Statistics

=> what is statistics?

It is a science of collecting, organizing analyzing data for better decision making

=> what is data?

facts or pieces of information that can be measured

Eg

Age of students

{24, 22, 21, - - - 50}

=> Types of stats:

there are two types:

(i) Descriptive stats: It consist of organizing & summari

(ii) Inferential stats: using data, we can make conclu using some techniques

Eg:

class - 20 students

1st sem maths {86, 70, 90, 55 - -}

1. what is the avg of class? [Descriptive]

2. 7th sem [inferential]

Sample and Population

Sample : A Sample is a subset of a Population selected
→ for analysis in statistics
→ Small subsets of data taken from Population

Population : whole dataset is known as Population

* Sampling techniques:
  method to select a subset from a Population

1. Simple Random sampling:
   Every member of population has an equal chance of getting selected in sample
   Equal chance selection from a populat

2. Stratified sampling:
   Splitting data into non-overlapping groups
   Picking samples from different groups fair

Eg:

Age group < 0-20, 20-40

Gender < M, F

3. Systematic Sampling:

Systematic Sampling selects every nth item from a list after a random starting point

4. Convenience Sampling:

Selecting Samples based on convenience

Variables:

* A variable is something that can change or have different value.
* It is a property than can hold / store/take any value

Eg: Age {8, 10, 15, 20 - - 25 - -}

Marks: {76, 80, 95 - -}

Types of variables:

1. Qualitative: [based on some characteristics we can derive categorical value

Eg: IQ :- 0 - 10 → low →

10 - 50 Avg

50 > 0 Good

2. Quantitative: Numerical value (measurable numerically)

Height : {162, 159, 155 - -}

weight : {59, 65, 79 - - }

Discrete (int) whole no

continuous (float) Decimal No

1. No. of students
2. No. of bank.

Height
{165.2, 167.9 - - }
weight
{165.5, 60.9 - - - }

1. Blood pressure — continuous/ Discrete
2. marital states — Qualitative
3. River length — continuous
4. Song length — continuous
5. Gender — Qualitative (category)

Variable measurement scales

order matter

1. Ordinal - ordered [rank, graduation
2. Nominal - categorical values [colors, classes, degree]
3. Interval + [No zero/absolute point] [order as well as value matter]
4. Ratio - zero means nothing.

→ categories data without a specific order (e.g. gender, color)

→ Data with meaningful order and equal differences between values but no true zero (e.g temperature in celsius)

→ Data with all the properties of interval data plus a true zero point (e.g. weight, height, income

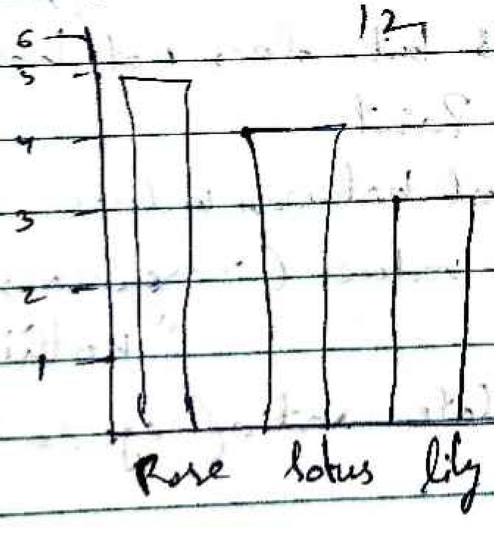
II what is frequency:

frequency refers to the number of times a specific value or event occur in a data set or observation

data: flowers

[Rose, lily, lotus, rose, rose, rose, rose, lotus lily, lily, lotus, lily lotus]

flowers	frequency	cumulative f
Rose	5	5
lily	3	8
lotus	4	12
	12	



bar graph chart

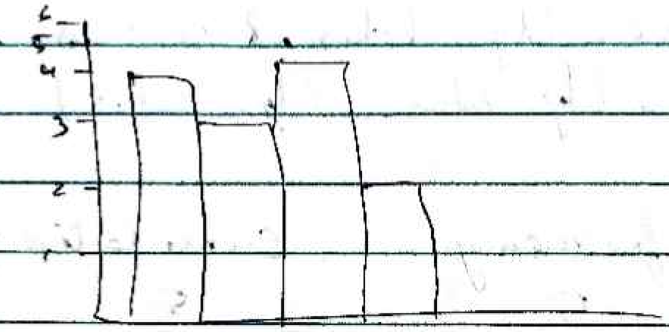
Histogram

A histogram is a bar chart showing data distribution in intervals.

Marks [12, 15, 15, 21, 28, 27, 35, 34, 36, 39, 42, 45]

bins

(10-20) - 4
(20-30) - 3 continuous
(30-40) - 4
(40-50) - 2



Interval scale -> It measures the difference b/w values but does not have a zero point.

