Statistics

Through some tools acc. to problem statement

Defn: Statistics is the science of collecting, organising and analysing the data

↓

Decision Making process

Data: "facts or pieces of information"

Eg: Heights of student in the class

{175cm, 180cm, 190cm - - }

[Data, it can be captured record and it can be measured.

IQ of the student

{85, 90, 100, - - }

conclusion

e.g. ATM (A) ——— 40km ——— (B) ATM

To open or not?

Does it can create an impact

Analysis parameters

Types of statistics:

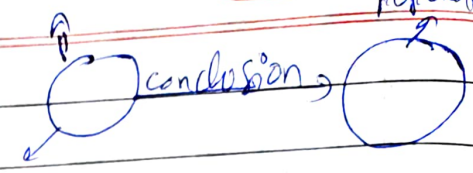
① Descriptive stats

Defn: It consists of organizing and summarizing the data

② Inferential stats

Defn: It consists of using data you have measured to form conclusion.

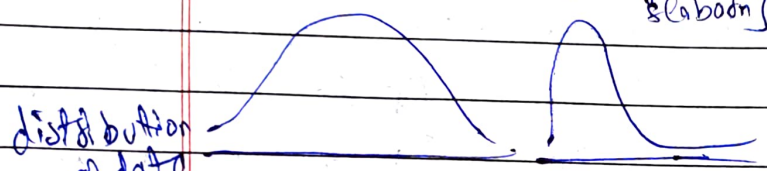
use sample to make conclusion on population data



(1) Measure of central tendency [Mean, Median, Mode]

(2) Measure of Dispersion [variance, std] ← Standard deviation

(3) Histograms, Bar chart, pie charts, [matplotlib, seaborn]



(1) z-test
(2) t-test
hypothesis testing, p value, Significance.

[conclusions]

→ All the industries uses these techniques for conclusions

Eg: let say there are 50 students in a math class in the university. we have collected the height of the students in the class [data]

[175cm, 180cm, 160cm, 140cm, 130cm, 140cm, 140cm, ___]

Descriptive Question:

"What is the average height of the students in the class" [(175+180+160+140+130+140+140)/50 = Ans]
↓ Mean

"what is the common height of the students" = 140cm

Inferential Question:

Sample data

"Are the Average height of the Students in the classroom Similar to what you expect in the entire college"
↓ Population

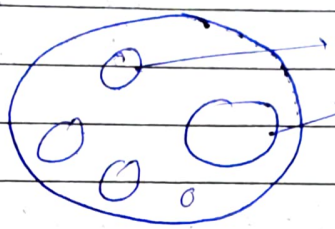
work on Sample data to come up with a soln for the Population data.

Date
Page

Sample Data And population DATA :

e.g Exit poll

party A 56%
party B 46%



ASK people to whom they vote.

Acc. to that collection of data Make conclusions.

Types of DATA :

DRC	DC	LSI	BUI	FWI	classes	Region
3.4	7.6	0.3			not fire	0
4.1	7.6	1			not fire	0
2.5	7.1	0.3			not fire	0
1.3	6.9	0			not fire	0
3	14.2	1.2			not fire	0
5.8	22.2	3.1			fire	0
9.9	30.5	8.4			fire	0
12.1	38.3	5.6			fire	0

Average value

=> 1st step in EDA

DATA

- numerical
  - Quantitative
    - Discrete
      - whole no.
      - [specific Range]
      - eg no. of bank acc
      - no. of children in a family
    - continuous
      - Any value
      - eg: weight, height, temp, speed
- categorical DATA
  - Qualitative
    - nominal
      - eg: gender M, F
      - blood grp
      - color
    - Ordinal
      - Rank.
      - eg: Good ①
      - ② Bad
      - ③ satisfactory

Assignment :

(1) what kind of variable Marital Status is ?
(2) what kind of variable Nile River length is ?
(3) what " " is Movie duration ?

