$F = \frac{MSC}{MSE}$, $MSC = \frac{SSC}{dfc}$, $MSE = \frac{SSE}{dfe}$

$df_C = C - 1$

~~$df_E = (C-1)(B-1)$~~ $df_e = (C-1)(B-1)$

$MS_B = SS_B / df_B(B-1)$

$df_T = N - 1$

Understanding Statistics

- mean $= \sum x \cdot P(x)$

  where x is the variable and $P(x)$ is the relative frequency (prob) of the variable.

- For continuous distribution, we do ~~summation~~ integration, but for discrete distribution we can do summation.

- Variance $= \sum (x - \bar{x})^2 \cdot P(x)$

. How Variance is $\sum (x-\bar{x})^2 \cdot P(x)$

- Mean of a binomial distribution is np.
- pmf → probability mass function.
- stats.binom.pmf (x, n, p) → Binomial distribution.
- cdf → cummulative density function.
  stats.binom.cdf will add all the pmf's.

$P(x) = {}^nC_x p^x q^{n-x}$

- Variance of the binomial distribution is npq. where $q = 1 - P$.
- stats.poisson.pmf (x, λ) → Poisson distribution.

  where λ is the average number of events that happen in a period Δt.

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

- $\sigma^2 = npq$

  standard deviation $\sigma = \sqrt{npq}$.

- The value you expect to get in statistical experiment is the mean. It is the highest distributed value in the distribution.

- Binomial Distribution is used when, we have categorical information, always we communicate this info interms of proportion (or) percentage. whenever we have percentage information for categorical variables, we can generate binomial distribution.

- Poisson Distribution is used to generate counting statistics (counting information).

- whenever we are dealing with discrete numerical quantity then, poisson distribution.

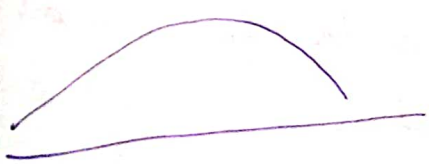
- whenever we are dealing with continuous numerical quantity, then ~~poisson~~ normal distribution (continuous distribution).

- In poisson distribution mean is λ.

	Binomial	Poisson
mean	np	λ
S.D	$\sqrt{npq}$	$\sqrt{\lambda}$

- Discrete distribution.

- Continuous distribution.



- Poisson distribution is independent of n. (mean doesn't effect the poisson distribution)
- Bernoulli is the collection of two binomials.

- Characteristics of Standard Normal Distribution:

$\bar{X} = 0$ } The mean of the variable becomes 0, SD becomes 1.
$SD = 1$ }

- Inverse survival function:

stats.norm.isf function is used in python.

It gives the cut off value (range) for the required area.

- Standard scalar, z scale, z score, (all are same).

- Important z score values,

95% → −1.96 to +1.96

90% → −1.64 to +1.64

99% → −2.58 to +2.58

- In stats.norm.cdf function (−∞, μ, σ).

- Gaussian distribution / Normal distribution

This kind of distribution will form a bell curve 

x (random variable) will form a Gaussian Distribution with some value of μ (mean) and ~~σ (standard deviation)~~ (variance σ (S.D).

There is an empirical formula in Gaussian Distribution,

- Probability of $(\mu - \sigma \leq x \leq \mu + \sigma)$ will be approx. equal to 68%.

It says that x which is a part of our random variable X.

That means 68% of the random variables will fall in this range.

Total number of elements from X present S.D to the between first SD in the left to first SD present in the right (i.e) b/w the first standard deviation is around 68% of the distribution.

68% of the elements of the random variable will be falling in the first standard deviation. 68% is the percentage of the distribution of data from the random in the range from the random variable X.

So, within first standard deviation, 68% of the data exists. (68% of the total distribution).

$\mu-\sigma \le x \le \mu+\sigma \Rightarrow 68\%$
$\mu-2\sigma \le x \le \mu+2\sigma \Rightarrow 95\%$
$\mu-3\sigma \le x \le \mu+3\sigma \Rightarrow 99.7\%$

} This percentage of the total distribution exists in the given range. (Confidence interval).

