Probability distribution

Random variable → A set of possible values from a random experiment.

→ A random variable value is unknown.

→ A function that assign values to each of experiment outcomes.

$$X = \{1, 0\} \quad \underline{\text{Tossing a coin}} - \underbrace{H, T}$$

↓

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

$$P(T) = \frac{1}{2}$$

$\Rightarrow \frac{1}{n}$ where n is Total no of outcome.

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

x dice – 1, 2, 3, 4, 5, 6

$\Rightarrow \frac{1}{6}, \frac{1}{6}, \ldots \frac{1}{6}$

$\Rightarrow \frac{1}{n}$ where $n = 6$

↓

function that can be used to get Probability.

Outcomes of an experiment

→ tossing a coin

→ throwing a dice

↓

<u>discrete outcomes</u>

(pmf)

Calculate the height of students in the class.

↓

160 cm, 160.1 cm

↓

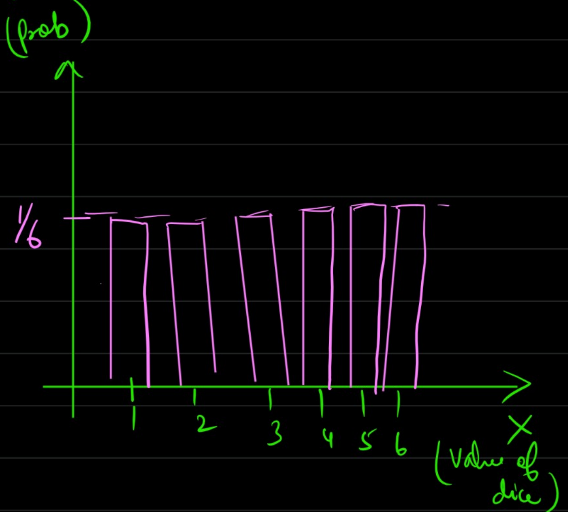
Continuous outcomes.

(pdf)

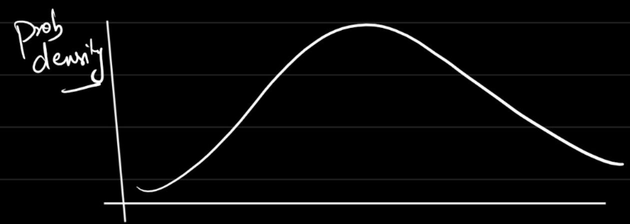
* irrespective of outcomes nature, draw outcomes in a form of distribution, probability distribution function.

throwing a dice $\Rightarrow$

$$\left.\begin{array}{l} P(1) = 1/6 \\ P(2) = 1/6 \\ P(3) = 1/6 \\ P(4) = 1/6 \\ P(5) = 1/6 \\ P(6) = 1/6 \end{array}\right\} \Rightarrow$$



Height of students



Two types of Experiment $\longrightarrow$ Prob distribution functions.

discrete $\Downarrow$ Prob mass fn

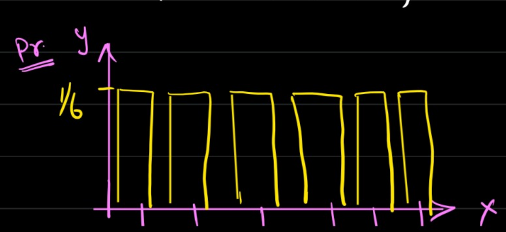
continuous $\Downarrow$ Prob density function

① Probability mass function $\Rightarrow$ Distribution of Discrete random variable

eg Rolling of a dice $\{1, 2, 3, 4, 5, 6\}$

$$Pr(X \le 3) = Pr(X=1) + Pr(X=2) + Pr(X=3)$$

$$= 1/6 + 1/6 + 1/6 \Rightarrow 3/6 = \underline{\underline{1/2}}$$

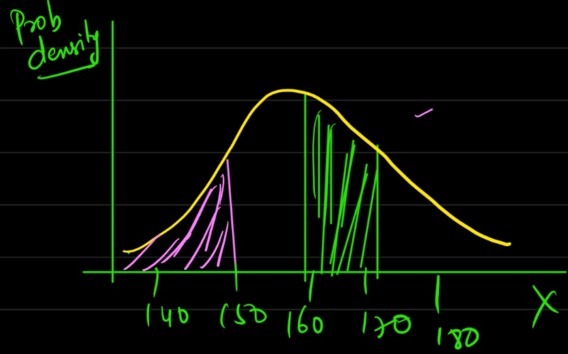


1 2 3 4 5 6

$$Pr(X \le 6) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6}$$

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

② Probability density fn (pdf) → Distribution of continuous data.

