non-Parametric

↓ make assumptions data dist

↓ They don't

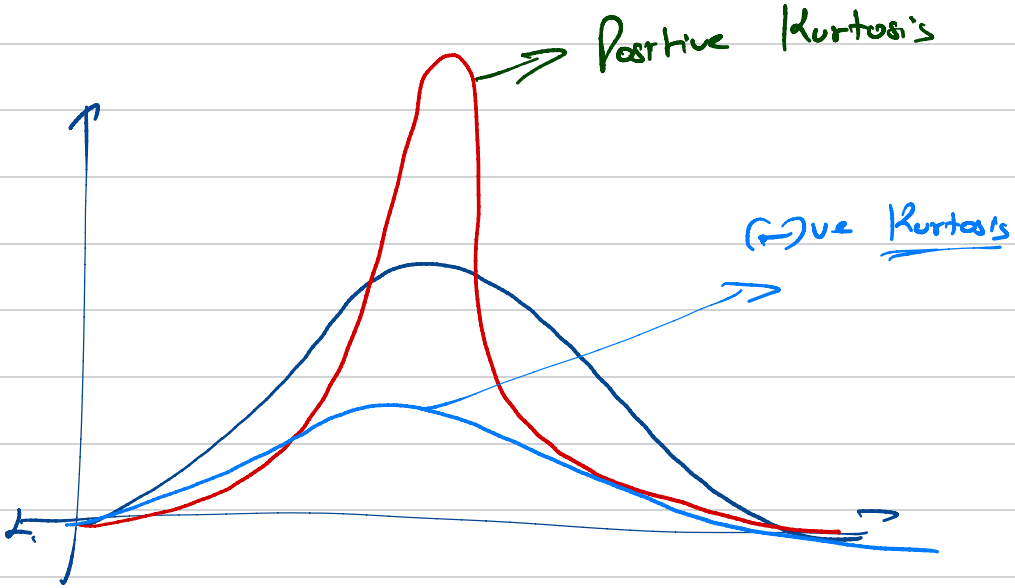
SKEWNESS



→ +ve Skewness median < mean

-ve " median > mean

KURTOSIS



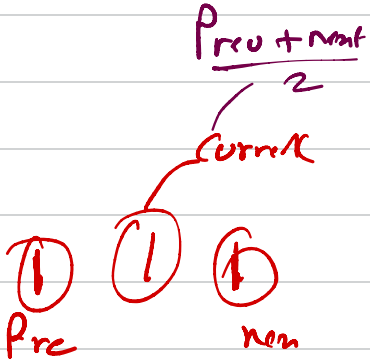
Leptokurtic → Pointy → +ve Kurtosis

Mesokurtic → normal → normal

Platykurtic → Flat one → -ve Kurtosis

→ Remove missing Value

1. Mean
2. Median
3. Mode
4. Interpolate



Simple Imputer

OUTLIERS

1. IQR → Q_1 percentile → 25%

Q_3 Percentile → 75%

IQR $(Q_3 - Q_1) \times 1.5$

upper $Q_1 - IQR$

lower $Q_3 + IQR$

} Points outside upper & lower outliers

(2) Z-Score → Compute z-score for all the values

$\frac{\bar{x} - \mu}{\sigma}$

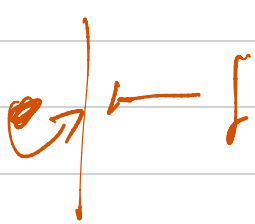
< 10 → Z-score

50 < 50 =50

upper — mean + 3×σ

lower — mean − 3×σ

} Remove them



Standardize Vs Normalize

Normalize → Numerical Column

→ [0, 1]

Scale of 0 to 1

a = [1, 2, 3, 4, 5]

$$\frac{a[i] - \min(a)}{\max(a) - \min(a)}$$ } Formula for normalizing

ML model → house price prediction

Feature → no of rooms

Ex-price → 120,000, 80,000

directly → ML model

Standardize → Reduce data

→ 0 mean

1 Variance

$a = [1, 2, 3, 4, 5]$

$$\frac{x - \text{mean}}{\text{std}}$$

When do we use which??

If you know Range → go for normalization

" " don't " " → Standardization

don't have gaussian dist

Have gaussian dist

Stock market