Log Normal Distribution

A random variable X usually belongs to the log normal distribution, if $\log(x)$ is normally distributed. ($\ln(x)$)

Generally this distribution is similar to bell curve, but right skewed.

- When Gaussian distribution is converted into standard Normal distribution, then ($\mu=0$) and ($\sigma=1$). When we do this, values will be converted according to the standard scalar (z score).

- Log normal distribution says that $x \sim$ Log Normal (x belongs to log normal dist) if if $\ln(x) \sim N(\mu,\sigma)$.
(If log of x belongs to Gaussian or normal distribution with some value of μ and σ.

- Covariance of x and x is variance of x.

- $\text{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$

Covariance

$x\uparrow \; y\uparrow$ then +ve → It cannot say how much positive.
$x\uparrow \; y\downarrow$ then -ve → It cannot say how much negative

Pearson Correlation Coefficient

This is used to overcome the above disadvantage.

- The probability of a variable that falls in between $\mu-\sigma \le x \le \mu+\sigma$ which is basically the range of first standard deviation will approximately equal to 68%.

- 68% of the data points belonging to the random variable x fall within the range of above mentioned range.

• stats.binom.pmf(x, n, p) where — This function gives the prob

x → depends upon the question asked.

n → number of sample size.

P → probability given.

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

60 - 90

20

95% → 1.96

68% → x

• stats.poisson.pmf(n, λ) where. — This function gives the prob

n → depends upon the question.

λ → mean

• stats.poisson cdf adds up everything

•

$\mu-\sigma \le x \le \mu+\sigma$ → 0.6828 → 68.28 %

$\mu-2\sigma \le x \le \mu+2\sigma$ → 0.9545 → 95.45 %

$\mu-3\sigma \le x \le \mu+3\sigma$ → 0.9974 → 99.74 %

} This is the percentage of data points spread within the region.

Somehow related to 2 scores

probability or the area

~~• stats.norm.cdf(x, μ, σ)~~

• stats.norm.cdf(x, μ, σ) — This function gives the probability

x → depends upon the question

μ → mean (average)

σ → standard deviation.

• isf → inverse survival function

stats.norm.isf(0.025) #95% area

→ 1.9599

(The answer gives the range of the 95% area and 99% area)

---

stats.norm.isf(0.005) #99% area

→ 2.575

The isf actually gives us the -2 score to +2 score value.

---

stats.norm.isf(0.05) #90% area

→ 1.64485

Sampling error → Difference b/w population mean and sample mean.

. Confidence interval is sample mean ± (margin of error)

. margin of error = $z_{\text{critical value}}$ (z score value) × $\frac{\text{population SD}}{\sqrt{\text{sample size } (n)}}$

. standard error = σ (SD) $/\sqrt{n(\text{sample size})}$

.

. In theoritical approach we take many trials to calculate mean of the mean.
. In practical approach we do only one round of sampling.

(. margin of error and standard error were used for practical approach)

Standard error of mean is calculated in practical approach.

. De-moivre's approximation of Discrete Distribution with Standard Normal Distribution.

. Generate Binomial distribution with $p=0.5$, $n=25$, and calculate $P(x \leq 14)$.

→ In discrete distribution, the histogram bars are centered on the numbers. This means $P(x \leq 14)$ in Discrete distribution is actually the area under the bars less than $x = 14.5$, we need to account for that extra 0.5, while calculating the same area in continuous distribution.

. with skew norm function, we can generate a skewed distribution.

. H_0 is null hypothesis (no effect hypothesis).
  H_a is alternative hypothesis (effect hypothesis)

. The equality goes to null hypothesis.

. not equal to goes for alternate hypothesis.

. ~~One tail sample means one sided check.~~

. Standard error of the mean $\sigma / \sqrt{n}$.