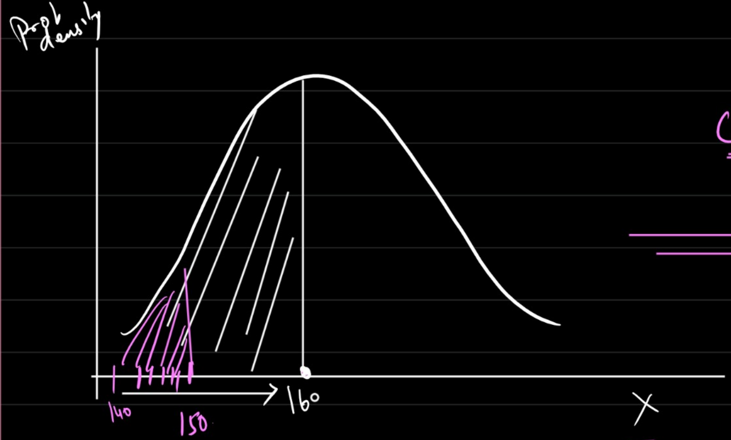


$Pr(X \le 150)$ = Area under curve

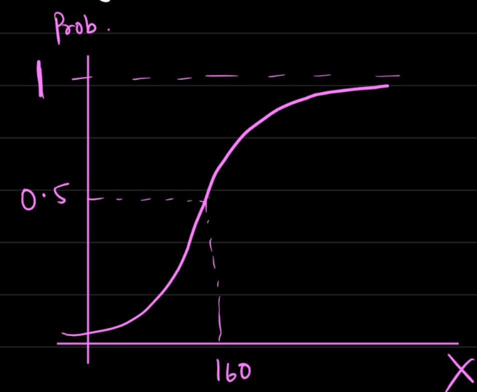
[0 to 1]

↑

③ Cumulative Distribution fn (cdf) → CDF is summation of all probabilities possible upto a given point



Cumulative



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

Different types of distribution

Prob dist. fn

Prob density fn

- → Normal distr / Gaussian distribution
- → Standard Normal distrib^n
- → Log Normal distr
- → Chi-square distribution
- → F - distribution.

Prob mass fn

- → Bernoulli
- → Binomial
- → Poisson distribution

Uniform distribution → discrete Prob
Uniform distribution → Cont. Prob.

* Probability distribution functions and Cumulative distribution function

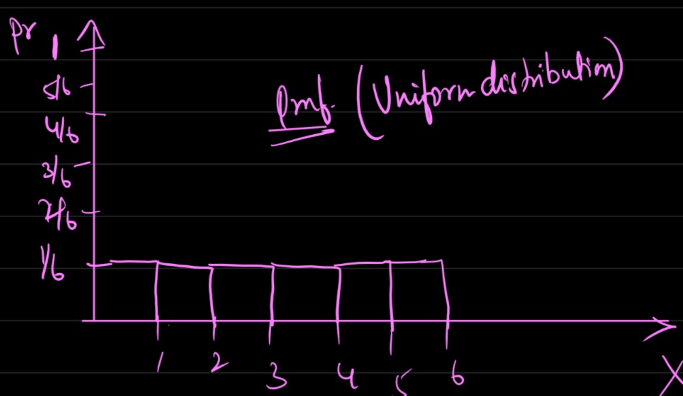
① Pmf.

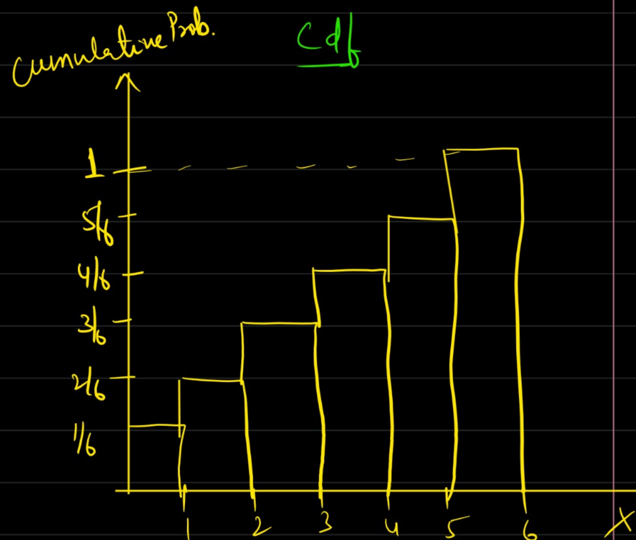
→ Discrete random variable.

eg. Rolling a dice

$\{1, 2, 3, 4, 5, 6\}$

$$\left.\begin{array}{l} Pr(1) = 1/6 \\ Pr(2) = 1/6 \\ Pr(3) = 1/6 \\ Pr(4) = 1/6 \\ Pr(5) = 1/6 \\ Pr(6) = 1/6 \end{array}\right\} \rightarrow \text{Uniform distribution}$$





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

$P(X \le 2) = P(X=1) + P(X=2)$

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

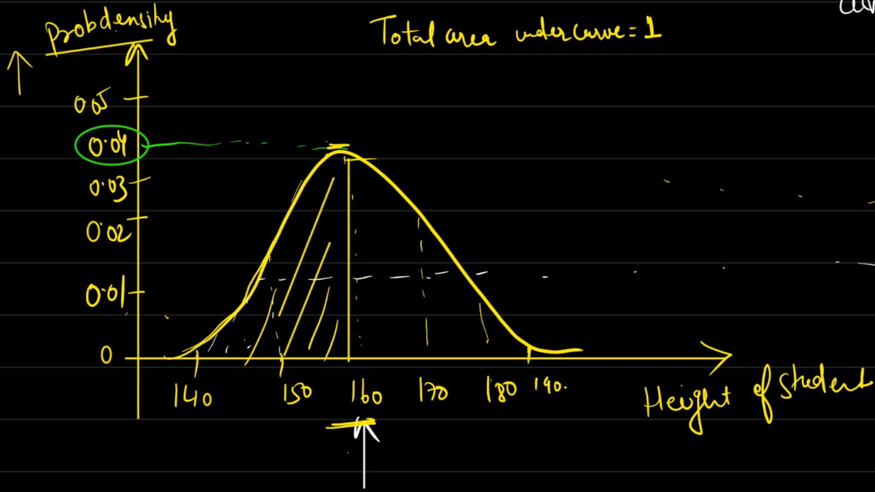
$P(X \le 6) = Pr(X=1) + P(X=2) + Pr(X=3) + Pr(X=4) + Pr(X=5) + P(X=6)$

