Random Variable

↳ A set of possible values from a random experiment.

Ex.,

Tossing a coin → Experiment is random, outcomes will be random.

↓

H, T

↳ → Quantify these random values

$X = \{ H, T \}$, $X = \{ 0 \rightarrow \text{Tail } (T), 1 \rightarrow \text{Head } (H) \}$

↓

Random variable (can take any value from the set of values)

Ex.,

Throwing a dice

$X = \{1, 2, 3, 4, 5, 6\}$ → Sample space

$P(X=1) = \frac{\text{No. of Possible outcomes}}{\text{Total No. of outcomes}}$

$= 1/6$

Different Types of Distribution

Probability distribution function

Prob. Density Function
- → Normal dist. / Bell shaped dist. / gaussian dist.
- → Std. Normal dist.
- → log Normal dist.
- → chi-square dist.
- → F-dist.

Prob. mass function
- → Bernoulli dist.
- → Binomial dist.
- → Poisson dist.

Inferential stats is based on these

Uniform dist.
- Discrete
- continuous

Uniform dist. is based on both functions.

PROBABILITY DISTRIBUTION

Random Variable :- A set of possible values from a random experiment.

The value of Random Variable is unknown.

A function that assigns values to each of experiment outcomes.

Ex., Tossing a coin. - {H, T}

$X = \{0, 1\}$

$P(x = H) = 1/2$

$P(x = T) = 1/2$

A function $= 1/n$ (where, n is the total no. of outcomes)

Outcomes of Experiment

- Tossing a coin
- Throwing a dice

(discrete outcomes)

⇓

Prob. mass func.

Calc. the prob. if a student's height is below 170 cm

(continuous outcomes)

⇓

Prob. density func.

① Prob. Mass Function (PMF)

→ Distribution of discrete random variable.

Ex.: Rolling a dice

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

$P(X \leq 3) \Rightarrow P(X=1) + P(X=2) + P(X=3)$

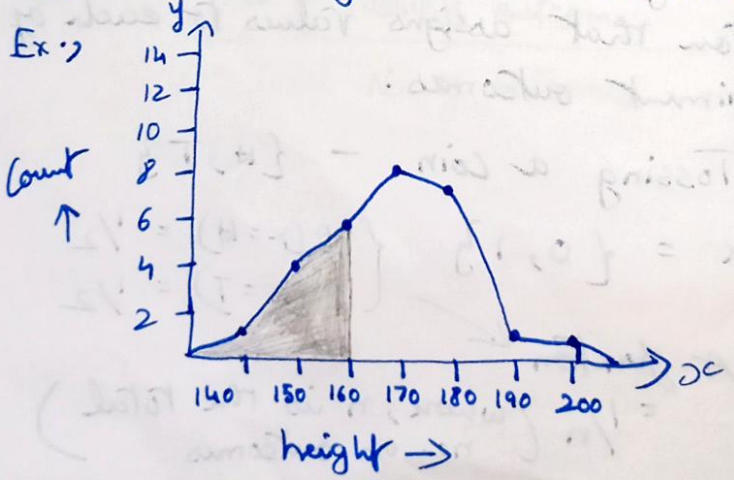
$= 1/6 + 1/6 + 1/6 = 3/6 = 1/2$

$0 \leq$ Probability value ≤ 1

② Prob. density function (PDF)

→ Distribution of continuous data.

Ex.:



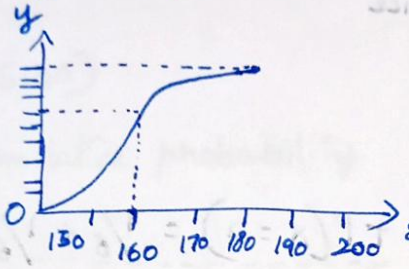
$P(X \leq 160)$ = Area under the curve in the shaded region.

$0 \leq$ Value of Area under curve in shaded region ≤ 1

Cummulative Distribution Function (CDF)

→ Summation of all probabilities upto a given point.

Ex.,



→ for Prob. density function

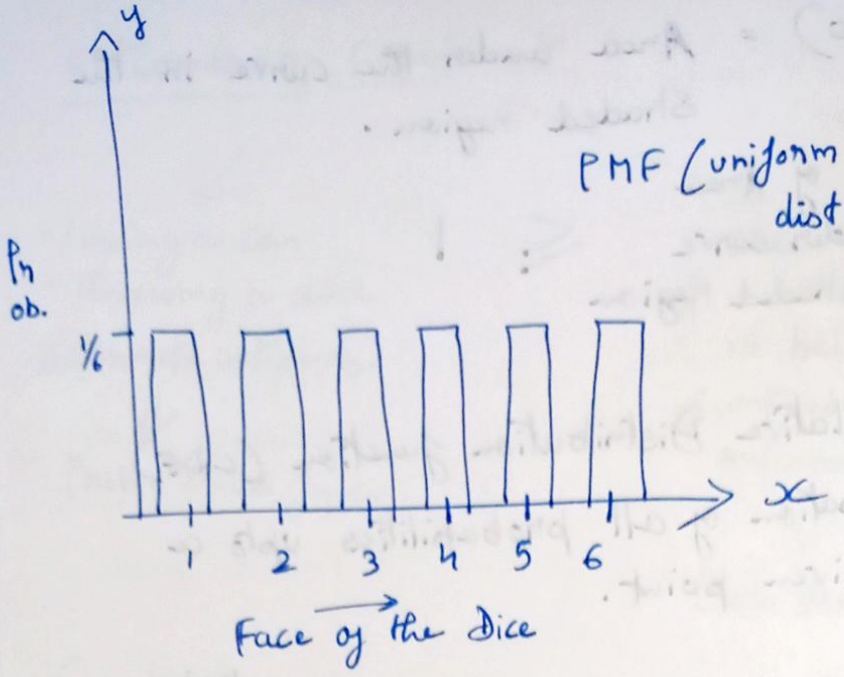
① Prob. Mass Func. (PMF)

→ Discrete Random variable

→ Rolling a dice {1, 2, 3, 4, 5, 6}

$P(1) = 1/6$
$P(2) = 1/6$
$P(3) = 1/6$
$P(4) = 1/6$
$P(5) = 1/6$
$P(6) = 1/6$

} → uniform distribution



$P(X \leq 1) = 1/6$

$P(X \leq 2) = P(X=1) + P(X=2) = 1/6 + 1/6$

$= 2/6 = 1/3$

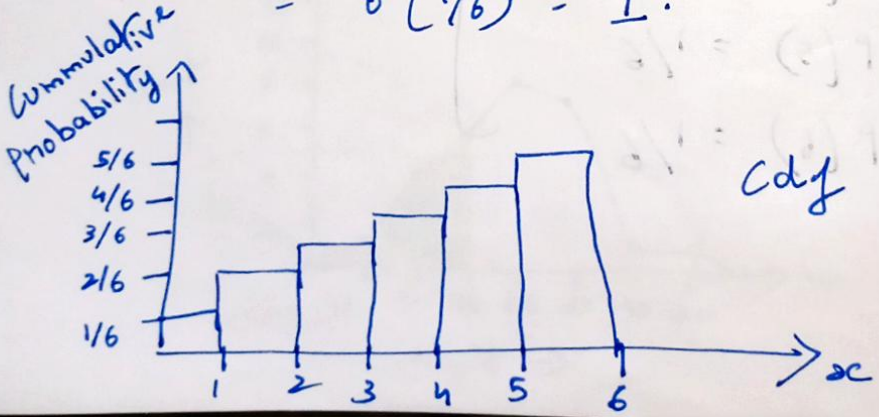
$P(X \leq 3) = P(X=1) + P(X=2) + P(X=3)$

