Statistics

collection, organising, collecting the data (statistical operation)

Example


→ person (Appearing for Exam)

$\mu_E = 85$ < English - 88/100

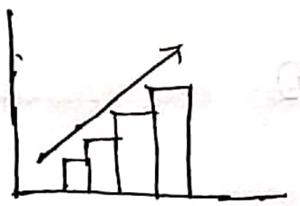
$\mu_m = 98$ > maths - 95/100

Types

Descriptive statistics

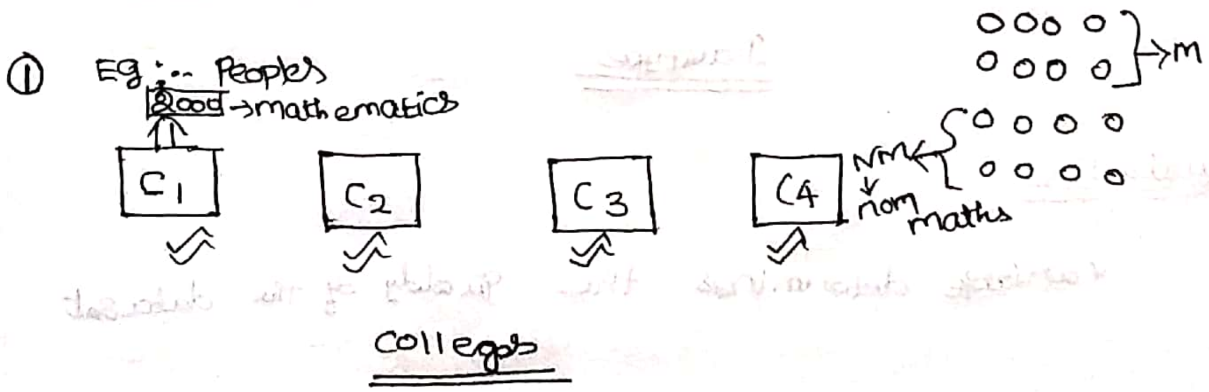
Collecting, organising and representing the data in a better way

Eg:-

covid-19 ⇒ 

Inferential statistics

① Eg:- Peoples



② Population

True population

Sample

Doing calculation by collecting demo from the various groups.

Variable in Statistics

Based on Descriptive Stats

Categorical

* Rating → 1, 2, 3, 4, 5 [Eg:- Zomato Delivery]
* Decision → yes or no [Switching towards Data -Science → yes or no]

Numerical

* Eg:- 200.91kg (weight)

Level of measurement

Datatype

Qualitative

Variable determines the quality of the data set

(i) Nominal :-

categorize the data based on the name

Eg:- Season :- winter, Summer like that
1, 2, 3, 4, 5

ordinal :- (ii)

ok, good, bad, Excellent, Amazing (order pair)

Quantitative

Describe the Quantity of the dataset

(i) Discrete (It ~~doesn't~~ give the absolute number)

Eg:- 1, 2, 3, 4, 5

(ii) Continuous (It doesn't give the absolute number)

Eg:- 75.5234110 (weight)

① Mean

Sample

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Population

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Eg:-

① 3, 9, 5, 10, 10, 1, 7, 8, 5, 5

$$\frac{3+9+5+10+10+1+7+8+5+5}{10}$$

$= 6.30$

② median

Eg:-

1, 5, 5, 5, 7, 8, 9, 10, 10 (Arrange in ascending order)

$$\frac{7+5}{2} = 6$$

③ mode :-

frequency of highest occurance of the data

Eg:- (we can take mode upto 2 and 3 only)

3, 9, 5, 10, 10, 1, 7, 8, 5, 5

5 → 3 (occurance of 5 will be more)

10 → 2

Skewness

Eg::

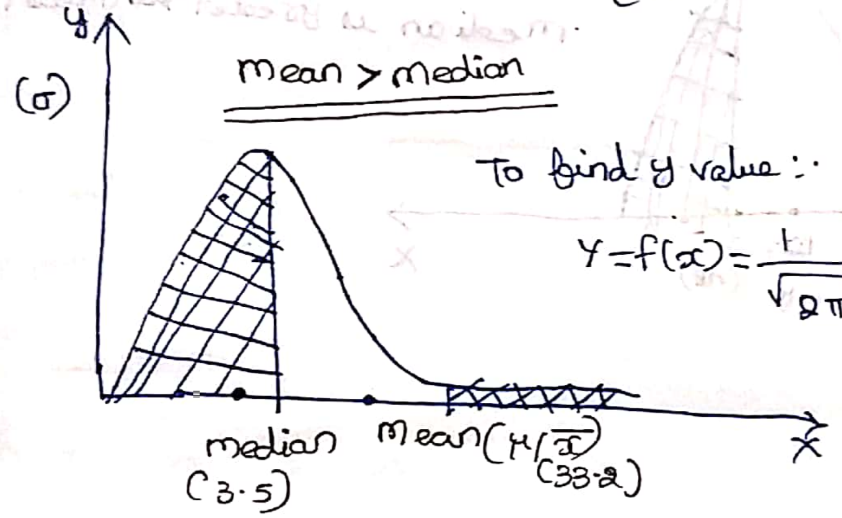
D = 1, 2, 3, 4, 1, 2, 8, $\boxed{100, 101, 110}$ → frequency is high

$\bar{x}$ (Sample) = 33.2 (outliers in Positive side)

median = 3.5

Data in Ascending order::

1, 1, 2, 2, 3, 4, 8, 100, 101, 110 [Mean is greater than medium]

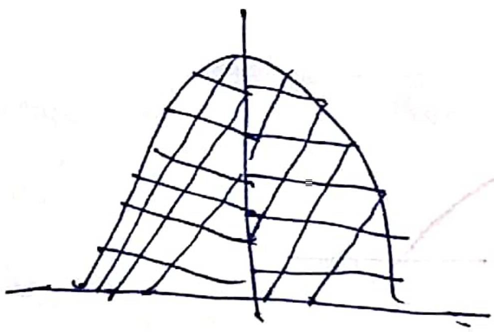


To find y value::

$$Y = f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

Positive skewness

Normal Distribution



Negative skewness [outliers in negative side]

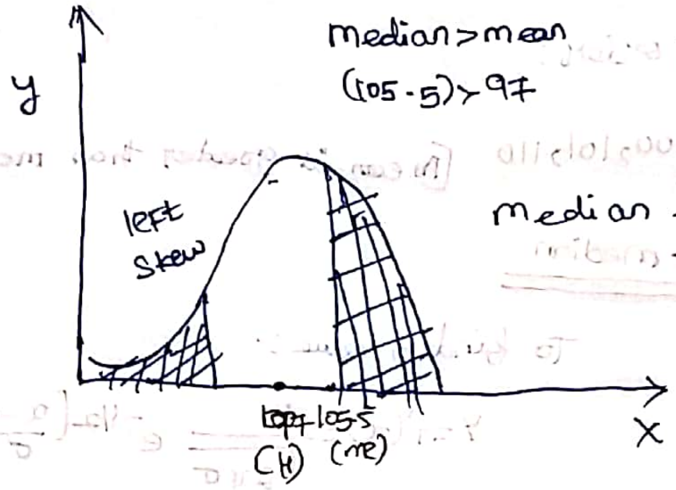
D = 2, 3, 4, 100, 101, 110, 200, 150, 160, 140, 110

