①

Type of statistics

Descriptive → Tells where the problem is
* for eg. how many century kohli scored

predictive → look at data & tell me what might happen. finding Relationship among object & predict
* for eg. how many centure kohli will score in wc

Prescriptive → action design to do something
* for eg. how do we get kohli to score more century in wc

Describtive → Describe the data without doing any prediction on it
Telling you the way it is

* Standard deviation std $\sqrt{\frac{1}{n-1}(x_1-\bar{x})^2 + -- + (x_n-\bar{x})^2}$

measure of spread
low std dev → clustered
high std dev → spread

$\bar{x} = \frac{1}{n}(x_1 + -- * x_n)$ avg.

it n-1 because diff. is already considered the element

* square is Done as we want to see both positive & negative deviation. Else it will neutralize it.

* mean absolute deviation (mad) $\frac{1}{n-1}(|x_1-\bar{x}| + -- + |x_n-\bar{x}|$

* Distribution $F(x) = Prob(X \leq x)$

$f(x) = F'(x)$

* distribution function is Integral of density fn
* density function is derivative of distribution fn

univariant bivariant

↓

sense of co-relation in two variables

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

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \rightarrow \text{covariance}(x,y) \quad \sigma_{xy}$$

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

↓ variance x

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

$$= \text{variance}(x)$$

$$= [\text{std dev}(x)]^2$$

<u>corelation</u> → cov(X,Y)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad -1 \leq \rho \leq 1$$

σ_x ↙ std dev x ; σ_y ↘ std dev y

covariance → nature of relationship between two Random variable but not the value of how strong the relationship (Because value is depend on unit of measurement)

corelation → it cancel out the unit of measurement due to devision of std. dev. It is measure of relationship between two Random variable.

③ variance → measure the dispersion of set of data point around their mean

std dev → $\sqrt{\text{variance}}$ → measure for spread of Data

relative std dev → coefficient of variation

↳ $\frac{\text{std. dev}}{\text{mean}}$

↓

used for comparing two data set

panda → df.corr() → will give correlation matrix

can be used Any Regression model

↳ Simply descriptive
- it predicts
- its prescriptive

Descriptive stat

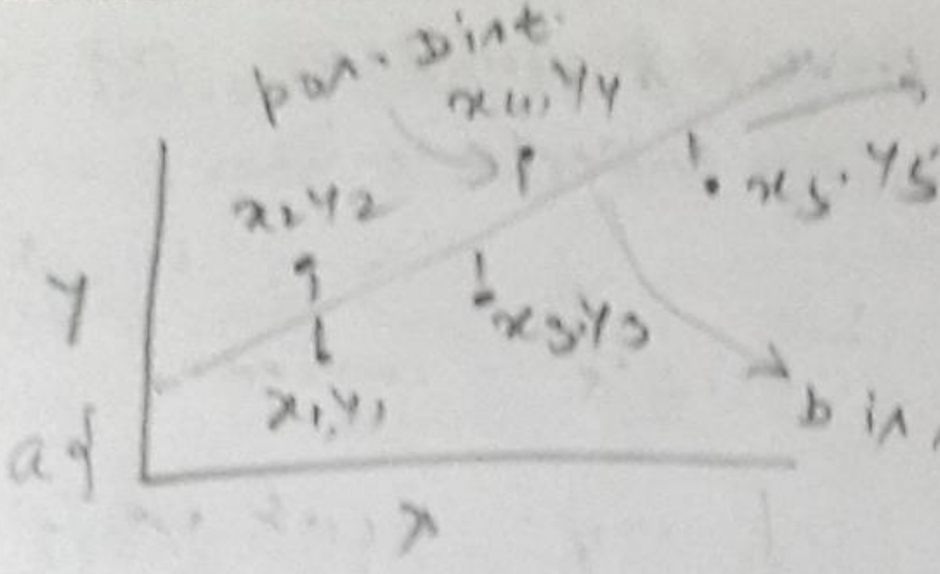
univariate
- location → mean, media, Q_3, Q_1 ← Quartile
- variation → std dev,, variance, Range
  Inter Quartile Range

bivariate → covariance. correlation

multivariate → linear Regression

$$y = \beta_0 + \beta_1 x_1 + -- + \beta_n x_n$$

it can be used for perdiction & prescriptio as well. but in Description. it u just to describe Relationship



$y = a + bx$

$$(y_1 - (a+bx_1))^2 + (y_2 - (a+bx_2))^2 + \cdots$$

↓

Tells how far the line is from the data → the smaller the closer

find the value of a, & b so that above eqn. is smallest i.e. choose a & b to minimize above eqn → This is what linear Regression Do

∝ least square algo

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - x)^2} = \frac{cov(x,y)}{var(x)}$$

coef-

$$y = \hat{a} + \hat{b}x$$

$$\hat{a} = y - bx$$

Intercept-

(5) Probability → likelihood of particular event

event → outcome of experiment

experiment → process performed to understand and observe possible outcome

Set of all outcome of experiment is called "sample space"

D → Defective
G → Good

3 component

$P(D) = 10\%$

sample space

{ GGG GGD GDG DGG
GDD DGD DDG DDD

4 chances of atleast 2 Defective component

$P(GDD \text{ or } DGD \text{ or } DDG)$

$= P(GDD) + P(DGD) + P(DDG)$ (OR)

mutually Exclusive or → Disjoint

if one happen other two can't happen

$3\ (.9 \times .1 \times .1)$

$= \underbrace{P(G) \times P(D) \times P(D)}$ → P(G) and P(D) and P(D)

independent event

when event are independent → Probability can be multiply

when events are disjoint → probability can be summed

mutually exclusive event → these are not independent but if one occur, other event can't

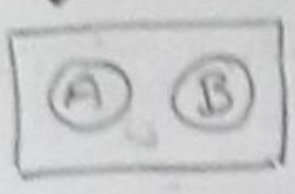
independent event → one event can't impact other event and both can occur simultaneously

Rules for computing Probability

1) addition Rule → mutually exclusive events

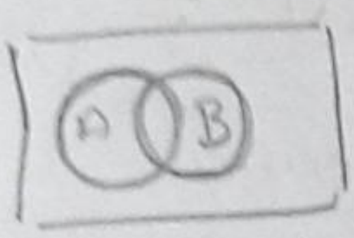
$P(A \cup B) = P(A) + P(B)$

A prevent occurence of B or vice versa



P(A or B)

i.e. A & B can not occur at same time



$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

addition Rule for not Mutually exclusive events

this part is 0 in case of mutually exclusive events

E.g.

* Probabily of getting king or Queen card

$P(K) = \frac{4}{52} = \frac{1}{13}$ $\quad P(Q) = \frac{1}{13}$

placed

$P(K \text{ or } Q) = \frac{1}{13} + \frac{1}{13} = 2/13$

* Porbability of K or Diamond card

$P(K) = \frac{4}{52}$ $\quad P(D) = \frac{13}{52}$ $\quad P(D \& K) = \frac{1}{52}$

(only one card can be king of Diamond)

$P(K \text{ or } D) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$

---

2) multiplication Rule → independent event

$P(A \cap B) = P(A) * P(B)$

Probability of Intersection

if occurance of A in in no way influenced by occurence of B or vice versa

* if events are not independent

Conditional Prob

$P(A \cap B) = P(A) * P(B/A)$ → if A has occure

or $P(B) * P(A/B)$ → if B has occu

① $P(B/A)$ = P(B given A)

find Probability of Event A occuring & then find whether event B will occur along with event A that will be a number of P(A)

↙ can also be written as $P(A \cap B) / P(A)$

e.g. Drawing 2 card from pack of card. what is probability of Drawing both cards as spade

let A = getting spade in first Draw

let B = getting spade in 2nd Draw if first card is not Replaced

(there are 12 spades & 51 cards Remaining

$P(A) = 13/52$ → $P(B/A) = 12/51$ → dependent Event as first card is not Re

if First card is Replaced $P(B) = 13/52$

↓ independent Event

it $P(B/A)$ → because both card has to be spade that possible if first Drawn card is already a spade. if yes. what the probability of 2nd card to be an spade

* marginal Probability →
* conditional probability

↳ $P(B/A) = \frac{P(A \cap B)}{P(A)}$

* joint probability

E.g.

