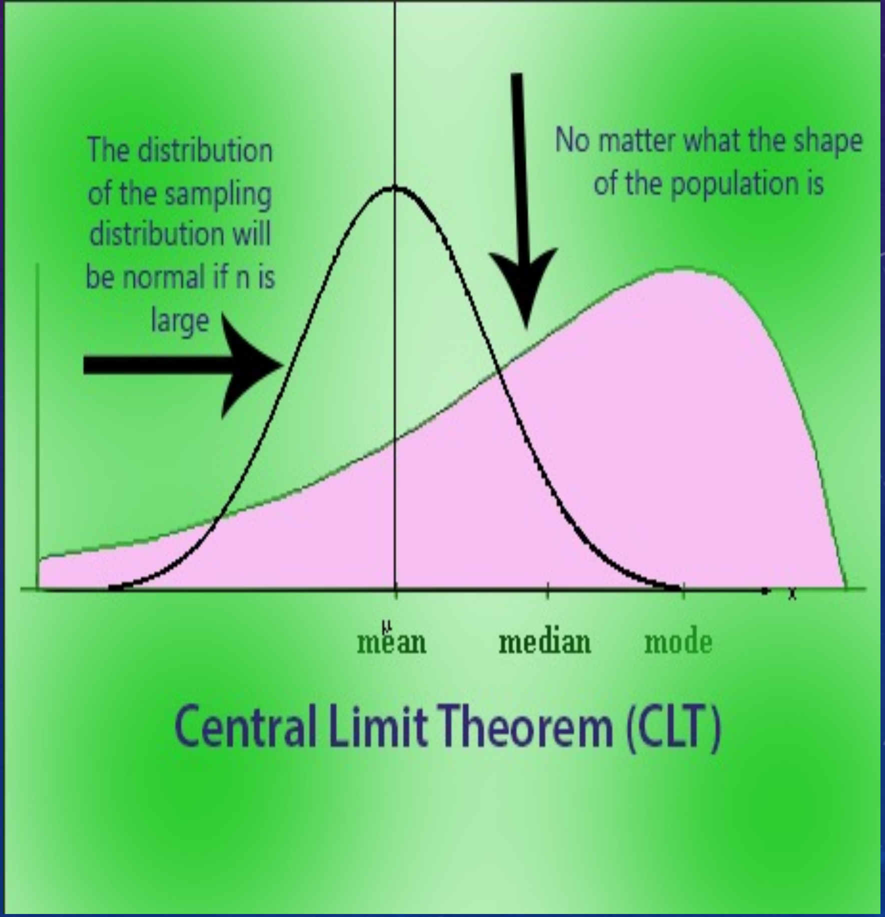
STATISTICS

INTERVIEW PREP

1. What is the Central Limit Theorem?

CENTRAL LIMIT THEOREM

- It states that the distribution of the mean of any independent, random variable can be approximated to normal if the sample size is large enough. Generally, the sample size of above 30 or sometimes 40 is taken as reference.
- This allows us to approximate bigger samples to normal distribution without having to take hundreds or thousands of distributions. Standard Normal distribution is preferred as such because mean is equal to zero and variance is one.
- (we sample from a population using a sufficiently large sample size, the mean of the samples (also known as the sample population) will be normally distributed (assuming true random sampling), the mean tending to the mean of the population and variance equal to the variance of the population divided by the size of the sampling. What's especially important is that this will be true regardless of the distribution of the original population.)


The distribution of the sampling distribution will be normal if n is large
No matter what the shape of the population is
x
mean
median
mode
Central Limit Theorem (CLT)

2. What is Sampling?

SAMPLING

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.

3. How many Sampling Methods do you know?

PROBABILITY DATA SAMPLING METHODS

- **Simple random sampling**: Software is used to randomly select subjects from the whole population.
- **Stratified sampling**: Subsets of the data sets or population are created based on a common factor, and samples are randomly collected from each subgroup.
- **Cluster sampling**: The larger data set is divided into subsets (clusters) based on a defined factor,
- then a random sampling of clusters is analyzed. For example, a researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.
- **Multistage sampling**: A more complicated form of cluster sampling, this method also involves dividing the larger population into a number of clusters. Second-stage clusters are then broken out based on a secondary factor, and those clusters are then sampled and analyzed.
- **Systematic sampling**: A sample is created by setting an interval at which to extract data from the larger population – for example, selecting every 10th row in a spreadsheet of 200 items to create a sample size of 20 rows to analyze.

NON-PROBABILITY DATA SAMPLING METHODS:

- **Convenience sampling**: Data is collected from an easily accessible and available group.
- **Consecutive sampling**: Data is collected from every subject that meets the criteria until the predetermined sample size is met.
- **Purposive or judgmental sampling**: The researcher selects the data to sample based on predefined criteria.
- **Quota sampling**: The researcher ensures equal representation within the sample for all subgroups in the data set or population (random sampling is not used).

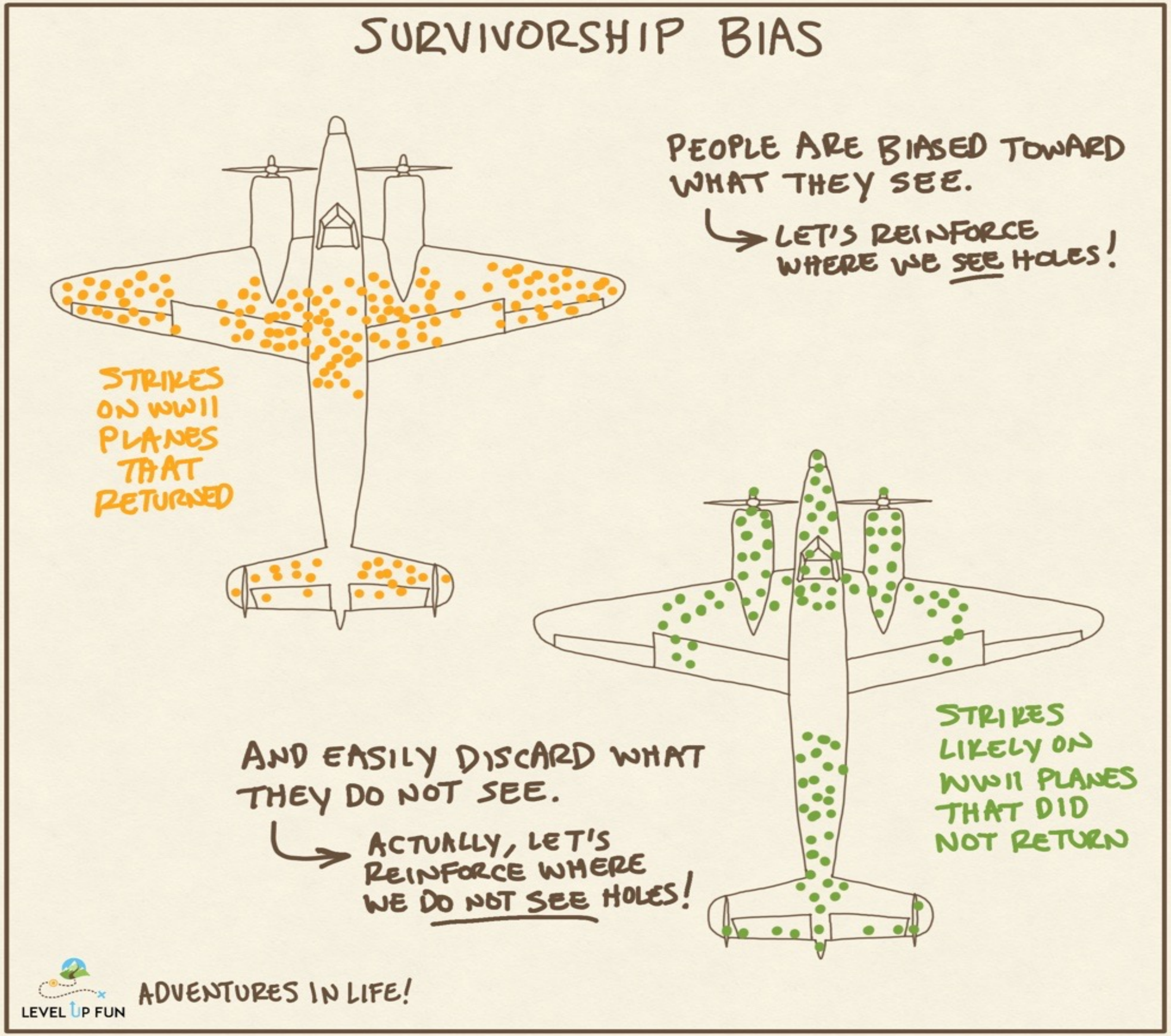
4. What Types Of Biases That Can Occur During Sampling?