Scale of Measurement of Data :

(1) Nominal scale Data
(2) Ordinal scale Data
(3) Interval scale Data
(4) Ratio scale Data

(1) Nominal scale Data :

(1) Qualitative/categorical variable
(2) e.g. Gender, Colors, Labels
(3) Order does not matter

e.g. Survey Log

RED → 5 → 50%
Blue → 3 → 30%
Yellow → 2 → 20%

(2) ~~Interval scale Data~~
(3) Ordinal scale Data :

(1) Ranking and order matter
(2) Difference cannot be measured

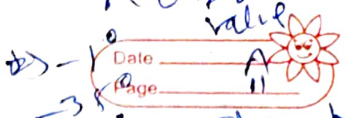
e.g. Qualification

		Marks	
PHD	1st	100	1st
BE	3rd	90	2nd
Masters	2nd	85	4th
BCom	5th	70	3rd
BSC	4th	30	5th

If Ranking is important

We can convert this into this.

i.e 0 is not a starting value



It can have any value

(3) Interval Scale data :

(a) The rank and order matters
(b) Difference can be measured (excluding ratio)
(c) Does not have "0" starting value

$\frac{60}{30} = \frac{2}{1}$ eg: temperature (°F)

30°f
60°f
80°f
90°f

60-30 = 30
-10°f
-20°f
-30°f
-40°f
-50°f

90-30 = 60

Whatever you are feeling on 30°
That does not mean you will feel
double or 60° [not considering Ratio]

(4) Ratio Scale Data :

Grades

0, 100, 90, 60, 35, 45, 50

(a) order and Rank matter
(b) Difference and ratio are measurable
(c) It does have a "0" starting point

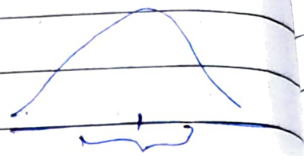
100 90 60 50

100 : 50 => 2 : 1

e.g:
(a) length of different rivers in the world ?
(b) Hospital stars ?
(c) ICU measurement ?

① Measure of central tendency :

① Mean
② Median { EDA and
③ Mode feature engineering }



① Mean

Population (N) Sample (n)

X : {1, 2, 2, 3, 3, 4, 5, 5, 6} → n = 10

Population Mean (μ) = $\sum_{i=1}^{N} \frac{x_i}{N}$ Sample Mean ($\bar{x}$) = $\sum_{i=1}^{n} \frac{x_i}{n}$

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

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

② Median ⇒ central element of the distribution

4, 5, 2, 3, 2, 1

Sort → 1, 2, 2, 3, 4, 5

Median even counts

1, 2, 2, 3, 4, 5

odd count

{1, 2, 2, 3, 4, 5, 7}

$\frac{2+3}{2} = 2.5$

Median = 2.5

Median = 3

Why Median? [outliers]

{1, 2, 3, 4, 5}

$S = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$ [Mean]

{1, 2, 3, 4, 5, 100} ← outlier

$S = \frac{1+2+3+4+5+100}{6} = \frac{115}{6} \simeq 18$ [Mean]

↳ this a very big shift in case of outlier

{1, 2, 3, 4, 5}

Median = 3

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

Median = 3.5

↳ slightly denoted

So use Median in case of outliers ✓

[This helps to use median in case of outliers]

(3) Mode: frequency Maximum

{2, 1, 1, 1, 4, 5, 7, 8, 9, 10}

Mode = 1

EDA and feature engineering

Type of Flower

Age

[To avoid loss of data]

Lilly	16
Rose	3
Rose	5

→ Rose — small flower

→ mean or median outliers

Date
Page

(2) Measure of Dispersion

(1) variance
(2) standard Deviation

(1) variance

Population variance

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

x_i = Data points
μ = population mean
N = population size

Sample variance

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

Bessel's correction

Why does the sample variance have $N-1$ in the Denominator?
⇒ The reason we use $n-1$ rather than n is so that the sample variance will be what is called an unbiased estimator of the population variance.

x_i = Data points
$\bar{x}$ → Sample Mean
n → Sample size

ex: {1, 2, 3, 4, 5} ⇒ Sample

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

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

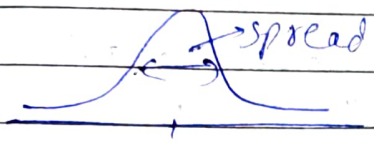
x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
		10

$$s^2 = \frac{10}{4} = 2.5$$

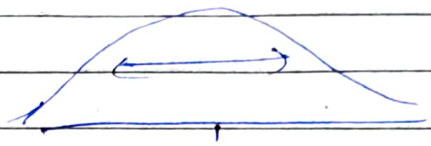
Date
Page

variance: Spread of the data.

$s^2 = 2.5$



$s^2 = 6.5$: ↑↑



More spreaded.

