UNIT - 3
Statistical Experiments and Significance Testing

SYLLABUS –

Statistical Experiments and Significance Testing, A/B Testing, Hypothesis Tests, Resampling, Statistical Significance and p-Values, Multiple Testing, Degrees of Freedom, ANOVA, Chi-Square Test, Multi-Arm Bandit Algorithm. Power and Sample Size

Statistical Experiments and Significance Testing

- Whenever there is a reference to statistical significance, t-tests, or p-values, it is typically in the context of the classical statistical inference "pipeline" as shown in Figure.

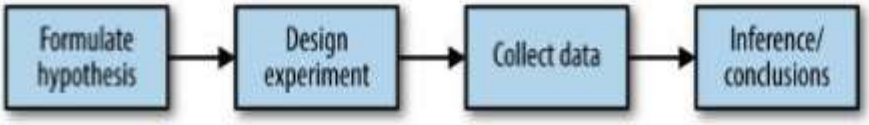


Figure - The classical statistical inference pipeline

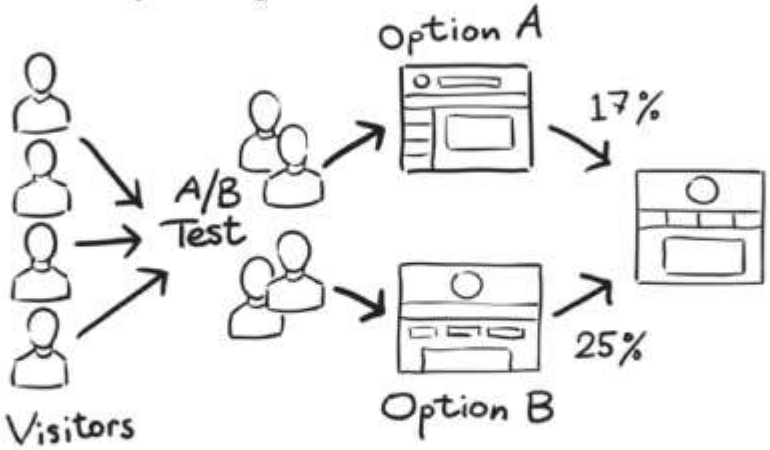
- This process starts with a hypothesis ("drug A is better than the existing standard drug," or "price A is more profitable than the existing price B").
- An experiment (it might be an A/B test) is designed to test the hypothesis—designed in such a way that it hopefully will deliver conclusive results.
- The data is collected and analyzed, and then a conclusion is drawn. The term inference reflects the intention to apply the experiment results, which involve a limited set of data, to a larger process or population.

What is A/B Testing? How A/B testing is useful in web design and marketing?
OR
Explain A/B Testing with suitable example?

- An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior.
- Often one of the two treatments is the standard existing treatment, or no treatment. If a standard (or no) treatment is used, it is called the control.
- A typical hypothesis is that a new treatment is better than the control.
- A/B testing data science is a methodical way to evaluate the performance of two variants of a website, app, or campaign. It also goes by the name "split testing."
- By dividing traffic into two groups and serving one group the A/B version while serving the other group the control, A/B testing seeks to determine what works and doesn't work for your business (the base version).
- Key Terms for A/B Testing are

Treatment- Something (drug, price, web headline) to which a subject is exposed.
Treatment group- A group of subjects exposed to a specific treatment.
Control group- A group of subjects exposed to no (or standard) treatment.
Randomization- The process of randomly assigning subjects to treatments.
Subjects- The items (web visitors, patients, etc.) that are exposed to treatments.
Test statistic- The metric used to measure the effect of the treatment.

- A/B tests are common in web design and marketing, since results are so readily measured. Some examples of A/B testing include:
  - Testing two soil treatments to determine which produces better seed germination.
  - Testing two therapies to determine which suppresses cancer more effectively.
  - Testing two prices to determine which yields more net profit.
  - Testing two web headlines to determine which produces more clicks.
  - Testing two web ads to determine which generates more conversions.
- 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.

How Hypothesis Test will help you to learn whether random chance might be responsible for an observed effect?
OR
Define Null Hypothesis and Alternative hypothesis with example?
OR
Explain One-Way Versus Two-Way Hypothesis Tests?

- Hypothesis tests, also called significance tests, are ubiquitous in the traditional statistical analysis of published research.
- Their purpose is to help you learn whether random chance might be responsible for an observed effect.

Key Terms for Hypothesis Tests

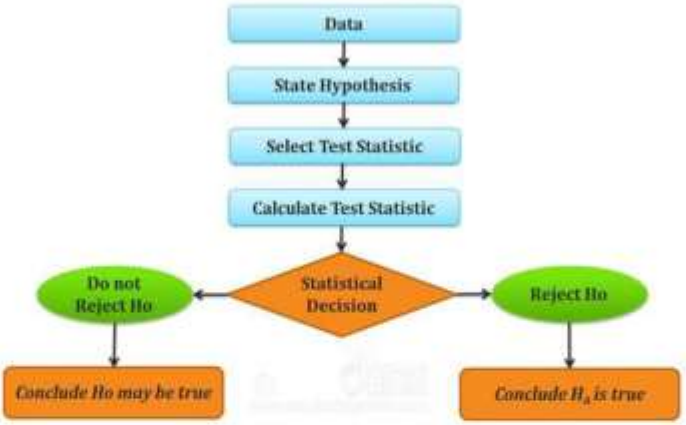
Null hypothesis
The hypothesis that chance is to blame.

Alternative hypothesis
Counterpoint to the null (what you hope to prove).

One-way test
Hypothesis test that counts chance results only in one direction.

Two-way test
Hypothesis test that counts chance results in two directions.

STEPS IN HYPOTHESIS TESTING



How does A/B Testing Work?

- 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 a hypothesis is. 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.

   a) **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".

   b) **Alternative Hypothesis or H_a:**

   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"**.

