# Unit 5 Stats

A/B testing, hypothesis tests, Null hypothesis & alternative hypothesis, 1-way vs 2-way hypothesis testing, resampling, permutation test, p-values, t-tests.

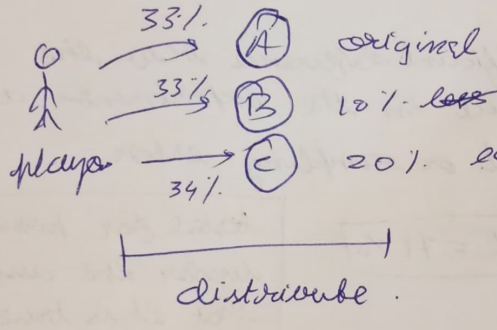
→ Question // assumption.

## A/B Testing:

is testing different versions at the same time & seeing which version gets the best response from players.

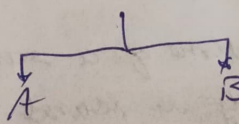
A version is compared with B version.

tests 1 variable
monitors one/two result




① SUTVA (stable unit treatment value assumption)

→ the potential outcome on 1 unit should be unaffected by particular assignment of treatments to the other units.




A purchase independent of B's purchase

② Sampling:

→ represents a population.

distribution → sample distribution is representative of population.

conversion rate

# Hypothesis Testing

→ the claims.

→ It is a test to test some hypothesis about parent population from which sample is drawn.

Class → 100 no. % age 70%.

↳ 20 no. % age 71%.

→ We use hypothesis testing to check the difference b/w this 70% & 71%.

→ is this 1% significant?

$\mu \neq \mu_0$ : two tailed
$\mu > \mu_0$ : right tailed
$\mu < \mu_0$ : left tailed

population has ($\mu_0$)

$\mu = \mu_0 : H_0$.

$H_1 : \begin{cases} H_1 : \mu \neq \mu_0 \\ H_1 : \mu > \mu_0 \\ H_2 : \mu < \mu_0 \end{cases}$

## NULL

There is no significant difference b/w the populations specified in the experiments, due to any experimental or sampling error.

$H_0$ 70% = 71%.

tested for possible rejection under the assumption that it is true.

## ALTERNATIVE

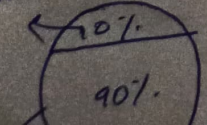
observations are easily influenced by some random cause.

$H_a$ or $H_1$

70% ≠ 71%.

Eg: The mean lifetime of a sample of 81 glucometer produced by a company is found to be 58 months with a SD of 10 months. Company claims that the mean lifetime of the glucometer is 60 months. Test the hypothesis @ 5% level of significance, whether the 58 months lifetime is acceptable or not?

diving by the irrelevance.




this is u Rasha.

$n = 81$

$\lambda_{mean} = 58$ $SD = 10$ ($\sigma$)

5%

company. $\lambda = 60$ False +ve

acceptance 58 or not?

Is the 2 month diff valid or not?

NULL

$58 = 60$.

ALTERNATIVE

$58 \neq 60$.

① Standard Error Mean (SEM).

$\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{81}} = 1.11$

② $Z\text{-score} = \frac{\text{Difference}}{SEM} = \frac{|58-60|}{1.11} = 1.804$

Null & alternative contain opposing viewpoints

NULL → = always .. ≥ ≤

alternative → < or >.

$1-\beta \approx 1$ Success.

## Errors in Sampling.

$1-\beta \Rightarrow$ reject null hypothesis when it is ($H_0$) is false.

Type 1 (

reject Null ($H_0$) when it is $H_0$ indeed True.

$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$

Type 2.

D.N.R ~~accepted~~ Null ($H_0$) when ~~it is~~ $H_1$ is true ~~wrong~~. ($H_0$ is false)

$\beta = P(\text{accept } H_0 \mid H_1 \text{ is true})$

Level of significance: probability of taking risk of accepting/rejecting a null hypothesis.

maximum $\alpha$ (producer) risk.

# Procedure for Hypotheses Testing

① set up hypotheses ↳ H0 (NULL)
↳ H1 (ALTERNATIVE)

② set up significance level. 0.1 | 5 10

③ Test criteria : Z test , < 30 students t test,
X test, F-te

④ computation.

⑤ decisions : TRUE / FALSE

| | Accept H0 | reject H0 |
|---|---|---|
| H0 is true. | correct | Type 1 error (α) |
| H0 is false. | Type 2 error (β) | correct. |

---

## ② Set up significance level.

$$p + q = 1$$

Significance level → p : probability of occurance (wrong decision ha
are)

q :

| Significance level | + | confidence level | = 1 |
|---|---|---|---|

# Procedure for Hypotheses Testing

① set up hypotheses ↳ H0 (NULL)
↳ H1 (ALTERNATIVE)

② set up significance level. 0.1 | 5 | 10

③ Test criteria:
Z test, < 30 students t test,
X test, F test

④ computation.

⑤ decisions: TRUE / FALSE

| | Accept H0 | Reject H0 |
|---|---|---|
| H0 is true. | correct | Type 1 error (α) |
| H0 is false. | Type 2 error (β) | correct. |

---

## ② Set up significance level.

$$p + q = 1$$

Significance level → p: probability of occurance (wrong decision done me)

q:

| Significance level | + | confidence level | = 1 |
|---|---|---|---|

# One-tailed & two tailed:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Eg: $H_0$ : 500 ml ($\mu$)

$H_1$ : $\mu \neq 500$ ml [2 tailed test]

or $H_1$ : $\mu < 500$ ml [Left test] } One tailed.

or $H_1$ : $\mu > 500$ ml [right test] }

when to use which $H_1$ ??

$H_1$

$\neq$

Eg: company claims that on average contains no more than 2 g.

Test whether there is a strong evidence that is greater than the claim.

