=> I am having a process whose outcomes are binary and we are repeating the process for n no of times
=> those kind of dist. we say as Binomial dist.

(1) Binomial Distribution :

=> In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes-no question, and each with its own Boolean-valued outcome: success (with prob. p) or failure (with prob. $q = 1-p$). A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment, and a sequence of outcomes is called a Bernoulli process; for a single trial, 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

$n = 10$

Eg: Tossing a coin for 10 times

	1st time	2nd	3rd	4th
$P(T) = p$	p	p	p	p
$P(H) = 1-p$	$1-p$	$1-p$	$1-p$	$1-p$

When we combine this we get Binomial Distribution.

→ Binomial dist. can combine various Bernoulli dist. in it.

Date
Page

Parameters $n \in \{0, 1, 2, 3, \ldots\}$ → no. of trials or experiment.

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

$q = 1 - p$

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

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

(3) Mean of Binomial distribution

Mean $= np$ (n → no. of experiments, p → prob)

(4) Variance And std

Variance $= npq$

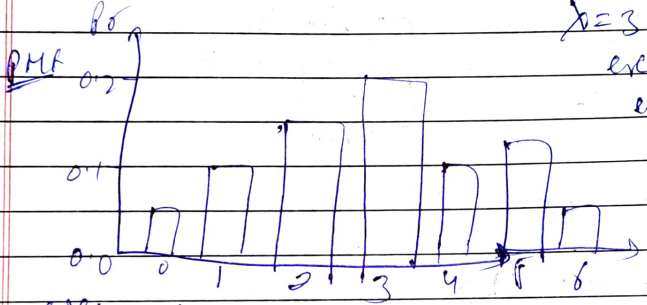
std $= \sqrt{npq}$

Poisson Distribution :

① Discrete Distribution (PMF)
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

parameter
$\lambda = 3$ ⇒ Expected no. of event to occur at every time interval



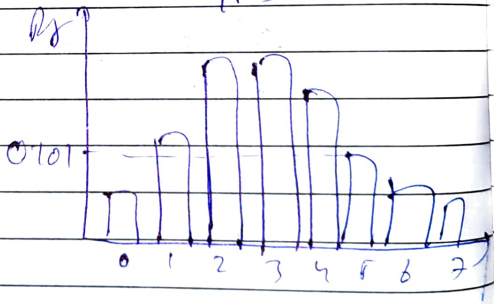
[λ helps to create the PMF curve in poisson Dist.]

PMF

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

$$\Rightarrow \frac{e^{-3} 3^5}{5!} = 0.101$$

$$\Rightarrow 10.1\%$$



[The probability of an event to occur at the 5th hour]

$P(X=4) + P(X=5) =$ ____

Mean and variance :

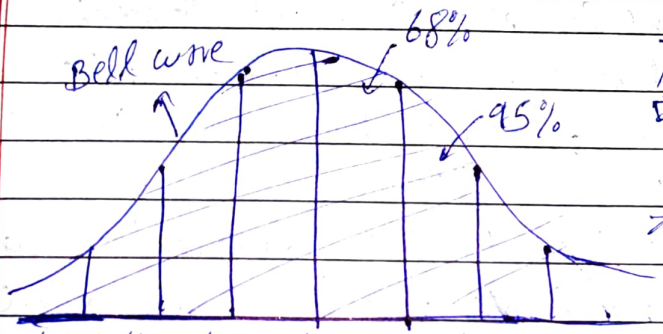
Mean $\Rightarrow E(X) = \lambda t$

Variance $\Rightarrow E(X) = \lambda t$

λ = expected no. of events to occur at every time interval.

t = time interval.

Normal / Gaussian distribution :



$X = \{$ $\}$

If X follows a normal dist then we get this curve. (Bell curve)

$\Rightarrow$ Symmetric Distribution
(50% each side)

If let say we have a random variable that follows normal / gau dist then 68% of the entire data falls on first std. assumption

1 std. to the right
2 " " " , etc

68 95% 99.7

$X = \{ \text{---------} \}$

Empirical Rule [3 Sigma Rule]

(with the mean, around 1 std deviation from the mean, around 68% of the data falls here) i.e blw $\mu-\sigma$ to $\mu+\sigma$

68 - 95 - 99.7%
1st std, 2nd std, 3 std.

Q-Q plot : whether a Distribution is Gaussian/Normal Distribution.

