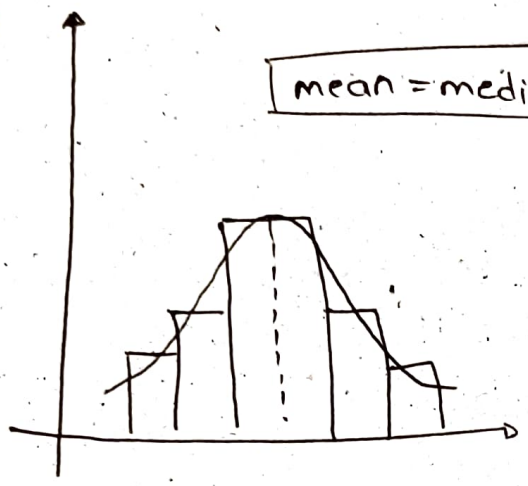
④ Skewness

1)

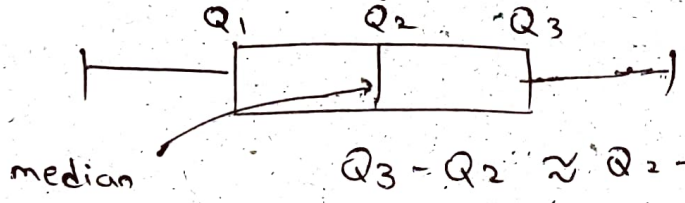


mean = median = mode

⇒ symmetrical distribution (NO SKEWNESS)

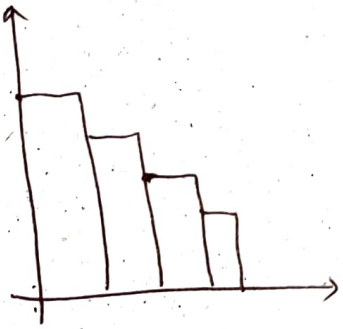
The mean, median and mode all are perfectly at the center.

Hence, boxplot for above distribut^n will look like:



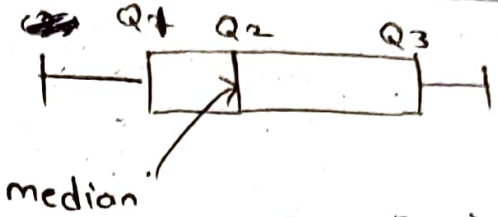
eg → Normal / Gaussian Distribution

2) Right skewed



⇒ positive skewed (right)

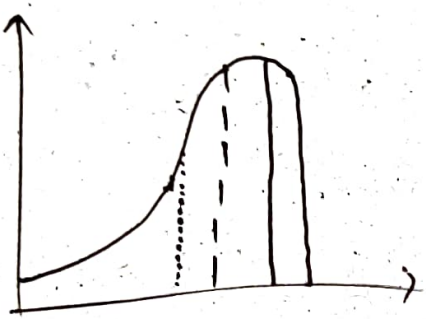
Boxplot for right skewed →



$Q_3 - Q_2 > Q_2 - Q_1$

mean > median > mode

3) Left skewed (-ve skewed)

