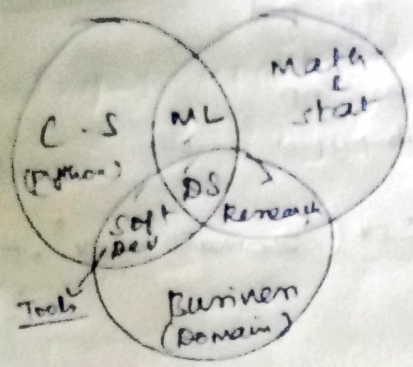
Statistics → Numerical Science



Data → collection of facts

Statistics

Statistics that "It is a discipline that concerns the collection, org, analysis & interpretation & presentation of data

collect, clean, analysis, Interpret (insights) (through vizual or mathematical (pandas, expln)

Data is X
Data ✓

What is Statistics

Data

↓

facts (raw fact), especially numerical facts, collected together for reference or information

Information

knowledge communicated concerning some particular fact

- categorical → Numerical → statistics

↓

(other than numerical)

Q A

Chi square → Categorical → Categorical

↓

can convert see relationship b/w categorical

Eg: Features (Q) Answer

X Y

Statistics in real world

Industry → supplements → product → mass, weight, color, price
(for body building) 1 kg

→ daily production
1 L products

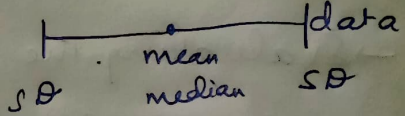
Industry → packet → 5 biscuits → 1 L product/box

① Population (whole) or group → 1 L product of biscuits
Other name of population: parameter

② Sample → Random selected products (i.e., 1000/1 L product)
Other name of sample: statistic

wt = 100 mg → ± 12 mg

S.D = Distance = √Var → range → spreadness of data



Central Tendency → mean, median, mode

SD, var → Measure of Dispersion (Spread)

→ 1 L → sample → weight → 95–105 sd.
(1000 products randomly selected)

Population → Whole set
→ We can't make analysis easily

Sample → Subset of population (lesser in groups when compare to the population)
→ Easy to analysis
hence it is a small group

Strategy (sample)

- Sample results always resemble the population approximately.

Why sample?

Population → Time, money waste & hard to do.

Sample → Easy to analyse

How to take sample :-

i) Random sampling

ii) Intervals

- At frequent intervals

iii) Population

Sample data should be similar to population

If sample is wrong, then it will not resemble the population

i.e., few products → Adopte - for Indians but not for Americans

Should also consider about climate conditions, day/night time, season

Types → Inferential → Resembles, confidence Interval

Descriptive → Describes above the data

Analysis → Interpretation

→ presentation

Types of data - i) Quantitative (numbers) data

ii) Qualitative data

↓

Categoricals.

Quantitative → discrete → whole nos

continuous → decimal nos

↓

Heights.

Fluctuates

↓

numerical

Qualitative → Categorical values

↓ no numerical values

→ Nominal data
Eg: Men, women
hair colour,
→ can be considered both as quantitative or qualitative because it can be converted into ordinal data

→ Ordinal data → Interconnection b/w data
Eg: First, second, third, Letter grades, Economic status: low, medium

- Nominal data & ordinal data has to be changed into number. (Numerical form)

4 Different types of Analytics :→ science of logical analysis

Tells about future
predicting or result division of whole into small components

i) Descriptive → describing above the problem
ii) Diagnostic → diagnosing based on descriptive
iii) Predictive → prediction
iv) Prescriptive

Difficulty increases ↓

Variable & Random Variable
↳ changes randomly (because of uncertainty)

Discrete → number
cont.

measured or counted, they can be weight, height etc.

i) Numerical variable ii) categorical variable

Descriptive

describing → avg, Median, Sd

1) Analysis data - helps to describe & summarize

ii) Organize, analyse & present data in a meaningful way

iii) Used to describe a situation

iv) Explain already known data & limited to a sample or population

Inferential

(Researcher start)

(conclusion) based on sample.

