another level of stats

Statistical methods

- Descriptive
  - Univariate
    - shape
    - center
    - Spread
  - Bivariate
    - correlation
    - regression
  - Multivariate
    - Heatmap
    - Multiple Linear Regr
- Inferential stats
  - applied to mean (Z-test, t-test
  - applied to Variance (F-test, ANOVA)

How many way to Sum... ?CT

Ratio → eg
- male female ratio mean 2:1
- class 2:1
- Temp → inside 30° out 60 → 1:2

→ Mukhiya ji Mean
→ In you in USA Calle indian (mean) all work around the Mukhiya

Z टेस्ट → mean = 3 — 1, 2, 3, 4, 5 — obj
mean = 3 — 3, 3, 3, 3, 3 — obj

→ Describe of data तो दोनों Both objs Same है? Now you stuck
→ So do you really think all info about data → No

that's why the another topics comes into picture

(2) Measure of Spread or disperstion →

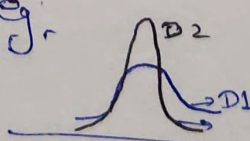
eg obj: O1 — 1 2 3 4 5
O2 — 3 Stines

So you understand why spread use bc2 alone CT not enough info of data

3 ← Then why go for Measure Symm

(3) Measure of Symmatricity:

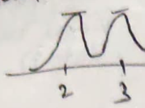
eg.



Spread of D1, D2 same for both but the shape is diffrance so thats why learn (Skewness kurtosis.) for high / shape

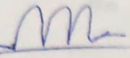
③

Model 1,2,2,3,3,4,5



→ called bi-model
↳ means (दो models) 2, & 3

② 2,2,3,3,4,4,5,6,7
2,3,4 called Multi Model



⇒ Arithmetic mean → $\sum_{i=1}^{n} \frac{x_i}{n}$

② **Geometric mean** (Ratio Same) → multiplies the no together and does a square root/cube root or of how many number are there.

eg:- 2 & 10 → GM

$\sqrt[2]{2\times10} = (2\times10)^{1/2} = \sqrt{20} =$

eg ① 2,2,5
3 no → $\sqrt[3]{2\times2\times5}$ or $(2\times2\times5)^{1/3}$

Usecase

eg If you go for buy a Camera
C1 has: Prob zoom 200, Review 8
" 250, Review 6
} → we can't use AM

eg C1 → $\frac{200+8}{2} = 104$ ① क्योंकि दो अलग-अलग Parameter हैं + दोनों में Large diffrence in Zoom & Review Vaue

C2 = $\frac{250+6}{2} = 128$ ✗ mean wrong नहीं है but meany ②

- **GM is used when we want to compare two different property**

eg C1 = $\sqrt[2]{200\times8} = 40$ → GM high so select
$\sqrt[2]{250\times6} = 38$

③ **Harmonic Means**:

$$H.M = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \dots \frac{1}{n}}$$

eg:- 2,4,5,100 ⇒ $\frac{4}{\frac{1}{2}+\frac{1}{4}+\frac{1}{5}+\frac{1}{100}} = \frac{4}{0.5+0.25+0.2+0.01} = 4$

H.M ↓ is the reciprocal of avg of all reciprocals

Means ↳ avg. of reci = $\frac{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}}{3}$ → recip recipt = $\frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}}$

Use in

① Confusion matrix ⇒ F. score (f-score is nothing but H.M of recall & Precision)

② we travelled 10k at 60km/h, then another 10km at 20km/hr. what is avg speed? ↳ may be asked in data Analy/BI

exam Cat/Aptitude ↳ $\frac{2}{\frac{1}{60}+\frac{1}{20}}$

④ **Frequency Means** (fx)

eg. Score frequency

Score	frequency
1	2
2	5
3	3
4	6
5	5
6	2

1 → 2 times

f	fx
1	1×2
2	2×5
3	3×3
4	4×8
5	5×5
6	6×2
Σ 29	Σ 150

⇒ $\frac{\Sigma fx}{\Sigma f}$

then what is mean of score (individual add time taken)

* Weighted Mean :-

हर Number का Weight importance deter करे

if calculate Mean = $\frac{1+2+3+4}{4}$ ⇒ what actually happen here → avg here but actual

So we can say ⇐ 1 has 1 frequ, 2 " 1, 3 " 1

All the values have equal weights → here equal wt so 1/4 equal all

↳ So we can write this in this form

$1\times\frac{1}{4} + 2\times\frac{1}{4} + 3\times\frac{1}{4} + 4\times\frac{1}{4}$ → equal weight

Now the Question is here

(Q) eg: 1, 2, 3, 4

weight ⇒ 0.1, 0.1, 0.7, 0.1

⇒ Calculate the mean How ⇒ $0.1\times1 + 2\times0.1 + 3\times0.7 + 4\times0.1$ = 2.8 means

Use Case / Application

- Weight mean can be used where Something are important and as compared to others.

eg: 0.1, 0.1, 0.7, 0.1 → what is Sum of weights → 1 if one then not need to divide we remove

→ what if W don't add to 1 ↳ divide the SUM by weights

eg(1) Camera example :- Preference (Real world example)

.. wants to Vanish buy → Image quality : 50%, Battery life : 30%, Zoom range : 20%.

Cam1 Sony → IQ 8 (Review), BL : 6, ZR : 7

Cam 2 ↳ IQ BL ZR: 9 4 6

do here W.Mean | Which camera should buy Refrence ↳ So what you do Mean or AM, GM, WM

Sony → $0.5\times8 + 0.3\times6 + 0.2\times7$ ⇒ 7.2 ✓ Prefere

Cam2 → $0.5\times9 + 0.3\times4 + 0.2\times6$ → 6.9

→ most quest Ask Mean because Inference में data needed होता है उस time handy होता है

eg(2) Dinesh had lunch → 7 times a week but some week he get only 1, 2 or 5 lunches. So

on 2 weeks → only 1 lunch for whole week

14 " – 2

8 " – 5

32 – 7

(Q). Ask → then what is the mean no of lunch that Dinesh have every week ??

Use Simple mean × not proper o/p

→ weight is more than 1 so → 2+14+8+32 → 56 ↳ so devide

$$\frac{2\times1 + 14\times2 + 5\times8 + 32\times7}{56}$$

⇒ $\frac{num}{56}$

recap → Summ ↳ descriptive

S1 – 1,2,3,4,5 ; 1,2,3, ; 3,3,3,3,3

MCT 3, 3 ; =M=3, 3

↓ MCT → not alone true

↳ Spread of data

(1) Variance, Range, Percentile / Quartile (IQR, Box Plot) ↳ Varia...

⇒ Dispersion

① Range → diffrence b/w max^m & min Value

↳ 1, 2, 3, 4, 5

range = 5 − 1 = 4 ⇒ It i knows from data so what new

Note:-
↳ outliers affect the range ← eg:- 1, 2, 3, 4, 1000

↳ range = 1000 − 1 ⇒ 999 } here calc is correct but ambigous Interpritions

② Percentage & Percentile :-

↳ It is a Value below which a certain % of observation lie.

eg: 1, 2, 3, 4, 5

odd % $\frac{3}{5} \times 100 = 60\%$.

→ df.describe() Shows percentiles

eg: data = 1, 2, 3, 4, 4, 6, 7, 8, 8, 10

(Q) What is the percentile rank of 3?

$$\text{ptile rank} = \frac{\text{No of Value below that number}}{\text{Total number}} \times 100$$

$$= \frac{2}{10} \times 100 = 20^{th} \text{ percentile}$$

↳ means 20% no. eq to or below 3 20% लोग उसके बराबर है या उससे नीचे

(Q) What Value exist at 75th percentile

$$\text{Value} = \frac{\text{percentile}}{100} \times (n+1) \quad \rightarrow \text{total no}$$

$$\Rightarrow \frac{75}{100} \times (10+1)$$

$$= \frac{75}{100} \times 11 = 8.25^{th} \text{ no.}$$

↳ So take avg of 8 & 9th no

8.5 → 8
8.7 → 9

The 8th no that is 7 is my 75th prctile

③ Quartile :- Quartiles are Values that divides a list of no into Quarters (4 equal part)

Solve → Page → Outliers

eg:- 6, 8, 5, 5, 7, 3, 9

↳ order the no :- 3, 5, 5, 6, 7, 8, 9

1st | 2nd | 3rd | 4th

Quartile1 Q_2 Q_3

$Q_1 = 5$
$Q_2 = 6$
$Q_3 = 8$

4 Quaters / 3 Quatiles

If bigger no. of datapts

eg:- Jan feb ... dec

eg:- 1, 1, 1, 1, 2, 2, 3, 3, 4

Q_1 Q_2 Q_3

what is Quartile ?

① order

② total count = 11

If count = odd

$\rightarrow Q_1 = \frac{n+1}{4}^{th} = \left(\frac{11+1}{4}\right)^{th} = 3^{rd}$ obs ⇒ 1

$\rightarrow Q_2 = \left(\frac{n+1}{2}\right)^{th} = \left(\frac{11+1}{2}\right)^{th} = \left(\frac{12}{2}\right)^{th} = 6^{th}$ obs ⇒ 2

$\rightarrow Q_3 = \frac{3(n+1)}{4}^{th} = \frac{3 \times 12}{4}^{th} = 3 \times 3 = 9^{th}$ ⇒

eg: 1, 2, 3, 3, 4, 4

$Q_1 = \frac{6}{4} = \frac{3}{2} = 1.5^{th} \rightarrow \frac{1+2}{2} = 1.5\ Q_1$

$Q_2 = 4.5^{th}$ → $\frac{4+4}{2} = 4$

If even

$\rightarrow Q_1 = \frac{n}{4}^{th}$ observation

$\rightarrow Q_2 = 3 \times \frac{n}{4}^{th}$

$\rightarrow Q_3 = \left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}$

→ Understand is import of Quatile:
↳ need

IN DS/DA

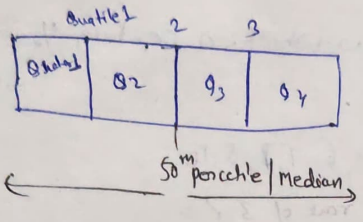
↳ df.describe() → FIVE Point Summary

Q_1 → min → 0th percentile
Q_2 → 25th → 25th "
Q_3 → 50th → 50 "
Q_4 → 75th → 75 "
max → 100 "percentile

Eg: Transcation amounts
500
1000
600
700
:
:

let say $Q_1 = 1000$
$Q_2 = 2000$ → 50% of the pop^n do the transcation of 2000 Rs or below 2000
$Q_3 = 3000$

⇒ Box Plot:



→ Now what is Outlier (extreme Value → much much higher, much much lower)

eg (-100) 1, 2, 3, 4, 5 (100)

So how I do to calculated the Range without effect the Outlier.

where is possibility to lie outlier in Quatile
ans → yes (1) & (4)th

↓ come here

Inter Quartile range = $Q_3 - Q_1$

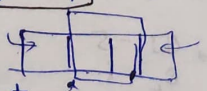
eg:- 2, 3, 3, (3) 3, 4, 4, 5, 5, 5, 6, (6) 6, 7, 8, 9, 9
Q_1 Q_3

$Q_1 = 25^{th}$ percentile $= \frac{25}{100} \times 16 = \frac{1}{4} \times 16 \Rightarrow 4^{th}$ obs $\Rightarrow Q_1 = 3$

$Q_3 = 75^{th}$ " $= \frac{75}{100} \times 16 = \frac{3}{4} \times 16^{th} = 12^{th}$ obs $= Q_3 = 6$

Points
→ Most data in $Q_1 - Q_3$ (50%)
ya

IQR = $Q_3 - Q_1$ = 6 - 3 = 3 Ans



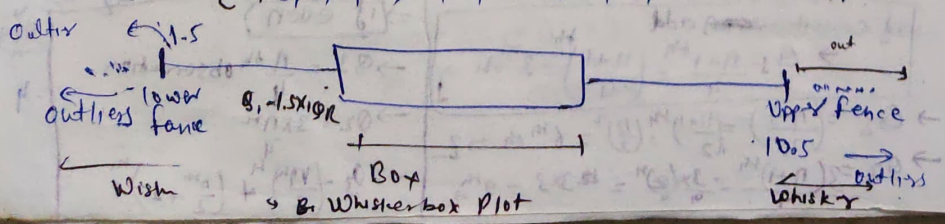
Useful to find Outlier
↳ In normal distribut $Q_1 - Q_3$ (68% lies)

How

↳ Lower fence = $Q_1 - 1.5 \times IQR$ ← अगर इससे Value से छोटा है तो
Upper fence = $Q_3 + 1.5 \times IQR$

why = $1.5 \times$?
any number
↳ but research conclude

eg:- {2, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 9}

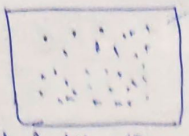


↳ Box Whisker box Plot

Topics
Mean deviation
Variance
Std dev
Measure of Symmetry
- Skewness
- Kurtosis
Covariance & Correlation
Probability

⇒ Measure of Spread → Variance & Std dev

eg:



30 students

In School Auditor/Parents come to School they don't know about class, and he wants to judgest the kusum. How she perform in class Compare to others.