[Quantile - quantile]

How do we find out whether the dist. is gaussian or not?

Probability :

$Pr(\mu-\sigma \leq X \leq \mu+\sigma) \simeq 68\%$

$Pr(\mu-2\sigma \leq X \leq \mu+2\sigma) \simeq 95\%$

$Pr(\mu-3\sigma \leq X \leq \mu+3\sigma) \simeq 99.7\%$

Eg: weight, height, IRIS DATASET [Domain Experts]

Date
Page

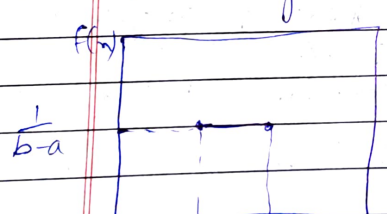
Uniform Distribution :

1. Continuous uniform distribution (pdf)
2. Discrete uniform Distribution (PMF)

1. Continuous uniform distribution : [Continuous Random variables]

In probability theory and statistics, the continuous uniform distribution or rectangular distribution is a family of symmetric probability distributions. The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds.
The bounds are defined by the parameters, a and b, which are the minimum and maximum values. The interval can either be closed (e.g. $[a,b]$ or open (eg. (a,b))

Uniform
prob. density fun^n



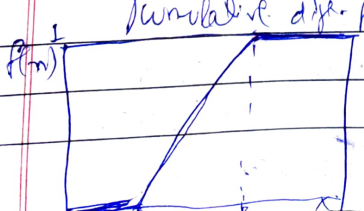
Notation : $U(a,b)$

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

$$Pdf = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

using Maximum convention

Cumulative dist. fun^n



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

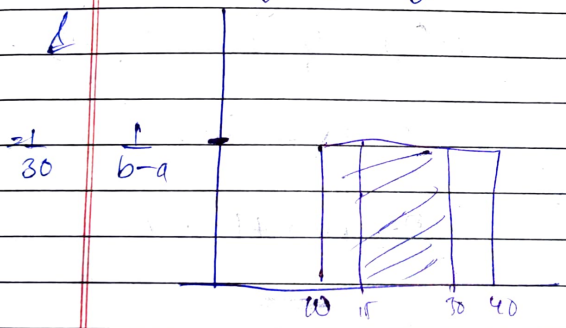
Mean $= \frac{1}{2}(a+b) = E(x)$

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

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

Eg: The no. of candies sold daily at a shop is uniformly distributed with a minimum of 10 and maximum of 40.

a) probability of daily sales to fall b/w 15 and 30



$P(15 \le x \le 30)$

$= (30-15) \times \frac{1}{30}$

$\Rightarrow \frac{15}{30} = 0.5$

$P(x \ge 20) = (40-20) \times \frac{1}{30}$

$\Rightarrow \frac{20}{30} = 0.66 = 66.66\%$

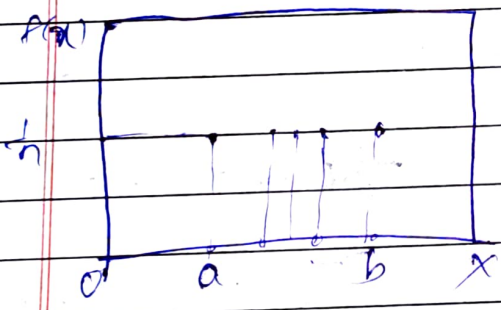
$\int_{x_1}^{x_2} (x_2 - x_1) \times \frac{1}{b-a}$

Date
Page

6 Discrete uniform Distribution { Discrete Random variable }

In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution where in a finite no. of values are equally likely to be observed; every one of n value has equal probability 1/n.
Another way of saying "discrete uniform distribution would be "a known, finite no. of outcomes equally likely to happen".