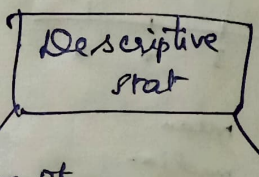
Used @ feature selection

1) Analysis of random sample of data & make inference about the population.

ii) Compare, test & predict data

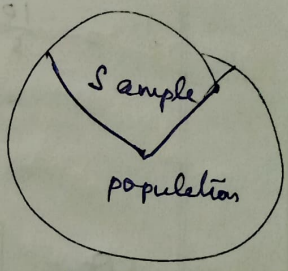
iii) Used to explain the chance of occurence of an event.

iv) Attempts to reach the conclusion



Measures of central tendency (mean, median, mode) (with the help of pandas)

Measures of Dispersion Variability (Variance, Range, Std. deviation)



- Estimation of parameters

- results shown in charts, graphs.

- results are shown with prob. scores

Inference abt population with the help of sample

Data should be in Numerical

central tendency → @ centre (mean, median)

↓

avg

data = 1, 2, 3, 4, 5

Sample, mean = $\bar{x}$

population, " = μ

Mean = Avg = $\frac{15}{5} = 3$

median = middle value

even = $\frac{\text{central two terms}}{2}$

Mode = frequency = repeated values

Standard Deviation

Measure of Dispersion → spread

$$\text{Var}, \sigma^2 = \frac{\Sigma (x_i - \mu)^2}{n}$$

$$\text{Std. deviation}, \sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{n}}$$

$$\text{Std. deviation} = \sqrt{\text{Var}}$$

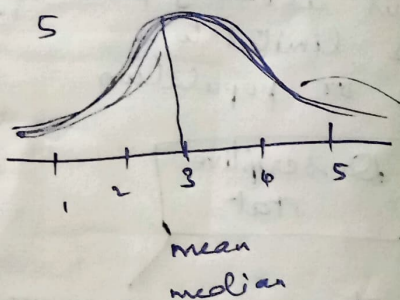
Data = [1, 2, 3, 4, 5]

