Introduction To Stats

7 days Session

7pm - 8pm → 8:30pm

① Basics To Advance

{Data Scientist, Data Analyst, Business Intelligence Tools}

2 days Basics

① Descriptive Stats

↓

{ ① Measure of Central Tendency
② Measure of Dispersion }

Summarizing the data.

Histograms, pdf, cdf, Probability, Permutation, Mean, Median, Mode, Variance, Standard deviation

i) Gaussian Distribution
② LogNormal Distr
③ Binomial Distr
④ Bernoulli's Distr
⑤ Pareto Distr {Power law}
⑥ Standard Normal Distr → Python
⑦ Transformation and Standardization → Python
⑧ Q-Q plot
⑨

② Inferential Stats

⑧ Z test ⇔ python
t test ⇔ python
ANOVA → F Test
CHISQUARE
HYPOTHESIS TESTING {P values}
Confidence Intervals

Z table, t table

What is Statistics?

Statistics is the science of Collecting, organizing and analyzing data. {Better Decision Making}

Dyn Data??

Data : Facts or pieces of information that can be measured

Eg: The IQ of a class

{98, 97, 60, 55, 75, 65}

Ages of students of a class

{30, 25, 24, 23, 27, 28} → DATA

Types of statistics

① DESCRIPTIVE STATS

It consists of organizing and summarizing data

② Inferential Stats

Technique where in we used the data that we have measured to form Conclusions

① Classroom of Maths student (20)

Marks of the 1st Sem

84, 86, 78, 72, 75, 65, 80, 81, 92, 95, 96, 97, - - - - -

Eg : Descriptive Stats

① What is the average marks of the students in the class?

Inferential Stats

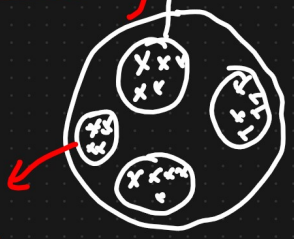
Sample [1] ← 5 → Population.

Eg : Are the marks of the students of this classroom similar to the age of the Maths classroom in the college?

Population And Sample

Elections → Goa, UP → Goa population

EXIT POLL → Maximum

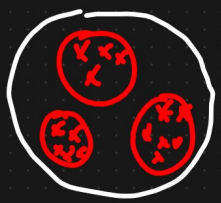


Sample data

Population (N) Sample (n)

Sampling Techniques

① Simple Random Sampling : Every member of the population (N) has an equal chance of being selected for your Sample (n)



② Stratified Sampling ÷ Where the population (N) is split into non-overlapping groups (Strata)

Eg: Gender → Male ✓ → Female ✓ Survey

Age group

(0-10) (10-20) (20-40) (40-100)

Eg: Profession : Dot net PHP Python

Doctors, Engineer

③ Systematic Sampling Thanos

(N) → n^{th} individual

Eg: Mall → Survey (Covid)

↳ 8th person → Survey

④ Convenience Sampling

↳ Survey ← Only those people

↳ DATA Science Ubi cax

Eg: EXIT POLL RBI → Household Survey

{Random Sampling} ⇓

Survey → Women

Eg: Drug → Tested ⇒

What Kind of Sample ??

Variables :

A variable is a property that can take on any value

Eg: Height = {182, 178, 168, 180, 160, 170}

Weight = {78, 99, 100, 60, 50, }

Two Kinds of Variables

① Quantitative Variable → Measured Numerically, {Add, Subtract, multiply, divide}

Eg: Age, Weight, Height

② Qualitative / Categorical Variables.

↳ Eg: Gender → M, F {Based on some characteristics we can denote Categorical Variables}

Eg: IQ

0-10 → Less IQ

10-50 → Medium IQ

50-100 → Good IQ

Blood Group: A+ve

Tshirt size: L, XL, M, S

Quantitative

- Discrete Variable
  - Eg: Whole number
  - No. of Bank Accounts
- Continuous Variables
  - Eg: Height : 172.5, 162.cm, 163.5cm,
  - Weight = 100kg, 99.5, 99.75

Eg: 2, 3, 4, 5, 6, 7, Rainfall = 1·1, 1·25, 1·35 - - -

② Total of children in a family

Eg: 2, 3, 4, 5,

Eg: What kind of variable Gender IS? Categorical

② What " " " Marital Status? "

③ River length? Continuous

④ population of the State is? Discrete

⑤ Song length? Continuous

Blood pressure? Continuous.

PIN CODE? { Discrete or categorical }

④ Variable Measurement Scales

4 types of Measured Variable — Colors, Gender, Type of flower

① Nominal data { Categorical data } → Classes

② Ordinal → Order of the data matters, value does not

③ Interval → Order matters, values also matter, natural zero is not present

④ Ratio.

Eg:

Students (Marks)	Rank
100	1
96	2
57	4
85	3

→ Ordinal Data

44 5

Temperatures

Fahrenheits

70 - 80 80 - 90 90 - 100 | 0 | Fahrenheit

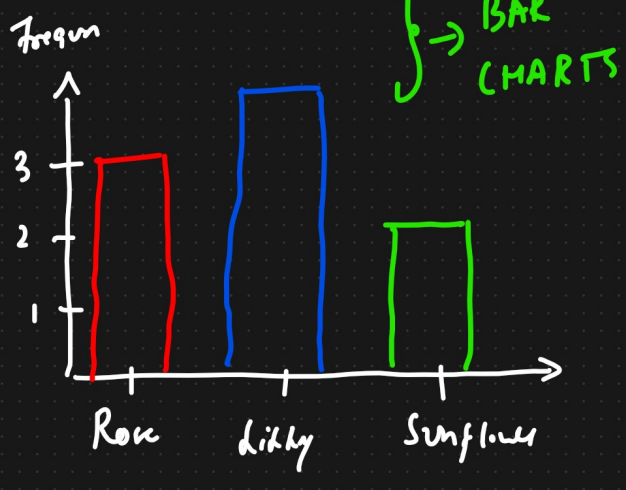
(F) Ratio data {Assignment}

Frequency Distribution

Sample dataset : Rose, Lilly, Sunflower, Rose, Lilly, Sunflower, Rose, Lilly, Lilly

Flower	Frequency	Cummulative Frequency
Rose	3	3
Lilly	4	7
Sunflower	2	9

(1) BAR GRAPH



} → BAR CHARTS

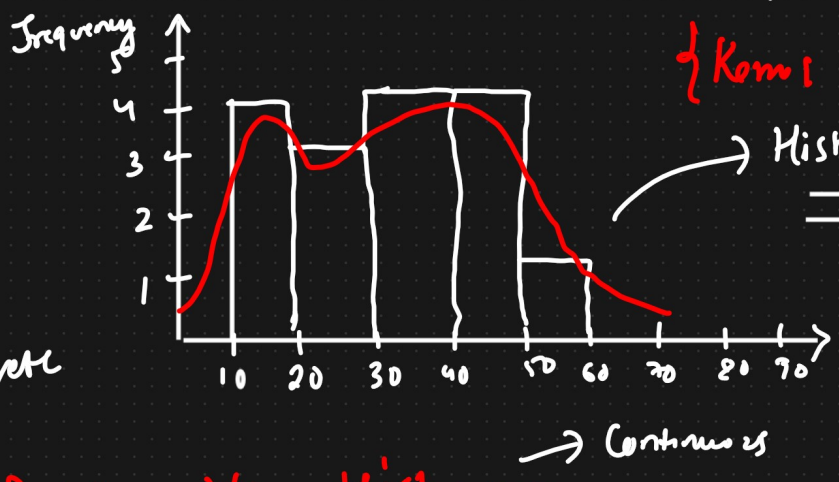
② Histograms : Continuous

Ages = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51 }

↳ Bins = 10

pdf : Smoothening of histograms

{ Kernel density Estimator }



Discrete ↑

BAR Vs Histogram

pdf : probability density function

BASICS To Intermediate Stats

1. Measure of Central Tendency
2. Measure of dispersion
3. Gaussian Distr
4. Z score
5. Standard Normal Distr

1. Arithmetic Mean for Population & Sample

Mean (Average)

Population (N) → $X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 3.2$$

Sample (n)

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$= 3.2$$

2. Central Tendency

1. Mean ✓ 2. Median ✓ 3. Mode ✓

Refers to the measure used to determine the centre of the distribution of data.