-> equal interval between values
-> No absolute value (zero does not mean "nothing")
-> Cannot calculate ratio (eg. 20 is twice of 10)

Eg: Temp
0°C does not mean "no temperature"
20°C is not "twice as hot as" 10°C

Calender
Diff b/w 2000 & 2020 is (20 years) but 2020 is not twice of 1010

Ratio Scale

It measures the data where there is zero point meaning zero represents complete absence

-> Equal interval b/w values
-> zero allows ratios
-> can perform all maths opn

Eg: Height & weight
Time Duration

=> Measure of central Tendency
=> Measure of Dispersion
=> Distribution

Measure of central Tendency:

Avg -> mean

Pop Sam

$\mu = \frac{\Sigma x_i}{N}$ $\bar{x} = \frac{\Sigma x_i}{n}$

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

$\frac{2+3+5+3+2+1+3}{7}$

Mean: It refers to the measure used to determine the center of the distribution of the data

$\{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$

outlier

$\frac{32}{10} = 3.2 \Rightarrow \frac{132}{11} \Rightarrow 12$

median → middle value

↳ ascending order

↳ data $<^{\text{even}}_{\text{odd}}$

Even :

$$\frac{\left(\frac{n}{2}\right)^{th} + \left(\left(\frac{n}{2}\right)^{th} + 1\right)^{th}}{2}$$

dataset : $\{11, 12, 13, 14, 15, 16\}$

$n = 6$

$$\frac{\left(\frac{6}{2}\right)^{th} + \left(\frac{6}{2}\right) + 1}{2}$$

$$\frac{3^{th} + (3+1)^{th}}{2}$$

$$\frac{3^{th} + 4^{th}}{2} = \frac{13 + 14}{2}$$

Median = 13.5

$$= \frac{27}{2}$$

$$= 13.5$$

odd :

$$\frac{(n+1)^{th}}{2}$$

$$\{11, 12, 13, 14, 15\}$$

$$n = 5$$

$$\frac{(5+1)^{th}}{2} = \frac{6^{th}}{2} \Rightarrow 3 \Rightarrow 13$$

$$\{11, 12, 13, 14, 15, 100\}$$

$$\frac{\left(\frac{n}{2}\right)^{th} + \left(\left(\frac{n}{2}\right)^{th} + 1\right)^{th}}{2}$$

$$\frac{\left(\frac{6}{2}\right)^{th} + \left(\left(\frac{6}{2}\right)^{th} + 1\right)^{th}}{2}$$

$$\frac{3^{th} + 4^{th}}{2} = \frac{13 + 14}{2} = 13.5$$

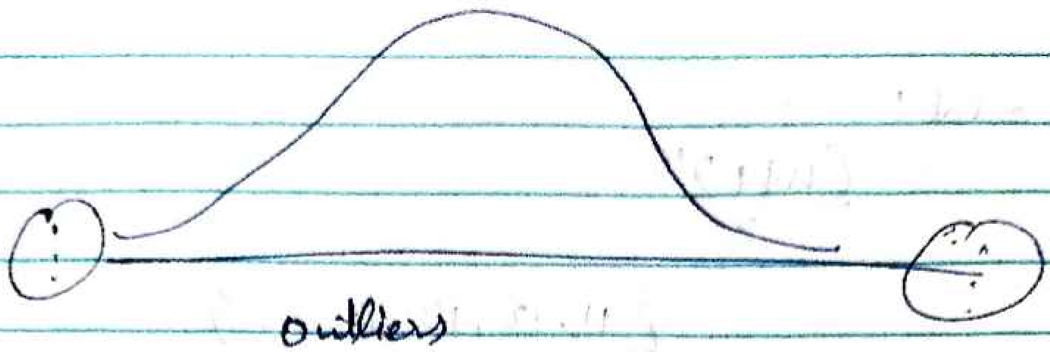
$$\{21, 23, 25, 29, 32, 100\}$$

$$\frac{25 + 29}{2} = \frac{54}{2} = 27$$

median works well with outliers

outlier :

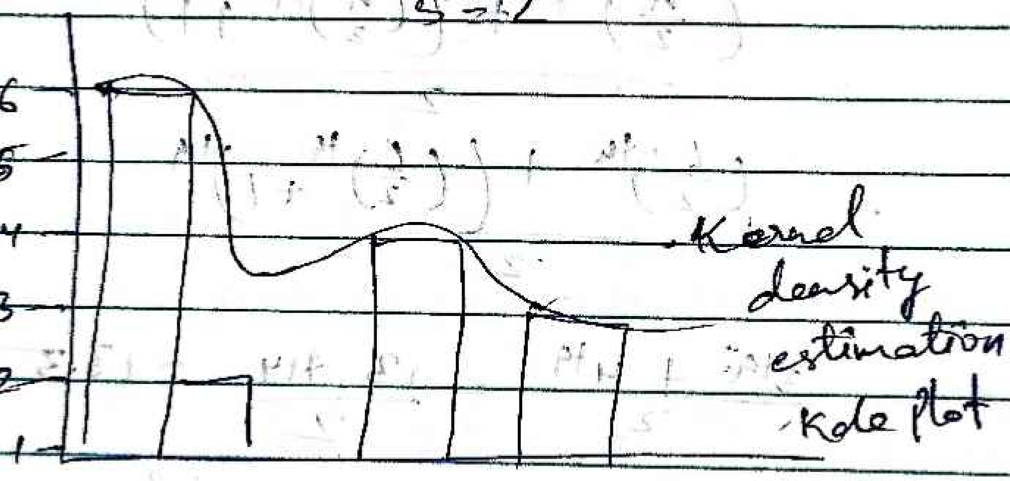
a datapoint(s) who doesn't follow pattern or trend of the dataset then it is considered as outlier [are extreme points]