- Initial → he give SMCQ → 30 Question

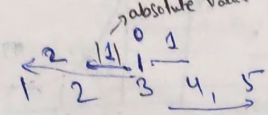
then calculate Avg Mark = 2.1

data 1, 2, 3, 4, 5 and take the no. of kusum if the 2.9 & good, if 1.5 & below

How away much data away from the mean?

Cal → mean = 3

→ away



distances from

$$\frac{1+2+1+2+0}{5} = \frac{6}{5} = 1.2 \rightarrow \text{each point Avg Away from Mean} = 1.2$$

Again → Reframe Question → on an Avg how much away each data point is away from mean Value ??

$$\sum_{i=1}^{n} \frac{|x_i - \mu|}{n}$$

→ mean deviation

Variance: The average of the squared differences from the mean.

$$\text{Var}_{pop} \ (\sigma^2) = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$ → pop mean

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

changes: $N - n$, $\mu - \bar{x}$, $N - n-1$ (denominators)

① Calculate Mean

→ for each no in data, substract the mean & no.

→ Square the differene

→ Cal the Avg squer of diff

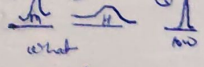
eg

① data = {1, 2, 3, 3, 4, 4}	② $\bar{x}$	③ $x - \bar{x}$	④ $(x - \bar{x})^2$
1	2.83	-1.83	3.34
2		-0.83	0.68
3		0.217	0.03
3		.17	0.03
4		1.17	1.37
4		2.17	1.37
$\sum \frac{x_i}{n} \ \frac{17}{3} \Rightarrow 2.83$			6.82

⑤ $\text{Var}_{sam} = \frac{6.82}{n-1} = \frac{6.82}{5} \Rightarrow 1.37$

What is actually it? find

What is spread data



→ Var ↑ spread ↑

→ Var ↓ spread ↓

* Standard deviation: Stand. de is a measure of how spread out numbers are.

↳ Square root of Variance

Std $= \sqrt{\text{Var}} = \sqrt{1.37} \Rightarrow 1.17$

SD of Pop $= \sigma = \sqrt{\text{Var}_p}$

SD of Sam $\Rightarrow S = \sqrt{\text{Var}_S}$

Number seven when compare to other

Use Case

Math is fun examp

$\sigma^2 = 21704$

$\sigma = \sqrt{21704}$

$= 147$

→ In Real life the Variance not use only show spread ⇓ Show understand ⇓ Useful SD compare so we compare the data

> Stand Dev is a Standard way of knowing how normal, large extra large is Something

→ mean = 394

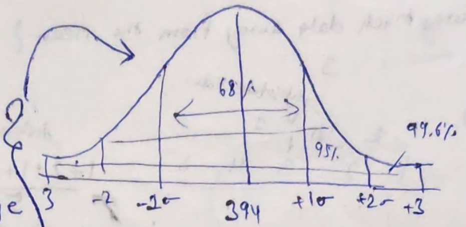
SD = 147

→ 1SD → 394 + 147 = 541 } normal
394 − 147 = 247

example

2SD ⇒ { 394 + 2×147 ⇒ Large
394 − 2×147 ⇒ Small

3SD → { 394 + 3×147 → extra large
394 − 3×147 → extra small



→ Why Square Varia

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

why sample ↓ because we cannot acces popln

Why (n−1)?? → Bessel correction / Unbiased estimation

We use n−1 rather than n is because Sample Variance will be unbiased estimators

→ How much bigger Sample Can be ??

pop → Sample →

pop → subset → Sample → 10 से ज्यादा नहीं ले सकते है और अगर 10 लेते है तो Sample become population

10−1 = 9 (n−1) इसलिए atleast 1 कम हो n−1 ✗

eg: N = 10
अगर Population size 10 है और 10 ही Sample में लेंगे तो Pop & Sample

but it is not Completely correct.

really eg:

Pop — $\bar{x}$, P-D, μ, N — Sample: $x-\bar{x}$, SD

→ यदि यह Pop है और Sample Variance Calculate करना है तो Sample 1st

→ pp Var $= \frac{x_i - \mu}{N}$ } → but you don't have a access pop)

Step → Pop Data → Pop Mean
↳ select → sample → then $\bar{x}$

↳ why एक साइड को sample लिया है?

→ चूंकि Pop Acces नहीं है तो Sample का M diff
→ तो PD ज्यादा है या SD
↳ कह सकते है → $(x_i - \mu)^2 > (x - \bar{x})^2$ → और यह Assume कर लेते यह सभी Point के लिए true होगा।

ex: $\frac{10}{5} = \frac{2}{1}$ Ratio होगा अगर 10 → 8 करते है तो Ratio same हो तो denominator को भी कम करेंगे $\frac{8}{4}$

Ratio Same रखना है

So that's why

$\frac{(x_i - \mu)^2}{N} > \frac{(x - \bar{x})^2}{n-1}$

Why Ratio Same → क्योंकि हमें यह 2 Constant रखना था इसलिए

→ चूंकि $x_i - \bar{x}$ के Value में कम आएगा क्योंकि यह Sample का है तो $x_i - \mu > x - \bar{x}$ रखने के Ratio Same रखने के लिए n−1 करते है।

∵ चूंकि PD, SD के ratio को maintain रखना है तो इसलिए S.Var में n-1 कर देते हैं

ताकि $\frac{(x_i - \mu)^2}{N} > \frac{(x_i - \bar{x})^2}{n-1}$ True हो।

but why then n-1 not n-2, n-3, ...

→ Concept here → **Degree of freedom** | **total no. − 1** max frequency

Maximum no of logical Independent Variables.

eg: 6 5 9 2 28 — avg of 5 no is 10

any no here Put → No → 28 So Avg 10 आएगा

Any Number — what no? but here — Constant

eg: 100 → 10 elem

→ तो यहाँ भी कितना No. को Atleast fix रखना होगा

No of possible are Changable

Atleast 1 element should not be Changable

So that n-1 come into picture

*** So max^m reduction you make n-1 to be called as a Sample

- you Remove 9 elemt and only 1 part called as a Sample

Measure of Symmetricity:

→ Symmetricity: Object sym ↔ mirrimage

① Why learn → what is

→ then how to measure the Symmetricity

help of

① Skewness

② Kurtosis (trade lots)

① **Skewness** → dataset's symmetricity

if when data is symmatric eg: — this is distribution

in nature then Skewness = 0.

called symatric ⇒ Skews

Mean = Median = mode

!! Why → because skewness is not symmatric.

Then how skew data looklike

Non-symatric

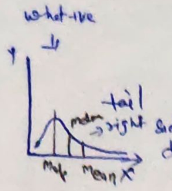
चाहे कि mirroring Not equal

→ If in boxplot →

→ histogram Vs density

(a) Positive Skewed :- (Right Skewed data)

what-ive



tail → right side of distribution

example : ① Distribution of Salary in Company

② Distribution of Marks in difficult exam

③ Avatar 2 movies → Release time lot book but time Reduces

So ↳ Where where is Mean Median Mode

↳ mean ≥ median ≥ mode

∴ median → Physical mid point

50 | 50

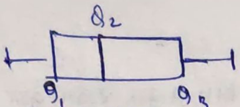
→ What is the reason of Right Skewed ?

- due to outliers (जरूरी नहीं है कि यहाँ Correct है या नहीं उसका कोई Sense है या नहीं !)

→ eg:

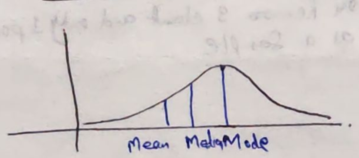
eg:

→ Box Plot



$Q_3 - Q_2 \geq Q_2 - Q_1$

(b) Left Skewed data



mode > median > Mean

Reason → outliers

eg ① Death Rate 50 age

② Marks in easy exam

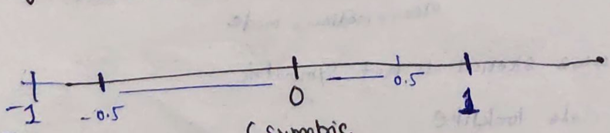
③ Wealth distribution

Q → How Skewness = 0 → Symmetric data

$$Sk = \frac{x_i - \mu}{n\,\sigma^3}$$

(μ → mean, n → data no, σ → Sigma)

if skewness



- −0.5 – 0, 0 – 0.5 } fairly Symmetric
- (−1)–(−0.5), 0.5 – (1) } → moderately skewed
- < −1 & > 1 } → highly skewed → worry about

df.Col.skewness

→ How to see Skewness | How to know

by ① Visualization

② Q,Q Plot

③ Skewness formula (Statistical way)

→ for ML data not be skewed ↓ because of outlier

Question → why Outliers bad in ML ↳ in some ML → that not give you expected Predictions

→ How to use in Industry

- Visualization

→ How to treat Skewness
↳ Transformation
(1) treat outliers
(2) Box-Cox transformation / yeo Johnson
(3) Exponential transf
(4) Reciprocal
(5) log transform

Why all we learn stats
↳ 1st moment
↳ mean $= \frac{\sum (x_i)^1}{N}$ → mean is computing taking 0 as reference

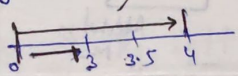
What common in these
↳ 2nd moment
↳ Varp $= \frac{\sum (x_i - \mu)^2}{N}$ → Here mean is the reference

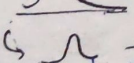
→ Movement about mean

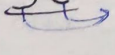
3rd Movement
↳ Skew $a_3 = \frac{\sum (x_i - \mu)^3}{N \sigma^3}$ → mean is reference

4th moment
Kurtosis → $\frac{\sum (x_i - \mu)^4}{N \sigma^4}$

What is actual meaning of mean?
↳ $\frac{\sum (x_i - 0)}{N}$ → On an avg how much a specific value or data point lie away from origin.