Family	income less than 10 lakh	income >= 10 L	total
Buyer of car	38	42	80
Non buyer	82	38	120
Total	120	80	200

Q1 → what is prob of randomly selected family buy a car → 80/200 → (marginal prob.)

Q2 → Prob of randomly selected family is both car buyer & income >= 10 L

$\frac{42}{200}$ (joint prob)

Q3 → randomly selected family belong to income >=10, find prob of this Family buying a car

$\frac{42}{80}$ (as per table)

$$P(\text{car}/>10L) = \frac{P(\text{car} \cap >10L)}{P(>10L)}$$

$$= \frac{42/200}{80/200} = 42/80$$

↓ conditional Prob.

$$\text{conditional prob} = \frac{\text{joint Prob}}{\text{marginal Prob.}}$$

④ **Bayes' theorem**

P(A and B)
= P(B) * P(A/B)
or P(A) ∝ P(B/A)

P(spam / word)

↳ P(word) * P(spam/word)

P(spam) ∝ P(word/spam)

$$P(spam/word) = \frac{P(word/spam) \propto P(spam)}{P(word)}$$

e.g. of Baye's Theorem

$$P(HIV/+) = \frac{P(HIV \text{ and } +)}{P(+)}$$

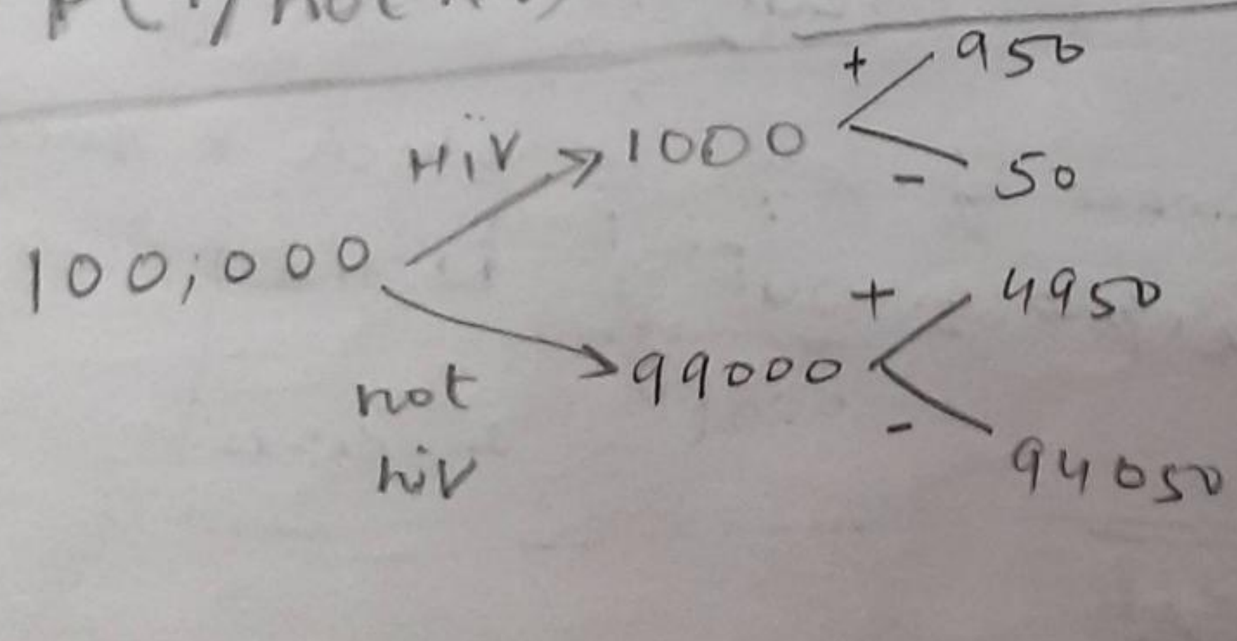
$$= \frac{P(+/HIV) * P(HIV)}{P(HIV \text{ and } +) + P(\text{not HIV and } +)}$$

$$= \frac{P(+/HIV) * P(HIV)}{P(+/HIV) \propto P(HIV) + P(+/\text{not hiv}) \propto P(\text{not hiv})}$$

P(HIV) = 1%
P(NHIV) = 99%
P(+/HIV) = 95%
P(-/not hiv) = 95%
↳ P(+/not hiv) = 5%

$$= \frac{.95 * .01}{(.95 * .01) + (.05 * .99)}$$

$$= 0.16$$



all folks tested positive has HIV

$$\frac{950}{950 + 4950}$$

= 16%

P(Spam) = 30%.

P(NSpam) = 70%.

P(congratulations/spam) = 75%

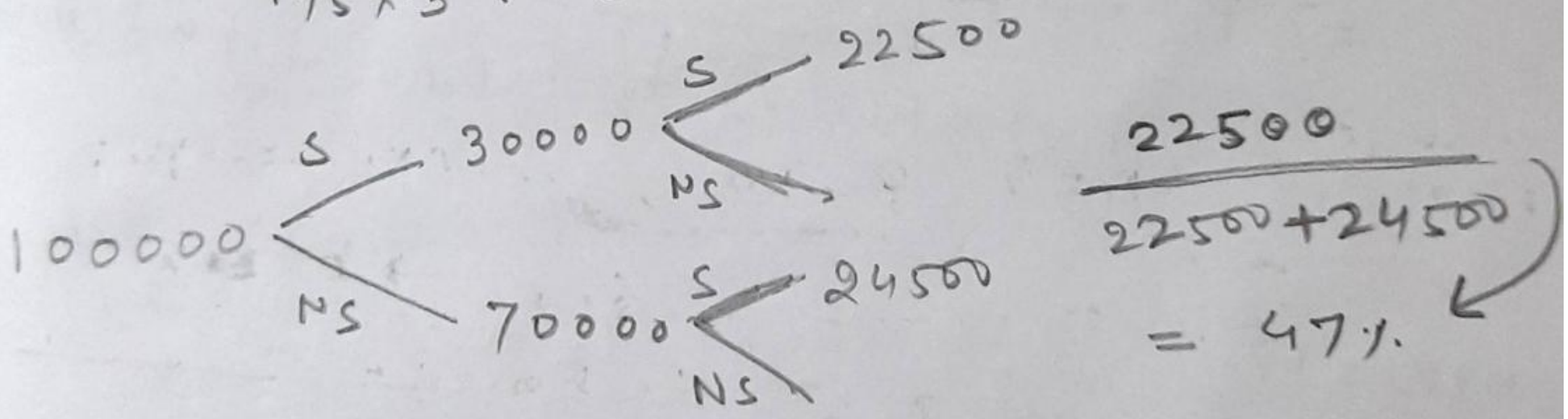
P(congratulations/NSpam) = 35%.