The Types Of Biases That Can Occur During Sampling

- **Selection (or ‘sampling’) bias**: occurs when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see.
- **Undercoverage bias**
- **Survivorship bias**: is the logical error of focusing on aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.

5. What Is Survivorship Bias ?

- **Survivorship bias**: is the logical error of focusing on aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.


SURVIVORSHIP BIAS
PEOPLE ARE BIASED TOWARD WHAT THEY SEE.
LET'S REINFORCE WHERE WE SEE HOLES!
STRIKES ON WWII PLANES THAT RETURNED
AND EASILY DISCARD WHAT THEY DO NOT SEE.
ACTUALLY, LET'S REINFORCE WHERE WE DO NOT SEE HOLES!
STRIKES LIKELY ON WWII PLANES THAT DID NOT RETURN
LEVEL UP FUN
ADVENTURES IN LIFE!

6. What Is Re-sampling & Why We Need It?

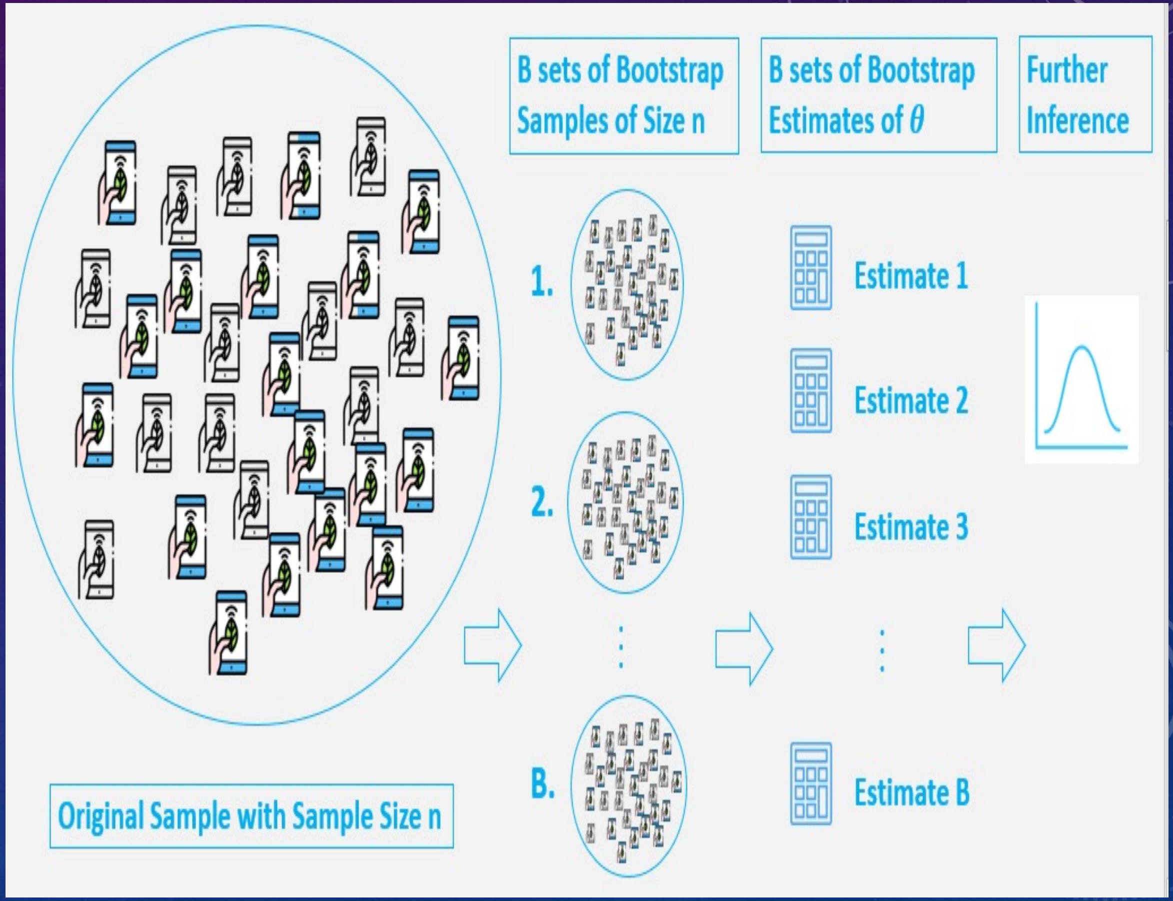
RE-SAMPLING:

- Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter. Resampling methods, in fact, make use of a nested resampling method.
- Two commonly used resampling methods that you may encounter are **k-fold cross-validation** and the **bootstrap**.

7. What Is Bootstrap & Any Usage of Bootstrap in ML?

BOOTSTRAP

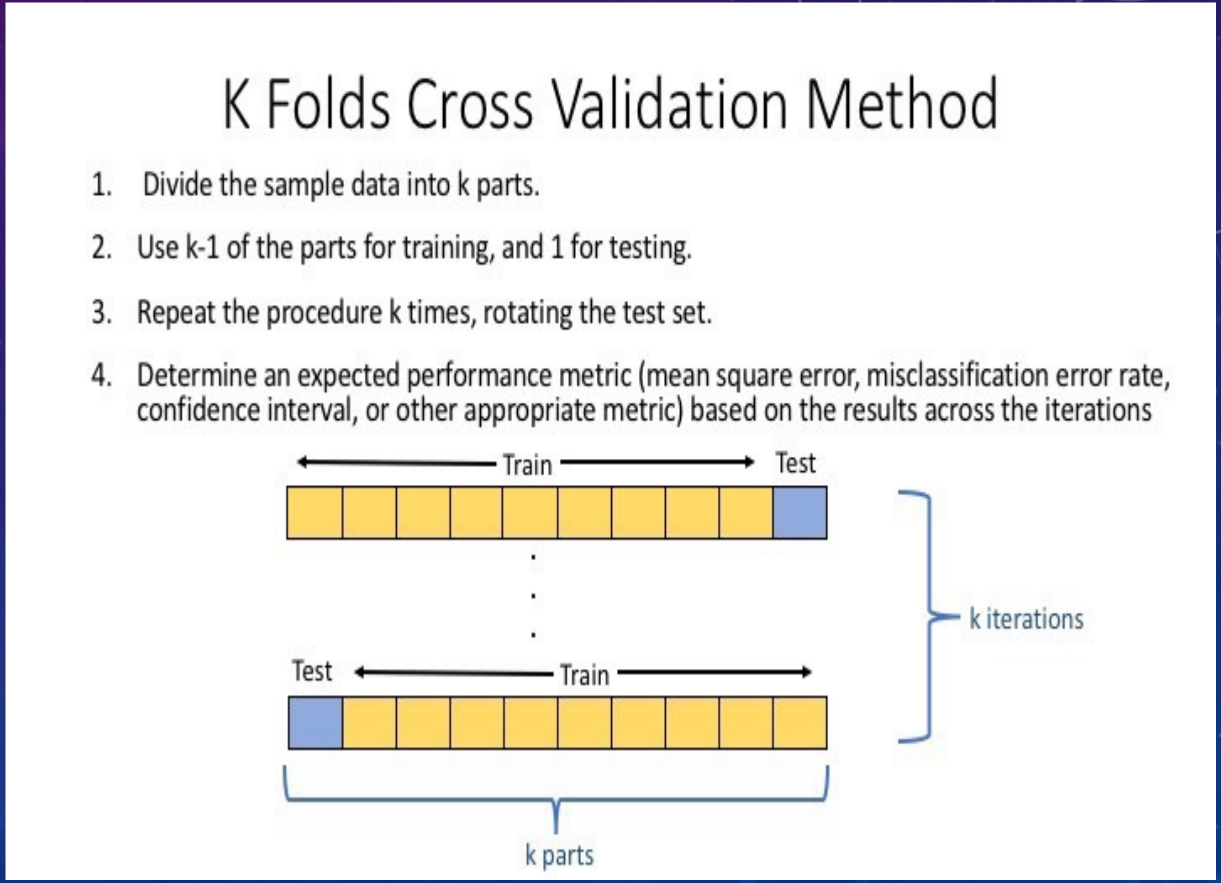
- Samples are drawn from the dataset with replacement (allowing the same sample to appear more than once in the sample), where those instances not drawn into the data sample may be used for the test set.