② **standard Deviation**

population std

$\sigma = \sqrt{\text{variance}}$

sample std

$s = \sqrt{\text{sample variance}}$

$s^2 = 2.5$

$\sqrt{s^2}$ = sample std

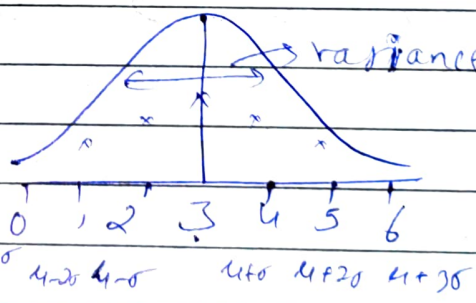
Consider

{1, 2, 3, 4, 5}

↳ $\mu = 3$ ⇒ mean

↳ $\sigma = 1$

↳ std



⇒ 1 std to the left and right

Date

③ Random variables &

$\begin{cases} x+5=7 \\ y+x=10 \end{cases}$ → here specific value

$x=2$

$y+2=10$

$y=8$

But in random variable

↓

Random variable is a process of mapping the output of a random process or experiment to a number.

Eg: Tossing a coin
Rolling a dice
Measure the temperature for the next day.

$X = \begin{cases} 0 & \text{if } H \\ 1 & \text{if } T \end{cases}$ ← Capital letter

↓ converting a process output or outcome into a specific number

Quantifying a Random Process.

$Y =$ { Sum of the rolling of dice 7 times

{4, 5, 6, 1, 2, 2} = 20

$P(Y \geq 15)$, $P(H)$

④ Sets

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

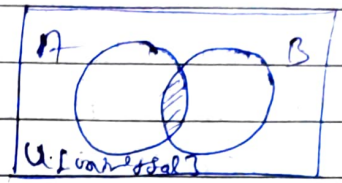
B = {3, 4, 5, 6, 7}

operations

Venn diagram

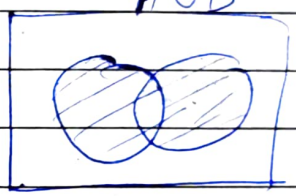
① Intersection :

$A \cap B = \{3, 4, 5, 6, 7\}$



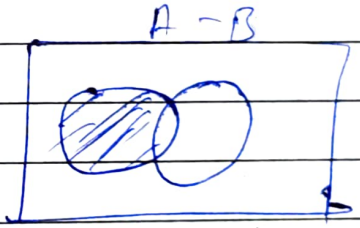
② Union :

$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$



③ Difference :

$A - B$: {1, 2, 8}



④ Subset :

A → B ⇒ FALSE

B → A ⇒ TRUE

⑤ Superset :

A → B ⇒ true

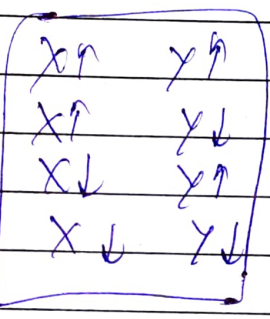
B → A ⇒ FALSE

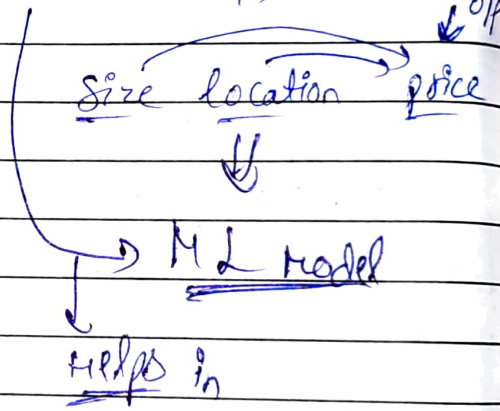
① Covariance And correlation

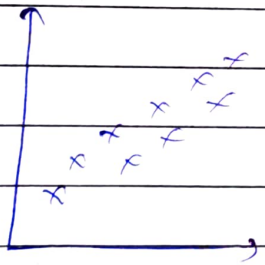
Covariance indicates the relationship of two variables whenever one variable changes.
If an increase in one variable results in an increase in the other variable, both variables are said to have a positive covariance.
Decrease in one variable also cause a decrease in the other.

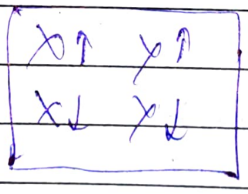
X	Y
2	3
4	5
6	7
8	9

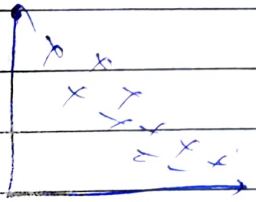
(Relationship b/w x and y)

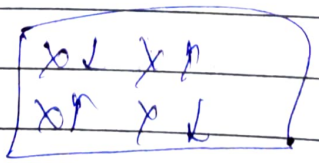












Diff. b/w Covariance and variance

⇒ It is the same thing
⇒ x relationship itself.