P(Spam/congratulations) = ?

$$\rightarrow \frac{P(cong/spam) \propto P(spam)}{P(cong) \longrightarrow P(cong \& spam) + P(cong \& NSpam)}$$

$$\frac{P(cong/spam) \propto P(spam)}{P(cong/spam) \propto P(spam) + P(cong/NSpam) \propto P(NSpam)}$$

$$= \frac{.75 * .3}{.75 \times .3 + .35 \times .7} = 0.47$$



Binomial Distribution

Probability of getting x success out of N trails

$$P(x) = \binom{n}{x} P^x (1-P)^{n-x}$$

$\binom{n}{x}$ is number of ways in which x success can take place out of N trails

$$\rightarrow \frac{n!}{x!(n-x)!}$$

(17) ** if P is prob of succes of a single trail then what is the prob of getting x success out of N trail

n → trail
P → success prob of each trail (based on past data)

$$P(x \text{ success}) = \binom{n}{x} p^x (1-p)^{n-x}$$

E.g → Bank issue cc to customer
→ bank has found 60% of all account paying on time
→ id 7 account is selected at Random
→ construct the Binomial prob Distribution of paying on time

for eg: $P(2 \text{ on time}) = \binom{7}{2} (.6)^2 (.4)^5$

if 2 account among 7 pay on time $\hookrightarrow \frac{7!}{2!5!} = 21$

Through panda

n=7 p=.6 , k = np.arrange(0,8)

k = 0,1,2,3,4,5,6,7 → because there can be 8 combination for 7 random account → 0 means none of 7 paid on time → 7 means all paid on time

```
import scipy.stats as stats
binomial = stats.binom.pmf(k,n,p)
```

Pmf → Prob mass function (P)

mean & std dev. of Binomial Dist

* mean μ of Bin. dist.

$$\mu = E(x) = nP$$

* std. dev. σ is given by

$$\sigma = \sqrt{nP(1-P)}$$

* for previous e.g. mean $= 7 \times .6 = 4.2$

$$\text{std} = \sqrt{4.2(1-.6)} = 1.3$$

Poisson Distribution

* another Discrete Distribution which also plays a critical Role in quality control in context of reducing the number of defect per std. unit

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

$P(x)$ = Porb of x succes given an idea of λ

λ = avg. number of succes

e = 2.71828

x = succes per unit

λ is parameter of poision Distri

mean of poision Distri $= \lambda$

std. dev. of Poision Distr. $= \sqrt{\lambda}$

$\lambda = nP \rightarrow$ P (Probability)
n (sample given)

(13) eg. of Poisson Distribution

* if on avg 6 customer arrive at every 2 min in bank during busy hours (3 customer in a min)

a) what is prob that exactly 4 customer arrive in a given min (8 cust at every 2 min)

b) wha is prob that more than 3 customer arrive in a given min

Diff w.r.t. binamial Distribution → sample space is not definite (there were 7 account) but here 6 is avg. number. It can be more or less than 6 also

from panda rate = 6

n = np.arrange(0,20) → can be set to any reason

Poisson = stats.poisson.pmf(n, rate)

do has been set from computational perspective & visualization perspec

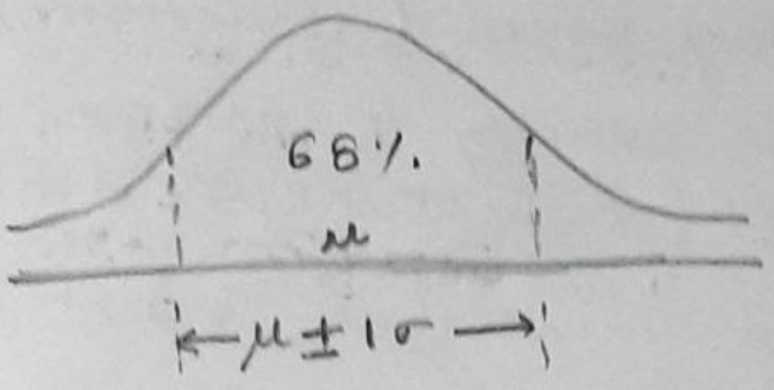
a) P(x=8)

b) 1 − P(x>=3) = 1 − P(x<3)

P(x>=3) = 1 − (P(0) + P(1) + P(2))

Gaussian / normal Distribution

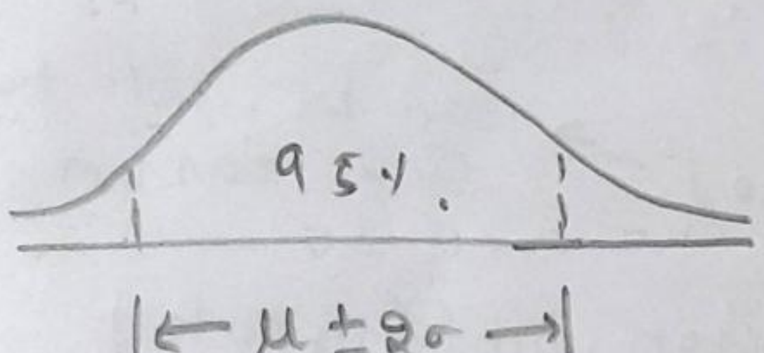
eg.



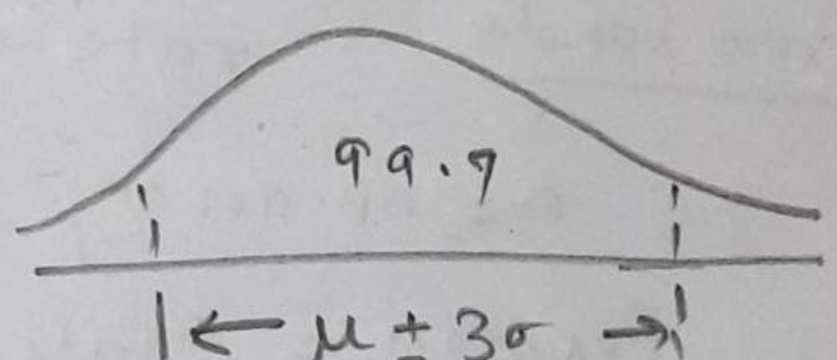
* derived from "central limit theorem"
* used when we don't have enough data but has mean & std dev.

* approximate 68% of the data in a bell shaped dist in with in 1 Std dev of the mean
  i.e. $\mu \pm 1\sigma$ (μ is mean & σ is Std. dev)

* This is emperical Rule approximates the variation of data in a bell shaped distribution