↳ if avg of 2 no. 3 & 4 → 3.5 → 

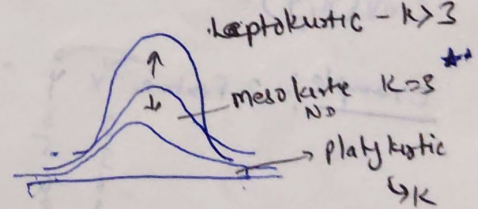
Question  me Say → Outlier

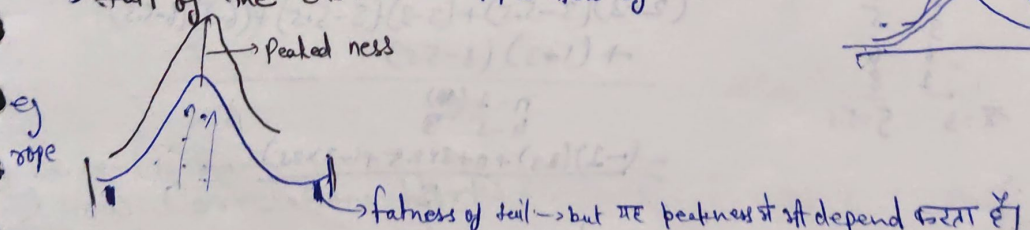
if ↳  → It can not have outliers ?! (outliers हो सकते हैं) ✗ not necessarily

if we can't find outliers in Visualization then use box plot.

⇒ Kurtosis called ↳ tail of the distribution / fatness of the tail.

$$a_4 = \frac{(x_i - \mu)^4}{N \sigma^4}$$





How Change
- K>3 → Leptokurtic → tail is fat → Many outliers
- K=3 → Meso ✓
- K<3 → Platykurtics ↳ also check is it ok.

k is reference point

Usecase!
- Buy Stock in Bulks
- Mutual funds house
- Sudden Spike in data (crypto) जहाँ K-3 we in safe pit

- Matrices / Vectors
  - Eigen Vector / Eigen
  - Probab. Value
  - Diff. of Calculus

Corr & Covar
Prob
Prob distribution

Correlation / Covariance :-

eg:- (1) House price feature → rooms

Cola ↳ Whatever you are going to predict / Price of a house, it should have some relation with the features

↳ Price of house ∝ Area of house, No of rooms, locality, relationship.

(2) No of match played → & total score

↳ How to measure these relationship. ?

(1) Covariance
(2) Correlations

(3) which feature we use or not in ML.

Ans :- Features which has high correlation value wrt target variable that going to be used.

* Covariance.

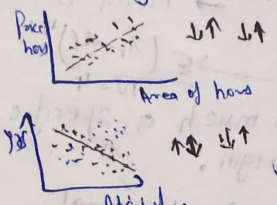
What is Variance (Spread of the data) → $\frac{\Sigma(x_i-\bar{x})^2}{n-1}$

Can write like $\frac{\Sigma(x_i-\bar{x})(\bar{x}_i-\bar{x})}{n-1}$

$\frac{\Sigma(x_i-\bar{x})(\bar{y}_i-\bar{y})}{n-1}$

Co - Variance

before it → Scenario 1 / Scenario 2



To determine relationship b/w 2 variables.

So we can also say Variance is nothing relationship itself

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

Example :- 2 feature

x	y
2	3
3	5
6	6
1	8
$\bar{x}=3$	$\bar{y}=5.5$

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{(2-3)(3-5.5)+(3-3)(5-5.5)+(6-3)(6-5.5)+(1+3)(8-5.5)}{4-1}$$

$$= \frac{(-1)(2.5)+0+3\times0.5+(-2\times2.5)}{3}$$

$$= -\frac{1}{3} \Rightarrow -0.33$$

What is the meaning of −0.33

it means 2 feature (x, y) is -vely related.

x	y
2	3
4	5
6	7
$\bar{x}=4$	$\bar{y}=5$

$Cov(x,y) = +ve$

x, y is +vely related

why not talking about Magnitude

$Cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1}$ Sq. kg

eg. Sq ft, lot

Name owner	Sqr Area	Wt/Area (Wtg per sqr)	time yrs	y
Ajay	1800	150	10	1.2 Cr

Let Cov ⇒ 0.5

Cov sq.kg (we do this) but Is it make any sense.

If $Cov(f_1, f_3) \Rightarrow$ Sqft. years $\overset{Let}{=} 0.6$ (Unit are different)

So we can make statements that
↳ The Strength of relationship b/w f_1 & f_3 $\overset{0.6}{>}$ f_1 & f_2 $\overset{0.5}{}$ → अगर unit अलग है तो Unit different and magnitude is not helpful

$\times$

① Since the dimensions are different we can not compare the magnitude.

And Question 2 → What would be the value of → $Cov(x, y)$ Answer → $-\infty$ to ∞

② If $cov(f_1, f_2) = 0.1$ → then we say $cov(f_1, f_3)$ is twice f_1 & f_2 → we say $\times$ (why because value is $-\infty$ to ∞ cov)
$cov(f_1, f_3) = 0.2$

→ Two Disadvantages of Covariance:
① Dimensions/unit is there
② $-\infty$ to $+\infty$

⇒ then what is use of CoV
↳ takes about direction of relationship only

Than what will be solution → in Magnitude
↓ come & Related

Correlation $= \frac{Cov(x, y)}{\sigma_x \times \sigma_y}$ → $\frac{kg.\,sqft}{kg.\,sqft}$ → So we say dimensionless quantity

σ_x → kg, σ_y → sqft

when divide $\sigma_{x,y}$ so the CoV Range in -1 to 1

Solve Problem ② also

↳ df.corr()

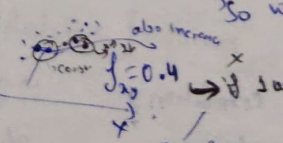
H.W Python Example ⇒

x	y
5	3
8	4
9	5
6	7
3	8

we need to know
⇒ How Correlations Works?

eg: $f_{x,y} = 0.4$ $f_{A,B} = 0.8$

↳ So we say $f_{A,B}$ is highly correlated than $f_{x,y}$

eg: 

→ if 1 unit change in x then 0.4 time change in y features. ↳ X

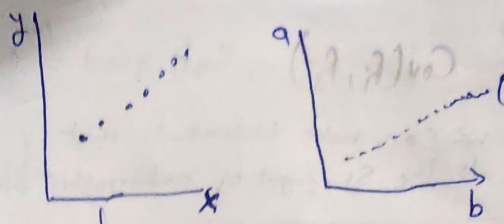
So Now
0.4
① +ve correlate

Correct Intrepretion is
↳ How much sure it is if there is 1 unit change in X the other unit will also change.
↳ here this change may be possibe in increase & decrease

Book
Shalto Rai

Solve 1 Problem

eg:- we say → -1 0 1
-ve corr +ve corr

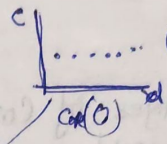


→ Correlation is not about slope ok

Q) which is more correlated ?
Ans → both are same Corr
↳ so why Previous definition work

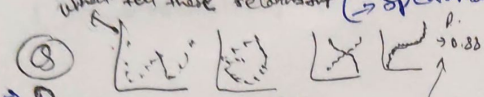
So → 1 ⇒ perfect correlation
if 1 unit ↑ then 100% chance y will change

$\rho_{xy} = 1$,
$\rho_{a,b} = 1$



Q) what is correlation

what tell these relationship (→ Spearman rank correlation



→ Pearson Correlation Coefficient ?

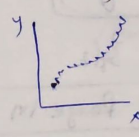
It ↳ Always measures linear relationship

$$Cor(x,y) = \frac{Cov(x,y)}{\sigma_x * \sigma_y}$$

Spearman rank Correlation → Non-linear relationship.

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

Rank(y) = R(y)



Spearman corr = 1
Pearson corr = 0.88

what Rank of x, y we have to arrange or sort Value

eg:

x	y	R(x)	R(y)
5	6	3	1/2
7	4	2	3
8	3	1	5
1	1	5	4
2	2	4	

Sort → Rank(x) ⇒ 1, 2, 5, 8, 7, 8
~~5 4 3 2 1~~

Rank(y) ⇒ 1, 2, 3, 4, 6
~~5 4 3 2 1~~

Use Case :-
↳ 1k features in house price predictions.

→ Use heatmap

→ How will you use correlation ?
Ans → Calc Corr to see relation if
Close to 1 or -1 then use in ML
if Corr(0) = 0 drop that after consult BI.
(good corr (0.3 to 1) or -0.3 to -1)

in datascience never use ↳ Spearman Correlation always use Pearson Correlation

Probability :-

Random variable

① Experiment → A repeatable procedure with a set of possible result & eg tossing a coin

② Sample space → All possible outcome of an experiments.
Sample space → Head & tail.

③ Event → One or more outcome of Experiment.

eg P(H) = 1/2 (0.5)
↳ why 0.5 Adv

Basic Counting principle :-

↳ m ways to do one thing and n way to do another thing then there are total m×n ways to do both the ~~probability~~ things.

Independent event
Exclusive

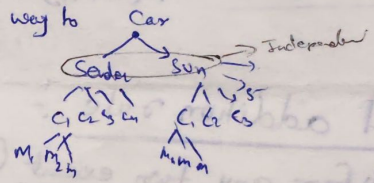
Eg ① 3 Shirts, 4 jeans → 3×4 → 12 using combo

② 3 icream, 4 cones → 4×3 → 4×3 = 12

More than two choices

option 1 → Sedan or SUV
color opt 2 → $C_1 C_2, C_3 C_4 C_5$
opt 3. - 3mode - (M_1, M_2, M_3)

How many way to buy car



2×5×3 = 30 ways

• these basic county principle in work in independent event. → Adv

⇒

⇒ Probability principles :-

• What is Probability?

Prob ⇒ Share of Success | Total no of Possible outcomes

Canwe say
It is Relative frequency → example: Prob Head → P(H) = 1/2 → Share of success (Coin) / Total outcom

We say Rf is what is freq of Head

• Histogram count
• PDF
Cont Dist/n

Principle

1. The exact outcome cannot be predicted (exception → person Shakunimama)
2. All possible outcomes are known.
3. Equally likely outcomes. → means → eg: if throw is fair H & T chance equally out
4. Repeatable under uniform condition. → (means eg → if you fix the angle of dice, friction, speed, rotation. → then come only 1 outcome eg → H) but Ideally practicaly not possible.

