**What is the difference between Bernoulli trials and Binomial distributions?**

A **binomial distribution** is a discrete probability distribution for an experiment (with **two possible outcomes**) that is repeated multiple times. For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail. The coin toss experiment is repeated 20 time

A **Bernoulli distribution**is a discrete probability distribution for a Bernoulli trial — a random experiment that has only two outcomes (usually called a “Success” or a “Failure”). For example, the probability of getting a heads (a “success”) while flipping a coin is 0.5. The probability of “failure” is 1 – P (1 minus the probability of success, which also equals 0.5 for a coin toss).

**It is a special case of the binomial distribution for n = 1. Binomial distribution is for multiple (n) number of trials.**

**What steps do you need to take while sampling from a population to make sure that you’re inferring correct results for the population?**

To make a good inference about the population and to ensure data is free of bias, we would need a random representative sample.

A Representative sample is a group or set chosen from a larger statistical population that adequately replicates the larger group according to whatever characteristic or quality is under study. Some examples include sex, age, education level, socioeconomic status (SES), or marital status.

**Why do we consider using inferential statistics when we have understood a data sample using descriptive statistics?**

Descriptive statistics describes or summarises (mean, std. dev etc. of) the **sample** data and finds if there are relations between variables.

while

Inferential statistics allows us to infer some parameters about the **population** data from the available sample data by making use of the Central Limit Theorem.

**Imagine, you have the height of 100 employees of your company. Using this information, how would you estimate the mean height of all the employees of your company?**

We have a **Sample** of size 100 from which we find the sample mean and std. dev. Using the Central Limit Theorem which allows us to assume a normal sampling distribution, we can estimate the mean height of the **population** - all the employees of the company for a particular confidence level y (say 95%). The sample should be a random representative sample, i.e. representative of the variability in the data.

So we can say the **population mean** will lie in the range *Sample mean – MoE* and *Sample mean + MoE* with y% confidence

Where MoE = Z\* (Z-score corresponding to the confidence level) X Sample std. dev/SQRT(sample size)

**How do you check the significance of the results obtained in hypothesis testing?**

As part of the hypothesis testing, we can test a claim or hypothesis (H0) about a population parameter – either reject it or fail to reject it (using a sample data) based on a considered significance value.

In the **critical value method**, we calculate the critical and acceptance regions from the critical value Zc (corresponding to the significance value).

Population parameter (from H0) +- (Zc x sample std.dev)

If the sample statistic falls in the critical region we reject the H0 and if it falls in the acceptance region we fail to reject the H0.

In the **p-value method**, we calculate the Z-score corresponding to the sample statistic. If the probability corresponding to the Z-score is above the alpha, then we reject the H0. Otherwise we fail to reject the claim.

**Let’s assume the following hypothesis experiment**

**Null hypothesis: Average battery life of life-saving equipment is less than or equal to 1 year**

**Alternate hypothesis: it is more than 1 year.**

**Which type of error would you like to minimise here - Type I or type II error? In this case, should you use a higher confidence interval or lower?**

Type1 error as it would mean we would reject the Null hypothesis – which says life-saving equipment < 1 year. Type1 error would mean that the life-saving equipment could be used for longer than 1 year and that could lead to loss of lives. More of these errors would mean more lives are at risk.

In the other scenario (Type2 errors) we would fail to reject Null hypothesis which means that we would use the equipment for lesser time than it can support. This will not put any lives at risk.

For this we should use a higher confidence level.

**Let's say you work for an e-commerce company and want to check whether changing the colour of the Go To Cart button of your website from blue to green increases the conversion of visitors to become customers or not. How would you assess this?**

Perform A/B Testing.

Control Group: No change. GTC button will be blue color

Target Group: Button color will change. GTC button will be green color.

In whichever group the conversion rate is higher that color would be the choice

REDO Include 2-sample proportion testing, hypothesis testing

**In a dataset containing admission information of college students, we have two categorical variables, the gender of a student -male and female and the college the student has got admission in i.e. College A and College B. How will you test whether there is a significant association between these variables - in other words, we wish to determine whether the gender of the student determines which college the student has got admission in?**

Take a representative sample of students (student, gender, College)

Hypothesis testing:

H0: College-Gender 4 combinations have equal chances

H1: College-Gender 4 combinations do not have equal chances

Calculate Chi-squared statistic = Sum of square of [(Actual – Expected)/Expected]

If the probability associated with the calculated chi-squared statistic < significance value, then we reject the H0 else we fail to reject H0

**If we want to check whether the mean salaries of data analysts working in three different companies A, B and C are equal or not, which test will we use and why?**

**Why do we not perform multiple independent t-tests to check if there is a statistically significant difference between three or more group means instead of an ANOVA?**