{ 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100 }

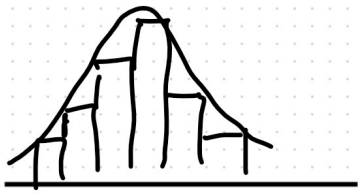
100 → Outliers

Mean = $\frac{32+100}{11}$

$= \frac{132}{11} = 12$

$\mu = 3.2$ + 100 → $\mu = 12$

Distribution



Median { 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112 }

1) Sort the numbers

12

$\mu = 3.2$

$\mu = 12$

Odd number = 11

$\mu = 12$

Median = 3

Median = 3.5

Avg = $\frac{3+4}{2} = 3.5$

Median = 3

Median 3.5

{ Median works well with outlier }

② Mode = { 1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 100, 100, 100 }

2 3

Mode = { Most frequent Element }

Mode : 6 → Measure of Central Tendency

Type of flower | petal length | petal width

Rose

Lilly

Sunflower

Mode

{ Missing Value → Most frequent occ element }

DATASET

10% Missing data

(Categorical Variable)

Ages of Students

Age
25
26
—
—
—
32
34
38

⇒ Mean? ✓
Median?
Mode ??

(F) Measure of Dispersion
① Variance
② Standard Deviation

{ Dispersion } ⇓ Spread

μ

{ 1, 1, 2, 2, 4 } → $\frac{10}{5} = 2$

{ 2, 2, 2, 2, 2 } → $\frac{10}{5} = 2$

① Variance ✓

Population Variance

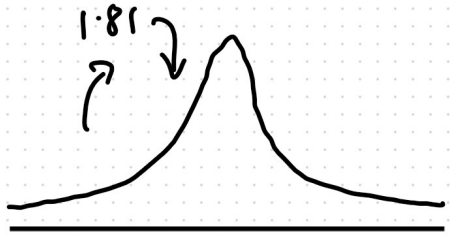
$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} = \frac{10.84}{6} = 1.81$$

Sample Variance

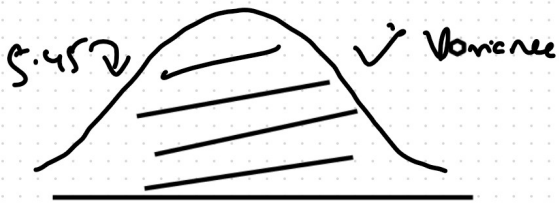
$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

x	μ	$x-\mu$	$(x-\mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	+0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\mu = 2.83$



Variance is more ??

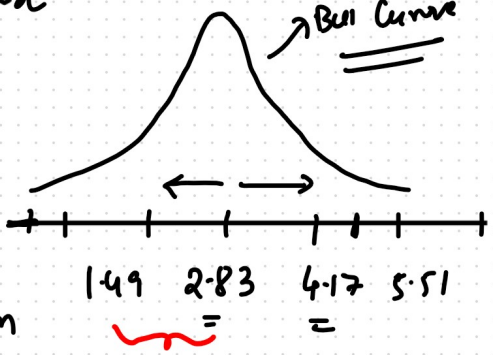


$\sigma = \sqrt{\text{Variance}} = \sqrt{1.81} = 1.345$

Variance = Spread

S.D = 5

↳ 1.5 σ from the μ



2.83 − 1.34 = 1.49

2.83 + 1.34 = 4.17

4.17 + 1.34 = 5.51

ⓨ Percentiles And Quartiles {Find outliers}

Percentage : 1, 2, 3, 4, 5

% of the numbers that are odd?

$\% = \frac{\#\text{ of numbers that are odd}}{\text{Total Numbers}}$

$= \frac{3}{5} = 0.6 = 60\%$

Percentiles (GATE, CAT, GMAT, SAT)

↳ Defn : A percentile is a value below which a certain percentage of observation lie

Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

5.25

11?

What is the percentile ranking of 10? n = 20

Percentile Rank of $x = \frac{\#\text{ of values below } x}{n} \times 100$ ($x = 10$)

$= \frac{16}{20} \times 100 = 80\%$

11?

$= \frac{17}{20} \times 100 = 85\%$

(2) What value exists at percentile ranking of 25%? 75%??

$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$

$= \frac{25}{100} \times (21) = 5.25$ → Index position

5 → 25%

$= \frac{75}{100} \times (21) = 15.75 \Rightarrow 9$ Answer

↳ index

Five Number Summary

1. Minimum
2. First Quartile (Q1)
3. Median
4. Third Quartile (Q3)
5. Maximum

Removing the outliers

Outlier ?? [27 > 13]

{1, 2, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

25% 75%

$\frac{25}{100} \times (19+1)$

$\frac{25}{100} \times 20 = 5 \rightarrow$ index

$= 3 \Rightarrow$

so

[Lower fence ⟷ Higher fence]

Q1 = 3

Q3 = 7

Lower fence = Q1 − 1.5(IQR)

Upper fence = Q3 + 1.5(IQR)

Q3 = (75%)

Q1 = (25%)

Interquartile Range (IQR) = Q3 − Q1

= 7 − 3 = 4

Lower Fence = Q1 − 1.5 (IQR)

= 3 − 1.5 (4)

= 3 − 6 = −3

[Lower Fence ⟷ Higher Fence]

[−3 ⟷ 13]

Higher Fence = Q3 + 1.5 (IQR)

= 7 + 1.5 (4)

= 7 + 6 = 13

Remaining data

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

Minimum = 1

Q1 = 3

Median = 5

Q3 = 7

Max = 9

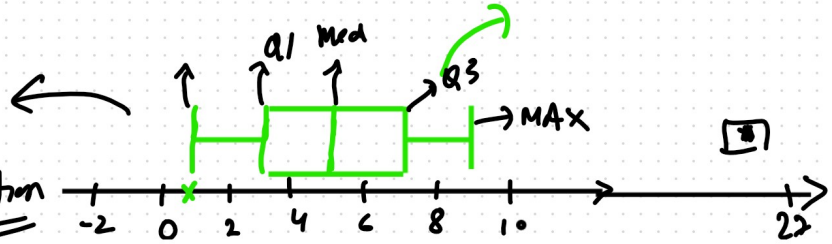
→ 5 Number Summary

⇓

Box plot

27

Box plot



DATA Visualization

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$n-1$ → Bessel's correction
↳ Degree of freedom

STATS

Why sample variance is $n-1$?

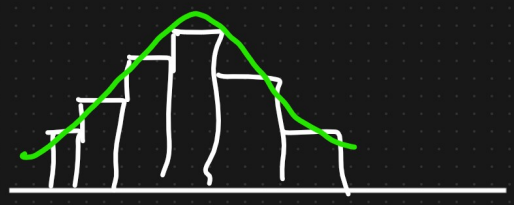
Day 3 - Advance Statistics

① Distributions
- Normal Distr
- Standard Normal Distr
- Z score
- Log Normal Distr
- Bernaullis Distr
- Binomial Distr

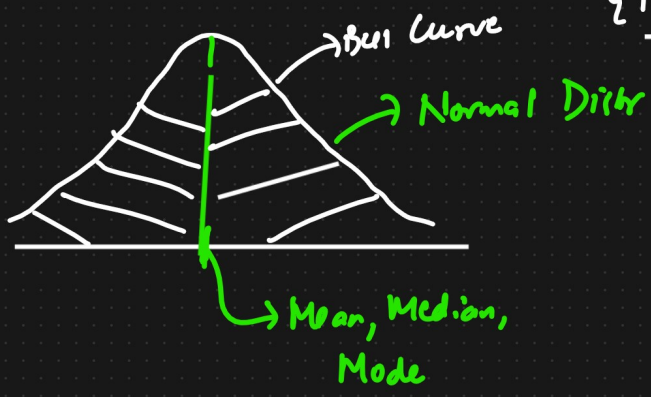
Practical
1. Mean, Median, Mode
2. Variance, Standard deviation
3. Histogram, pdf, Bar plot, Violin plot
4. IQR
5. Log Normal Distribution

① Distributions

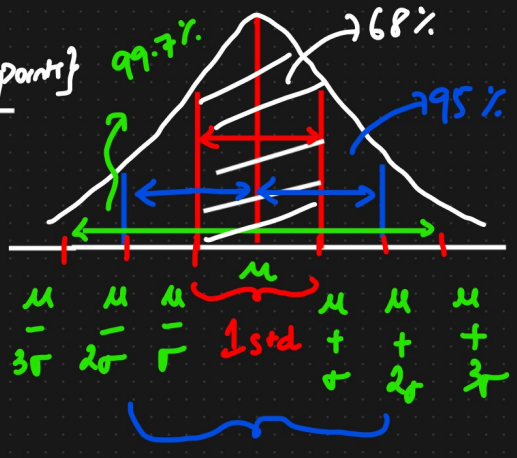
Ages = { 24, 26, 27, 28, 30, 32 - - - - - - - - - }