$= 1/6 + 1/6 + 1/6 = 3/6 = 1/2$

$P(X \leq 6) = P(X=1) + P(X=2) + P(X=3) +$

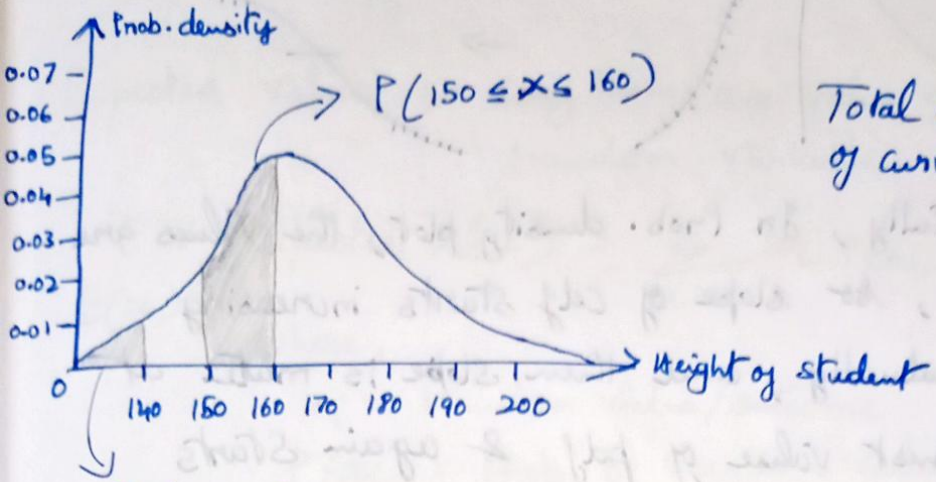
$P(X=4) + P(X=5) + P(X=6)$

$= 6(1/6) = \underline{1}.$

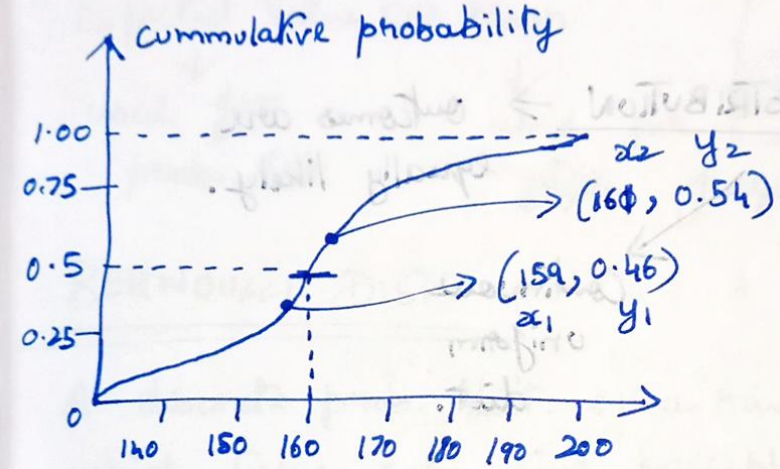


② Prob. Density func. (PDF)

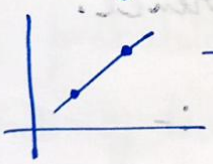
→ Random variable is continuous in nature



$P(x \leq 140)$



Prob. density of a PDF plot is the slope/gradient of CDF at a given point.

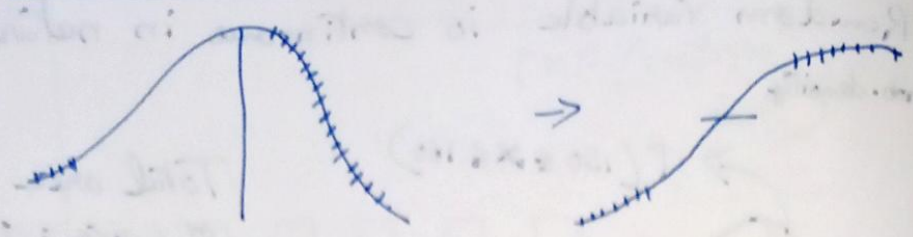


$$\rightarrow \frac{y_2 - y_1}{x_2 - x_1}$$

for the above graph,

$$\text{Slope} = \frac{0.54 - 0.46}{161 - 159} = \frac{0.08}{2} = 0.04$$

Observation



Initially, In Prob. density plot, the values are less, so slope of cdf starts increasing gradually, and then slope is max. at highest value of pdf & again starts declining.

UNIFORM DISTRIBUTION → outcomes are equally likely.

↓

Discrete uniform dist.

Continuous uniform dist.

1) **Discrete Uniform dist.** :- The outcomes are discrete and have equal probability of occurence.

Notation of Uniform dist. :-

$U(a,b)$ | $Unif\{a,b\}$

Mean of discrete Uniform dist. :-

$\frac{a+b}{2}$ (a => min. value
b => max. value)

Variance of discrete Uniform dist. :-

$$\frac{n^2-1}{12}$$

Expected value => Long term avg. value of a random variable.

E.V (Expected value) = $\sum_{i=1}^{n} x_i P(x_i)$

$E(x)$

where,

x_i = random value/outcome

$P(x_i)$ = Prob. of random value/outcome.

Expected value <=> Mean

↓ used for prob. dist.

↓ used for ~~Prob.~~ frequency dist.

$Var(x) = E(x^2) - (E(x))^2$

BERNOULLI DIST.

A discrete prob. dist. of a random variable which takes only two possible outcomes, typically labelled as success (1) and failure (0).

Sum of probabilities = 1.

Ex., Prob. of tossing a coin

$P(x=H) = 1/2$

$P(x=T) = 1 - 1/2 = 1/2$

$p+q=1$, $p=1-q$, $q=1-p$.

The probabilities need not be 1/2 & 1/2.
It can be 0.7 & 0.3, 0.9 & 0.1,
0.8 & 0.2, ~~& 0.5~~ 0.6 & 0.4, etc.

PMF $\Rightarrow$ $P(x=k) = \begin{cases} p & \text{if } k=1 \\ 1-p & \text{if } k=0 \end{cases}$

$\Downarrow$

Combining together, we get

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

i) If $k=1$,

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

ii) If $k=0$,

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

$\therefore$ The total prob. should be 1, we add both the outputs above.

$$p + 1 - p = 1 \;//$$

Conditions of Bernoulli Distribution

1. No. of trial should be finite.
2. Each trial should be independent.
3. only two possible outcomes.

④ Prob. of each output should be same in every trial.

Ex., Pass or fail in a test,
win or lose a match, rain today or not.

Mean & Variance of Bernoulli distribution.

$\Downarrow$ $\Downarrow$

p $p(1-p)$

BINOMIAL DIST.

Bi → two → 2 outcomes

Binomial dist. is 'n' bernoulli trials.

PMF of Binomial dist. $= {}^{n}C_{k}\, p^{k}(1-p)^{n-k}$

$${}^{n}C_{k} \Rightarrow \binom{n}{k} = \frac{n!}{n-k!\cdot k!}$$

Ex., with 3 tosses, what is the probability of getting exactly 2 heads?

$n = 3$, $k = 2$

$$P(x=2) = {}^{n}C_{k}\, p^{k}(1-p)^{n-k}$$

$$= {}^{3}C_{2}\,(0.5)^{2}(0.5)^{3-2}$$

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

$$= \frac{3\times 2}{1\times 2} \times (0.5)^3$$

$$= 3\times(0.5)^3 = 0.375.$$

Ex., when you toss a coin 10 times, what is the prob. that you will get head 3 time?

$n = 10, \quad k = 3$

$$P(x=3) = {}^{10}C_3 (0.5)^3 (1-0.5)^{10-3}$$

$$= \frac{10!}{(10-3)!\cdot 3!} \times 0.5^3 \times 0.5^7$$

$$= \frac{10!}{7!\cdot 3!} \times 0.5^{10}$$

$$= \frac{10\times 9\times 8\times 7!}{7!\times 3!} \times 0.5^{10}$$

$$= 5\times 3\times 8 \times 0.5^{10}$$

$$= 120\times 0.5^{10} = 0.12.$$

Mean of Binomial dist. $\Rightarrow np$

Variance of Binomial dist. $\Rightarrow npq$ or $np(1-p)$

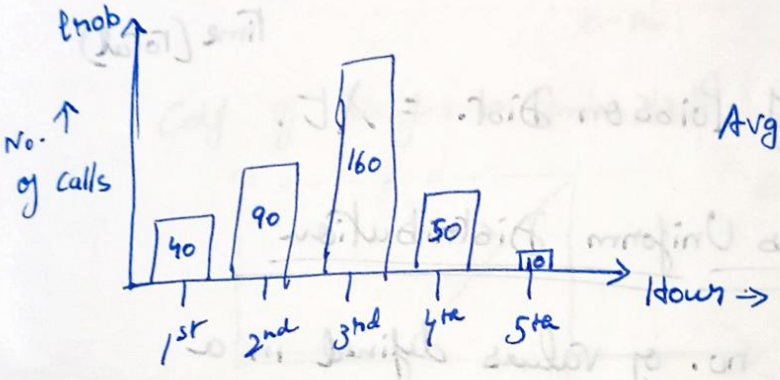
POISSON DISTRIBUTION

A discrete probability distribution that describes the no. of events that occur within a fixed interval of time or space, given a known average rate of occurence.