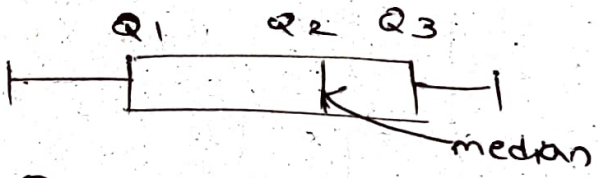


..... mean
– – median
—— mode

mean < median < mode

box-plot for left skewed

↳



$Q_2 - Q_1 > Q_3 - Q_2$

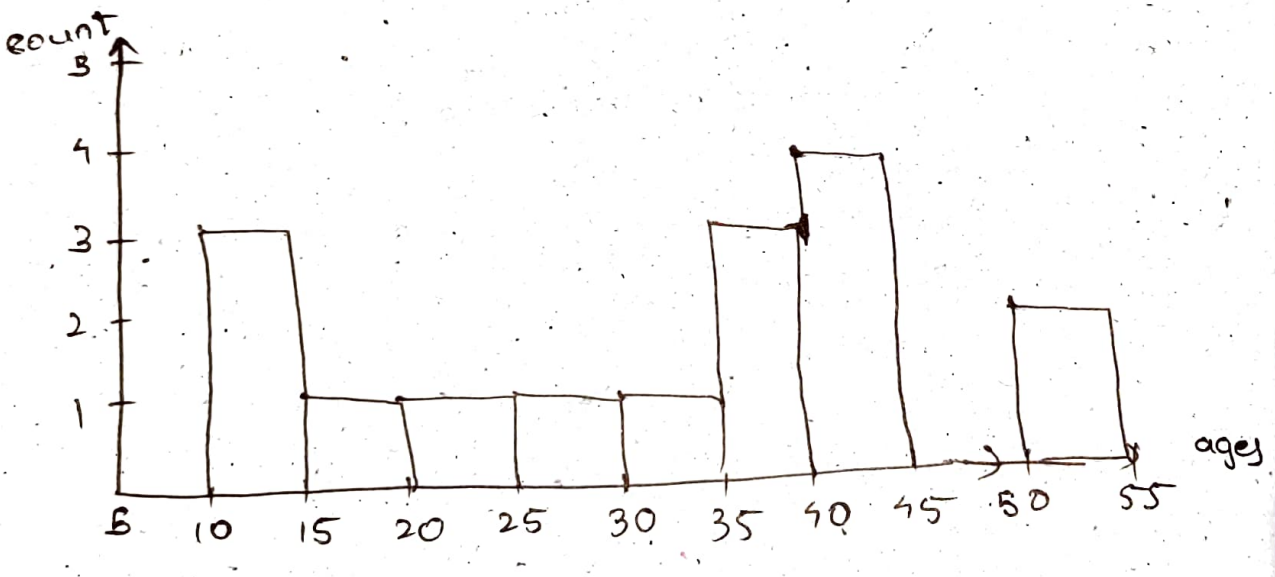
Histograms - It is nothing but representation of frequency of elements.

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

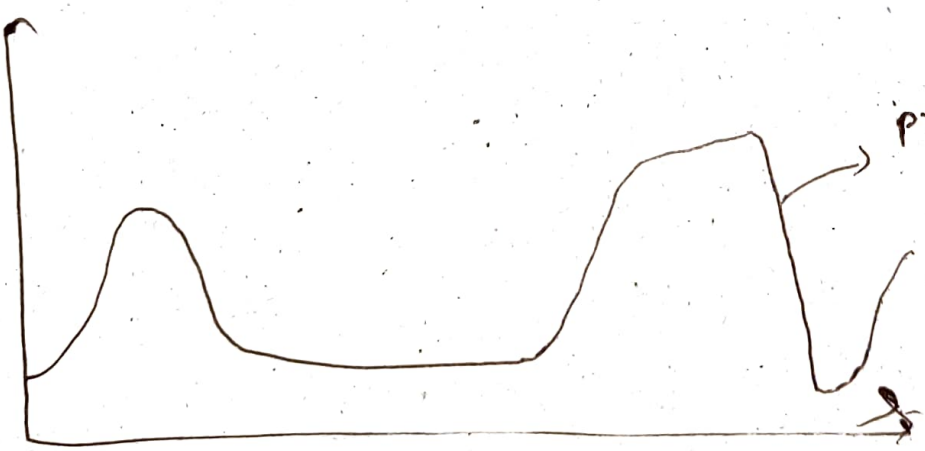
Suppose bins = 10

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

Note: while checking range (say 10 to 15), we check for ≥ 10 but < 16.