① Gaussian/Normal Distribution



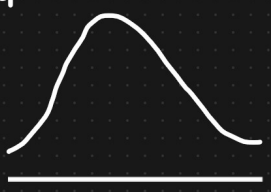
Dataset
{ 100 datapoints }



Empirical Formula

68 - 95 - 99.7 % Rule

Eg: ① Height → Normally Distributed

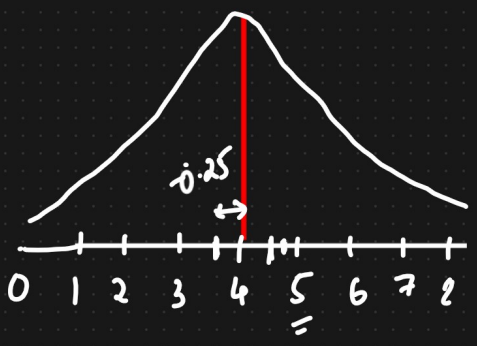


↓

Domain Expert → {Doctor}

② Weight ③ IRIS DATASET

Eg: $\mu = 4 \quad \sigma = 1$



4.5 → Standard deviation

+0.5 sd

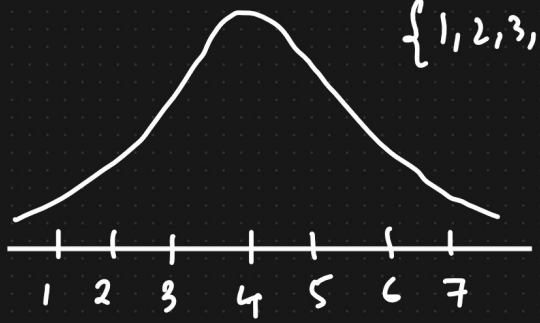
4.75 ??

$$Z_{score} = \frac{x_i - \mu}{\sigma}$$

$$= \frac{4.75 - 4}{1} = 0.75 \text{ sd}$$ (+ve)

3.75

$$Z_{score} = \frac{3.75 - 4}{1} = -0.25$$



$\mu = 4$, $\sigma = 1$

{1, 2, 3, 4.5, 6, 7}

$$\Leftarrow Z_{score} = \frac{x_i - \mu}{\sigma} \quad \mu = 4, \ \sigma = 1$$

{ -3, -2, -1, 0, 1, 2, 3 }

$$Z(1) = \frac{1-4}{1} = -3 \qquad Z(3) = \frac{3-4}{1} = -1$$

$$Z(2) = \frac{2-4}{1} = -2$$

$y \sim SND(\mu = 0, \sigma = 1)$

{ 1, 2, 3, 4, 5, 6, 7 } → Normal Distr

↓

Z score

↓

Standard Normal Distr ($\mu = 0, \ \sigma = 1$)

{-3,-2,-1, 0, 1, 2, 3} ← Satisfying this property

Practical Application

DATASET

Age (years)	Salary (RS)	Weight (Kg)
24	40K	70
25	80K	80
26	60K	55
27	70K	45

Units

Zscore ← Age

$\mu = 0, \sigma = 1$ ⇓ Standard Normal Distr

Standardization ← Zscore

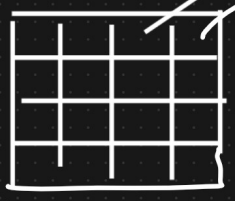
Normalization $\{\mu = 0, \sigma = 1\}$

(-1 to 1) ↳ (0 to 1)

→ MinMax Scaler → (0 to 1)

CNN → Image classification

Pixels



Divide by 255

0-255 ⇓ MinMax Scaler

Normalization ↗ 0-1

Practical Eg {India vs SA}

① ODI series ↓ 2021 (CRICKET)

Scores Average 2021 = 250

Standard Deviation = 10

Compared to both the scores in which year Rishabh pant final

Series < Team final score = 240

Score was better??

2020

Series Average 2020 = 260

Standard Deviation = 12

Team final score = 245

2021

$$Z\ score = \frac{x_i - \mu}{\sigma} = \frac{240-250}{10} = \frac{-10}{10} = -1$$

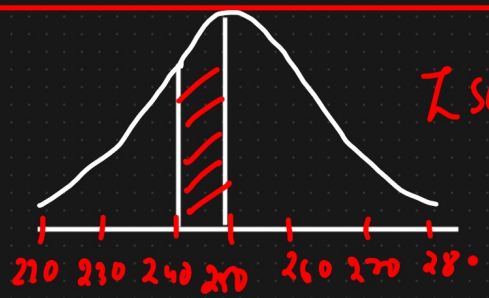
2020

$$Z\ score = \frac{x_i - \mu}{\sigma} = \frac{245-260}{12} = \frac{-15}{12} = -1.25$$

Final Match

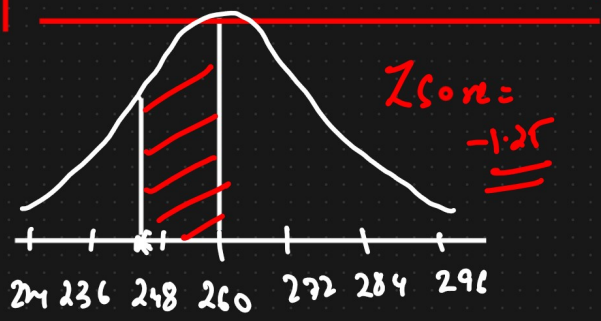
In 2021 ✓

$\mu = 250 \quad x_i = 240 \quad \sigma = 10$



In 2020 ✓

$\mu = 260 \quad x_i = 245 \quad \sigma = 12$



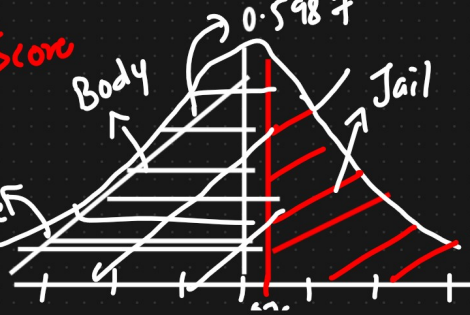
Assignment

Stats Interview Question 3/7

Question: What percentage of scores falls above 4.25?

X → Score

Area of Body curve



$\mu = 4$

$\sigma = 1$

1 - Left Area

{1 2 3 4 0.25 6 7} $1 - 0.5987$

$$Z = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4.0}{1} = 0.25 = 0.4013$$

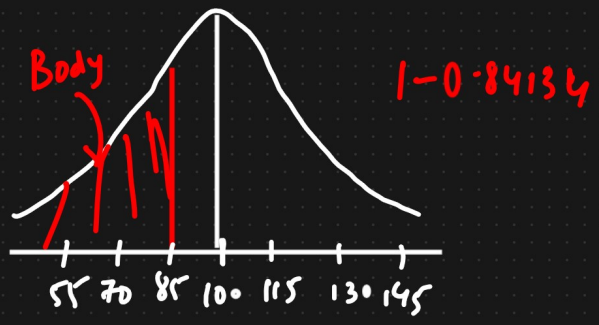
$\Downarrow$ 40%.

④ In India the average IQ is 100, with a standard deviation of 15. What percentage of the population would you expect to have an IQ lower than 85?

(Ans) $Z = \frac{85 - 100}{15} = \frac{-15}{15} = -1$

and

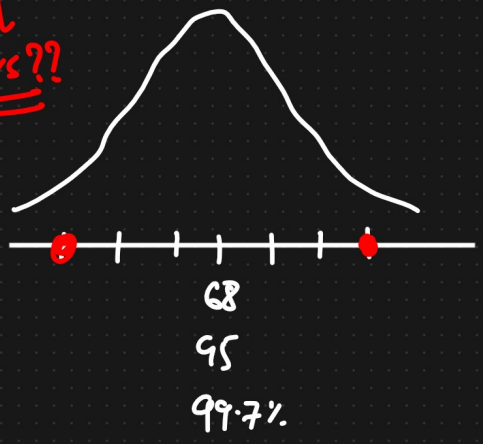
IQ 90 to 120



DAY 4 — STATS

1. IQR - python ✓
2. Probability ✓
3. Permutation And Combination ✓
4. Confidence Intervals ✓
5. P value ✓
6. Hypothesis Testing ✓

After 3rd sd outliers??



$$Z_{score} = \frac{x_i - \mu}{\sigma}$$ ✓

4. **Probability** : Probability is a measure of the likelihood of an Event

Eg : Roll a dice {1, 2, 3, 4, 5, 6}

$$Pr(6) = \frac{\#\text{ of way an event can occur}}{\#\text{ of possible outcome}}$$

$$= \frac{1}{6}$$

Toss a coin {H, T}

$$Pr(H) = \frac{1}{2}$$

② Addition Rule (Probability, "or")

Mutual Exclusive Event

Two Events Are mutual exclusive if they cannot occur at the same time

Eg: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$\{1, 2\}$

Tossing a coin $\{H$

$T\}$

Non Mutual Exclusive

Multiple events can occur at the same time

Eg: Deck of cards $\{Q, ♡\}$

① If I Toss a coin, what is the Probability of the coin landing on heads or tails?

