STATISTICS FOR THE DATA SCIENCE

Part - 2

- HISTOGRAME AND SKEWNESS
- COVARIANCE AND CORRELATION
- ADVANTAGE VS DISADVANTAGE OF VARIANCE
- PEARSON CORRELATION COFFICIENT
- PDF / PMF / CDF
- BERNOULLY DISTRIBUTION
- BINOMIAL DISTRIBUTION

#Value_freeContent

@Krishan kumar

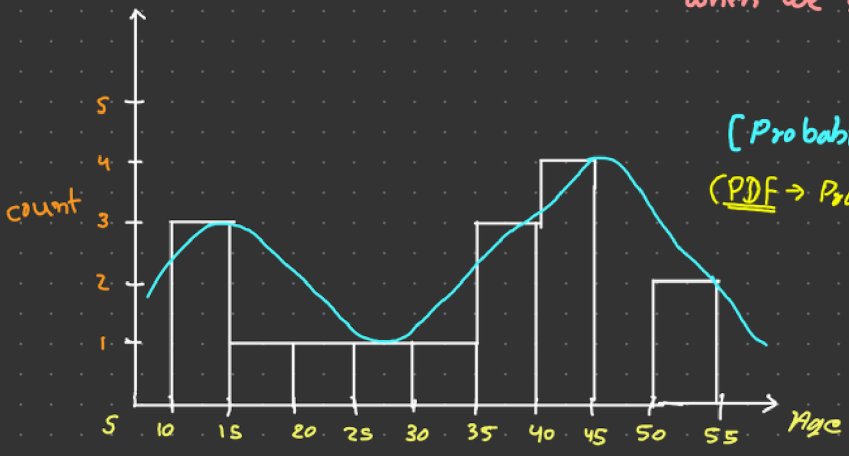
Histograms And Skewness

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

0 to 50 value

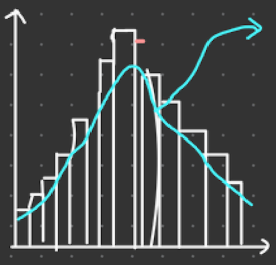
↳ $\frac{50}{10}$ = (5) bin size (No. of bins → 10)

↳ $\frac{50}{20}$ = (2.5) bin size (No. of bins → 20)



when we smoothen our histogram
↓
Technique
↓
[Probability distribution fun.]
(PDF → Probability density function)

Skewness



Normal / Gaussian destribution
↓
These types of destribution is called → Symmetrical destribution.

* Symmetrical Data

No Skewness
(No curve)
Equal from both side

→ Boxplot :- To finding outliers

$$Q_3 - Q_2 \cong Q_2 - Q_1$$

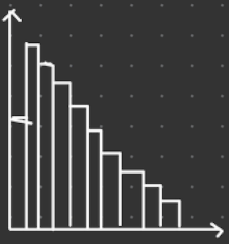
The mean, median, mode are all perfectly at the centre.

Mean = Median = Mode

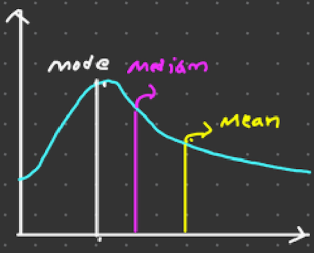
There are no skewness in the symmetrical distribution data and just because there are no any maxi value and minimam value in Right OR Left corner of data.

② Right Skewed data

Means longitute in Right side (Mejority data on right side present)



⇒ Positive / Right skewed → Distribution in PDF →



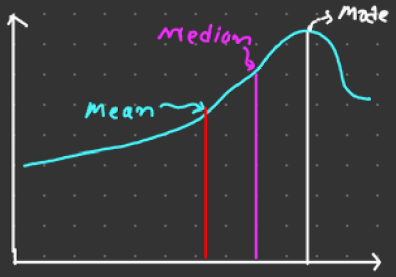
Relation btw mean, median ⇒ mode

Mean > Median > Mode

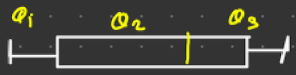
* Box plot for Right Skewed data

$\Rightarrow \quad Q_3 - Q_2 \geq Q_2 - Q_1$

* Left Skewed data



* Box. plot



$\Rightarrow \quad Q_2 - Q_1 \geq Q_3 - Q_2$

Relation

Mean < Median < Mode

Covariance and Correlation

[To rectify Relation btw X and Y]

