STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
   Answer: (a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
   Answer: (a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
   Answer: (b) Modeling bounded count data

4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
   Answer : (d)All of the mentioned

5. _____ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer: (c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer : (b) False

7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
   Answer: (b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
   Answer: (a) 0

9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned
   Answer: (c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

Answer: Normal distribution

The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the "bell curve."

A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the *mean* of the distribution.
The normal distribution is also known as a *Gaussian distribution* or *probability bell curve*. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

Why Is the Normal Distribution Called "Normal?"

The normal distribution is technically known as the Gaussian distribution, however it took on the terminology "normal" following scientific publications in the 19th century showing that many natural phenomena appeared to "deviate normally" from the mean. This idea of "normal variability" was made popular as the "normal curve" by the naturalist Sir Francis Galton in his 1889 work, *Natural Inheritance.*

Normal Distribution Definition

The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

$f(x) \geq 0 \; \forall \; x \in (-\infty, +\infty)$

And $\int_{-\infty}^{+\infty} f(x) = 1$

Normal Distribution Formula

The probability density function of normal or gaussian distribution is given by;

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

Normal Distribution Curve

The random variables following the normal distribution are those whose values can find any unknown value in a given range. For example, finding the height of the students in the school. Here, the distribution can consider any value, but it will be bounded in the range say, 0 to 6ft. This limitation is forced physically in our query.

Whereas, the normal distribution doesn't even bother about the range. The range can also extend to $-\infty$ to $+\infty$ and still we can find a smooth curve. These random variables are called Continuous Variables, and the Normal Distribution then provides here probability of the value lying in a particular range for a given experiment.

Parameters of normal distribution

Since the mean, mode and median are the same in a normal distribution, there's no need to calculate them separately. These values represent the distribution's highest point, or the peak. All other values in the distribution then fall symmetrically around the mean. The width of the mean is defined by the standard deviation.
In fact, only two parameters are required to describe a normal distribution: the mean and the standard deviation.

1. The mean

The mean is the central highest value of the bell curve. All other values in the distribution either cluster around it or are at some distance away from it. Changing the mean on a graph will shift the entire curve along the x-axis, either toward the left or toward the right. However, its symmetricity will still be maintained.

2. The standard deviation

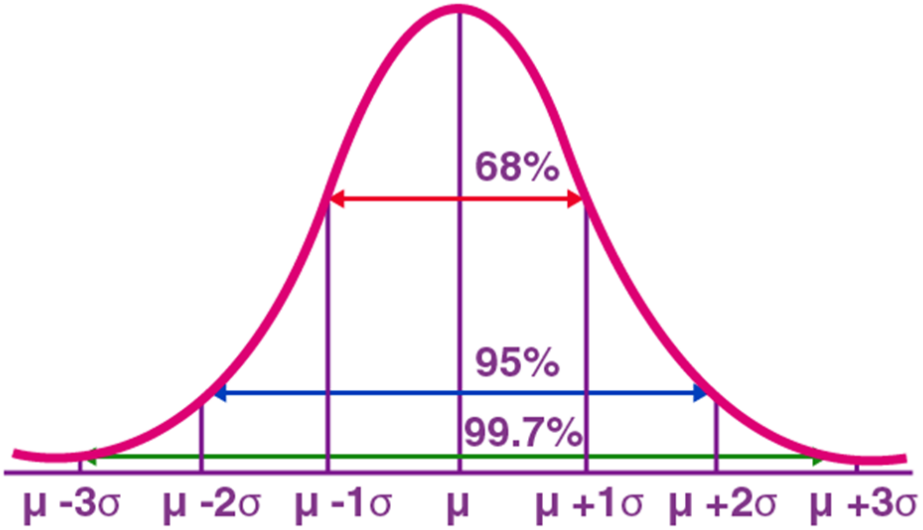
In general, standard deviation is a measure of variability in a distribution. In a bell curve, it defines the width of the distribution and shows how far away from the mean the other values fall. In addition, it represents the typical distance between the average and the observations.
Changing the standard deviation will change the distribution of values around the mean. A smaller deviation will reduce the spread -- tightening the distribution -- while a larger deviation will increase the spread and produce a wider distribution. As the distribution gets wider, it becomes more likely that values will be farther away from the mean.