$\bar{x}$ = 97 (mean)

median = 105.5

median > mean

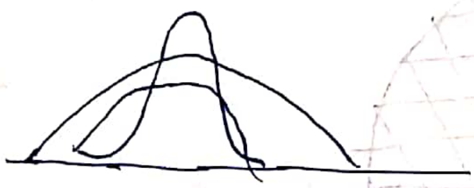
(105.5) > 97



median is greater than mean

Measure of dispersion

① Range :-



Eg:- laptop price varies from 60k to 99k

dispersion

1, 3, 100, 110, 101, 2, 8, 100

= 109

(Summation of data From mean)

② Variance or Standard deviation

Population Dataset :- (Variation of datas From the actual mean)

(4, 5, 10, 8, 2, 10), 100, 110

↳ $\mu = 6.5$

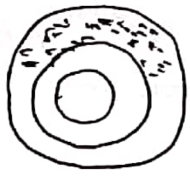
$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$= \frac{(4-6.5)^2 + (5-4.5)^2 + (10-6.5)^2 + (8-6.5)^2 + (2-6.5)^2 + (10-6.5)^2}{6}$$

$\sigma^2 = 11.1$

* IF variance is high then this will happen,

Eg :-



[Driver is not steady, he will apply many brakes]

[0 - 140]

* IF variance is less then this will happen

Eg :-

[1, 1, 1, 1,] → no distance

opposite

Variance :

Standard deviation (It is used to check the Fluctuation - of the data)

$$\sigma = \sqrt{\left(\sum_{i=1}^{N} \frac{x-\mu}{N}\right)}$$

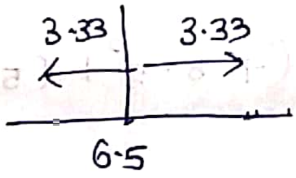
absolute mean
$\downarrow$
Standard deviation

4, 5, 10, 8, 2, 10

$\mu = 6.5$

$\sigma^2 = 11.1$ (variance)

$\sigma = 3.33$ (Standard deviation)



Eg:-

4, 5, 10, 8, 2, 10

6.5 ± 3.33

(i) $6.5 + 3.33 = 9.83$

(ii) $6.5 - 3.33 = 3.17$

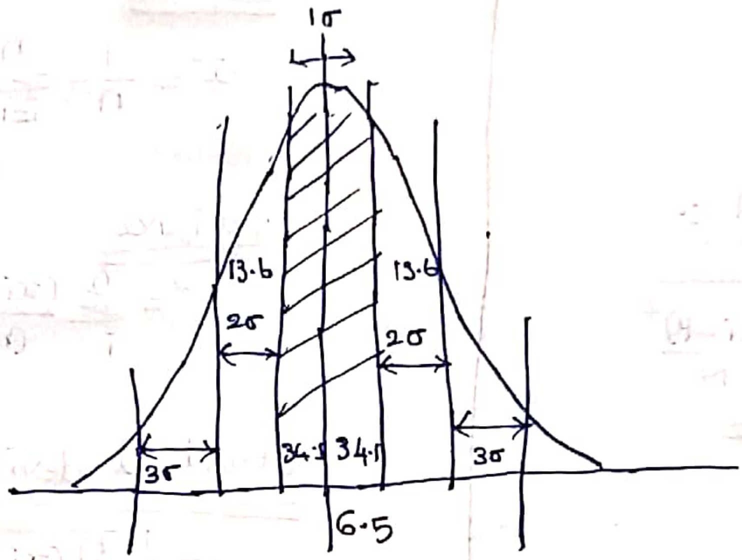
$2 \leftarrow$ Far $\quad 3.17 \leq J \leq 9.83 \quad \rightarrow 10$ Far

$$\left.\begin{matrix} 1\sigma \\ 2\sigma \\ 3\sigma \end{matrix}\right. \begin{matrix} \rightarrow 68\% \\ \rightarrow 95\% \\ \rightarrow 99\% \end{matrix}$$

working of Standard deviation

Diagram

D= 4, 5, 10, 8, 20



1st case:-

$6.5 \pm 3.33 \; (\mu \pm \sigma)$

68 %.

2nd case:-

$6.5 \pm 2 \times 3.33 \; (\mu \pm 2 \times \sigma)$

95 %.

3rd case:-

$6.5 \pm 3 \times 3.33 \; (\mu \pm 3 \times \sigma)$

99.7 %.

Population	Sample
Mean :-	**Mean** :-
$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Standard deviation :-	**Variance**
$\sigma^2 = \sqrt{\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N}}$ Variance	$s^2 = \sum_{i}^{n} \frac{(x_i-\bar{x})^2}{(n-1)}$
S.d $= \sqrt{\sigma^2}$	**Standard deviation**
	$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{n-1}}$

Probablity Density Function

Probablity of occurance of the particular outcome

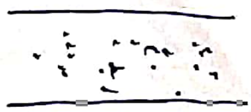
$\uparrow$

Binomial Distribution

Eg :-

(i) 1, 2, 3, 4, 5, 6 (1-Dice)

(ii) H-T (coin)

(iii)  $\left\{\begin{matrix} \text{EMS} \\ \text{Fire} \end{matrix}\right\}$ (How many vehicle that we can deploy)

9-9

PDF

- Normal Distribution
- Bernoullis Distribution
- Binomial Distribution
- uniform Distribution
- Student T Distribution
- Poisson Distribution

Binomial Distribution

Formula :-

$$P(n/N) = \binom{N}{n} p^n q^{N-n}$$

$$= \left(\frac{N!}{n!\,(N-n)!}\right) p^n (1-p)^{N-n}$$

$n \to$ Successes

$N \to$ trail

Eg

① 100 – 75 Die (Population)
 cancer

(6 Person) – (4 recover) (sample)

$N = 6$ (trail)

$n = 4$ (Success)

P(recovery) = 0.25

$q = 1 - p$ (Failure) = 0.75

$$= \frac{6!}{4!(6-2)!} (0.25)^4 (0.75)^{6-4}$$

$$= \frac{6!}{4! \cdot 4!} (0.25)^4 (0.75)^2$$

$= 0.032959$

(2) Die = 3 times N = trail
n = Success

(i) P_5 (Probablity of not getting any Five)

(ii) P_5 (probablity of getting only Five)

(iii) P_5 (probablity of getting Five all the time)

(i)

$N = 3$

$n = 0$

$\frac{3!}{3!} (1/6)^3 (5/6)^3$

$P_5 = (1/6)$ (Success)

$P_5' = 1 - \frac{1}{6} (5/6)$ (Failure)

$P(n=0) = C_n^N p^n q^{N-n}$

$= 3C_0 (1/6)^0 (\frac{5}{6})^{3-0}$

$= 0.57$

$P(n=1) = 3C_1 (1/6)^1 (5/6)^{3-1}$ N = 3 (trail)
n = 1 (success)

$= 3C_1 (1/6) \times (5/6)^2$

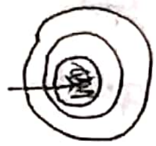
$= 0.3477$

$P(n=3) = {}^{3}C_{3} \left(1/6\right)^{3} \left(5/6\right)^{5-3}$

$= 4.62 \times 10^{-3}$ $\quad N = 3$ (trail)

$n = 3$ (Success)

$= 0.00462$

3)  (5 trail) (4 hit) 

i) Probablity of getting (hit) more than 2 times when he hit 4 times

$N = \cancel{5}$ (trail) (4) $\quad N = 5$ (trail)

$n = \cancel{4}$ (Success) (3)