X	Y
2	3
4	5
6	7
8	9

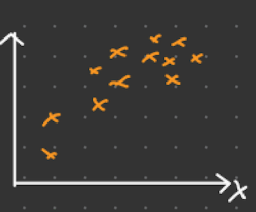
In this scenario →

X ↑	Y ↑
X ↓	Y ↑
X ↑	Y ↓
X ↓	Y ↓

↳ For example →

$x\uparrow$	$y\uparrow$
$x\downarrow$	$y\downarrow$

→ Plotted like →



Covariance →

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

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

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

$\Rightarrow Cov(x,x) = Var(x)$

x_i → Data point of x

$\bar{x}$ → Sample mean of x

y_i → Data point of y

$\bar{y}$ → Sample mean of y

Q What are the difference between covariance and variance?

Ans Variance of x $var(x)$ is nothing but $cov(x,x)$, whenever we talk about variance(x) it is specifically told about spread of the data

$\Rightarrow$ $Cov(x,y)$

$x\uparrow$	$y\uparrow$
$x\downarrow$	$y\downarrow$

→ Then the output will be +ve covariance

$x\uparrow$	$y\downarrow$
$x\downarrow$	$y\uparrow$

→ Then output will be -ve covariance

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

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

$$\Rightarrow \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)}{3-1}$$

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

positive covariance = +ve value

So, X and y are having a positive covariance.

Advantage of covariance

* Relation between X and y +ve OR -ve value of covariance

* Disadvantage of covariance

* We don't conclude weather which value is covariance with any other value Because ↓

Covariance does not have a specific limit value.

ex 

[there is no limit of Difference]

Note: To fix the disadvantage of covariance we use :-

↳ Pearson correlation coefficient

Pearson correlation cofficient

limition → +1 to -1

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma x, x\ \sigma y}$$

* The more value towards +1 the more +ve correlated it is. (x,y)

* The more value towards -1 the more -ve correlated it is (x,y)

Spearman Rank correlation

Rang → -1 to +1

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

Rank → Highest element in the values makes → 1st Rank

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

* Feature selection

+ve	+ve	+ve	≈ 0	-ve
Size of House	No. of Rooms	Location	No. of people staying	Hounte ⇒ Price ↑↑

PDF / PMF

* PDF = Probability Density Function
* PMF = Probability mass function.

↳ PDF is denote a distribution of data. That helps to understand that how our data will distributed.

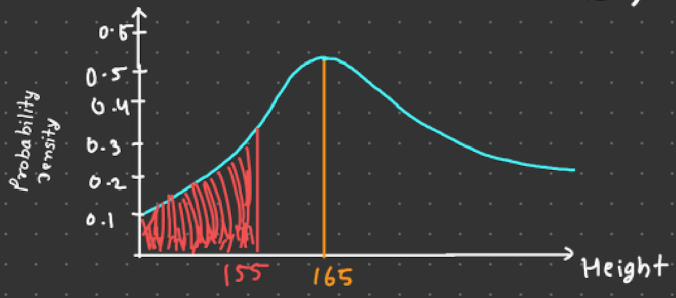
Types of Probability Distribution function

1. Pdf
2. Pmf
3. cdf

Probability density function (Pdf)

↳ Continous Random variables, ex → Height of students in classroom

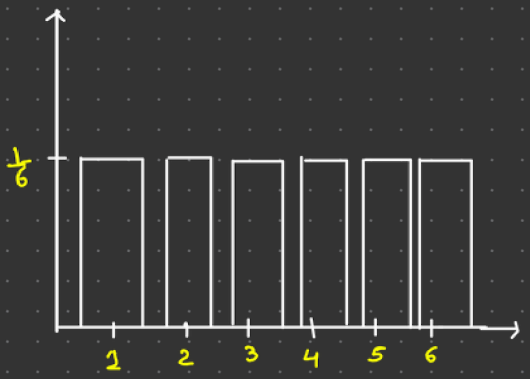
↳ $P(X \leq 155)$



Probability mass function (Pmf)