Normal Distribution Standard Deviation

Generally, the normal distribution has any positive standard deviation. We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out. If the standard deviation is smaller, the data are somewhat close to each other and the graph becomes narrower. If the standard deviation is larger, the data are dispersed more, and the graph becomes wider. The standard deviations are used to subdivide the area under the normal curve. Each subdivided section defines the percentage of data, which falls into the specific region of a graph.

Using 1 standard deviation, **the Empirical Rule** states that,

- Approximately 68% of the data falls within one standard deviation of the mean. (i.e., Between Mean- one Standard Deviation and Mean + one standard deviation)
- Approximately 95% of the data falls within two standard deviations of the mean. (i.e., Between Mean- two Standard Deviation and Mean + two standard deviations)
- Approximately 99.7% of the data fall within three standard deviations of the mean. (i.e., Between Mean- three Standard Deviation and Mean + three standard deviations)



Thus, the empirical rule is also called the 68 – 95 – 99.7 rule.

Skewness and kurtosis in a normal distribution

Skewness represents a distribution's degree of symmetry. Since the normal distribution is perfectly symmetric, it has a skewness of zero. In other distributions with a skewness less than or greater than zero, the left tail (left skewness) or the right tail (right skewness) will be longer, respectively.
Kurtosis measures the thickness of each tail end of a distribution vis-à-vis the tails of a normal distribution. For a normal distribution, kurtosis is always equal to 3. In a distribution with kurtosis greater than 3, the tail data will exceed the tails of the normal distribution, resulting in a phenomenon called *fat tails*. In financial markets, fat tails describe tail risk -- the chance of a loss due to some rare event. Distributions with kurtosis less than 3 show tails that are skinnier than the tails of a normal distribution.

11. How do you handle missing data? What imputation techniques do you recommend?

Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. In the dataset, the blank shows the missing values.
In Pandas, usually, missing values are represented by **NaN**. It stands for **Not a Number**.

Types of Missing Values

Formally the missing values are categorized as follows:

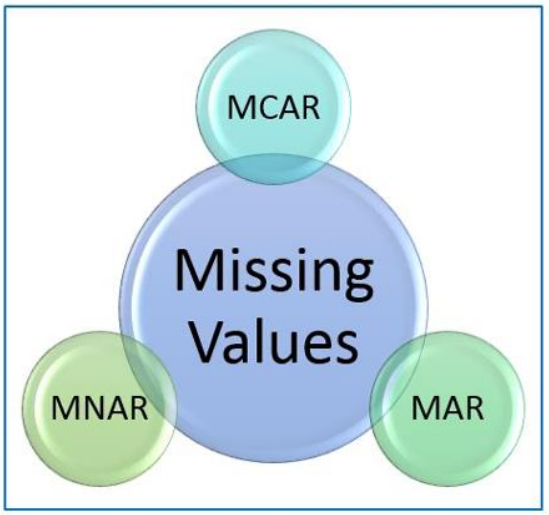


Figure 1 - Different Types of Missing Values in Datasets

Missing Completely At Random (MCAR)

In MCAR, the probability of data being missing is the same for all the observations. In this case, there is no relationship between the missing data and any other values observed or unobserved (the data which is not recorded) within the given dataset. That is, missing values are completely independent of other data. There is no pattern.

In the case of MCAR data, the value could be missing due to human error, some system/equipment failure, loss of sample, or some unsatisfactory technicalities while recording the values. For Example, suppose in a library there are some overdue books. Some values of overdue books in the computer system are missing. The reason might be a human error, like the librarian forgetting to type in the values. So, the missing values of overdue books are not related to any other variable/data in the system. It should not be assumed as it's a rare case. The advantage of such data is that the statistical analysis remains unbiased.

Missing At Random (MAR)

MAR data means that the reason for missing values can be explained by variables on which you have complete information, as there is some relationship between the missing data and other values/data. In this case, the data is not missing for all the observations. It is missing only within sub-samples of the data, and there is some pattern in the missing values.