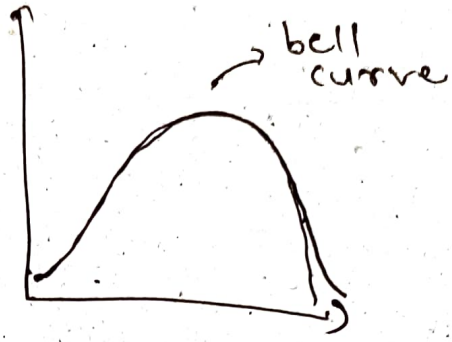


Smoothening above histogram

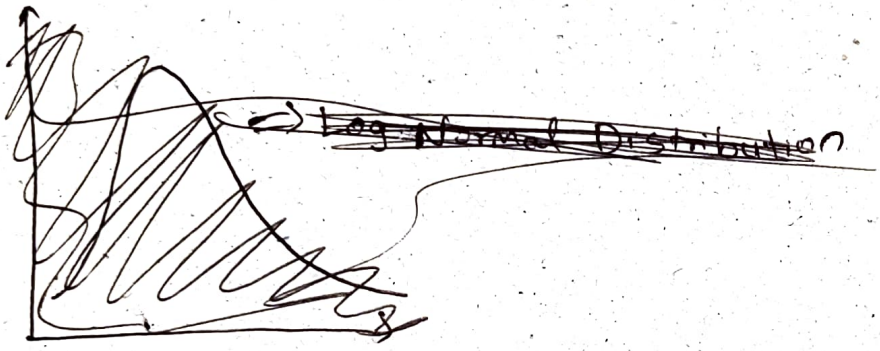


histogram

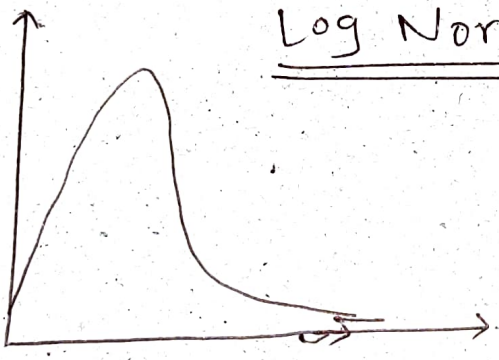
smoothen



normal / gaussian distribution



Log Normal Distribution



⊛ Probability Distribution Functions

discrete

- Probability mass function (pmf)
- probability density function (pdf)

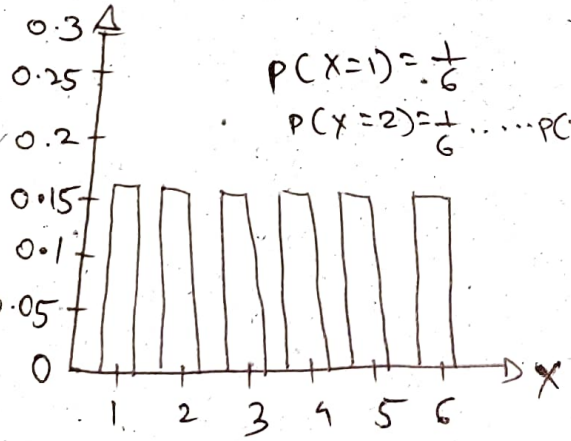
→ cumulative distribution function (cdf)

continuous

① Probability mass function (pmf) : Distribution of discrete random variable

eg → rolling a die {1, 2, 3, 4, 5, 6}

prob mass $\frac{1}{6} = 0.16$



$P(X=1) = \frac{1}{6}$

$P(X=2) = \frac{1}{6}$ ..... $P(X=6) = \frac{1}{6}$

$P(x) = P(X=x)$

conditions:

- $P(x) \geq 0$
- $\Sigma P(x) = 1$

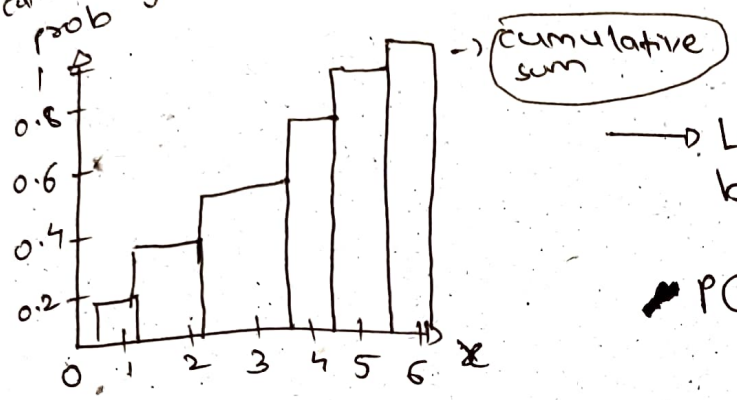
Note: ① Value of pmf at specific point indicates the probability of the random variable taking on that value

② Sum of pmf over all values of random variable is 1

② cumulative distribution function (cdf)

For Discrete Data

eg → rolling a die {1, 2, 3, 4, 5, 6}



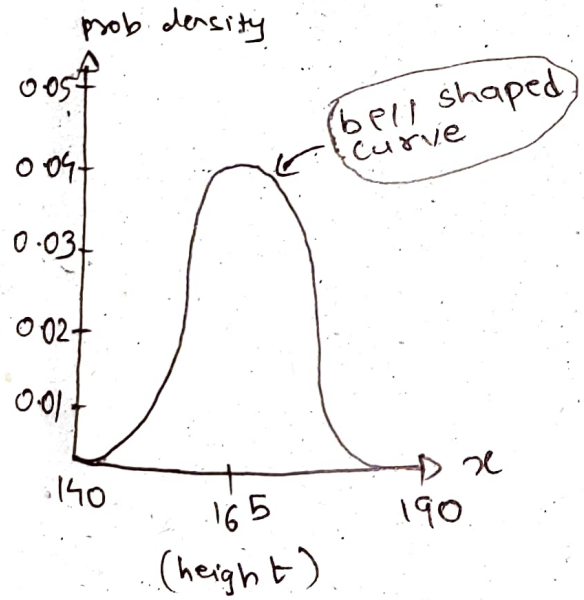
→ Last bar will always be 1.

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

Note: ① Value of cdf at a specific point indicates probability that random variable will take on a value less than or equal to that point.

③ probability density function (pdf): Distribution of continuous data / random variable.

eg → height of students



- Not that students with height of 140 or 190
- max students near height of 165.

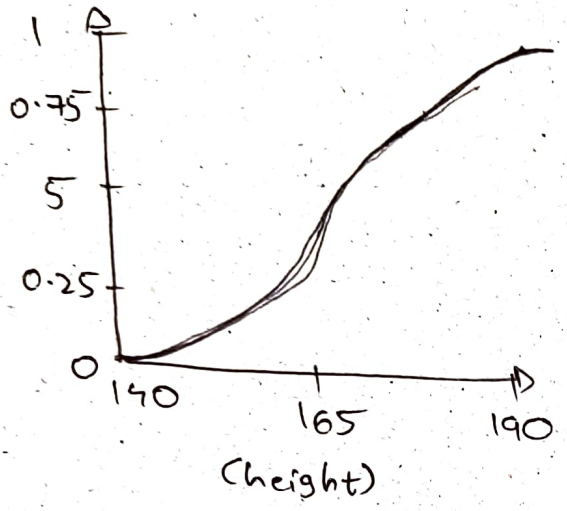
Note:
You can see that the maximum height of curve is at 0.04.
what does that mean?
will see in some time

⑤ Cumulative distribution function : (cdf)