Alternative Hypothesis

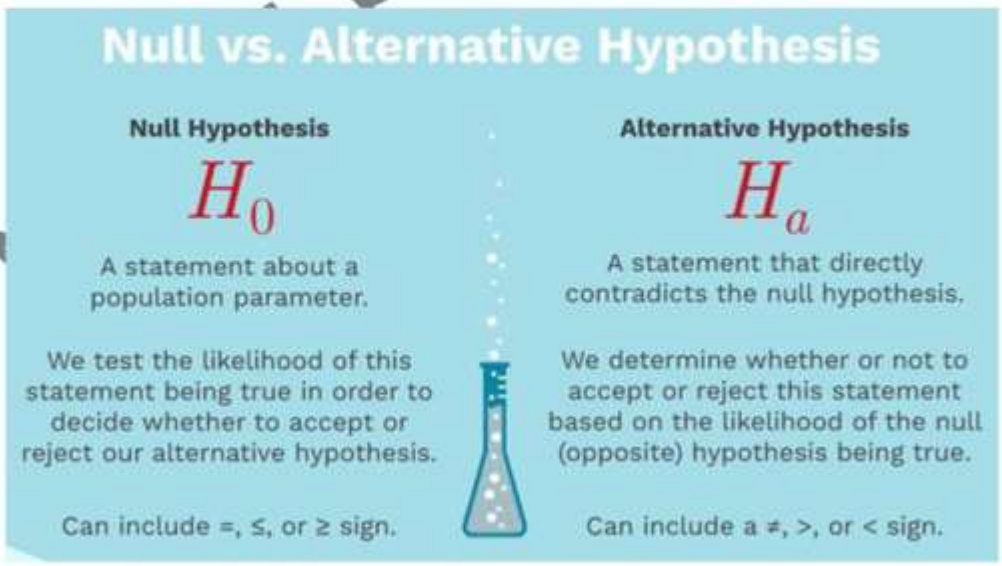
Hypothesis tests by their nature involve not just a null hypothesis but also an offsetting alternative hypothesis. Here are some examples:

- Null = "no difference between the means of group A and group B"; alternative = "A is different from B" (could be bigger or smaller)
- Null = "A ≤ B"; alternative = "A > B"
- Null = "B is not X% greater than A"; alternative = "B is X% greater than A"

Taken together, the null and alternative hypotheses must account for all possibilities. The nature of the null hypothesis determines the structure of the hypothesis test.

One-Way Versus Two-Way Hypothesis Tests

- Often in an A/B test, you are testing a new option (say, B) against an established default option (A), and the presumption is that you will stick with the default option unless the new option proves itself definitively better.
- In such a case, you want a hypothesis test to protect you from being fooled by chance in the direction favoring B. You don't care about being fooled by chance in the other direction, because you would be sticking with A unless B proves definitively better.
- So you want a directional alternative hypothesis (B is better than A). In such a case, you use a one-way (or onetail) hypothesis test.
- This means that extreme chance results in only one direction count toward the p-value.
- If you want a hypothesis test to protect you from being fooled by chance in either direction, the alternative hypothesis is bidirectional (A is different from B; could be bigger or smaller).
- In such a case, you use a two-way (or two-tail) hypothesis. This means that extreme chance results in either direction count toward the p-value.



What is Resampling and permutation test?

- Resampling in statistics means to repeatedly sample values from observed data, with a general goal of assessing random variability in a statistic.
- It can also be used to assess and improve the accuracy of some machine-learning models (e.g., the predictions from decision tree models built on multiple bootstrapped data sets can be averaged in a process known as bagging.)
- There are two main types of resampling procedures: the bootstrap and permutation tests. Permutation tests are used to test hypotheses, typically involving two or more groups.

Key Terms for Resampling

Permutation test
The procedure of combining two or more samples together and randomly (or exhaustively) reallocating the observations to resamples.

Synonyms
Randomization test, random permutation test, exact test

Resampling
Drawing additional samples ("resamples") from an observed data set.

With or without replacement
In sampling, whether or not an item is returned to the sample before the next draw.

Permutation Test

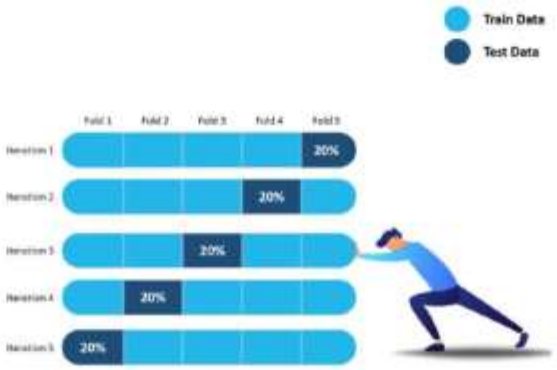
- In a permutation procedure, two or more samples are involved, typically the groups in an A/B or other hypothesis test.
- Permute means to change the order of a set of values. The first step in a permutation test of a hypothesis is to combine the results from groups A and B (and, if used, C, D,).
- This is the logical embodiment of the null hypothesis that the treatments to which the groups were exposed do not differ.
- We then test that hypothesis by randomly drawing groups from this combined set and seeing how much they differ from one another.
- The permutation procedure is as follows:
  1. Combine the results from the different groups into a single data set.
  2. Shuffle the combined data and then randomly draw (without replacement) a resample of the same size as group A (clearly it will contain some data from the other groups).
  3. From the remaining data, randomly draw (without replacement) a resample of the same size as group B.
  4. Do the same for groups C, D, and so on. You have now collected one set of resamples that mirror the sizes of the original samples.
  5. Whatever statistic or estimate was calculated for the original samples (e.g., difference in group proportions), calculate it now for the resamples, and record; this constitutes one permutation iteration.
  6. Repeat the previous steps R times to yield a permutation distribution of the test statistic.