For example, if you check the survey data, you may find that all the people have answered their 'Gender,' but 'Age' values are **mostly** missing for people who have answered their 'Gender' as 'female.' (The reason being most of the females don't want to reveal their age.)

So, the probability of data being missing depends only on the observed value or data. In this case, the variables 'Gender' and 'Age' are related. The reason for missing values of the 'Age' variable can be explained by the 'Gender' variable, but you can not predict the missing value itself.

Suppose a poll is taken for overdue books in a library. Gender and the number of overdue books are asked in the poll. Assume that most of the females answer the poll and men are less likely to answer. So why the data is missing can be explained by another factor, that is gender. In this case, the statistical analysis might result in bias. Getting an unbiased estimate of the parameters can be done only by modeling the missing data.

Missing Not At Random (MNAR)

Missing values depend on the unobserved data. If there is some structure/pattern in missing data and other observed data **can not explain** it, then it is considered to be Missing Not At Random (MNAR).

If the missing data does not fall under the MCAR or MAR, it can be categorized as MNAR. It can happen due to the reluctance of people to provide the required information. A specific group of respondents may not answer some questions in a survey.

For example, suppose the name and the number of overdue books are asked in the poll for a library. So most of the people having no overdue books are likely to answer the poll. People having more overdue books are less likely to answer the poll. So, in this case, the missing value of the number of overdue books depends on the people who have more books overdue.
Another example is that people having less income may refuse to share some information in a survey or questionnaire.
In the case of MNAR as well, the statistical analysis might result in bias.

Missing data can cause the below issues: –

1. **Incompatible with most of the Python libraries used in Machine Learning:-** Yes, you read it right. While using the libraries for ML(the most common is skLearn), they don't have a provision to automatically handle these missing data and can lead to errors.
2. **Distortion in Dataset:-** A huge amount of missing data can cause distortions in the variable distribution i.e it can increase or decrease the value of a particular category in the dataset.
3. **Affects the Final Model:-** the missing data can cause a bias in the dataset and can lead to a faulty analysis by the model.

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

1. Deleting the Missing value

Generally, this approach is not recommended. It is one of the quick and dirty techniques one can use to deal with missing values. If the missing value is of the type Missing Not At Random (MNAR), then it should not be deleted.
If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted (In the analysis, all cases with available data are utilized, while missing observations are assumed to be completely random (MCAR) and addressed through pairwise deletion.)
The disadvantage of this method is one might end up deleting some useful data from the dataset.
There are 2 ways one can delete the missing data values:

(a) Deleting the entire row (listwise deletion)

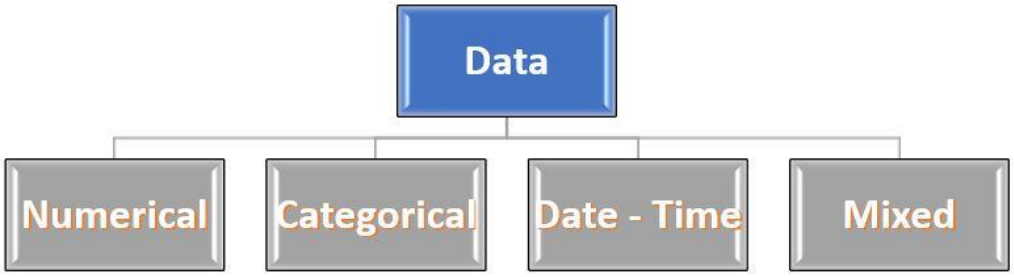
If a row has many missing values, you can drop the entire row. If every row has some (column) value missing, you might end up deleting the whole data.

(b) Deleting the entire column

If a certain column has many missing values, then you can choose to drop the entire column.