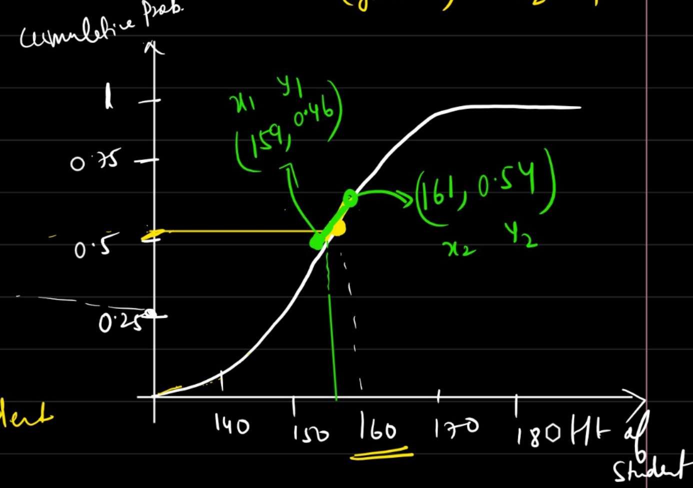
$= \underline{1}$

② Prob density fn (Pdf)

→ Random Variable is continuous in nature.

Slope (gradient) $= \frac{y_2 - y_1}{x_2 - x_1}$





Gradient (slope) $= \frac{0.54 - 0.46}{161 - 159}$

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

Prob density of a pdf = slope of cdf.

gradient of cdf.

Discrete Uniform distribution

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

→

Uniform distribution
- Discrete Uniform Dist (pmf)
- Continuous Uniform distribution (pdf)

→ In a discrete Uniform distribution, the outcomes are discrete and have the same prob.

- eg. rolling a dice
- eg. tossing a coin
- eg. picking up a card from well shuffled deck

1, 2, 3, 4, 5, 6

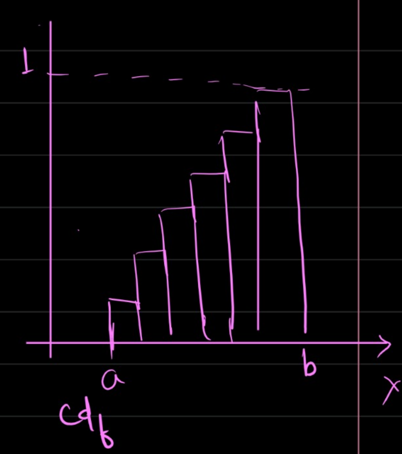
$P(X=1) = 1/6$

Notation of Uniform distribution ⇒ $U(a,b)$ | $Unif\{a,b\}$

Possible values when you throw a die → 1 – 6

pmf → $\frac{1}{n}$

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



Q what is prob of getting 3 when you throw a dice?
$\rightarrow$ $P(X=3) = 1/6$

mean of discrete Uniform distribution $\Rightarrow \frac{a+b}{2}$

Variance " " " $\Rightarrow \frac{n^2-1}{12}$

* mean = Sum of nos divided by number of numbers.
* Expected value $\rightarrow$ the long run avg value of repetitions of experiment it represents.

$\rightarrow$ long term avg value of a random value.

dice = 1, 2, 3, 4, 5, 6

$$\text{mean} = \frac{1+2+3+4+5+6}{6} = \underline{3.5}$$

$$E.V = \sum_{i=1}^{6} x_i \, P(x_i)$$

x	prob	$x \cdot p$
1	$1/6 \Rightarrow 0.167$	0.17
2	0.167	0.33
3	0.167	0.50
4	0.167	0.67
5	0.167	0.83
6	0.167	1

Sum(px) = 3.50

$\rightarrow$ mean is used freq. distribution
$\rightarrow$ Expected value is used for prob distribution

$$E(X) / \mu \Rightarrow \sum_{i=1}^{i} x_i \, p(x_i)$$

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

$$E(X) = \sum x \, p(x)$$

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

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

$$\Rightarrow \frac{N+1}{2}$$

Sum of first 'N' no. $= \frac{N(N+1)}{2}$

$1+2+3 \cdots 10$

$\Rightarrow \frac{10(10+1)}{2}$

Sum of first N squares $= \sum_{i=1}^{N} x^2 = \frac{N(N+1)(2n+1)}{6}$

$1^2+2^2+ \cdots 10^2$

$\frac{10(10+1)(2 \times 10+1)}{6}$

$$E(X^2) = \sum x^2 \, p(x)$$

$$\Rightarrow \sum_{i=1}^{N} x^2 \cdot \frac{1}{N}$$

$$\Rightarrow \frac{1}{N}\left(1^2 + 2^2 + 3 + \cdots N^2\right)$$

$$= \frac{1}{\cancel{N}}\left(\frac{\cancel{N}(N+1)(2N+1)}{6}\right)$$

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

$$Var(x) = E(X^2) - (E(x))^2$$

$$= \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{2}$$

$$\Rightarrow \frac{2N^2+3N+1}{6} - \frac{N^2+2N+1}{4}$$

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

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

Bernoulli distribution → Binary < two outcomes

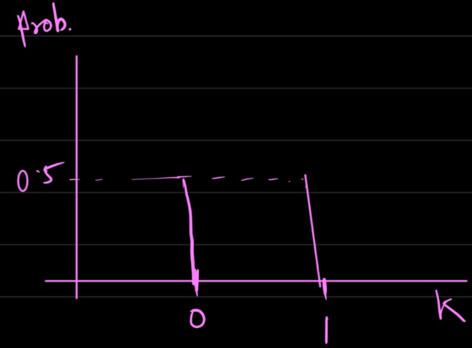
A discrete prob distribution of a random variable which takes only two possible outcomes, typically labelled as success (coded as 1) and failure (coded as 0) with a fixed prob of success and failure 1.