B sets of Bootstrap Samples of Size n
B sets of Bootstrap Estimates of θ
Further Inference
1.
2.
B.
Estimate 1
Estimate 2
Estimate 3
Estimate B
Original Sample with Sample Size n

8. What Is K-fold Cross-Validation?

K-FOLD CROSS-VALIDATION

- A dataset is partitioned into k groups, where each group is given the opportunity of being used as a held out test set leaving the remaining groups as the training set.
- The k-fold cross-validation method specifically lends itself to use in the evaluation of predictive models that are repeatedly trained on one subset of the data and evaluated on a second held-out subset of the data.
- There is an alternative in Scikit-Learn called **Stratified k fold**, in which the split is shuffled to make it sure you have a representative sample of each class and a k fold in which you may not have the assurance of it (not good with a very unbalanced dataset).



K Folds Cross Validation Method

1. Divide the sample data into k parts.
2. Use k-1 of the parts for training, and 1 for testing.
3. Repeat the procedure k times, rotating the test set.
4. Determine an expected performance metric (mean square error, misclassification error rate, confidence interval, or other appropriate metric) based on the results across the iterations

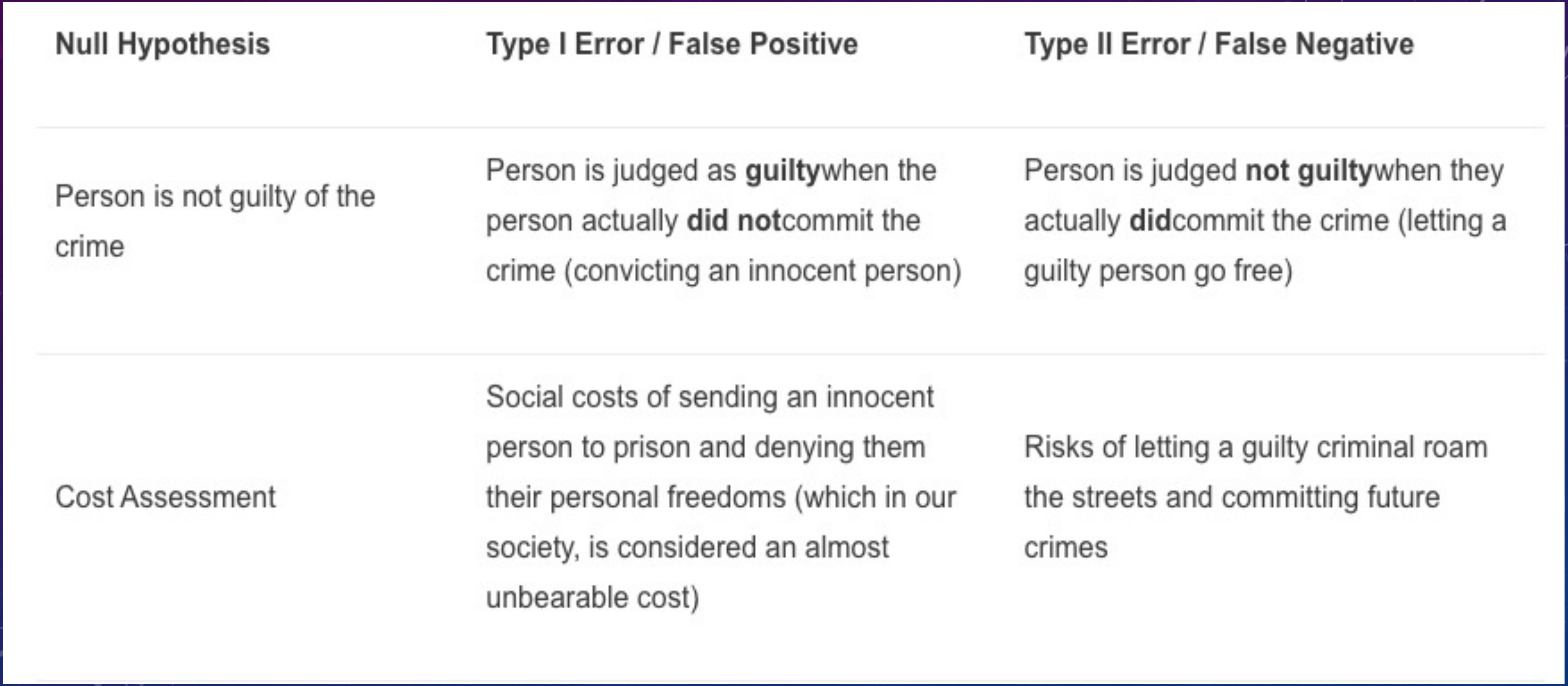
9. Differentiate Between Univariate, Bivariate, And Multivariate Analysis

- **Univariate**: Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.
- **Bivariate**: Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.
- **Multivariate**: Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate but contains more than one dependent variable.

10. What Is The Difference Between Type I Vs Type II Error?

- In statistical test theory, the notion of statistical error is an integral part of hypothesis testing. The statistical test requires an unambiguous statement of a **null hypothesis** (H_0), for example, “this person is healthy”, “this accused person is not guilty” or “this product is not broken”. The result of the test of the null hypothesis may be **positive**(healthy, not guilty, not broken) or may be **negative**(not healthy, guilty, broken).

- If the result of the test corresponds with reality, then a correct decision has been made (e.g., person is healthy and is tested as healthy, or the person is not healthy and is tested as not healthy). *However, if the result of the test does not correspond with reality, then two types of error are distinguished:* ***type I error****and* ***type II error***.