discrete uniform prob. mass fun^n



n = 5 above n = b - a + 1

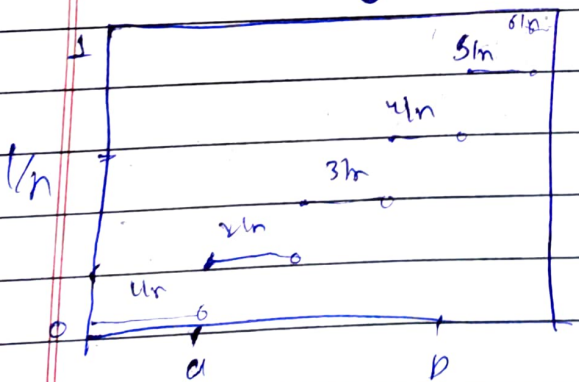
e.g. Rolling a dice {1,2,3,4,5,6}

P(1) = 1/6 P(2) = 1/6 P(3) = 1/6

a = 1, b = 6

n = b - a + 1 = 6

Cumulative dist. fun^n



Notation U(a, b)

parameters a and b b ≥ a

PMF $\frac{1}{n}$

$$\left.\begin{matrix}\text{Mean}\\ \text{Median}\end{matrix}\right\} = \frac{a+b}{2}$$

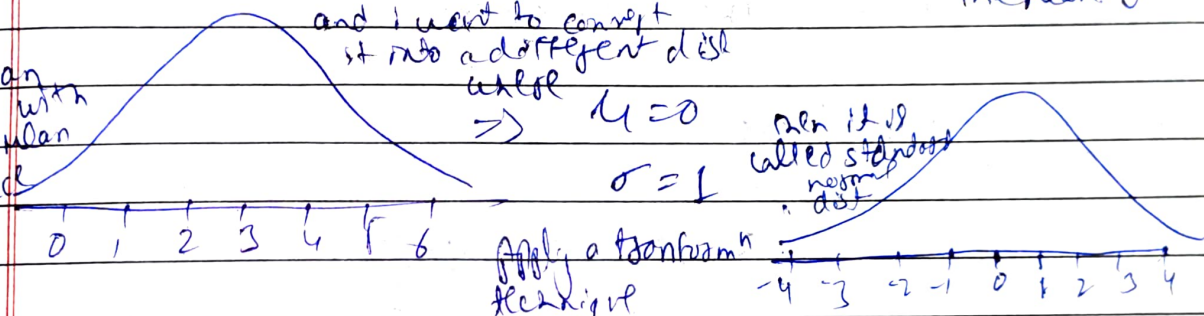
In uniform Distribution the prob. of getting the output is Equal.

Standard Normal Distribution & z-score

$X = \{1, 2, 3, 4, 5\}$ $\mu = 3$

$\sigma = 1.414 \approx 1$

z-score helps to identify that any value how much std is away from the mean.



$X = \{1, 2, 3, 4, 5\}$ $\quad z\text{-score} = \frac{x_i - \mu}{\sigma}$ $\quad y = \{-2, -1, 0, 1, 2\}$

(1) $\frac{1-3}{1} = -2$ (3) $\frac{3-3}{1} = 0$

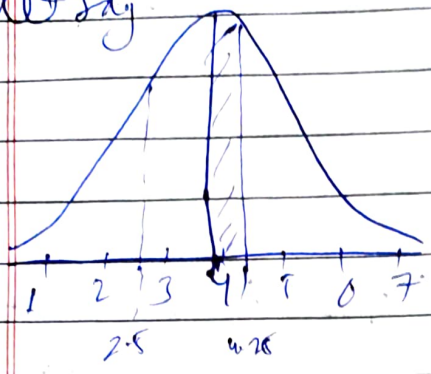
(2) $\frac{2-3}{1} = -1$ (4) $\frac{4-3}{1} = 1$

$X \sim SND(\mu = 0, \sigma = 1)$

z-score helps to ...
z-score helps to apply standardization

Date
Page

Eg: 101 Say



$\mu = 4$

$\sigma = 1$

z-Score importance

Q. How many standard deviation 4.25 is away from the Mean?

$x_i = 4.25$

$$z\text{-score} = \frac{4.25 - 4}{1} = 0.25$$

$x_i = 2.5$

$$z\text{-score} = \frac{2.5 - 4}{1} = -1.5$$

Eg: Dataset

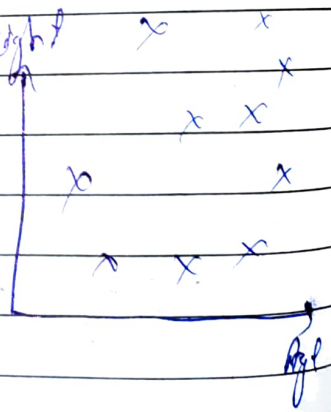
Age	weight	height	Salary
years	kg	cm's	INR or $
24	70	175	45K
25	60	160	50K
26	55	150	60K
27	40	130	30K
30	30	175	20K
31	25	180	70K

units are different here

when we solve ML problem, we bring all the units from same unit scale

For ease to bring it into same unit scale

We use a technique called as standardization i.e. z-score



1. clustering algorithm
2. linear Regression
3. logistic Regression

$$z\text{-score} = \frac{x_i - \mu_{age}}{\sigma} \quad \frac{x_i - \mu_{weight}}{\sigma} \quad \frac{x_i - \mu_{height}}{\sigma} \quad \frac{x_i - \mu_{salary}}{\sigma}$$