* To model any experiment with only two possible outcome.

eg: Tossing a coin

Head or Tail $(K \in \{0,1\}$

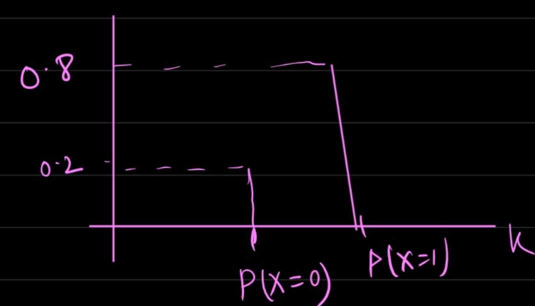
$$P(X=0) = 0.5 = p$$
$$P(X=1) = 1-0.5 = q$$
$$(1-p)$$



pmf ⇒

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

$$P(X=k) = \underline{p^k (1-p)^{1-k}}$$



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

① if $k=1$

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

② if $k=0$

$$\boxed{p+(1-p) = 1}$$

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

Conditions of Bernoulli dist

① No of trial should be finite
② Each trial should be independent
③ Only two possible outcome < pass / fail
④ Prob of each output should be same in every trial.

1 2 3 4 5

$P(H)=0.5$ $P(H)=0.5$ $P(H)=0.5$ $P(H)=0.5$ $P(H)=0.5$

Examples

→ tossing a coin
→ Someone has committed a fraud or not
→ getting a six or not
→ pass | fail
→ win | or not
→ Customer Conversion
→ rain or not

Q Bumrah bowls 6 balls at wicket, with prob of 0.6 at hitting the stump with each ball. What is prob of not hitting a wicket?

→ Bernoulli distn.

$p(\text{hitting a wicket}) = 0.6$
$p(\text{hitting not a wicket}) = 1 - 0.6 = \underline{0.4}$

* Mean and Variance of Bernoulli distn.

Mean ⇓ p Variance ⇓ $p(1-p)$

① Mean

$$E(x) = \sum_{i=1}^{k} x \, \underline{P(x)}$$

k can take two values 0 & 1

$$= x=1 \quad + \quad x=0$$
$$= 1 \times 0.6 + 0 \times 0.4$$
$$= 0.6 \Rightarrow \underline{p}$$

$P(x=1) = 0.6 = p$
$p(x=0) = 0.4 = q$
$(1-p)$

$\underline{E(x) = p}$

② $var(x) \Rightarrow E(x^2) - E(x)^2$

$E(x) \hookrightarrow p$

$$E(x^2) = \sum x^2 p(x)$$

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

$$= p.$$

$$Var(x) = p - (p)^2$$

$$= p(1-p)$$

* **Poisson distribution**

→ The Poisson distribution 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.

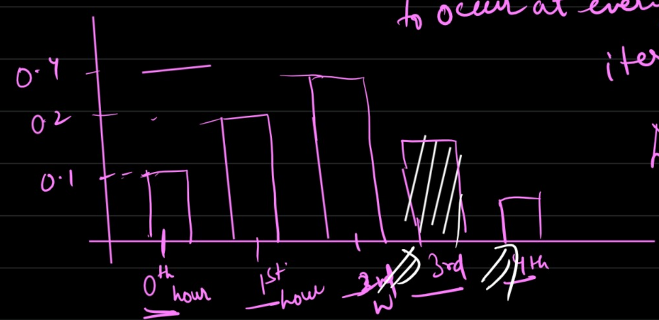
* No of events occurring in a fixed time interval.

ex. No of calls recieved by a customer care every hours.

$\lambda \to$ Expected No of events to occur at every time interval.

ex. No of people visiting temple/hospital/banks/airport every hour.

pmf



$\lambda = 100$

eg. No of accident every hour at a busy route

eg. No of emails recieved every hour.

$$\text{pmf} \Rightarrow p(X=x) \Rightarrow \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

e is -2.71828

λ – avg rate of events every interval

$\lambda = 10$

P visiting at 5th hour

$$P(X=5) = \frac{e^{-\lambda}\lambda^{x}}{x!} \Rightarrow \frac{(2.718)^{-10} \cdot 10^{5}}{5!} \Rightarrow$$

eg: The avg number of customers entering a store in an hour is 5. What is the prob of exactly 3 customers will enter the store next hour?