For continuous data

cumu Prob

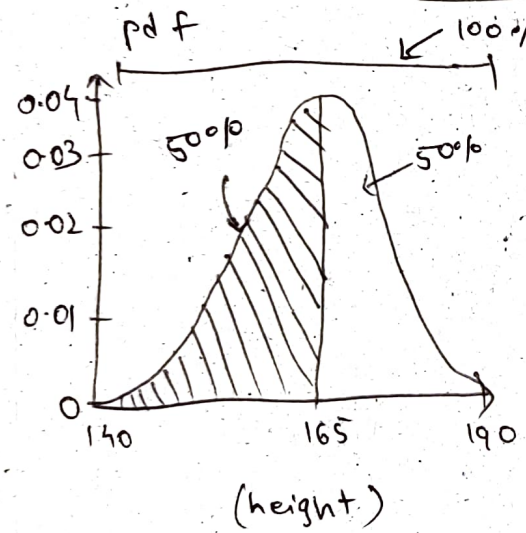
eg. → height of students

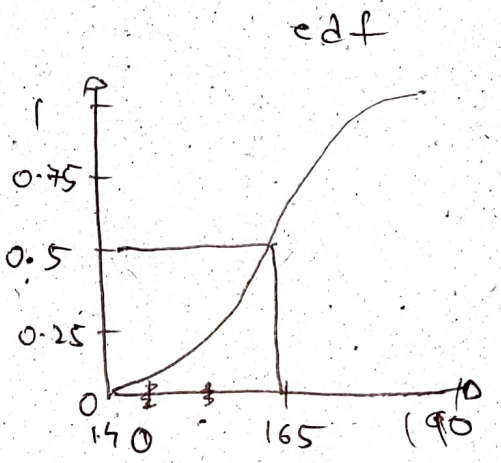


→ s curve

→ always till 1

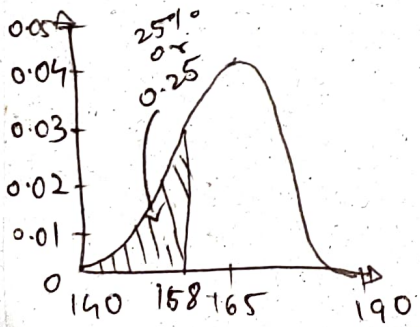
Comparing pdf & cdf



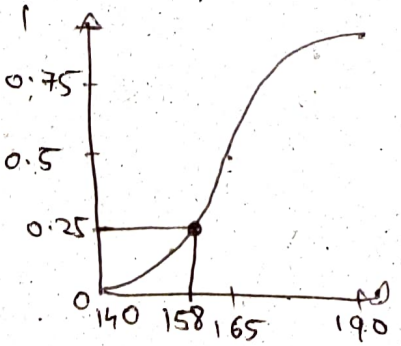


Till 165, area under curve is 50% → 0.5

Here, at 165, we can see it matches to 0.5 at Y-axis.



→

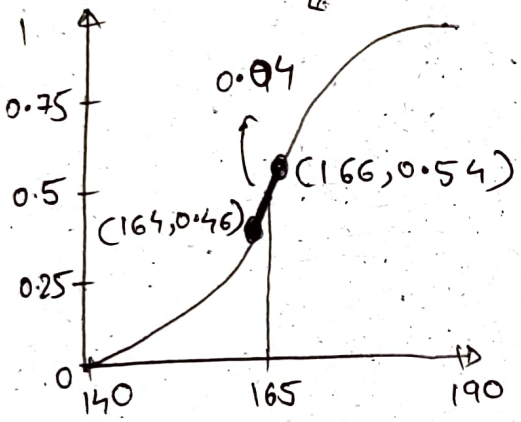


Hence, we can see that, the y-axis in cdf represents the area under curve for a specific value of random variable in pdf.

Hence in this way, we can derive cdf from pdf (pdf $\xrightarrow[\text{under curve}]{\text{area}}$ cdf)

Now, for cdf ⟶ pdf.