Types of Resampling:

- Resampling is the method that consists of drawing repeatedly drawing samples from the population.
- It involves the selection of randomized cases with replacement from sample.
- Two common method of Resampling are:
  K-fold Cross-validation
  Bootstrapping

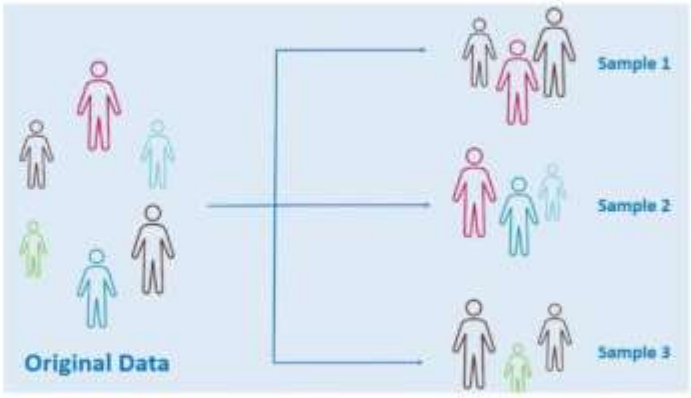
K-fold cross-validation:

- In this method population data is divided into k equal sets in which one set is considered as the test set for the experiment while all other set will be used to train the model. In first experiment, first set is considered as the test set and all other as trained set. Process will be repeated k-time by choosing different sets as a test set.



Bootstrapping:

- In bootstrapping, samples are drawn with replacement (i.e. one observation can be repeated in more than one group) and the remaining data which are not used in samples are used to test the model.



Discuss about the Statistical Significance and p-Values. Write 6 principles by ASA Statements.

- Statistical significance is how statisticians measure whether an experiment (or even a study of existing data) yields a result more extreme than what chance might produce.
- If the result is beyond the realm of chance variation, it is said to be statistically significant.

Key Terms for Statistical Significance and p-Values

p-value
Given a chance model that embodies the null hypothesis, the p-value is the probability of obtaining results as unusual or extreme as the observed results.

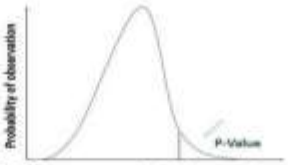
Alpha
The probability threshold of "unusualness" that chance results must surpass for actual outcomes to be deemed statistically significant.

Type 1 error
Mistakenly concluding an effect is real (when it is due to chance).

Type 2 error
Mistakenly concluding an effect is due to chance (when it is real).

- P-Value, or the Probability Value, is the determining factor on a null hypothesis for the probability of an assumed result to be true and being accepted or rejected and acceptance of the alternate result in case of rejection of the assumed result.

P-Value



- P-Value calculation also includes a probability of other results' occurrence. However, statisticians refer to this value for more relevant results.
- In most cases, it lies within a range of 0 – 0.05 (5%) and has a negative result, which means the alternate result would be considered, and a value higher than 0.05 signifies that it will accept the desired result.
- However, this will not be hard and fast for all cases and will depend upon the conditions and product.
- Always a probability of the occurrence of a required result when in a scenario made a **null hypothesis**.
- For example, in a hypothetical situation, we survey a new appliance in the market, and results assume that 60% of females will accept the appliance, with an alternate result expected that 60% of males will accept the appliance.
- With the help of the p-value chart, we try to determine the results. A higher value will signify that the assumed expected result is "True," which means 60% of females accept the appliance.
- Consequently, a lower would imply acceptance of the alternate results, which means 60% of males accept the appliance. Hence, it determines the acceptance or rejection of an assumed result.

Formula

It can be calculated using z analysis (z test) where:

$$Z = \frac{(P1 - P0)}{\sqrt{(P0(1 - P0))/n}}$$

Where,
P1 = sample proportion of the whole population
P0 = Assumed proportion for the result to occur
n = size of the population
The Z-value is predicted from previous calculations. If the **p-value statistics** is equal to or less than the calculated z value, then the sample can be approved for the desired result (null hypothesis). Else gets rejected, and the alternate result gets approved.

P-value

- In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.
- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.
- A p-value of 0.05 or lower is generally considered statistically significant.
- P-value can serve as an alternative to—or in addition to—preselected confidence levels for hypothesis testing.
- The ASA (American Statistical Association) statement stressed **six** principles for researchers and journal editors:
    1. P-values can indicate how incompatible the data are with a specified statistical model.
    2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
    3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
    4. Proper inference requires full reporting and transparency.
    5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
    6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

What is t-test? What are the types of t-test?

- There are numerous types of significance tests, depending on whether the data comprises count data or measured data, how many samples there are, and what's being measured.
- A very common one is the t-test, named after Student's t-distribution, originally developed by W. S. Gosset to approximate the distribution of a single sample mean.
- A ***t* test** is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.
- A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related.
- T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.
- The t-test is a test used for hypothesis testing in statistics and uses the t-statistic, the t-distribution values, and the degrees of freedom to determine statistical significance.

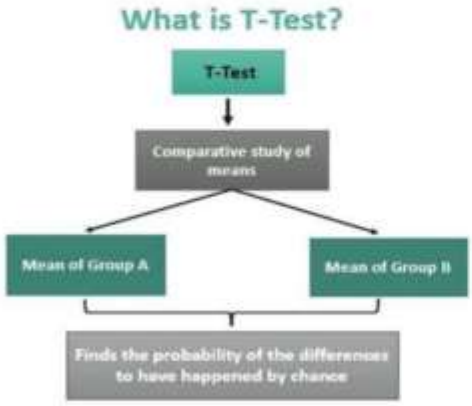
Key Terms for t-Tests

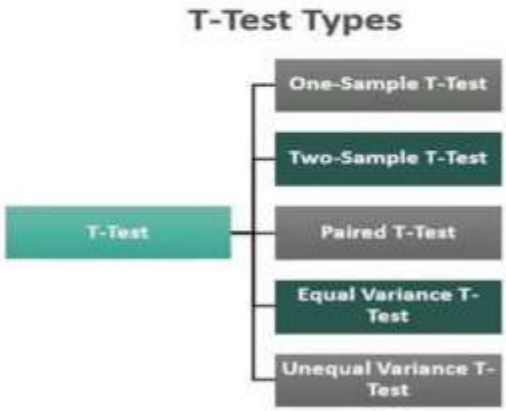
Test statistic
A metric for the difference or effect of interest.

t-statistic
A standardized version of common test statistics such as means.

t-distribution
A reference distribution (in this case derived from the null hypothesis), to which the observed t-statistic can be compared.