$$\text{Var} = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

(5 → mean value) → because the distance shouldn't be negative

$$= \frac{10}{5} = 2$$

$$\pm 2$$



Data is good

↓

Because, this data helps us to do remaining ones

$$SD = \sqrt{\frac{4+1+0+1+4}{5}}$$

Real world data ⇒ falls on distributions

↓

Each data follows a pattern

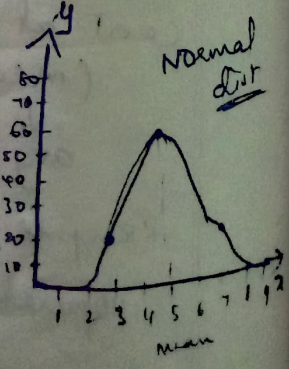
Normal Distribution

- Data analysis

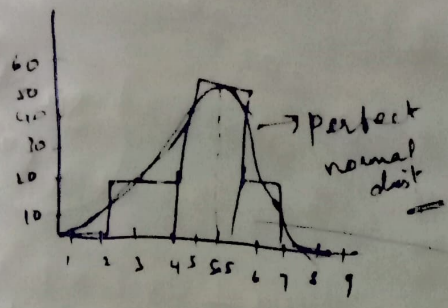
Eg

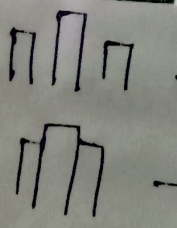
i) Height of the class

	Small	Normal	Big
Count	20	60	20
Height	3	5	7



histogram





→ histogram

max^m pattern in the world follows normal dist

SD

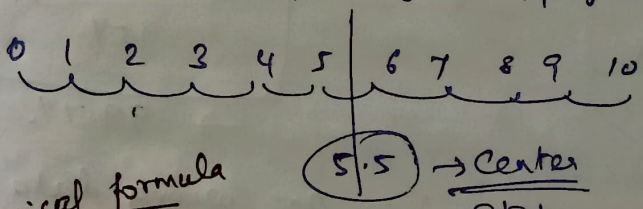
sample

A statistic that measures the dispersion of a dataset → relative to its mean

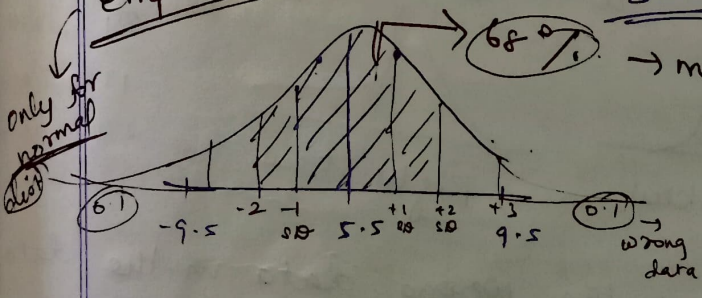
$\frac{10}{2} = 5$ math

$\frac{10}{2}$ (5.5) Stat (Numerical. Science)

5.5-5.6 => probably b/w => physicist



Empirical formula



Use of studying Central Tendency & Measure of Dispersion

Empirical formula

Normal Distribution

Central Tendency

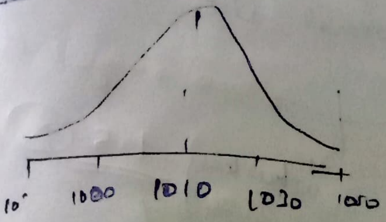
(Mean, median & mode)

↓ avg

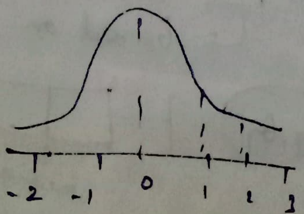
↓ All will be in one straight line @ centre (co-r-to dispersion)

Measure of Dispersion

Variance, sd, Range



normal dist
(No range for its values)



Std. normal dist
↓
mean = 0
std. deviation = 1

F test → Fisher Test

Statistics

Inferential: Sample → Popu

Descriptive Stat

- Central Tendency
  - Mean
  - median
  - mode
- Measure of Dispersion
  - S.D
  - Variance
  - Range

Inferential Stat

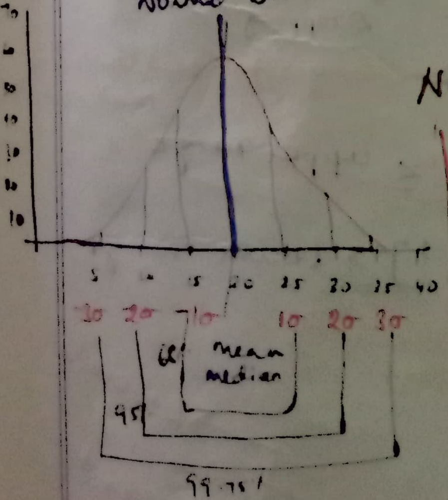
- Z-test → 2 column comparison
- t-test → 2 column → num-to num
- correlation / covariance
  - Spearman rank corr
  - pearson correlation
- Chi² → categorical-categorical
- Annova (f-test) → Fisher

→ Only in feature selection (X), not in Y (Ans)

Distribution

Eg mean = 20 std = 2

Normal Dist



68% of data fall under 1st std

Normal Dist } = Mean = median = mode

⇒ To find out outlier (>3σ) → boxplot
⇒ Handle null value

→ Ofc care → Outlier Boxplot

distplot → histogram + kde
↳ (kind = kde)

Types of Distribution

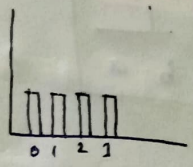
Discrete

i) Uniform distribution ③
ii) Binomial ② " → multiple trial
iii) Bernoulli ① " → only 2 possible outcomes
↳ In a single trial
iv) Poisson distribution ④
↳ In a controlled environment

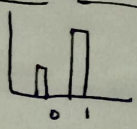
Continuous

i) Normal Distribution (or) Gaussian dist
ii) Standard Normal distribution — Z test
iii) Student T distribution — T test
iv) Chi square distribution

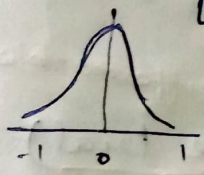
Uniform.



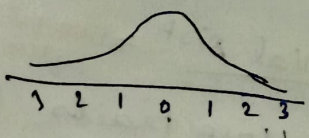
Binomial Bernoulli:



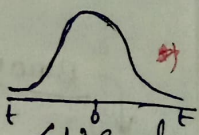
Std normal dist
↳ 0 to 1



normal dist.