→ No. of events occuring in a fixed time interval.

Ex:, Customer care

Expected no. of calls to occur every time (1hr) interval is 100.



Another Ex:, No. of people visiting a temple in any hour.

$$pmf \Rightarrow P(X=x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

where, $e \Rightarrow$ euler no.

$= 2.71828$

$\lambda \Rightarrow$ avg. rate of events every interval

Ex.; $\lambda = 10$ (avg. people visiting temple in any hour)

Question ⇒ Person visiting at 5th hour?

$$P(x=5) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

$$= \frac{(2.718)^{-10} \cdot 10^{5}}{5!}$$

$$=$$

Mean of Poisson Dist. $= \lambda \times t$ (t: time (total))

Variance of Poisson Dist. $= \lambda t$.

ii) Continuous Uniform Distribution

→ Infinite no. of values defined in a specified range / bound.

→ Random variable is continuous.

→ Rectangular distribution.

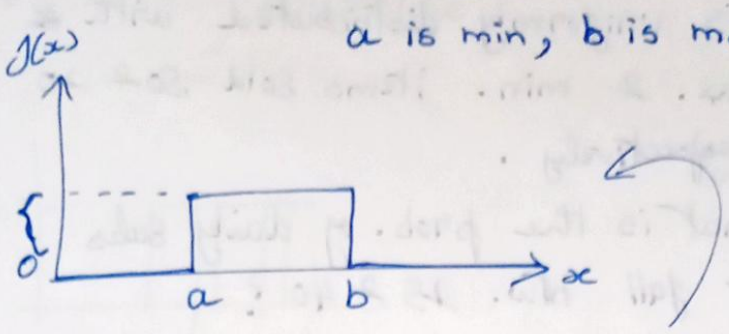
Ex.; OTP generation (random & infinite possibilities),
waiting time at a bus stop.

Notation : $U(a,b)$

Parameter : $-\infty < a < b < \infty$,

$b >> a$,

a is min, b is max.



Pdf of cont. uniform dist.

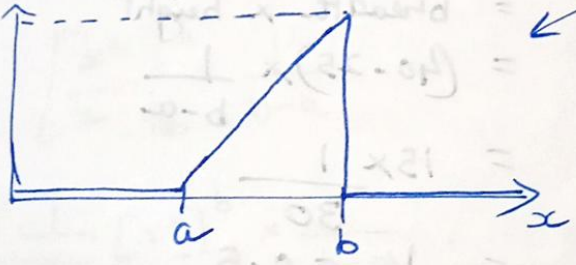
Prob. = Area of rectangle (max. possible value is 1)

= base x height

$b-a \times (f(x)) = 1$

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

cdf of uniform cont. dist.



$$cdf = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$$

line eq"

mean/median $= \frac{1}{2}(a+b)$

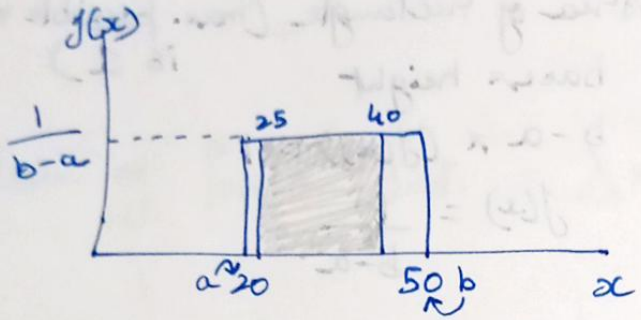
avg. of pdf / center of distribution

$$\text{Variance} = \frac{1}{12}(b-a)^2$$

Ex, 1. The no. of items sold at a shop daily is uniformly distributed with a max. & min. items sold 50 & 20 respectively.

what is the prob. of daily sales to fall btw. 25 & 40?

$$P(25 \le x \le 40) = ?$$



Prob. = area of □le

= breadth × height

$$= (40-25) \times \frac{1}{b-a}$$

$$= 15 \times \frac{1}{30}$$

$$= \frac{1}{2} = 0.5.$$

∴ 50% chance that the no. of items sold is btw. 25 & 40.

for a continuous Random Variable with prob. density function, $f(x)$,

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

$\sum x \cdot f(x)$ → for pmf discrete = $\sum$, continuous = $\int$

$$E(x) = \int_a^b x \cdot f(x)\, dx = \int_a^b x \cdot \frac{1}{b-a} \cdot dx$$

$$= \frac{1}{b-a} \int_a^b x\, dx$$

$$= \frac{1}{b-a} \left[\frac{x^2}{2}\right]$$

$$= \left[\frac{x^2}{2}\right]_a^b \times \frac{1}{b-a}$$

$$= \frac{1}{2}\left[x^2\right]_a^b \times \frac{1}{b-a}$$

$$= \frac{1}{2}(b^2 - a^2) \times \frac{1}{b-a}$$

$$= \frac{1}{2}(b-a)(b+a) \times \frac{1}{b-a}$$

$$= \frac{b+a}{2} \longrightarrow \text{mean (Expected value)}$$

$$Var(x) = E[x^2] - (E[x])^2$$

$E[x]$ → $\frac{a+b}{2}$

$E[x^2]$ ↙

$$= \int_a^b x^2 f(x)\,dx$$

$$= \frac{1}{b-a} \int_a^b x^2\,dx$$

$$= \frac{1}{b-a} \left[\frac{x^3}{3}\right]_a^b$$

$$= \frac{1}{3(b-a)} (b^3 - a^3)$$

$$= \frac{1}{3(b-a)} \times (b-a)(b^2 + ab + a^2)$$

$$= \frac{b^2 + ab + a^2}{3}$$

$$Var(x) = \frac{b^2 + ab + a^2}{3} - \left(\frac{a+b}{2}\right)^2$$

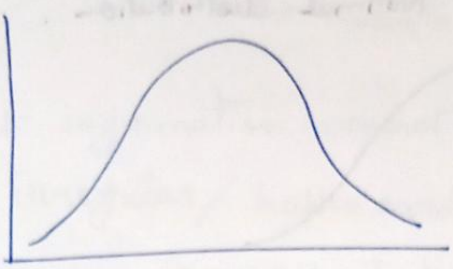
$$= \frac{b^2 + ab + a^2}{3} - \frac{a^2 + b^2 + 2ab}{4}$$

$$= \frac{4b^2 + 4ab + 4a^2 - 3a^2 - 6ab - 3b^2}{12}$$

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

Normal / Gaussian distribution

→ A continuous prob. distribution.



* Most of the real world data follows a Normal dist.

characteristics of Normal Dist.

→ Symmetrical about mean.

→ mean = median = mode.

→ skewness = 0.

Empirical rule of a Normal dist.

65% of data of a Normal dist. lies within 1 standard deviation from mean.

95% → within 2 std. dev. from mean of data

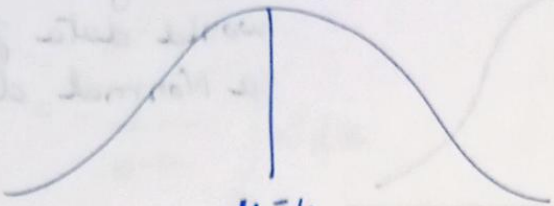
99.7% → within 3 std. dev. from mean of data

Standard Normal Distribution

→ SND is continuous prob dist.