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

$$\text{Cov}(X,X)$$

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

X ↑	Y ↑
X ↓	Y ↓

+ve cov.

var of x is nothing but its own relationship

X ↑	X ↓
X ↓	Y ↑

-ve cov.

spread of the data. ⇐ $\text{Cov}(X,X) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{n-1}$

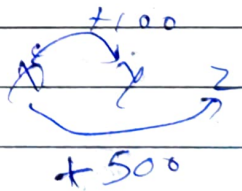
$$\text{Cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

X	Y
2	3
4	5
6	7

$\bar{x}=4 \quad \bar{y}=5$

$$= \frac{(2-4)(3-5)+(4-4)(5-5)+(6-4)(7-5)}{2}$$

$$\Rightarrow \frac{4+0+4}{2} = 4 \text{ +ve}$$



X & Y are having a positive variance.

Advantages

(1) Relationship b/w x & y +ve or -ve value

Disadvantage

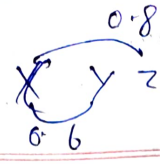
(1) Covariance does not specific limit value

↓

It can be any +ve or -ve value

To avoid this disadvantage we use another kind of correlation technique

x-z is more highly correlated than xy.



value Ranging between

Page

(2) Pearson correlation coefficient [-1 to +1]

The pearson coefficient is a type of correlation coefficient that represents the relationship b/w two variables that are measured on the same interval or ratio scale. The pearson coefficient is a measure of the strength of the association b/w two continuous variables

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = [-1 \text{ to } +1]$$

(1) The more the values towards +1 the more +ve correlated it is (x,y)

(2) The more the values towards -1 the more -ve correlated it is (x,y).

use case

Dataset 1000 features (DL Models)

Independent features

Dependent feature

Size of the House	no of Rooms	Locations	no of people staying	Price
imp	imp	imp	not imp	

Haunted → -ve corr → not imp

Feature selection & near to 0 ⇒ Drop no of people staying feature

Date
Page

(3) Spearman Rank correlation :

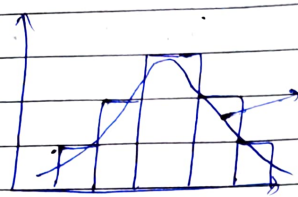
$$r_s = \frac{cov(R(x), R(y))}{\sigma_{R(x)} * \sigma_{R(y)}}$$

Here we don't focus on values instead we compute the rank of x & y. and use them to calculate cov.

x	y	R(x)	R(y)
1	2	8	5
3	4	4	4
5	6	3	3
7	8	2	1
0	7	6	2
8	1	1	6

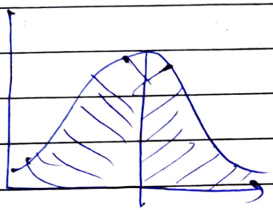
```
import seaborn as sns

df = sns.load_dataset('taxis')
df.head()

import numpy as np

np.cov

df.corr()          | df.corr(method = 'spearman')

df.corr(method = 'pearson')

df = sns.load_dataset('penguins')

df.corr()
```

Skewness :



→ Normal / Gaussian Distribution

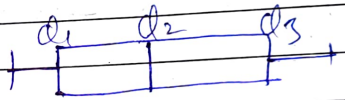
Symmetrical Distribution



① No skewness

[In Symmetrical Distribution no skewness]

Box plot



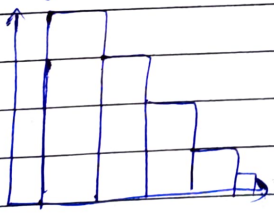
$Q_3 - Q_2 \simeq Q_2 - Q_1$

⇒ Normal / Gaussian Dist.

⇒ The Mean, Median and Mode all are perfectly at the center.