Ans) Mutual Exclusive Addition Rule

$$Pr(A \text{ or } B) = P(A) + P(B)$$

$$= \frac{1}{2} + \frac{1}{2}$$

$$\boxed{Pr(A \text{ or } B) = 1}$$

Roll a Dice

$$Pr(1 \text{ or } 3 \text{ or } 6) = Pr(1) + Pr(3) + Pr(6)$$

$$= 1/6 + 1/6 + 1/6$$

$$= 3/6 = 1/2 = 0.5 //.$$

$\frac{52}{52}$ $\frac{2}{}$

Non Mutual Exclusive

You are picking a card randomly from a deck. What is the probability of choosing a card that is queen or a heart? → (52)

Ans) Non mutual Exclusive

$$P(Q) = \frac{4}{52} \qquad P(\heartsuit) = \frac{13}{52} \qquad Pr(Q \text{ and } \heartsuit) = \frac{1}{52}$$

Addition Rule for non mutual exclusive Events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(Q \text{ or } \heartsuit) = P(Q) + P(\heartsuit) - P(Q \text{ and } \heartsuit)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$= \frac{16}{52} \text{ a} \quad //$$

③ Multiplication Rule

{ Independent Events }

Eg: Rolling a dice {1, 2, 3, 4, 5, 6}

1, 1, 2, Each & Every are independent → Red marble ●

Dependent Events → $Pr(Red) = \frac{3}{5}$ → $Pr(Green) = \frac{2}{4}$

Naive Bayes {Conditional Probability}

Independent Events

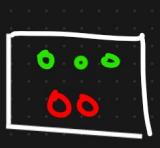
④ What is the probability of rolling a "5" and then a "4" in a dice?

Ans) Independent Event

Multiplication Rule

$Pr(A \text{ and } B) = P(A) * P(B)$

$Pr(5 \text{ and } 4) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$ //.



3/5

$\frac{2}{4}$

⑤ What is the probability of drawing a Queen and then a Aces from a deck of cards?

$P(Q \text{ and } R) = P(Q) * P(R/Q)$ → Event

Ans) Dependent

$P(A \text{ and } B) = P(A) * P(B/A)$ → conditional probability

↓

Bayes theorem

$P(Q \text{ and } A) = P(Q) * P(A/Q)$

$= \frac{4}{52} * 4/51$

⑥ Permutation and Combination

Permutation

School trip {choclate factory} → Dairy, 5 star, Milky bar, Eclairs, Gem, Silk

Student {Assignment}

Student → $\underline{6} \times \underline{5} \times \underline{4} = \underline{120}$

⑤

Dairy, Gem, Milky
Milky, Gem, Dairy

$n = 6$
$r = 3$

Permutation

$$^{n}P_{r} = \frac{n!}{(n-r)!} = \frac{6!}{(6-3)!}$$

$$= \frac{6 \times 5 \times 4 \times 3!}{3!}$$

$$= 120$$

⑥ Combination

Dairy	Gem	Eclair
—	—	—

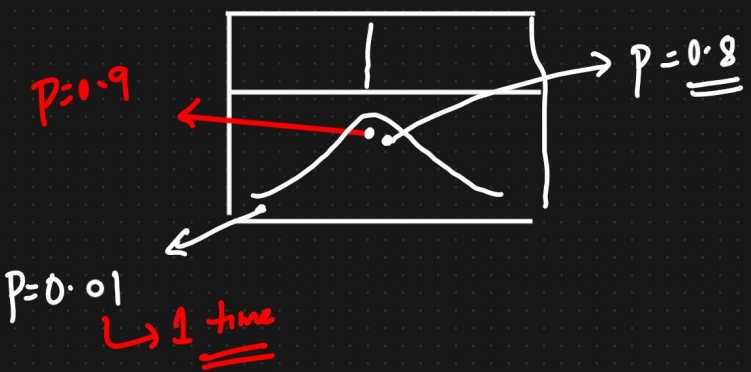
$$^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{6!}{3!(6-3)!}$$

$$= \frac{6 \times 5 \times 4 \times 3!}{3 \times 2 \times 1 \times 3!}$$

$$= 20$$

① p value { Many people gets confused }

Every 100 time I touch the mouse pad 80 times I touch this specific region



Hypothesis Testing, Confidence Interval, Significance Value, - - - -

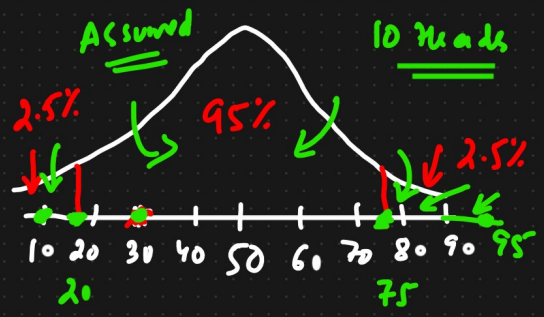
Coin → Test whether this coin is a fair coin or not by performing 100 tosses

Shady coin P(H) = 100%

P(H) = 0.5 P(T) = 0.5

100 → 50 times Head (The coin is fair)

30 Times Head 95 heads



Hypothesis Testing

100% − 5%

⇓

95%

CI

① Null Hypothesis : Coin is fair

✓② Alternate Hypothesis :- Coin is unfair

③ Experiment

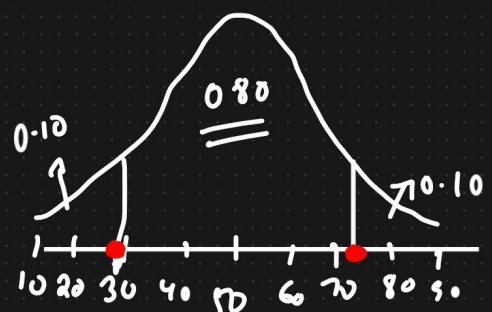
④ Reject or Accept the Null Hypothesis

Significance Value

α = 0.05 { Domain Expert }

α = 0.20 α = 0.3 C.I = 70%

80%



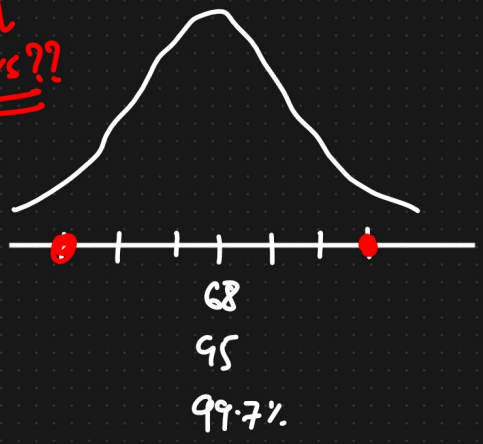
1

25

DAY 4 — STATS

1. IQR - python ✓
2. Probability ✓
3. Permutation And Combination ✓
4. Confidence Intervals ✓
5. P value ✓
6. Hypothesis Testing ✓

After 3rd sd outliers??



$$Z_{score} = \frac{x_i - \mu}{\sigma}$$ ✓

4. **Probability** : Probability is a measure of the likelihood of an Event

Eg : Roll a dice {1, 2, 3, 4, 5, 6}

$$Pr(6) = \frac{\#\text{ of way an event can occur}}{\#\text{ of possible outcome}}$$

$$= \frac{1}{6}$$

Toss a coin {H, T}

$$Pr(H) = \frac{1}{2}$$

② Addition Rule (Probability, "or")

Mutual Exclusive Event

Two Events Are mutual exclusive if they cannot occur at the same time

Eg: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$\{1, 2\}$

Tossing a coin $\{H$

$T\}$

Non Mutual Exclusive

Multiple events can occur at the same time

Eg: Deck of cards $\{Q, ♡\}$

① If I Toss a coin, what is the Probability of the coin landing on heads or tails?

Ans) Mutual Exclusive Addition Rule

$$Pr(A \text{ or } B) = P(A) + P(B)$$

$$= \frac{1}{2} + \frac{1}{2}$$

$$\boxed{Pr(A \text{ or } B) = 1}$$

Roll a Dice

$$Pr(1 \text{ or } 3 \text{ or } 6) = Pr(1) + Pr(3) + Pr(6)$$

$$= 1/6 + 1/6 + 1/6$$

$$= 3/6 = 1/2 = 0.5 //.$$

$\frac{52}{52}$ $\frac{2}{}$

Non Mutual Exclusive

You are picking a card randomly from a deck. What is the probability of choosing a card that is queen or a heart? → (52)

Ans) Non mutual Exclusive

$$P(Q) = \frac{4}{52} \qquad P(\heartsuit) = \frac{13}{52} \qquad Pr(Q \text{ and } \heartsuit) = \frac{1}{52}$$

Addition Rule for non mutual exclusive Events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(Q \text{ or } \heartsuit) = P(Q) + P(\heartsuit) - P(Q \text{ and } \heartsuit)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

$$= \frac{16}{52} \text{ a} \quad //$$

③ Multiplication Rule

{ Independent Events }

Eg: Rolling a dice {1, 2, 3, 4, 5, 6}

1, 1, 2, Each & Every are independent → Red marble ●

Dependent Events → $Pr(Red) = \frac{3}{5}$ → $Pr(Green) = \frac{2}{4}$

Naive Bayes {Conditional Probability}

Independent Events

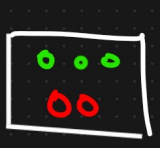
④ What is the probability of rolling a "5" and then a "4" in a dice?