Standardization → ML models

[To bring accuracy in models]

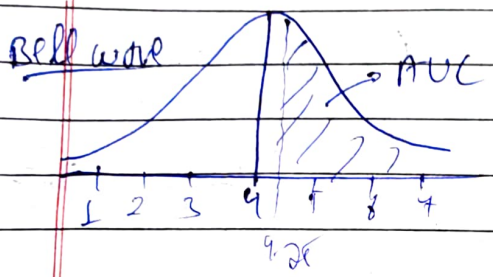
z-table [+ve z-table & -ve z-table]

Problem statement on z-score [z table]

$X = \{1, 2, 3, 4, 5, 6, 7\}$ $\quad$ z-score $= \frac{x_i - \mu}{\sigma}$

let, $\mu = 4$ $\quad \sigma = 1$ [using these we can construct the curve]

Bell curve



[Area under curve]

Total area = 1 => Symmetric Distribution

z-table

Q what percentage of scores fall above 4.25?

$x_i = 4.25$

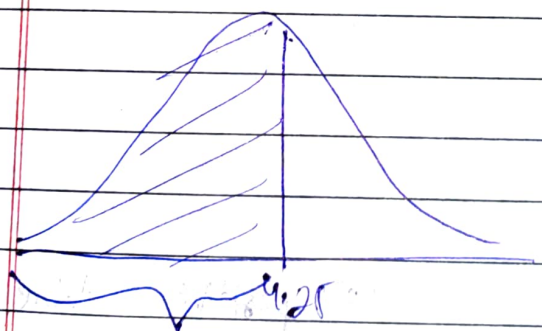
$\mu = 4$

$\sigma = 1$

z-score $= \frac{x_i - \mu}{\sigma}$

$= \frac{4.25 - 4}{1} = 0.25$ => std

Search z-table



0.59871

but we want Right area [or above curve]

Q what % of scores falls above 4.25?

=> using z-table;

$1 - 0.59871$

=> 0.4013

=> 40.13%

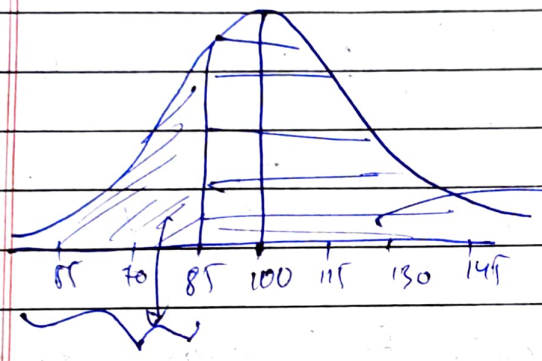
[In -ve z-table values are ...]

Q In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population would you expect to have an IQ lower than 85?

A $\mu = 100 \quad \sigma = 15$

(1) z-score $= \frac{x_i - \mu}{\sigma}$

$\Rightarrow \frac{85 - 100}{15} \Rightarrow -1$



0.15866 = 15.88%

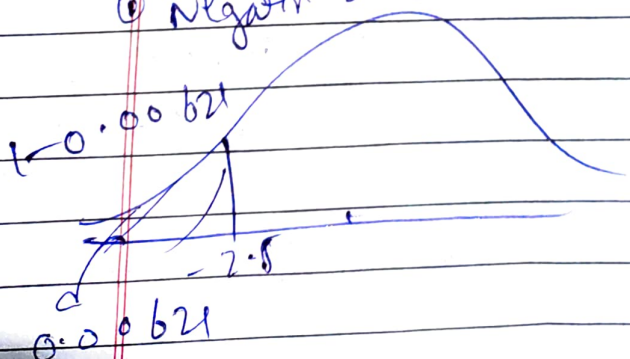
(2) $IQ \geq 85$

$\Rightarrow 1 - 0.15866$

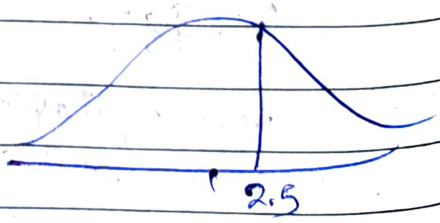
$\Rightarrow 84.13\%$

$75 \geq IQ \leq 100$ → Internal Assignment

(1) Negative z table



(2) Positive z-table



→ It is imp. to understand some concepts like Sampling distribution or the properties of sampling dist.
→ to calculate prob.