Probability Rules

1. for any event A, $0 \le P(A) \le 1$ → why no get 1 → because it is (toss coin) / or you divide by total number
2. Sum of all Probability = 1.
   eg:- if toss a coin — Success → (H, T) → total no. of outcomes (H, T) → 2 → $\frac{1}{2}$
   eg: $P(H) + P(T) = 1$
   $P(H) = 1 - P(T)$ {Rule of subtraction}
3. Complement rule → $P(\text{not } A) = 1 - P(A)$
4. General addition rule:-
   → for any two event
   $$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$
   Special case of disjoint events → $P(A \text{ or } B) = P(A) + P(B)$
   → those event cannot occur together
   $P(A) + P(B)$ eg:- move your head in one go left & Right
   eg:- travelling by train & Airplane same time

   P(A or B)

   → A or B → A+B − A∩B
   → common part

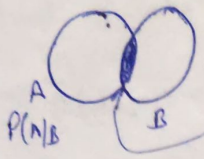
5. General multiplication Rule:-
   - $P(A \text{ and } B) = P(A) * P(B/A)$
     $\Rightarrow P(B) * P(A/B)$

   Special A and B are independent → (independent event → does not affect each other, may be occur together)
   $P(A \text{ and } B) = P(A) * P(B)$
   $P(A/B) \Rightarrow 0$ (Prob A when B has occured)
   eg:- I having breakfast & Mr Modi in Bill in Lok
   eg:- I having breakfast & wanting youtube. 10 pm, 10 pm

⇒ Conditional Probability :

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

P(A) when event B is happen

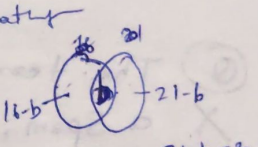


→ Already occur
अगर पहले से गया then
A निकलना है तो वो सिर्फ
Intersection Part में ही आएगा।।
और total क्या है B ही
इसलिए यहाँ

$\frac{P(A \cap B)}{P(B)}$ ← और इसका Probability निकलना है। जब Already हुआ है।
← total यहा B है।

Now go to → Rule (4) for Better Understanding

⇒ Q. 16 people study french, 21 study spanish there are 30 people altgether



$16 - b + b + 21 - b = 30$
$-b + 37 = 30$
$b = 7$

(a) work on Probability
$P(F) \rightarrow \frac{16}{30}$
$P(S) \rightarrow \frac{21}{30}$
$P(f \text{ only}) \rightarrow \frac{9}{30}$
$P(S \text{ only}) \rightarrow \frac{14}{30}$
$P(f \text{ or } Sp) \rightarrow \frac{30}{30} = 1$, $P(f \text{ and } S) = \frac{f \cap S}{\text{total}} = \frac{7}{30}$

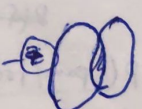
Special case of Condition

⇒ Bayes theorem

we know Conditional Probability $\Rightarrow P(A/B) = \frac{P(A \cap B)}{P(B)}$ — (1)

or

$P(B/A) = \frac{P(B \cap A)}{P(A)}$



$A \cap B$ or $B \cap A$ → are same ye

we say

$P(A \cap B) = P(A/B) * P(B)$ — (1)

$P(B \cap A) = P(B/A) * P(A)$ — (2)

equality (1) & (2)

$P(A/B) \cdot P(B) = P(B/A) * P(A)$

Bayes theorms

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

→ Naive Bayes
ML
→ Most classification

Revision

1. A dice throw prob 6 ?
2. A bag 5 red & 3 blue → Random. Prob of yellow ⇒ $\frac{5}{16}$
   8 yellow

3. 52 card → Prob heard/Red ? $\frac{13}{52} \Rightarrow \frac{1}{4}$

→ Likelood → (Prob of reverse happening)
→ Prior Prob

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

P(B) → Evidence

Posterior Probability
→ What is updated prob of A if some evidence of event B is there.

think on
P(A|B) → Its meaning that evidence of B given if prior prob of A was P(A).

→ wrong value
✗ not good

Q) The term free occurs in 20% of the emails marked as
✗ a Spam. 0.1% of non-spam emails include the term free.
50% of the emails are spam.
Calculate prob that an email is spam if the word free appears in it.
→ P(spam/free)

Ans
P(spam) = 0.5
P(free/spam) = 0.2
P(free/nospam) = 0.1% ⇒ 0.01%
or
~~P(free)~~

✗

$$P(spam/free) = \frac{P(free|spam) \times P(spam)}{P(free)}$$

$$= \frac{0.2 \times 0.5}{0.01} \Rightarrow \frac{0.10}{0.01} \Rightarrow \frac{10}{1} = 10\%.$$

Q.) 10% of patients in a clinic have liver disease. 5% of the clinic patient are alcholics. Among these patients diagnosed with liver disease 7% are alcoholics. You are interested in knowing the probaly having liver disease given that he is an alcoholic.

P(A) = Prob of having liver disease = 0.10
P(B) = Prob of alcoholm = 0.05
P(B/A) = having liver dis who has alcohism ⇒ 0.07

P(A/B) → Prob drink alcoh / % of alcohol having liver disease

$$\frac{P(B|A) \cdot P(A)}{PB} \Rightarrow \frac{0.07 \times 0.10}{0.05}$$

⇒ 0.14 ⇒ 14%.

Clear नहीं है कितना % A का है total का 50, use → law of total probability

Q.2 In a factory 30% of the products are defective. If a defective product is randomly selected, what is the probability that it was produced by Machine A, given that Machine A produces 60% of defective products?

Adv

↳ ~~P(defective) = 30%~~

$$P(A/D) = \frac{P(D/A) \cdot P(A) + P(D/B) \cdot P(B)}{P(D)}$$

→ Law of total Prob

$$P(D) = P(D/A) \cdot P(A) + P(D/B) \cdot P(B)$$

Use Case:

- AI
- Medical testing
- financial Risk
- Marketers (feedback)
- Medical diagnostics +ve, -ve

eg: ① eg: House Price Pr

f_1 room, f_2 Area, f_3 Loyalty | Price

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

so we write like

$$P(f_1 f_2 f_3 / y) = \frac{P(y/f_1 \cdot f_2 \cdot f_3) \cdot P(f_1 f_2 f_3)}{P(y)}$$

break

$$= \frac{P(y/f_1) \cdot P(y/f_2) \cdot P(y/f_3) \cdot P(f_1) \cdot P(f_2) \cdot P(f_3)}{P(y)}$$

eg: if what is price if room area loyal 3 3 3

↓

Naive Bayes Model

↓

In ML

Called NB

eg ② Play/not Play

Outlook	temp	Play
Sunny	Hot	No
Rainy	Cold	Yes
overcast	Mild	No
Sunny	H	Yes
=		

We say

	Yes	No
Sunny	2	3
overcast	4	0
Rain	3	2

	P(Y)	P(N)
	2/9	3/5
	4/9	0/5
	3/9	2/5
	100%	100%

P(Y) → 9/14

→ P(N) = 5/14

③ Today (Sunny / Hot) → Play or not

S H

↙ so here use bayes theorem

$$P(Yes / Today) = \frac{P(Sunny/Yes) \cdot P(Hot/Yes) \cdot P(Yes)}{P(Today)}$$

(Today = S, H)

or

$$P(No / Today) \Rightarrow \frac{P(Sunny/No) \cdot P(Hot/No) \cdot P(No)}{P(Today)}$$

(Today = Sunny, Hot)

→ Constant for both / Same

④ Use Temp

	Yes	No	P(Y)	P(No)
hot	2	2	2/9	2/5
cold	4	2	4/9	2/5
mild	3	1	3/9	1/5
	9	5	100%	100%

$P(Yes) = 0.031$

$P(No) = 0.085$

Normalize these

$$\frac{P(Yes)}{P(Yes) + P(No)}, \quad \frac{P(No)}{P(Yes) + P(No)}$$

$$\frac{0.03}{0.03 + 0.085} \quad \frac{0.085}{0.03 + 0.085}$$

Random Variable

Start → (1) Tossing a Coin →

Experiment is Random → (H), (T)

the Outcome will be Random

How Represent in Maths

$X = \begin{cases} 0 & \to H \\ 1 & \to T \end{cases}$] → quantified each of the event

X called Random Variables

(2) $X_{dice} = \{1, 2, 3, 4, 6\}$

↳ discrete → this is Called discrete random Variable

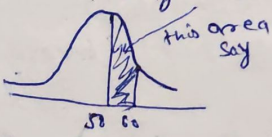
~~Which~~ This is discrete or Cont...

→ Continous Random Variables

age
hight

distribution of Age → Continous → then How to Calculate the probability here

↳ (1) If what is prob of Age 50 & 60 Age



- A Random Var Can take any Values
- random Variable Value is Unknown. $x \{1, 0\}$

eg: Toss {H, T)

$P(H) = 1/2$
$P(T) = 1/2$
} → give me a generic function
↳ give me a graph to calculate Prob of a event while tossing a Coin

$$P(x) = \frac{1}{n}$$; n = total no. of outcome

$P(x=0) = 1/2$
$P(x=1) = 1/2$

(Q.2) dice {1, 2, 3, 4, 5, 6)

$P(2) = P(1) = P(x) = \boxed{1/n}$ → we say 1/n function that can be use to get probability.

$P(3) = P(1/6) \Rightarrow P$

★★ Conclusion → Outcome of Experm

Discre → Prob mass function

Continous → Prob density fun

P(H = 0.5) → Law of large Num (1000)
↓ converges to average
↓
50% → Head
50% → Tail

in Algebra:
$x + 5 = 0$
$x = -5$
X → value is fixed

- Experiment → Random → Outcome → Random → discrete / Continous

'But what → Probability distribution functions'

→ Distribution use
→ data Analysis & Summarize
→ Understand datasc

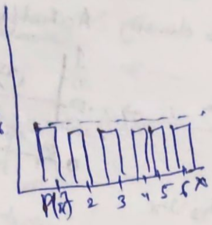
- irrespective of outcome nature, draw outcome in a form of distribution → Prob distribution fn.

Prob distribution fn (↓ → Prob mass fn (discrete) ; ↘ Prob density fn (Pdf))

eg: trow dice

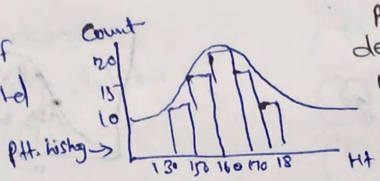
$P(1) = 1/6$
$P(2) = 1/6$
$P(3) = 1/6$
$P(4) = 1/6$
$P(5) = 1/6$
$P(6) = 1/6$

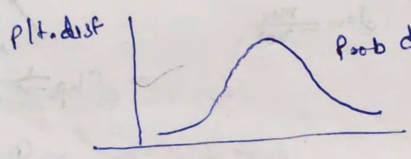
⇒



discrete → So bar graph ~~histogram~~ or PMF

② Continous = Pdf
Height of Std.





PDF is a Statistical term that described the probability distributn of the continues random Var

PDF use case (continous)

1. Anomaly detection — fraud detection → spending behavior that fall into region may be fraud
2. KDE (Kernal Density P) → Smooth histograms → Understand data
3. Bayesian Inference → update the Distb of new eding

PMF use case (Discrete)

1. Classification Proble → multi non Logistic Regs
2. NLP → model the prob of words or seq of words eg. n grams
3. Recommeder Sys → to model user Preference → Collaborative filtering

Deep dive

eg. Rolling of dice ⇒ PMF

$P(x=1) = 1/6$

$P(x=2) = 1/6$

$P(x \le 3) = P(x=1) + P(x=2) + P(x=3)$

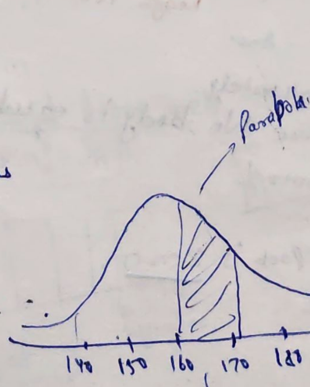
$\Rightarrow \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \Rightarrow \frac{3}{6} \Rightarrow \frac{1}{2}$

$P(x \le 4) = \frac{4}{6}$

$P(x \le 6) = 1$

eg. PDF → continous
Height of Std

Ⓠ What is Prob its Height $P(160 \le x \le 170)$?
or
if ask Prob(≤ 150)



Parabola Curve, has a eq

Note
- Area under Curve = 1 - always → because Prob always = 1 → use by Calculus Integration

⇓
→ How actually work

$\int_{160}^{170} h\,dx$ → sum (all small rectangles)

dx — all are small □

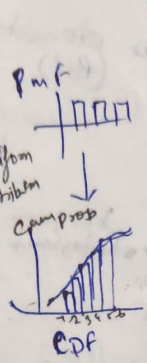
↳ How you will get this area / (without Integ..