→ A special case of Normal distribution.

Ex:,



$\mu = 4$

$\sigma = 1$

How many standard deviation 4.5 is away from mean?

$$Z_{score} = \frac{x - \mu}{\sigma} = \frac{4.5 - 4}{1} = 0.5.$$

∴ 4.5 is 0.5 std. dev. away from mean.

Use case:- Many of the machine learning algorithms like linear regression, logistic regression, clustering requires scaling

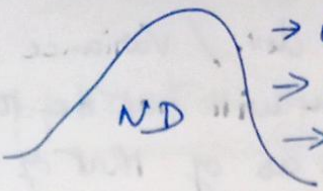
⇓

Standardization

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

This way the computation is faster.

Central Limit Theorem



→ 68% - 95% - 99.7%.
→ symmetric
→ mean = median = mode

But information cannot be obtained for irregular / haphazard distributions.

This can be solved by Central Limit Theorem (CLT).

CLT (Central Limit Theorem)

→ If you have a population with a mean μ & std. dev. σ and take sufficiently large no. of random samples from the population with replacement, then the distribution of sample means will be approximately normally distributed.

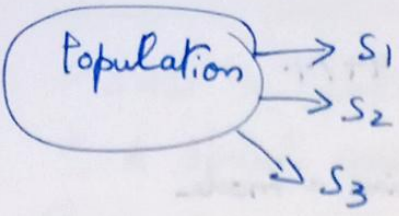
→ Sampling mean of a population (μ, σ) will ~~proba~~ approximately be a normal distribution $(\mu, \sigma/\sqrt{n})$.

$n \Rightarrow$ sample size

Two Conditions of CLT :-

1) The no. of samples should be large.
2) The sample size should be ≥ 30.
   (Except the pop. dist. which is already a normal dist.)

Standard error = $\sigma/\sqrt{n}$



The std. dev. / Variance of Sample will not be the same as of that of the population.

$\frac{\sigma}{\sqrt{n}}$ → Higher the sample size, standard error will be low.

$SE \propto \frac{1}{n}$, $SE \downarrow n \uparrow$

Standard Error is used to minimize the variability of btw. sample & population.

Problem. 1. You have a population with $\mu = 100$ & std. dev. $\sigma = 20$. If you have sample size = 50 from the population, what is prob. that sample mean will be less than 105?

$\mu = 100$, $\sigma = 20$, $n = 50$, $\bar{X} = 105$

(∵ sample is mentioned, we have to use central Limit Theorem).

$$Z_{score} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{105 - 100}{20/\sqrt{50}}$$

$$= \frac{5 \cdot \sqrt{2}}{4} = 1.7675$$

From the Zscore, we have to use the Z-table to find the probability value.

Application of Z-Score

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

Ex:-1.

$N(\mu = 50\,cm,\ \sigma = 20\,cm,\ D(data\ point) = 110)$

How many std. dev. D is away from mean?

$$Zscore = \frac{x_i - \mu}{\sigma} = \frac{110 - 50}{20} = \frac{60}{20} = 3.$$

Ex.,2.

$X = \{1, 2, 3, 4, 5, 6, 7\}$

$\mu = 4,\ \sigma = 1$

what % of score will fall above 4.5 ?

(or)

what is the prob. that score is more than 4.5 ?

$$Zscore = \frac{x - \mu}{\sigma} = \frac{4.5 - 4}{1} = 0.5.$$

↓

4.5 is 0.5 std. dev. away from mean.

For each Zscore, There is a prob. value.

From Z-table,

Prob. for Zscore of 0.5 = 0.6915.

0.6915 is area under the curve below 0.5 (z-score of 4.5).

For more than 4.5,

$1 - 0.6915 = 0.31.$

$\therefore$ 31% chance is there that marks scored will be greather than 4.5.

Ex.,3. The avg. IQ is 100 with $\sigma = 15$. what % of people lower than IQ of 80 ?

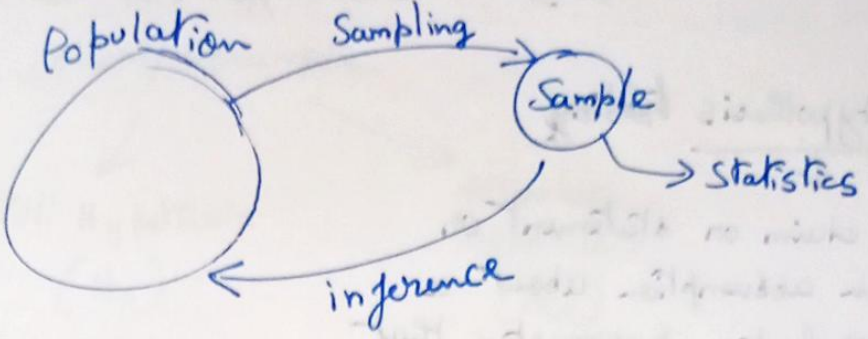
$$Z\text{score} = \frac{80 - 100}{15}$$

$$= \frac{-20}{15} = -1.33$$

Z-table value for $-1.33 = 0.0918$.

$\therefore \sim 9.2\%$ of people have $IQ < 80$.

Point Estimate

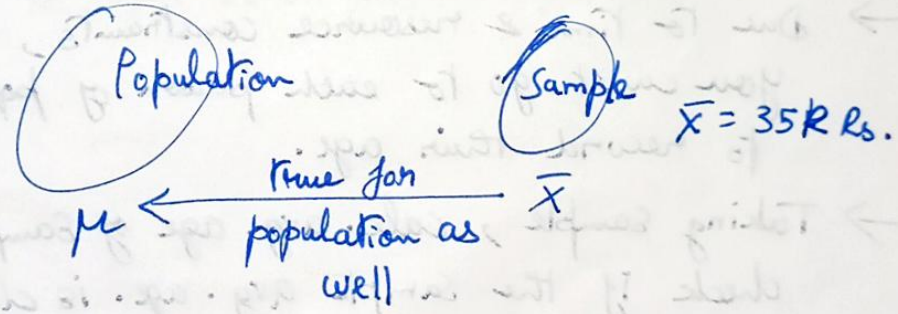


Estimate

↳ A specified observed value used to estimate an unknown population parameter using sample.

(1) Point Estimate ⇒ A single value used to estimate the true value of a population parameter.

Ex.: Avg. salary of IT employee



(2) Interval Estimate → Range of values used to estimate the unknown population parameter.

Ex.: I will score btw. 70-80 marks in the exam I wrote.

HYPOTHESIS TESTING AND MECHANISM

Hypothesis

↓

A claim or statement or an assumption about a population parameter that can be tested using statistical methods.

1. Null Hypothesis : The initial or default assumption.
2. Alternate Hypothesis : The opposite of Null hypothesis.

Hypothesis Testing

→ You made a claim about the population (say avg. age is 45 years) (hypothesis).

→ Due to time & resource constraints, you can't go to each person of population to record their age.

→ Taking sample, calc. avg. age of sample & check if the sample avg. age. is close to the claim that you have made about population ⇒ Trying to verify Hypothesis

Hypothesis Testing

mechanism of Hypothesis Testing

(1) Frame the hypothesis

claim/Assumption

↙ Null Hypothesis (H_0)

↘ Alternate Hypothesis (H_A)

Ex.; claim/Assumption ⇒ Avg. age of people in DS skills is 45 years

$H_0 : \mu_{age} = 45$

$H_A : \mu_{age} \neq 45$

[IMP. :- H_0 will have equality sign].

(2) Statistical Analysis (Pvalue, Significance level)

(3) Conclusion

(To reject H_0 or fail to reject H_0)

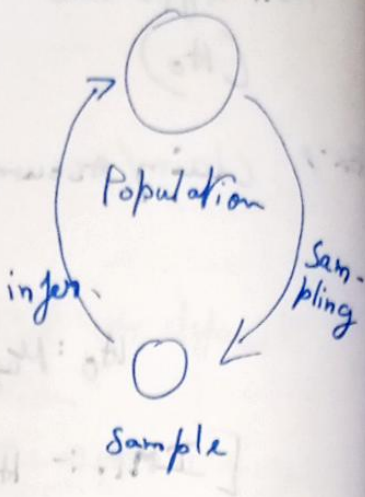
P- Value

→ Probability value calculated from a statistical test, corresponding to z-score of a sample.