Null Hypothesis	Type I Error / False Positive	Type II Error / False Negative
Person is not guilty of the crime	Person is judged as **guilty**when the person actually **did not**commit the crime (convicting an innocent person)	Person is judged **not guilty**when they actually **did**commit the crime (letting a guilty person go free)
Cost Assessment	Social costs of sending an innocent person to prison and denying them their personal freedoms (which in our society, is considered an almost unbearable cost)	Risks of letting a guilty criminal roam the streets and committing future crimes

11. What Is P-value?

P-VALUE

- P-Value measures the possibility of getting outcomes equal to or greater than those obtained under a certain hypothesis, provided the null hypothesis is true (It measures the strength of evidence in support of null hypotheses). This indicates the likelihood that the observed discrepancy happened by coincidence.
- The lower the p-value, the more important is the variable in predicting the target variable. Usually we set a 5% level, so that we have a 95% confidentiality that our variable is relevant.
- The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

12. What Is Confidence Intervals ?

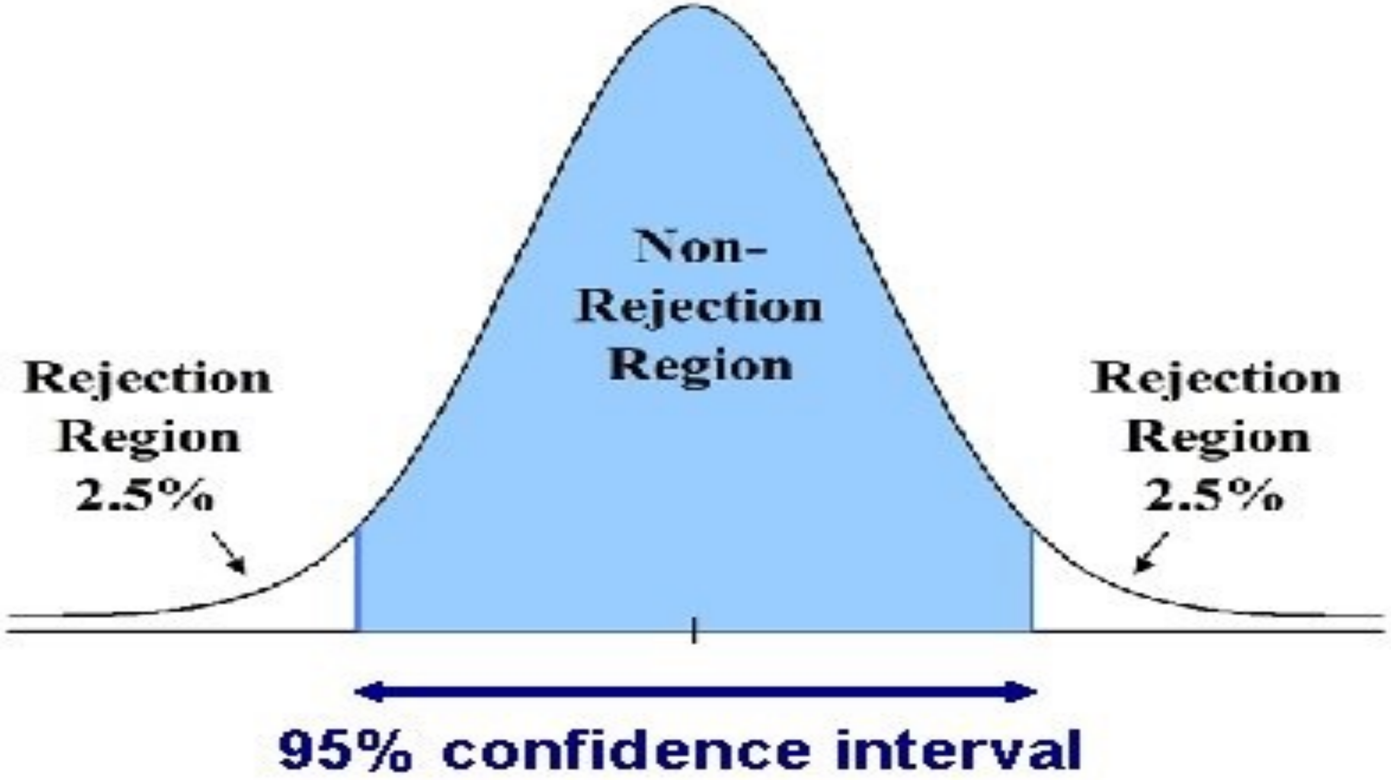
CONFIDENCE INTERVALS & ALPHA

- **Confidence Interval** indicates the boundaries in which the mean value will fall.
- It is a range of scores constructed, such that the population mean will fall within.
- They are limits constructed such that for a certain percentage of the time the true value of the population mean will fall within these range.
- **Alpha** is defined as 1-confidence interval. This implies probability that the true value remains outside the confidence interval. If confidence interval is 99% then alpha is 1-99% which is 0.01.

P Values vs. Confidence Intervals

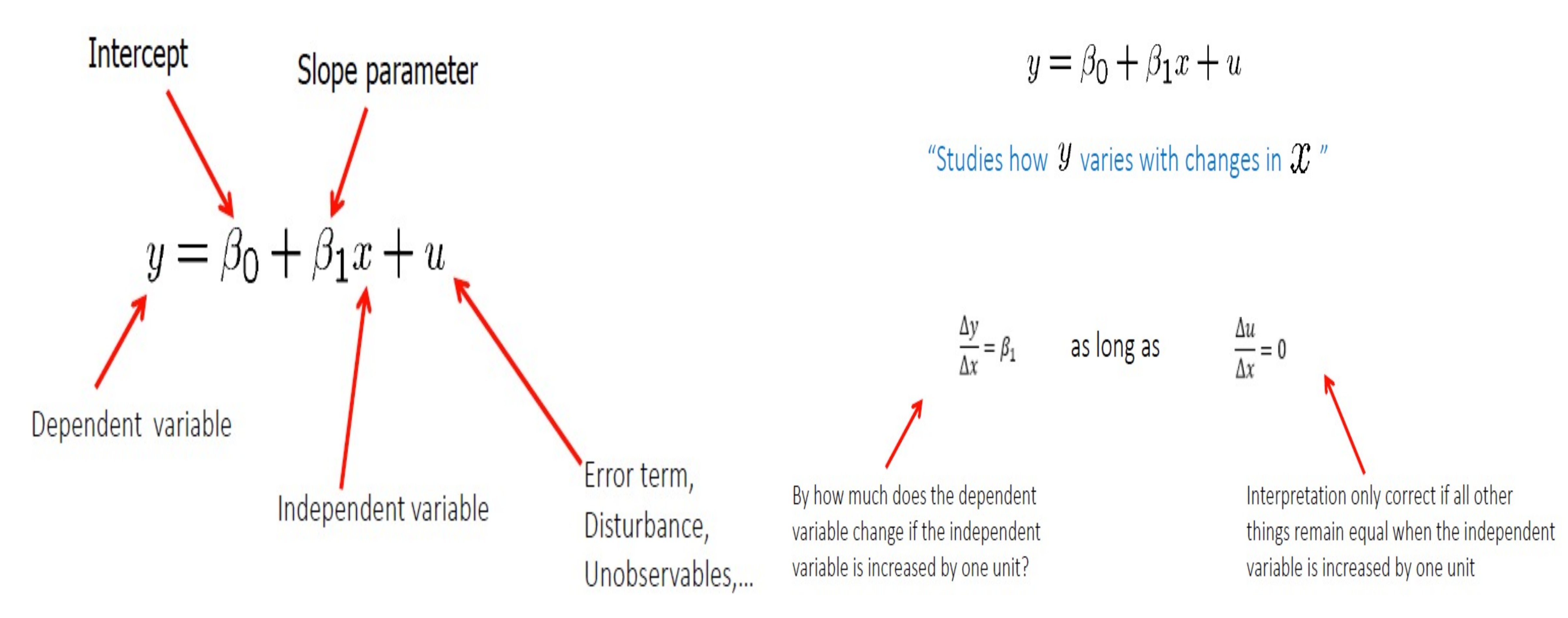
- **There is a direct relationship between levels of alpha set for a statistical test and the level set for constructing a confidence interval.**

For example, alpha = 0.05 for a 2-sided statistical test is equivalent to a 95% confidence interval



13. What Is Linear Regression?

THE SIMPLE REGRESSION MODEL



Intercept

Slope parameter

$$y = \beta_0 + \beta_1 x + u$$

Dependent variable

Independent variable

Error term,
Disturbance,
Unobservables,...

$$y = \beta_0 + \beta_1 x + u$$

"Studies how y varies with changes in x "

$\frac{\Delta y}{\Delta x} = \beta_1$ as long as $\frac{\Delta u}{\Delta x} = 0$

By how much does the dependent variable change if the independent variable is increased by one unit?

