**1. What is central limit theorem and why is it important?**

**Ans: -** The CLT is a statistical theory that states that - **if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean**.

The Central Limit Theorem is important for statistics because **it allows us to safely assume that the sampling distribution of the mean will be normal in most cases**. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will see in the next section.

**2. What is sampling? How many sampling methods do you know?**

**Ans: -** Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.

**Simple random sampling**: This is the purest form of probability sampling. In simple random sampling, individuals are chosen from a whole population at random. These individuals could be assigned numbers and then a random number generator selects from among these numbers. This is effectively how telephone surveying works.

**Systematic sampling**: Another type of probability sampling, systematic sampling picks respondents from a larger population by choosing them at regular intervals. This is the method of sampling used when a researcher picks every "nth" person in a group to be part of a study. By creating a consistent sampling interval (or studying every seventh person in a group, for example), the statisticians can get a manageable sample size that should still be representative of the entire population.

**Cluster sampling**: Cluster sampling begins by dividing a total population into clusters that have similar characteristics. Researchers then pick a small number of randomly selected clusters to study further. For instance, if researchers using cluster sampling wanted to study elementary schoolers in a district, they could make each individual school a cluster and pick three of those schools to study. They can also enact multistage sampling where individual students from those schools get randomly selected for further analysis.

**3. What is the difference between type1 and type II error?**

**Ans: -** A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

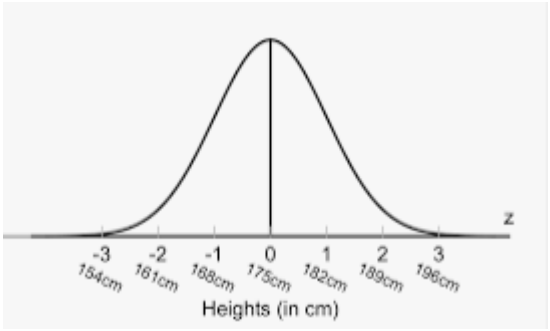
In statistical hypothesis testing, a type I error is the mistaken rejection of an actually true null hypothesis (also known as a "false positive" finding or conclusion; example: "an innocent person is convicted"), while a type II error is the failure to reject a null hypothesis that is actually false (also known as a "

Specifically, they can make either Type I or Type II errors. As you analyse your own data and test hypotheses, understanding the difference between Type I and Type II errors is extremely important, because **there's a risk of making each type of error in every analysis, and the amount of risk is in your control**.

**4. What do you understand by the term Normal distribution?**

**Ans: -** 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.

In a normal distribution, **data are symmetrically distributed with no skew**. Most values cluster around a central region, with values tapering off as they go further away from the center. The measures of central tendency (mean, mode, and median) are exactly the same in a normal distribution.




**5. What is correlation and covariance in statistics?**

Ans: - Correlation is **a statistical measure that expresses the extent to which two variables are linearly related** (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

Covariance **provides insight into how two variables are related to one another**. More precisely, covariance refers to the measure of how two random variables in a data set will change together. A positive covariance means that the two variables at hand are positively related, and they move in the same direction.

**6. Differentiate between univariate, Bivariate and multivariate analysis.**

**Ans: - Univariate Analysis**

Univariate analysis is the simplest form of data analysis where the data being analysed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

You can think of the variable as a category that your data falls into. One example of a variable in univariate analysis might be "age". Another might be "height". Univariate analysis would not look at these two variables at the same time, nor would it look at the relationship between them.

Some ways you can describe patterns found in univariate data include looking at mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation. Additionally, some ways you may display univariate data include frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.

**Bivariate Analysis**

Bivariate analysis is used to find out if there is a relationship between two different variables. Something as simple as creating a scatterplot by plotting one variable against another on a Cartesian plane (think X and Y axis) can sometimes give you a picture of what the data is trying to tell you. If the data seems to fit a line or curve then there is a relationship or correlation between the two variables. For example, one might choose to plot caloric intake versus weight.