→ $\lambda = 5$

$$P(X=3) \quad \frac{e^{-\lambda} \cdot \lambda^{k}}{k!} \Rightarrow \frac{e^{-5} \cdot 5^{3}}{3!}$$

$$\Rightarrow \frac{(2.718)^{-5} \times 125}{6}$$

$$\Rightarrow \frac{0.00674 \times 125}{6} \approx \underline{\underline{0.14}}$$

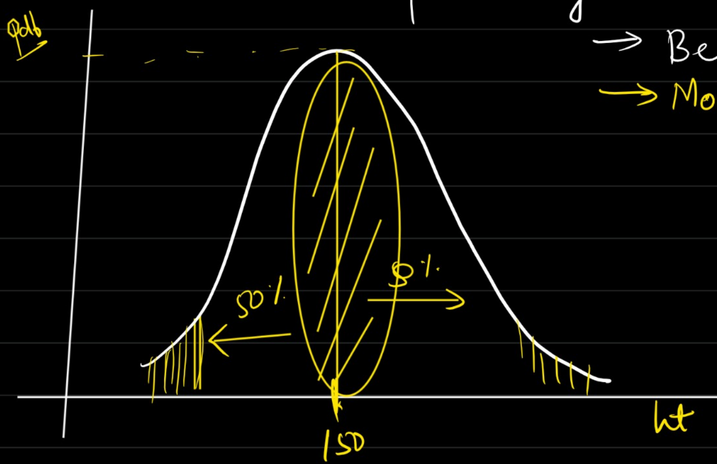
Avg no of events at every interval.

$\underline{\text{Mean}} \rightarrow \lambda \times t$

$\underline{\text{Variance}} \rightarrow \underline{\lambda \times t}$

Normal / Gaussian distribution

→ A continuous probability distribution



→ Bell-Shape distribution.

→ Most of the real world data follows a normal distribution

Example:-

→ Height of a population

→ IQ of a population

→ Measurement errors

→ Exam scores

→ Blood Pressure.

→ Size of things

* Characteristics of N.D

→ Symmetrical about mean

→ mean = median = mode.

→ Skewness = 0

Empirical rule of a Normal distribution.

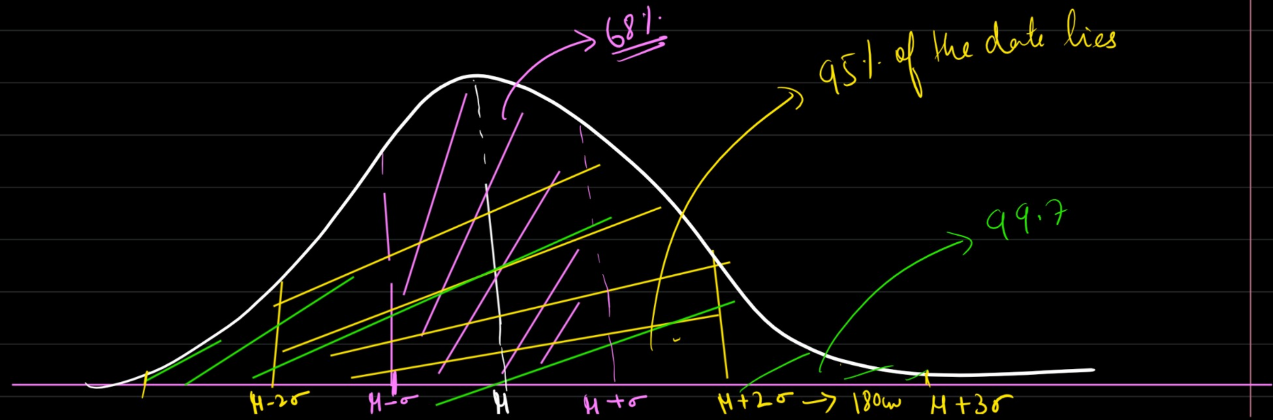
68% - 95% → 99.7% rule.

→ 68% of the data values lies within 1 standard deviation from the mean value

$x = \{160, 161, 162.5 \ldots\}$

→ 95% ⟶ within 2 SD

→ 99.7% ⟶ within 3 SD



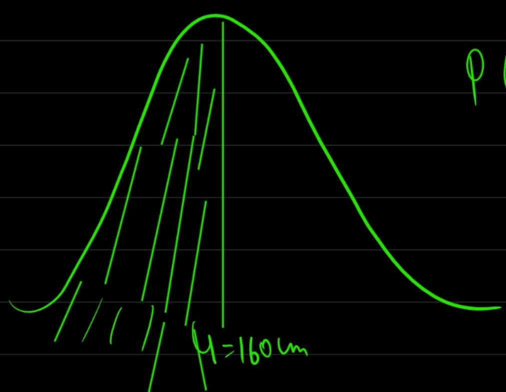
$\mu - 3\sigma$ ↓ 140   ↓ 150

$\mu = 160\,cm$, $\sigma = 10\,cm$ $\Big\} \rightarrow 68\% \rightarrow 160-10$ and $160+10$

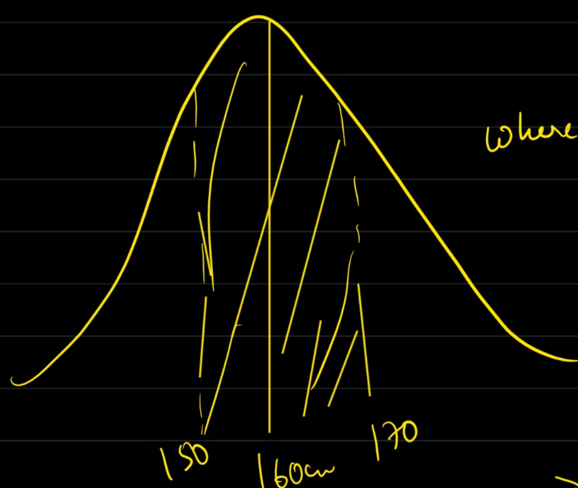
$\rightarrow 150$ to $170\,cm$.