Interpretation only correct if all other things remain equal when the independent variable is increased by one unit

14. WHAT ARE THE ASSUMPTIONS REQUIRED FOR LINEAR REGRESSION?

THE ASSUMPTIONS REQUIRED FOR LINEAR REGRESSION

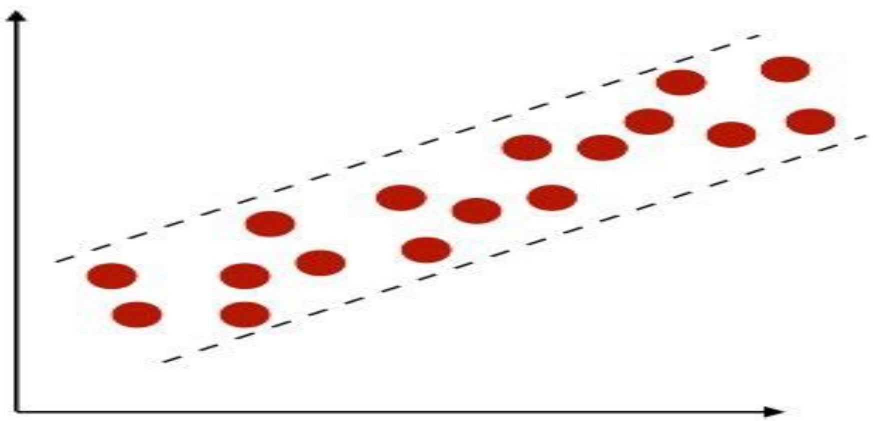
- There are four major assumptions:

1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data,
2. **Normality**: For any fixed value of X, Y is normally distributed,
3. There is minimal multicollinearity between explanatory variables, and
4. **Homoscedasticity**. This means the variance around the regression line is the same for all values of the predictor variable. The variance of residual is the same for any value of X.

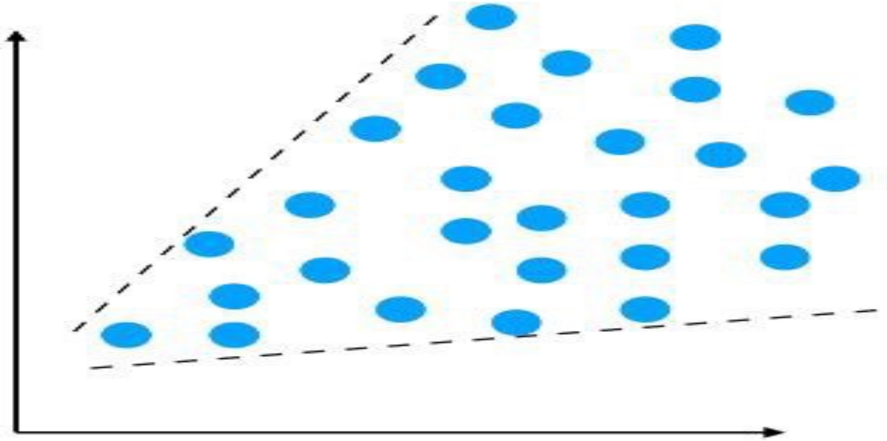
15. What Is Heteroscedasticity?

HETEROSCEDASTICITY:

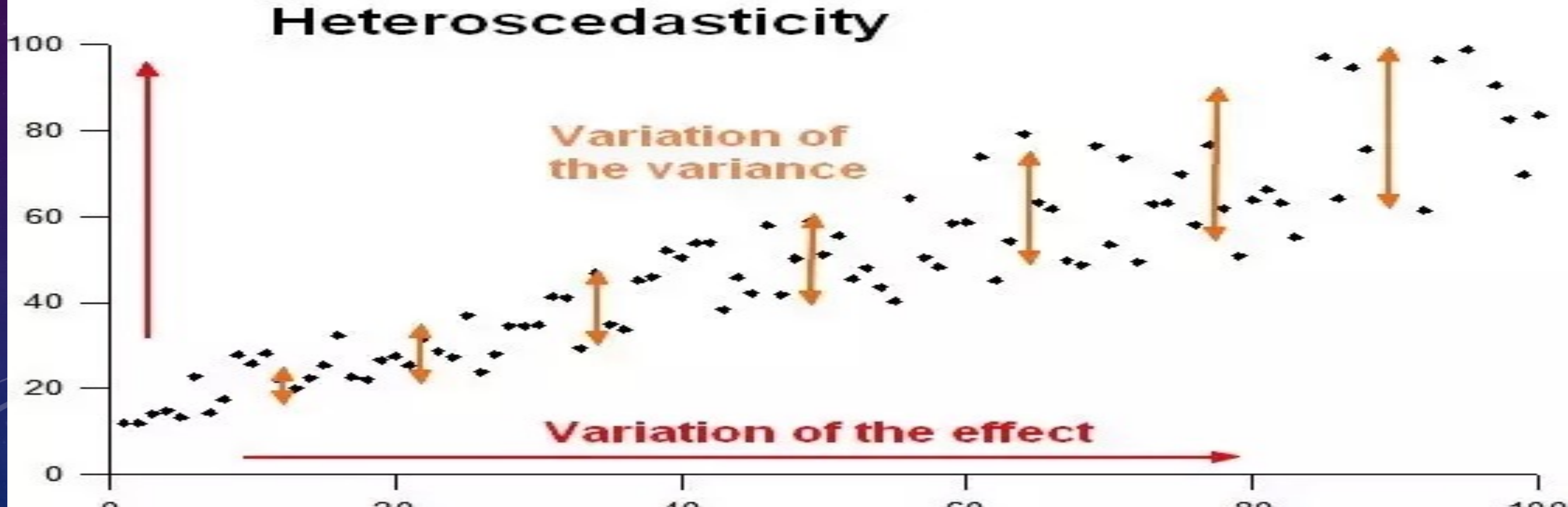
- Heteroscedasticity means unequal scattered distribution. In regression analysis, we generally talk about the heteroscedasticity in the context of the error term. Heteroscedasticity is the systematic change in the spread of the residuals or errors over the range of measured values. Heteroscedasticity is the problem because Ordinary least squares (OLS) regression assumes that all residuals are drawn from a random population that has a constant variance.
- There are many reasons why heteroscedasticity can exist, and a generic explanation is that the error variance changes proportionally with a factor.



Homoscedasticity



Heteroscedasticity



16. What Is Measurement Error ?

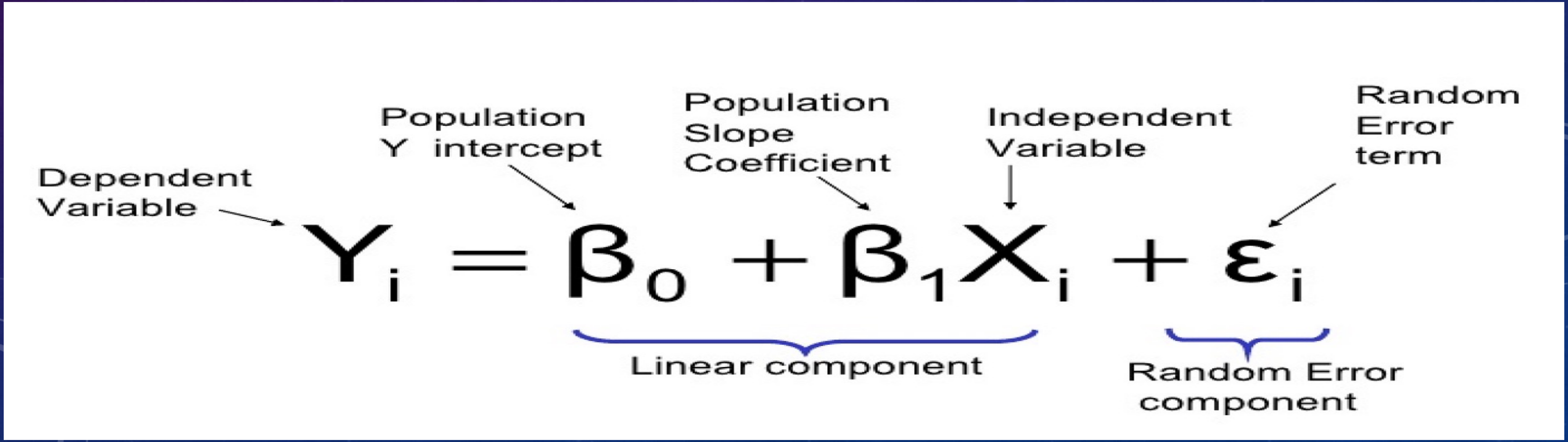
MEASUREMENT ERROR:

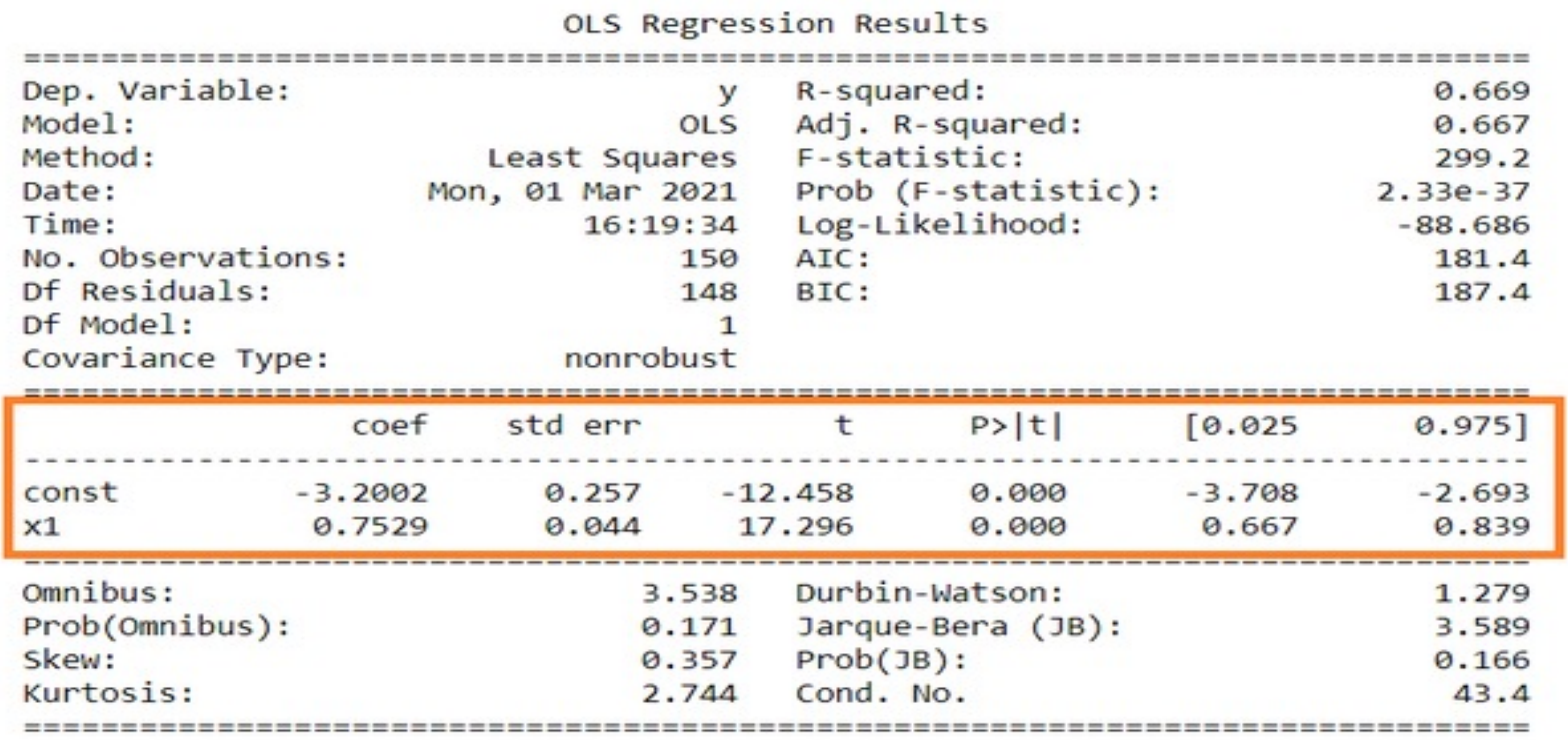
- The discrepancy between the measured value and actual value in terms of number is called measurement error.

17. What Is Coefficient Value ?

COEFFICIENT VALUE

- The **coefficient value** signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. This property of holding the other variables constant is crucial because it allows you to assess the effect of each variable in isolation from the others