Ans) Independent Event

Multiplication Rule

$Pr(A \text{ and } B) = P(A) * P(B)$

$Pr(5 \text{ and } 4) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$ //.



3/5

$\frac{2}{4}$

⑤ What is the probability of drawing a Queen and then a Aces from a deck of cards?

$P(Q \text{ and } R) = P(Q) * P(R/Q)$ → Event

Ans) Dependent

$P(A \text{ and } B) = P(A) * P(B/A)$ → conditional probability

↓

Bayes theorem

$P(Q \text{ and } A) = P(Q) * P(A/Q)$

$= \frac{4}{52} * 4/51$

⑥ Permutation and Combination

Permutation

School trip {choclate factory} → Dairy, 5 star, Milky bar, Eclairs, Gem, Silk

Student {Assignment}

Student → $\underline{6} \times \underline{5} \times \underline{4} = \underline{120}$

⑤

Dairy, Gem, Milky
Milky, Gem, Dairy

$n = 6$
$r = 3$

Permutation

$$^{n}P_{r} = \frac{n!}{(n-r)!} = \frac{6!}{(6-3)!}$$

$$= \frac{6 \times 5 \times 4 \times 3!}{3!}$$

$$= 120$$

⑥ Combination

Dairy	Gem	Eclair
—	—	—

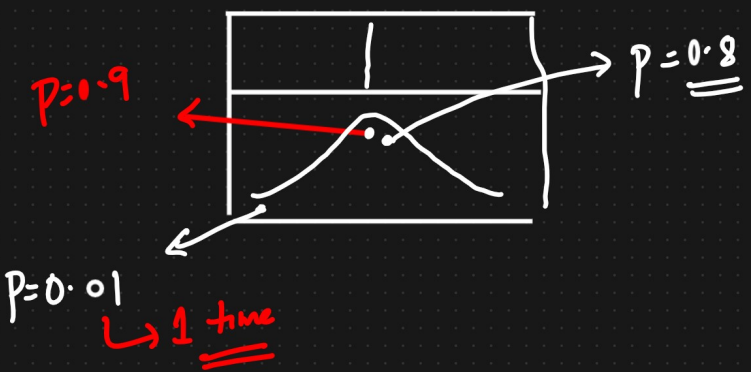
$$^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{6!}{3!(6-3)!}$$

$$= \frac{6 \times 5 \times 4 \times 3!}{3 \times 2 \times 1 \times 3!}$$

$$= 20$$

① p value { Many people gets confused }

Every 100 time I touch the mouse pad 80 times I touch this specific region



Hypothesis Testing, Confidence Interval, Significance Value, - - - -

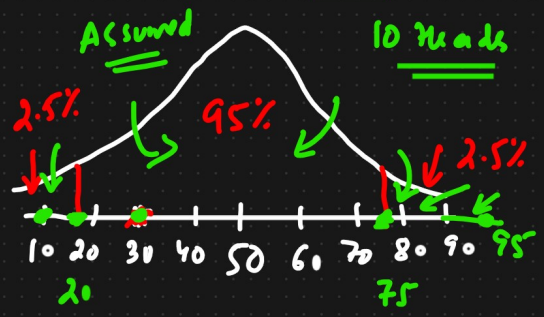
Coin → Test whether this coin is a fair coin or not by performing 100 tosses

Shady coin P(H) = 100%

P(H) = 0.5 P(T) = 0.5

100 → 50 times Head (The coin is fair)

30 Times Head 95 heads



Hypothesis Testing

100% − 5%

⇓

95%

CI

① Null Hypothesis : Coin is fair

✓② Alternate Hypothesis :- Coin is unfair

③ Experiment

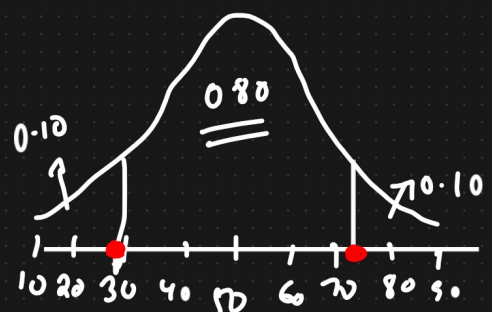
④ Reject or Accept the Null Hypothesis

Significance Value

α = 0.05 { Domain Expert }

α = 0.20 α = 0.3 C.I = 70%

80%



1

25

Today Topics

1. Type 1 and Type 2 Error ✓
2. One Tailed and 2 Tailed Test ✓
3. Confidence Interval ✓
4. Z-test, t-test, Chi-Square Test

1. Type 1 and Type 2 Error

Null Hypothesis (H_0) = Coin is fair

Alternate Hypothesis (H_1) = Coin is not fair

Reality check

Null Hypothesis is True or Null Hypothesis is False

Decision

Null Hypothesis is True or Null Hypothesis is False

Outcome 1 :
We reject the Null Hypothesis, when in reality it is false → Yes

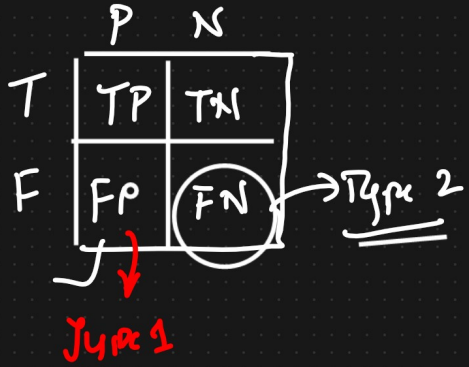
Movie

Person - Death Sentence

Outcome 2 : We reject the Null Hypothesis, When in reality it is true → Type 1 Error

Outcome 3 : We Accept/retain the Null Hypothesis when in reality it is false → Type 2 Error

Outcome 4 : We Accept the Null Hypothesis when in reality it is true → Good

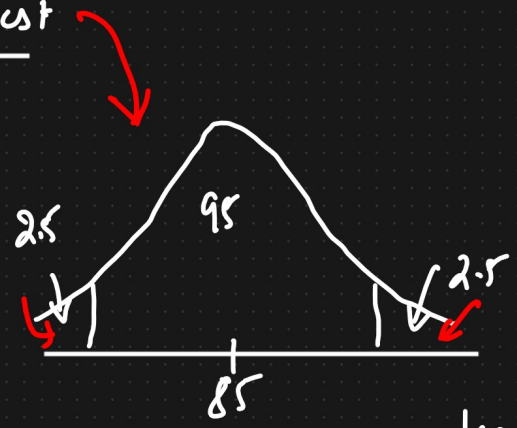


② 1 Tail and 2 tail Test

Eg: Colleges in Karnataka have an 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%. With a standard deviation 4%. Does this college has a different placement rate ? $\alpha = 0.05$

↳ 85%.

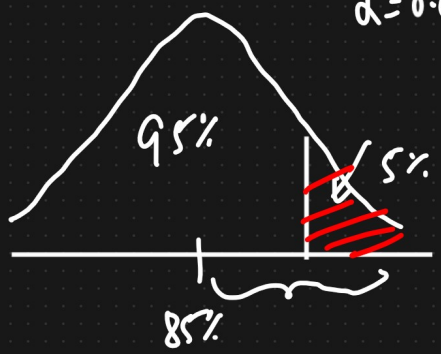
2 tailed Test



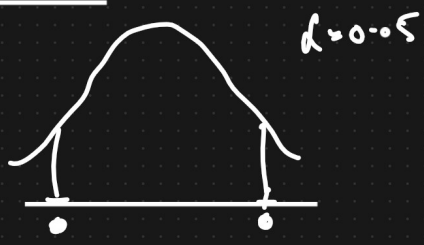
lenes

Does this college have a placement rate greater than 85% ?