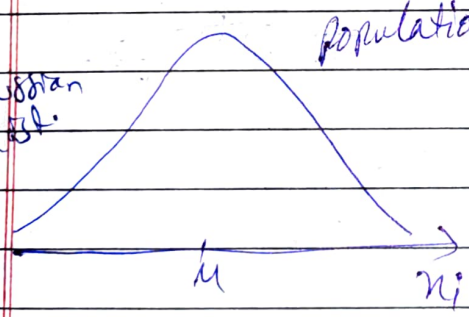
Date
Page

Central limit theorem

The central limit theorem relies on the concept of a sampling distribution, which is the probability distribution of a statistic for a large no. of samples taken from population.

The central limit theorem says that the sampling distribution of the mean will always be normally distributed as long as the sample size is large enough. Regardless of whether the population has a normal, poisson, binomial or any other distribution, the sampling distribution of the mean will be normal

(1) $X \simeq N(\mu, \sigma)$



n = sample size → any value

sample data

$S_1 = \{x_1, x_2, x_4, \ldots x_n\} = \bar{x}_1$

$S_2 = \{x_2, x_3 \ldots x_n\} = \bar{x}_2$

S_2 — — $\bar{x}_3$

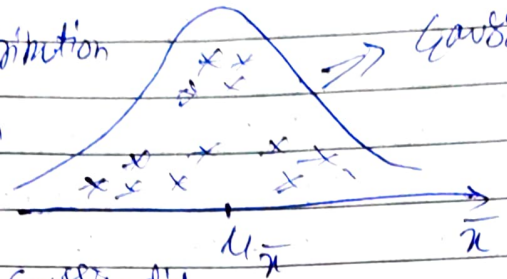
S_4

S_r $\bar{x}_m$

⇓

CLT

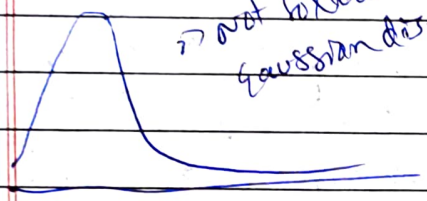
Sampling distribution of the mean



⇒ Gaussian/Normal distribution

⇒ It will follow Gaussian dist

① $X \not\sim N(\mu, \sigma)$ ⇒ $n \geq 30$ ⇒ Sample size



⇒ not follow Gaussian dist.