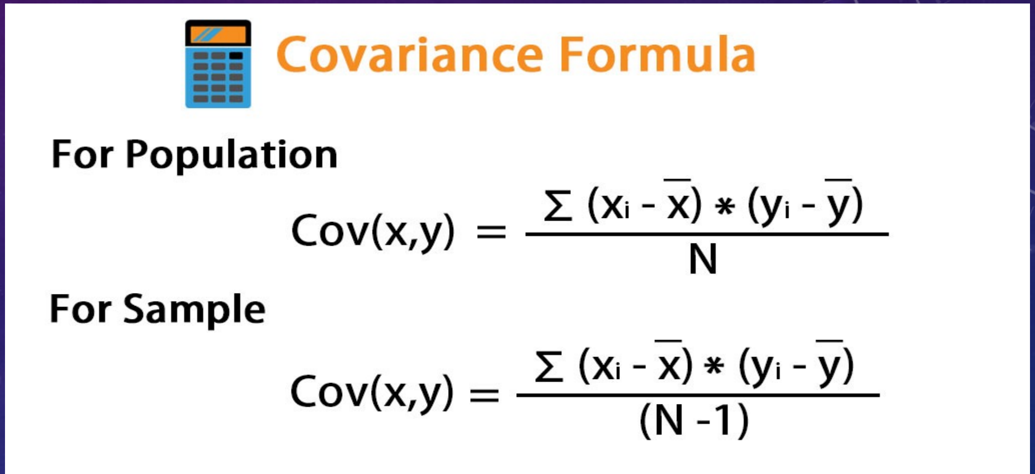


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.669
Model:                            OLS   Adj. R-squared:                  0.667
Method:                 Least Squares   F-statistic:                     299.2
Date:                Mon, 01 Mar 2021   Prob (F-statistic):           2.33e-37
Time:                        16:19:34   Log-Likelihood:                -88.686
No. Observations:                 150   AIC:                             181.4
Df Residuals:                     148   BIC:                             187.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.2002      0.257    -12.458      0.000      -3.708      -2.693
x1             0.7529      0.044     17.296      0.000       0.667       0.839
==============================================================================
Omnibus:                        3.538   Durbin-Watson:                   1.279
Prob(Omnibus):                  0.171   Jarque-Bera (JB):                3.589
Skew:                           0.357   Prob(JB):                        0.166
Kurtosis:                       2.744   Cond. No.                         43.4
==============================================================================
```

18. What Is Covariance & Correlation ?

COVARIANCE AND CORRELATION

- Both Correlation and the Covariance establish the relationship and also measures the dependency between the two random variables.
- **Correlation**: It is the statistical technique that can show whether and how strongly pairs of variables are related.
- **Covariance**: It measures the directional relationship between the returns on two assets. The positive covariance means that asset returns move together while a negative covariance means they move inversely. Covariance is calculated by analyzing at-return surprises (standard deviations from the expected return) or by multiplying the correlation between the two variables by the standard deviation of each variable.



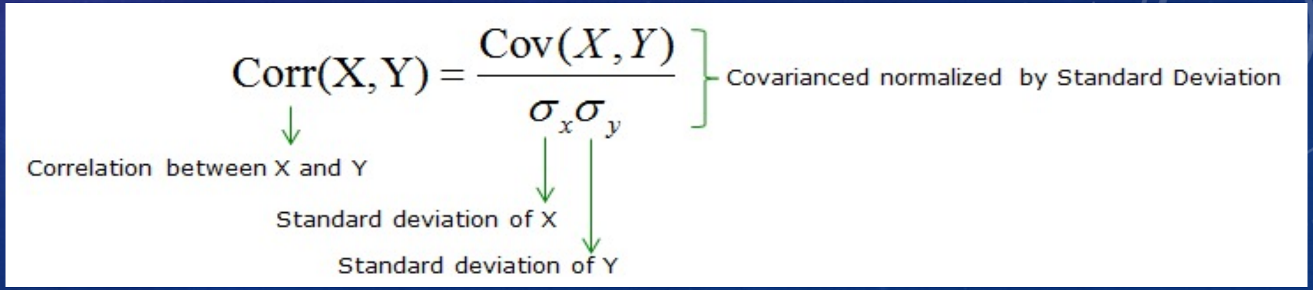
Covariance Formula

For Population

$$\mathrm{Cov}(x,y) = \frac{\Sigma\,(x_i - \overline{x}) * (y_i - \overline{y})}{N}$$

For Sample

$$\mathrm{Cov}(x,y) = \frac{\Sigma\,(x_i - \overline{x}) * (y_i - \overline{y})}{(N - 1)}$$



$$\underset{\downarrow}{\mathrm{Corr}(X,Y)} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Covarianced normalized by Standard Deviation

Correlation between X and Y

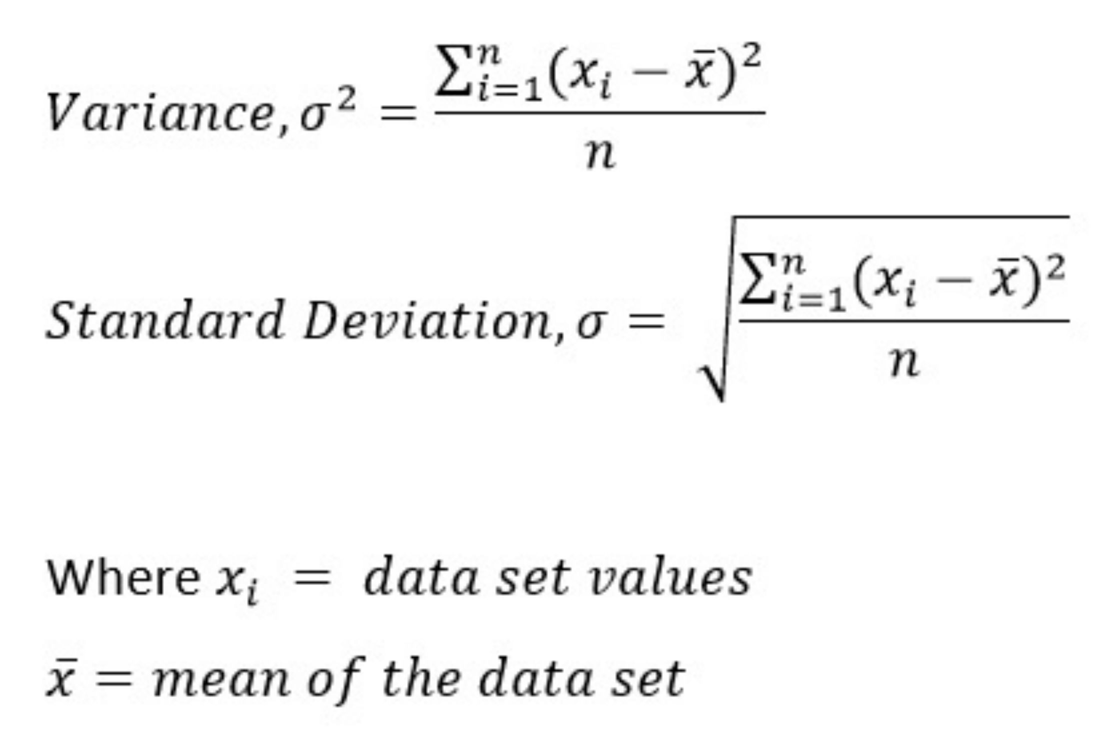
Standard deviation of X

Standard deviation of Y

19. What Is Variance & Standard Deviation ?

VARIANCE & STANDARD DEVIATION

- **Variance** is the average error between the mean and the measured values. It indicates the difference between the average value calculated and the observed value as such. It is an indication of how different individuals in group differ or vary from each other.
- Square root of the variance is also called **standard deviation**. This is done to keep the measurement same as original one. They indicate the nearness of the points measured w.r.t. mean. Smaller the standard deviation will be nearer to mean and vice versa.



$$Variance, \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$Standard\ Deviation, \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Where $x_i \;=\; data\ set\ values$

$\bar{x} = mean\ of\ the\ data\ set$

20. What Is Standard Error ?

STANDARD ERROR

- Standard error indicates how well a sample represents the original population. When we break the original population into various small samples, we would like to know the difference between the sample considered and original population. This is represented by standard error.
- The smaller the standard error, the closer or true representation of original population.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Standard deviation

Number of samples

21. What Is R-Square & Why We Need Adjusted R-Square

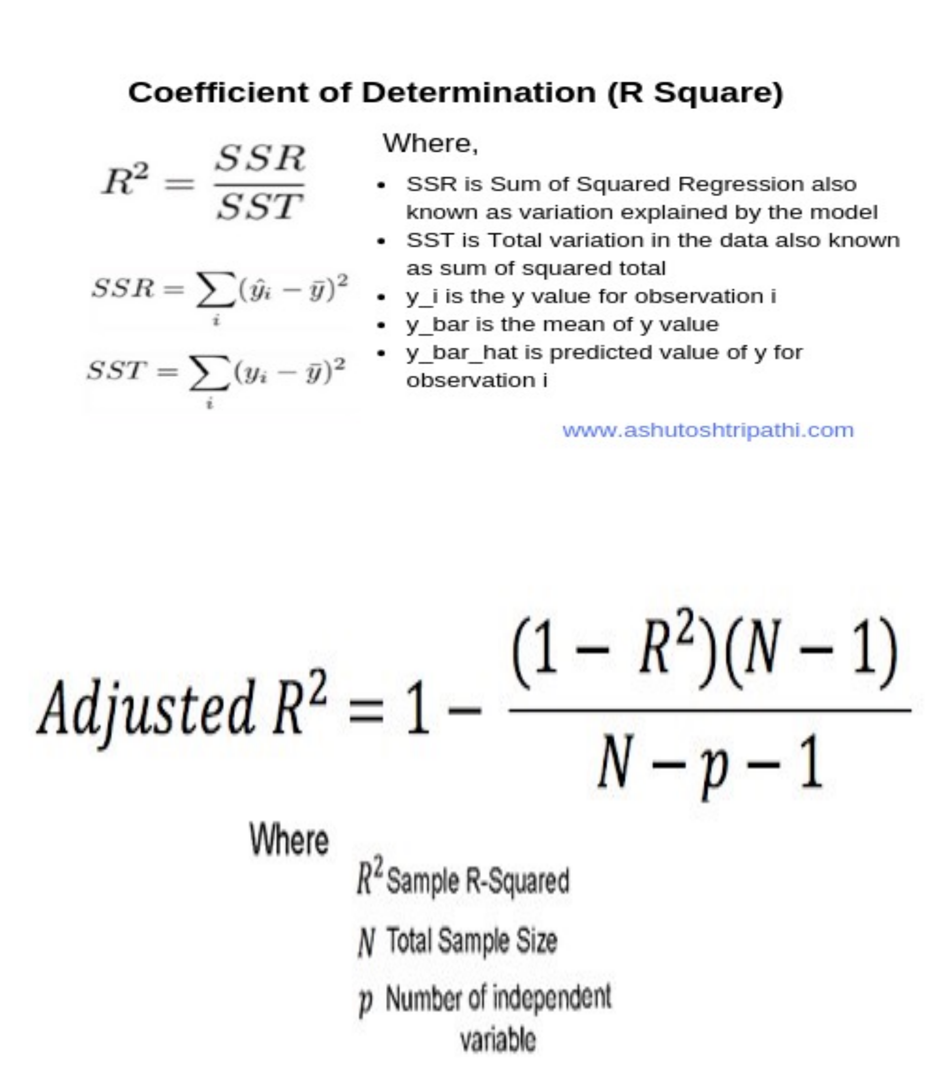
R-SQUARE & ADJUSTED R-SQUARE

- **R-squared** is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