$\alpha = 0.05$



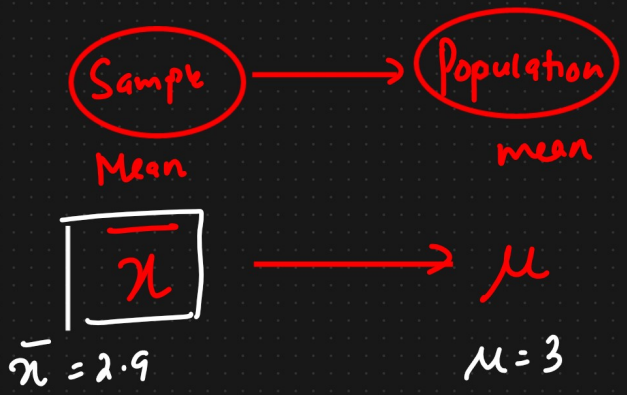
③ Confidence Intervals



Point Estimate

The value of any statistic that estimates the value of a parameter

Influential Stats



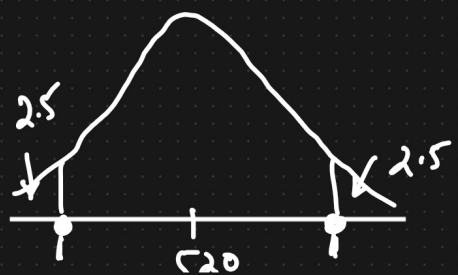
Confidence Intervals

Point Estimate ± Margin of Error

Q) On the quant test of CAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520 score. Construct a 95% CI about the mean?

Ans) $\sigma = 100$ $n = 25$ $\alpha = 0.05$ $\bar{x} = 520$

$\alpha = 1 - 0.95 = 0.05$



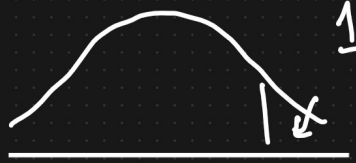
{ ① Population std is given } → Z test
{ ② $n \geq 30$ }

Point Estimate ± Margin of Error

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$ → Standard Error

$$Z_{\frac{0.05}{2}} = Z_{0.025}$$

$$1 - 0.025 = 0.975$$



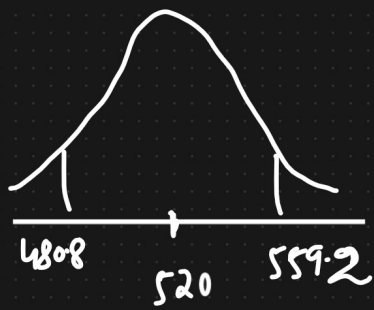
1.96

Upper bound $= \bar{x} + Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

Lower bound $= \bar{x} - Z_{\frac{0.05}{2}} \frac{100}{\sqrt{25}}$

Upper $= 520 + 1.96\,(20) = 559.2$

Lower $= 520 - 1.96\,(20) = 480.8$



Stats

Find the average size of the sharks throughout the world?

σ, $\bar{x}$, n, $\alpha = 0.05$

Q) On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a standard deviation of 80. Construct 95% confidence interval about the mean?

Ans) Condition $\quad n = 25 \quad \bar{x} = 520 \quad S = 80 \quad \alpha = 0.05$

Here population sd is not given → t-test

Point Estimate ± Margin of Error

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \rightarrow \text{Standard Error}$$

$$t_{\frac{0.05}{2}} = 2.064$$

$$\text{Upper bound} = \bar{x} + t_{0.05/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$\text{Degree of freedom} = n - 1 = 25 - 1 = 24$$

$$= 520 + 2.064\left(\frac{80}{5}\right)$$

$$= 553.024$$

$$\text{Lower bound} = \bar{x} - t_{0.05/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$= 520 - 2.064\left(\frac{80}{5}\right)$$

$$= 486.97$$

$$[486.97 \longleftrightarrow 553.024]$$

① One Sample Z-Test

① Population sd is given

② Sample size $n \geq 30$

(*) In the population, the average IQ is 100 with a sd of 15. Researchers wants to test a new medication to see if there is positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect the intelligence?

$\alpha = 0.05$ (CI = 95%)

→ 110 ✓

Ans) 1) Define Null Hypothesis

$H_0 = \mu = 100$

② Alternate Hypothesis H_1 $\mu \neq 100$

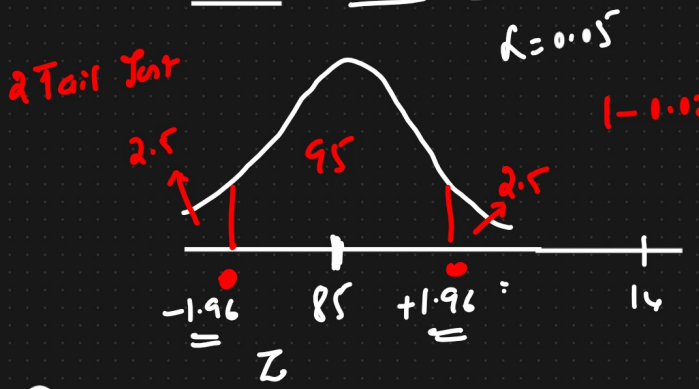
③ State Alpha

$\alpha = 0.05$

④ State Decision Rule

Z table

$1 - 0.025 = 0.975$



$P \leq 0.05$

⑤ Calculate Test Statistics

Sample data

$$Z = \frac{\bar{x} - \mu}{\left\{\frac{\sigma}{\sqrt{n}}\right\}} \rightarrow \text{Standard Error}$$

$$= \frac{140 - 100}{\frac{15}{\sqrt{30}}} = \frac{40}{15} \times \sqrt{30} = 14.60$$

State our Decision

$\{Z = -0.2\}$

$Z = 14.60$

$14.60 > 1.96$

If Z is less than -1.96 or greater than 1.96, reject the null hypothesis

Medication improve the intelligence or decrease ?!

Improve the intelligence

② One Sample t-test

Z-test ⇒ population std

t-test ⇒ unknown population std

EQ) Population the average IQ = 100

$n = 30 \quad \bar{x} = 140 \quad s = 20$

Did the medication affect intelligence?

$\alpha = 0.05$

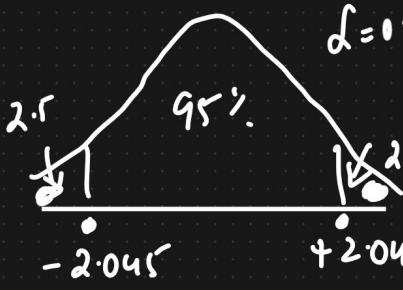
Ans) ① $H_0 = \mu = 100$

② $H_1 \quad \mu \neq 100$

③ Calculate the degree of freedom

$n - 1 = 30 - 1 = 29$

④ State Decision Rule



$\alpha = 0.05 \quad t = 10.96 > 2.045$

Reject Null Hypothesis

$P \leq$ significance value

⇓

{ Increased the intelligence. }

⑤ T Test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\bar{x} = 140$

$\mu = 100$

$s = 20$

$n = 30$

$= 10.96$

Reject the Null Hypothesis

Real World Problem

6th DAY Live SESSION

1. CHI SQUARE ✓
2. Covariance ✓
3. Pearson Correlation Coefficient ✓
4. Spearman Rank Correlation ✓
5. Practical Implementation

   Z-test, t-test, chi square test
6. F Test (ANOVA)

CHI SQUARE TEST

1. Chi Square Test claims about population proportions

   It is a non parametric test that is performed on categorical (nominal or ordinal) data.

Q) In the 2000 Indian Census, the age of the individual in a small town were found to be the following:

less than 18	18-35	>35
20%	30%	50%

In 2010, age of n=500 individuals were sampled. Below are the results

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the population distribution of ages has changed in the last 10 years?

Ans)

<18	18-35	>35
20%	30%	50%

{Population} 2000

Expected

<18	18-35	>35
121	288	91
100	500 × 0.3	500 × 0.5
	150	250

n=500

Observed

Expected

<18	18-35	>35
121	288	91
100	150	250

Observation

Expected

{Chi Square table}

① H_0 = The data meets the distribution 2000 census

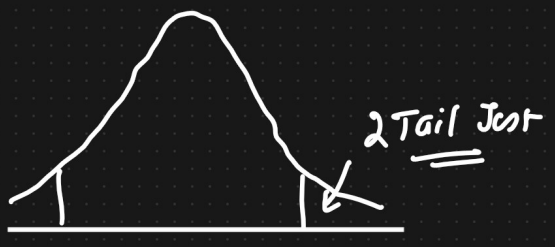
H_1 = The data does not meet " " "

② $\alpha = 0.05$ (95% C.I)

df: 2, $\alpha = 0.05$

③ Degree of freedom = $n - 1 = 3 - 1 = 2$

④ Decision Boundary



If χ^2 is greater than 5.99 reject H_0

5) Calculate Test Statistics

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$= 232.94$$

$\chi^2 = 232.494 > 5.99$ {Reject the Null Hypothesis}

$\alpha = 0.05$ ↙ 0.01, 0.10 } Domain.