Probablity of getting Success (Ps) $= \frac{4}{5}$

$= 0.8$

Probablity of getting failure (PF) $= 1 - \frac{4}{5}$

$= \frac{5-4}{5}$

$= \frac{1}{5}$

$= 0.2$

$N = 4$ (trail)

$n = 4$ (Success)

$P(T) = P(3) + P(4)$

$$= \left(\frac{4!}{3!(4-3)!}\right)(0.8)^{3}(0.2)^{4-3} + \left(\frac{4!}{4!(4-4)!}\right)(0.8)^{4}(0.2)^{(4-4)}$$

$= 0.8192$

(ii) Probablity of getting atleast 3 times going to fail

(Failure) 3 miss 1 hit (Success) (i)

(Failure) 4 miss 0 hit (Success) (ii)

↓

(i) $P(T) = P(T3) + P(T4)$

(i) $(PT3) = {}_NC^n \; p^n \; q^{N-n}$

$N \to$ trail (4)

$n \to$ Success (1)

$(PT3) = 4C^1 \; (0.8)^1 \; (0.2)^3$

$= \left(\frac{4!}{1!\,(4-1)!}\right) (0.8)^1 (0.2)^3$

(ii)

$(PT4) = {}_NC^n \; p^n \; q^{N-n}$

$N \to$ trail (4)

$n \to$ Success (0)

$= 4C^0 \; p^0 \; q^{4-0}$

$= \frac{4!}{0!\,(4-0)!} (0.8)^0 \; (0.2)^4$

$P(T) = \left(\frac{4!}{1!\,(4-1)!}\right) (0.8)^1 \; (0.2)^3 + \left(\frac{4!}{0!\,(4-0)!}\right) (0.8)^1 \; (0.2)^3$

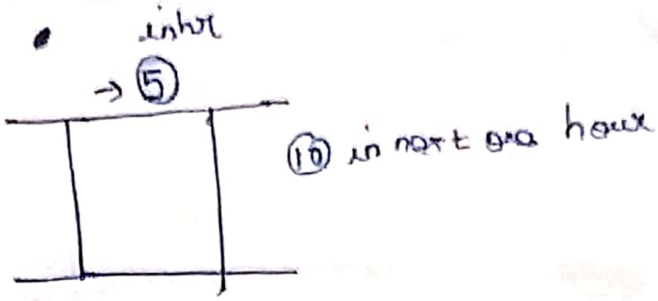
$P(T) = 0.02$

Scanned with CamScanner

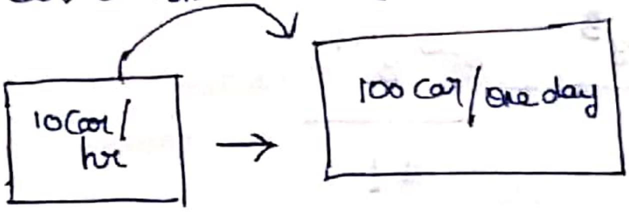
Poisson Distribution

eg :-

①



② In car wash company



Q-A

① Sales man = 3 insurance /week [find the Probablity of not getting any insurance]

$\mu = 3$

Probablity of selling insurance is (1) [1 − P(x0)] (1 insurance)

Probablity of not selling any insurance is (0) (P(x0)) (0 insurance)

$$P(x0) = \frac{e^{-\mu} \mu^{x}}{x!}$$

$$P(x0) = \frac{e^{-3}\, 3^{0}}{0!}$$

$= 0.0497$

$= 1 - P(x0)$

$= 1 - 0.0497$

$= 0.95031$

② Probablity of Selling 2 or more than 2 insurance but less than 5 insurance

$P(2 \leq x < 5)$

$P(2 \leq x < 5)$

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

$$= \frac{e^{-3} 3^2}{2!} + \frac{e^{-3} 3^3}{3!} + \frac{e^{-3} 3^4}{4!}$$

$= 0.61$

③ 100 → 1.5 mistake (average)

100 pages → New Book

(i) Probablity there won't be any mistake

$$P(0) = \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-1.5} (1.5)^0}{0!} = 0.223$$

④ 400 (page) and 0 - mistake)

$$\boxed{P(0) = \frac{e^{-6} (6)^0}{0!}}$$

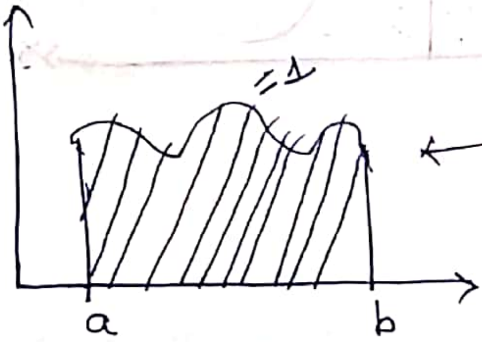
Scanned with CamScanner

Normal Distribution

Probablity Density Function

(Here $\mu = ?$ and $\sigma = ?$)

$f(x) \geq 0 = y$ $\quad a \leq x \leq b$



Here (y) value can be choosen from a and b but it is positive

Normal Distribution

Eg:-

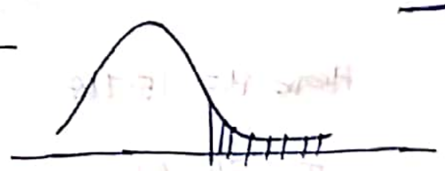
Data = 1, 2, 3, 4, 1, 2, 100, 110

$\sqrt{1}$ → $\sqrt{1}$ → ①

$\sqrt{100}$ → $\sqrt{10}$ → 3.33 → Applying square rooting

↓ log(1)

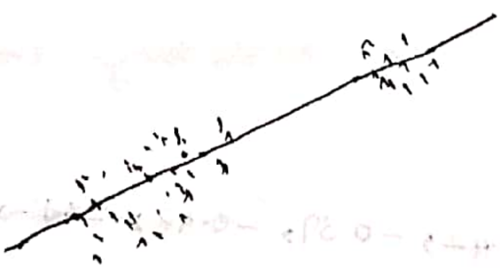
↓ log(100)



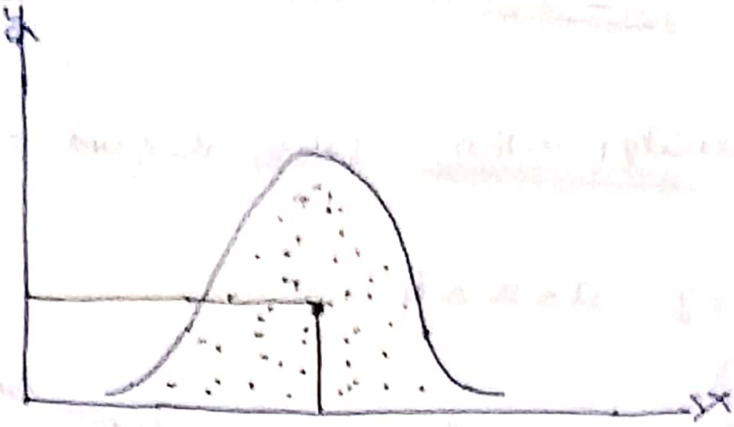
Positive skewness

For unnormalized data

unnormalized data



Normalised data



To find y value :-

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty \le x \le \infty$$

Standard normal distribution

Here $\mu = 0$, $\sigma = 1$ (Population Data)