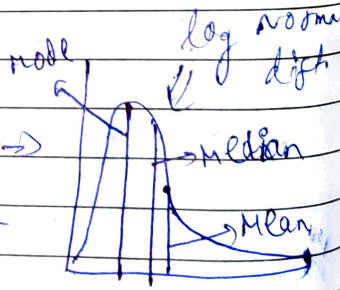
⇒ Mean = Median = Mode

② Right skewed :



⇒ positive skewed

⇒ Tend towards Right



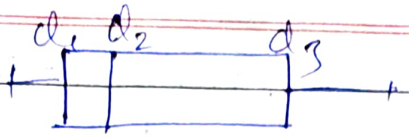
Relationship b/w Mean, Median, Mode → Mean > Median > Mode

Mode [Peak]

Median not much effected

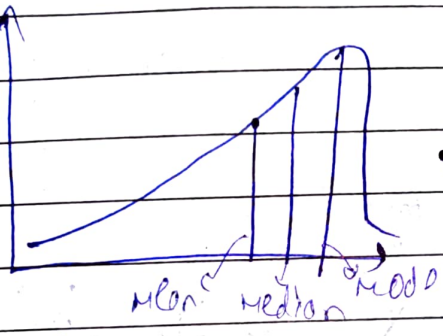
Mean get effected by larger

Box plot



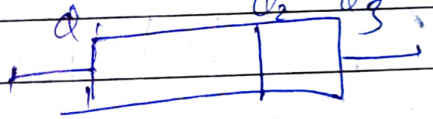
$Q_3 - Q_2 \geq Q_2 - Q_1$

(3) left skewed distribution :



=> Negative skewed

Relationship : Mean ≤ Median ≤ Mode



$Q_2 - Q_1 \geq Q_3 - Q_2$

(4) what is a Relationship b/w Mean, Median & Mode

Histogram: [frequency] ⇒ Count

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

⇒ Buckets

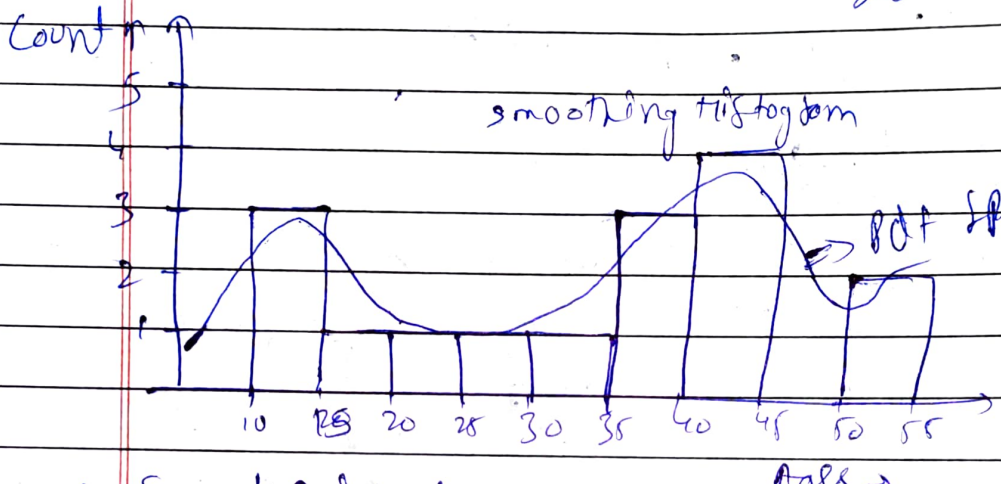
no. of Bins = 10

bin size = 5 = $\frac{50-0}{10}$ = 5

⇒ How many elements are present in a Bucket

$\frac{50}{20}$ = 2.5 → bin size

∴ {20 no. of bins}

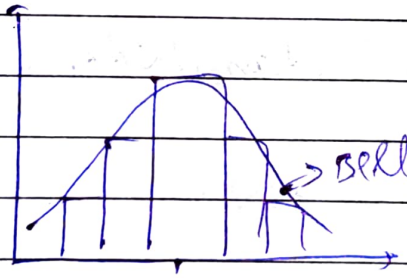


pdf {probability distribution function}

{prob density fun}

⇒ Smoothening histogram

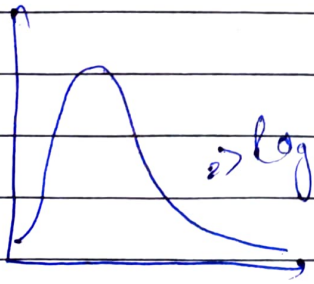
{Kernel density estimation}.