You don't know anything about population.

You made a claim/ hypothesis

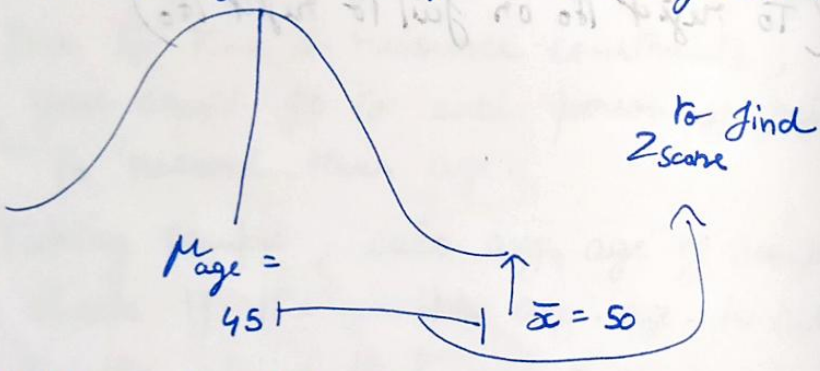
$\mu_{age} = 45$ years



→ Claim/hypothesis : μ_{age} of employee is 45 years (Population)

→ sample is taken

→ Avg. age of the sample → 50 years.



50 years is far from 45 years, that's why the claim can be rejected?

Yes.

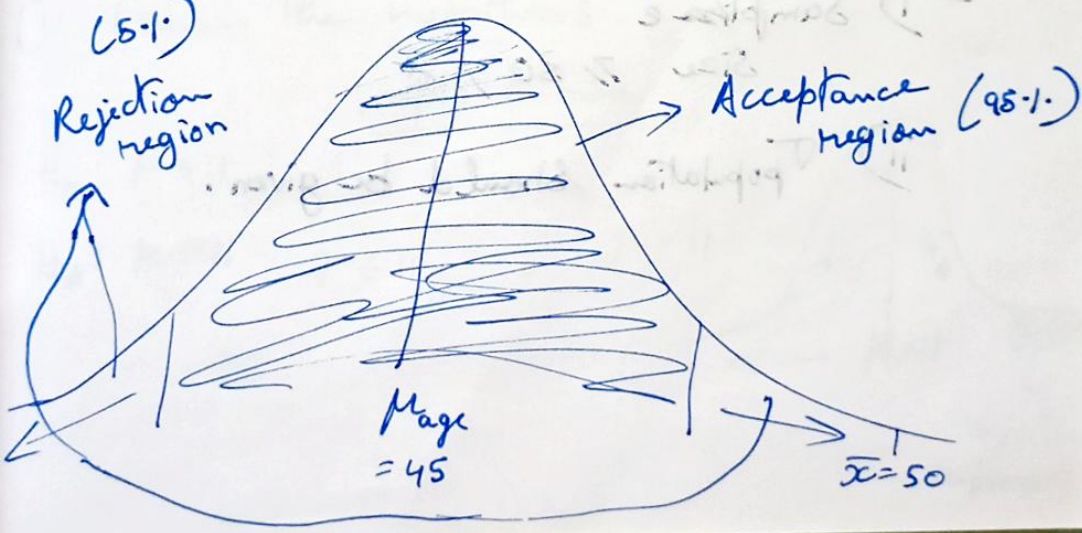
To know how far a data pt. is away from the mean, we calc. Zscore.

$$Z_{score} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Corresponding to Zscore, there is a probability value. That prob. value is called

p-value

Clinical trial of a medicine of 1000 people
- ↳ 950 got cured with this medicine
- ↳ medicine is working 95%
- ↳ Out of 100 times, 95% will work
  [95% confident]
- ↳ 5% of the time medicine doesn't work
  [5% margin of error]
  ⇓
  Level of Significance



if, p-value < Significance value (α),
    reject the H_0

else,
    fail to reject H_0

TYPES OF HYPOTHESIS TESTING

① z-test ② t-test ③ chi-square test
④ ANOVA test

① z-test

Criteria for z-test:

CLT → $\mu = \mu$, $\sigma' = \sigma/\sqrt{n}$
↳ sampling mean distribution → Normal dist.
→ Sample size ≥ 30

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

i) Sample size ≥ 30, &
ii) $\sigma_{population}$ should be given.

Ex.,

(1) Suppose a child Psychologist says that the average time a working mother spends talking to their children is upto 11 mins. To test the hypothesis, you conducted an experiment with random sample of 100 working mothers and find that they spend 11.5 mins. per day talking with their children. Assume prior research suggests that the population standard deviation is 2.3 mins.. Conduct the test with 5% level of significance ($\alpha = 0.05$).

Soln.

$S.S \geq 30$ ✓ (100)

σ population is given (2.3 mins) ✓

Z test criteria is satisfied.

$\mu = 11$ mins. $\bar{x} = 11.5$ mins. $\sigma = 2.3$ mins.

$n = 100$ $\alpha = 0.05$

Step

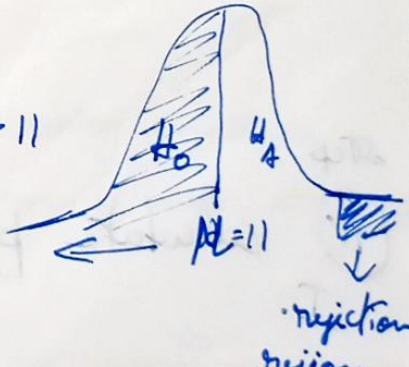
(1) Frame the hypothesis.

upto 11 mins

$H_0 : \mu \leq 11$

$H_A : \mu > 11$

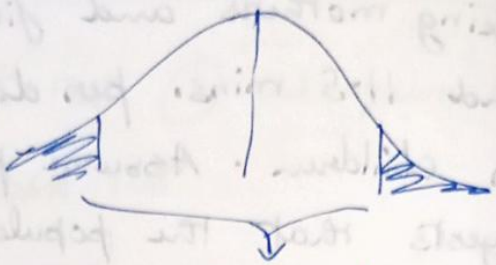
$\mu \leq 11$ $\mu > 11$



rejection region is at H_A side.

One tail test: (rejection at one side of the dist.). One tail test occurs because of ($\mu <$) or ($\mu >$) condition.

For ($\mu \neq$) condition two tail test occurs.



Two rejection areas for two-tail test.

step (ii) $\alpha = 5\%$, one tail test

type of test = Z test

step (iii) Z score $\bar{X}$ (test statistics)

$\Downarrow$

Z statistics

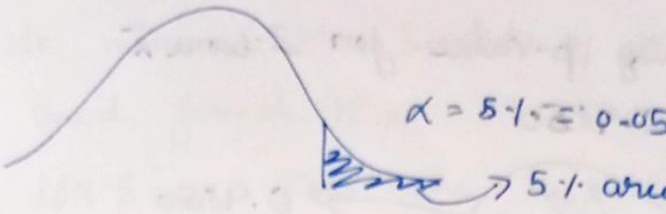
$$Z_{score\ of\ \bar{x}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{11.5 - 11}{23/\sqrt{100}} = 2.17$$

step (iv) calculate p-value on Z critical I.

Zcritical → α to Zscore

p-value → Zscore to p-value



$\alpha = 5\% = 0.05$

→ 5% area

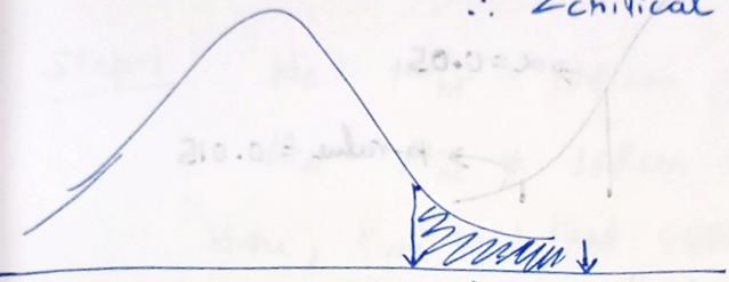
⇓

To be converted into Zscore

Z table

value for 0.05 = −1.64

∴ Zcritical = 1.64.