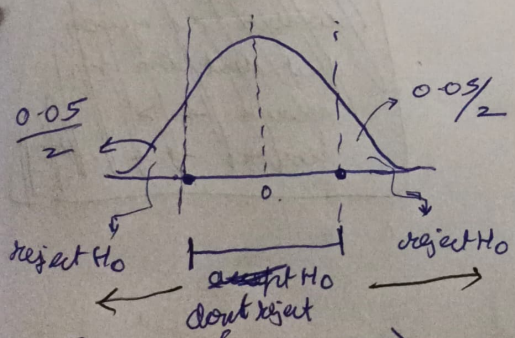
$H_0$ : $\mu = 2$

$H_1$ : $\mu > 2$

| $H_1$ | |
|---|---|
| greater | $\mu > 2$ |
| lesser | $\mu < 2$ |
| differs | $\mu = 2$ |

Consider $\alpha = 0.05$.

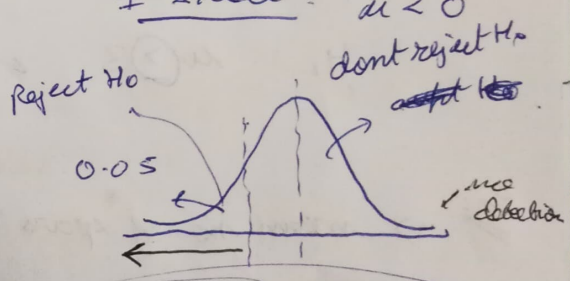
2 tailed test. ($\mu \overset{H_1}{\neq} 0$)




Z score ($\alpha = 0.025$)

$= -1.96$ & $1.96$.

1 sided: $\mu < 0$




$Z(\alpha = 0.05) = -1.645$.

1- sided:

① area of detection of other side.

② directional power is gained.

decided on alternative testing $H_0$ is either ~~accepted~~ or rejected

F.T.R

Eg: Want to test if clg students take less than 5 years to graduate from clg on average. Then,

$H_0: \mu \geq 5$ (position starting from)

$H_1: \mu < 5$ (where aiming?)

Eg: Con test, 40% pass. We want to test if more than 40% pass that day.

$H_0: \mu = 40$ (0.4)

$H_1: \mu > 40$ (0.4)

If the $H_0$ is not given, & suppose you are testing is next connection more than 3 mbps, then obviously $\mu$ is testing if you think it is less than or equal to 3.

This becomes Non hypothesis then.

Eg: Testing mean speed is more than 3 mbs.

$H_0: \mu \leq 3$

$H_1: \mu > 3$

TRICK:

whatever is given just what is in $H_1$

more $\rightarrow >$

less $\rightarrow <$

Eg: 1) mean no. of years is 34

$H_0: \mu = 34$

$H_1: \mu \neq 34$

2) at most 60% of americans vote.

$H_0:$

$H_1: \mu$

NULL:

assume $H_0$ is true (claim)

↓ reject $H_0$ — Fail to reject $H_0$

IF you want to support a claim YOU MUST state it as $H_1$ (not $H_0$)

want to prove it? $H_1$

NOTE:

① The original claim could either be $H_0$ or $H_1$

↳ depends where EQUALITY IS

Eg: (mean) of fluid is AT LEAST 12 oz in a can.

CLAIM: $\mu \geq 12$ $H_0$

OPP: $\mu < 12$ $H_1$.

Which is $H_1$ or $H_0$?

Wherever there is equals ... put it there.

---

Eg 1: mean no of years americans work is 34

$\mu = 34$ $H_0$

$\mu \neq 34$ $H_1$

Eg 2: at most 60% vote in elections.

$p \leq .6$ $H_0$

$p > .6$ $H_1$

Eg 3: mean starting salary is at least $100000 per year

$\mu \geq 100000$ $H_0$

$\mu < 100000$ $H_1$

Eg 5 21% get drunk

$p = 0.29$ $H_0$

$p \neq 0.29$ $H_1$

$\bar{x} = 0.49$

$x \rightarrow 0.44$

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

Eg 5 fewer than 5% of adults ride the bus.

$p < .05$ $H_1$

$p \geq 0.05$ $H_0$

Eg 6 mean no. is not more than 10

~~$\mu = 10$~~ $\mu \leq 10$ $H_0$

$\mu > 10$ $H_1$

Eg 7. about half prefer to stay outside.

~~$p = 0.5$~~ $H_0$

$p \neq 0.5$ $H_1$

Eg 8. chance of developing breast cancer is under 11% of w

$p < 0.11$ $H_1$

$p \geq 0.11$ $H_0$

$P(X = 10 \cap X = 8 | X =$

Eg 9. mean cost is more than 2k

$\mu > 2k$ $H_1$

$\mu \leq 2k$ $H_0$

Eg: $n = 273$ $x = 6$

$n = 63$ thin. $x = 4$

Is there a good evidence to suggest more than 30% ($p$) teen girls stay thin by smoke? $H_1 = ?$

$H_0$:

$p > 0.3$ $H_1$

$p \leq 0.3$ $H_0$

(4)

$p = \frac{63}{273} \times 23 =$ 23

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

s : population S.D

STEP 1. [Claim
Opp.

STOP 2 [ whoever = put as $H_0$ (NULL)

Eg: sample of 706 found (61%) were MALE. Claim: most CEO are male

CLAIM : $p > 0.5$ ⇒ $H_1$ ~~$p > 0.5$~~

OPP : $p \leq 0.5$ ⇒ $H_0$ $p = 0.5$

$\hat{p} = 0.61$ $p = 0.5$ $q = 1 - p$

$Z = 5.84$ ← UNUSUAL.

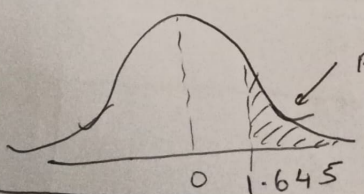
$H_0$ is rejected.

so $H_1$ is accepted.

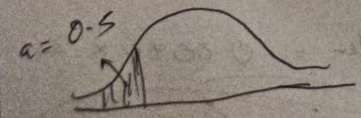
HOW TO MAKE DECISION

significance level:

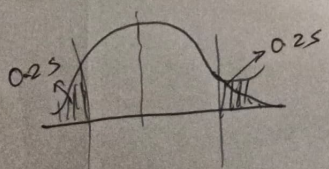
$\alpha = 0.05$.