→ 0 to ∞
T-dist



< 30 samples
small sample size

Student-t distribution → degree of freedom → less sample (0 to ∞)

Binomial

Fair coin → Tail & Head

Bernoulli's distribution → 2 possible outcomes (Success, Failure)
→ single trial
only two outcomes

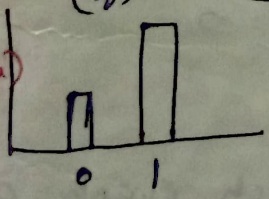
pmf

probability need not to be equal

Success = (p) = 0.5 (P)
Failure = (1-p) = 1 - 0.5 = 0.5 (q)

Mean = p

X → random variable



Success ≠ Failure

mean $\Rightarrow E(x) = \sum_{i=1}^{n} x \cdot p(x)$

Variance = p(1-p) = pq

S.D = $\sqrt{pq}$

prob(failure)

$$P(x=x) = p^x (1-p)^{1-x}, \quad x = 0, 1$$

$$p(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

↑ prob(success)

pmf

↓

Because outcome is a discrete

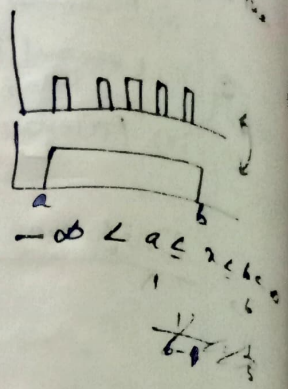
$$PMF = \begin{cases} q = 1-p, & \text{if } x = 0 \\ p, & \text{if } x = 1 \end{cases}$$

Uniform Dist

- more possible outcomes

Eg: Dice (1 to 6)

possible outcomes depends upon the material.



- Same probability

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

Distribution by the Analog voltage

Expected value * orginal value

$$(30-15) * \frac{1}{40-10} = \frac{15}{30} = \frac{1}{2} = 0.5$$

Binomial Dist

→ Two potential outcomes per trial → Difference b/w Bernoulli & Binomial

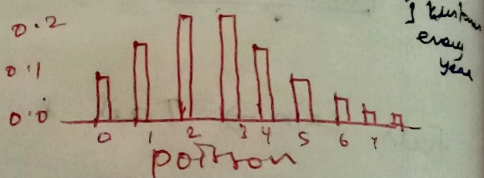
derived from Bernoulli

- More trails (multiple trial)
- Each trial is independent
- Trial (only two possible outcomes)

P(Success) is same across all trials

Poisson Dist

$\lambda = 3$ → Eg: 3 times every year



- discrete distribution
- Describes the no. of events occurring in a fixed time interval or region of opportunity
- Requires only one parameter, λ
- Bounded by $0 \leq \infty$

Assumptions

- The rate at which events occur is constant
- Occurrence of one event does not affect the occurrence of a subsequent event (i.e., events are independent)

Q) Coin - 1 time toss → Bernoulli's
" - more time toss → Binomial

Poisson

- parameter, λ (Expected rate of occurrence)
- prob. of given no. of events in a fixed time interval. (within a given time interval)

Eg: no. of calls attend in 1 hour at call center.

Rules

① 1 Succes → 2 succes (not influence)

1 hr (30 calls) 35 calls 2nd hr.

② Small Time → Succes rate
long time → Succes rate

Small time (1 hr) =	Ram	Raja
	25	30 ✓
	10	20 ✓
✓	35	20

P(S) over a short interval is equal to the P(S) over a long time interval

Raja → day succes
↓
Because attended more calls.

③ Interval becomes smaller because the interval approaches zero.