0.11 > 0.05 ⇓ Accept the Null

0.11 ✗ 0.0 ⇓ Reject the Hypothesis

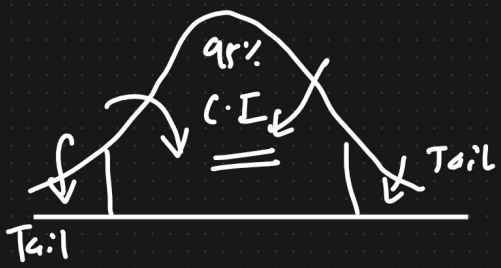
0.002 < 0.05 {Reject the null hypothesis}

I Was Correct Here

{ p-value < Significance value ⇓ Reject the Null Hypothesis. }

OR

{Accept the Null Hypothesis}



p 0.11 > 0.05 ⇓ Accept

p = 0.002 < 0.05 ⇓ Reject the null hypothesis

② Covariance

X	Y
Weight	Height
50	160
60	170
70	180
75	181

{ X↑ Y↑ ; X↓ Y↓ }

No. of hour study	play
2	6
3	4
4	3

X↑ Y↓ ; X↓ Y↑

Quantity relationship between X & Y

Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

+ve ⇒ positive Correlation

X↑ Y↑ ; X↓ Y↓

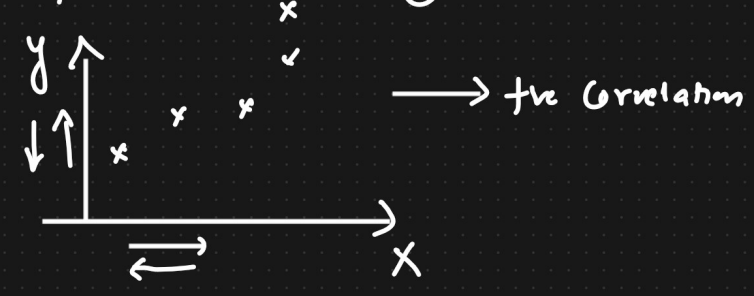
-ve ⇒ negative Correlation

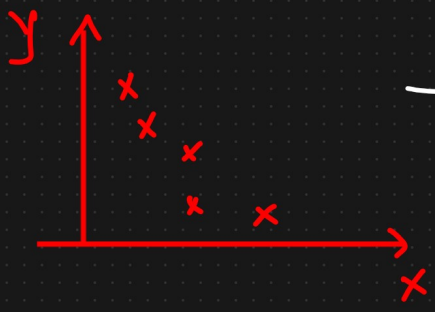
X↓ Y↑ ; X↑ Y↓

= +ve or -ve

X ⇕ & Y

O

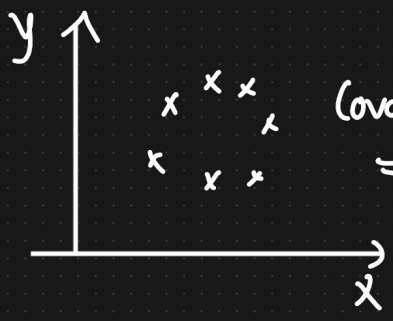




→ -ve Correlation

⇓

-ve Covariance



Covariance = 0

Disadvantage of Covariance

① Positive or Negative ✓

+100

-200

-2000

+1000 {Direction}

② Pearson Correlation Coefficient

(-1 to 1)

The more towards +1 more positively correlation

The more towards -1 more negative correlated

$$\rho_{(x,y)} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \quad \{-1 \text{ to } 1\}$$

$$Spear(x,y) = \frac{Cov(R(x), R(y))}{R_{\sigma_x} \times R_{\sigma_y}}$$

X Height	Y Weight	R(x)	R(y)
170	75	2	2
160	62	3	3
150	60	4	4
145	55	5	5
180	85	1	1

Non linear properties

Coding

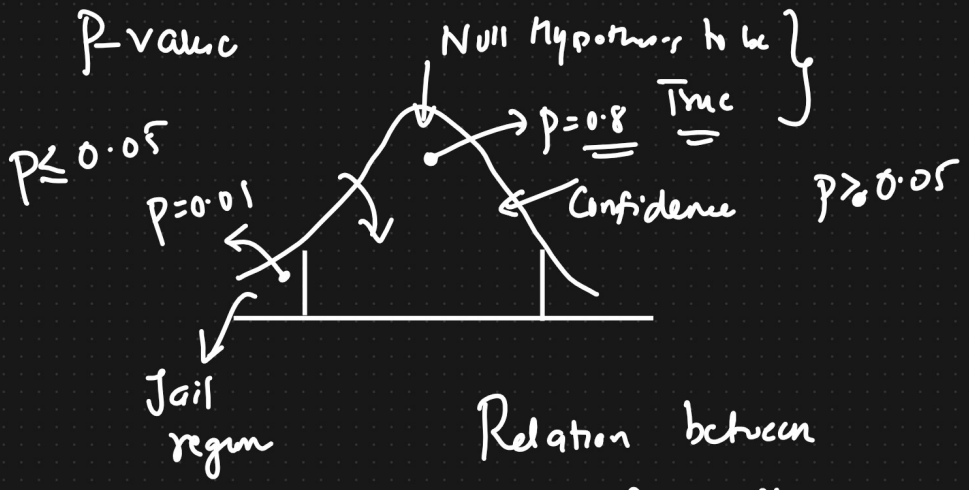
$P \leq 0.05 \rightarrow$ Reject the NULL Hypothesis

probability

5% probability the null Hypothesis is Correct

$\alpha = 0.05$

$P \geq 0.05 \rightarrow$ Accept the NULL Hypothesis



Relation between pvalue & significance

P-value ≤ Significance [C.I]

↳ Reject the Null Hypothesis

P≥α ↳ Accept the Null

① P value and Significance value

7 day

① P value & Significance value

② Distribution

③ Central Limit Theorem

④ Bernoulli's Distr Log Normal Distr

⑤ Binomial Distr

⑥ Pareto's Distr {Power Law}

⑦ F Test (ANOVA) → 1 hr → upload a seperate video.

{ 7080 ⇓ KRISHIO → 10% }

① P value & Significance

↳ Derive the p value

Q) The average weight of all residents in Bangalore city is 168 pounds. With a standard deviation 3.9. We take a sample of 36 individuals and the mean is 169.5 pounds. C.I : 95%

Ans) $\mu = 168$ $\sigma = 3.9$ $\bar{X} = 169.5$ $n = 36$ $\alpha = 0.05$

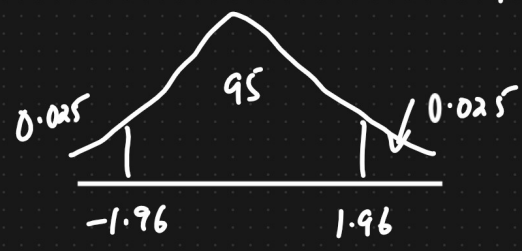
① $H_0 = \mu = 168$

$H_1 : \mu \neq 168$

② $\alpha = 0.05$

③ Decision boundary

$1 - 0.025 = 0.9750$ ⇓



(5) Z Test

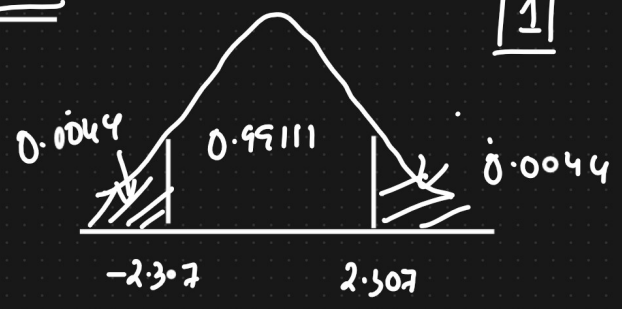
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{169.5 - 168}{\frac{3.9}{\sqrt{36}}}$$

$$= \frac{1.5}{3.9} \times 6$$

$$= 2.307$$

(6) $Z = 2.307 > 1.96$ Reject the Null Hypothesis

P value



$1 - 0.9828$

$1 - 0.99111 =$

P value $= 0.0044 + 0.0044$

$= 0.0088$

P value < 0.05

$0.0088 < 0.05$ → Reject the Null Hypothesis

② P value $\leq$ Significance value

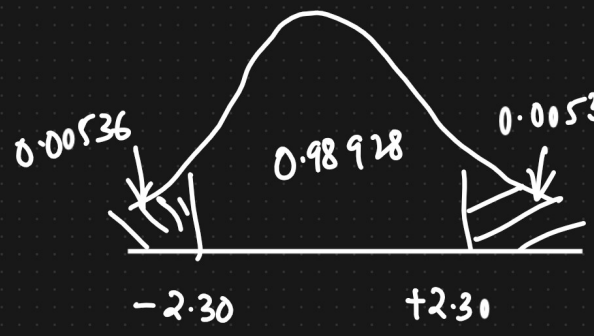
⇓

Reject the Null hypothesis

P value $>$ Significan value

⇓

Fail to Reject the Null Hypothesis



$Z = 2.30$

$1 - 0.98928 =$

$2.30 > 1.96$ {Reject the Null Hypothesis}

P value $= 0.00536 + 0.00536$

$= \_\_\_\_\_ \leq \alpha$ ⇒ Reject the Hypothesis

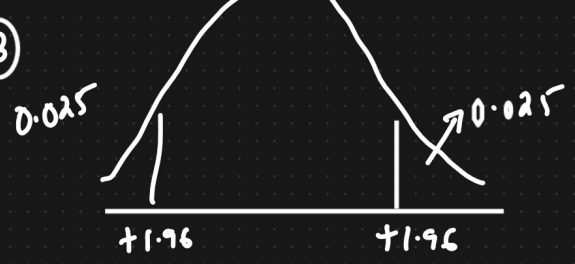
② Average age of a College is 24 years with a standard deviation 1.5. Sample of 36 student students mean is 25 years. With $\alpha = 0.05$ C.I : 95%, do the age vary?