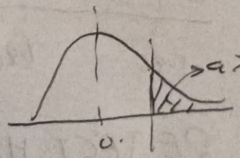

$H_1$: $p < 0.5$ $H_1$: $p = 0.5$ $H_1$: $p > 0.5$




Left Tail.




Two tail




Right tail.

# P-Value


REJECT $H_0$ IF P-value $\leq \alpha$

F.T.R $H_0$ IF P-value $> \alpha$

$p < 0.2$ $H_1$

$p \geq 0.2$ $H_0$

$\frac{11}{84}$ $\frac{20}{100} \times 84 = 16.8$

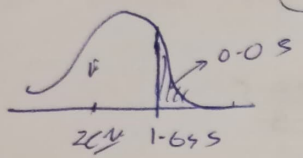
Eg: Find P-Values

① $\alpha = 0.05$ $H_1: p > 0.25$

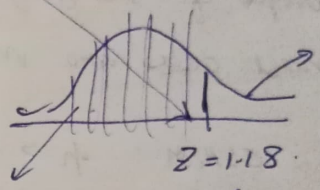
$Z = 1.18$

F.T.R

TRAP




P-Value




0.8810

Test stat.

F.T.R

$H_0$ ~~Rejected~~

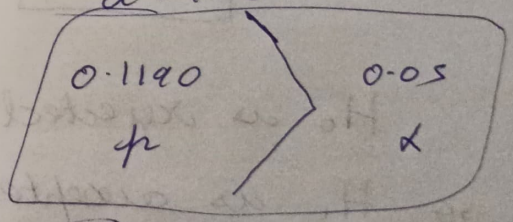
not enough evidence to support the claim.

~~$a = 0.8810$ ($p$)~~

OOPS!!

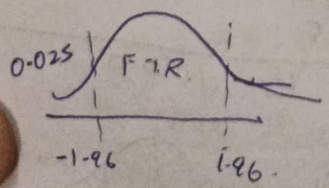
table gives left area to Z.

$a = 1 - 0.8810 = 0.1190$



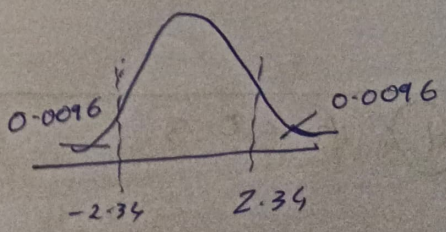

② $\alpha = 0.05$ $H_1: p \neq .25$ $Z = 2.34$

Trad.




REJECT $H_0$.

P-value

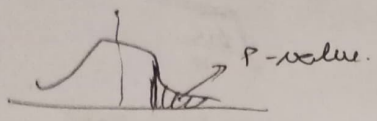



P-value = $0.0096 \times 2$

$0.0192 < 0.05$

reject $H_0$.

# P-Value




REJECT $H_0$ IF P-value $\leq \alpha$

F.T.R $H_0$ IF P-value $> \alpha$

Eg: Find P-Values

$p < 0.2$ $H_1$

$p \geq 0.2$ $H_0$

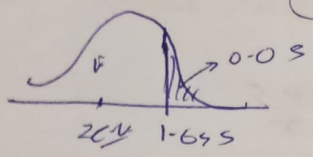
$\frac{11}{84}$ $\frac{20}{100} \times \frac{84}{100} = 1.68$

① $\alpha = 0.05$ $H_1 : p > 0.25$

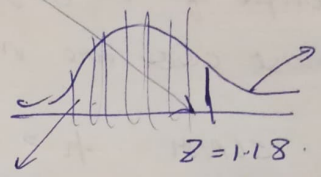
$Z = 1.18$

F.T.R

TRAD




P-value




0.8810

Test stat.

F.T.R

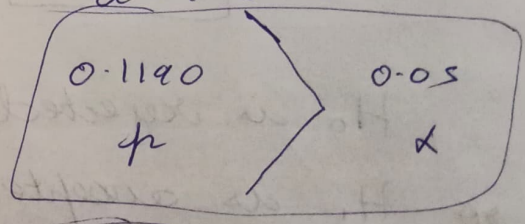
$H_0$ ~~Rejected~~

not enough evidence to support the claim.

~~$\alpha = 0.8810$ ($p$-value)~~

OOPS!! table gives left area to Z.

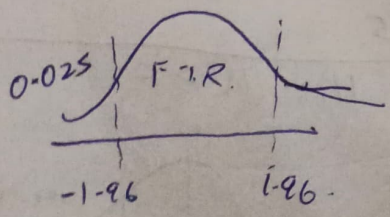
$a = 1 - 0.8810 = 0.1190$




② $\alpha = 0.05$ $H_1 : p \neq .25$

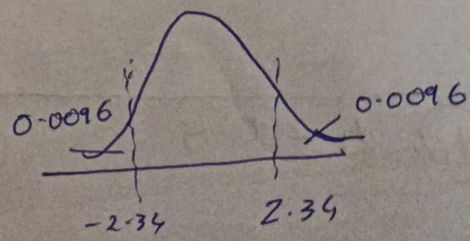
$Z = 2.34$

Trad.




REJECT $H_0$.

p value.




P-value = $0.0096 \times 2$

$0.0192 < 0.05$

reject $H_0$.

In a sample of 300, ... that most CEO are male use significance level as 0.05

claim $p > 0.05$ $\Rightarrow$ $H_1$ $\quad \alpha = 0.05$.

opp $p \leq 0.05$ $H_0$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = 3.81$$

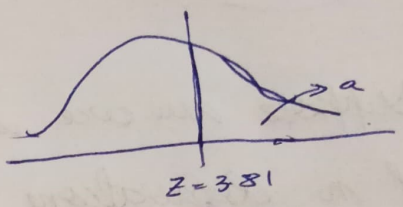
$\hat{p}$ = sample $p$ $= \frac{183}{300} = 0.61$

$p > 0.05$

$p = 0.05$

$q = 0.05$

$n = 300$ (sample size)




$p$-value $= 0.0001$

accept $H_1$

support claim.