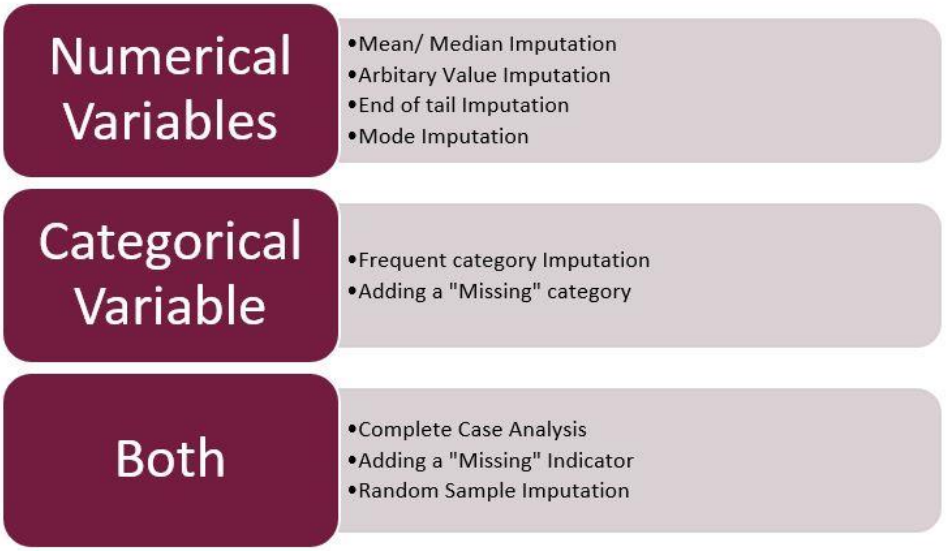
2. Imputing the Missing Value

Imputation
Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.
Most data is of 4 types:- Numeric, Categorical, Date-time & Mixed. These names are quite self-explanatory so not going much in-depth and describing them.



There are many imputation methods for replacing the missing values.

Techniques used In Imputation



(a) . **Complete Case Analysis(CCA):-**

This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. This method is also popularly known as “Listwise deletion”.

Assumptions:-

Data is Missing At Random(MAR).

Missing data is completely removed from the table.

Advantages:-

Easy to implement.

No Data manipulation required.

Limitations:-

Deleted data can be informative.

Can lead to the deletion of a large part of the data.

Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.

The production model will not know what to do with Missing data.

When to Use:-

Data is MAR(Missing At Random).

Good for Mixed, Numerical, and Categorical data.
Missing data is not more than 5% – 6% of the dataset.
Data doesn't contain much information and will not bias the dataset.

(b) **Arbitrary Value Imputation**

This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or "Missing" or "Not defined" for numerical & categorical variables.

Assumptions:-
Data is not Missing At Random.
The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data.

Advantages:-
Easy to implement.
We can use it in production.
It retains the importance of "missing values" if it exists.

Disadvantages:-
Can distort original variable distribution.
Arbitrary values can create outliers.
Extra caution required in selecting the Arbitrary value.

When to Use:-
When data is not MAR(Missing At Random).
Suitable for All.

(c) **Frequent Category Imputation**

This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as **Mode Imputation.**

Assumptions:-
Data is missing at random.
There is a high probability that the missing data looks like the majority of the data.

Advantages:-
Implementation is easy.
We can obtain a complete dataset in very little time.
We can use this technique in the production model.