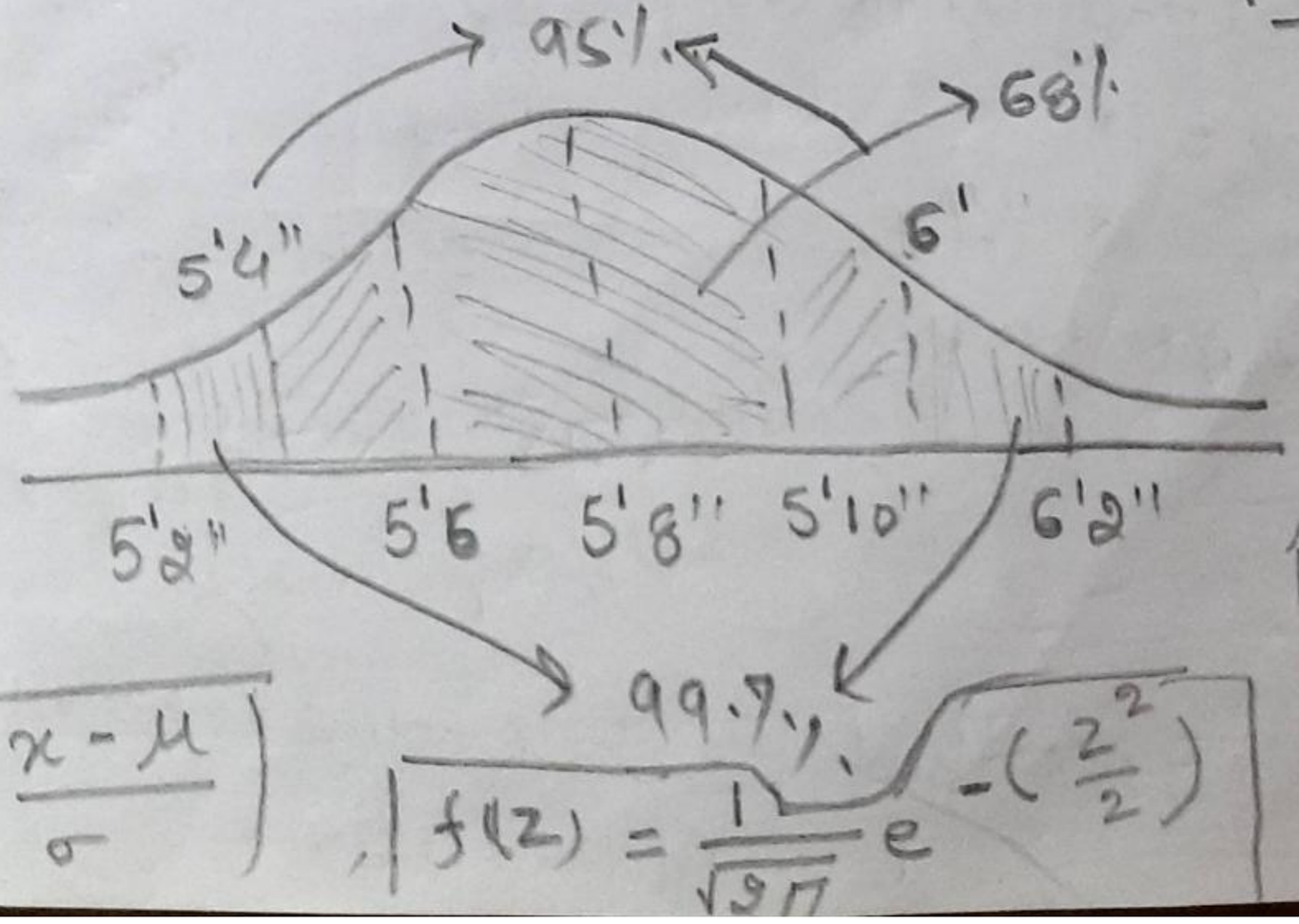
chances of being 2 std dev



'chance of being 3 Std dev.

for eg:

* mean heigh = 5 ft 8 inch
* std dev = 2 inch



library

stats.norm.cdf

cdf = cumu

Dist fun

z is number independent v

$$z = \frac{x - \mu}{\sigma}$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{z^2}{2}\right)}$$

(15)

E.g. mean of morning breakfast pack in .295 kg
a Std. dev in 0.025 kg.

a) what is prob. that pack wt. in less than .28 kg
b) " " " " more than .350 kg.
c) " " " " between .26 to .34kg

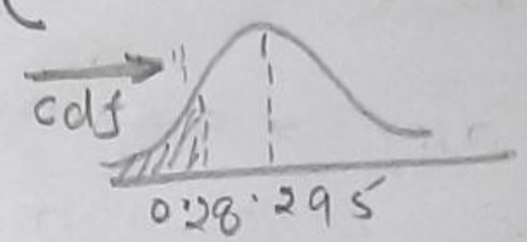
a) $z = (.28 - .295/0.025) = -0.6$

stats.norm.cdf(z) = 27%.

OR

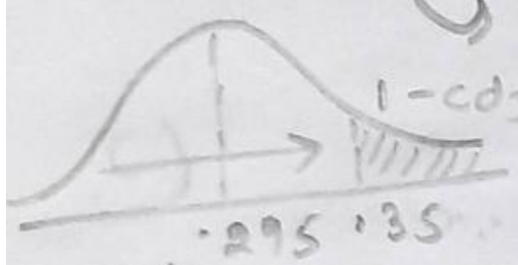
stats.norm.cdf(.28, loc = .295, scale = 0.025)
(loc = mean, scale = std dev)

↳ = 27%.



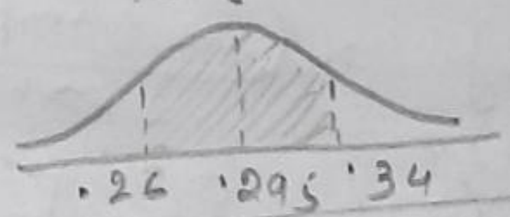
b) 1 - stats.norm.cdf(.350, .295, .025)

↳ 1.39% (whole prob in 1, so for more than we have to do 1 - cdf)



c) stats.norm.cdf(.340, .295, .025)
- stats.norm.cdf(.260, .295, .025)

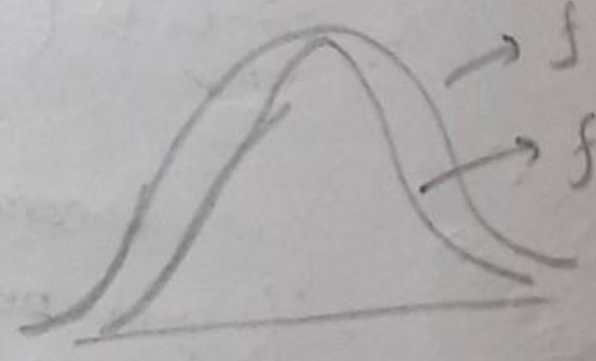
↳ = 88%.



cdf(.34) - cdf(.26)

Sampling Distribution → $\bar{x}$

$\begin{cases} N = 60 \rightarrow \text{Total number} \\ \mu = 70 \rightarrow \text{mean} \\ \sigma = 4 \rightarrow \text{std dev.} \\ n = 4 \rightarrow \text{sample taken} \end{cases}$



$\sigma(\bar{x}) = \frac{\sigma(x)}{\sqrt{\text{sample size}}}$

Sampling Distribution: in → take 4 Random sample (${}^{60}C_4$) → take mean of each sample

Set → plot Distribution of each sample mean
→ called sampling Distribution

Sampling Error $|\bar{x} - \mu|$

* if $x_1, x_2, x_3, \ldots x_n$ are n independent random samples drawn from Normal population with mean = μ, std dev σ, then sampling distribution $\bar{x}$ follows a normal distribution with mean μ & std dev $\frac{\sigma}{\sqrt{n}}$. $\frac{\sigma}{\sqrt{n}}$ is known by term Std. Error.