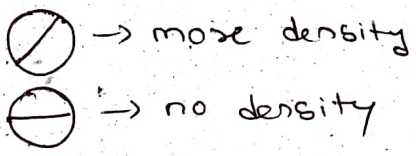
gradient = $\frac{\text{rise}}{\text{run}} = \frac{0.54-0.46}{166-164} = \frac{0.08}{2} = 0.04$



Find gradient at point which intersects with 165.

gradient → slope.

Higher gradient = higher density



0.04 is approx gradient at 165.

You can verify this from pdf graph.

Maximum height (at 165) is at 0.04 on y-axis.

This means, the y-axis in pdf represents the gradient for a specific value of random variable using the s-curve in cdf.

Hence, in this way we can derive pdf from cdf.

pdf $\underset{\text{gradient}}{\overset{\text{area under curve}}{\rightleftharpoons}}$ cdf

Probability density function → f(x)

Cumulative distribut function → F(x)

Note: ① differentiation of cdf = pdf

② integration of pdf = cdf.

$$\frac{d}{dx} F(x) = f(x) \quad \& \quad \int_{-\infty}^{x} f(x)dx = F(x)$$

④ Types of Probability Distribution.

① Normal / Gaussian distribution. (pdf)

② Bernoulli Distribution (pmf): Outcomes are only two → {success, failure}

③ uniform distribution

④ Log normal distribution (pdf)

⑤ Poisson distribution. (pmf)

⑥ Power log distribution. (pdf) → 80-20% rule

⑦ Binomial Distribution (pmf)

die.

→ uniform distribution

① Bernoulli Distribution. ← pmf (Binary Outcomes)

It 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).

For a set of possible outcomes, it asks a **Yes-No** question.

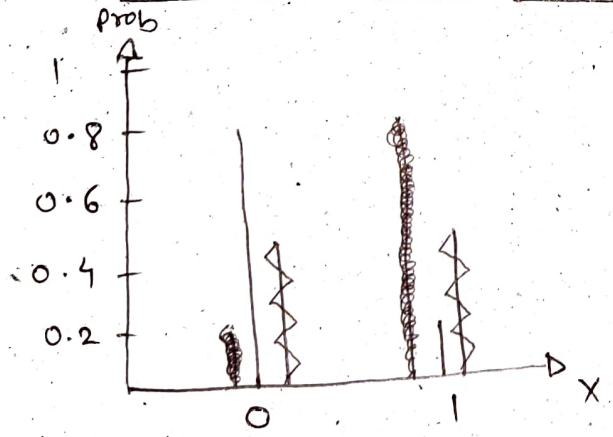
success / yes / true / one → p
failure / no / false / zero → q → (1-p)

eg → Tossing a fair coin {0, 1}

P(H) = **0.5** → p

P(T) = 1 - 0.5 = 1-p = **0.5** → q

probability mass function



P(x=0) = 0.2 & P(x=1) = 0.8
P(x=0) = 0.8 & P(x=1) = 0.2
P(x=0) = 0.5 & P(x=1) = 0.5

eg → whether the person will pass/fail.

mathematically;

Pmf → $k = 0$ or 1

$$\boxed{P(x=k) = p^k (1-p)^{1-k}}$$

$$P(x=1) = p^1 (1-p)^{1-1} = (p)$$

$$P(x=0) = p^0 (1-p)^{1-0} = (1-p) \rightarrow q \;//$$

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

Mean, Variance, Std dev for Bernoulli distribn:

mean:

$k = 1$ or 0

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

Say $P(K=1) = 0.6 \rightarrow p$

$P(K=0) = 0.4 \rightarrow q$

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

$$= \boxed{0.6} \rightarrow p$$

$$\therefore E(k) = p$$

median :

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

Variance & std

Variance $= p(1-p) = pq$

Std $= \sqrt{pq}$

← pmf

⑧ Binomial Distribution :

Binomial distribution with parameters 'n' and 'p' is discrete probability distribution of number of successes in a sequence of (n) independent experiments, each asking a yes-no question and with its own Boolean-valued outcome : (success with probability (p)) or (failure with probability (q = 1-p)). A single experiment is also called a Bernoulli trial or Bernoulli experiment. A sequence of outcomes is called Bernoulli process.