$p < \alpha$ reject

reject $H_0$.

Eg: previously organization 4.5 on average. currently thinks the mean is higher. random. is chose

mean $= 4.75$, $\sigma_{sample} = 2$ $\quad \alpha = 0.05$.

$\mu > 4.5$ $\quad H_1$

$\mu \leq 4.5$ $\quad H_0$

$$Z = \frac{4.75 - 4.5}{\frac{2}{\sqrt{15}}}$$

F.T.R

value.

About Pop Mean ($\sigma$ known)

1. random sample.
2. $\sigma$ is given.
3. $n > 30$ OR pop is normally distributed. → chap 6

# T-tests

for population mean $\mu$.

→ when sampling from normally distributed population

$\mu = \mu_0$ ← (Hypo.)

↗ population mean

Suppose we are drawing a simple random sample of n observations from normally dist. population.

To test $H_0 : \mu = \mu_0$ — either z-test or t-test.

**$\sigma_{pop}$ is known**

$$z = \frac{\bar{x} - \mu}{\frac{\sigma_{pop}}{\sqrt{n}}}$$

**$\sigma_{pop}$ is NOT known**

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

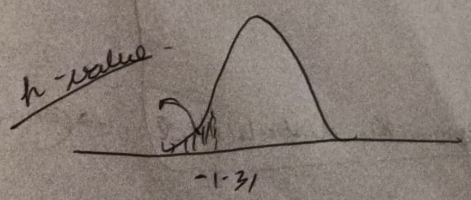
$s$ : sample

$s/\sqrt{n}$ → **Standard error**

estimated standard dev of sampling distribution of sample mean.

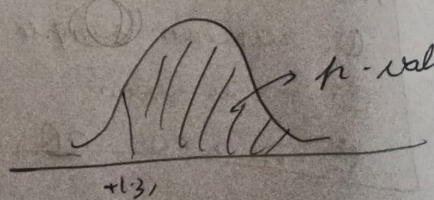
$H_0 : \mu = \mu_0$

$\mu < \mu_0$




$t = -1.31$

$\mu > \mu_0$




$t = -1.31$

Eg: 25 delegates women. over Rxn = 1 s

$r = 0.18$

speed stn ... mean rxn is differe fro 0.9 s sec.

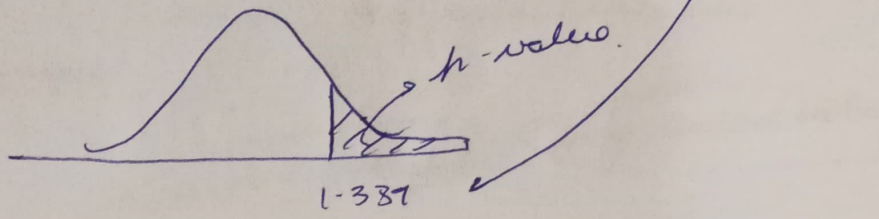
$\mu = 0.95$

$\mu \neq 0.95$

$\alpha = 0.05$

$$t = \frac{1 - 0.95}{0.13/\sqrt{25}} = 1.389$$

25-1 DOF




Eg: $\mu = 3.5$

$\mu \neq 3.5$

both side → 0.05

$\bar{x} = 3.3$ $s = 0.851$

DF = 10 - 1 = 9

$$t = \frac{3.3 - 3.5}{0.851/\sqrt{10}} = -0.743$$

2.262 → critical t

(−2.262 2.262)

If it falls in that region

we fail to reject $H_0$.

## Module 1

EDA, structured data ... rectangular data
DF, indexes, NDs, Estimation of location.
Mean, Med + weighted trimmed robust estimates.
variability (deviations, variance, SD, range, quartiles) percentile.

**Data:**
- **structured**: pre-defined items / specific concepts
  rows, obs. units / variables.
- **unstructured**:

**Descriptive Statistics:** Organizing & summarizing data
(sample & population) (freq, central tendency)

**Inferential Statistics:** formal methods for drawing some good conclusions from data
(based on samples)
→ uses probability to determine how confident we can be that conclusions are correct.
(regression, hypothesis)

**Mathematical Models:** → describe phenomena that occur in real world.

① **deterministic:**
↳ when value is precisely determined from another value.
↳ v, u, a
↳ high degree precision

**Statistical Models:** → used to predict life's more uncertain situations
↳ based on idea that one value affects another value.
↳ weather prediction (probability)

Probability: → study randomness, mathematical tool.
→ chance of occurence of event.

Population → collection.

↓

SAMPLING

→ gain info about population

Categorical / qualitative
numerical / quantitative.

%
Pie Chart → 100% matching data pts & total.

Bar graph → ~~DO NOT~~ consider when data points are missing.

↳ exact no.
↳ size of category

NO MISSING DATA.
→ sorted.

## Sample:

↳ shud have same characteristics as the population

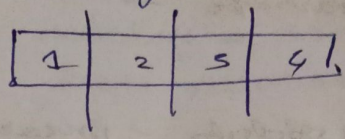
### Random Sampling

1. Simple
2. Stratified
3. cluster
4. systematic.
(pattern selection).

① Simple Random

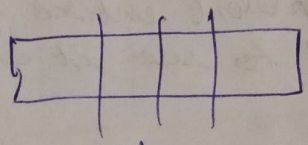
↳ each has same chance of being selected
↳ random no. generated - 9, 3, 6 ②

② Stratified Sampling




now perform random sampling for each category.
(equal dist.)
may / may not.

③ Cluster



clusters

↓

→ select random clusters as a whole.
→ may NOT contain equal no. of randomly chosen students from each class.

Non Random

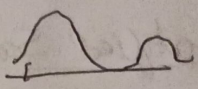
① Convenience Sampling
→ using results already available.
→ sending process.

Sampling Errors
↳ sample size
↳ defective device } non sampling errors.

Sampling bias: → a sample is collected from population where ~~each~~ some members of population do not have an equally likely chance of being chosen.

## Problems

1. sample - Biased
2. self selected samples (internet etc.)
3. size sample
4. influential
5. removal of participants →
6. self control
7. Misleading use of data - incorrect graphs etc.