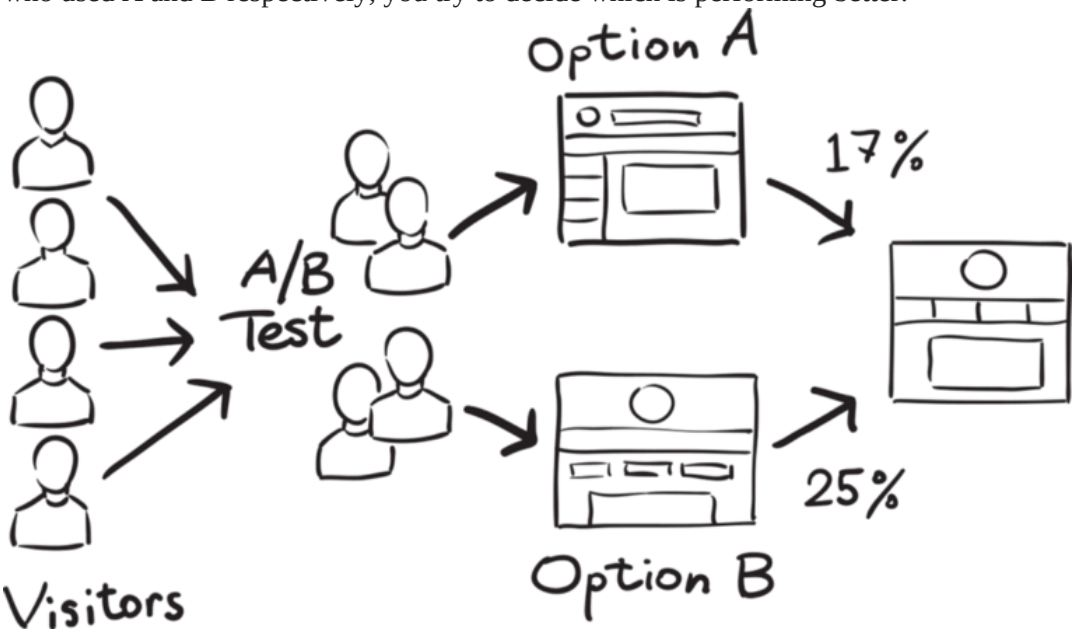
Disadvantages:-
The higher the percentage of missing values, the higher will be the distortion.
May lead to over-representation of a particular category.
Can distort original variable distribution.

When to Use:-
Data is Missing at Random(MAR)
Missing data is not more than 5% – 6% of the dataset.

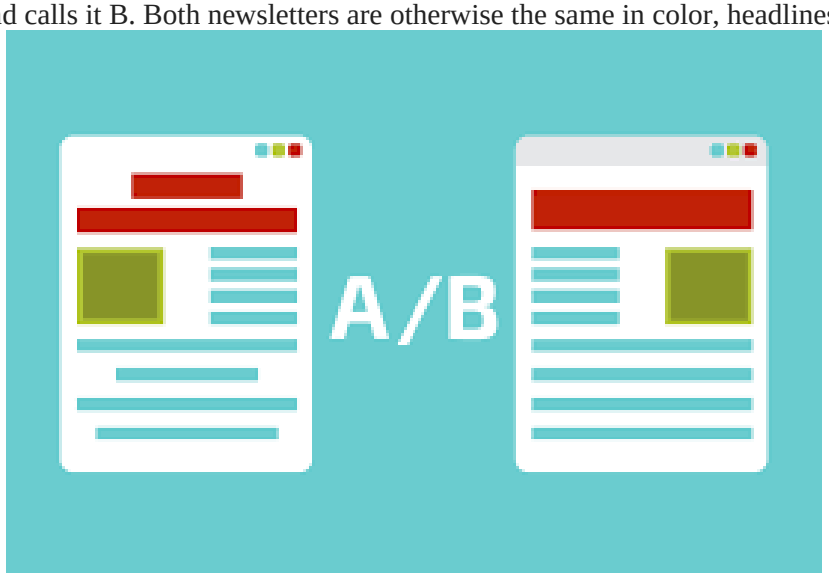
12. What is A/B testing?

A/B testing is a popular way to test your products and is gaining steam in the data science field. **A/B testing** is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.
For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.
In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.
Let's say there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in color, headlines, and format.



Objective

Our objective here is to check which newsletter brings higher traffic on the website i.e the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

1. Make a Hypothesis

Before making a hypothesis, let's first understand what is a hypothesis. A hypothesis is a tentative insight into the natural world; a concept that is not yet verified but if true would explain certain facts or phenomena.
It is an **educated guess** about something in the world around you. It should be testable, either by experiment or observation. In our example, the hypothesis can be "By making changes in the language of the newsletter, we can get more traffic on the website".
In hypothesis testing, we have to make two hypotheses i.e Null hypothesis and the alternative hypothesis. Let's have a look at both.

1. *Null hypothesis or H_0:*

   The **null hypothesis** is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is **no** difference between the control and variant groups. It states the default position to be tested or the situation as it is now, i.e. the status quo. Here our H_0 is " there is no difference in the conversion rate in customers receiving newsletter A and B".

2. ***Alternative Hypothesis or H_0:***

The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.
In our example, the H_a is- "**the conversion rate of newsletter B is higher than those who receive newsletter A**".
Now, we have to collect enough evidence through our tests to **reject the null hypothesis**.