- The definition of R-squared is the percentage of the response variable variation that is explained by a linear model. **R-squared = Explained variation / Total variation**

- There is a problem with the R-Square. The problem arises when we ask this question to ourselves.** Is it good to help as many independent variables as possible?**

- R-Squared value will always increase. It will never decrease with the addition of a newly independent variable, whether it could be an impactful, non-impactful, or bad variable, so we need another way to measure equivalent R Square, which penalizes our model with any junk independent variable.



Coefficient of Determination (R Square)

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

Where,

- SSR is Sum of Squared Regression also known as variation explained by the model
- SST is Total variation in the data also known as sum of squared total
- y_i is the y value for observation i
- y_bar is the mean of y value
- y_bar_hat is predicted value of y for observation i

www.ashutoshtripathi.com

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

R^2 Sample R-Squared

N Total Sample Size

p Number of independent variable

22. What Is The Difference Between Bernoulli And Binomial Distributions

BERNOULLI AND BINOMIAL DISTRIBUTIONS:

- The Bernoulli distribution simulates one trial of an experiment with just two possible outcomes, whereas the binomial distribution simulates n trials.

23. What Is The Difference Between T-test & Z-test

T-TEST & Z-TEST

- **t-test** : A t-test is the type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances.

- **z-test:** It is a statistical test used to determine whether the two population means are different when the variances are known, and the sample size is large. The test statistic is assumed to have the normal distribution, and nuisance parameters such as standard deviation should be known for an accurate z- test to be performed.

- **Z-tests are related to t-tests**, but t- tests are best performed when an experiment has the small sample size. Also, T-tests assumes the standard deviation is unknown, while z-tests assumes that it is known. If the standard deviation of the population is unknown, then the assumption of the sample variance equaling the population variance is made.

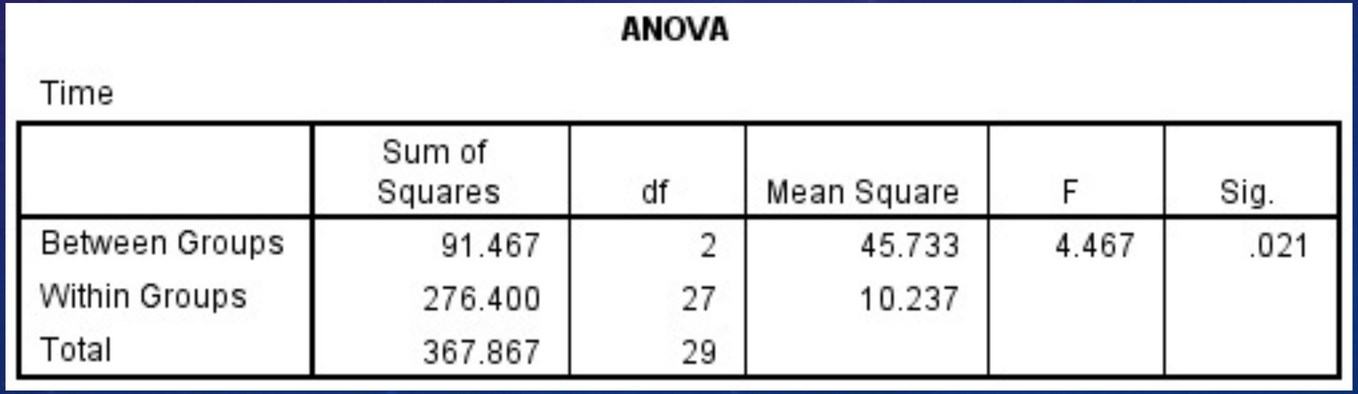
24. WHAT IS ANOVA & WHERE TO USE?

ANOVA

- It stands for “Analysis of Variance ”. It is a statistical method to compare the population means of two or more groups by analyzing variance. The variance would differ only when the means are significantly different.

- **One-way ANOVA** is the hypothesis test in which only one categorical variable or the single factor is taken into consideration. With the help of F-distribution, it enables us to compare means of three or more samples.

- **Two-ways ANOVA** examines the effect of two independent factors on a dependent variable. It also studies the inter-relationship between independent variables influencing the values of the dependent variable, if any.



ANOVA

Time

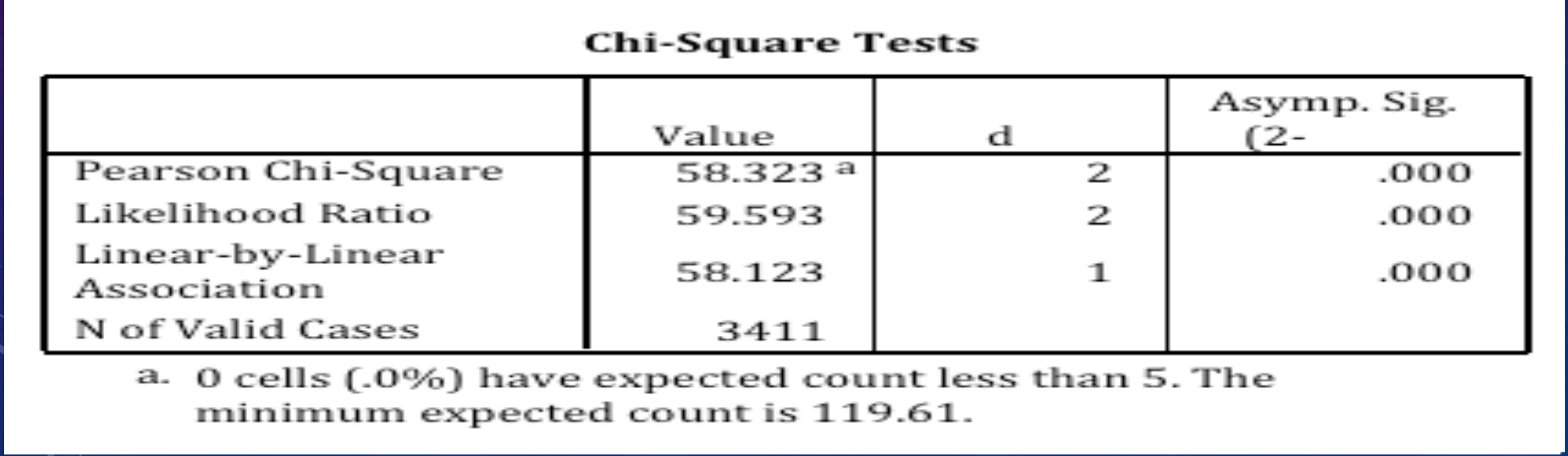
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	91.467	2	45.733	4.467	.021
Within Groups	276.400	27	10.237		
Total	367.867	29			

25. WHAT IS CHI-SQUARE & WHERE TO USE?

CHI-SQUARE STATISTIC:

- It is a test that measures how expectations compare to actual observed data (or model results).

- Chi-square test is intended to test how it is that an observed distribution is due to chance. It is also called the "goodness of fit" statistic because it measures how well the observed distribution of the data fits with the distribution that is expected if the variables are independent. Chi-square test is designed to analyze the categorical data.



Chi-Square Tests

	Value	d	Asymp. Sig. (2-
Pearson Chi-Square	58.323 [a]	2	.000
Likelihood Ratio	59.593	2	.000
Linear-by-Linear Association	58.123	1	.000
N of Valid Cases	3411		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 119.61.