Central limit theorem

min sample size theoritically to meet central limit theorem

$x \not\sim GD/ND(\underset{\text{mean}}{\mu}, \underset{\text{Variance}}{\sigma^2})$ $\quad n \geq 30$

Random Variable → may or may not in normal Distribution

Sample, $x_1 \; x_2 \; \ldots \; x_{30}$ $\quad \bar{x}_1$ (mean of sample S_1)

S_2 $\quad \bar{x}_2$

$\vdots$ $\quad \vdots$

S_{100} $\quad x_{100}$

if I take all above & try to plot $\bar{x}$, it will follow normal distribution irrespective what is distribution of x (as sample size becomes larger & variance larger)

$$\bar{x} \simeq GD\left(\mu, \frac{\sigma^2}{n}\right)$$

mean (will be approx same as mean of x)

$\left(\frac{\text{Std dev}}{\text{or}}{\text{Std err}}\right) \frac{\sigma}{\sqrt{n}}$

$\bar{x} - \mu \rightarrow$ sampling

when Std dev is known

$\bar{x} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$

margin of error

when proportion is known

$\tilde{P} \pm Z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$

Z → confidence coficient

2.58 → 99% confidence

1.96 → 95% confidence

Refer calculate Z-Score video on youtube

(17) Types of probability

1) Priori classical probability → apply Rule to find prob
2) Empirical prob → Data Driven prob or Relative freq Prob
3) Subjective prob
   ↳ start without data & ask experts
   ↳ learn & Revise based on data availability over a period

$P(A) = \frac{m}{n}$ → favrable outcome
n → total number of outcome

for eg. in KBC

50-50 option → classical prob
Audience poll → Data. Driven prob (Empirical)
Phone a friend → Subjective prob

Baye's theorem Exercise

Drilling company

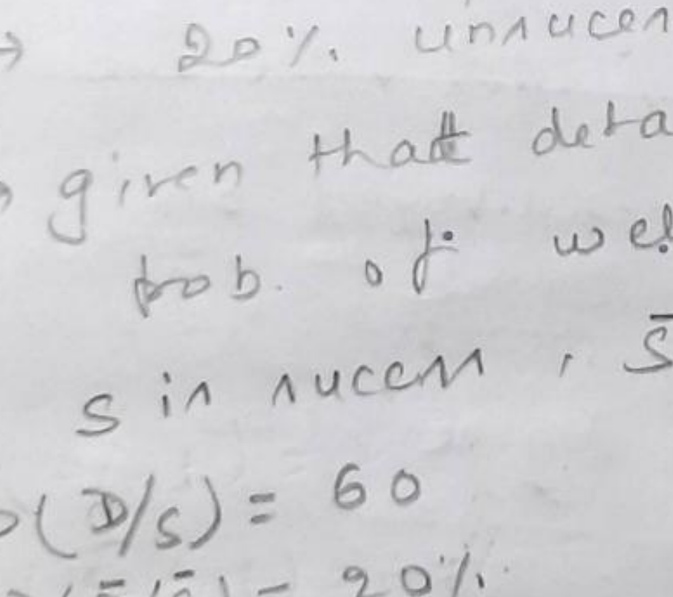
→ estimated 40% chance of striking oil for new well
→ detailed test has been scheduled for more info
→ Historically. 60% successful wells had detailed test
→ 20% unsuccessful well had detailed test
→ given that detailed test is scheduled. what is prob. of well will be successful.

* S in succes, $\bar{S}$ in unsuccess,

$P(D/S) = 60$
$P(\bar{D}/\bar{S}) = 20\%$
$P(S) = 40\%$
$P(\bar{S}) = 60\%$

$$P(S/D) = \frac{P(D/S) * P(S)}{P(D)}$$

$$= \frac{P(D/S) * P(S)}{P(D/S) * P(S) + P(\bar{D}/\bar{S}) * P(\bar{S})}$$



$\frac{24}{24+12} = (67\%)$

$\frac{.6 \times .4}{(.6 \times .4) + (.2 + .6)} = \frac{.24}{.24 + .12}$

$= 0.67\%$

Prob. sample que

Q1 Out of 37 men and 33 women
36 are completely refrain from alcoholic beverages
9 women are non-smoker
16 men smoke but do not drink
13 men & 7 women drink but do not smoke

How many both drink & smoke, what is associated prob?

solution → (using marginality principal)

m → men
$\bar{m}$ → women

$m + \bar{m} = 70$

$D + \bar{D} = 70 \rightarrow D = 34$
$\bar{D} \rightarrow 36$

$D = DS + D\bar{S} \rightarrow DS = 34 - D\bar{S}$

$D\bar{S} = mD\bar{S} + \bar{m}D\bar{S} = 13 + 7 = 20$

$DS = 34 - 20 = 14$

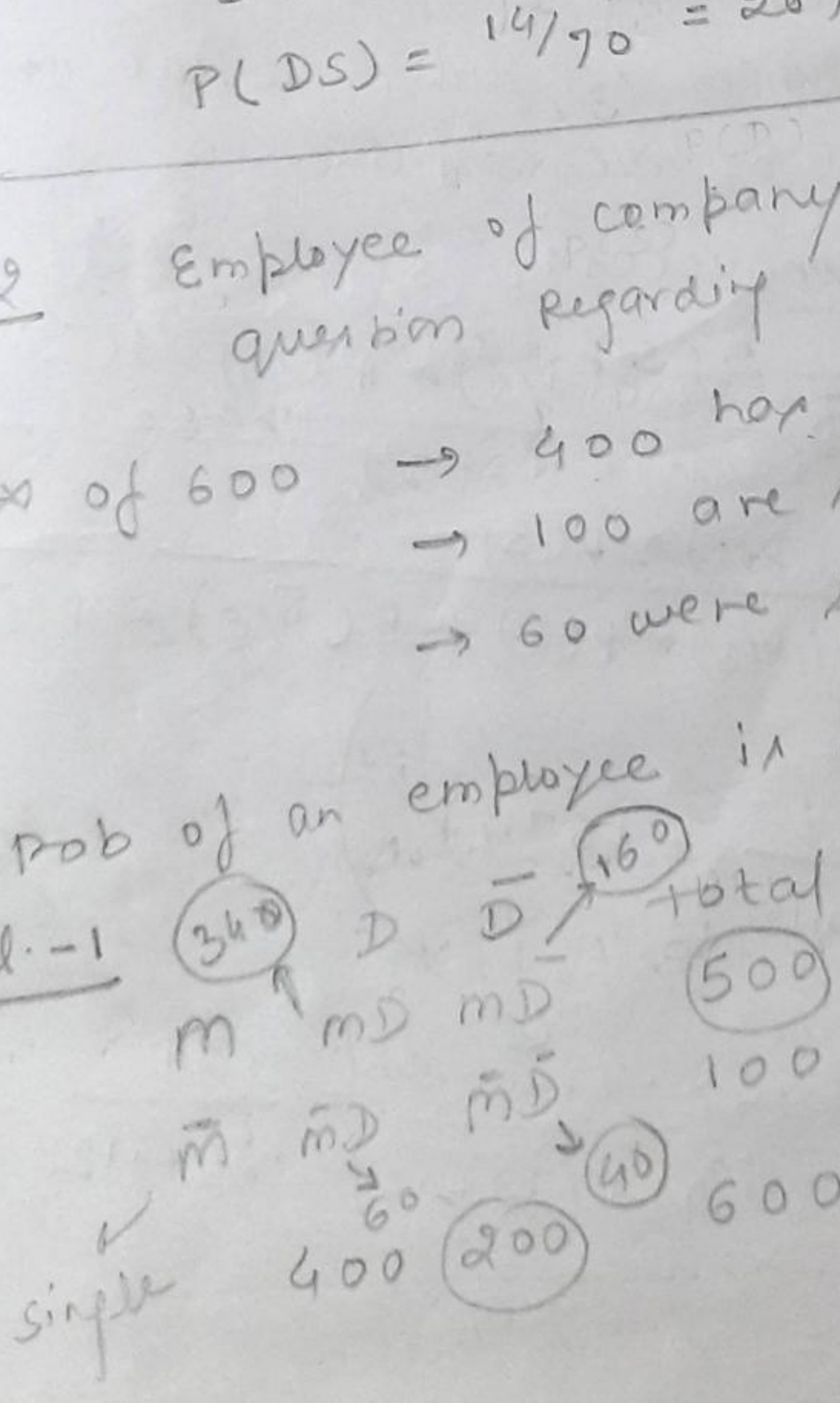
$P(DS) = 14/70 = 20\%$

Q2 Employee of company were surveyed on question regarding college degree

% of 600 → 400 has college degree
→ 100 are single
→ 60 were single college student