- A t-test is an inferential statistic used to determine if there is a statistically significant difference between the means of two variables.
- The t-test is a test used for hypothesis testing in statistics.
- Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.
- T-tests can be dependent or independent.





- When choosing a *t* test, you will need to consider two things: whether the groups being compared come from a single population or two different populations, and whether you want to test the difference in a specific direction.

One-sample, two-sample, or paired *t* test?

- If the groups come from a single population (e.g., measuring before and after an experimental treatment), perform a **paired *t* test**. This is a within-subjects design.
- If the groups come from two different populations (e.g., two different species, or people from two separate cities), perform a **two-sample *t* test** (a.k.a. **independent *t* test**). This is a between-subjects design.
- If there is one group being compared against a standard value (e.g., comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample *t* test**.

One-tailed or two-tailed *t* test?

- If you only care whether the two populations are different from one another, perform a **two-tailed *t* test**.
- If you want to know whether one population mean is greater than or less than the other, perform a **one-tailed *t* test.**

While the **T values** indicate the chances of the **difference** between the sample means being a result obtained by chance, **p-values** reflect the probability of having sufficient proof to negate the **indifference** between the mean of the two samples.

Explain Multiple Testing with its key terms?

- Multiple testing refers to any instance that involves the simultaneous testing of more than one hypothesis.
- If decisions about the individual hypotheses are based on the unadjusted marginal p-values, then there is typically a large probability that some of the true null hypotheses will be rejected.
- In supervised learning tasks, a holdout set where models are assessed on data that the model has not seen before mitigates this risk.
- In statistical and machine learning tasks not involving a labeled holdout set, the risk of reaching conclusions based on statistical noise persists.
- In statistics, there are some procedures intended to deal with this problem in very specific circumstances.
- For example, if you are comparing results across multiple treatment groups, you might ask multiple questions. So, for treatments A–C, you might ask:

i. Is A different from B?
ii. Is B different from C?
iii. Is A different from C?

- This gives rise to lots of opportunities to find something interesting in the data, including multiplicity issues such as:
  1) Checking for multiple pairwise differences across groups
  2) Looking at multiple subgroup results ("we found no significant treatment effect overall, but we did find an effect for unmarried women younger than 30")
  3) Trying lots of statistical models
  4) Including lots of variables in models
  5) Asking a number of different questions (i.e., different possible outcomes)
- Key Terms for Multiple Testing are

Type 1 error
Mistakenly concluding that an effect is statistically significant.
False discovery rate
Across multiple tests, the rate of making a Type 1 error.
Alpha inflation
The multiple testing phenomenon, in which alpha, the probability of making a Type 1 error, increases as you conduct more tests.
Adjustment of p-values
Accounting for doing multiple tests on the same data.
Overfitting
Fitting the noise.

What is Degree of Freedom? Explain with example?

- Degrees of freedom are the maximum number of logically independent values, which may vary in a data sample.
- Degrees of freedom are calculated by subtracting one from the number of items within the data sample. Degrees of freedom are calculated by subtracting one from the number of items within the data sample.
- Degrees of freedom are commonly discussed in various forms of hypothesis testing in statistics, such as a chi-square.
- Degrees of freedom can describe business situations where management must make a decision that dictates the outcome of another variable.
- Typically, the degree of freedom equals your sample size minus the number of parameters you need to calculate during an analysis. It is usually a positive whole number. Degree of freedom is a combination of how much data you have and how many parameters you need to estimate.
- The degrees of freedom formula is straightforward. Calculating the degrees of freedom is often the sample size minus the number of parameters you're estimating:

$$DF = N - P$$

Where:
N = sample size
P = the number of parameters or relationships

- **For example** :-
  If we have 100 independent samples and we want to calculate a statistic of the sample, like the mean. All 100 samples are used in the calculation and there is

one statistic, so the number of degrees of freedom for the mean, in this case is calculated as:

DF=N-P
DF=100-1 (N=100, P=1)
DF=99

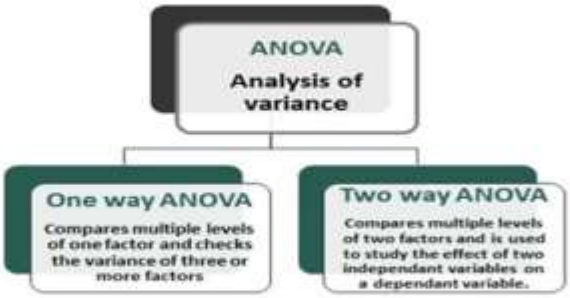
- Degree of freedom is very important in data distributions and Statistical Hypothesis Test.
- Several statistical tests use the concept of degrees of freedom, including t-tests, F-tests, chi-squared tests, and ANOVA. Here are details:
  1) **The T-test** is a statistical test that is used to determine whether two groups are significantly different from each other. The degree of freedom in a t-test is the number of observations minus the number of parameters estimated, which is usually one for a two-sample t-test.
  2) In an **F-test**, degrees of freedom refer to the number of independent observations that are available to estimate the variance of a population. The formula for calculating degrees of freedom for the numerator is **n1 – 1**, and denominator is **n2 – 1**, where n1 and n2 are the number of observations in the two groups (being compared) belonging to numerator and denominator respectively.
  3) The **chi-squared test** is used to determine whether there is a significant association between two categorical variables, and the degree of freedom in this test depends on the number of categories. The formula for calculating degrees of freedom in a chi-square test is (r – 1) x (c – 1), where r is the number of rows in the contingency table and c is the number of columns.

Explain Analysis of Variance (ANOVA) test with example? State the procedure used to ANOVA test?