## NOTE:

1. sampling types may cause variations.
2. If it doesn't represent entire population it is not useful to use it as sample

@ music store → convenient sample

200 → 175 ✓

→ 46 ✗

NO NOT GD

## Sampling Variability

→ samples taken by different parameters.

→ large sample size → BETTER!!

**Confidence intervals.**

## Scales

↳ Nominal: Y/N questions No 1/2

↳ ordinal: top 5; ordered hierarchical

↳ interval:

↳ Ratio:

CRV → a RV whose outcomes are measured.

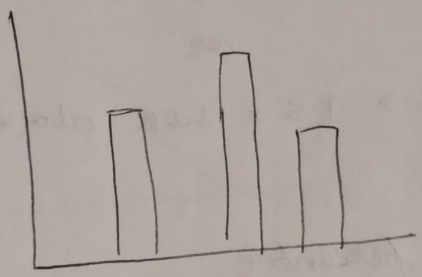
DISCRETE : result of counting

CONTINUOUS : result of measuring

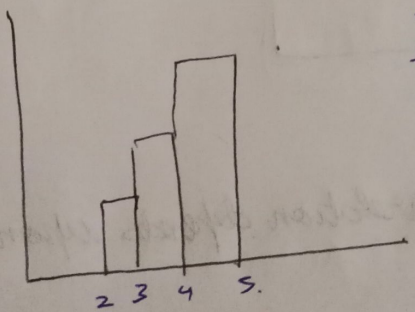
Sampling error – natural variation that results from selecting a sample to represent a larger population.

## Bar-graphs

→ categorical data
→ discrete
→ frequencies on y



## Histogram




→ continuous
→ intervals
→ tells the distribution actually
→ values within interval

## Measures of Location of data:

① quartiles
② percentile

1st Q 25th percentile

Median 2: 50th percentile.

$Q_3$ 75th percentile.

$90^{th}$ percentile ⇒ 10% of the test scores are the same as or greater than your test score.

IQR → $Q_3 - Q_1$

Outlier is determined by IQR.

$x \leq 1.5 \times IQR$ below $Q_1$

or

$x > 1.5 \times IQR$ above $Q_3$

Finding $k^{th}$ percentile.

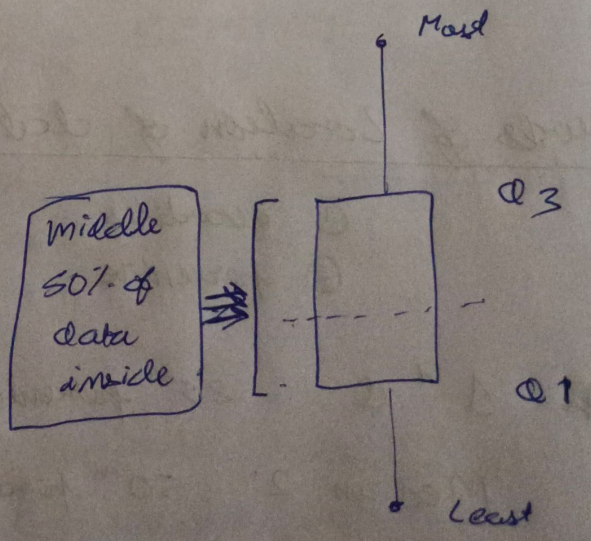
$$i = \frac{k}{100}(n+1)$$

n: no. of data points

Percentile interpretation depends upon usage.....

Box Plots: (5)

→ median
→ $Q_1$
→ $Q_3$
→ min
→ max.



# Measures of centre of the data

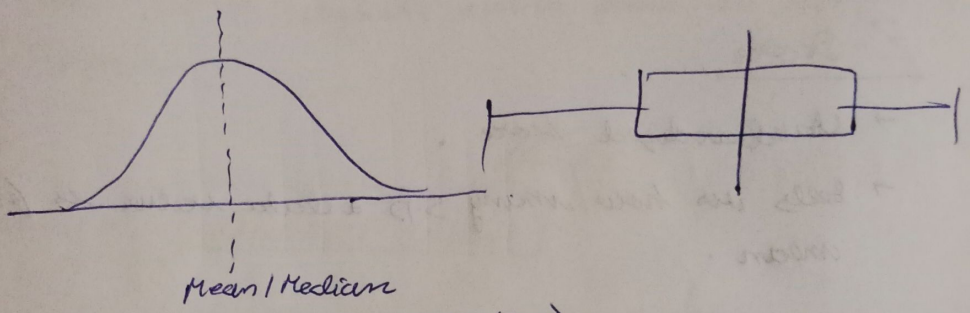
mean, median

median → better when there are outliers

sample mean: $\bar{x}$

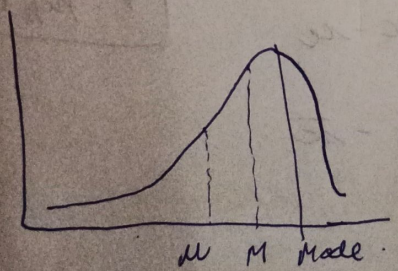
population mean: $\mu$

## Skewness



Mean / Median
(symmetric distribution)

## LEFT

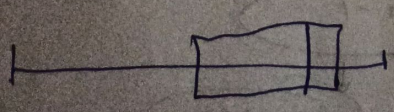



left to bend pull.

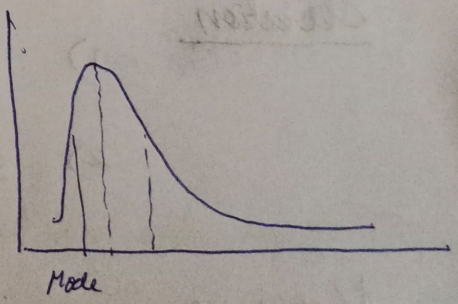
→ more higher values.

→ Mean < Median.

$\mu$ < Median < Mode



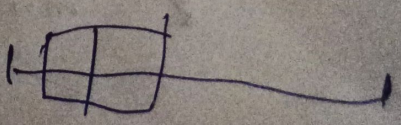
## RIGHT




→ more lower values

→ Mean ≥ Median

Mean > Median > Mode



# Measures of Spread of Data

→ variation in data

## STANDARD deviation

① measure how far data values are from their mean.