For a single trial/experiment i.e. n = 1, the binomial distribution is a bernoulli distribution.

eg → Tossing a coin {Bernoulli distribution}

P(H) = 0.5 ↓ (1-p) → 0

P(T) = 0.5 ↓ (p) → 1

Now, eg → Tossing a coin for 10 times.

① P(T) = p P(H) = 1-p
② P(T) = p P(H) = 1-p
⋮
⑩ P(T) = p P(H) = 1-p

} ⇒ Binomial distribution. (combine various Bernoulli)

Parameters :

$n \in \{0, 1, 2, \ldots\}$ → no. of trials/exp.

$p \in [0, 1]$ → 0 to 1 → prob. for each trial (success)

$q = 1 - p$

Pmf for binomial distribution:

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

or

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

$K \in \{0, 1, 2, \ldots\}$ → no. of successes

Mean of Binomial dist:

$$\text{mean} = np$$

Variance & std:

$$\text{Var} = npq$$

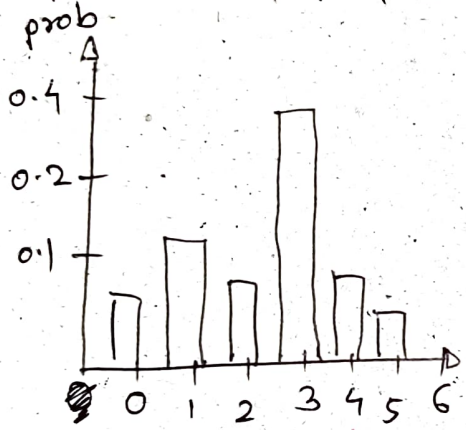
$$\text{Std} = \sqrt{npq}$$

⑧ Poisson Distribution: ← pmf

① Discrete distribution

② Describes the number of events occuring in a fixed time interval.

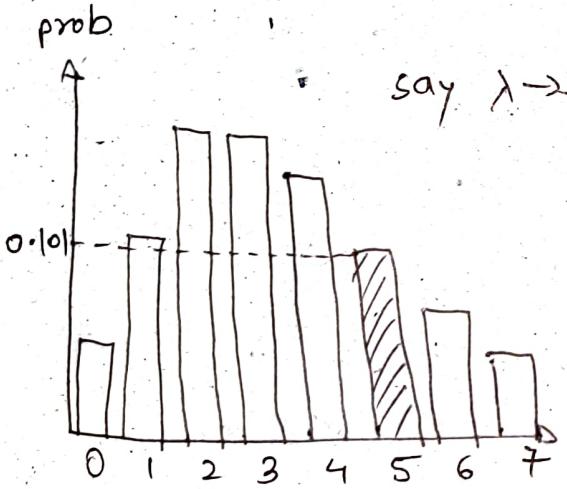
eg → • no. of people visiting hospital every hour.
• no. of people visiting banks every hour
• no. of people visiting airport every hour.



$\lambda = 3$

↳ mean no. of events occuring every hour.

λ → no. of events within a given event / interval of time or space



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

$$\therefore P(X=5) = \frac{e^{-3} \cdot 3^{5}}{5!}$$

$$= 0.101 \Rightarrow 10\%$$

∴ prob that event '5' will occur

Similarly, you can find

$P(X=3) + P(X=5) =$ ---

Mean of poisson distribution.

$$\text{mean} \Rightarrow E(X) = \mu = \lambda * t$$

$\lambda \rightarrow$ expected no. of events to occur at every time interval.