Data = 1, 2, 3, 1, 100

Here μ = 15.714

σ = 34.41

After using z-stats :-

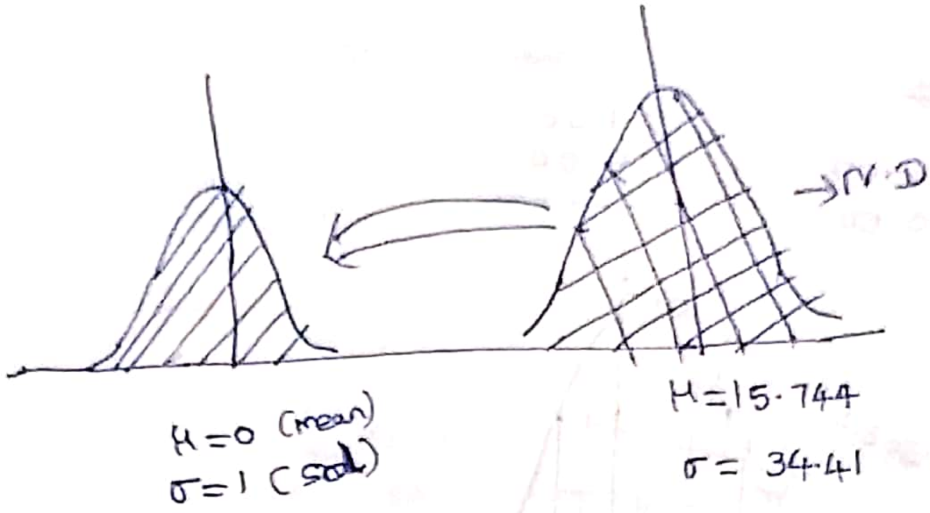
$$z = \left(\frac{x-\mu}{\sigma}\right) =$$

our Data become

−0.42, −3.9, −3.6, −0.42, −0.39, −0.42, 2.44 → SD

μ = 6.334e−17

σ = 1



Problem Statement

① $45 < x < 60$ (what is the Probability of completing assignment with in (45 to 60 minutes)

$\sigma = 10$
$\mu = 50$ } These are normal Distribution

$$P\left(\frac{45-50}{10}\right) \le \frac{x-\mu}{\sigma} \le \left(\frac{60-50}{10}\right)$$ For SND $\left\{\begin{matrix}\mu \ne 0 \\ \sigma \ne 0\end{matrix}\right)$

$$Z = \frac{x-\mu}{\sigma}$$ $x = 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 48, 59, 60$

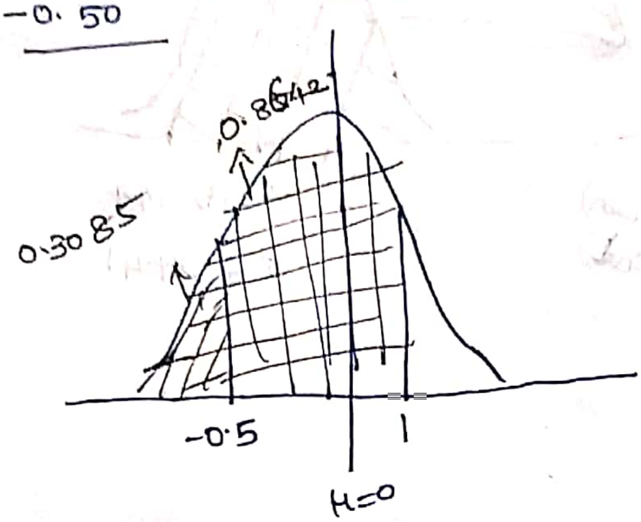
$P(-0.5 \le z \le 1)$

※ Here we are converting normal Distribution in to Standard normal Distribution

$P(-0.5 \le z \le 1)$

Eg :-

$$\begin{array}{r} 0.00 \\ -0.50 \\ \hline \end{array} \qquad \begin{array}{r} 1.00 \\ 0.00 \\ \hline \end{array}$$



Probablity = 0.8643 − 0.3085
= 0.5328

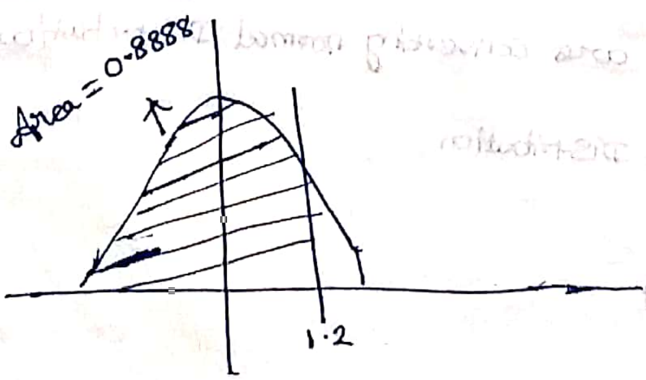
(2) Z statistics For 1.22 ?

Z = 1.22 Z table is valid from 0 to 3.4 only

$$\begin{array}{r} 1.20 \\ 0.02 \\ \hline 1.22 \end{array}$$

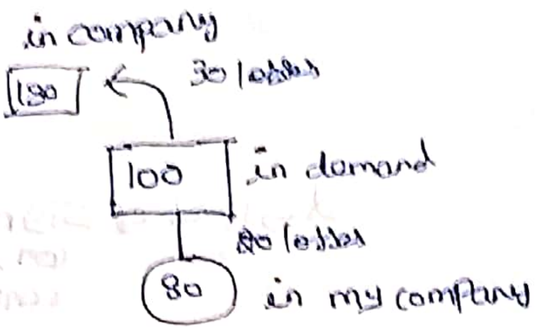
Z value = 0.8888

Area enclosed by the curve



Estimation Problem

Eg::



Point estimation

eg:: complete my task in 10 days

Interval estimation

eg:: complete my task in 10-20 days

Eg::

101, 120, 120, 130, 150, 160, 110, 120, 115, 170, 118, 120, 119

$\mu = 126.38$

$\sigma = 19.46$

$n = 13$

$$\boxed{\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

Estimation Formula

(i) For 90% of level of significance

$\alpha = 1 - 90\%$

$\alpha = 0.10$

$\alpha/2 = 0.05$

$\boxed{Z_{0.05} = 1.645}$

$\alpha \rightarrow$ level of significance (or) confidence

$\boxed{117 \leq E \leq 135}$

(ii) For 95% level of significance

$\alpha = 1 - 95\%$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$\boxed{Z_{0.025} = 1.96}$

$\boxed{116 \leq E \leq 137}$

(iii) For 99% of level significance

$\alpha = 1 - 99\%$

$\alpha = 0.01$

$\alpha/2 = 0.005$

$\boxed{Z_{0.005} = 2.525}$

$\boxed{112 \leq E \leq 140}$

Decreasing

$117 \leq E \leq 135 \rightarrow 90\%$

$116 \leq E \leq 137 \rightarrow 95\%$ while confidence level increasing

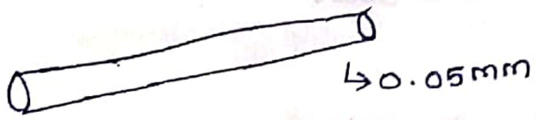
$112 \leq E \leq 140 \rightarrow 99\%$

lower part will decrease

Hypothesis Testing

IF we are making an assumption as Fact without proving the assumption is true or False

Eg :-

① 

Factory

Null hypothesis ← $H_0 : \mu = 0.05$ (Correct Statement)

Alternate hypothesis ← $H_1 : \mu = \neq 0.05$ (We are Disapproving)