$95\% \begin{cases} \mu - 2\sigma \rightarrow 160 - 2\times10 \Rightarrow 140\,cm \\ \mu + 2\sigma \rightarrow 160 + 2\times10 = 180\,cm \end{cases}$

$99.7\% \begin{cases} \mu - 3\sigma - 160 - 3\times10 \Rightarrow 130\,cm \\ \mu + 3\sigma - 160 + 3\times10 \Rightarrow 190\,cm \end{cases}$



$P(X \le 160\,cm) = 50\% = 0.5$



where $\mu = 160$
$\sigma = 10$

$\rightarrow Pr(\mu - \sigma \le X \le \mu + \sigma) \approx 68\%$

$\Rightarrow \underline{0.68}$

$\rightarrow Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 95\%$

$Pr(140 \le X \le 180) = 0.95$

$\rightarrow Pr(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx 99.7\%$

$Pr(130 \le X \le 190) = \underline{0.997}$

Continuous Uniform distribution

- A uniform distribution refers to a type of prob distribution in which outcomes are equally likely.

→

Uniform distribution

→ Discrete Uniform Dist (Pmf)

→ Continuous Uniform distribution (pdf)

* A continuous uniform prob distribution is a dist that has an infinite no of values defined in a specified range/bound.

→ r.v is continuous.

→ rectangular dist.

Examples – A perfect Random no generator.)

- Prob of guessing exact time at any moments

→ Waiting time at a bus stop (consistent bus arrival)

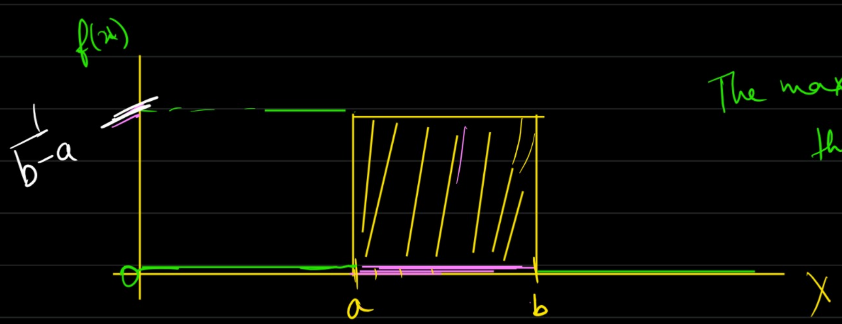
→ Temp variation in a day [if a temp fluctuating b/w a min and max value)

Notation: $U(a, b)$

parameter: $-\infty < a < b < \infty$

$b >> a$ a is min, b is max.

pdf



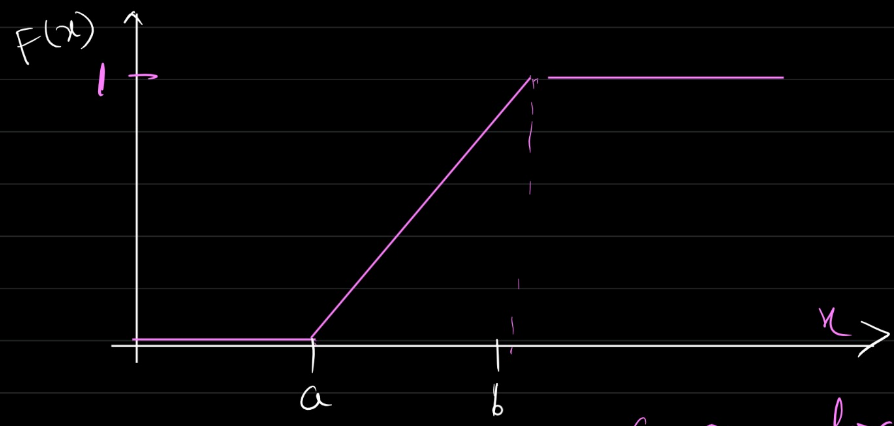
The maximum prob of the variable X is 1.

Area of rectangle = base × height

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

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

cdf

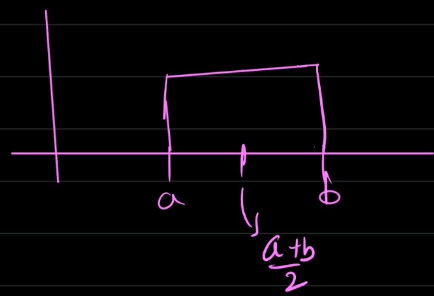


$$\text{cdf} = \begin{cases} 0 & \text{for } x < a \\ \dfrac{x-a}{b-a} & x \in [a, b] \\ 1 & x > 1 \end{cases}$$

mean/median $= \frac{1}{2}(a+b)$

↓

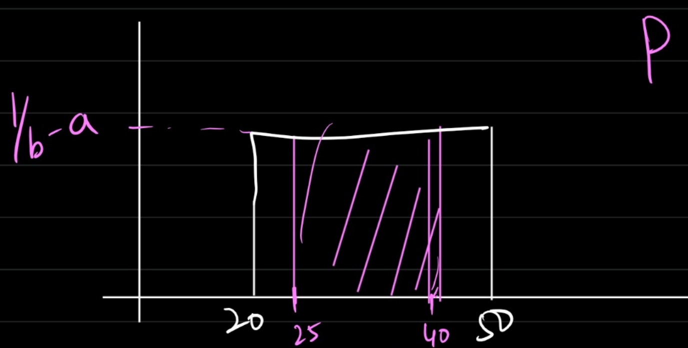
Avg of pdf | Center of distribution.