- Suppose that, instead of an A/B test, we had a comparison of multiple groups, say A/B/C/D, each with numeric data.
- These tests allow us to determine if **two population or sample means are statistically significantly different**. However, what if we wanted to test the means between **three** samples?
- One would have to carry **three** different T-Tests in this scenario and if there were **four** groups, we would need **six** tests. The number of tests needed quickly explodes as the number of groups increases.
- The statistical procedure that tests for a statistically significant difference among the groups is called analysis of variance, or ANOVA.
- The ANOVA test **allows a comparison of more than two groups at the same time to determine whether a relationship exists between them**.
- ANOVA is to test for differences among the means of the population by examining the amount of variation within each sample, relative to the amount of variation between the samples.
- ANOVA test, in its simplest form, is used to check whether the means of three or more populations are equal or not.
- The ANOVA test applies when there are more than two independent groups.
- The goal of the ANOVA test is to check for variability within the groups as well as the variability among the groups.
- The ANOVA test statistic is given by the f test.

- An ANOVA test can be either one-way or two-way depending upon the number of independent variables.

Classification of ANOVA Test



Key Terms for ANOVA

Pairwise comparison
A hypothesis test (e.g., of means) between two groups among multiple groups.

Omnibus test
A single hypothesis test of the overall variance among multiple group means.

Decomposition of variance
Separation of components contributing to an individual value (e.g., from the overall average, from a treatment mean, and from a residual error).

F-statistic
A standardized statistic that measures the extent to which differences among group means exceed what might be expected in a chance model.

SS
"Sum of squares," referring to deviations from some average value.

Example –

- Table shows the stickiness of four web pages, defined as the number of seconds a visitor spent on the page.
- The four pages are switched out so that each web visitor receives one at random.
- There are a total of five visitors for each page, and in Table, each column is an independent set of data. (Stickiness is **the metric that measures the number of people that are highly engaged with a product**.)
- We must take the visitors as they come. Visitors may systematically differ **depending on time of day, time of week, season of the year, conditions of their internet, what device they are using**, and so on. These factors should be considered as potential bias when the experiment results are reviewed.

Stickiness (in seconds) of four web pages

	Page 1	Page 2	Page 3	Page 4
	164	178	175	155
	172	191	193	166
	177	182	171	164
	156	185	163	170
	195	177	176	168
Average	172	185	176	162
Grand average				173.75

- When we were comparing just two groups, it was a simple matter; we merely looked at the difference between the means of each group. With four means, there are six possible comparisons between groups:
  - Page 1 compared to page 2
  - Page 1 compared to page 3
  - Page 1 compared to page 4
  - Page 2 compared to page 3
  - Page 2 compared to page 4
  - Page 3 compared to page 4

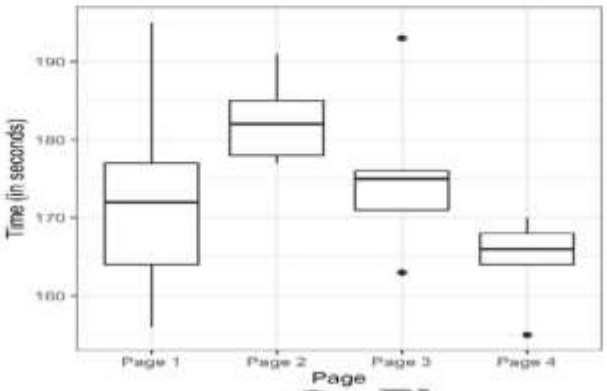


Figure-Boxplots of the four groups show considerable differences among them.

The **procedure used to test this is ANOVA.** The basis for it can be seen in the following resampling procedure (specified here for the A/B/C/D test of web page stickiness):

1. Combine all the data together in a single box.
2. Shuffle and draw out four resamples of five values each.
3. Record the mean of each of the four groups.
4. Record the variance among the four group means.
5. Repeat steps 2–4 many (say, 1,000) times.

ANOVA Assumptions

- The population the groups are sampled from is a **normal distribution.**
- Groups are sampled **independently.**
- The populations that are used for the sample have **equal variances.**

Variance

The main concept in the ANOVA tests is **variance**, which a measure of the **spread/dispersion** of the data. For the normal distribution, the variance is defined as:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Where $\bar{x}$ is the mean of the data, x_i are the individual data points, n is the number of data points and σ^2 is the variance.

Sum of Squares

The variance is used to calculate the **Sum of Squares (Error) (SSE)**. This is just the numerator of the variance equation above:

$$SSE = \sum (x_i - \bar{x})^2$$

In the ANOVA test, you carry out three different SSE:

- **Sum of Squares within groups (SSW):** This is the just the SSE within each individual sample.
- **Sum of Squares between groups (SSB):** This is SSE between the mean of each sample and the global/grand mean (the mean of the means of each group!).
- **Sum of Squares Total (SST):** This is the SSE of the whole dataset which is done by combining all the samples together.

A known result is that *SST = SSB + SSW*.

Write steps to perform one way ANOVA test?

- The one way ANOVA test is used to determine whether there is any difference between the means of three or more groups. A one way ANOVA will have only one independent variable. The hypothesis for a one way ANOVA test can be set up as follows:

  Null Hypothesis, H_0**:** $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

  Alternative Hypothesis, H_1**:** The means are not equal

  Decision Rule: If test statistic > critical value then reject the null hypothesis and conclude that the means of at least two groups are statistically significant.

The steps to perform the one way ANOVA test are given below:

- **Step 1:** Calculate the mean for each group.
- **Step 2:** Calculate the total mean. This is done by adding all the means and dividing it by the total number of means.
- **Step 3:** Calculate the SSB.
- **Step 4:** Calculate the between groups degrees of freedom.
- **Step 5:** Calculate the SSE.
- **Step 6:** Calculate the degrees of freedom of errors.
- **Step 7:** Determine the MSB and the MSE.
- **Step 8:** Find the f test statistic.
- **Step 9:** Using the f table for the specified level of significance, α, find the critical value. This is given by $F(\alpha, df_1. df_2)$.
- **Step 10:** If f > F then reject the null hypothesis.

Limitations of One Way ANOVA Test

- The one way ANOVA is an omnibus test statistic.
- This implies that the test will determine whether the means of the various groups are statistically significant or not.
- However, it cannot distinguish the specific groups that have a statistically significant mean.
- Thus, to find the specific group with a different mean, a post hoc test needs to be conducted.