③ Cummulative Density function (CDF)

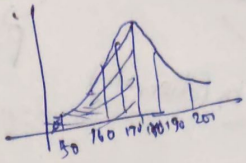
adding values to current point

eg:- Pmf → discrete

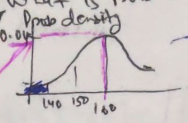
x	P(x)	cdf
1	1/6	→ 1/6
2	1/6	→ 2/6
3	1/6	→ 3/6
4	1/6	→ 4/6
5	1/6	→ 5/6
6	1/6	→ 6/6

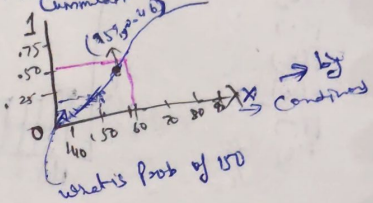


④ Pdf



What is Prob density? Actually Cummulative Prob of



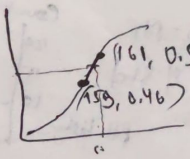


Ans → ① Statistical Answer

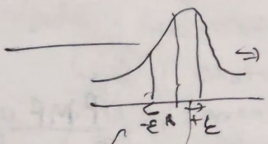
So we say that

→ Prob density of Pdf
↳ ⇒ Slope of CDF

PDF Probbu



this value



$\Rightarrow \frac{-\varepsilon - 160 - \varepsilon}{\text{Total no. Probability}}$

plane $\Rightarrow \frac{\text{mass}}{\text{Vol}}$

How to cal Slope

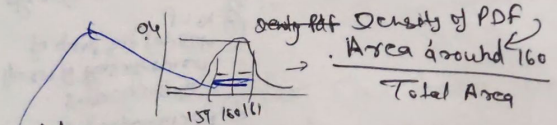
$\text{Slope} \Rightarrow \frac{y_2 - y_1}{x_2 - x_1}$ (Gradient)

$\Rightarrow \frac{(0.54 - 0.41)}{161 - 159}$

$\Rightarrow \frac{0.08}{2} \Rightarrow 0.04$

Where is 0.04

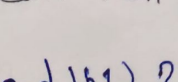
why



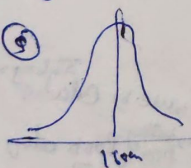
Note

Prob density can be greater than 1

↳ Why took value of around 159 and 161) ?



⑤ some one ask what is prob a person will have height 160 cm

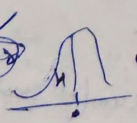


① what P(x = 160 cm) — Ans

Ans = 0 → why o think mathmaticly
↳ because No body is equals to 160 cm

So use interval 160.000000 1

2 Answer



Area of Point = 0 so the 160 Prob is Zero.

→ Use Case → ① Risk Management
② Calc the Prob of diffrnt risk level :- eg (help determine the Prob that an asset's value will fall below a certain threshold, which is crucial for Risk assesments

② Quantile Calculation
↳ Setting thresold for decision-making, ex high risk asstes

③ Hypothesis testing
↳ Calc the P-Value in Statistical tests
eg A/B testing →

→ Probability distribution

Prob distrn
- Discrete
  - (1) Binomial only 2 outcomes
  - (2) Poisson: Center No of exp → Intime for 5 visitor/hr
  - (3) Bernoulli: 2 outcomes single trial only
- Continu
  - (1) Normal (Gaussian distribution) mean
  - (2) Exponential dist: time between two events in a Poisson

we studying → Random Variables

- Discrete
  - If outcome is discrete than Probability → Use fn
  - ↓ Probability mass funct
  - ↓ distrib
    - (1) Discrete Uniform Discrete
    - (2) Bernoulli
    - (3) Binomial
    - (4) Poisson distribut
    - (5) Geometric distribut
- Continous
  - ↓ Probability function
    - (1) Normal distribution
    - (2) Standard Normal distribution
    - (3) Log-Normal distribution
    - (4) Chi-Square dist
    - (5) F dist
    - (6) Exponential dist
    - (7) Continous uniform dist
    - (8) t-distribution.

→ Prob functn: PMF, PDF

↳ What we see in individual Distribu
- (1) Notation
- (2) Shape → & what factor Shape depend on
- (3) PMF eq expression
- (4) CDF eq expr

⇒ In Uniform → Discrete
→ Continous

(1) Discrete Uniform distribution ($U(a, b)$)

- Discrete

→ A uniform distribution refers to a type of probability distribution in which outcomes are equally likely.

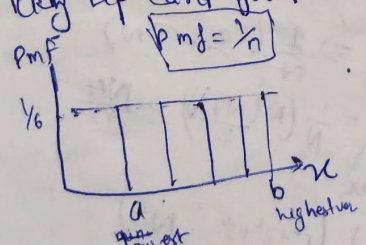
→

ex: (1) rolling a dice (dice {1, 2, 3, 4, 5, 6}) • The outcomes are independent

(2) tossing a coin

(3) Picking up card from well

PMF: $P_{mf} = 1/n$



mathematically write

$n = b - a + 1$
$= 6 - 1 + 1$
$= 6$

eg: {1, 2, 3, 4, 5, 6}
a = 1, b = 6

what is CDF?

CDF (Probability till the end)

→ eg: dice {1, 2, 3, 4, 5, 6}, a = 1, b = 6

↳ for each of the distribution is mean

mean of discrete uniform distribution = $\frac{a+b}{2}$

Variance of discrete Uniform distribution $= \frac{n^3-1}{12}$

↳ Here why we calculate the mean & Variance

we know mean & Variance give the information

early learn → mean = $\frac{\text{Sum of nos}}{\text{total no}}$ ok

so related to mean there is terms/or concept

Expected Value → In the long run avg value of represented of experiments

$\sum_{i=1}^{n} x_i p(x)$

g.

X	P(x)	$x_i P(x)$
1	1/6	$1 \times 1/6 = 0.167 \times 1$
2	"	$2 \times 1/6 = 0.167 \times 2$
3	"	$3 \times 1/6 = 0.167 \times 3$
4	"	$4 \times 1/6 = 0.167 \times 4$
5	"	0.167×5
6	"	0.167×6

mean of Uniform distrib

1, 2, 3, 4, 5, 1

$\mu = \frac{\text{sum of terms}}{6} = \frac{21}{6} = 3.5$

$\sum_{i=1}^{n} x_i P(x)$ 3.5 around 3.5

what you observes here

⇒ So we can say that mean & the Expected Value is same

$E(x) = \sum_{i=1}^{n} x_i p(x)$

Variance of Uniform distrib

$\text{Var}(x) = \text{Var}(a, b) = \frac{n^2-1}{12}$

eg Variance $= \frac{6^2-1}{12} = \frac{36-1}{12} = \frac{35}{12} = 3.$

1 — 3.5 — 6

Prove

g Sum of n natural numbe $= \frac{n(n+1)}{2}$

1 to N

Sum of n^2 term $= \frac{N(n+1)(2n+1)}{2}$

$1^2, 2^2, \ldots$

we have $\text{Var}(x) = E(x^2) - (E(x))^2$

Variance of any Number

$E(x) = \sum_{i=1}^{n} x\, p(x)$

$= \sum_{i=1}^{N} x \cdot \frac{1}{N} \Rightarrow \frac{1}{N}(1, 2, 3, \ldots + N)$

$= \frac{1}{N}\left(N\left(\frac{N+1}{2}\right)\right) = \frac{N+1}{2}$

$E(x^2) = \sum_{i}^{n} x^2 \cdot px$

$= x^2 \cdot \frac{1}{N} \Rightarrow \frac{1}{N}(1^2 + 2^2 + 3^2 \ldots N^2)$

$= \frac{(n+1)(2n+1)}{6}$

$\text{Var}(x) = E(x^2) - (E(x))^2$

$\Rightarrow \frac{(n+1)(2n+1)}{6} - \left(\frac{N+1}{2}\right)^2$

$\text{Var}(x) = \frac{N^2-1}{12}$

$\sigma = \sqrt{\frac{N^2-1}{12}}$

SD

Continous Uniform distribution

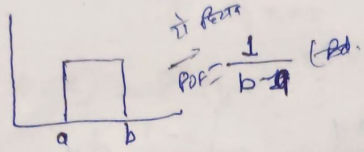
- That has infinite no of values in a specific range
- R.V is continuous
- It is rectangular distributed

Eg: ① A perfect random no. generator
② Prob of guessing exact time at any movement
③ Waiting time at a bus stop
④ temp variation in a day

① Notation: $U(a,b)$

Parameter $-\infty < a < b < \infty$

② Pdf

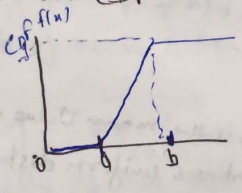


what is Rect area $= b-a \times f(x)$ | we know area under give the prob which is 1

$1 = b-a \times f(x)$

$$f(x) = \frac{1}{b-a}$$

How to make CDF



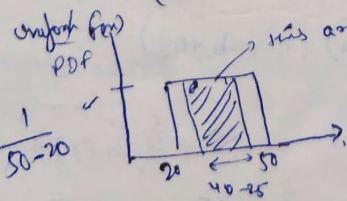
$$Cdf = \begin{cases} 0 & f(x)\ x < 0 \\ \frac{x-a}{b-a} & x \in (a,b) \\ 1 & x > 1 \end{cases}$$

→ So we see Prob lie in CDF/PDF ₹ a & 3 के बीच होगा।

↳ The mean of Conti Uniform dist $= \frac{1}{2}(a+b)$
Variance " " " $= \frac{1}{12}(b-a)^2$

③ The no. of items sold at a shop daily is Uniformaly distributed with max and min sold 50 & 20 respectly
so what is prob of daily sales to fall b/w 25 to 40 ??

$$P(25 \le x \le 40) = ??$$

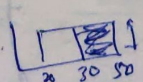


$\Rightarrow \frac{1}{b-a}$ } height बना रहा था 50

$\frac{1}{50-20} = \frac{1}{30}$

$\text{Prob}(x \le 40) = \frac{1}{30} \times 15 = 0.5$ Ans

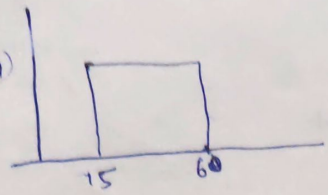
④ $P(x \ge 30)$



$\Rightarrow 50-30 = 20$

$\Rightarrow 20 \times \frac{1}{30} = 0.66$ Ans

Q. The amount of time for pizza delivery is Uniformly distribution b/w 15 & 60 mins. what is standard deviation of the amount of time it takes for pizza to be delivered

⇒ 

$Var = \frac{1}{12}(b-a)^2$

$= \frac{1}{12}(60-15)^2 = \frac{(45)^2}{12} = 168.75$

$SD(\sigma) = \sqrt{Var} = \sqrt{168.75} \approx 13$ min

optional

→ for a continuous random var prob density fn (fx)

→ by def $E(x) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$ ← for Discrete before $\sum x \cdot p(x)$

By Basic Sum → integral ↑ Summation for contin variable

↳ Uniform distribution 

⇒ $\int_a^b x f(x)\,dx \Rightarrow \int_a^b x \cdot \frac{1}{b-a}\,dx$

Basic Integration formula ⇒ $\int x\,dx \Rightarrow \frac{x^{1+1}}{1+1} \Rightarrow \frac{x^2}{2}$

$\Rightarrow \frac{1}{b-a}\int_a^b x\,dx \Rightarrow \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \Rightarrow \frac{1}{2(b-a)}\left[x^2\right]_a^b$