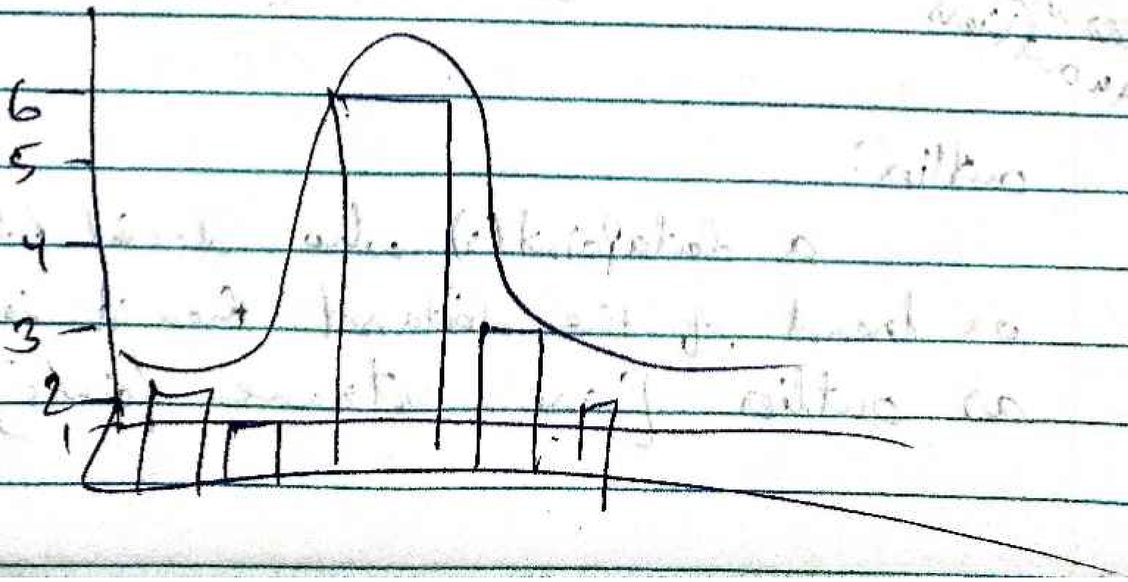
→ Mode most frequent value (repeated)

[1, 1, 2, 3, 5, 1, 1, 3, 13, 5, 1]

1 = 6 3 = 3
2 = 1 5 = 2



{1, 1, 2, 5, 9, 5, 7, 8, 9, 5, 9, 5]



for categorical missing data

0 - 5% → } mode
→ }

< new category "missing"
"unknown"
"random"

species

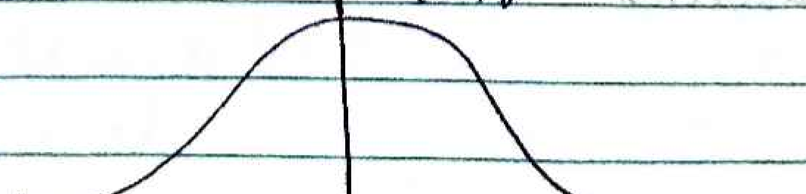
Rose	lily	lily	
lotus	lily		
lily	lily	Rose	null
Rose	Rose	Rose	NA
lily	lotus		15-10
Rose	Rose		

for numerical values

Gaussian — skewed

Normal Distribution
↓
mean

skewed
↓
median
0-5%
10%



Mean = median = mode

Measure of Dispersion → spread

1, 5

[5, 1, 1, 1, 2] → $\frac{5+5}{5} = 2$

[2, 2, 2, 2, 2] $= 2$

2

variance

It measures how far the numbers in a dataset are from the mean (avg) (how each value differs from a dataset in a mean)

High variance → more spread (far from the mean)

low variance → closer to mean

Pop

$$\sigma^2 = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{N}\right)^2 \qquad s^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{n-1}\right)^2$$

Summation

Bessel's correction

x	$x_i - \mu$	
1	-1.83	3.34
2	-0.83	0.69
2	0.17	0.69
3	0.17	0.02
4	2.17	1.36

$$\mu = \frac{1+2+2+3+4+5}{6} = 2.83$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N} = \frac{10.8}{6}$$

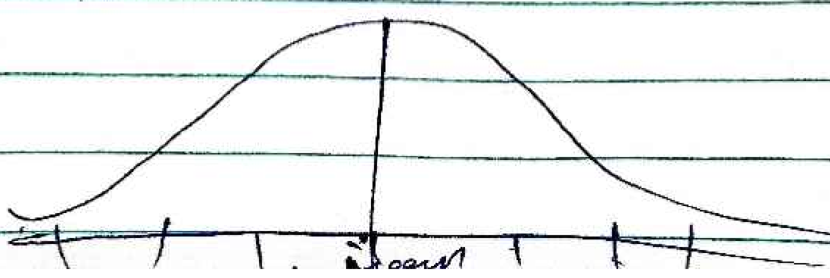
$$\sigma^2 = 1.8$$

x_i → every data pt
μ → mean of pop
N → ~~to~~ total pop
Σ — summation

Standard Deviation [unit same easily comparable]

Pop

$$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}} \qquad S = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}$$



68%
95%   99.7%

$\mu \pm 1\sigma = 68\%$
$\mu \pm 2\sigma = 95\%$
$\mu \pm 3\sigma = 99.7\%$

-> Squareroot of variance
-> It gives measure of spread that is in the same units as the original data, making it easier to interpret