Write steps to perform two way ANOVA test?

- The two way ANOVA has two independent variables. Thus, it can be thought of as an extension of a one way ANOVA where only one variable affects the dependent variable.
- A two way ANOVA test is used to check the main effect of each independent variable and to see if there is an interaction effect between them.
- To examine the main effect, each factor is considered separately as done in a one way ANOVA.
- Furthermore, to check the interaction effect, all factors are considered at the same time. There are certain assumptions made for a two way ANOVA test.

These are given as follows:

- The samples drawn from the population must be independent.
- The population should be approximately normally distributed.
- The groups should have the same sample size.
- The population variances are equal.

Suppose in the two way ANOVA example, as mentioned above, the income groups are low, middle, high. The gender groups are female, male, and transgender.

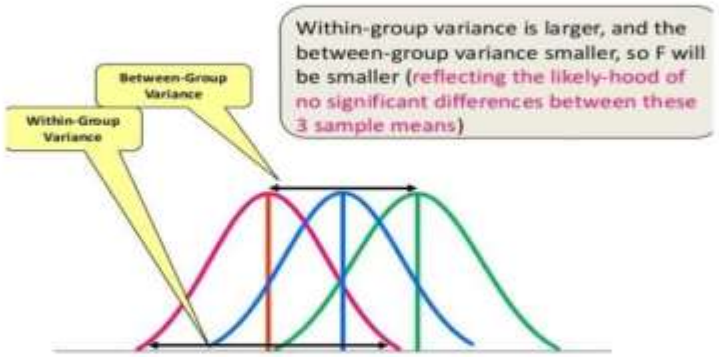
Then there will be 9 treatment groups and the three hypotheses can be set up as follows:

H01: All income groups have equal mean anxiety.
H11: All income groups do not have equal mean anxiety.
H02: All gender groups have equal mean anxiety.
H12: All gender groups do not have equal mean anxiety.
H03: Interaction effect does not exist
H13: Interaction effect exists.

Explain F-Statistics in detail?

- F test is a statistical test that is used in hypothesis testing to check whether the variances of two populations or two samples are equal or not.
- In an f test, the data follows an f distribution. This test uses the f statistic to compare two variances by dividing them.
- An f test can either be one-tailed or two-tailed depending upon the parameters of the problem.
- The f value obtained after conducting an f test is used to perform the one-way ANOVA (analysis of variance) test.
- The statistic that measures whether the means of different samples are significantly different is called the F-Ratio. The lower the F-Ratio, the more similar will the sample means be. In that case, we cannot reject the null hypothesis.

F = Between-group variability / Within-group variability



The test statistic for the ANOVA test is the F-Statistic, which is calculated as follows:

$$F = \frac{SSB/n_1}{SSW/n_2}$$

where *n_1* and *n_2* are the **degrees of freedom** for each Sum of Squares (between and within groups):

$$n_1 = m - 1$$
$$n_2 = n - m$$

Where m is the number of groups and n is the total number of data points.

Compare one way ANOVA and Two way ANOVA test?

BASIS FOR COMPARISON	ONE WAY ANOVA	TWO WAY ANOVA
Meaning	One way ANOVA is a hypothesis test, used to test the equality of three of more population means simultaneously using variance.	Two way ANOVA is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
Independent Variable	One	Two
Compares	Three or more levels of one factor.	Effect of multiple level of two factors.
Number of Observation	Need not to be same in each group.	Need to be equal in each group.
Design of experiments	Need to satisfy only two principles.	All three principles needs to be satisfied.

Explain how chi-square test is used with count data to test some expected distribution?

- The chi-square test is used with count data to test how well it fits some expected distribution.
- The most common use of the chi-square statistic in statistical practice is with r × c contingency tables, to assess whether the null hypothesis of independence among variables is reasonable.
- The chi-square test was originally developed by Karl Pearson in 1900.
- r × c means "rows by columns"—a 2 × 3 table has two rows and three columns.

Key Terms for Chi-Square Test

Chi-square statistic
A measure of the extent to which some observed data departs from expectation.

Expectation or expected
How we would expect the data to turn out under some assumption, typically the null hypothesis.

- Suppose you are testing three different headlines—A, B, and C—and you run them each on 1,000 visitors, with the results shown

Table Web testing results for three different headlines

	Headline A	Headline B	Headline C
Click	14	8	12
No-click	986	992	988

- The headlines certainly appear to differ. **Headline A returns nearly twice the click rate of B.** The actual numbers are small, though.
- A resampling procedure can test whether the click rates differ to an extent greater than chance might cause.
- For this test, we need to have the **"expected"** distribution of clicks, and in this case, that would be under the **null hypothesis assumption** that all three headlines share the same click rate, for an **overall click rate of 34/3,000.**
- Under this assumption, our contingency table would look like Table

Table Expected if all three headlines have the same click rate (null hypothesis)

	Headline A	Headline B	Headline C
Click	11.33	11.33	11.33
No-click	988.67	988.67	988.67

The *Pearson residual* is defined as:

$$R = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

R measures the extent to which the actual counts differ from these expected counts (see Table 3-6).

Table 3-6. Pearson residuals

	Headline A	Headline B	Headline C
Click	0.792	−0.990	0.198
No-click	−0.085	0.106	−0.021

The chi-square statistic is defined as the sum of the squared Pearson residuals:

$$X = \sum_{i}^{r} \sum_{j}^{c} R^2$$

where r and c are the number of rows and columns, respectively.

When to use chi-square test?

- Chi-Square test is designed for a specific set of data types, and that is a categorical variable.
- This means the test could **not be applied to continuous data types.**
- If it is to be applied on a continuous data type, the data needs to be divided into buckets, and frequency or count for each bucket needs to be provided.
- difference between categorical and continuous data types

Continuous Data Type – Continuous data types are ones that are infinite numerical value between any two values. For example, salary, time.

Categorical Data Type – Categorical data types are ones that contain a finite set of distinct categories or groups. For example, gender, marital status.

Chi-Square Distribution

- When we consider, the null speculation is true, the sampling distribution of the test statistic is called as chi-squared distribution.