Prob of an employee is (single or has college

sol.-1



$\bar{m}$ or D = $\bar{m}\bar{D} + \bar{m} + mD$

$mD + \bar{m}D$

$\bar{m}D$ is already counted

(19)

$\bar{m} \cap D = mD + \bar{m}D + \bar{m}\bar{D}$

$= 360 + 60 + 40 = 440$

$P(m \cap D) = \frac{440}{600} = 77.3\%$

approach - 2 (not mutually exclusive approach)

$P(m \cap D) = P(m) + P(D) - P(m \cap D)$

$= \frac{400}{600} + \frac{100}{600} - \frac{60}{600}$

$= \frac{440}{600} = 77.3\%$

Q3. ∝ prob of house sale will increase next 6 month is .25

∝ prob of interest rate going up in same period is .74

∝ prob of house sale or interest rate go up in next 6 month is 0.89.

find prob. of both house sale & Interest Rate will Increase in next 6 month is ?

approach-1

	I	$\bar{I}$	
S	SI	$S\bar{I}$	0.25
$\bar{S}$	$\bar{S}I$	$\bar{S}\bar{I}$	(.75)
	.74	(.26)	1

$SI + S\bar{I} + \bar{S}I = .89$

$SI + (.25 - SI) + (.74 - S$

$SI = .99 - .89 = .1$

approach 2

$P(S \cap I) = P(S) + P(I) - P(S \cup$

$P(S \cap I) = P(S) + P(I) - P(S \cap I)$

$= .25 + .74 - .89 = .1$

Q4. company has 2 machine

∝ old machine produce 23% defective widget

∝ new machine produce 8% defective widget

∝ in addition new machine produce 3 times as widget as older machine

give the randomly choosen widget, what is prob it's produced by new machine? → tested △ found d

approach-1

	N	$\bar{N}$	
	75	25	→ 100 (volume)
D	.08	.23	

$$D = DN + D\bar{N}$$

$$= .08 \times 75 + .23 \times 25$$

$$P(DN) = \frac{DN}{DN + D\bar{N}} = \frac{.08 \times 75}{.08 \times 75 + .23 \times 25} = 0.51$$

approach → 2

	D	$\bar{D}$	
N	.08×75 = 6	69	75
$\bar{N}$	.23×25 = 5.75	19.25	25
	11.75	88.25	100

$$P(DN) = \frac{6}{11.75} = 0.51$$

$Q_5 \rightarrow$ in a used car lot

1) 70% have A/c
2) 40% CD player
3) 20% have Both

	A.C	$A\bar{C}$	total
CD	.2	(.2)	.4
$C\bar{D}$	(.5)	(.1)	(.6)
	.7	(.3)	1

1) how many car. have A/c but no CD → .5 i.e. 50
2) what is prob of car has CD given it has A/c also → $\frac{.2}{.7} \Rightarrow \frac{20}{70} = .2857$

$$P(CD/AC) = \frac{P(AC/CD) * P(CD)}{P(AC)}$$

↓ by Bayes's theorem

$$= \frac{\frac{.2}{.4} \times .4}{P(AC/CD) * P(CD) + P(AC/C\bar{D}) \times P(C$$

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

↓ .2857

P(CD) & P(AC) → marginal
P(CD) & P(AC) → joint

P(CD/AC) → Conditional Prob

(21) Sample size problem

$Q_1 \rightarrow$ marketing manager of food chain wishes to estimate avg. yearly amount that family spend on fast food He wants an estimate to be within 100 with a confidence level of 99%. It is know from earlier plot study that std dev. of family expenditure on fast food is 500. How many family must be chosen for this problem.

for 99% confidence → value of $z = 2.58$

& std. dev. is known

$$100 = 2.58 * \frac{\sigma}{\sqrt{n}} \Rightarrow 100 = 2.58 * \frac{500}{\sqrt{n}}$$

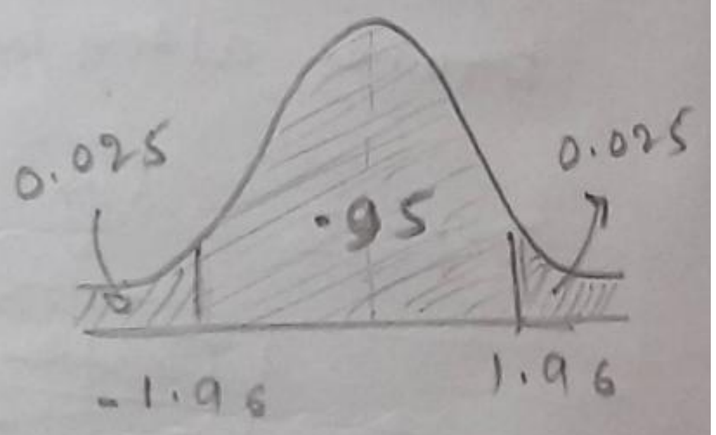
called margin of error

calculate for n → will be sample size number of above problem

$Q_2 \rightarrow$ sports good manufacturer want to estimate the propation of criket player among high school student in India. the company wants the estimate to be with in .03 with a confidence level of 95%. A plot study done earlier reveals that out of 80 high school student, 36 play cricket. what should be sample size for study.

for 95% confidence → $z = 1.96$

$$1.96 * \left(\pm\sqrt{\frac{P(1-P)}{n}}\right) = 0.03$$



$$P = \frac{36}{80} = 0.45$$

$$1.96 * \sqrt{\frac{0.45 * 0.55}{n}} = 0.03$$

calculate n → will be sample size for above solution

↓

(1050)

Q3→ wat if in Q_2, Plot study is not given as part of problem statement. How to calculate sample size in such case ie. when P is unknown

take max value of P & 1-P i.e. .5 each

$$1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}} = .03$$

calculate n → will be my sample size

↓

(1067)

Hypothesis (statistical)

a statistical hypothesis is a statement about a population parameter. It may & may not be true. The manger has to ascertain the truth of hypothesis.

H

H_0 null — H_1 Alternative

* formulate null hypothesis in such a way that its Rejection will Result into Alternate Hypothesis

* A Null hypothesis is status quo.

* Researcher & Decision makers generally want before alternate hypothesis.

Type I and Type II error (Null Hypothesis)

	True	False
Reject	Type I error (α)	No Error
Accept	No Error	Type II error (β)

* Rejecting Null hypothesis when its True Type I error.

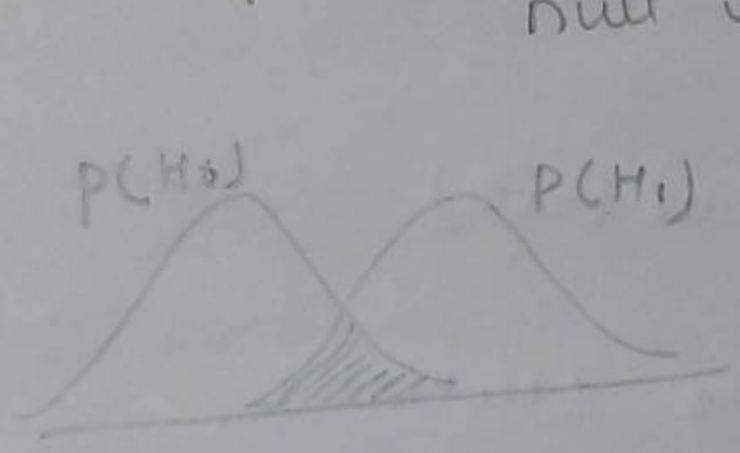
* Accepting the null hypothesis when it false is Type 2 error.