* variance formula
Population : $\frac{\Sigma (x-\mu)^2}{N}$ (N = total of data pts)

when we have data from the entire population we use 'N' in the denominator. This gives us an exact measure how the data pts vary around the population mean (μ)

* sample

$$\frac{\Sigma (x_i - \bar{x})^2}{n-1}$$

when we're working with a sample, we only have sample mean $\bar{x}$, which is an estimate of the population mean μ
using sample mean in calculation tend to make the variance slightly smaller than the true population variance
To correct this bias (underestimating variability) we divide by 'n-1' instead of n, this makes variance estimate larger & accurate

* Percentage:

maths	- 88	100
SS	- 75	
Sci	- 90	
Eng	- 80	
GK	- 99	

$\frac{88+75+90+80+99}{500} \times 100$ 86.4

* Percentile → It is a value below which a certain % of observation lie.
ascending
dataset = 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8
n = 20 10, 11, 11, 1
Percentile rank of 10 = $\frac{\text{no. of values below } x}{n} = \frac{16}{20}$
= 80

80% of values are below 10

" $= \frac{17}{20} \times 100 = 85$

85% value are below 11

what value exists at Percentile rank of 25?

$$\text{value} = \left(\frac{\text{Percentile}}{100} \times n\right) + 1$$

$$\left(\frac{25}{\cancel{100}} \times \cancel{20}\right) + 1$$

$5 + 1 = 6$ index values

$$75 = \left(\frac{75}{\cancel{100}} \times \cancel{20}\right) + 1$$

$= 15 + 1$

$= 16 \rightarrow 9$

five Number summary

1. Minimum Q_0
2. 25 Percentile → first Quartile $[Q_1]$
3. Median → 50 Percentile Q_2
4. 75 Percentile Third Quartile $[Q_3]$
5. Maximum Q_4

[1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 15, 27]