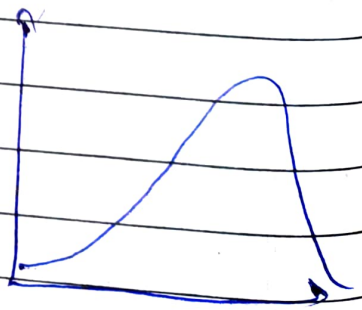
Bell wave

Normal / Gaussian Distribution

{When you know about data dist you can make a lot of assumptions about data}



⇒ log normal dist.



⇒ When we smoothen it we can get different distribution.

Prob. Density fn is a part of Prob. Distribution function

Date
Page

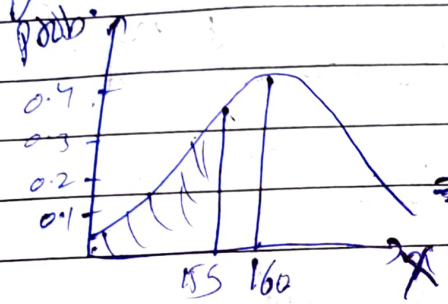
① Probability Distribution function / Density function:

② Probability Density function (PDF) ↔ Distribution of the continuous data

eg: - Height of students

[0 to 1]

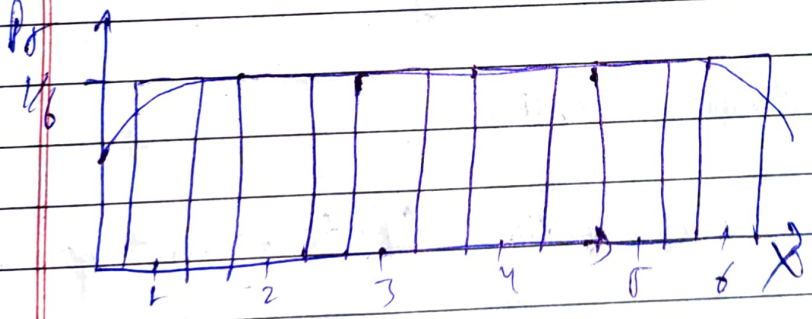
Pr(X < 155) = Area under the curve 0.3, 0.4, 0.5



⇒ Kernel Density estimator

③ Probability Mass function (PMF) of Distribution of a Discrete Random variable

eg: Rolling a Dice {1, 2, 3, 4, 5, 6}

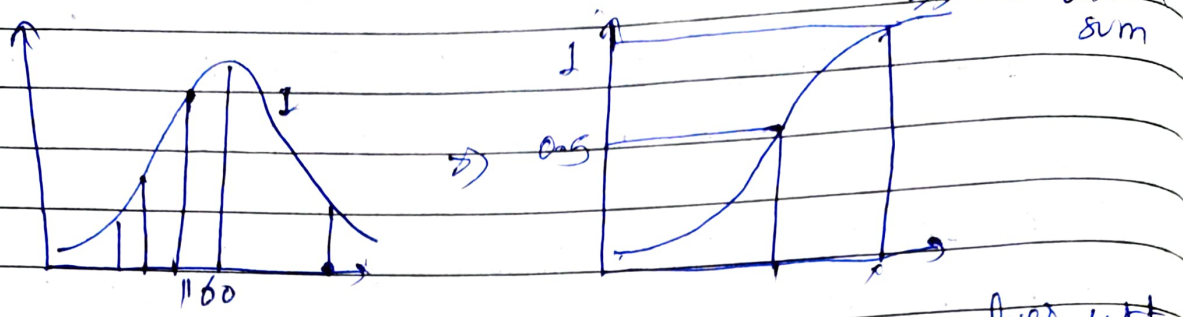


$Pr(X<4) = Pr(X=1) + Pr(X=2) + Pr(X=3) + Pr(X=4)$

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

$\Rightarrow \frac{4}{6} = \frac{2}{3}$

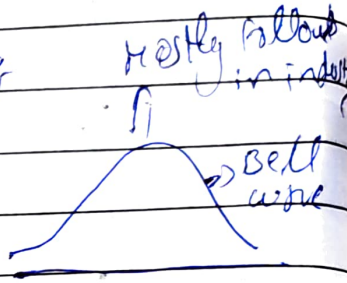
(5) Cumulative Distribution function (CDF) :



⇒ we are doing a cumulative sum of all the values upto the data index we have

Types of the probability Distribution :

(1) Normal / Gaussian Distribution (PDF)



(2) Bernoulli Distribution (PMF) → outcomes are 2 [success or failure]

(3) Uniform Distribution

(4) Log Normal Distribution (PDF)

(5) poisson Distribution (PMF)

(6) power law Distribution (PDF) → 80-20% Rule

(7) Binomial Distribution (PMF) ⇒ discrete values

⇒ some dist are of various types, it may belong to a PDF or PMF

Date
Page

probability Density function And probability Mass function.

cumulative Distribution function (CDF)

using the PDF & PMF we practically create the visualization graphs

[what is the example used in the actual dist.]