(23) $\alpha \rightarrow$ prob of committing the type I error
also called the level of significance of the test (statistical level of significance)

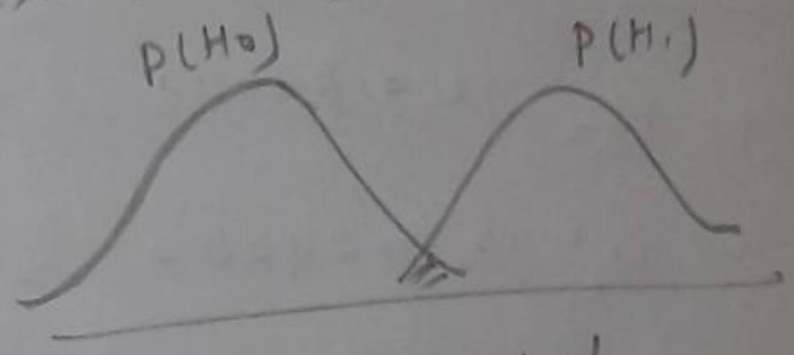
$\beta \rightarrow$ prob of committing the type II error

$1-\alpha \Rightarrow$ confidence level of the test (Prob of accepting null when it's true)

$1-\beta \Rightarrow$ Power of the test (Prob of Rejecting null when it's False)



more overlap → test is not Powerful



less overlap → test is powerful

$\alpha \uparrow \beta \downarrow$ .. $\alpha \downarrow \beta \uparrow$ its tradeoff

(1-β can also be written as)

$1-\beta \rightarrow$ Prob of accepting alternative when it's true

e.g.: launching a product line in a new market

- a Survey of Random Sample of 400 household in that market showed a mean Income per household of $30000. Std. dev. based on a pilot study of 400 household is $8000.
- Karen believes the product line will be adequately profitable only in market where mean household income is greater than $99000
- Should Karen introduce product line into new market?

$n = 400$

$\bar{x} = \$30000$

$\sigma = \$8000$

* alternate hypothesis can be $>, <, \neq$ but can never have equality sign
* Null hypothesis is always tested for threshold as should have equality sign

Step-1 formulate Null & alternate hypothesis

H_0 $\mu = 29000$

H_1 $\mu > 29000$ (Right tail test)

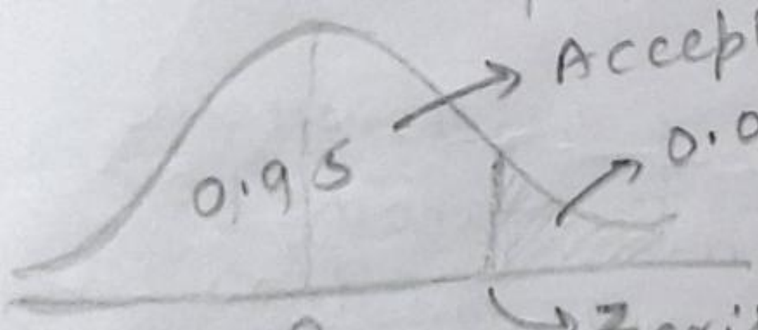
Step-2 Select the appropriate statistic

$$Z_{stat} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \begin{matrix} \rightarrow \text{sampling error} \\ \rightarrow \text{std. err} \end{matrix}$$

Step-3 compute Z_{stat}

$$Z_{stat} = \frac{30000 - 29000}{8000/\sqrt{400}} = 2.5$$

Step-4 mark in diag. acceptance Region & Rejection Region



mean (μ) is 0 when $\bar{x} = 30000$ → calculate using Z score calculation as μ is 29000, and $\bar{x}$ is 30000, it is Right tail test.

---

theoritically when nothing is mention → take α as 5%. that means 95% confidence and level of significance is 5%.

that means

$P(Z) > 1.645 = 0.05$ i.e. 5%.

(25)

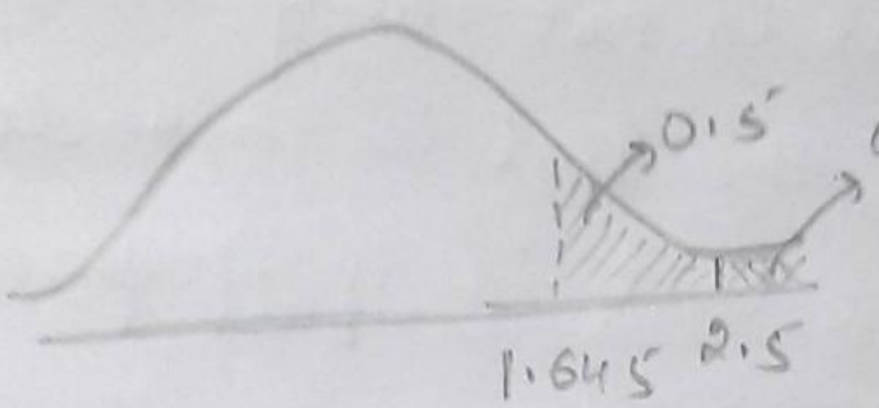
steps → find out where computed Z falls

2.5 falls in Rejection Region as it greater than 1.645

* hence Null hypothesis is Rejected

* if $P(Z) \geq 1.645 = 0.5$ means, we are allowed to take Risk of 5%.

* calculate Z is 2.5 ie. actual Risk is much smaller than 5%.



definetly it less than 0.5

P(2.5) → called P-value of statistics (actual Risk)

step 6 → As Null hypothesis is Rejected, Alternate hypothesis is Accepted, hence Keren can lauch the product in new market.

∞ desired Risk is 5%., Actual Risk is 0.62%., means I am 99.38% confident on the decesion made.

* if Pval ≤ α (in above case 0.62% ≤ 5%.)
Null hypothesis is Rejected
(another way to approach above problem)

8.9.2 A retailer is weighing strawberry to sell as 250 g. punnets. customer complained that what he bought was less than 250 g. retailer decided to check weight of 36 punnet He finds avg wt. is 248.5 g. with std dev. of 4.8 g. In using significance test to judge whether he is selling under wt. punnet, which of following is true.

a) 5 %. level he is selling underwt.

b) at 5 %. level. he is not selling under wt

c) at 5 %. level the test is inconclusive

d) A significance test is inappropriate in this case.

Step 1

$\bar{x} = 248.5$ $\mu = 250$

$n = 36$

$\sigma = 4.8$

H_0 $\mu = 250$

H_1 $\mu < 250$ (left tail test)