→ Prob. Density fn
PDF → continuous → parametric (normal dist)
outcomes → cont. value
non-parametric
↳ chi² dist

PMF → Discrete
outcomes → discrete value
↓
Prob. Mass fn

CDF ↓ Cumulative density fn

Student t test

Annova → Z test + t test
(f test) 1 on 1 1 on 1
↓
Compares more than 1

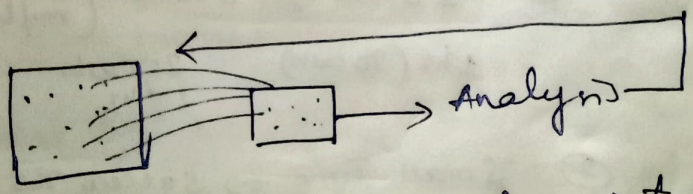
Inferential statistics

To draw Inferencing about the population from sample data

Inference → Conclusion reached on the basis of evidence and reasoning.

Hypothesis

To draw Inference about the population data to be from sample data.



Describing data → descriptive stat

Taking sample & concluding from population → Inferential stat

more than 3 value comparison → ANOVA

ANOVA compares 3 values

t-test → better than z-test

population → sample → Analysis → Result

Inferences about population

It should resembles the population.

z test, t test, chi²

Study Buddy

1st time result check → correct → Null hypothesis

2nd time " " → incorrect → Alternate hypothesis (Researcher hypothesis)