2. Create Control Group and Test Group

Once we are ready with our null and alternative hypothesis, the next step is to decide the group of customers that will participate in the test. Here we have two groups – **The Control group**, and **the Test (variant) group**.
The Control Group is the one that will receive newsletter A and the Test Group is the one that will receive newsletter B.
For this experiment, we randomly select 1000 customers – 500 each for our Control group and Test group.
Randomly selecting the sample from the population is called **random sampling**. It is a technique where each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and **it's important to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.**
Another important aspect we must take care of is **the Sample size.** It is required that we determine the minimum sample size for our A/B test before conducting it so that we can eliminate **under coverage bias.** It is the bias from sampling too few observations.

3. Conduct the A/B Test and Collect the Data

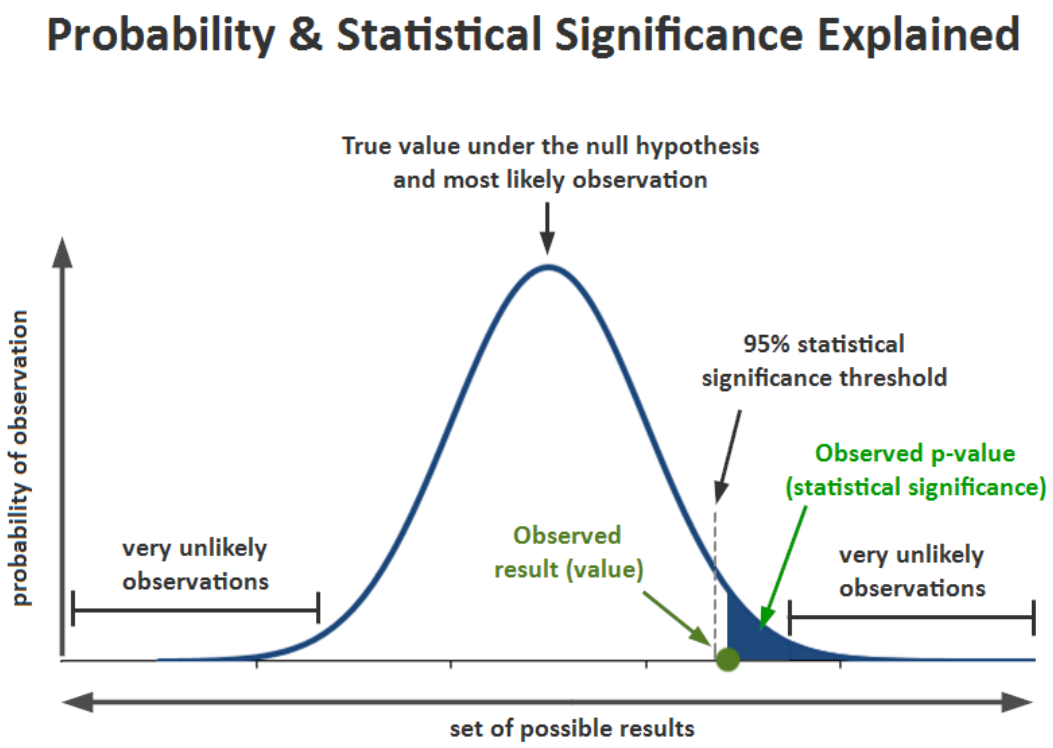
One way to perform the test is to calculate **daily conversion rates** for both the treatment and the control groups. Since the conversion rate in a group on a certain day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period.
When we run our experiment for one month, we noticed that the mean conversion rate for the Control group is 16% whereas that for the test Group is 19%.

Statistical significance of the Test

Now, the main question is – Can we conclude from here that the Test group is working better than the control group?
The answer to this is a simple No! For rejecting our null hypothesis we have to prove the **Statistical significance** of our test.
There are two types of errors that may occur in our hypothesis testing:

1. **Type I error**: We reject the null hypothesis when it is true. That is we accept the variant B when it is not performing better than A
2. **Type II error**: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A

To avoid these errors we must calculate the statistical significance of our test.
An experiment is considered to be statistically significant when we have enough evidence to prove that the result we see in the sample also exists in the population.
That means the difference between your control version and the test version is not due to some error or random chance. To prove the statistical significance of our experiment we can use a two-sample T-test.
The **two–sample t–test** is one of the most commonly **used** hypothesis **tests**. It is applied to compare whether the average difference between **the two** groups.