s_1 $\bar{x}_1$

s_2 $\bar{x}_2$

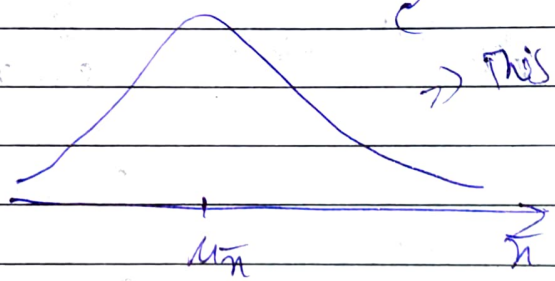
⋮ ⋮

$\bar{x}_m$

↓

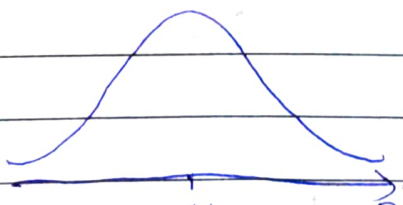
CLT

↓



⇒ This follows Gaussian dist

② Normal Distribution



$X \approx N(\mu, \sigma)$

↓

rmal
tion

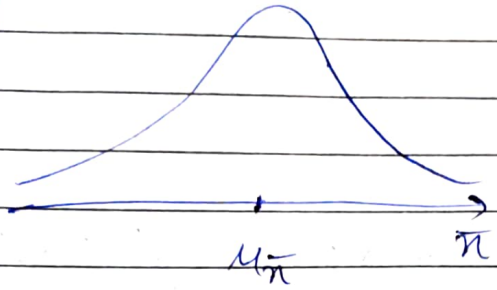
Sample size

$\bar{x}_1$

$\bar{x}_2$

$\bar{x}_m$

Sampling distribution of Mean



σ = population Std
μ = population Mean
n = sample size

$$\bar{X} \simeq N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

⇒ Inferential Statistics

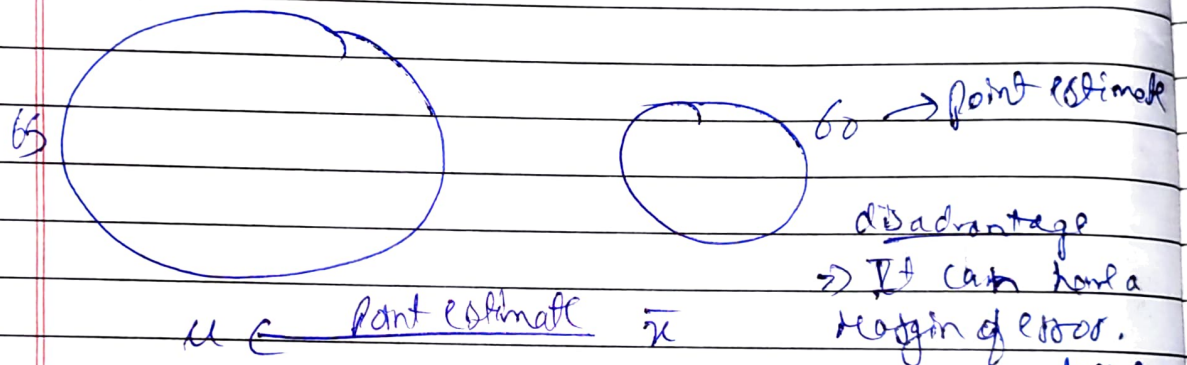
Isse pehle descriptive stats tha

Estimate : It is a specified observed numerical value used to estimate an unknown population parameter.

Types :

(1) **Point estimate** : Single numerical value used to estimate an unknown population parameter

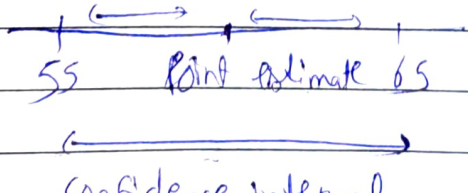
Ex : Sample mean is a point estimate of a population mean



disadvantage
⇒ It can have a margin of error.
⇒ So we can define interval.

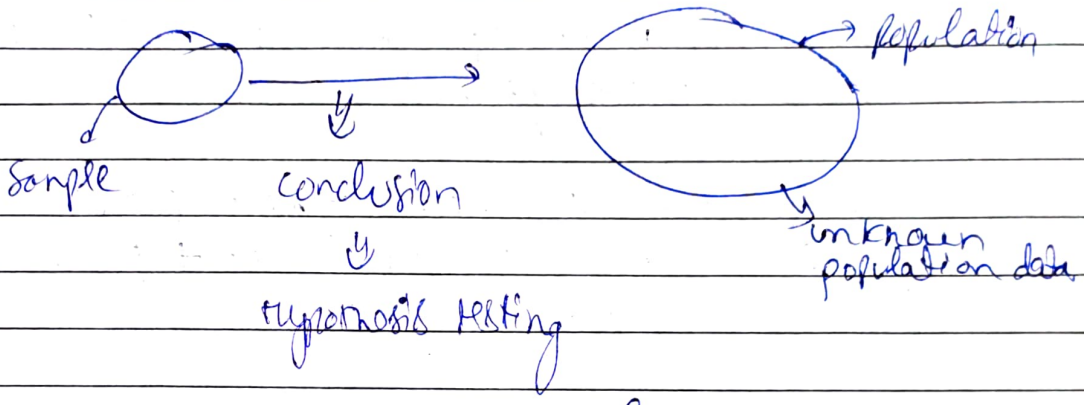
(2) **Interval estimate** : Range of values used to estimate the unknown population parameter

55 – 65 ⇒ Sample mean



A Hypothesis And Hypothesis Testing Mechanism

Inferential Stats → conclusion or inference



Hypothesis testing Mechanism

Eg: person → crime → court
not guilty until proven

(1) Null Hypothesis (H_0) — person is not guilty
- The assumption you are beginning with.