- The chi-squared test helps to determine whether there is a notable difference between the normal frequencies and the observed frequencies in one or more classes or categories. It gives the probability of independent variables.
- Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.

Finding P-Value

- P stands for probability here. To calculate the p-value, the chi-square test is used in statistics. The different values of p indicates the different hypothesis interpretation, are given below:

P≤ 0.05; Hypothesis rejected
P>0.05; Hypothesis Accepted

- Probability is all about chance or risk or uncertainty. It is the possibility of the outcome of the sample or the occurrence of an event.
- But when we talk about statistics, it is more about how we handle various data using different techniques. It helps to represent complicated data or bulk data in a very easy and understandable way.
- It describes the collection, analysis, interpretation, presentation, and organization of data. The concept of both probability and statistics is related to the chi-squared test.

Formula

The chi-squared test is done to check if there is any difference between the observed value and expected value. The formula for chi-square can be written as;

$$X^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

or

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$

where O_i is the observed value and E_i is the expected value.

Write steps to perform chi-square test with an example?

- The chi-squared test helps to determine whether there is a notable difference between the normal frequencies and the observed frequencies in one or more classes or categories. It gives the probability of independent variables.
- Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.

Steps to perform the Chi-Square Test:

1. Define Hypothesis.
2. Build a Contingency table.
3. Find the expected values.
4. Calculate the Chi-Square statistic.
5. Accept or Reject the Null Hypothesis.

Steps for Chi-Square Test with an example:

- Consider a data-set where we have to determine why customers are leaving the bank, let's perform a Chi-Square test for two variables.
- Gender of a customer with values as Male/Female as the predictor and Exited describes whether a customer is leaving the bank with values Yes/No as the response.
- In this test we will check is there any relationship between Gender and Exited.

1.Define Hypothesis

Null Hypothesis (H0): Two variables are independent.
Alternate Hypothesis (H1): Two variables are not independent.

2. Contingency table

A table showing the distribution of one variable in rows and another in columns. It is used to study the relation between two variables.

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Contingency table for observed values

Degrees of freedom for contingency table is given as (r-1) * (c-1) where r,c are rows and columns. Here df = (2–1) * (2–1) = 1.

In the above table we have figured out all observed values and our next steps are to find expected values, get the Chi-Square value and check for relationship.

3. Find the Expected Value

- Based on the null hypothesis that the two variables are independent.
- The formula for estimated value for each cell is the total for rows multiplied by the total for the columns, divided by the total for the table, or simply.
- Expected values in each cell = (Row total * Column total)/Table total

E = (82*216) / 400
E = 44.28≈44

E = (82*184) / 400
E = 37.72 ≈38

E = (318*216) / 400
E= 171.72 ≈172

E = (318*184) / 400
E= 146.28 ≈146

Exited\ Gender	Yes	No	Total
Male	38	178	216
Female	44	140	184
Total	82	318	400

Exited\Gender	Yes	No
Male	44	172
Female	38	146

4. Calculate Chi-Square value

Summarizing the observed values and calculated expected values into a table and determine the Chi-Square value.

Gender,Exited	O	E	O-E	Square of O-E	(Square of O-E) / E
Male,Yes	38	44	-6	36	0.818181818
Male,No	178	172	6	36	0.209302326
Female,Yes	44	38	6	36	0.947368421
Femal,No	140	146	-6	36	0.246575342
Chi Square Value					2.22142987

In the above table,
O – Observed values
E – Expected values

We can see Chi-Square is calculated as 2.22 by using the Chi-Square statistic formula.

5. Accept or Reject the Null Hypothesis

With 95% confidence that is alpha = 0.05, we will check the calculated Chi-Square value falls in the acceptance or rejection region.

Having degrees of freedom =1(calculated with contingency table) and alpha =0.05 the Chi-Square value is 3.84.

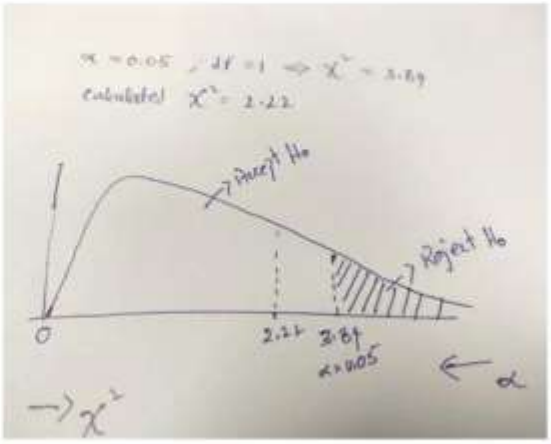
The Chi-Square values can be determined with the Chi-Square table.

The chi-square distribution is the right side since the difference in Observed and Expected is large.

In the fig, we can see Chi-Square ranges from 0 to inf and alpha ranges from 0 to 1 in the opposite direction.

We will reject the Null hypothesis if Chi-Square value falls in the error region (alpha from 0 to 0.05).

So here we are accepting the null hypothesis since the Chi-Square value is less than the critical Chi-Square value.



Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Chi Square Table

Limitations

- Chi-Square is sensitive to small frequencies in cells of tables. Generally, when the expected value in a cell of a table is less than 5, chi-square can lead to errors in conclusions.
- The chi-square test is sensitive to sample size. Relationships may appear to be significant when they aren't simply because a very large sample is used.
- In addition, the chi-square test cannot establish whether one variable has a causal relationship with another. It can only establish whether two variables are related.

Explain Multi-Arm bandit algorithm?

- Multi-arm bandits offer an approach to testing, especially **web testing**, that allows explicit optimization and more rapid decision making than the traditional statistical approach to designing experiments.
- A traditional A/B test involves data collected in an experiment, according to a specified design, to answer a specific question such as, **"Which is better, treatment A or treatment B?"** The presumption is that once we get an answer to that question, the experimenting is over and we proceed to act on the results.

Key Terms for Multi-Arm Bandits

Multi-arm bandit
An imaginary slot machine with multiple arms for the customer to choose from, each with different payoffs, here taken to be an analogy for a multitreatment experiment.