Zcritical = 1.64 | Zscore = 2.17 ($\bar{x}$)

if Zscore of $\bar{x}$ (Zstatistics) lies in Zcritical region, you reject the H_0.

Step (V) Conclusion.

I. ∵ Zstatistics > Zcritical,

we reject the H_0

∴ H_A is not rejected & working mothers spend more than 11 mins. with their children.

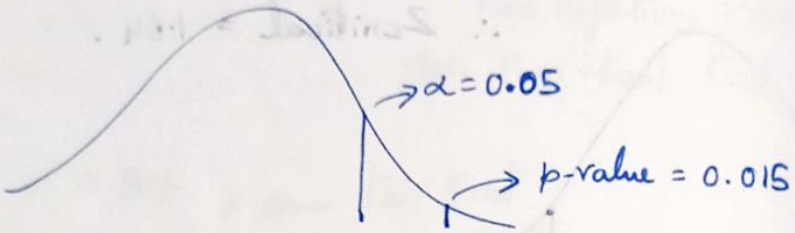
Step (IV)

II. Calculate p-value to find hypothesis

$Z_{score}\ \bar{x} = 2.17$

corresponding p-value for Z-score $\bar{x}$

$= 0.9850$



$\because$ p-value $< \alpha$,

we reject the hypothesis.

Step (V)

II. $\because$ p-value < 0.05,

we reject the hypothesis.

This means that H_A is not rejected & working mothers spend more than 11 mins with their children.

Ex.,(2) The avg. height of all residents in a city is 168cm with a σ of 3.9cm. One researcher believes that mean is different. We measure the height of 36 individuals and found that the avg. height is 169.5 cm. Test the hypothesis at 95% confidence interval. $\alpha = 5\%$.

Soln.

$\mu_{ht} = 168\text{cm}$, $\sigma = 3.9\text{cm}$, $n = 36$,

$\bar{x} = 169.5\text{cm}$ $\alpha = 5\%$

Step-1: $H_0 : \mu_{ht} = 168\text{ cm}$

$H_A : \mu_{ht} \neq 168\text{ cm}$.

Here, Two tail test occurs due to ($H_A \neq$) condition.

Step-2: $\alpha = 5\%$

due to Two tail test,

$\alpha = 5/2 = 2.5\%$ on 0.025.

Step-3: $Z\text{statistics} = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{169.5 - 168}{3.9/\sqrt{36}}$

$= 2.31$.

Step-4: calculating p-value

Corresponding p-value of 2.31 obtained from Z table is 0.9896.

0.9896

$\Rightarrow 1 - 0.9896$

$\Rightarrow 0.01$

Two tail test $\Rightarrow$ Total areas

$= 0.01 + 0.01$

$= 0.02.$

Step-5:

p value < 0.05

$\therefore$ reject the H_0.

$\therefore \mu_{ht} \neq 168cm.$

Main condition

if (p-value $< \alpha$) or ($Z_{critical} < Z_{stats}$):

reject H_0

Type-I and Type-II error

Type-I Error : Rejection of null hypothesis when it's actually true.

Ex.; An innocent person is convicted.

Type-II Error : Failure to reject the null hypothesis when it's actually false.

Ex.; A guilty person may not be convicted.

	H_0 is True	H_A is True
Support H_0	Correct conclusion	Type II Error
Support H_A	Type I Error	Correct Conclusion

Scenario-I : Reject the null hypothesis, when ~~actually~~ it's false. ⇒ good

Scenario-II : Reject the null hypothesis, when it's true. ⇒ Type I error

Scenario-III : Failure to reject null hypothesis, when it's false ⇒ Type II error

Scenario-IV : Failure to reject null hypothesis, when it's true. ⇒ good

Ex., A person has committed a crime or is suspected to have committed a crime.

H_0 : Person is innocent

H_A : Person is guilty.

Scenario - 1 : An innocent player is charged for a crime & in the court he is declared guilty of the crime

Type - I error.

Scenario - 2 : A guilty person is charged for a crime but the court declares him as innocent.

Type - II error.

Student's t-distribution

(2) t-test

Criteria for t-test

i) Sample size < 30.

ii) σ (population std. dev.) not given.

Sample size ≥ 30 (any no.) ↓ it follows std. Normal distribution.

degree of freedom = sample size - 1

For Normal Distribution, tails are thin.

As the d.o.f decreases, tails become thicker.

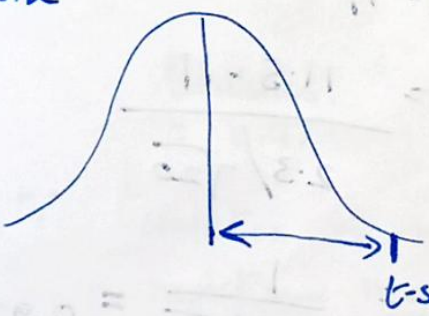
The t-distribution depends on degree of freedom. (d.o.f).

t-distribution → t-statistics

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

t-score

[where, s => Sample std. dev.]



t-distribution → degree of freedom

↓

t-Table

degree of freedom (d.o.f) = Sample size - 1

Ex:- Degree of freedom denotes the unconstrained positions existing in a dataset.

T-Test Problems

Q.1. Suppose a child Psychologist say that the avg. time working mother spend every day talking to their children is upto 11 minutes per day. To test the hypothesis, you took a random sample of 20 working mother & found avg. time they spent is 11.5 minutes. The sample std. dev. is 2.3 mins. conduct test with 5% level of significance.

Soln: $n = 20$ $\bar{x} = 11.5$ step - 3 t-test

$s = 2.3$ $\mu = 11$

$$t\text{-test} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{11.5 - 11}{2.3/\sqrt{20}}$$

$$= \frac{0.5}{2.3/2\sqrt{5}} = \frac{1}{2.3/\sqrt{5}} = 0.97$$

Step - 1

$H_0 : \mu \leq 11$,

$H_A : \mu > 11$

Step - 2

$\alpha = 0.05 \rightarrow 5\% \rightarrow$ one tail test

$dof = 20 - 1 = 19$

Step - 4

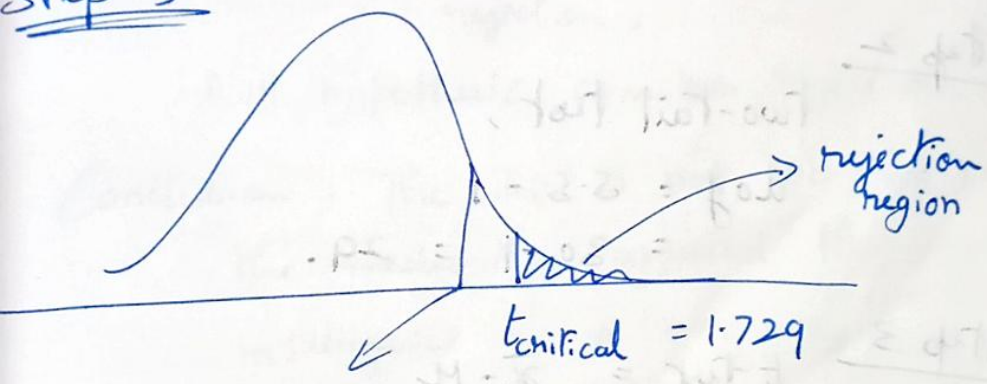
$t_{critical}$ corresponding to 0.05

$dof = 19$, one tail test.

From t-table,

$t_{critical} = 1.729$

Step - 5



$t_{statistics} = 0.97$

$\therefore$ $t_{statistics}$ is not falling in the rejection region, we fail to reject the null hypothesis

Conclusion : The avg. time working mother spend talking to their children is $\leq$ 11 mins.

Q.2. In population, avg. IQ is 100. A team of researchers want to test a medicine to check the +ve on -ve effect on intelligence. A sample of 30 participants who took medicine has an mean IQ of 140 with std. dev. of 20. did the medication affect the intelligence? Test hypothesis with 5% alpha.