Lower fence $= Q_1 - 1.5\,(IQR)$

Higher fence $= Q_3 + 1.5\,(IQR)$

IQR → $Q_3 - Q_1$ → Inter Quartile Range

$Q_1 = 3$

$Q_3 = 8$

$IQR = 8 - 3 = 5$

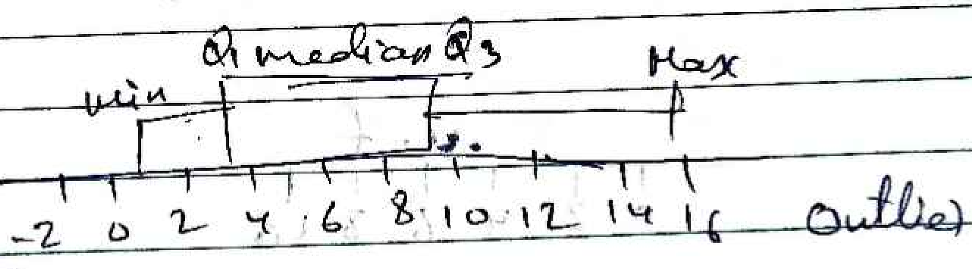
lower fence $= 3 - 1.5(5)$

$= -4.5$

higher fence $= 8 + 7.5$

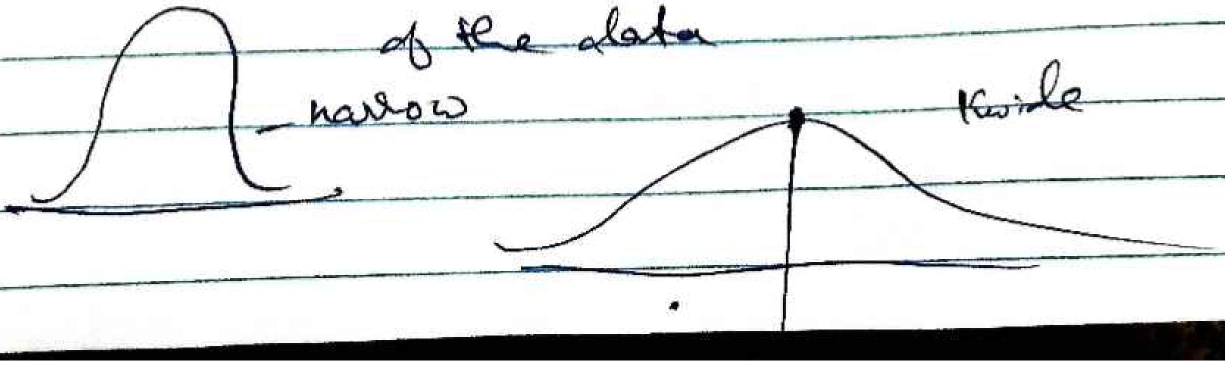
$= 15.5$

1. min = 1
2. $Q_1 = 3$
3. median = 5
4. $Q_3 = 8$
5. max = 15

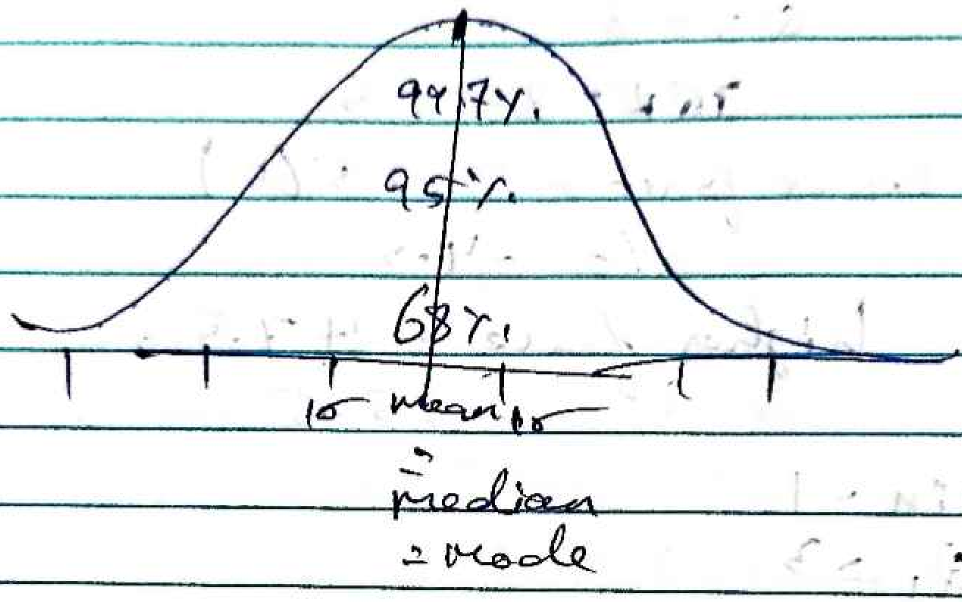


Data Distribution:

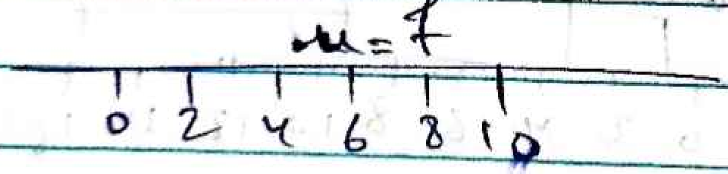
* It refers to a way in which values data pts are spread or arranged
* It shows how after different values occur in data set describes the overall pattern of the data



Gaussian / Normal Distribution



$\mu \pm 1\sigma = 68\%$
$\mu \pm 2\sigma = 95\%$
$\mu \pm 3\sigma = 99.7\%$
} Empirical Rule



Z-score

$$Z\text{-score} = \frac{x_i - \mu}{\sigma} = \_\_\_\_$$

$\mu = 0 \quad \sigma = 1$

Standard Normal Distribution

$\mu = 0 \quad \sigma = 1$

Height	weight	$h_i - \mu_n$	
169	60	3.2	10.24
172	65	6.2	38.44
150	45	-15.8	249.64
168	70	2.2	4.84
170	71	4.2	17.64
829			320.8

$$\mu_n = \frac{829}{5} = 165.8$$

$$\sigma_n^2 = \frac{320.8}{5} = 64.16$$

$$\sigma_n = 8.009$$

$$z_1 = \frac{169 - 165.8}{8.009} = \frac{3.2}{8.009} = 0.399 = 0.4$$

$$z = 6.2 = 0.7$$

$$z_2 = \frac{6.2}{8.009} = 0.7$$

$$\frac{15.8}{8.009} = -1.9$$

$$\frac{2.2}{8.009} = 0.27$$

$$\frac{4.2}{8.009} \quad 0.52$$

Normalization

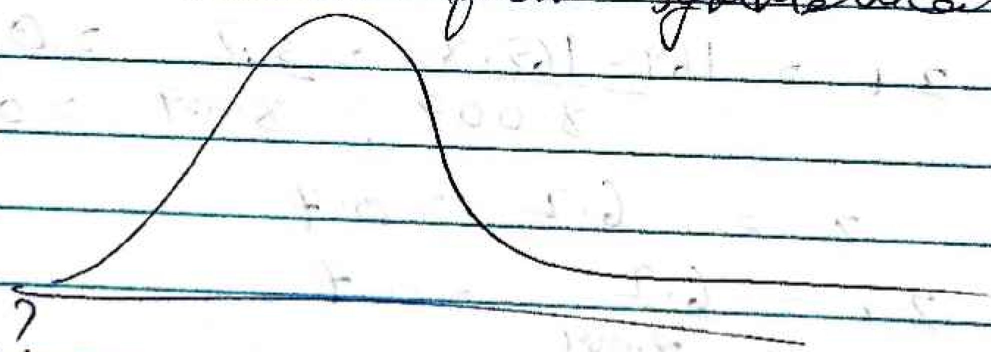
$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}}$$ — min max scaler

* Positively skewed distribution

what?

In positively skewed distribution most values are concentrated on the lower end with a long tail extending to the right. A few high values pulls the average to the right of the median

Skewness: A distribution or asymmetry that deviates from symmetrical bell curve

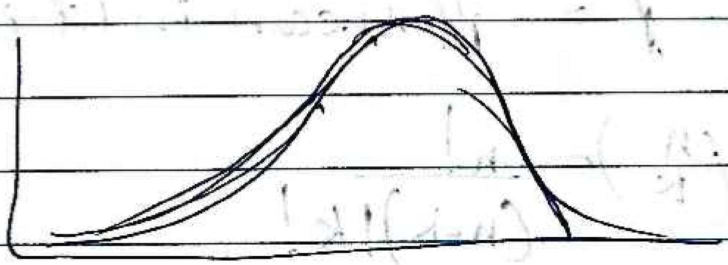


when?