Based on z test, t test, chi²

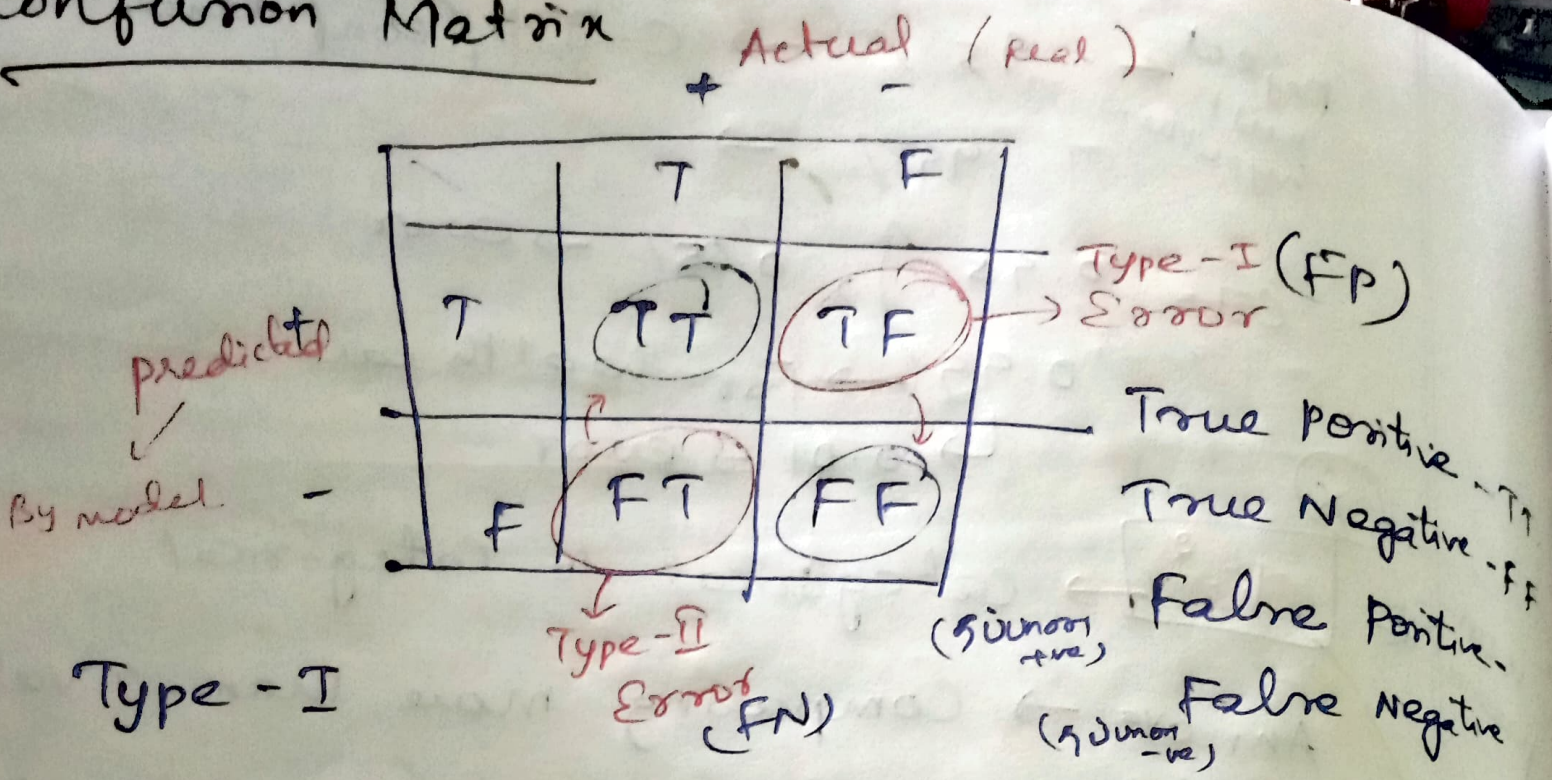
i) Reject null hypothesis

ii) Failed to reject null hypothesis

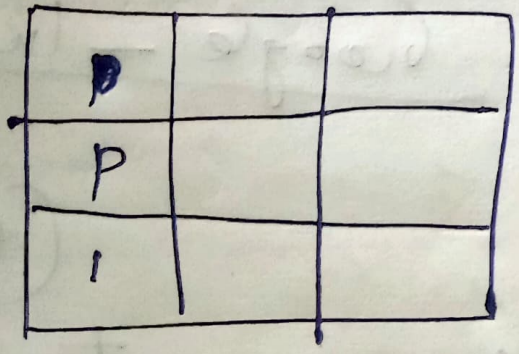
Assumption → No Differences → Null hypothesis

Difference → (Accept / Failed to reject Null hypothesis)

Confusion Matrix



Type - I



Actual	predict → (should consider this one).
T	T = T T
F	F = F F
F	T = Falre T → Falre +ve (negative 2y Error)
T	F = Falre F → Falre negative (positive 2y Error)

Statisticshowto.com

19/11 (✗) I.Q

hypothesis testing in feature selection
↳ Refer in google (analytics vidhya)

Feature Selection

Null hypothesis [Relationship] ⇒ same (i.e, relation)

- There's no significant difference b/w pop or feature
- Assume that the mean of two sample is equal
- Both are same (i.e, relationship)

Alternative hypothesis (no relationship)

Based on p value

Critical value — ↑ higher the prop

P value → At which pt, prob. of i.e. null hypothesis being true

↓ = Stronger the evidence of rejecting the null hypothesis
lower p value

prob - is always 100's
↓
0's & 1's.

0   0.05   1

Reject null hypo	Failed to reject (Accept Null hypothesis)
< 0.05	>= 0.05
p value is ↓	p value is ↑

A1 → Target	A2 → Target	A3 → Target
0.2	0.5	0.8
no relation	relationship	relationship

Failed to reject null hypothesis

1. One - on one prediction → for Public
2. Event prediction → Industries, hospitals.
3. Batch "
Auto Lweekly once

4. streaming → @ Stock Market

minimizing for scaling the data → curse of dimentionality

multi-collinearity → features Corion

∴ multi-collinearity Bias error.

Confidence Interval = 100 − p value

Annova → can compare more than two variables

Hypothesis Testing

Statistics for data science

→ p value, NCbi → study abt hypothesis, Type-I & Type-II.

→ Biostatistics

Type-I error :

(False-positive) → Actual False, predicted True

→ Occurs if an investigator rejects a null hypothesis that is actually true in the population.

Type-II Error (False-Negative)

- occurs if the investigator fails to reject a null hypothesis that is actually false in the population

Inferential → hypothesis testing

Feature → Target ⇒ should have relationship

B/w two features ⇒ NO relationship should be there.

p value ⇒ 0.05 ⇒ may get varied

z test ⇒ Converting normal dist into std. normal dist.