(2) Alternate Hypothesis (H_1) - The person is guilty
- opposite of Null Hypothesis

(3) Experiments → Statistical Analysis
→ collect proof (DNA, Fingers Test)
(Z-test, t-test, CHI)
{p value, significance value}

(4) Accept the Null Hypothesis or Reject the Null Hypothesis

Eg: Colleges at District A states its average passed percentage of students are 85%. A new college is opened in the district and it was found that a sample of student 100 have a pass percentage of 90% with a standard deviation of 4%.
Does this college have a different passed percentage.

Null hypothesis (H_0) = $\mu = 85\%$
Alternate hypothesis (H_1) = $\mu \neq 85\%$

It helps us to draw some conclusions about the population data.

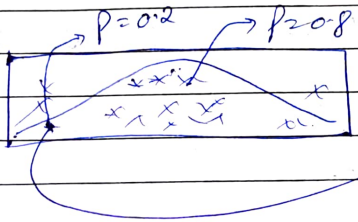
→ It is used in inferential statistics.

Date
Page

p value

The p value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observation if the null hypothesis were true.

p values are used in hypothesis testing to help decide whether to reject the null hypothesis.



Eg: Out of 100 touches, we touch around 20 times in this region

Hypothesis testing: Eg: coin is fair or not {100 times}

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

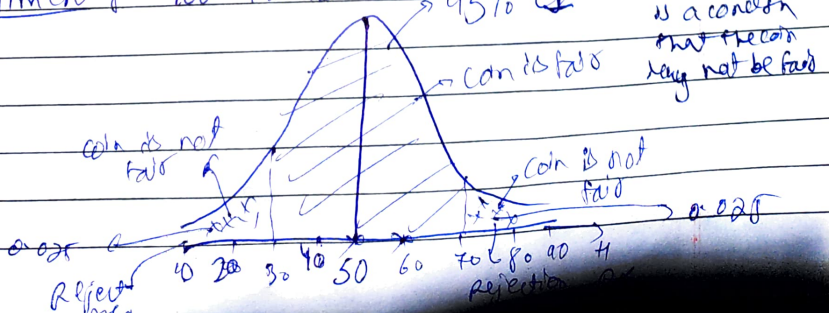
$P(H) = 0.6$ $P(T) = 0.4$

$P(H) = 0.7$ $P(T) = 0.3$

(1) Null hypothesis:
H_0 : coin is fair

(2) Alternate hypothesis:
H_1 : coin is not fair.

(3) Experiment: 100 times



If it goes above to take that is a condition that the coin may not be fair

Domain expert decide

(4) Significance value : $\alpha = 0.05$

$CI = 1 - 0.05 = 0.95$

(5) Conclusion : $P < \text{significance}$

Reject the null hypothesis
else
fail to Reject the Null hypothesis

Output is the p value &.

$P <$ Significance value
Then Reject. the null hypo.

$P >$ Sig.
Then Accept or fail to Reject the null hypo

Hypothesis testing and Statistical Analysis

(a) Z-test } Dealing with Average data → Average → z table → z score and p value
(b) t test }
(c) CHI SQUARE => categorical Data
(d) ANNOVA => variance of data

Z-test :

conditions
i) population std ii) $n \geq 30$.

(1) The average heights of all residents in a city is 168cm with a $\sigma = 3.9$. A doctor believes the mean to be different. He measured the height of 36 individual and found the average heights to be 169.5cm.

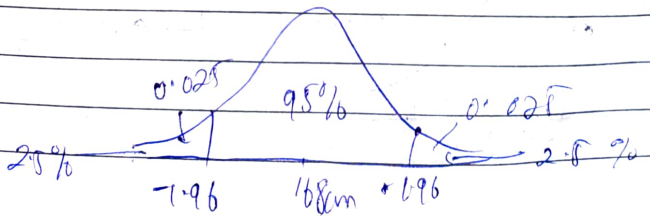
(a) state null and alternate hypothesis
(b) At a 95% confidence level, is there enough evidence to reject the null hypothesis

$\mu = 168cm \quad \sigma = 3.9 \quad n = 36 \quad \bar{x} = 169.5$
$C.I = 0.95 \quad \alpha = 1 - C.I = 1 - 0.95 = 0.05$

(a) Null hypothesis $H_0 = \mu = 168cm$

(b) Alternate hypothesis $H_1 = \mu \neq 168cm$.