useful for data with rare but significant high values such as income levels few individuals earn much more that rest

negatively skewed distribution:
most values are ~~ch~~ clustered at higher end, with a few values creating a long tail to the left

when?
used for dataset where values are typically higher, but a few lower values exist (retirement age)



Exponential Distribution:
λ = constant rate

It describes the time b/w events in a process where events occur independently & at a constant rate λ

$$f(x) = \lambda e^{-\lambda x} \quad x \geq 0$$

Bernoulli distribution (Discrete)
It models a single experiment with two possible outcomes

success($X=1$), failure($X=0$)

Binomial distribution (Discrete)

Binomial dist extends Bernoulli to n independent trials

$$P(X=k) \Rightarrow \binom{n}{k} p^k (1-p)^{n-k}$$

X = random variable
n - total no of trial
k - no of success $(0 \le k \le n)$
p = P(success in 1 trial 0.5

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

$n=5 \quad k=3 \quad p=0.5$

$$P(3) = \frac{5!}{2!\,3!} \times 0.5^{(3)} \times 0.5^{(2)}$$

uniform distribution: Continuous

$[a.b]$

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

the probability of any value within the range $[a,b]$ is same uniform distribution (Discrete)

all outcomes are equally likely

$P(x) = \frac{1}{n}$ — total no.

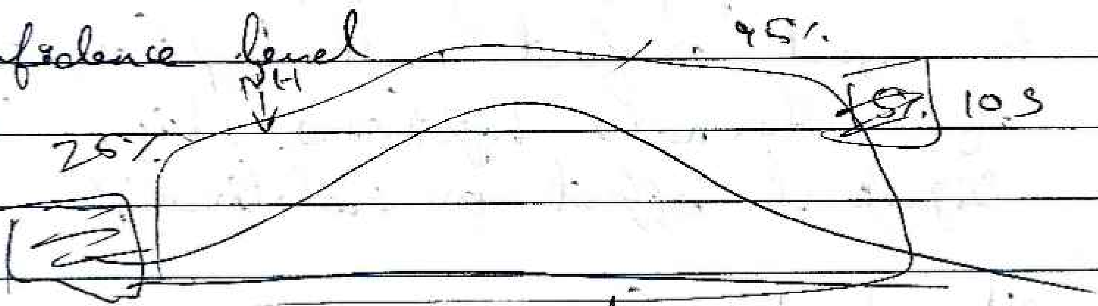
confidence Interval

$\bar{x} = 50 \rightarrow$ point estimate $\Rightarrow \mu$

It is a range of values within which we expect a particular population parameter to fall

confidence Interval = point estimate ± margin of error

confidence level



∴ accept

LOS $\rightarrow \alpha$

Hypothesis Testing

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand is sufficient to support a particular hypothesis

Hypothesis testing allows us to make probabilistic statements about population parameters

Null Hypothesis H0

the null hypothesis assumes that there is no significant relationship or effect b/w two variables. [In simpler terms -> it says nothing new is happening]

It serves as a starting point for HT & represent static que or the assumption of no effect until proven otherwise

the purpose of HT is to gather evidence to reject or fact null hypothesis in favour of altternate hypothesis which claims is significant effect or relationship

* Alternate hypothesis : Ha or H1
It is a statement, that contradicts the NH & claims there is significant effect or relation.

Rejection Region Method

5%
0.5

1. Ho & Ha Los Loss of Significance
2. α -> value
   -> significance level -> 95%
   ND -> $n \geq 30$
   σ -
3. assumptions
4. decide test z-test -
   t-test -
5. value ↓
6. Test conduct
7. Reject / Accept
8. State ~~results~~ results

1. 50 -> units per day
   $\sigma = 5$ 45-55
   training ->
   4
   30 emp -> 53 units per day

1. $\mu = 50$ -> Ho
2. $\alpha = 0.05$ (5%
3. data normal, σ random
   $n = 30$

4. z-test

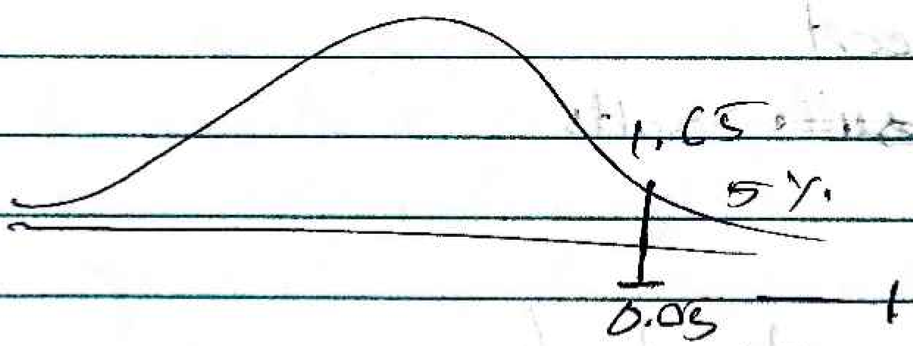
5. z-score = $\frac{x - \mu}{\sigma}$ → Population

$$\frac{x - \mu}{s/\sqrt{n}} = \frac{53 - 50}{5/\sqrt{30}}$$

$$= \frac{3}{5} \times \sqrt{30}$$

$$= \cancel{\frac{9}{25}} \times \cancel{\frac{30}{5}}$$

$\alpha = 5\%$ $Z = 3.28$

6. 