Arm
A treatment in an experiment (e.g., "headline A in a web test").

Win
The experimental analog of a win at the slot machine (e.g., "customer clicks on the link").

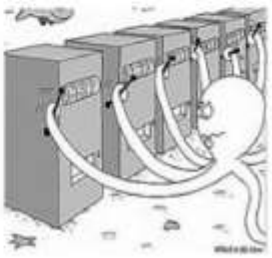
- **Bandit algorithms**, which are very **popular in web testing**, allow you to test multiple treatments at once and reach conclusions faster than traditional statistical designs.
- They take their name from slot machines used in gambling, also termed one-armed bandits (since they are configured in such a way that they extract money from the gambler in a steady flow).
- If you imagine a slot machine with more than one arm, each arm paying out at a different rate, you would have a multi-armed bandit, which is the full name for this algorithm.
- Here is one simple algorithm, the **epsilon-greedy algorithm for an A/B test**:
  1) Generate a uniformly distributed random number between 0 and 1.
  2) If the number lies between 0 and epsilon (where epsilon is a number between 0 and 1, typically fairly small), flip a fair coin (50/50 probability), and:
     a. If the coin is heads, show offer A.
     b. If the coin is tails, show offers B.
  3) If the number is ≥ epsilon, show whichever offer has had the highest response rate to date
- Epsilon is the single parameter that governs this algorithm. If epsilon is 1, we end up with a standard simple A/B experiment (random allocation between A and B for each subject). If epsilon is 0, we end up with a purely greedy algorithm one that chooses the best available immediate option (a local optimum). It seeks no further experimentation, simply assigning subjects (web visitors) to the best-performing treatment.
- Bandit algorithms can efficiently handle 3+ treatments and move toward optimal selection of the "best."
- For traditional statistical testing procedures, the complexity of decision making for 3+ treatments far outstrips that of the traditional A/B test, and the advantage of bandit algorithms is much greater.

- **Examples;**

Multi-Armed Bandit Problem (MABP)?

- A bandit is defined as someone who steals your money. A one-armed bandit is a simple slot machine wherein you insert a coin into the machine, pull a lever, and get an immediate reward. But why is it called a bandit? It turns out all casinos configure these slot machines in such a way that all gamblers end up losing money!
- A multi-armed bandit is a complicated slot machine wherein instead of 1, there are several levers which a gambler can pull, with each lever giving a different return. The probability distribution for the reward corresponding to each lever is different and is unknown to the gambler.

The task is to identify which lever to pull in order to get maximum reward after a given set of trials. This problem statement is like a single step Markov decision process. Each arm chosen is equivalent to an action, which then leads to an immediate reward. The below table shows the sample results for a 5-armed Bernoulli bandit with arms labelled as 1, 2, 3, 4 and 5:



Arm	Reward
1	0
2	0
3	1
4	1
5	0
3	1
3	1
2	0
1	1
4	0
2	0

This is called Bernoulli, as the reward returned is either 1 or 0. In this example, it looks like the arm number 3 gives the maximum return and hence one idea is to keep playing this arm in order to obtain the maximum reward (pure exploitation).

Use Cases

Bandit algorithms are being used in a lot of research projects in the industry.

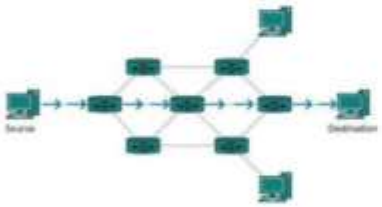
Clinical Trials

The well being of patients during clinical trials is as important as the actual results of the study. Here, exploration is equivalent to identifying the best treatment, and exploitation is treating patients as effectively as possible during the trial.

Clinical Trials

Network Routing

Routing is the process of selecting a path for traffic in a network, such as telephone networks or computer networks (internet). Allocation of channels to the right users, such that the overall throughput is maximised, can be formulated as a MABP.



Network Routing

Online Advertising

The goal of an advertising campaign is to maximise revenue from displaying ads. The advertiser makes revenue every time an offer is clicked by a web user. Similar to MABP, there is a trade-off between exploration, where the goal is to collect information on an ad's performance using click-through rates, and exploitation, where we stick with the ad that has performed the best so far.

Online Ads

Game Designing

Building a hit game is challenging. MABP can be used to test experimental changes in game play/interface and exploit the changes which show positive experiences for players.

Game Designing

Explain Power and Sample Size with its key terms?

- **If you run a web test, how do you decide how long it should run** (i.e., how many impressions per treatment are needed)? Despite what you may read in many guides to web testing, there is no good general guidance—it depends, mainly, on the frequency with which the desired goal is attained.

Key Terms for Power and Sample Size

Effect size
The minimum size of the effect that you hope to be able to detect in a statistical test, such as "a 20% improvement in click rates."

Power
The probability of detecting a given effect size with a given sample size.

Significance level
The statistical significance level at which the test will be conducted.

- Sample size refers to the number of participants or observations included in a study. This number is usually represented by n. The most common use of power calculations is to estimate how big a sample you will need.
- The size of a sample influences two statistical properties:
    1) the precision of our estimates and
    2) the power of the study to draw conclusions.
- Power is the probability of correctly rejecting the null hypothesis that sample estimates (e.g. Mean, proportion, odds, correlation co-efficient etc.) does not statistically differ between study groups in the underlying population.
- Large values of power are desirable, at least 80%, is desirable given the available resources and ethical considerations.
- Power proportionately increases as the sample size for study increases. Accordingly, an investigator can control the study power by adjusting the sample size and vice versa.
- Power analysis involves taking these three considerations, adding subject-area knowledge, and managing tradeoffs to settle on a sample size.
- During this process, you must rely heavily on your expertise to provide reasonable estimates of the input values.
- Power analysis helps you manage an essential tradeoff. As you increase the sample size, the hypothesis test gains a greater ability to detect small effects. This situation sounds great.
- However, larger sample sizes cost more money. And, there is a point where an effect becomes so minuscule that it is meaningless in a practical sense.