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

Q1 The number of items sold at a shop daily is uniformly distributed with max and min item sold 50 and 20 respectively.

→ Prob of daily sales to fall between 25 & 40



$$P(25 \le X \le 40)$$

⇓

breadth × ht

$$(40-25) \times \frac{1}{50-20}$$

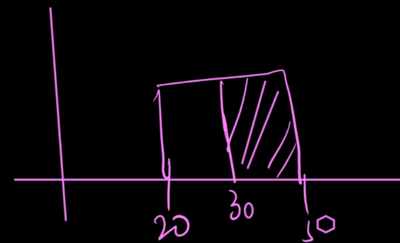
$$\Rightarrow 15 \times \frac{1}{30} = \frac{1}{2} = 0.5$$

50% chance the no of items sold [25, 40]

→ prob. that sales of ≥ 30

$$P(X \geq 30) = (50-30) \times \frac{1}{50-20}$$

$$\Rightarrow 2\cancel{0} \times \frac{1}{3\cancel{0}} = \underline{66.66}$$



Q The amount of time for Pizza delivery is Uniformly distributed b/w 15 and 60 mins. What is standard deviation of the amount of time it takes for a pizza to be delivered?

$\Rightarrow$ $a = 15$ min
$b = 60$ min

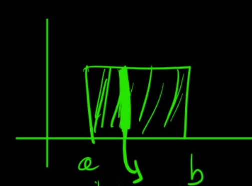
$$\text{var}(X) = \frac{(b-a)^2}{12} = \frac{(60-15)^2}{12} = \frac{45^2}{12} = 168.75 \text{ min}$$

$$\sigma = \sqrt{\text{var}} = \sqrt{168.75} \approx \underline{13 \text{ mins}}$$

---

for a cont^n random var with prob density fn $f(x)$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$



$\int_a^b dx$

$$E(X) = \int_a^b x\, \underline{f(x)}\,dx$$

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

$$\Rightarrow \frac{1}{b-a}\int_a^b x\,dx$$

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

$$= \frac{1}{2}\frac{1}{\cancel{b-a}}\cancel{(b-a)}(b+a)$$

$$\Rightarrow \frac{a+b}{2}$$

if $X \sim U(a,b)$

$\int_a^b x\,dx$

$= \frac{x^2}{2}$

$= \frac{1}{2}\left[x^2\right]_a^b$

$\frac{1}{2}\left(b^2 - a^2\right)$

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

$$\text{Var}(X) = E[X^2] - \left[E[X]\right]^2$$

$E[X] \to \frac{a+b}{2}$

$E[x^2]$ $\int_a^b x^2 \underline{f(x)}\,dx$ $\hookrightarrow \frac{1}{b-a}$

$\Rightarrow \frac{1}{b-a}\int_a^b x^2\,dx \Rightarrow \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b$

$\Rightarrow \frac{1}{3(b-a)}\left(b^3-a^3\right)$

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

$= \frac{b^2+ab+a^2}{3}$

$Var(X) = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2$

$\Rightarrow \frac{b^2+ab+a^2}{3} - \frac{a^2+b^2+2ab}{4}$

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

$Var(X) \Rightarrow \frac{(b-a)^2}{12}$ $\left((a-b)^2 = a^2-2ab+b^2\right)$

$\sim U(a,b)$

Hypothesis testing and statistical analysis

1. Z test
2. t test

] → Average

3. Chi-square → Categorical data
4. Annova → variance

Z test criteria → CLT → $\mu = \mu$, $\sigma = \sigma/\sqrt{n}$ $n \geq 30$

→ S.S ≥ 30

→ σ population should be given

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

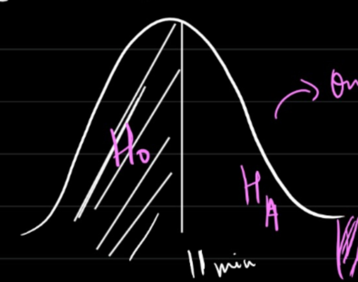
Q. Suppose a child Psychologist says that the average time working mother spend talking their children is upto 11 minutes per day.
To test the hypothesis you conducted an experiment with random sample of 100 working mother and find that they spend 11.5 minutes per day talking with their children. Assume prior research suggests the population standard deviation is 2.3 mins. Conduct the test with 5% level of significance ($\alpha = 0.05$)

Ans

1. frame the hypothesis

$H_0 : \mu \leq 11$

$H_A : \mu > 11$



→ One tail test ↓ rejection region is at one side of the distribution

2. level of significance = 5%

3. Type of test → Z test

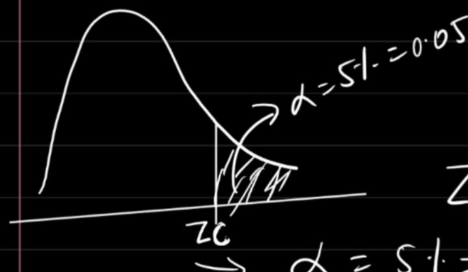
* To identify if its two tail test H_A has $\boxed{\mu \neq 11}$