7. rejection of H_0

8. $\mu > 50$

2. avg → 50 g 40 g →

σ - 4 g $\bar{x}$ $n = 49$

σ - 4 g $\bar{x}$ n = 49

1. $H_0: \mu = 50g$ $\quad H_a: \mu \neq 50g$

2. $\alpha = 0.05$

3. $n \Rightarrow 30$ z-test

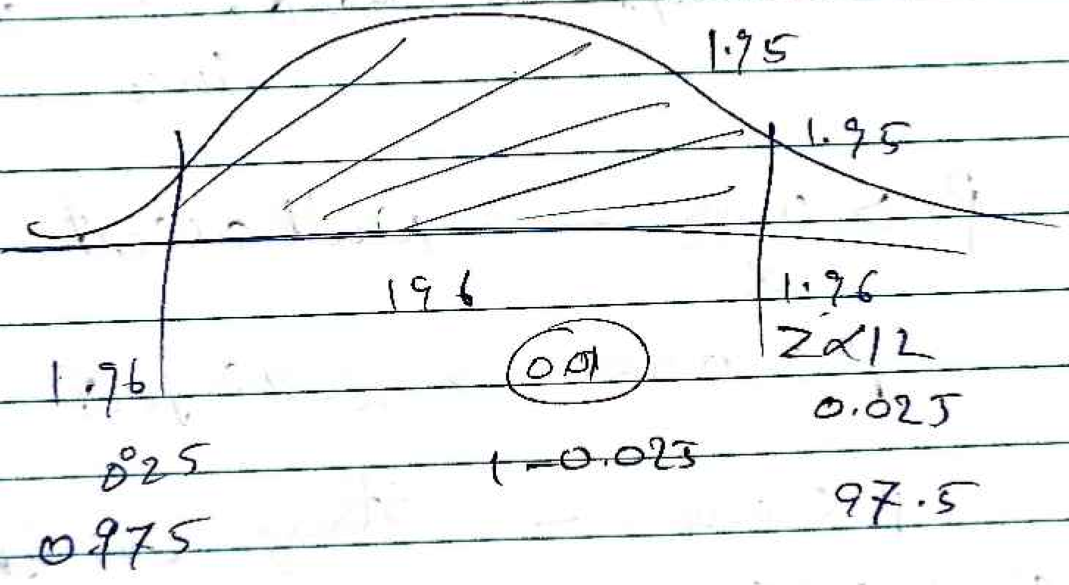
4. z-test

5. $= \frac{49-50}{4} \times \sqrt{40} = -\frac{1}{4} \times \sqrt{40}$

$= -\frac{\sqrt{40}}{4}$

$z = -1.58$

6.



7. Null Hyp Accept

8. $\mu = 50g$

2. Errors

	Type 1	Type-2
	H_0 true	H_0 false
reject H_0	Type-1	correct
accept H_0	correct	Type 2

Type-1 false +ve

$\hookrightarrow$ Ho reject $\rightarrow$ ~~Ho~~ $\hookrightarrow$ Ho True (correct)

rejecting Ho when Ho is actually correct

Type-2 false -ve

accept Ho when Ho is actually incorrect

$P \geq \underline{0.05}$ → Null accept
α

$P < 0.05$ → Null reject

$P < 0.01$ – Strong evidence

$0.01 \leq P < 0.05$ – moderate ev

$0.05 \leq P < 0.1$ → weak evidence

$P \geq 0.1$ → No evidence

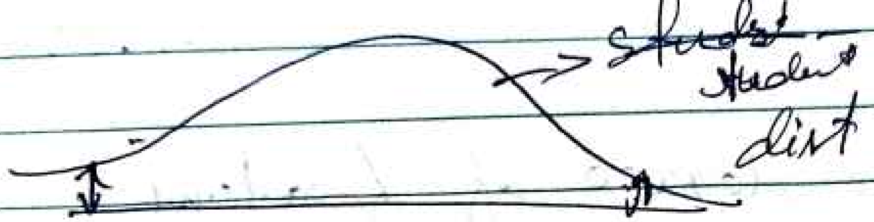
$P < \alpha$ – Ho reject

$P > \alpha$ → Ho accept

P- ~~Value : Test~~ :

P-value -> It is a measure of the strength of the evidence against the Null Hypothesis

T-Test



3 Types of T-test

~~1. ...~~

1. One sample t-test
compares the mean of a single sample to a known μ

$\mu = 80g$ $n = 30$ $\bar{x} = 49.76$
$s = 1.8$

[when we σ - Z-test (Pop - standard dev)
s - t-test (sample - std)]

$\mu = 50g$ $n = \cancel{25}$ $\cancel{std} = 1.2g$
$\bar{x} = 49.7g$

$H_0 : \mu = 50$ $\cancel{H_a : \mu \neq 50}$ $H_a : \mu \neq 50$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{49.7 - 50}{1.2/\sqrt{25}} = \frac{0.3}{0.2} \times 5$$

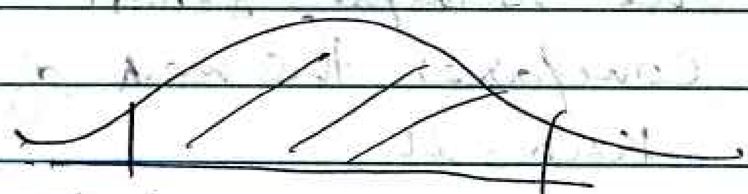
$$= \frac{-1.5}{1.2}$$

$$= -1.25$$

df = degree of freedom

$= n-1 \Rightarrow df = 24$

t critical $= 2.064$



-2.064   2.064

H_0 accept $\rightarrow$ 80%

2- Independent two sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\rightarrow$ standard error