Ans) ① H_0 $\mu = 24$ $\sigma = 1.5$ $n = 36$ $\bar{x} = 25$ $\alpha = 0.05$

H_1 $\mu \neq 24$

② $\alpha = 0.05$

③

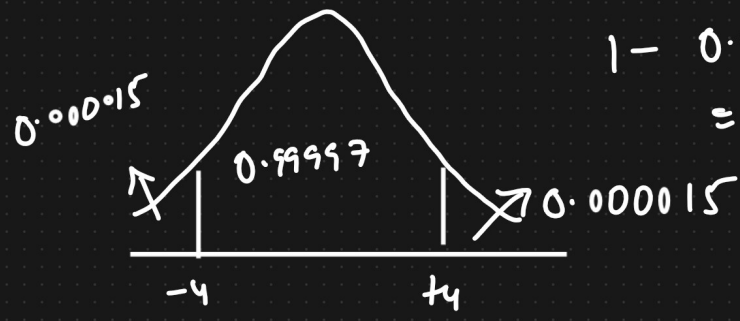


④ $Z\text{-score} = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{25 - 24}{1.5} \times 6$

$= \dfrac{1 \times 6}{1.5}$

$= 4$

$4 > 1.96$ Reject Null Hypothesis



$1 - 0.99997$

$=$

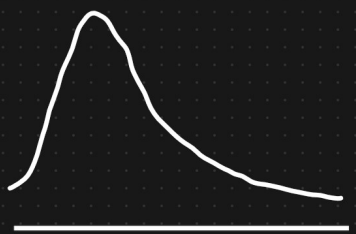
0.000015

$p\text{value} = 0.000015 + 0.000015$

$= 0.00003$

pvalue $\leq$ Significance {Reject the Null Hypothesis}

② Log Normal Distribution



Eg:- Wealth Distribution

② people writing big comments

{$y \sim$ LogNormal. Distr}

$\Downarrow$

$\log(y) \longrightarrow$ Normal Distribution

④ Bernoulli's Distribution

2 Outcomes | 0 or 1 | Single Trial

Tossing a coin

$P = 0.5$

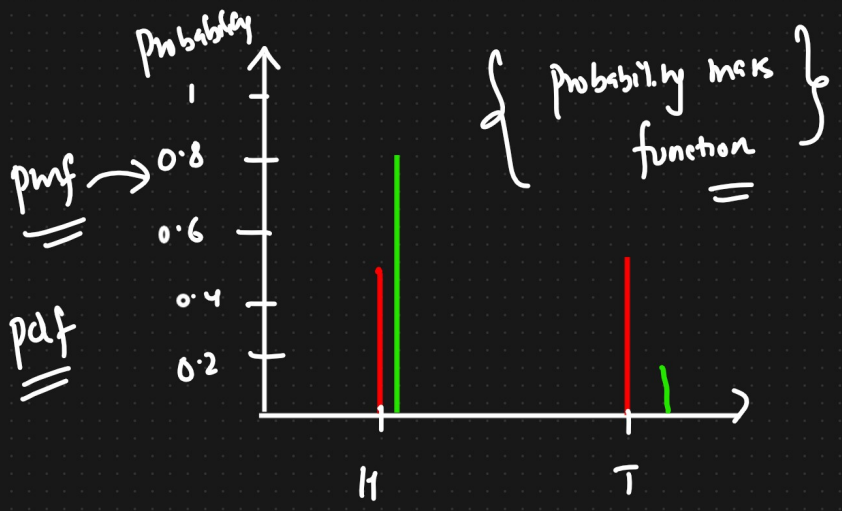
$q = 1 - P = 0.5$

$P(H) = 0.5 = P$

Do not have a fair coin

$P(H) = 0.3 = P$

$P(T) = q = 1 - p = 1 - 0.3 = 0.7$



⑤ Binomial distribution

Every → Bernoulli distribution

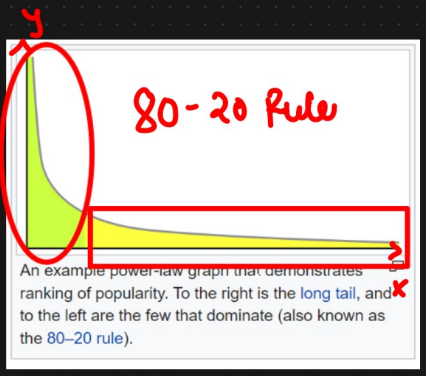
Multiple Trail

↓

$P(H) = 0.5$ $P(H) = 0.6$ - — - - - - .

$P(T) = 0.5$ $P(T) = 0.4$

Pareto distribution



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

① Eg: 80% of the wealth is distributed with 20% of the people

② 80% of the company project are done by 20% of the people in a team

(K) 80% of sales is done by 20% of famous product.

(K) 80% of the Cricket match is won by 20% of the team.

Power law distribution

Log Normal Distribution

Lognormal Distribution

→ Transformation

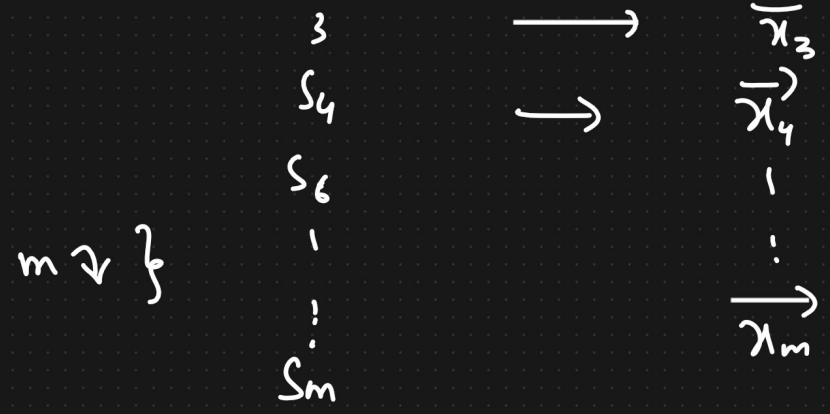
(K) Central Limit Theorem

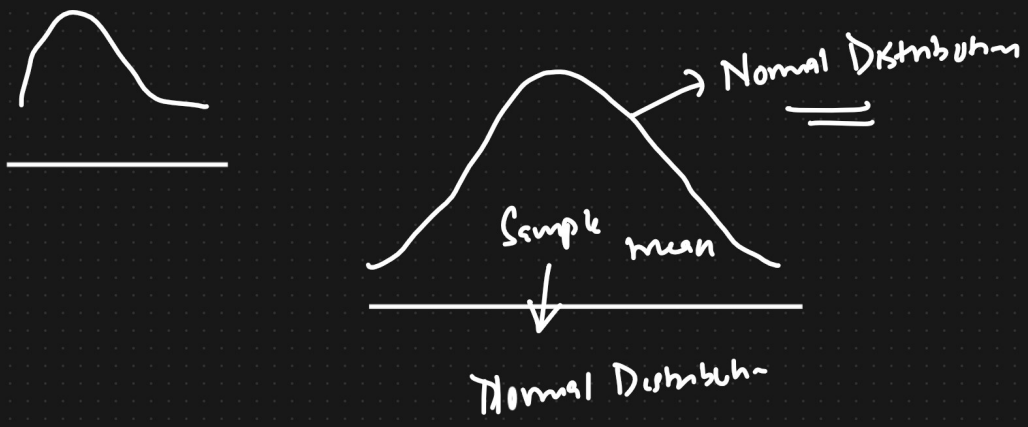
Sample size → $n \geq 30$

S_1 $\bar{x}_1$

S_2 → $\bar{x}_2$

} → Sample mean





Poisson Distribution

→ 2 Algorithms

① Machine Learning Algorithms ✓
② Deep Learning Algo ✓
③ FLASK & DJANGO ✓
④ Mongodb, Sql ✓
⑤ Blockchain Session ✓