To understand this, we must be familiar with a few terms:

1. **Significance level (alpha):** The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. Generally, we use the significance value of 0.05
2. **P-Value:** It is the probability that the difference between the two values is just because of random chance. P-value is evidence against the null hypothesis. The smaller the p-value stronger the chances to reject the H_0. For the significance level of 0.05, if the p-value is lesser than it hence we can reject the null hypothesis

3. **Confidence interval:** The confidence interval is an observed range in which a given percentage of test outcomes fall. We manually select our desired confidence level at the beginning of our test. Generally, we take a 95% confidence interval

Next, we can calculate our t statistics using the below formula:

$$T - statistic = \frac{Observed\ value - hypothesized\ value}{Standard\ Error}$$

$$Stamdard\ Error = \sqrt{\frac{2 * Variance(sample)}{N}}$$

13. Is mean imputation of missing data acceptable practice?

The easiest way to impute is to replace each missing value with the mean of the observed values for that variable. imputing the mean preserves the mean of the observed data. So, if the data are missing completely at random, the estimate of the mean remains unbiased. Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. It is only reasonable idea if only few values are missing or there is small dataset size. But overall, it is not good idea to replace missing data with mean of dataset due to following reason:

• Mean of data is much more sensitive to outliers, so it can lead to underestimate or overestimate mean than actual mean of dataset if all data is present. Replacing missing values with median would be much more idea in that case.

• For large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations.

• Any statistic that uses mean from mean-imputed data that you would have gotten without the imputations will have a standard error that's too low. Because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.

• This strategy can severely distort the distribution for this variable, leading to underestimates of the standard deviation. Moreover, mean imputation distorts relationships between variables by "pulling" estimates of the correlation toward zero.

14. What is linear regression in statistics?

Regression analysis employs a model that describes the relationships between the dependent variables and the independent variables in a simplified mathematical form. • In statistics, linear regression is a technique to model or find linear relationship between dependent and independent variable. • When there is only one independent variable in the linear regression model, the model is generally termed as a simple linear regression model. When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.

• Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. There can be many line possible passing through data point but the best-fitting line is the line that minimizes the sum of the squared errors of prediction.

• In simple linear regression model equation of line is given by y = ß0 + ß1x where y is termed as the dependent or study variable and x is termed as the independent or explanatory variable. The terms ß0and ß1 are the parameters of the model. The parameter termed as an intercept term, and the parameter ß1 is termed as the slope parameter. These parameters are usually called as regression coefficients.

• Although we minimize the sum of the squared distances of the actual y scores from the predicted y scores (y'), there is a distribution of these distances or errors in prediction which is important to

discuss. We will define these directed (signed) distances (residuals) as e = (y-y'), where y' is our predicted value. Clearly both positive and negative values occur with a mean of zero.
• The square root of this value is the standard deviation and is known as the standard error of estimate

15. What are the various branches of statistics?

The field of statistics is divided into two major divisions: descriptive and inferential. Each of these segments is important, offering different techniques that accomplish different objectives.

(1) **Descriptive statistics:** Descriptive statistics is the part of statistics that deals with presenting the data we have. Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data. This can take two basic forms – presenting aspects of the data either visually or numerically. Descriptive statistics is responsible for summarizing a statistical sample (set of data obtained from a population) Rather than learning about population Which represents the sample. Some of the measures commonly used in descriptive statistics to describe a set of data are the measures of central tendency and the Measures of variability or dispersion.

(2) **Inferential Statistics:** Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates. Most predictions of the future and generalization on a population study of a smaller specimen are in the scope of the inference statistics. Different Techniques use to examine the relationships between variables and draw conclusion & predication in inferential statistics. Some of techniques are linear regression analyses, logistic regression analyses, ANOVA, correlation analyses, structural equation modeling, and survival analysis.