3. Paired t-test

$$t = \frac{\bar{d}}{sd/\sqrt{n}}$$

$\bar{d} \Rightarrow 2 \rightarrow$ mean diff

sd = std $\rightarrow$ diff

Chi - Square test

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

O - observed frequency

E - Expected frequency

$$E = \frac{\text{Row Total} \times \text{Column total}}{\text{Grand Total}}$$

$$df = (r-1) \times (c-1)$$

r = no. of rows

c - no of cols

	Satisfied	Not satisfied	Total
High school	50	70	120
college	90	60	150
P.G	20	10	30
	160	140	300
	S	NT	Total

EF

HS $\frac{120 \times 100}{300} = 64$ $\frac{120 \times 140}{300} = 56$

$\frac{150 \times 160}{300} = 80$ $\frac{150 \times 40}{300} = 70$

$\frac{160 \times 30}{300} = 16$ $\frac{140 \times 30}{300} = 14$

S	64	56	120
C	80	70	150
GP	16	14	30

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

$S.S = \frac{(50-64)^2}{64} = \frac{196}{64} = 3.06$

$S.NS = \frac{(70-56)^2}{56} = \frac{196}{58} = 3.5$

$C.S = \frac{(90-80)^2}{80} = \frac{100}{80} = 1.25$

$C.NS = \left(\frac{60-70}{70}\right)^2 = \frac{100}{70} = 1.42$

$P.G.S = \frac{(20-10)^2}{16} = \frac{16}{16} = 1$

$P.G.NS = \frac{(10-14)^2}{14} = \frac{16}{14} = 1.14$

$\chi^2 = 3.06 + 3.5 + 1.25 + 1.42 + 1 + 1.14$

$= 11.37$

$df = (3-1) \times (2-1)$

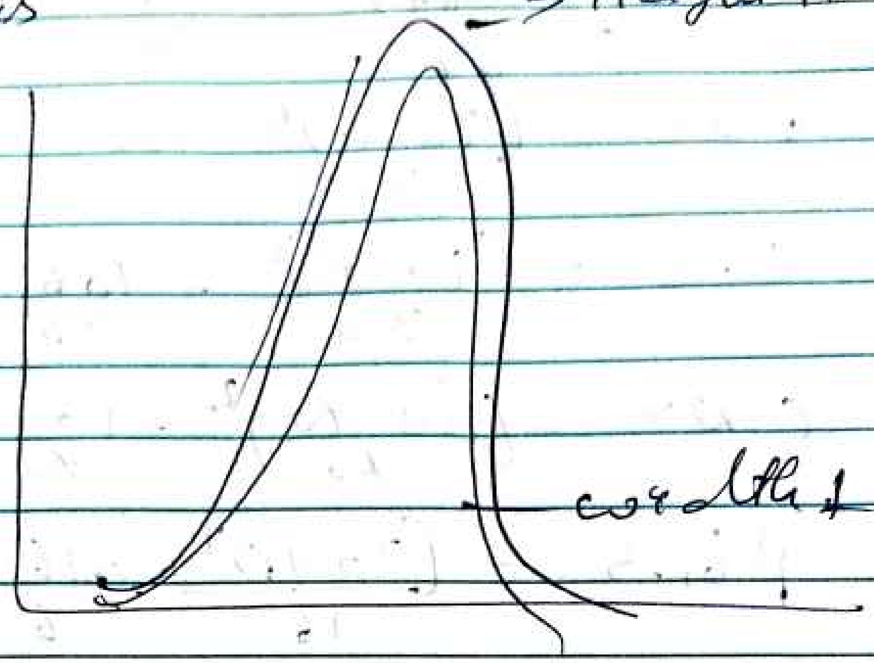
$= 2$

$\alpha = 0.05$

χ^2 critical $= 5.991$

$\chi^2 = 11.37$

Kurtosis



- Kurtosis measures "tailedness" of a distribution or how extreme the outliers are.

1. Mesokurtic
   - Tails are similar to NO
   - distribution with kurtosis ≃ 3
   eg. Standard Normal distribution

2. Leptokurtic
   - A distribution with Kurtosis > 3
   - Heavy tails (with more/extreme Outliers)
   - eg. T-distribution with very small df

3. Platykurtic
   - distribution with Kurtosis < 3

- Light tails (fewer extreme outliers)
- Light tails (fewer extreme outliers)

Eg: uniform dist

Excess Kurtosis $K-3$

K – Kurtosi

$EK > 0 \rightarrow$ Leptokurtic

$EK < 0 \rightarrow$ platykurtic

$$K = \frac{n \cdot \Sigma (x_i - \bar{x})^4}{(\Sigma (x_i - \bar{x})^2)^2} \cdot \frac{1}{n}$$

n = no. of observation

x_i – each data point

⇒ High Kurtosis (Leptokurtic)

- More extreme outliers
- Higher likelihood of rare, extreme values
- ~~Higher likelihood of~~

Eg. financial returns during market crisis/crashes

→ Lower Kurtosis (platykurtic

- fewer extreme outlier
- Data is evenly spread

→ Kurtosis near 3 (mesokurtic)

- Similar to normal distribution