PMF & CDF

- Discrete Random variable

PMF

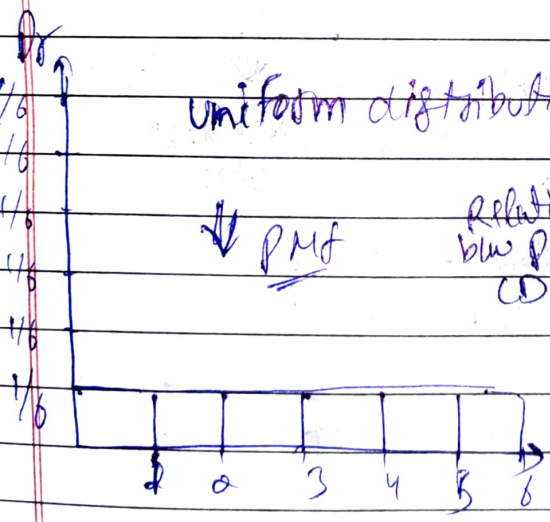
CDF

eg: Rolling a dice

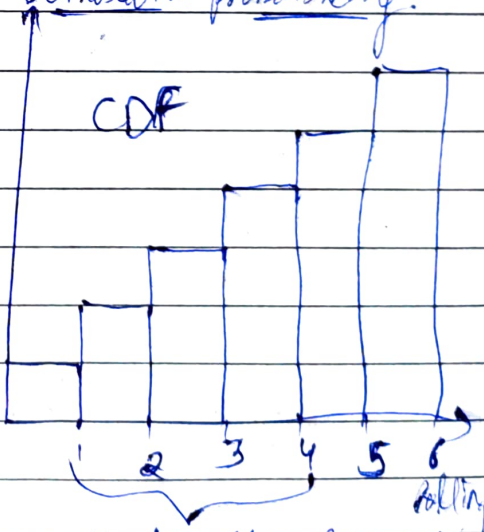
[cumulative basically means combining]

$\{1, 2, 3, 4, 5, 6\}$ $P(x) = 1/6$

Fair dice $P(2) = 1/6$



Relationship b/w PMF & CDF



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

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

→ pretty simple

→ we try to add more events and try to cumulatively add more events

$P(X \leq 4) = P(X=1) + P(X=2) + P(X=3) + P(X=4)$

we have to add here

$\Rightarrow 1/6 + 1/6 + 1/6 + 1/6 = 2/3$

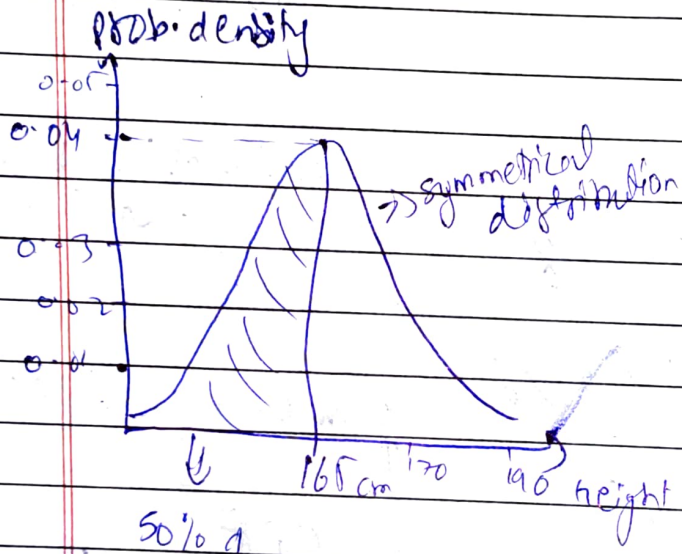
(1) $P(X \leq 6) = P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5) + P(X=6)$