Step 2

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

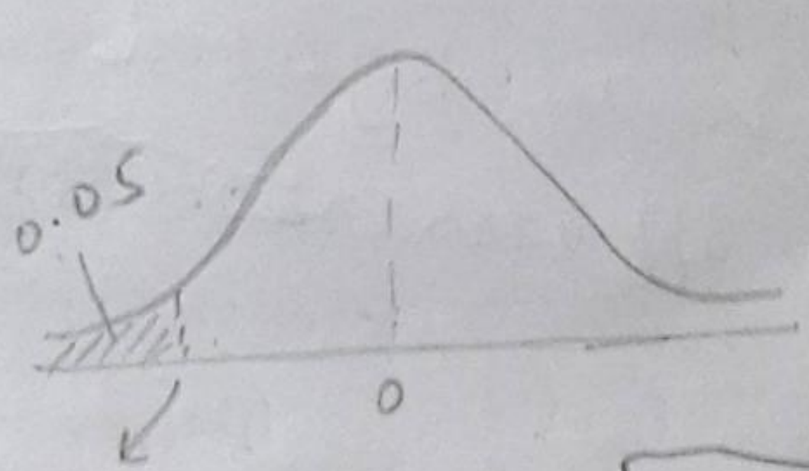
Step 3

$$z_{stat} = \frac{248.5 - 250}{4.8/\sqrt{36}} = -1.875$$

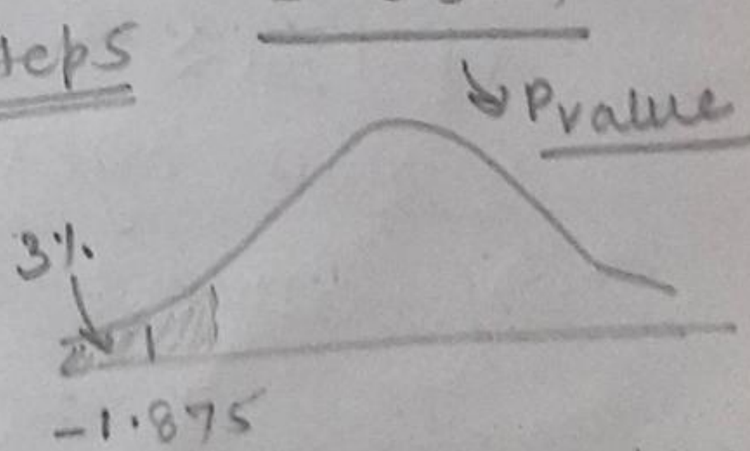
$P(z_{stat}) < -1.875$

$= .0304$

Step 4



Step 5



P value $< \alpha$

↳ falling in Rejection region

Hence at 5 % confidence level, he is selling underwt. (Null hypothesis Rejected)

(27) State null & Alternative hypothesis for the following

a) Is the avg. waiting time for the customer of super market at checkout greater than 15 min

$H_0 \rightarrow \mu = 15$; $H_1 \rightarrow \mu > 15$ $\left(Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)$

b) is proportion of household owning color TV in Chennai is less than 0.4 (proportion is always in % and μ is always a number)

$H_1 \rightarrow \pi > 0.4$ $\quad H_0 \rightarrow \pi = 0.4$

$Z = (p - \pi) / \sqrt{\pi(1-\pi)/n}$

c) Is the avg. expenditure per household on eating out significantly higher in Bangalore than in calcutta.

↳ it involves comparing two population

$H_0 \Rightarrow \mu_1 = \mu_2$ $\quad \mu_1 \rightarrow$ Bangalore eating

$H_1 \Rightarrow \mu_1 > \mu_2$ $\quad \mu_2 \rightarrow$ calcutta eating

$H_0 \rightarrow \mu_1 - \mu_2 = 0$

$H_1 \rightarrow \mu_1 - \mu_2 > 0$ } T-test compare

d) Two Random sample survey conducted with 2 month gap to assess public opinion on outcome. The question that was posted was "If the general election going to take place tomorrow, would you cast vote for or against the ruling party"

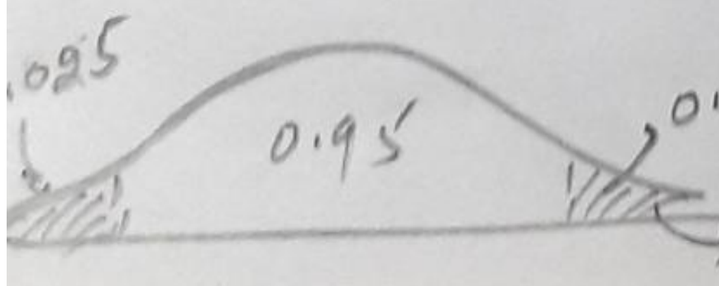
↳ (this is a opinion poll problem)

$H_0 \quad \pi_1 = \pi_2$

$H_1 \quad \pi_1 \neq \pi_2$ } Two tail test

$\pi_1 \rightarrow$ Proportion of people in first poll vote for Ruling party

$\pi_2 \rightarrow$ proportion of people in 2nd poll



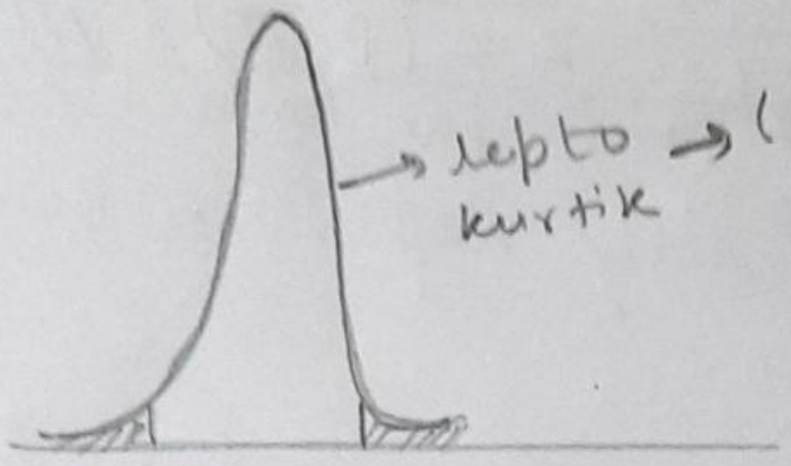
Use case (for eg. in case of scientific tests)

if $n < 30$ & σ is unknown (Population std dev)

* To do T-test, assumption is distribution is normal

* Use T-test → $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$s \rightarrow$ sample std. dev.



→ (tail is flatter & more points are in the center)

T-test application - one sample

Experian marketing Services reported that the typical American spend a mean of 144 min per day accessing Internet via mobile. In order to test validity of the name, you select a sample of 30 friends & family. The result for time spent per day accessing Internet via mobile device are stored in InternetMobileTime. → Excel file name

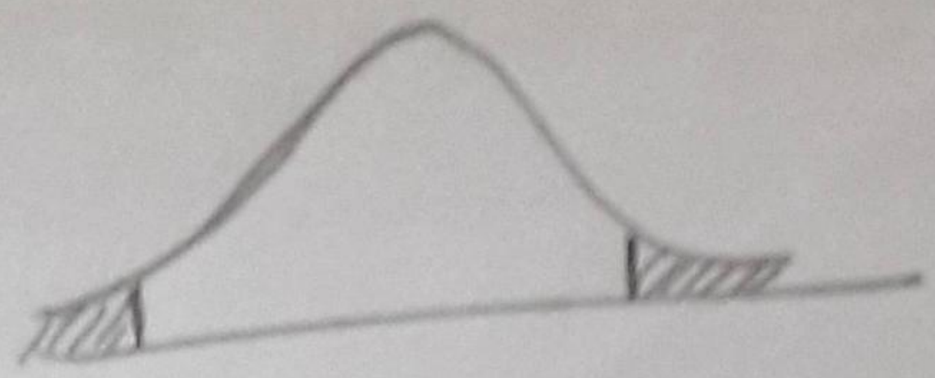
a. Is there evidence that the population mean time per day is diff from 144 min. Use p-value approach and level of significance is 0.05.

b. what assumption about population distribution is needed in order to conduct the t-Test in (a)?

(29)

$H_0 \rightarrow \mu = 144$

$H_1 \rightarrow \mu \neq 144$ ↳ Two tail Test



$\bar{x}$ = mean (of all entries of given excel sample)
→ 175.2667

s = std dev (of all entries of given excel sample)
→ 139.8368

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \rightarrow 1.2246674$$

Pvalue = 2 × Pvalue(t) → (due to two tail test)
left tail + Right tail.

↓
calculated as 0.2305

Pvalue > α → Null hypothesis is accepted
(23%) (5%)

↓
means it's not diff from 144 minute