↳ Variable → Discrete Random variable

ex → Rolling a dice { 1,2,3,4,5,6 }



$P(X \leq 4)$

$\Downarrow$

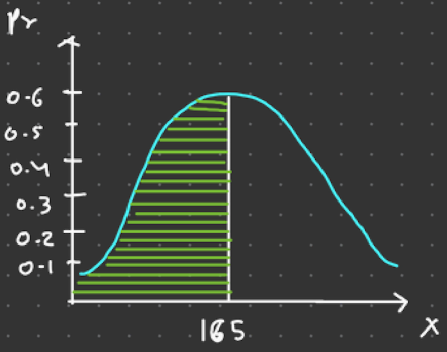
$P(x=1) + P(x=2) + P(x=3) + P(x=4)$

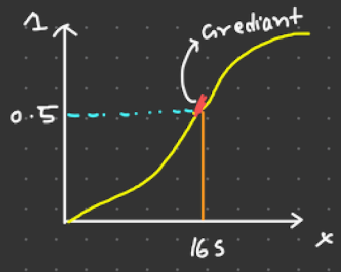
$\Rightarrow \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \rightarrow \frac{4}{6} = \frac{2}{3}$

$P(1) \rightarrow 1/6$

$P(2) \rightarrow 1/6$

Cumulative Distribution fun. (cdf)





PDF Vs PMF Vs CDF

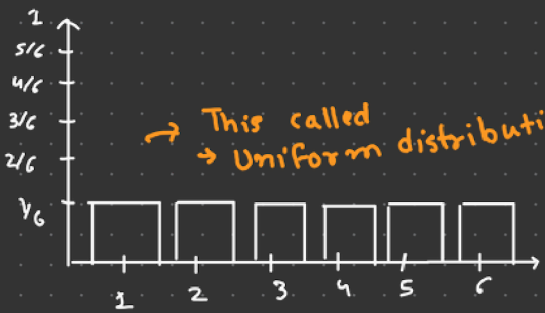
Relation and difference

① Pmf

* Descrete Random variable.

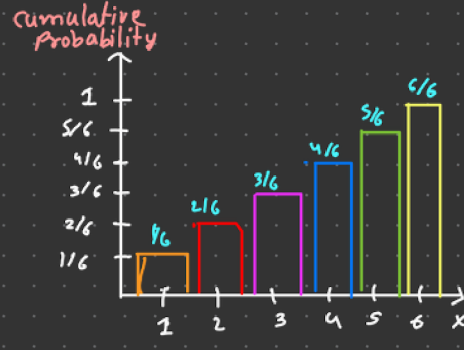
Ex → Rolling a dice ⇒ [1,2,3,4,5,6]

P(1) → 1/6
⋮
P(6) → 1/6



→ This called → Uniform distribution

CDF



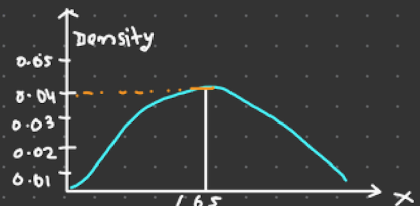
Ex → cdf for P(X ≤ 2)

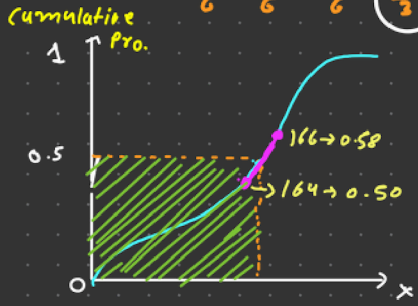
↳ cdf = P(X ≤ 2) = P(X=1) + P(X=2)