$\Rightarrow \frac{1}{2(b-a)}(b^2-a^2) \Rightarrow \frac{1}{2(b-a)}(b-a)(b+a)$

(b: upper value, a: lower value)

$\Rightarrow \boxed{\frac{b+a}{2}}$ this is the mean value of Continous uniform distribution

⇒ Variance of any Continous distribution → It help to derive any variance of distribution

$Var(x) = E(x^2) - (E(x))^2$ ↳ $\frac{a+b}{2}$

$E(x^2) \Rightarrow \int_a^b x^2 f(x)\,dx \Rightarrow \int_a^b x^2 \cdot \frac{1}{b-a}\,dx \Rightarrow \frac{1}{b-a}\int x^2\,dx$

$\Rightarrow \frac{1}{b-a}\left[\frac{x^{2+1}}{3}\right]_a^b \Rightarrow \frac{1}{3(b-a)}(b^3-a^3) \Rightarrow$

$b^3-a^3 = (b-a)(b^2+ab+a^2)$

$\Rightarrow \frac{1}{3(b-a)}(b-a)(b^2+ab+b^2)$

$= b^2+ab+a^2$

Put $= E(x^2) - (E(x))^2 \Rightarrow \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 \Rightarrow \frac{b^2+ab+a^2}{3} - \frac{a^2+b^2+2ab}{4}$

$\Rightarrow \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} \Rightarrow \frac{b^2-ab+a^2}{12} \Rightarrow \boxed{\frac{(b-a)^2}{12}}$

It is a Variance distribution of the Continous distribution

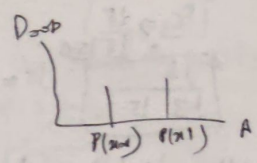
Types Prob distrib

② Bernoulli distribution :

- A discrete prob dist of a random var which takes only two possible outcomes typically labbled as success (1) and failure (0)

→ To model a problem statement with only two possible outcome

eg: Toss a coin (head or tail)



write Mathematically
$P(x=0) = 0.5 = p$
$P(x=1) = 1-p = 1-0.5 = q$
$= 0.5$

PMF ⇒ $P(x=k) = \begin{cases} p & \text{if } k=1 \\ 1-p & \text{if } k=0 \end{cases}$

↳ $\boxed{P(x=k) = p^k (1-p)^{1-k}}$

if $k=1 \Rightarrow p^1 (1-p)^{1-1}$
$\Rightarrow p^1 (1-p)^0$
$\Rightarrow p \cdot 1 \Rightarrow p$

if $k=0 = p^0 (1-p)^{1-0}$
$\Rightarrow 1 \cdot (1-p) = (1-p)$

⇒ Condition of BD

1. No of trial should be finite
2. Each trial should be independent
3. Only two possible outcome — 1, 0
4. Prob of each output should be same in every trial

Example : ① tossing a coin (both bernoul or Discret dist

② prob of getting 3 and not 3 while throwing dice
∴ 2 outcome

In ML :
↳ In classification

dataset → | Transaction amt | no of loan | Loan | Salary | fraud/no fraud |
1500	2	5	3	1
				0
				1

So use in Bern → 2 outcome see

⇒ If this classificat prob you get a cost fn which is actuall comp from bernoulli → PMF

eg:- defaltor / Not defaulter
Pass / fail in exam
diabetic / not diabetics

and convert into logistic regression using bernoulli dist

⇒ Mean and Variance of Bernoulli distribution

↳ p ~~(1-p)p~~

mean $E(x) = \sum_{i=1}^{n} x\, p(x)$

let Pr = 0.6 → $x = 1$ + $x = 0$ ← let 0.4

$\Rightarrow 1 \times 0.6 + 0 \times 0.4 = 0.6 \Rightarrow p$

Varian $E(x^2) - E(x)^2 \rightarrow (p)^2$

$E(x)^2 = \Sigma x^2 (p(x))$ → (1-p)

$\Rightarrow 1^2 \times p + 0 \times (1-p)$

$\Rightarrow p$ ⇒ $E(x^2) - (E(x))^2 \Rightarrow p - (p)^2 \Rightarrow p(1-p)$ proved

(3) Binomial distribution

2 → 2 outcom

- Binomial distributing is 'n' → Bernoulli trials

So PMF ⇒ $\boxed{{}^{n}C_{k}\, p^{k}(1-p)^{n-k}}$ for Bernoulli $\boxed{p^{k}(1-p)^{1-k}}$

$n, k \neq 0$

↳ no of trial

How to intrepret ⇓

eg:- 1st toss 2nd ... 10th toss

⇓

what say binomial

$\frac{\lfloor n}{\lfloor n-k \rfloor k}$ ⇒

${}^{5}C_{3} \Rightarrow \frac{\lfloor 5}{\lfloor 5-3 \lfloor 3}$

$\frac{15}{13 \quad 15-3}$ $15 = 5\times4\times3\times2\times1$

↳ what is prob of getting 'k' heads out of n trails

n → total no of trail's

k → the no of event that you are get.

eg with 3 tosses what is prob of getting exactly 2 head ??

${}^{n}C_{k}\, p^{k}(1-p)^{n-k}$

⇒ $n = 3, \; k = 2$ prob of head = 1/2

$P(x=2) = {}^{n}C_{k}\, p^{k}(1-p)^{n-k}$

$\Rightarrow {}^{3}C_{2}\left(\frac{1}{2}\right)^{2}\left(1-\frac{1}{2}\right)^{3-2} = \frac{\lfloor 3}{\lfloor 2 \lfloor 3-2} \cdot (.5)^{2}(.5)$

$\Rightarrow \frac{3\times2}{2\times1}(0.5)^{3} = 3\times(0.5)^{3}$

$= 3\times0.125$

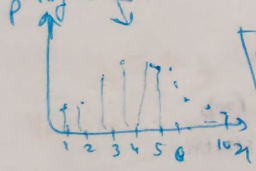
$= 0.375$

(Q) What you toss a coin 10 times what is prob that you will get head exactly 3 times

⇒ $n = 10, \; k = 3, \;$ Toss $= 1/2$

What is ${}^{10}C_{3}$ means

P Prob



Binomial

mean = np

Var = np

(4) Po

→ (5) Poisson distribution

(4) Geometric distribution

↳ It represent the probability of getting first success after having a consecutive no of failure (after n no of tails)

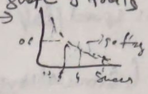
eg: I am interest in getting (H)

1st toss, 2nd toss, 3 - 4th toss

T T T H

↳ you interest and now stop the trail → this is Called as G.D.

→ shape of tails



⇓

$$PMF \Rightarrow P(x=x) = (1-p)^{x-1} \cdot p$$

eg G.D → jumping ball → ...

→ Not so much usecase

$$mean = 1/p$$

$$var = \frac{1-p}{p^2}$$

(5) Poisson distribution

3.27

← important for BI, data Analyst.

It is a discrete prob distribution that describes the no of events that occur within a fixed interval of time or space given a known average rate of occurrence.

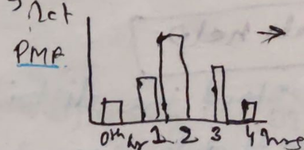
eg: No of events occuring in a fixed time interval.

(1) No of calls recieved by a customer care every hours.

Expected of calls every hr is $\lambda = 100$

↳ these type of Buss Model use Possion distribution

↳ Act PMF



↳

(2) No of people going to templ

" " " " to hospital/airport every hours -

$$PMF = P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

λ - avg rate of events every hr/interval

$e = 2.718$

(Q) if $\lambda = 10$

Prob of a person visiting at 5th hr

$$P(x=5) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} =$$

(Q) The Avg no of customers entering a store in an hours is 5. What is the prob exactly 3 customes will the store next hrs

end 3:44

$\rightarrow \lambda = 5 \quad P(x=3) \Rightarrow \frac{e^{-\lambda} \cdot \lambda^x}{x} \Rightarrow \frac{e^{-5} \cdot 5^3}{3!} \Rightarrow \frac{(2.718)^{-5} \times 125}{6} \Rightarrow \frac{0.00674 \times 125}{6}$

≈ 0.14 Ans

Ans 14% change to come 3 person.

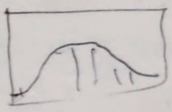
No. of Probability desity funtion (for continouse random variable)

① Normal distribution | Gausian distribution | Bell-shape distribution

why most see in data
↳ most of the naturally occuring | human generated follow a normal distributions.

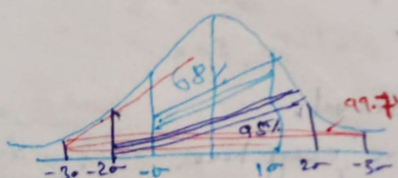
→ The distribution helps in hypothesis testing how will you do HT if you don't not what is the nature of of distributions,

Why
example → ① Quincunx (math is fun)
→ ley.
↳ we say → wandering walk
no matter what is speed | Ball
② Galton board (simulations
What is point g show
↓
let again →
Experiment · Measure of height



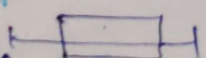
Why learn?
↳ ① natural data follow.
↳ ② Statisticaly → by its characterstics
① Symatric in nature
② skewness is Zero
③ kurtosis = 3 (mesokurtic)
④ mean = median = mode
→ but not so much impact bcz also find these of other

why wonder ③ - 68-95-99.7% imperical rule.
↳ why emprical.
↳ bcz it is true & found it with experiment | Also no any statistical proof also was. [?] of experiment



how it formula help.

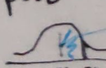
It helps to Understand the distribution of data
① Outlier of data (show) → Box Plot



y why we use Q1 to Q3
↳ How say most data lie in Q1 to Q3 ← come 68% (?)
We know -1σ to 1σ 68% data lie

we say extreme values (low, high) low in persposition or high
Ans → low in persposition (→ (normal data)
↓ occurance ↳ means ratio) or majority
beyond -3σ to 3σ

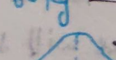
Q) ③ Talk about Prob density
① ↳ Calculate the prob → ① people will fall b/w 60kg & 70kg



↳ how find
↳ by calculate the are (AUC)
↳ not manualy,
↓ by use table

• Book ↳ No

② What is Prob if the wt is 60kg →



→ Ans is Prob is 0 because the AUC g point is zero

Real Example: Ht weight, Mean of error : Score, Blood Presure,

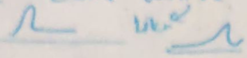
Can we mean = 0 } → for normal distribut std = 1. } → Ans → wrong

if we say it for Standard Normal Usecas

→ Industry use case of normal distribution

- Beyond 3rd Std → outliers
- Some of ML / Statistical models require the data to be normally distributed → why → ① Most of data fall natural
  - ② outlier identification (to deal with outliers)
  - ③ Prediction will be wrong

If data is skewed or not normal dist, the data will be like



or get Prediction

Something in skewed data Prediction will be wrong (Unbias or Unbalanced)
↳ why

→ if like this data → use transformation in data

Blog (feature Transformation) in ds — in depth

1. Log transformation → for Right skew data → to normal distr ↳ data into log of data
2. Square
- reciprocal
- Boxcox

and etc

Q) Which transformation use when?
- there is rule → but also do (hit & trail) ↳ do & visualize if look normal then go. → if not normal do feature & go with nearest one

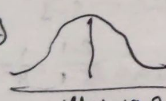
→ if you are modeling statistical ML models then data should follow normal distrib and if not do transformation — which use