For eg :- $0.045 < 0.05$
$0.051 > 0.05$

② $H_0: \mu = 0.05mm$

$H_1: \not\leq < \quad < \not\geq$

For eg:

one tail test

$\mu \not\leq 0.05mm$

Two tail test

$\mu < 0.05mm$ or $\mu \geq 0.05 mm$

③ Renting an Apartment

Budget → 15000

Apartment rate → 7000 – 30000

H_0: I will rent an apartment

H_1: I won't rent an apartment

$H_0: \leq 15k$

$H_1: < 15k$ or $\geq 15k$

Critical region

$H_0: \mu = 0.05$

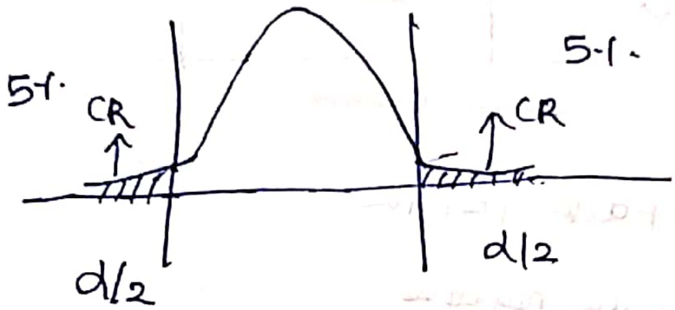
$H_1: \leqq < \quad < \geqq$

$H_0: \mu = 0.05 = 90\%$

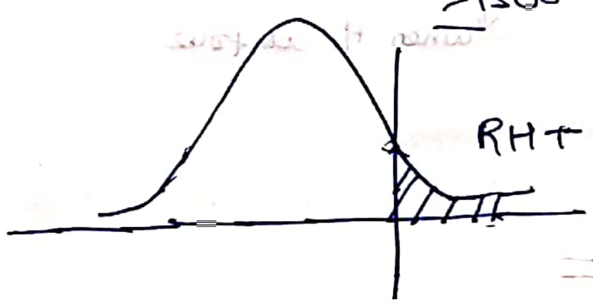
$H_1: 0.05 \leq$ or $\geq 0.05 = 10\%$

Two tail test

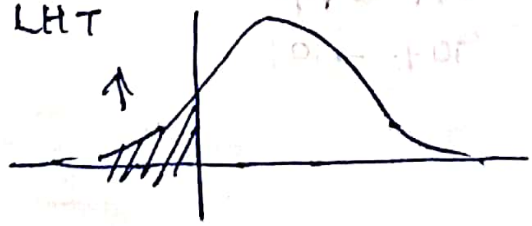
$\alpha = 10\%$ (critical value)



one tail test



$\alpha = 10\%$



≤ 15000

$\alpha = 10\%$

Confusion matrix

Decision	Innocent Ho (True)	not Innocent Ho (False)
Reject Ho	Type 1 Error	✓
Accept Ho	✓	Type 2 Error

Type 1 Error – False positive

Type 2 Error – False negative

$\alpha = P$[rejecting Ho when Ho is true]

$\beta = P$[not rejecting Ho when Ho is false]

↳ Power of test — when H_1 is true

level of significance (α)

H_0: clear exam 95%

H_1: Not clear 5%

eg :-

95% → 5%

99% → 1%

90% → 10%

Prerequisites For Hypothesis Testing

* Null hypothesis & Alternate hypothesis

* level of significance i.e $= \alpha =$ 5%, 1%

* Z test, t-test, chi square test, Anova test.

  ↓ For one and two tail test

Eg

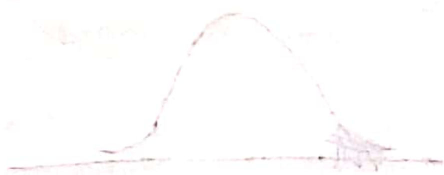
① Average Pointer 500 min
↓
Average

100 Pointer → sample

Average ↑ (490 ± 30)
↓
S.deviation

H_0: Buy $500 \geq$

H_1: Not Buy $500 \leq$



Std = 30

$\mu = 500$

$\bar{x} = 490$

S.Std = 30
↓
Sample

$\alpha = 5\%$

Sample size = 100

SE (Standard Error) $= \left(\frac{s}{\sqrt{n}}\right)$

$= \dfrac{\text{std of Sample}}{\sqrt{\text{sample size}}}$

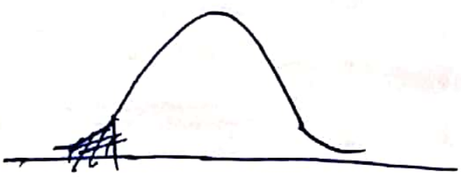
$= \dfrac{30}{\sqrt{100}} = 3$

$Z(\text{test}) = \dfrac{\text{Sample mean} - \text{Population mean}}{SE}$

$= \dfrac{490 - 500}{3}$

$= \dfrac{-20}{3}$

$= -3.33$



$\alpha = 5\%$

$Z(5\%) =$

$Z(\alpha) =$

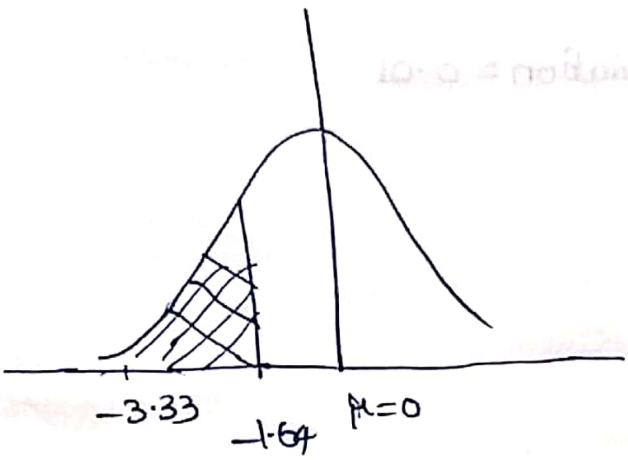
[From table -1.60
-0.04
-1.64]

$Z(0.05) = -1.64$

$Z(test) = -3.33$

$Z(0.05) = -1.64$

$Z(test) < Z(0.05)$



IF Z test lies in between Z table prediction means we have to accept (H_0)

H_0: Reject

H_1: Accept

Problem Statement

① A company used a specific brand has approached the company has an average life of 1000 hours. A new brand has approached the company with new Tube lights with same power at a low price. A sample of 120 light bulbs were taken for testing which yielded an average of

1100 hours with standard deviation of 90 hours. Should the company give the contract to this new company at a 1% significance level.