Soln:

Step 1 : $H_0 : \mu = 100$

$H_A : \mu \neq 100$

Step 2

two-tail test,

$dof = S.S - 1$

$= 30 - 1 = 29.$

Step 3 $t\text{-test} = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

$\bar{x} = 140 \quad s = 20$

$\mu = 100 \quad n = 30$

$$t\text{-Test} = \frac{140 - 100}{20/\sqrt{30}}$$

$$= \frac{40}{20/\sqrt{30}} = \sqrt{29} \times 2$$

$$= 10.9545.$$

Step 4

$t_{critical}$ at $\alpha = 0.05$, d.o.f = 29

$= 2.045$.

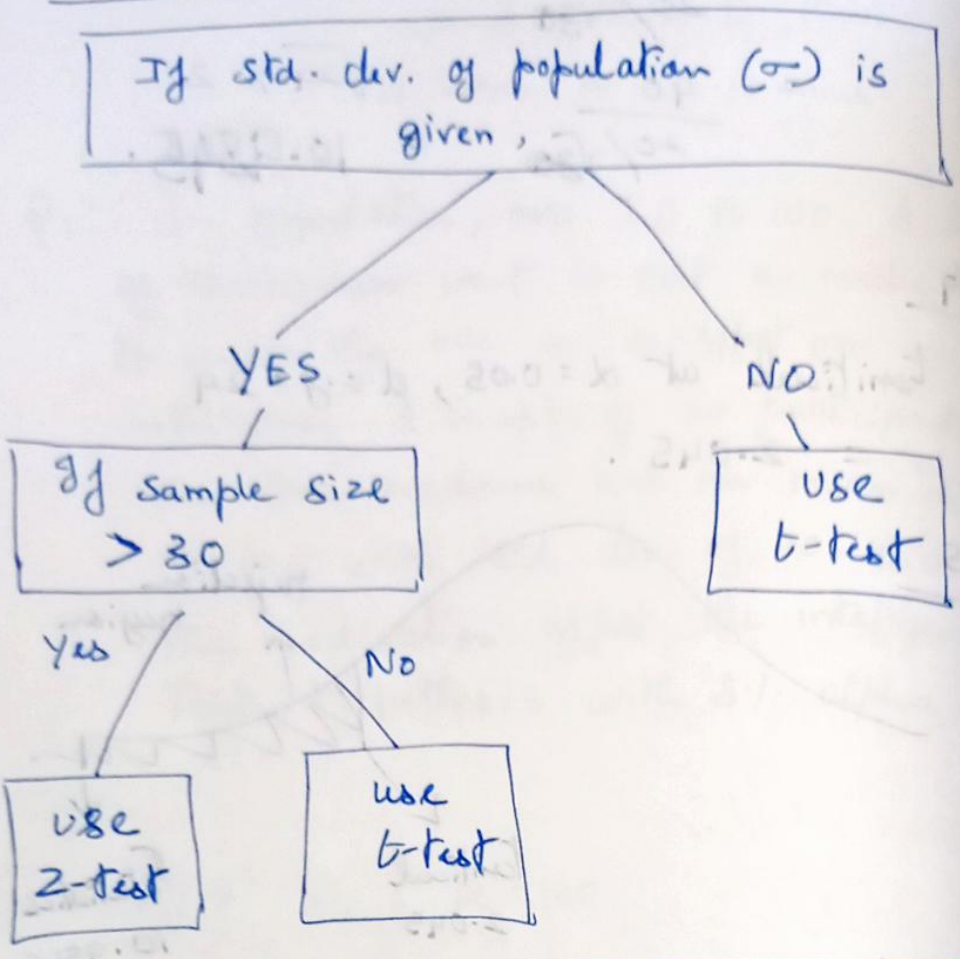
Step 5

$\therefore$ $t_{statistics}$ lies in region of rejection,

$\therefore$ Null hypothesis can be rejected.

Conclusion: The IQ is not 100 (avg.) & the medication affected the intelligence is a +ve way.

T-test vs Z-test



Confidence Interval & Margin of Error

Ex.; what score will you get in exam?

80 → chances of not getting are high due to single data point prediction.

chances of getting are high due to interval ← 75-85
(more CONFIDENCE)

80 is a point Estimate.

CONFIDENCE INTERVAL → interval (range of values)

↓

point estimate ± error

Ex.; Confidence Interval ⇒ 80 ± 5

⇒ 75 — 85

Point Estimate is located exactly in the middle of confidence interval.

CI = point estimate ± margin of error

(in hypothesis testing, margin of error is α)

In Z test (two tail test),

$$CI = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where,

$\bar{x}$ ⇒ Sample mean,

n ⇒ Sample size,

σ ⇒ population std. dev.,

$Z_{\alpha/2}$ ⇒ Z score corresponding to given $\alpha/2$

α ⇒ Level of significance (or) m

In Z test (one tail test),

$$CI = \bar{x} \pm Z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

In t-test (two-tail test),

$$CI = \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

(s → Sample std. dev.)

§ [illegible]

[illegible]

Probability & Baye's Theorm

Probability rules :

1) For any event A → $0 \leq P(A) \leq 1$
2) Sum of all probabilities of all possible outcome is 1
3) Complement rule :-

$$P(\text{not } A) = 1 - P(A)$$

4) General addition rule :-

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

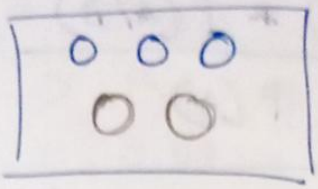
5) Multiplication rule :-

Events

- Independent events
  - → Tossing a coin
  - → Throwing a dice
- Dependent events.

Dependent Events

A box contains 3 Blue balls & 2 Grey balls.



$P(Gr) = 2/5$ → one grey ball is taken out → $P(Gr) = 1/4$.

$P(B) = 3/5$ → one ball is taken out → $P(B) = 3/4$.

The probability is changing based on the last event..

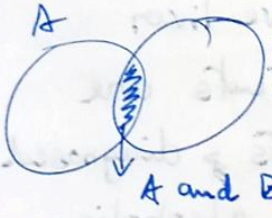
These are called dependent events.

If Events are dependent,

$$P(A \text{ and } B) = P(A) \times P(B/A)$$

$P(A)$: Prob. of A

$P(B/A)$: Prob. of B when A has already occured.



A and B (or) B and A

$$P(A \text{ and } B) = P(A) * P(B/A) - (1)$$

$$P(B \text{ and } A) = P(B) * P(A/B) - (2)$$

Equating eqn's (1) & (2),

$$P(A) * P(B/A) = P(B) * P(A/B)$$

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

(or) BAYES THEORM

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

$P(B/A)$ = Prob. of event B, given/when event A has already occured.

$P(A/B)$ = Prob. of event A, given/when event B has already occured.

Ex:,

(1) 10% of patients in a clinic have liver disease. 5% of clinical patients are alcoholics. Among these patients diagnosed with liver disease, 7% are alcoholic. what is prob. of patients having liver disease, given that he/she is an alcholic?

Soln.

$P(A)$ ⇒ Prob. of having liver disease
$= 0.10$

$P(B)$ ⇒ Prob. of alcoholism
$= 0.05$

$P(B/A) = 0.07.$

$P(A/B) = ?$

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

$$= \frac{0.10 \times 0.07}{0.05}$$

$$= 0.14.$$

$$= 14\%.$$

∴ The prob. of patient having liver disease given that he/she is an alcoholic is 14%.

Use Cases

↳ Naive Bayes classifier (Machine Learning Model)

↳ Bayesian Statistics (data analysis & parameter estimation based on bayes theorem).

Chi-Square Distribution & chi-Square Test

chi-square distribution is a probability dist. that describes the distribution of a sum of squares of k random variables.

→ degree of freedom (k) = n-1

what is chi-square dist. ?

S_1	S_2	S_3
5	2	5
8	3	6
6	d.o.f =	8
2	2-1 = 1	d.o.f
3		= 3-1 = 2

d.o.f = 5-1
= 4

→ If you square the no. of any sample, it will closely follow chi-square dist.