↓

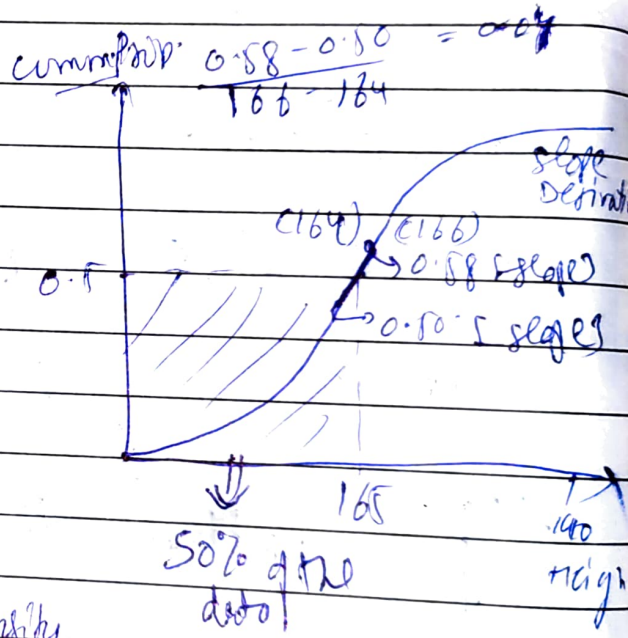
Cumulative Prob. $= 1$

(2) Probability Density function (PDF)

① ^continuous Random variable



50% of the entire distribution



50% of the data

Probability density
↳ Gradient descent of CDF

~~Probability density at 185cm = 0.04~~

Date
Page

Bernoulli Distribution [Binary outcomes]

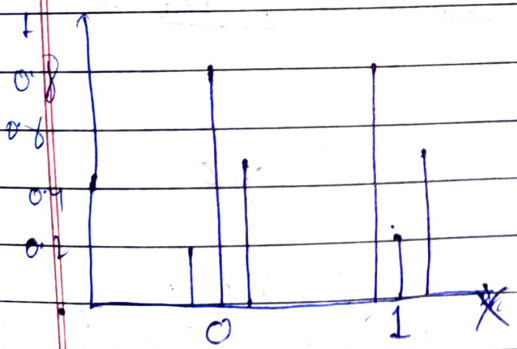
In probability theory and statistics, the Bernoulli distribution, named after the Swiss mathematician Jacob Bernoulli, is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability q = 1-p.

Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes-no question.

Such questions lead to outcomes that are boolean-valued: a single bit whose value is success/yes/true/one with probability p and failure/no/false/zero with probability q.

Bernoulli distribution

Prob. mass function.



Three examples of Bernoulli distr.

$P(X=0) = 0.2$ and $P(X=1) = 0.8$

$P(X=0) = 0.8$ and $P(X=1) = 0.2$

$P(X=0) = 0.5$ and $P(X=1) = 0.5$

Outcomes are binary

e.g. Tossing a fair coin

H ↓ {0, 1} ↓ T

$P(T) = 0.5 = p$

$P(H) = 1 - 0.5 = 1 - p = q$

e.g. whether the person boss/fail.

↓ PMF

PMF vs Pdf

↓ Discrete Random variable

↓ Continuous random values

PMF vs PDF

↓ ↓

Discrete random variable | Continuous random variable

Mathematically - PMF fun^n eqn

PMF $K = 0$ or 1

$$P(X=K) = p^K (1-p)^{1-K} \Rightarrow \text{PMF}$$

simplified eqn of PMF

$$\text{PMF} = \begin{cases} q = 1-p & \text{if } K=0 \\ p & \text{if } K=1 \end{cases}$$

(1) $P(X=1) \Rightarrow p^1 (1-p)^0$

$P(X=1) \Rightarrow p$

(2) $P(X=0) \Rightarrow p^0 (1-p)^{1-0}$

$= (1-p) = q$

(3) Mean, variance And standard Deviation

Mean

$K = 1$ or 0

$P(K=1) = 0.6 = p$

$P(K=0) = 1-0.6 = q = 0.4$

$$E(K) = \sum_{i=1}^{K} K \cdot P(K)$$

Expected value q^n

$E(K) = 1 \times 0.6 + 0 \times 0.4$

$\Rightarrow 0.6 = p$

Median of Bernoulli Distribution :

$$\text{Median} = \begin{cases} 0 & \text{if } p < \frac{1}{2} \\ [0,1] & \text{if } p = \frac{1}{2} \\ 1 & \text{if } p > \frac{1}{2} \end{cases}$$

Variance & Std :

$p = 0.5 \quad q = 0.5$

$$\text{variance} = p(1-p) = pq$$

$$\text{Std} = \sqrt{pq}$$