---

mean = 0, Std 1 ⇒ not

mean & Std any yes can be any

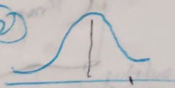
eg

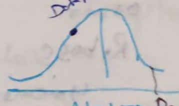


σ = 20 } what is the interpretation →
↳ on Avg the data point is 20 unit away from mean

⇒

eg ②

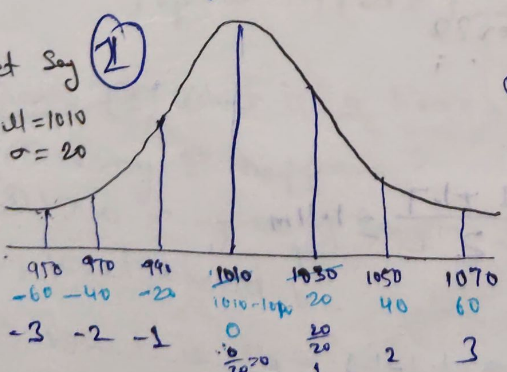




Q) which datapoint (D) among (D1, D2) is closer to μ (mean)?
↓
we can't say
↓
why → both dist Scale is not same. by Std

Let say ①

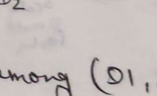
μ = 1010
σ = 20



Step ① → Can we subtract all with μ =

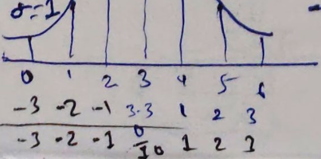
Step ② if we then divide by Std

→ $\frac{x - \mu}{\sigma}$ →

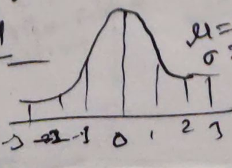


now we compare easy and it is called Standardisation of data → $\mu = 0$, $\sigma = 1$
↳ Standard normal distribution

②



→ it also follow = $\frac{x_i - \mu}{\sigma}$ → Z-score



Calcul

(Eg)

1:53

Which Z-score is higher that datapoint will be far from Mean

2:80 Industry usecase of Std N.D

(1) Standarization := Z-score = $\frac{x_i - \mu}{\sigma}$

Eg. house Price Predict

Room No.	Area of	Price
1	2000	
2	900	
3	3000	
2	1500	
1		

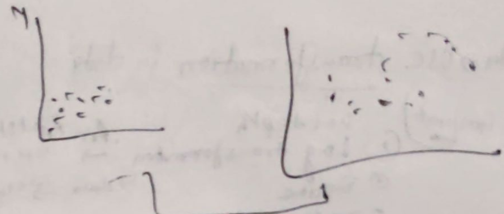
Work also

Both are in different Scale

→ unit is different also

→ for ML no problem in Prediction

→ but what is problem



Both data have different Scale

⇓

Learning Parameter will also vary much in terms of Scale

⇓

Scaling (Optional)

↳ for Small range

Type

When to use which: Outlier → std; No outlier → normalize

(1) Standarization ($\mu=0, \sigma=1$)

(2) Normalization (Range 0 to 1)

(Min-Max Scaler)

$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$

→ eg $\frac{1-1}{5-1} = 0$

$\frac{2-1}{5-1} = \frac{1}{4} = 0.25$

do Scaling why.

↳ Better Implem

→ faster optimizat (Gradient descent)

↳ Lk Lk by Scale small faster

other

Robust Scaler

↳ remove of effect of outlier

2:25

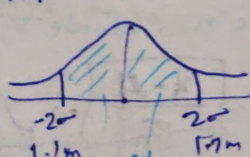
(Q) 95% of Students at School are b/w 1.1m and 1.7m tall.

Can you calculate mean and std dev??

95% → 1.1m and 1.7m tall

↓

2 Std dev



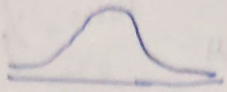
⇒ $\frac{1.1+1.7}{2} = 1.4m$ μ

4 std total deviat ⇒ 1.7 − 1.1 ⇒ $\frac{1.7-1.1}{4} = 0.15m$ Std

2150 17 mar

Central limit theorem (8th wonder of world)

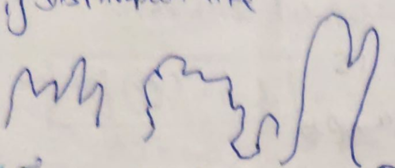
We know the → Normal Distribution:



→ Symmetric
→ Skewness = 0
→ Kurtosis = 3
→ mean = median = mode
→ 68-95-99.7 rule

How powerful
→ normal
exponential
family

If Distribution like



Haphazard/irregular Distribution ← We cannot define statistic for irregulars !

↳ It not useful for us → by statistical find a way

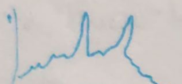
CLT → A way to convert any form of distribution to Normal distribution.

How works/or CLT says : eg S1 (50) → $\bar{x}_{s1}$
S2 (50) → $\bar{x}_{s2}$
S_{20k} (50) → $\bar{x}_{s20k}$

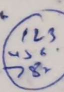
$\mu = \mu$

$\sigma = \frac{\sigma}{\sqrt{n}}$

If



→ Sample with replacement
↳ a datapoint selected in S_1 → may or may not in S_2



$S_1(1,2)$
$S_2(3,4)$
$S_3(2,4)$
$S_4(3,4)$

→ The CLT states that if you have any pop with μ, σ and you take sufficiently large random sample from population with replacement normally distributed.

- Sampling mean of population (μ, σ) will approximate to a normal distribution

Population (μ, σ) → large no. of sample → Sampling → ND
$\mu = \mu$
$\sigma = \sigma/\sqrt{n}$

Questions: (1) What is n here?
↳ Sample size eg Sample = 20 ↳ in Sample size 50
(2) Why it happens?
(3) How many no. of Sampling I take? → Should be large → may b 60k, 1L until → The distribution look ND approx
(4) What should be value of n? ≥ 30 / if ND (exp pop in ND)
Why ≥ 30 → intuitions
(5) What is the usecase.

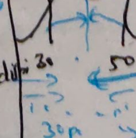
→ why n ≥ 30 → So sampling mean follow Z-distribution

Pop → Z-score = $\frac{x-\mu}{\sigma}$

⇓ CLT

$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ → Z-distribution

criteria n ≥ 30
if less → T-distribution

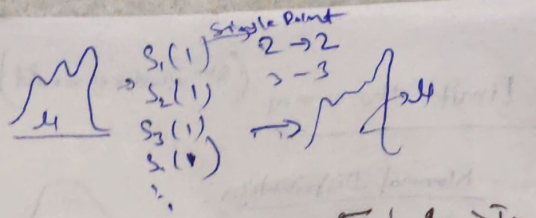


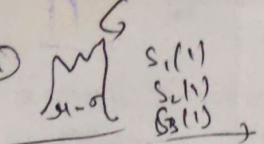
$\bar{x}$ → 30 8 50
↳ Avg - 40

If ND dis → (peaked) → so need less sample
(flat) → large sample

→ N.D → most (68%) → will lie in the center -1 std → 1 std

→ Why μ is μ after CLT ?
Why σ is $\sigma/\sqrt{n}$ after CLT??



① 

$\mu = \mu$

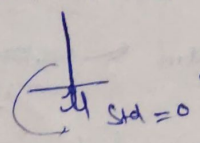
$\sigma = \frac{\sigma}{\sqrt{n}} \xrightarrow{n=1} \sigma = \sigma$ unit var

$= \frac{\sigma}{\sqrt{2}}$ $\quad 30 \geq 30$

$= \frac{\sigma}{\sqrt{3}}$

② $n = \sigma$

$S(\infty)$ ↳ whatever no of elements are there is Population you take all element in sample

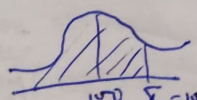


$\sigma = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\infty} = 0$

③ You have a population with $\mu = 100$ and Std dev $\sigma = 20$
If you have sample size of 50 from this population, what is the prob that sample mean will be less than 105.

$\Rightarrow \mu = 100, \sigma = 20, n = 50 \quad \bar{x} = 105$

$Z_{score} = \frac{x - \mu}{\sigma} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{105 - 100}{\frac{20}{\sqrt{50}}} =$



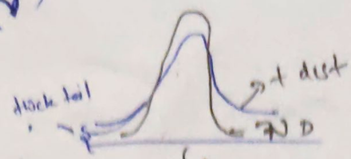
1.7124.

$\Rightarrow \frac{5\sqrt{2}}{4} = 1.7625 \Rightarrow$ area $z_{score} = 0.89$

→ It comes under set 3 → /sets/clt/clt_sum

t-distributions :- (ν)

Student t-distributions



↳ non why distribution look ND
ND

- t-distribution
- log-normal dist
- Exponential dist
- Power dist
- Chi-Square
- F distribution
- Pareto

2) Pop not given
- What if Sample size < 30 ?
  ↓
  Student t distributions

→ dof (ν) = n-1 →

(ν = +∞) → As dof increase the dist tend to Standard Normal distribution.

↳ became SND

↳ Z statistic = $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ (Z-Score)

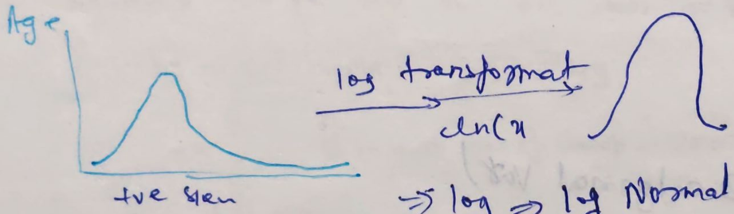
t-Statistics (Score) = $\frac{\bar{x} - \mu}{S/\sqrt{n}}$

Usecase → in T-test

→ **Intutions**
- due to Sample Size
- Thick tail means the data is present at extreme

log Normal distribut :- A continous distn whose logarithm is Normally distributed.

$x \to \log(x)$



⇒ log ⇒ log Normal distribution

- weal

UseCase :- eg example wealth distribution of world (rich attract)

self → Ain't real & not be drop these as outlier

↳ where, outliers can not be dropped, you use log N.D to make it a N.D

eg :- time spend in reading Articles (1st post.....)
:- people write comments on your post

• Powers law dist (Pareto Principle) :- eg:

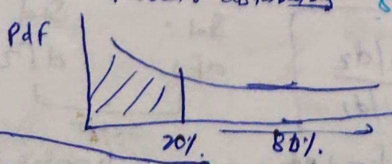
Pareto distributions — 80:20 rule

Party eg Party = 100 — 20% classe, 80% norm

20% employe — 80% sale will by 20% customer
- 80% of Rich 20 people

→ 80 thing done by top 20% people

- Iple → 2-3 player to bests



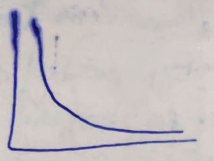
20 — 11 times → 80
Called power law

20% of x will Cover 80% of y

Pdf = $\frac{\alpha x_m^{\alpha}}{x_m^{m+1}}$ α:- parameter

Power law
Exp Pareto

(4) Exponential distribution:



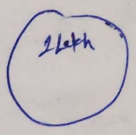
Pdf = $\lambda e^{-\lambda x}$ $\lambda =$.

eg:→ Radioactive decay
• life span of battery

(5) CHI-SQUARE DISTRIBUTION: The CS Dist is a prob that describes the distribution of sum of squares of K random variables.