Observations

→ chi-square dist. shape is determined by 'k'.
→ Non-negative distribution.
→ right-skewed dist.

chi-square test (X^2 test)

→ follows chi-square distrib.
→ goodness of fit test ⇒ used to compare the observed & expected categorical data

→ Test of independence ⇒ To determine the relationship btw. two categorical variables.

→ It tests the claims about population proportions.

→ χ^2 test is non-parametric test.

Parametric test ⇒ some assumption about population.

Many articles says χ^2 distrib. is parametric, but, in true sense, it is not assuming anything about population. (only depends on d.o.f).

Ex. ① In a class of 75 students, 11 are left-handed. Theoretically, 12% of people are left-handed. Prove or disapprove the theory. $\alpha = 5$.

Soln.

Step - 3

	Observed	Expected
Left-handed	11	$\frac{12}{100} \times 75 = 9$
Right-handed	64	66
	75	75

$$\chi^2_{\text{statistics}} = \sum \frac{(\text{observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(11-9)^2}{9} + \left(\frac{64-66}{66}\right)^2$$

$$= \frac{4}{9} + \frac{4}{66} = \frac{4}{9} + \frac{2}{33}$$

$$= 0.44 + 0.06$$

$$= 0.5.$$

Step-1

$H_0 : \mu = 12\%$

$H_A : \mu \neq 12\%$

Step-2

$\alpha = 5\%$, chisquare dist.,

1 tail test (because, the values are squared leading to no negative values)

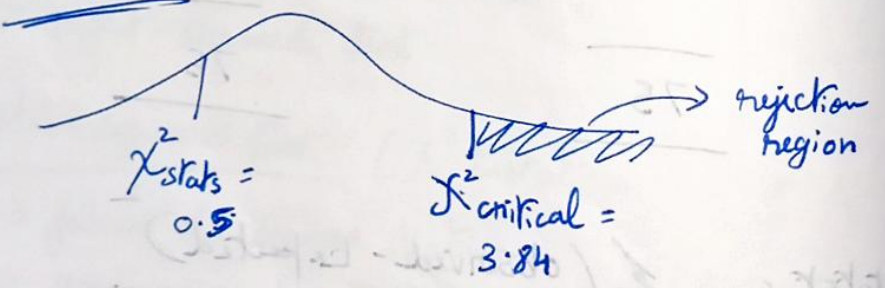
Step-4

$\chi^2_{critical}$ for $\alpha = 0.05 = 3.841$ (from χ^2 dist. table).

d.o.f = No. of groups − 1

= 2 − 1 = 1.

Step-5



if $X^2_{stats} > X^2_{critical}$, reject the H_0.

∴ 12% of people are lyt-handed with 95% confidence.

but here, $X^2_{stats}(0.5) < X^2_{critical}(3.84)$

which means, we fail to reject the null hypothesis.

Conclusion

12% of people are lyt-handed with 95% confidence.

F - distribution (Fischer - Snedecor dist.)

→ Comparison of variances of two or more samples.

→ It's right-skewed and takes only non-negative values.

The f dist. with d_1 & d_2 (degree of freedom) is the dist. of x given by ~~x~~ = $\frac{S_1/d_1}{S_2/d_2}$.

where, S_1 & S_2 are independent random variables ~~std. dev.~~ with χ^2 dist.

d_1 & d_2 are degrees of freedom.

$F_{statistics} = \frac{S_1^2}{S_2^2}$ (Variance ratio test)

σ => population std. dev.

S => sample std. dev.

Observations

→ Since S_1^2 & S_2^2 is there, it's a non-negative dist.

→ As dof1 & dof2 is ≥ 30, the f dist. behaves approximately like a normal dist.

F-Test (Variance ratio test)

Ex., 1. The foll. data is about the no. of bulbs produced daily by two workers A & B.

A ⇒ 40 30 38 41 38 35

B ⇒ 39 38 41 33 32 39 30 34

$\alpha = 0.05$.

Can we consider (based on above data) that worker B is more stable & efficient?

Soln.

Step 1 : $H_0 : S_1^2 = S_2^2$; $H_A : S_1^2 \neq S_2^2$

Step 2 :

F test, one tail test, $\alpha = 0.05$.

Step 3 :

$$\text{F statistics} = \frac{S_1^2}{S_2^2}$$

Worker B

X_2	$\bar{X}_2$	$(X_i - \bar{X}_2)^2$
39	37	4
38	37	1
41	37	16
33	37	16
32	37	25
39	37	4
40	37	9
34	37	9
$\bar{X}_2 = 37$		$\Sigma = 84$

$$S_2^2 = \frac{84}{n-1} = \frac{84}{8-1} = \frac{84}{7} = 12$$

Worker A

X_i	$\bar{X}_i$	$(X_i - \bar{X}_i)^2$
40	37	9
30	37	49
38	37	1
41	37	16
38	37	1
35	37	4
$\bar{X}_i = 37$		$\Sigma = 80$

$$S_1^2 = \frac{80}{n-1} = \frac{80}{6-1} = \frac{80}{5} = 16.$$

$$F_{statistics} = \frac{S_1^2}{S_2^2} = \frac{16}{12} = 1.33.$$

Step 4:

$F_{critical}$ for $\alpha = 0.05$

$dof_1 = 6 - 1 = 5$

↳ in numerator of F-table

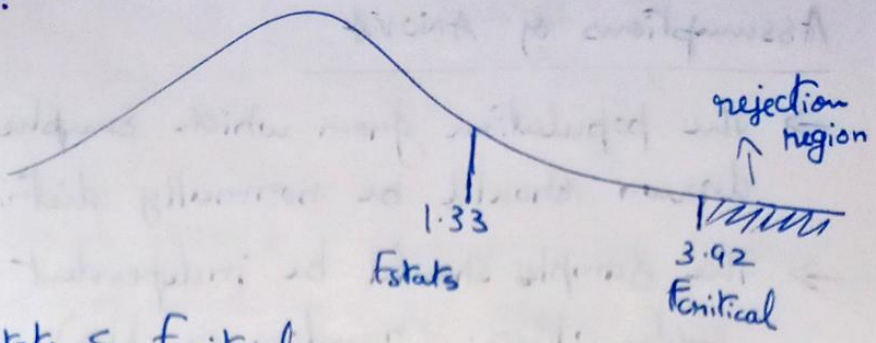
$dof_2 = 8 - 1 = 7$

↳ in denominator of F-table

From F-table,

$F_{critical} = 3.972$

Step 5 :



∵ Fstats < Fcritical,

we fail to reject the null hypothesis.

Conclusion :- Worker B is not more stable & efficient. Both workers work at the same level of stability & efficiency.

ANOVA (Analysis of Variance) & it's Assumption

Definition : ANOVA is a statistical method used to compare the means of 3 or more groups.

Generally, ANOVA is used for more than 2 groups.

Reason : For 2 groups only, F-test is used.

ANOVA Terms :

* Factors / Variables / Criteria
* Levels

Ex., medicine → factor / variable / criteria

1mg : 10mg : 20mg

Levels

Assumptions of ANOVA

→ The population from which samples are drawn should be normally distributed.

→ The samples should be independent of each other (random samples).

→ Absence of outliers.

→ Homogeneity of Variance :- Homogeneity means that the variance among the groups should be approximately equal.

TYPES OF ANOVA

① One Way ANOVA : One factor with atleast two levels & levels are independent.

Ex.: Stress level of employees

Dept. A	Dept. B	Dept. C
8	9	5
5	3	7
1	7	1
2	6	2

② Repeated measure ANOVA : One factor with atleast 2 levels but levels are dependent.

Ex:- No. of hours studied

Day 1 ←	Day 2 ←	Day 3 ←	Day 4
10	8	4	6

(3) Factorial ANOVA : Two or more factors each of which contains atleast 2 levels. The levels can be dependent, independent or both.

If only two factors are present, it is called TWO WAY ANOVA.

Ex:-

	medicine (factor 1)		
	0mg	10mg	20mg
M	5	2	3
	4	6	9
	1	9	4
F	9	1	7
	8	2	6
	7	3	4

gender (factor 2)