4. $$Z_{score}\ \bar{x}\ (\text{test statistics}) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

↓

Z statistics

$$= \frac{11.5 - 11}{2.3/\sqrt{100}} = 2.17.$$



Zcritical

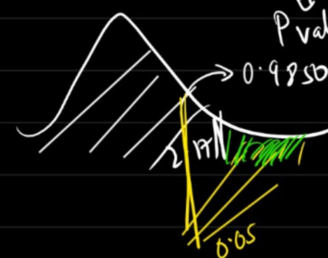
$\rightarrow \alpha = 5\% = 0.05$

$Z_{critical} = 1.64$

p-value

$Z_{statistics} = 2.17$

$P_{value} \rightarrow 0.9850$



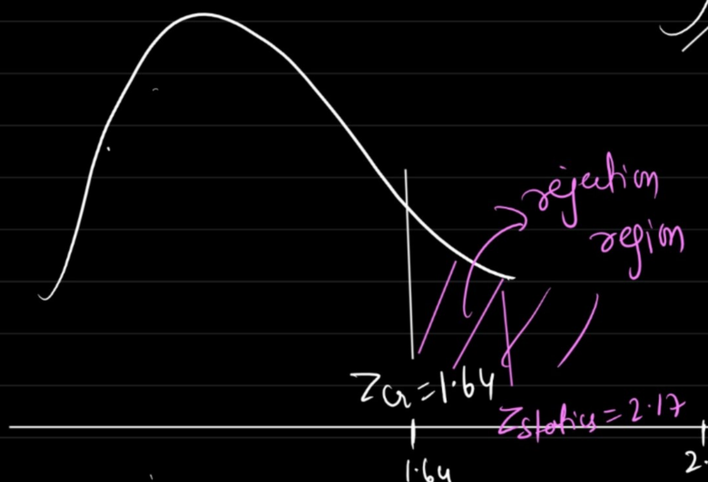
$1 - 0.9850$

$\Rightarrow 0.015$

$p_{value} < 0.05$

$\downarrow$

reject the H_0



if $Z_{statistics}$ lies in Z_{cr} region, you reject the H_0

Step-5 - Conclusion $\rightarrow$ $Z_{statis} > Z_{cr}$

reject the H_0

meaning > 11 mins

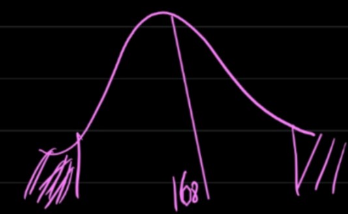
Q. The average heights of all residents in a city is 168 cm with a $\sigma = 3.9$ cm. One researcher believes the mean to be different. He measured the height of 36 individuals and found the average height to be 169.5 cm. Test the hypothesis with 95% confidence interval. $\alpha = 5\%$

Ans Step-1 $H_0 : \mu_{age} = 168$ cm

$H_A : \mu_{age} \neq 168$ cm

Step 2 level of significance $\Rightarrow$ Its a two tail test $\alpha = 5/2 = 2.5\%$ or 0.025

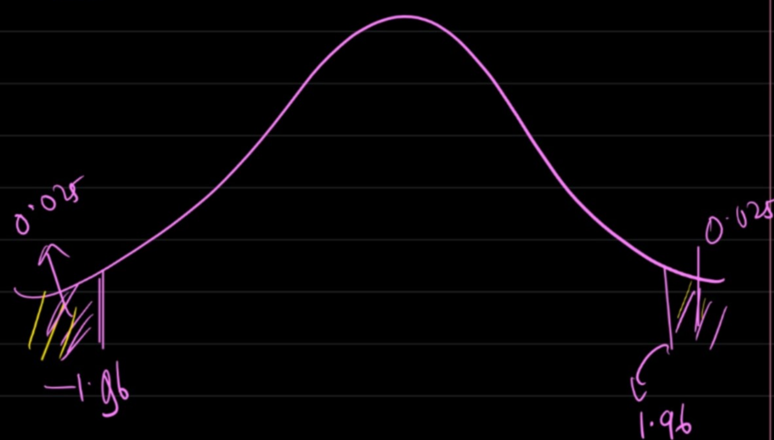
Step-3 type of test — Z test.



Step-4 Z statistics → $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{169.5-168}{3.9/\sqrt{36}}$

Z statistics = 2.31

Step-5 Z critical → 1.96

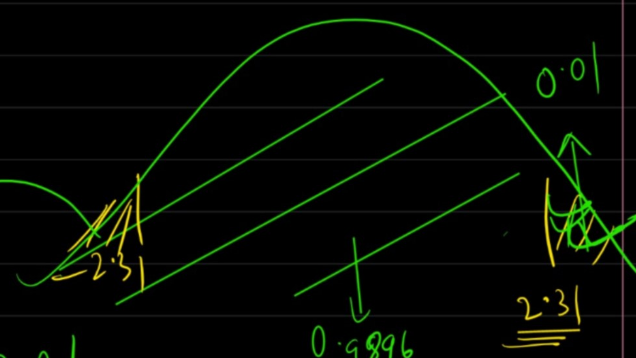


reject the H_0.

Conclusion: the Avg height $\neq$ 168 cm

2nd approach — p value

$\alpha = 5\%$



pvalue = 0.01 + 0.01
$\Rightarrow$ 0.02

1 − 0.9896 $\Rightarrow$ 0.01

pvalue < 0.05
→ Reject the H_0