$t \rightarrow$ time interval

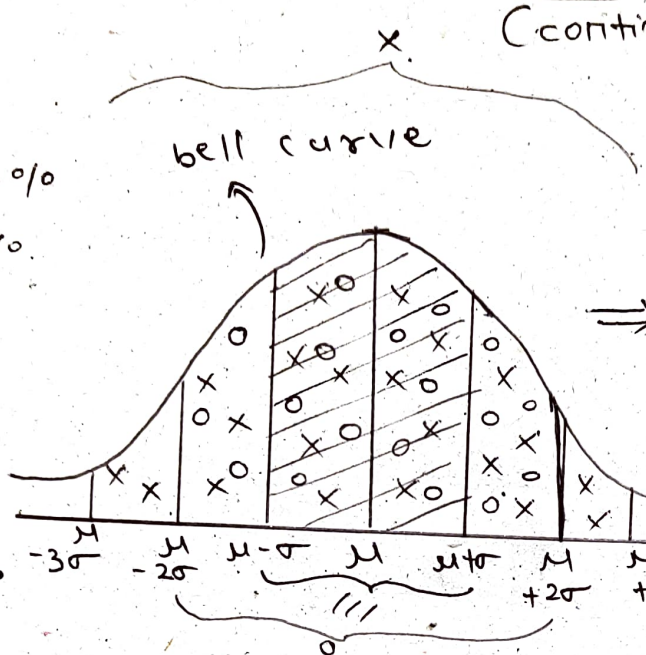
Variance of poisson distribution.

$$\text{Variance} \Rightarrow E(X) = \mu = \lambda * t$$

Hence, In a poisson distribution, Mean & Variance are equal.

(XV) Normal / Gaussian Distribution.: ← pdf.

(continuous random var)



In symmetric distribution, μ is in center of x-axis
hence, area on left = area on right

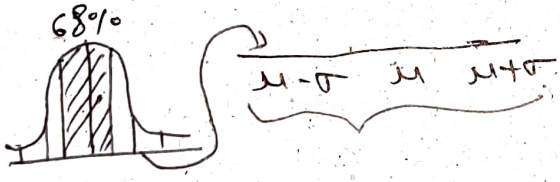
$\mu + \sigma$ → 'μ' + one std dev. (right) →

$\mu - 3\sigma$ → 'μ' − three std dev (left) ←

Empirical Rule: [3 sigma rule]

68 - 95 - 99.7 %

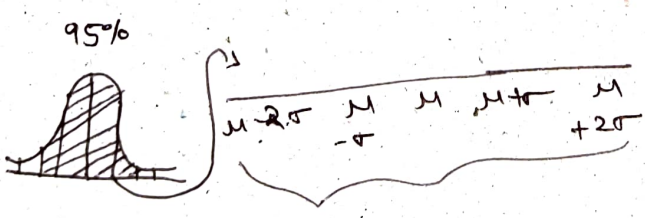
(68) →



⟹ 68% of data will be present in this area ($\mu-\sigma$ to $\mu+\sigma$) first std dev on b.s.

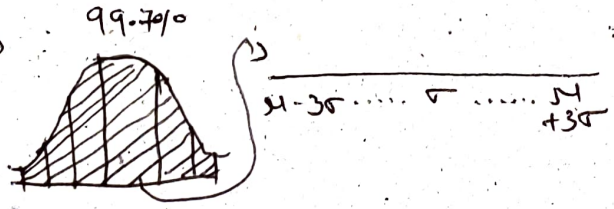
---

(95) ⇒



⟹ 95% of data will be present in this area ($\mu-2\sigma$ to $\mu+2\sigma$) two std. dev on b.s.

(99.7) →



⟹ 99.7% of data will be present within three std. dev. on b.s from the mean ($\mu-3\sigma$ to $\mu+3\sigma$)

How to find out if a distribution is normal dist or not?

Q-Q plot ⇒ we can find out whether a dist is normal/gaussian distribution or not.

Probability — X → random var

$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68\%$

$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\%$

$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.7\%$

eg → weight, height, iris dataset

Note: Anything begond $\mu - 3\sigma$ & $\mu + 3\sigma$, we generally consider as outliers