$\mu = 1000$

$\bar{x} = 1100$

$S = 90$

Size $= 120$

Standard deviation $= 0.01$

Ho: avg >= 1000

H1: avg < 1000

$$SE = \frac{S.d}{\sqrt{n}}$$

$$= \frac{90}{\sqrt{120}} = 8.22$$

$$Z(\text{test}) = \frac{SM - PM}{SE}$$

$$= \frac{1100 - 1000}{8.22}$$

$$= 12.16$$

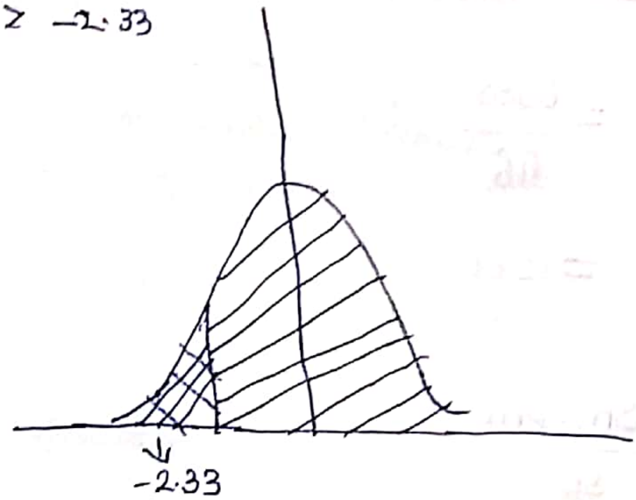
$$Z(0.01) = -2.33 \begin{bmatrix} -2.30 \\ -0.03 \end{bmatrix} \rightarrow 0.0099\,[0.01]$$

$\Downarrow$

From Ztable

$Z(\text{test}) > Z(0.01)$

$12.16 > -2.33$



H_0: Accept

H_1: Reject

T-Statistics

- multiple samples
- Population is not given
- small sample size < 30 (n)
- unknown variance

① A tyre manufacturer claims that the average life of a particular category of its tyre is 18000 km when used under normal driving conditions. A random sample of 16 tyres was tested. The mean and SD of life of the tyres in the sample were 20000 km and 6000 km respectively. Assuming that the life of the tyres is normally distributed, test the claim of the manufacture at 1% level of significance.

H_0: Population mean = 18000 km

H_1: Population mean $\neq$ 18000 km

$$SE = \frac{s.d}{\sqrt{n}} = \frac{6000}{\sqrt{16}}$$

$$= 1500$$

$$t(\text{Test}) = \frac{Sm - pm}{SE}$$

$$= \frac{20000 - 18000}{1500}$$

$$= 1.33$$

$\mu = 18000$

$S = 20000$

$s.d = 6000$

Sample size = 16

In t Table

dF = Sample size − 1

= 15

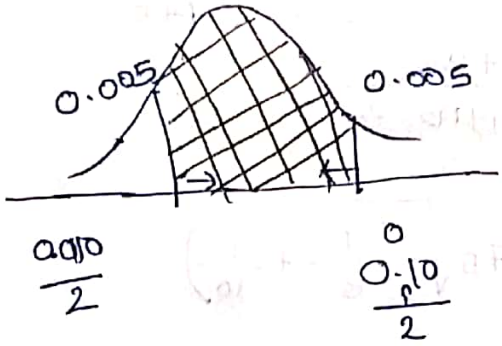
H_1

>

<

$$\frac{0.01}{2} = 0.005$$

2.947



$t(0.005) = 2.947$

$t(-0.005) = -2.947$

$-2.947 < 1.33 < 2.947$ [t Test lie in between these region]

Accept H_0

Reject H_1

② The means of two random samples of sizes 10 and 8 from two normal population is 210.40 and 208.92. The sum of squares of deviation from their means is 26.94 and 24.50 respectively. Assuming population with equal variances can we consider the normal populations have equal mean (significance level = 5%)

$P_1 \neq P_2$

$n_1 = 10$

$n_2 = 8$

n_1(mean) $= 210.40 (\bar{x})$

n_2(mean) $= 208.92 (\bar{x})$

$$Std = \sqrt{\frac{26.94 + 24.50}{10+8-2}} = 1.745$$

$$= \sqrt{\frac{\mu_1 + \mu_2}{n_1+n_2-2}}$$

$$SE = \frac{Std}{\sqrt{n}} = 1.75\sqrt{\left(\frac{1}{8} + \frac{1}{10}\right)}$$

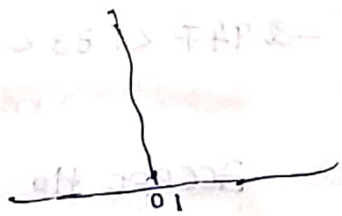
$$= \frac{1.75}{\sqrt{8+10}}$$

$$= 0.84$$

$$DF = n_1+n_2-2$$

$$= 10+8-2$$

$$\boxed{DF = 16}$$



$$t = \frac{26.94 - 24.50}{0.84} = \frac{Sm - pm}{SE}$$

$$\boxed{t = 1.76}$$

$H_0 : \mu(P(n_1)) = \mu(P(n_2))$

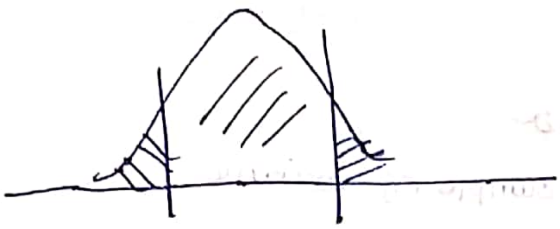
two populations are equal.

H_1 : Population mean are not equal

$<$ $>$

Significance level = 5%

$$\frac{0.05}{2} = 0.025$$



-2.120 2.120

$t = 1.76$

$t(5\%) = 2.120$

$-2.20 \le 1.76 \le 2.120$

Accept H_0

Reject H_1

P-value

It is less Stable than z-Statistics

chisquared test

→ For Variance
→ For Expected and Actual one

* It deals with high significant statistical analysis
* To compare the two independant data

Formula

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$$

n → sample size
s^2 → sample of variance
σ^2 → Population of Variance

Problem statement

① The variance of a certain size of towel Produced by a machine is 7.2 over a long period of time. A random sample of 20 towels gaves a Variance of 8. You need to check the variability for towel has increased at 5% level of significance assuming a normally distributed sample.

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$$

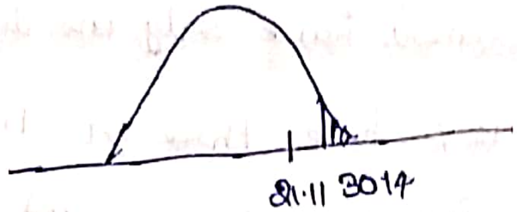
$n = 20$
$S.V = 8$
$P.V = 7.2$

$$= \frac{(20-1)\; 8}{7.2}$$

$\chi^2 = 21.11$

$H_0: V <= 7.2$

$H_1: V > 7.2$



critical value (5%) $= 30.14$

Second definition for chi squared test

The chi-square test is widely used to estimate how closely the distribution of a categorical variable matches an expected distribution (the goodness-of-fit test), or to estimate whether two categorical variables are independent of one another (the test of independence)

To calculate goodness-of-fit-test by chi squared test

Formula

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O → actual count in each category

E → Expected count in each category.