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

② PDF

↳ Distribution of continous Random variable





→ slope ⇒ $\frac{0.58 - 0.5}{166 - 164} \rightarrow \frac{0.08}{2}$

⇒ 0.04

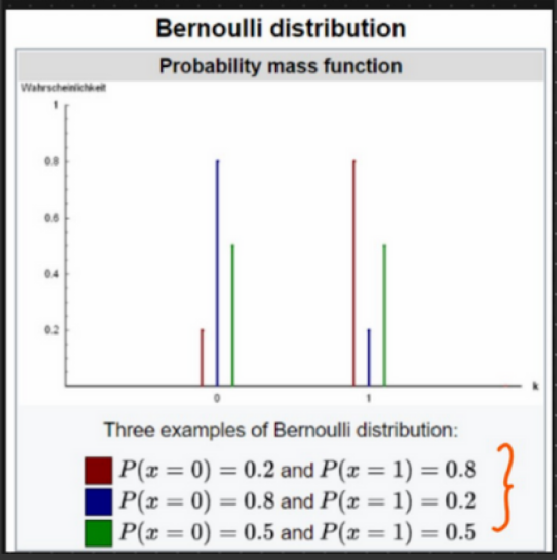
Types of Probability Distribution

- Normal / Gaussian distribution (Pdf)
- Bernoully distribution (Pmf)
- Uniform distribution (Pmf)
- Poisson distribution (Pmf)
- Log normal distribution (Pdf)
- Binomial distribution (Pmf)

Bernoully distribution

In 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 though us a model for the set of possible outcomes of any single expriement that asks a yes-no question.

Pmf → To see the distribution



Points

① Discrete random variable (PmF)

② Out comes are binary

ex: Tossing a coin (H,T)

P(H) → 0.5 → p (Pro. of Succuss)

P(T) → 0.5 → 1-p ⇒ q

(q = Pr. of failure)

③ When there person will fail / Pass

↳ P(Pass) → 0.7 ⇒ p → Succus value

P(fail) → 1-0.7 ⇒ 0.3 = q → Failare value

* **PMF** $= p^k * (1-p)^{1-k}$, $k \in \{0,1\}$

if $k = 1$

Pmf → $P(k=1) = p^1 * (1-p)^{1-1}$

$= p$

if $k = 0$

$Pmf = P(k=0) = p^0 * (1-p)^{1-0}$

$= (1-p) = q$

$\Rightarrow q$

Simplified

$$Pmf = \begin{cases} q = 1-p & , \text{if } k = 0 \\ p & , \text{if } k = 1 \end{cases}$$

⇒ Mean of bernaully distribution,

$$\langle \varepsilon(k) = \sum_{j=1}^{k} k.P(k)$$

, $P(k=1) = 0.6 \Rightarrow p$

, $P(k=0) = 0.4 \rightarrow 1-p \Rightarrow q$

→ $[0 \times 0.4 + 1 \times 0.6] \Rightarrow 0.6 = p$

Whenever we abstract mean in bernaully distribution, then we get P-Value.

⇒ Median of Bernoully dis,

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

⇒ Variance of bernoully dis.,

$Var = p(1-p)$

$\boxed{var = pq}$

⇒ Std of bernoully dis,

$\boxed{Sdtd = \sqrt{pq}}$

Binomial Distribution

In Statistic the binomial dis. with parameter n and P is the discrete Pro. Dis of the number of success in a syuence of n independent experiments. Each asking Yes or No. question. and each its own boolen-value outcome. success (with Pro. P) or failare ($q = 1-P$).

A single success / failare experiement is also called a bernoulli trial OR Bernoully experiment and a sequence of outcome is called a bernoully process.

For a single trial ex⇒ $n=1$ the binomial dis. is a bernoully dis. the binomial dis. is the basis for the popular binomial test of statistial singnificant.

$B(n,P)$

Notation: $B(n,P)$

- Every experiment outcome is binary
- This experiment is Perfoms for n trial
- Group of bernoully dis → Binomial Distribution
- Discrete random variable.

Parametters

- $n \in \{0,1,2,\dots\}$ → No. of trial
- $P \in [0,1]$ → Success Pro. of each trial
- $q = 1-P$
- Ex⇒ Tossing a coin 10 time

Support:- $k \in \{0,1,2,3,\dots n\}$
→ Number of Success

* pmf :-

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

for $k = 0, 1, 2, 3, \ldots, n$ where

$$^{n}C_{k} = \frac{\lfloor n}{\lfloor k \, \lfloor n-k}$$

⇒ **Mean** :-

$$\text{Mean} = np$$

⇒ **Variance** :-

$$\text{Var} = npq$$

⇒ **Std**

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