Degree of freedom

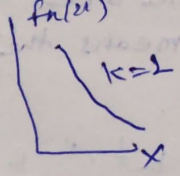
Population



Sample (2)
$S_1(1,2)$
$S_2(2,3)$
$S_3(4,5)$

$K = 2-1 = 1$

Square the sample
$(2-1)^2$
$(3-1)^2$
$(6-5)^2$



(6) Why always the CHI

→ Usecase
① in ML (to compare 2 categorical var)
② In hypothesis testing

(5) f-distribution (fisher Snedecor distribution)

• The f dist is a probability dist that is useful in context of comparing variables of two or more samples

→ It is right skewed takes only +ve values
→ f dist d1 & d2 is dist give by $X = \frac{S_1/d_2}{S_2/d_2}$

S_1 Std df 1 — S_2 Std df 2 → ratio → ratio of variance called F distribution

Usecase: compare two sample Variance using f-test
more than two

2:30 [23 march]

Hypothesis testing :-

Q) Avg Salary of It Employee is 15 lakh.
↓
Statement
↳ claim (Hypothesis)

define

> • Hypothesis is a claim | predictive | generic statement, capabile of being tested by scientific method.

eg: ① Avg age of std in a class of IIT is 23 yrs
② Consumption of Ice-creame sales increase in Summer
③ Student who recieve consulling will show greater increare in performance.
× ④ There is tons of water on Pluto.
× ⑤ Swarglok/(naraan) we will go after we die.
× ⑥ In Patal lok Shesh Naag lives.

Example : Sachin : Avg age of Deloitte emplyee is 45 yrs.

test Alok → Sample 20 people → Sample stats 49 yr
Agrim → 20 pe → Sample stats = 62 yrs
★ Ruchika → 2000 people SrS = 49.
Considers
↳ Highest no of samp ⇒ 2000,
$\bar{x}$ = 49 yrs
μ = 45 yrs

• framing of Hypothesis

Null H (H_0) = H_0 is a statement that is hold true unless we have sufficient statistical evidence to reject it.

Alternate (H_a) :- It is opposite of Null Hypoth

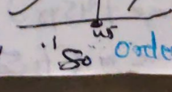
H_0 : Avg Age 45
H_1/H_A : Avg Age Not = 45

Q) Avg Age of people in deloitte is greater than 45 yr
$\mu_{age} > 45$, $\mu_{age} \le 45$ yrs

H_0 - $\mu_{age} > 45$ H_1: μ age ≤ 45 ✗

Why → H_0: $\mu_{age} \le 45$ yrs , H_a: $\mu_{age} > 45$ yrs.

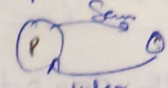
Whichever statement has equal sign that will be your Null Hypothesis



greater than 45
H_0 (45 not include)
So we need to include the central point in order disprove something

Story → Inferential Statistics
↳ we estimate something about population using sample.

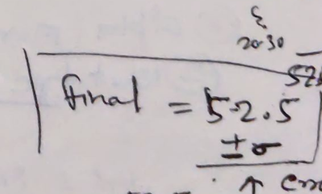
eg: Avg Salary of IT People in India is 15 lakh



we are infering something about popn paramenters using sample statist.
⇓
Q) How much this inference is true??
⇓
• you will pose some doub
• you will avoid saying a specific no → 15 lakh

Senerio
Sharma
↓
Point 90
Estime
① (80, 95) ← less more confi
② Interl → 80-95 - more confidn

> actually HT all about Studying how confident you are with this Particular inference



final = 52.5 ± σ
↑ error

52.5 ± 6
46 - 58

Probability Distribution

→ General Practical use Case
→ Social Trial !

4 mar
+1 if graph

Level of Significance & Confidence !

→ We can not be 100% sure (Confident) to Conclude something for the population using sample.

not confident the chances of errors

eg! If a pharma company wants to known if a vaccine will work or not.

→ 100 samples → 95 → people well
↓ so
Vaccine works 95% of the time.

Out of 100 times, if I conduct a test how many times same type of Conclusion can be made.

95% ⇒ make some threshold

⇓ then what is 5%

⇓

⇒ 5% ⇒ level of significance

II. The percentage of risk involved in testing. alpha → define by business team

⇓

then the confidence is

1 − α → Confidence

100 − 5% = 95%

α = 1, 5, 10

CI = 99%, 95%, 90%

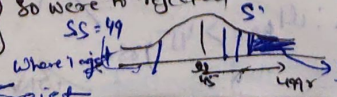
→ How much margin of error you are ready to accept.

Ex: If I say I am ready to accept 5% margin of error then it simply out of 100 experiment 95% of the true

Steps

1. framing of Hypothesis
2. alpha (margin of error)
3. What type of test that you will you → ① S.S ≥ 30 or σ pop → SND → Z test
   Sample Statistic (test statistic)
   ↳ Z, T, F, chi-sq
   ② S.S < 30 or σ pop not → t distribution → t test
   ③ Variance of two sample/or more than two → f-distribution → f test, ANOVA
   ④ Two Categorical → CHI-square distribution → chi-square test
4. Test Critical
   ↳ Z critical
   ↳ t critical
   ↳ f "
   ↳ Chi-sq "

eg Sachin's H_0 age = 45 yrs
H_A age ≠ 45 yrs
SS = 49



we need for reject

→ Two approach →

↳ Convert $\alpha = 5\%$ into Z score

Z Critical

Z table → Z score & corresponding Prob Value will be give

Convert Sample Statistics to Probability Value

P-Value

5 Step 5 Conclution

1 approach

accept

reject

1.64

S.S > 1.64 ⇒ rej 5%

Statistics table = 5% = 1.64

SS > 1.64 ⇒ reject region you reject the H_0

2nd approach

45 5%

↳ Prob Value Correspondy to Sample Score

0.03

P-Value ≤ 0.05

reject the H_0

→ P-Value : The P-Value is the probability Value, calculated from a test statistics

→ use to devide whether to reject a H_0 or not.

*One tail test, two tail test :

$H_0 : \mu_{age} = 45$

$H_A : \mu_{age} \neq 45$

H_0 45

≠ → Alternate Hypothesis two tail test

↳ two tail test

↳ alpha/2

α → whole

Scnesio-1 one tail

$H_0 : \mu_x = \mu_0$

$H_1 : \mu_x < \mu$

accept

Rejection region 5% ← Left tail

μ_0 (One tail test)

Scenario-2 Two Tail

$H_0 : \mu_x = \mu_0$

$H_A : \mu_x \neq \mu_0$

Accept

α/2 α/2 0.025

α = 5%

Scenario 3 1 tail (Right tail)

$H_0 : \mu_x = \mu$

$H_1 : \mu_1 > \mu$

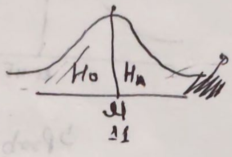
Rejection 5%

24 mar
02:04

(Q) Suppose a child Psychologist says that the avg. time

Statistic test
(1) Z test] → deal with Avg Var most
(2) T test
(3) CHI-Square → Categorical data
(4) ANOVA → Variance
(5) F-test

→ Step 1 : $H_0 : \mu \leq 11$
$H_A : \mu > 11$



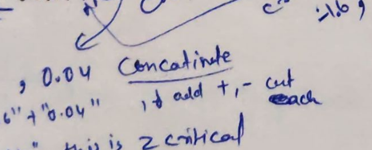
Rejection region (on side of dist as its 1 tail test)

Step 2 - L of sig = α = 5% = 0.05

Step 3 → S.S ≥ 30 & σ pop give
↓ Z-test
→ $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ ⇒ $\frac{11.5 - 11}{2.3 / \sqrt{100}}$ = 2.17 (+ve)

Step 4
↳ (1) Z critical (test critical) | (2) P-value

(1) α = 0.05
↳ Z table → using α value
Z ... .04
-1.6 ← 0.0505
Check α is = 0.0505 ↓ corresponding value -1.6, 0.04
-1.6, 0.04 Concatinate (+ add +, - cut each)
⇒ "1.6" + "0.04"
⇒ "1.64" this is Z critical
-1.64
↳ so Z score is +ve in nature so we also do 2.17 is in +ve side so comp & our α
-1.64 → +ve → because of Symmetric in nature



Zcritical = 1.64

(2) α = 0.05
Z statics = 2.17 → Z table → +ve table
Z statics ⇌ p-value
2.1, 0.07 → 0.9850 total prob
p-vale ⇒ 0.9850 corresp 2.17
1 -
0.9850
p-val 1 - 0.9850
⇒ 0.015

So
Step 6 ⇒ 2.17 > 1.64 → Conclusion!
reject the H_0 : $H_0 \neq 0$ ↳ The working mother spend > 11 mins;

3:03

(Q2) Sachin

3:15 $x \approx N(4, 2)$

Q How may

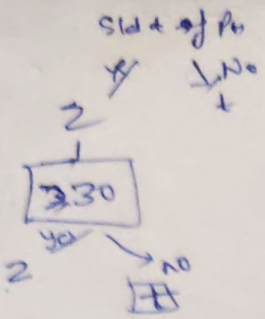
Q 3:22

Q what percentage

Q 3:28

Q 3:31

T test vs Z test



Types of Errors

→ Type-I error — is the rejection of the null hypothesis when it is actually true.

→ Type-II error → fail to reject the null hypothesis that is actually false.

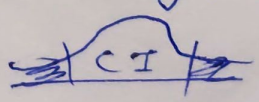
⇒ Confidence Interval & Margin of Error

estimate :-

Point — Single Value

Interval — Point ± error

800-1000 → more accurate representation of reality → C.I → What value the sample statistics with ~~table~~ take can be known through C-I.



⇒

→ CHI SQUARE TEST

(8) 12% of people are left handed. To Verify this theory you took a sample of 75 students, 11 are left handed.
With 5% level of significance, Conduct the test.

(1) H_0 : 11 = 12%.
H_1 11 ≠ 12%.

	Observed	Expect
→ left	11	9
→ Right	64	66

⇒ this diff by Chance or by Statistics

(2) $\alpha = 5\%$.

given : left, Rgth
observed, Expect
: Std dev not given
↳ then use CHI test

(3) Chi Sq test
↳ CHI square Statistics

$$\chi^2_{Stat} = \sum \frac{(O-E)^2}{E}$$

$$\chi^2_{Stat} = \frac{(11-9)^2}{9} + \frac{(64-66)^2}{66} \Rightarrow \frac{2^2}{9} + \frac{2^2}{66} = 0.505$$

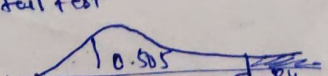
$\Rightarrow \chi^2_{Critical}$ for $\alpha = 0.05$

2 categorial Varible
↳ So dof = 2 − 1 = 1 (no of categories = 2)

↳ χ^2 table ⇒ dof 1, $\alpha = 0.05$

$\chi^2_{critical} = 3.84$

Step 5 — 1 tail test



if $\chi^2_{stat} > \chi^2_{critical}$ → Reject H_0
$0.505 < 3.84$ → fail to reject H_0

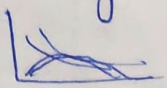
Conclusion
12% of people are left handed with 95% confidence.

3rd Math
Anova

f-test & Implementation

→ f-test:

- follow f-distributions
- Comparing variances
- $x = \frac{s_1/d_1}{s_2/d_2}$ where
- right skewed
- Non-negative



- its depends on DOF

(8) The following data is about the ~~number~~ of bulbs produced daily by two workers A and B.

A	B
40	39
30	38
38	41
41	33
38	32
[illegible]	39
[illegible]	40
35	34

$\alpha = 0.05$

can we consider based on data that workers B is more stable and efficient?

- Why not mean can be used for test?

→ mean is same for the both the sample so we will compare variance.