$H_0: O = E$

$H_A: O \neq E$

① A Survey conducted by a pet Food company determin[ed] that 60% of dog owners have only one dog, 28% have two dogs and 12% have three or more. You were not convinced by the survey and decided to conduct your own survey and have collected the data below,

Data : Out of 129 dog owners, 73 had one dog and 38 had two dogs. Determine whether your data supports the results of the survey by the pet

$$x^2 = \Sigma \frac{(O-E)^2}{E}$$

Significance level

<u>Expected one</u>

$E(1 dog) = 60\% = 0.60 \times 129 = 77.4$

$E(2 dog) = 28\% = 0.28 \times 129 = 36.12$

$E(3 dog) = 12\% = 0.12 \times 129 = 15.48$

<u>Actual one</u>

$O(1 dog) = 73$

$O(2 dog) = 38$

$O(3 dog) = 18$

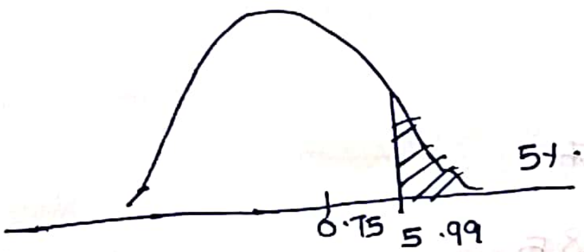
$$x^2 = \frac{(73-77.4)^2}{77.4} + \frac{(38-36.12)^2}{36.12} + \frac{(8-15.48)^2}{15.48}$$

$$= \frac{19.36}{77.4} + \frac{3.53}{36.12} + \frac{6.35}{15.48}$$

$x^2 = 0.75$

$df = 3-1 \quad (n-1)$
$\quad = 2$

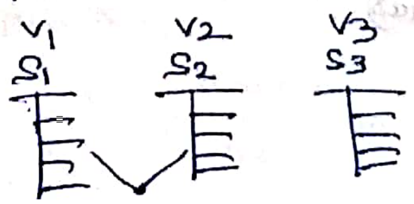
$x^2(\text{table}) = 5.99$



ANOVA Test (To calculate variance in between and with in the dataset)

(Analysis of variance)

It is used to mean of two or more samples are different from each other not.



1 Factor
one-way ANOVA

Assumptions

* Samples should independent
* data should be normally distributed
* Variance should be equal

Problem Statement

① In a survey conducted to test the knowledge of mathematics among 4 different schools in city - The sample data collected for the marks of students out of 10 is below.

School marks

School 1 8 6 7 5 9
School 2 2 6 4 6 5 6 7
School 3 3 6 5 5 6 7 8 5
School 3 4 5 6 6 7 6 7

Ho : All schools are performing equally

H_1 : It's not performing equally

K = 4 (Total no of samples)

N = 24 (no of datas)

School	Marks						
S_1	8	6	7	5	9		
S_2	6	4	6	5	6	7	
S_3	6	5	5	6	7	8	5
S_4	5	6	6	7	6	7	

Sum of Squares of mean between the sample (S_1 S_2 S_3 S_4)

$$(SSb) = \sum_{i=1}^{K} n_i (\overline{x_s} - x_g)^2$$

sample total no of data in a

$\overline{x_s}$ – Sample mean

x_g – grand mean

Mean between

mean b = (SSb / (k-1))

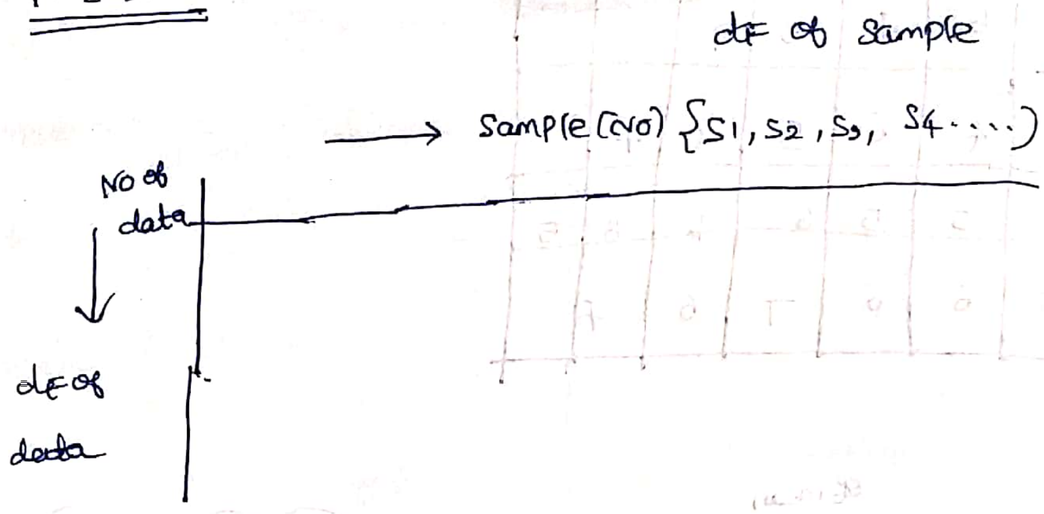
Sum of square Mean with in

$$SS_{within} = \sum_{i=1}^{N} (x - \bar{x})^2$$

Mean with in

$$\text{Mean w in} = \frac{SS_{within}}{24-4} = \frac{SS_{within}}{N-K}$$

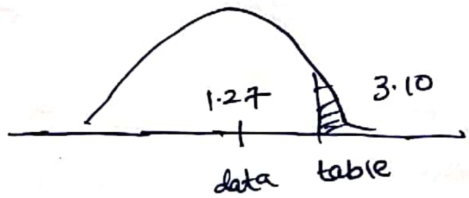
F-table



F = 3.10 (From table)

F = 1.27 (From data)