- Central limit Theorem
- CLT states that for a large sample drawn from a population with mean μ and standard deviation σ, the sampling distribution of mean, follows an approximate normal distribution with mean, μ and standard ~~deviation~~ error $\sigma/\sqrt{n}$ irrespective of the distribution of population for large sample size.
- As a general rule, statisticians have found that for any population distribution, when the sample size is at least 30 and above, the sampling distribution of the mean is approximately normal / Gaussian.

- Confidence interval is also known as acceptance zone for H_0.
- t_{stat} is the distance between two means.

. A hypothesis is a testable statement made in support of a finding or a claim about something in the world around you.
. It should be capable of being tested, either by experiment (or) observation.
. Hypothesis testing is an objective scientific method that is used in making statistical decisions using observational or experimental data.
. Hypothesis testing is an assumption that we make about the population parameter, a premise or claim we want to test.

. For population data. General formula for SD is :-

$$\sigma_{pop} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

. For sample data, General formula for SD is :-

$$s_{samp} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

. The numpy function np.std by default divides it by n.
. ddof (delta degrees of freedom)
(ddof = 1) means the denominator is n-1.

. df['vol'].std() by default divides by n-1.
. np.std[ ] by default divides by n.

. ttest_1samp (samp_array, 300) will give us the t-stat score, p-value.
. tstat is the hypothesis testing formula.
. Every value has its own confidence interval.
. tstat is actually the distance between two means.
. Practically point estimate is impossible, we can do a closer prediction with range estimate with different confidence interval levels.

dependent (paired) / Independent (unpaired)	(Samples) 2 groups	>2 groups
Mean	Two Sample 't' Test	ANOVA
Proportion	Two Sample proportion test	Chi Square Test

Full information to Trade

One Sample Test — Two tail / left tail / Right tail

Two Sample Test — Dependent / Independent

One sample T test (or) Proportion test

One sample mean test → one source of data
Two sample mean test → two source of data

One sided check means one Tailed sample

Formula for One sample T test

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

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

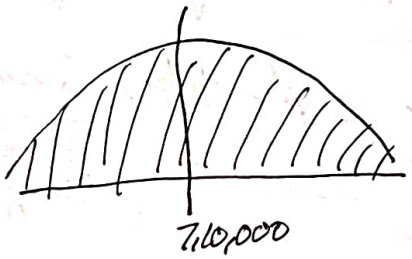
Two Sample T Test formula

$n = 500$
$\bar{x} = 7,10,000$

Lets assume $\mu = 7,10,000$

Both are same

$$t_{stat} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = 0$$



Pvalue = 100%. here Null hypothesis will be true above

- whenever $P < 5\%$ ∴ we reject Ho
- when $P > 5\%$ ∴ Ho becomes true

→ when we reject Ho, it implies our sample doesn't represent the population.

	To reject Ho: tstat	Pvalue (area)
95%	>1.96	<0.05
99%	>2.58	<0.01
90%	>1.64	<0.1

$H_a > $ — (Right tailed)

$H_a < $ — (Left tailed)

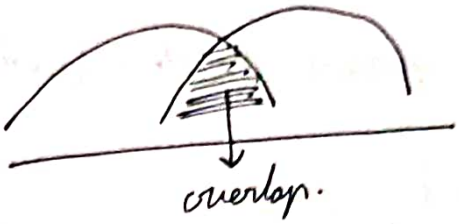
- Two tail problem statements in One Sample T test actually have equal or not equal.
- Left tail (or) right tail generally have less than or greater than in the problem statement.

Pg 247 (Problem statement)

Two Sample I Test:

- Two sample groups are taken (different groups) in the problem statement.
- If H_a must be true, two means must be 1.96 SD's away.
- Overlap happens, then p value increases.

Diagramatic representation of overlap:-



- Two sample t test
  - Independent (unpaired) — ttest_ind(g1, g2)
  - Dependent (paired) — ttest_rel(g1, g2)

g1 and g2 are two group of arrays.
rel means relative.