② determine if data pt is close / far from mean.

→ It is ALWAYS +ve or zero.

## Z-Score:

→ standardized score.

→ tells us how many S.D a data value is from the mean.

$$Z = \frac{x - \mu_x}{\sigma_x}$$

x : particular score
$\mu$ : population mean
$\sigma$ : population S.D

S : sample S.D
$\sigma$ : pop SD

## Deviation:

1) Population : $x - \mu$

2) Sample : $x - \bar{x}$

## Variance:

: average of squares of deviations.

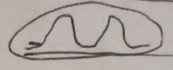
$$s = \sqrt{var}$$

$$var = \frac{\Sigma (x - \bar{x})^2}{n-1} \approx \frac{\Sigma (x - \bar{x})^2}{n-1}$$

sample

For population → $n$

Bimodal



$$S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N-1}}$$ → Sample ka S.D

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$ → population ka S.D

## Sampling Variability of a statistic

∴ how much statistic varies from on sample to another.

measured by **STANDARD ERROR**

is the average S.D resulting from repeated sampling

Standard error of mean : $\frac{\sigma}{\sqrt{n}}$

$\sigma$ : SD population

$n$ : size of sample

NOTE:

① mean affected the most by skewing.

Mean absolute dev = $\frac{\Sigma|x_i - \bar{x}|}{N}$

M.A.D = Median $(|x_1 - m|, |x_2 - m| \cdots)$

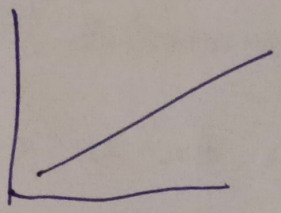
# Probability

$$-1 \leq r \leq 1$$

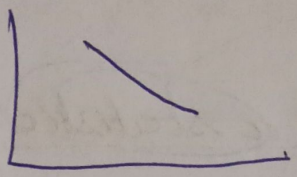
$r = 0$ No corr.

## Expected value:

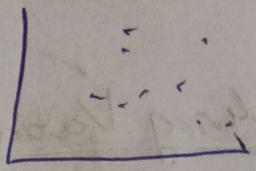
$$E[x] = \sum_{x=1}^{n} h(x) \cdot x$$



+ve



-ve.

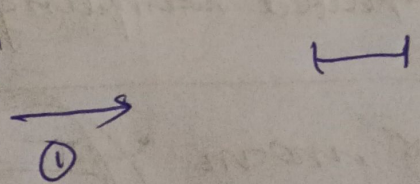


ONLY LINEAR

## Hexagonal Binning:

explores 2 or more categorical variable. as well.

Violin Plot



**Probability**: certain we are of results....

**experiment**: operation carried under controlled conditions..

↓ result.

**Outcome**

**sample Space**: set of possible outcomes....

**Conditional**

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

## Independent & M.E.

→ **INDEPENDENT**

Two events are independent if knowledge of occuring of one doesn't affect chance of other.

$$P(A \cap B) = P(A) \cdot P(B)$$
$$P(A/B) = P(A)$$
$$P(B/A) = P(B)$$

→ **Mutually exclusive**

either A or B.

$$P(A \cap B) = 0$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

t-dist

N > 30 normal
N < 30 t-dist.

A = first attempt success. 0.65

B = second attempt 0.65

P(B/A) = 0.90.

length

A ∩ B = ?

0.90 × 0.65.

↓

both goals.

either P A ∪ B

any door is 1/3

Eg: 3 doors.

| | caught | not caught. | |
|---|---|---|---|
| 1 out → | 1/5 | 4/5 | 5/15 |
| 2 out → | 1/4 | 3/4 | 3/12 |
| 3 out → | 1/2 | 1/2 | 1/6 |
| | 19/60 | 41/60 | 1 |

$\frac{\frac{1}{15} + \frac{1}{12}}{19/60} = 9/60$

Eg: P(biker) = 0.2 no bike = 0.8

P(A/B) = P(A ∩ B) / P(B)

P(worker / biker) = 3/4

P(worker / no bike) = 1/4

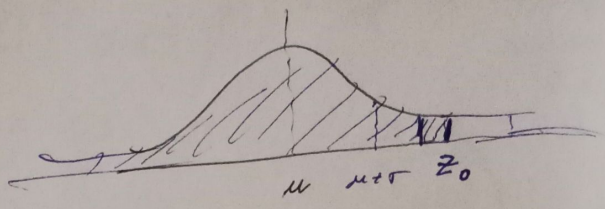
P(biker / worker) = ?

P(worker ∩ biker) = 3/4 · 0.2 = P(biker / worker) · P(worker)

P(worker ∩ nobike) = 1/4 · 0.8 = P(nobike / worker) · P(worker)

# Normal distribution

$X \sim N(\mu, \sigma^2)$




→ depends only on mean & ~~median~~ S.D

→ area = 1

## Standard Normal distribution

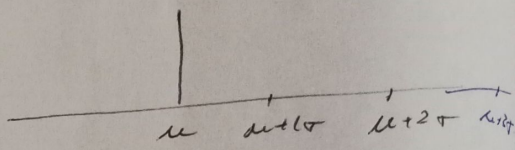
$\mu = 0$

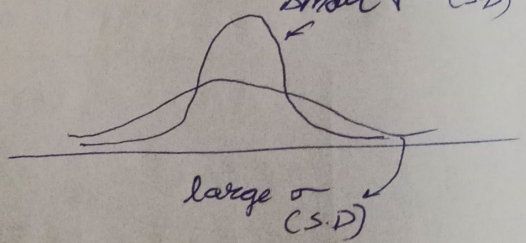
S.D = 1

$N(0, 1)$

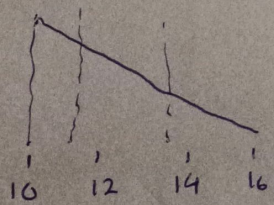
Z-score: indicates no. of S.D a score falls above/ below the mean.