Level of Significance - α (alpha)

$\alpha = 5\%$ risk ⇒ Okay to take 5% risk

$= 0.05$ error are allowed.

vary for medical data ($\alpha = 0.01$)

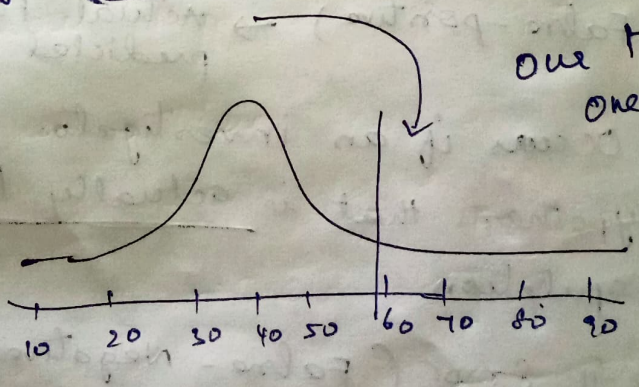
Critical value (C) $= 100 - \alpha$

p-value →

Single-tail test → Focusing on single side

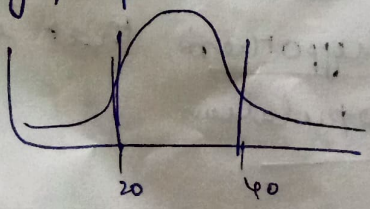
two-tail test

i) ppl got 60+ marks → score high ⇒ null hypothesis

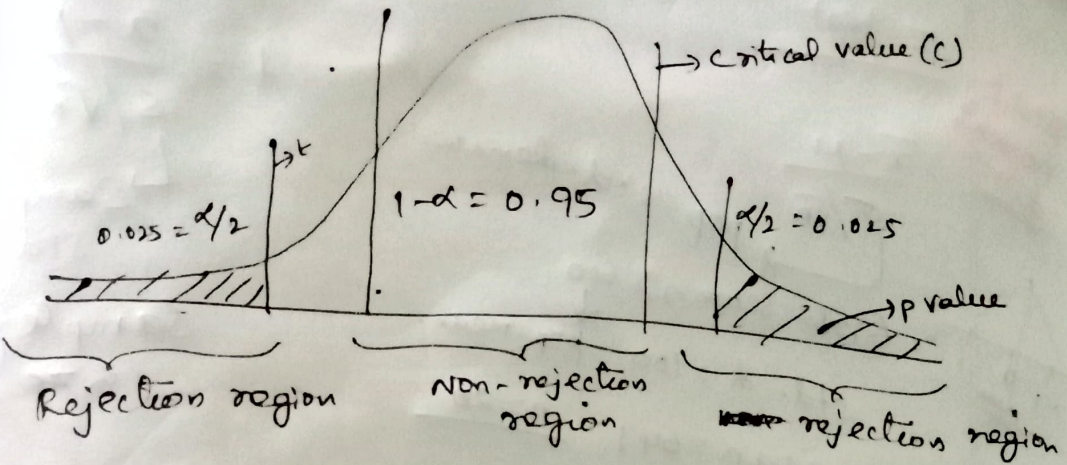


one target is one side alone

ii) age gap of 20-40 smokers for



⇒ two tail test



$t > c$, → reject H_0

$t < c$ → failed to reject H_0

$P > \alpha$ → fail to reject H_0 $P \leq \alpha$ → reject H_0

Hypothesis

1) Define null, Alternative

2) α → allowed error, $CI = 1 - \alpha$

$= 1 - 0.05$

$= 0.95$

→ % proportion or mean?

→ one tail or two tail?

3) score → which test have to choose?

- critical value

z, t, f, chi, ...

Eg: 16 ± 2 => 16 20

↓

B/w this H_0 is true

4) p value

- prob of being H_0 is true

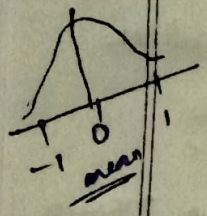
	H_0	H_a
H_0	NO error	Type II error (β)
H_a	Type I error	NO error

z test

Std. normal dist

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

σ → pop. std. deviation



Girls score => > 600

std. d = 100 (pop. mean)

mean = 641

$n = 20$

$$\text{Score} \Rightarrow \frac{641 - 600}{100/\sqrt{20}}$$

$$= 1.8336.$$

In z table,

p value = 0.033

↓

we reject H_0

(∵ p value < 0.033)

↳ Hence the given values are false.

T test

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

↳ Sample std. deviation

Sample size => < 30

If it is greater than 30, Use z-test