(c) Based on C.I we will draw Decision boundary



2 Tail test
=> Can be greater than or less than
i.e. in both direction of 168

remaining left-side area

$\Rightarrow 1 - 0.025 = 0.9750$ $\Rightarrow$ z-score

$\Downarrow$

Area $= \pm 1.96$

Conclusion If z is less than -1.96 or greater than $+1.96$,

Reject the null hypothesis.

(x) z-test

$$z_d = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$ [for sample data]

$$z\text{-score} = \frac{\bar{x} - \mu}{\sigma}$$

$\Downarrow$

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

$$\Rightarrow \frac{169.5 - 168}{3.9/\sqrt{36}}$$

$$z_d = \frac{1.5}{0.65} = \boxed{2.31}$$

Conclusion

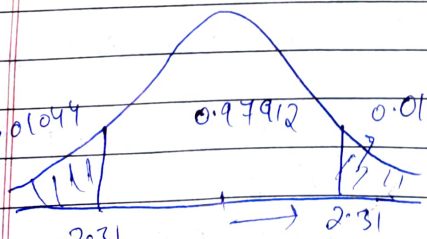
$2.31 > 1.96$ Reject the null hypothesis

$p < 0.05$

2.31

$\Downarrow$ z-table

0.98956



$1 - 0.98956 = 0$

(x) final conclusion the Average ± 188cm
Dd average height stems to increasing based on sample heights

(1) p value = 0.01044 + 0.01044
= 0.02088

$p < 0.05$

$0.02088 < 0.05 \Rightarrow$ Reject the null hypothesis

(2) A factory manufactures bulbs with a average warranty of 5 years with standard deviation of 0.50. A worker believes that the bulb will malfunction in less than 5 years. He tests a sample of 40 bulbs and find the average time to be 4.8 years

(a) state null and alternate hypothesis

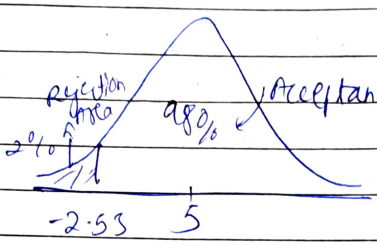
(b) At a 2% significance level, is there enough evidence to support the idea that the warranty should be revised ?

$\mu = 5 \quad \sigma = 0.50 \quad n = 40 \quad \bar{x} = 4.8$

(a) Null hypothesis $H_0 : \mu = 5$

Alternate hypothesis $H_1 : \mu < 5$ {1 tail test}

(b) Decision Boundary



(c) z-test:

$$z_1 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.8 - 5}{0.50/\sqrt{40}}$$

$$= -2.53$$

Area under curve with z-score $-2.53 = 0.057$

p-value = 0.057

$\alpha = 0.02$

compare p-value with α

$0.057 < 0.02 \rightarrow$ false

we accept the Null Hypothesis

we fail to Reject the Null Hypothesis

Student t distribution :

In z stats when we perform any analysis using z score we require σ (population standard deviation) → is already known

★ How do we perform any analysis when we don't know the population standard deviation?

↓

Student's t distribution — t-stats

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

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

s = sample standard deviation

0.05 % → z table

t table ⇒ t test

★ Degree of freedom

$$dof = n - 1 \Rightarrow 3 - 1 = 2$$

Eg let consider
3 people & three chairs in a room

when the first person comes inside a room, he/she has 3 diff. options, then after this 2nd person comes it has 2 options
but the third person doesn't have any option
so degree of freedom is 2

T-Stats ÷ T Test → One Sample t-test

(1) In the population the average IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all.
A sample of 30 participants who have taken the medication has a mean of 140 with a standard deviation of 20.
Did the Medication affect Intelligence?
CI = 95% α = 0.05

(2) $\mu = 100$ $n = 30$ $\bar{x} = 140$ $S = 20$
C.I = 95% $\alpha = 0.05$

(3) Null hypothesis : $H_0 \div \mu = 100$

(4) Alternate hypothesis : $H_1 \Rightarrow \mu \neq 100$ { 2 tail test

(5) $\alpha = 0.05$

(6) Degree of freedom

$df = n - 1 = 30 - 1 = 29$

(7) Decision Rule