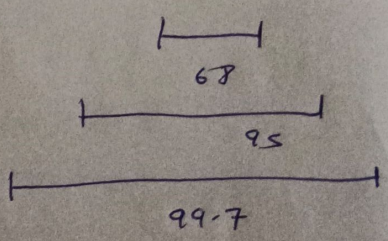
$$Z = \frac{x - \mu}{\sigma}$$




$P(Z \leq z_0)$ = area until $z_0$



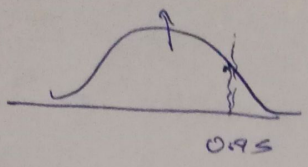




$$Z_{11} = \frac{X - \mu}{\sigma} = \frac{11 - 10}{2} = 0.5$$

$$Z_{13.6} = \frac{3.6}{2} = 1.8$$

$$P = a_{z=1.8} - a_{z=0.5}$$

Eg: Top 5%. $N(500, 100^2)$ $X = ?$




area = 0.95

$\Rightarrow z = 1.645$

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

$1.645 \times 100 + 500 = X$

664

## Sampling Distribution

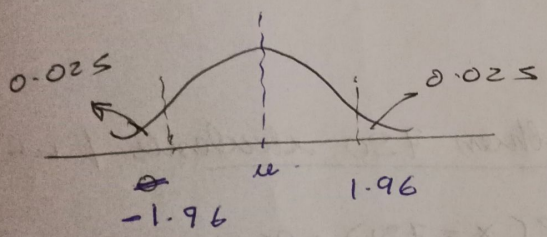
Standard dev of mean

$$\sigma_{\bar{x}} = \frac{\sigma_{Pop.}}{\sqrt{n}}$$

$\sigma$ : population

$n$ : sample size.

If z's levels within what lies sample average P = 0.95




$X = \mu - 1.96$ (S.D mean)

$\mu + 1.96$ (S.D mean)

In Normal distribution Mean = Median

In continuous distribution $P(x < 1) \approx P(x \leq 1)$

# Very Very IMP

Eg: 6000 weekly ...
10% defective ...
random sample $n = 100$
$x$ : no. of defective in a
$x = ?$

discrete
↓
Binomial & Normal
↓
continuous

$n = 100$
$p = 0.1$
$q = 0.9$

$\mu = np$
$\sigma = \sqrt{npq}$

Eg: 75% → favour.
find probability more than 780 students / 1000 will be in favour. $P(x = 781) + P(x = 782 \cdots)$

$p = 0.75$
$q = 0.25$
$n = 1000$.

$\mu = np = 750.$
$\sigma = \sqrt{npq} = 13.7.$

continuous.

$N(750, 13.7^2)$

$P(780.5 < x < 1000.5)$

$\approx P\left(\frac{780.5 - 750}{13.7} \leq z \leq \frac{1000.5 - 750}{13.7}\right)$ → z-score

Eg:

defective 10%.

$n = 200$

P of defective 24 & 30

$p = .10$

$q = .9$

$n = 200$

$\mu = np = 20$

$\sigma = \sqrt{npq} = 4.24$

$P(25 \leq X \leq 29)$

$\approx P(24.5 \leq X \leq 29.5$

$P\left(\frac{|20 - 24.5|}{4.25} \leq Z \leq \frac{|29.5 - 24.5|}{4.24}\right)$

# Central Limit Theorem

↓ Draw a finite sample size n from population

↓ $\mu$, $\sigma$ is know

↓ Normal. $n \geq 30$.

$\mu_x$ : mean of population

$\sigma_x$ : S.D of population then

$$X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

$$Z = \frac{x - \bar{x}}{\frac{\sigma_x}{\sqrt{n}}}$$

sample ka

sample means form their ow

$n \geq 30$

Eg: $\mu_{pop} = 90$ $\quad n = 25$

$\sigma_{pop} = 15$.

a)

for Normal distribution

$$Z = \frac{x - \bar{x}}{\sigma}$$

for sample di

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

Eg: $\mu = 74$, $\sigma = 6.8$. } normal distribution.

a) selected @ random $x < 65$.

$$P\left(Z < \frac{65 - 74}{6.8}\right) \leftarrow \text{area}$$

d) If sample of 50 selected @ random $P(\bar{X} > 75)$

mean exam score.

$n = 50$

$$P\left(Z > \frac{75 - 74}{\frac{6.8}{\sqrt{50}}}\right)$$

MEAN → CLT.

# CIE

Eg: Deviation data ... length of deviation vector ....

$x =$ [ . . . . . ] some 10 pts.

① find mean

② $x = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$ ; $\bar{x} =$ Mean $\begin{bmatrix} \\ \\ \\ \\ \\ \\ \\ \end{bmatrix}$.

③ Deviation vector $d = x - \bar{x} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$.

$|d| = \sqrt{a^2 \ldots} = \sqrt{54}$

Histogram → tells the distribution .... can be plotted in boxplot.

→ continuous data. (median, quart ....)

Bargraph → categorical

Eg: Disjoint are M.E.

# Hypothesis Testing Qs.

Eg:

$\mu = 1.2$ (not) — 24.5

$\mu \neq 1.2$

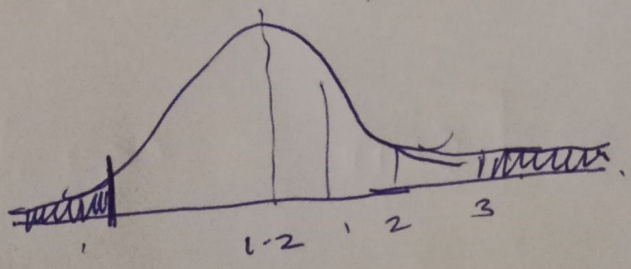
The mean of 100 injected in 1.05 sec w S.D = 0.5

Remember.

$$\bar{x} = \frac{s}{\sqrt{n}} = 0.05.$$

$$z = \frac{1.2 - 1.05}{0.05}$$

$z = 3$



# Random Variable

DISCRETE : counting.

continuous : ht., temp. measuring.

## Probability Distribution Function (PDF)

Discrete.