- When the population standard deviation is not known to us, then we use T-test.
- If it is known, we go for z test.
- Whenever we are calculating mean for two samples (or) groups, we call it Two Sample 't' test (or) Two sample mean test.
- ttest and z test are same for $n = 500$.
- Considering one sample,

$$t\text{test} \Rightarrow \frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad z\text{test} \Rightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- According to statistics, when $n \geq 30$, $s \simeq \sigma$ and tstat = Zstat.

- $\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \Leftarrow$ tstat

  - where n_1 and n_2 are the sample sizes
  - $\bar{x}_1$ and $\bar{x}_2$ are the sample means.
  - s_1^2 and s_2^2 are sample variances

· Pg(247) Problem statement

Speedy delivery is the effect we wanted to study.

$\mu_{dTime} < 3$ hrs → Ha (speedy delivery is the effect we wanted to study)

$\mu_{dtime} \geq 3$ hrs → Ho (we don't want to study late delivery)

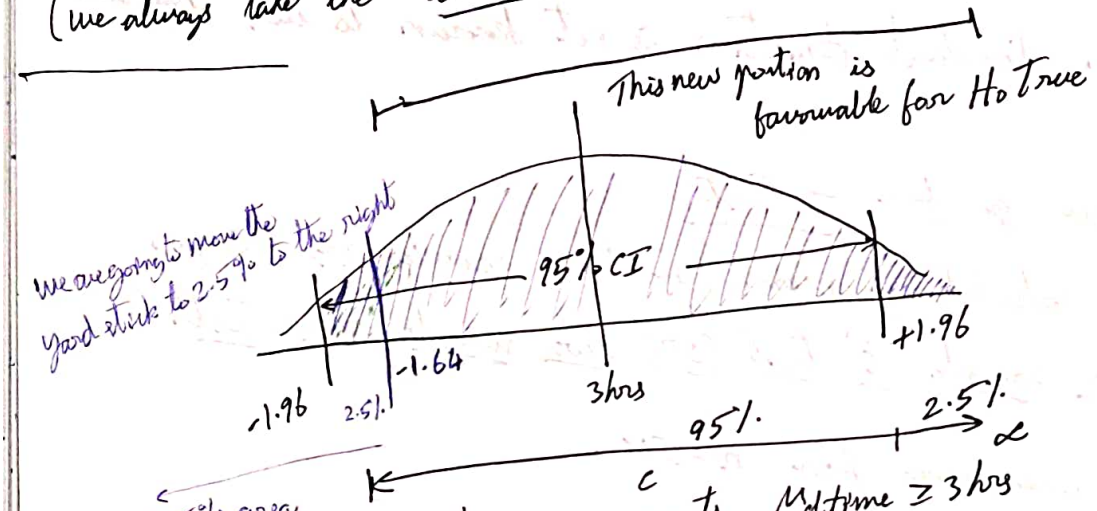
(another example)

Pharmaceutical company is claiming their medicine works within 10min.

$\mu_{Time} < 10$ min → Ha (we want to study this less time effect)

$\mu_{Time} > 10$ min → Ho

(we always take the contradiction statement as Ho)



· In the jupyter sir has created a bad score in the last. He has taken only 50 samples

---

· How to decide Ha and Ho

Ha is alternate hypothesis

Ho is null hypothesis

· Our research question is always on alternate hypothesis whereas the null hypothesis is the status quo.

· Null hypothesis is the -ve version of our research question.

· Alternate hypothesis is what we are trying to question.

One sample mean test

It has three different flavours,

- Two tail, left tail and right tail.

Two Tailed one sample I test

. Company claims they fill cans with 300ml. we need to verify whether the expected population mean is 300ml.

$\mu = 300ml \Rightarrow H_0$ } Equal to is always given to H_0

$\mu \neq 300ml \Rightarrow H_a$

↳ they can ① overfill (or) ② underfill (these are effects)

Left tail and Right tail

. In these problems we use greater than or less than ($<$ or $>$).

. Let us take the delivery time of a courier company.

Delivery company claims $\mu < 3$ hrs.

$\mu_{dtime} < 3$ hrs $\rightarrow H_a$ → we want to study the effect hypothesis

$\mu_{dtime} \geq 3$ hrs $\rightarrow H_0$ → we don't want to study late delivery

. We always take the contradiction statement as H_0 which is the negative version of our research question.