Diagram



Accept Ho

reject H_1

$$SSb = 5(7-6.20)^2 + 6(5.66-6.20)^2 + 7(6-6.20)^2 + 6(6.16-6.20)^2$$

$$= 4.99$$

Scanned with CamScanner

$SS\ with\ in = 10 + 5.33 + 8 + 2.83$

$= 26.16$

$$Mean\ with\ in = \frac{SS\ with\ in}{N-k}$$

$$= \frac{26.16}{20}$$

$$= 1.30$$

$$Mean\ bet = \frac{SSb}{k-1}$$

$$= \frac{4.99}{4-1}$$

$$= 1.66$$

$$Ftest\ (data) = \frac{mssb}{mss\ w}$$

$$= \frac{1.66}{1.30}$$

$$= 1.27$$

Sampling distribution of mean

x	p(x)
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

mean

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

variance

$$\sigma^2 = \sum_{n=1}^{n=x} (x - \mu)^2 \, p(x)$$

$$= (1-3.5)\frac{1}{6} + (2-3.5)\frac{1}{6} + \cdots + (6-3.5)\frac{1}{6}$$

$$= 2.92$$

Std deviation

$$\sigma = \sqrt{2.92}$$

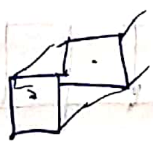
$$= 1.71$$

Scanned with CamScanner

sample

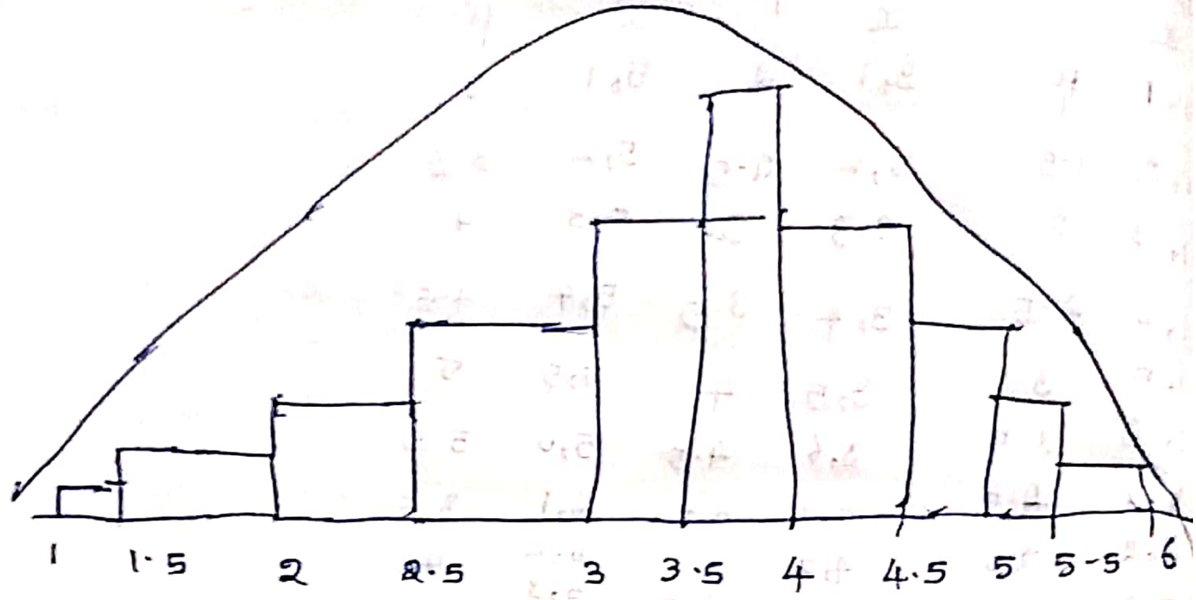
x	μ	x	μ	x	μ
1,1	1	3,1	2	5,1	3
1,2	1.5	3,2	2.5	5,2	3.5
1,3	2	3,3	3	5,3	4
1,4	2.5	3,4	3.5	5,4	4.5
1,5	3	3,5	4	5,5	5
1,6	3.5	3,6	4.5	5,6	5.5
2,1	4.5	4,1	2.5	6,1	3.5
2,2	2.0	4,2	3	6,2	4
2,3	2.5	4,3	3.5	6,3	4.5
2,4	3	4,4	4	6,4	5
2,5	3.5	4,5	4.5	6,5	5.5
2,6	4	4,6	5	6,6	6

If two dice were thrown

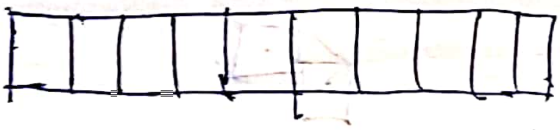


x	P(x)
1	1/36
1.5	2/36
2	3/36
2.5	4/36
3	5/36
3.5	6/36
4	5/36
4.5	4/36
5	3/36
5.5	2/36
6	1/36

Plotting



For normalised data



For un normalised data

Central limit theorem

IF you have taken a population, μ, s.d, if you have taken a large amount of sample of population. then distribution of sample mean will always be close to a normal distribution.

A/B Testing

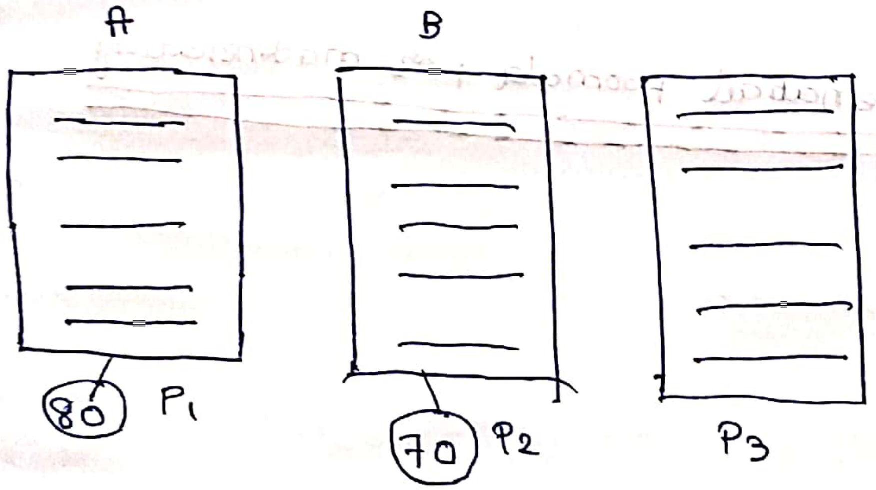
To compare diff version of experiment

Othernames

* Bucket testing
* split testing

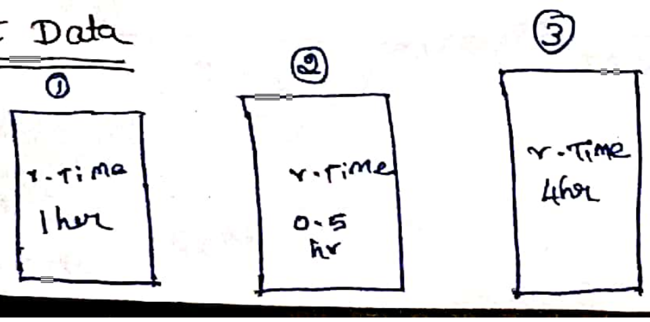
Diagram

Total 150 pages



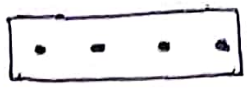
Steps for proceeding A/B Testing

* collect Data



* Goals

Buttons For the ad



* Hypothesis

IF we get some profit means we will scan

or else not.

* Variations

comparing different pages

* running Experiment
* Analyze result

Some mathematical Formula For machinelearning

* $\frac{dy}{dx}$

  x^2

  $y = x^2$

  $\frac{dy}{dx} = 2x$

* $y = x^2$

  $y = 2^2$

  $y = 4$

$y = 3^2$

$= 9$

$y = 4^2$

$y = 16$

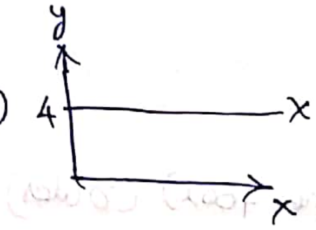
$y = 5^2$

$y = 25$

Formula

* $y = c$

$\frac{dy}{dx} = 0$ { if $(y=0)$



$(y)\ \frac{dy}{dx} = x^n$ *

$\frac{dy}{dx} = nx^{n-1}$

* $y = x^2$

$\frac{dy}{dx} = 2x$

* $y = \sin x$

$\frac{dy}{dx} = \cos x$

* $y = \cos x$

$\frac{dy}{dx} = -\sin x$

* $y = \tan x$ $\frac{dy}{dx} = \sec^2 x$

* $y = x^2$

$\frac{dy}{dx} = 2x$

* $y = \ln(x)$

$\frac{dy}{dx} = \left(\frac{1}{x}\right)$

* $y = e^x$

$\frac{dy}{dx} = e^x$

Tools — (Sine and cosine tool) (rules)

* $y = f(x) = h(x) \cdot g(x)$

$f'(x) = y' = h'(x) \times g(x) + g'(x)\, h(x)$

Eg :-

* $y = (1-x2)(1-x3)$

$\frac{dy}{dx} = -2x(1-x^3) + -3x^2 \cdot (1-x)$

cosine tool

$f(x) = y = \left(\frac{h(x)}{g(x)}\right) = \left(\frac{h'(x)\,g(x) - h(x) - g'(x)}{(g(x))^2}\right)$

Eg

* $y = \left(\frac{1-x^2}{1+x^3}\right)$

$$= \left(\frac{-2x(1+x^3) + 3x(1-x^2)}{(1+x^3)^2}\right)$$

chainrule

It is used mainly in neural networks.

* $\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx}$

*

* $y = 2x$

$y = mx + c$

when $x = 0$

$y = c$

* $y = mx$

$\frac{y}{x} = m$

$$y = 2x$$

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

* $y = mx + c = 0$

$$\frac{dy}{dx} = m$$