① $0 \leq P(x) \leq 1$

② $\Sigma P(x) = 1$

} conditions.

## Expected Value: $E(x)$ Mean.

$$E(X) = \Sigma x P(x)$$

long term average

## Binomial distribution.

probability distribution used to model behaviour of a random variable

3 characteristics.

① fixed no. of trials. ($n$)

② ONLY two possible outcomes

$p$ : success

$q$ : failure

$p + q = 1$

↳ should be consistent for all trials.

③ the trials are independent

$${}^nC_k \, p^k (1-x)^{n-k}$$

If $n = 1$

↳ Bernoulli trial + ② + ③

Mean, $\mu = np$

variance $\sigma^2 = npq$

S.D $\sigma = \sqrt{npq}$.

$$P(X = x) = p^x (1-p)^{1-x}$$

$x = 0, 1$

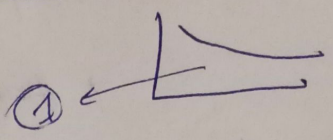
# Geometric Distribution "first"

$1 - \frac{7C_2}{10}$ $\frac{7 \times 6}{2 \times 10}$ 2.1

① repeating independent Bernoulli trials until success is obtained.

② atleast 1 trial

③ $p, q \ldots$




$$p(x) = (1-p)^{x-1} p^1$$

$$\text{Var} = E(x^2) - [E(x)]^2$$

↳ $p$: suess.

# Poisson

① no. of occurence of an event in a given unit of time, distance / area / volume.

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

$\lambda$: mean time.

$$P(x \leq 5) = P(x=1) \ldots P(x=5)$$
$$= q^0 p^1 + q^1 p^1 + q \ldots$$

Eg: 2% defect → 100 bytes

a) $P(X < 7)$ within first 6 bytes → defective piece.

b) $P(X > 12)$

c) $P(X \geq 18)$

d) $P(15 \leq x \leq 30)$

e) $P(20 < X \leq 45)$

a) $P(X < 7) \rightarrow$ ~~1~~ to 6

$$P(X \leq x) = 1 - q^x$$

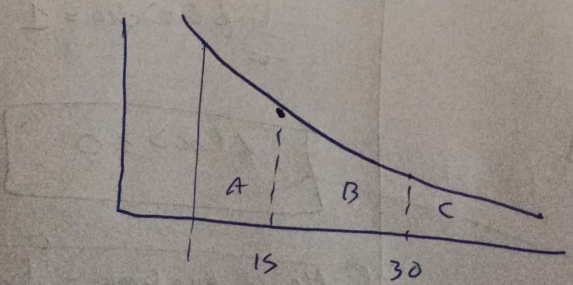
$P = 1 - (0.98)^6$

b) $X \leq 12$ $\quad 1 - (0.98)^{12}$ $\quad$ BUT $x > 12$

$\Rightarrow 1 - (1 - 0.98)^{12}$

c) $X \leq 17$ $\quad 1 - (0.98)^{17}$

$(0.98)^{17} =$




$P(x \leq 14) + P(15 \leq x \leq 30) + P(x \geq 31) = 1$

$P(X \leq 1) + P(x \geq 1) = 1$

$\mu = 22$ $\sigma_n = 2.3$ $n = 160$

Eg: 12 customers / day

a) what is probability exactly 8 customers in 1 day

→ freq given

Poisson!!

$\mu = 12$

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

$$P(X = 8) = \frac{12^8 e^{-12}}{8!}$$

Eg:

average → 7 text in 2 hrs.

a) exactly 9 text in 2 hrs.

$\mu = 7$ $\quad x = 9$

$$P(X = 9) = \frac{7^9 e^{-7}}{9!}$$

b) exactly 24 text in 8 hr

$$\mu = 7 \times \frac{8}{2} = 28$$

$$P(X = 24) = \frac{28^{24} e^{-28}}{24!}$$

## PMF

$\sum P(x) = 1$

$P(x) \geq 0$

ALWAYS

## PDF

$\int_{-\infty}^{\infty} f(x)\,dx = 1$

$f(x) \geq 0$

Mean life time $= \int_{-\infty}^{\infty} x f(x)\,dx$ $(E[X])$

## Valid PDF.

$$\int_{6}^{8} f(x) = 1$$

| | mean | S.D |
|---|---|---|
| +3 | +3 | – |
| ×2 | ×2 | ×2 |
| x-1 | x-1 | – |

Bayes

Eg: given day  parepting = 3

a) $P(X=5) = \frac{3^5 e^{-3}}{5!}$

b) $P(X=5 \mid X \geq 4) = \frac{P(X=5)}{P(X=4)+P(X=5)} \ldots\ldots$

$= P(X=5 \mid X=4) + P(X=5 \mid$

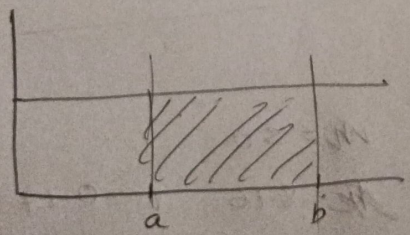
Covariance

# CRV

↳ uniform distribution

↳ exponential distribu

# Uniform distribute




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

$P(x = \frac{a+b}{2}) = 0$.

Mean $= \frac{a+b}{2}$

var $= \frac{(b-a)^2}{12}$

You flip a coin .. $p(H) = 50\%$ is known but think it is less .. Type of test?

$n = 50$ $\bar{x} = 0.46$ $\alpha = 0.01$

$p = 0.5$ Null

$p < 0.5$ Alternative

single tailed.

verify - given

2.58

$$Z = \frac{\bar{x} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.46 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{50}}}$$

IMP.

$$Var(X - 2Y) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X, Y)$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

ab

$H_0 : \mu = 8.21$

$H_1 : \mu < 8.21$

$n = 50$

$\bar{x} = 8.16$ $\sigma = 0.17$

$\alpha = 0.01$

$$Z = \frac{8.16 - 8.21}{\frac{0.17}{\sqrt{50}}} = -2.08$$

$Z_{-0.01} = -2.33$

FTR $H_0$.