. Let us consider a college is claiming average marks above 98.5.

$\mu_{avg} > 98.5 \rightarrow H_a$

$\mu_{avg} \leq 98.5 \rightarrow H_0$

. Pharmaceutical example (previous page)

. In the left tail and right tail, for the H_a statements,

If $>$ (greater than) inside H_a statement then (Right Tailed)

If $<$ (less than) inside H_a statement then (Left Tailed).

(we can remember which side we keep the 5%.)

. This is how we figure out H_a and H_0 in One sample mean test.

- Only one sample t test (or) one sample proportion test (or) one sample mean test has three flavours (two tail, left tail and right tail.
- Going further we will be doing only two tail.
- randn will generate values in standard normal scale.

Complete till session 4.

KRISH NAIK STATISTICS PLAYLIST (statistics for machine learning)

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of the data.

- Descriptive statistics

→ Analyzing data, summarizing data, organizing data in the form of number of graphs.
→ BAR PLOT, HISTOGRAM, Piechart, PDF, CDF, Normal distribution.
→ Measure of Central Tendency. (mean, median, mode)
→ Variance, standard deviation.

- Inferential statistics

Kind of predictive analysis

- The gaussian/normal distribution can be converted into standard normal distribution by using standard scalar (z score).
- If we have a column with a few values and we already know that these values are forming a log normal distribution. We need to convert these values into standard normal distribution.

Convert the values into log → scale them using standard scalar for standard normal distribution.

- $cov(x,x) = var(x)$ (Covariance) (Variance)
- Covariance helps us to quantify the relationship between features, between random variable in a particular data set.

Central Limit Theorem :-

It specifies that, if the sample size is $n \geq 30$ selected randomly. Let us consider this sample as S_1. Similarly S_2 has 30 (or) more sample. similarly we take till S_{100} (Samples).

We calculate all the 100 means. If we take all the means and try to plot them, it will follow normal distribution and the sample mean will be approx. equal to population mean.

$$\bar{X} \approx G.D\left(\mu, \frac{\sigma^2}{n}\right)$$

Chebyshev's Inequality

- When a random variable x belongs to Gaussian distribution, we can tell the percentage of the distribution. like 68%, 95% and 99.7%.
- When a random variable x doesn't belong to gaussian distribution, to tell the distribution percentage, Chebyshev's Inequality is used.

$$Pr(\mu - \sigma < x < \mu + \sigma) \geq 1 - \frac{1}{k^2}$$

$k \rightarrow$ the number of standard deviations

$$Pr(\mu - k\sigma < x < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

If K is 2, then,

$$Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \geq 1 - \frac{1}{4}$$

$$\geq \frac{3}{4} \text{ (which is 75\%)}$$

So, if the Y doesn't follow any Gaussian distribution, then 75% of the data points belonging to Y random variable will be falling within the range of second standard deviation.

Pearson Correlation Coefficient

This concept will also be used for feature selection process.

- Covariance helps us to find the direction of the relationship.

Pearson Correlation Coefficient

$$\rho_{(x,y)} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

- In covariance we don't know how much positive or negative the variation is.
- Pearson Correlation Coefficient states that, based on the variance of x and y, it will be able to tell that strength (how strong it is correlated) and Direction of the relationship.
- The value ranges from −1 to +1. $(-1 \leq \rho \leq 1)$

Spearman's rank Correlation Coefficient

It is some what better than Pearson Correlation Coefficient.

- Instead of x and y, we use rank of x and rank of y.

$$\Rightarrow \frac{cov(rg_x, rg_y)}{\sigma_{rg_x}\, \sigma_{rg_y}}$$

Even if we have non-linear data points, we can use this method.

$$r_s = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

Outlier:- An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the data set.

- Reasons for an outlier to exist are:-
  - → Variability in the data
  - → An experimental measurement error.

• Impacts of having outliers in a data set :-

→ It causes various problems during our statistical analysis.

→ It may cause a significant impact on the mean and the standard deviation.

Interview Questions :-

• Tell us a few examples about left skewed and right skewed distribution.

• What is the relationship b/w mean, median and mode in skewed dist.

Right skewed :- Wealth distribution, length of comment

Normal Distribution :- Age, weight, height.

Left skewed :- Life span of a human being.

Bernoulli Distribution

We only have two outcomes in Bernoulli distribution.

Example of a Coin :

$x = 1$ (Head) $\Rightarrow$ Success

$x = 0$ (Tail) $\rightarrow$ Fail

$$P(X = x) = p^x (1-p)^{1-x}$$ (Probability mass function) (outcome is discrete in pmf)

• Probability density function (outcome is continuous value)

• x will usually have zero and one.

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

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

[x = 0 or x = 1] [P(x=0) = 0.4 = 1-p, P(x=1) = 0.6]

Mean $E(x) = \sum_{i=1}^{x} x \cdot P(x)$

$= 0(0.4) + 1 \times (0.6) = 0.6 \Rightarrow p$

The mean of the Bernoulli distribution is p.

Variance $= p(1-p) = pq$ and S.D is $\sqrt{pq}$

What is p-value :-

It is the probability for the "null hypothesis" to be true.

$p \leq 0.05$ → Significance value

Different Sampling Techniques :-

- Suppose we have to do an exit poll, we collect the data randomly from random places. This is Random Sampling Technique.
- Stratified Sampling Technique :-

What if the exit polls are wrong. The sampling might be wrong. May be when 1000 samples were taken, there were 700 men and 300 women. It is completely biased. We may not get proper output. 50:50 ratio is better.

- Systematic Sampling :-

We select the n^{th} person from every group and try to record the information.

- Cluster Sampling Technique :-

It is more domain related. If we are researching medical domain. The person we select randomly are familiar with this domain knowledge.

- Continuous Variable means mean test and others is proportion Test.
- Whenever we are building a prediction model to predict a numerical variable, that is called Regressor model.

Model is going to say Diabetic / Healthy which is a classification model.

Two Sample Z-Test for Difference in proportions

$$Z_{data} = \frac{P_1 - P_2}{\sqrt{P_{pooled}\,(1 - P_{pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where} \quad \boxed{P_{pooled} = \frac{x_1 + x_2}{n_1 + n_2}}$$

ANOVA :-

Consider $\bar{A}, \bar{B}, \bar{C}$ are means of ages.

$H_0 : \bar{A} = \bar{B} = \bar{C}$

If null hypothesis H_0 holds good (meaning all the means are overlapping), then age is an useless feature.

If there is a significant difference of mean in the feature, then we can consider it.

H_a should hold good for the feature to be considered.

For a strong feature, MSTR must be very high

- magnitude and direction
  - magnitude → number
  - direction → - or + (positive or negative)
- Covariance and Coefficient of Correlation
  - Covariance → $-\infty$ to $+\infty$
  - Coefficient of Correlation → -1 to $+1$

~~Pearson Correlation~~

Covariance :-

$$\boxed{Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_y)}$$

$$\boxed{Cov(x, x) = Var(x)}$$

Covariance helps us to quantify the relationship between features, between random variable in a particular data set.

ANOVA (Session 6) points :-

- Whenever we reject the null hypothesis, whatever the test it might be, we have some interesting story to tell from Ha. This story is known as post hoc analysis.
- in mod = ols ('Income ~ Age-Group', data = data). fit()
  (First parameter must be numeric variable)
- Tukey Test will do the pair wise analysis.
  It performs a pair wise two sample T test.

- Let us look at the example of a confusion matrix :-

ACTUAL	PREDICTION: Healthy (-ve)	PREDICTION: Covid 19 (+ve)
Healthy (-ve)	True Negative	False positive → Type I error (or) α error
Covid 19 (+ve)	False Negative → Type II error (or) β error	True Positive

- Predicting Healthy as Healthy is a true negative
- Predicting positive case as a positive is true positive
- Type II error is the costly error in the above example.