**Multivariate Analysis**

Multivariate analysis is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on your goals. Some of these methods include:

- Additive Tree
- Canonical Correlation Analysis
- Cluster Analysis
- Correspondence Analysis / Multiple Correspondence Analysis
- Factor Analysis

**7. What do you understand by sensitivity and how would you calculate it?**

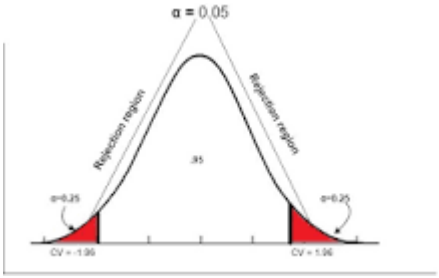
**Ans: -** The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as **sensitive analysis**. It's usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond's price.

The sensitivity is calculated by **dividing the percentage change in output by the percentage change in input**.

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Ans: - Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis. Hypothesis testing is formulated in terms of two hypotheses: **H0: the null hypothesis; • H1: the alternate hypothesis**.

What is H0 and H1 What is H0 and H1 for two tail test?



Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In our case it will be the average height of students in the batch is 100. Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.

**9. What is quantitative data and qualitative data?**

**Ans: -** Quantitative data is **data expressing a certain quantity, amount or range**. Usually, there are measurement units associated with the data, e.g. metres, in the case of the height of a person. It makes sense to set boundary limits to such data, and it is also meaningful to apply arithmetic operations to the data.

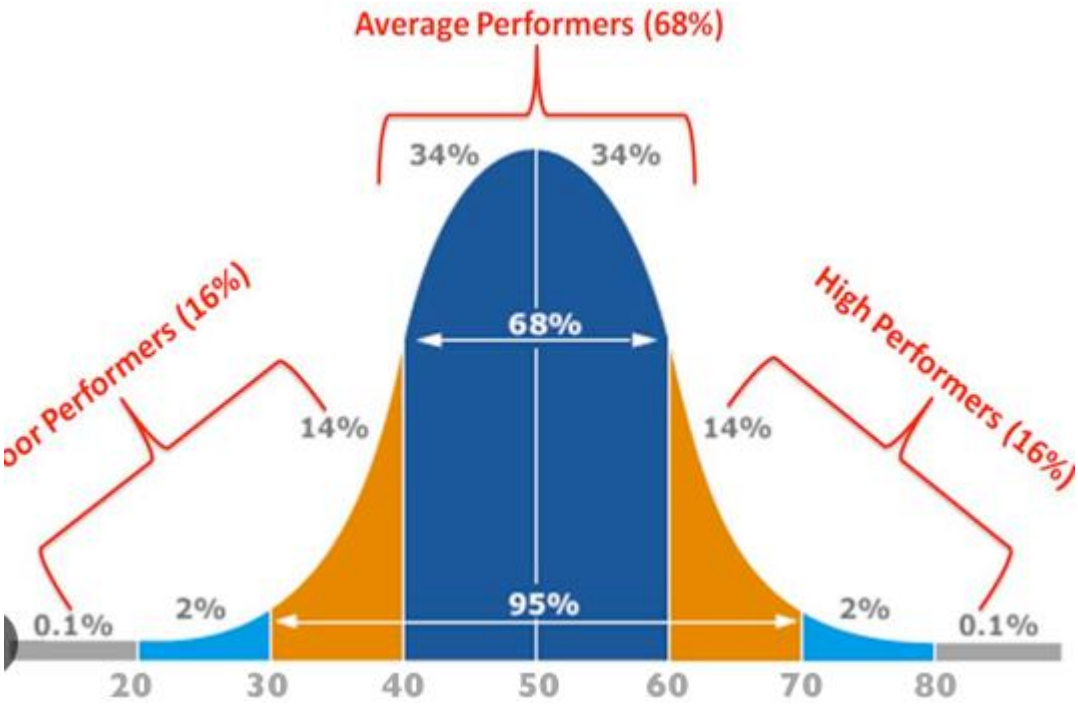
What is Qualitative Data? Qualitative data is the **descriptive and conceptual findings collected through questionnaires, interviews, or observation**. Analyzing qualitative data allows us to explore ideas and further explain quantitative results.

**10. How to calculate range and interquartile range?**

**Ans: -** The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), **first find the median (middle value) of the lower and upper half of the data**. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

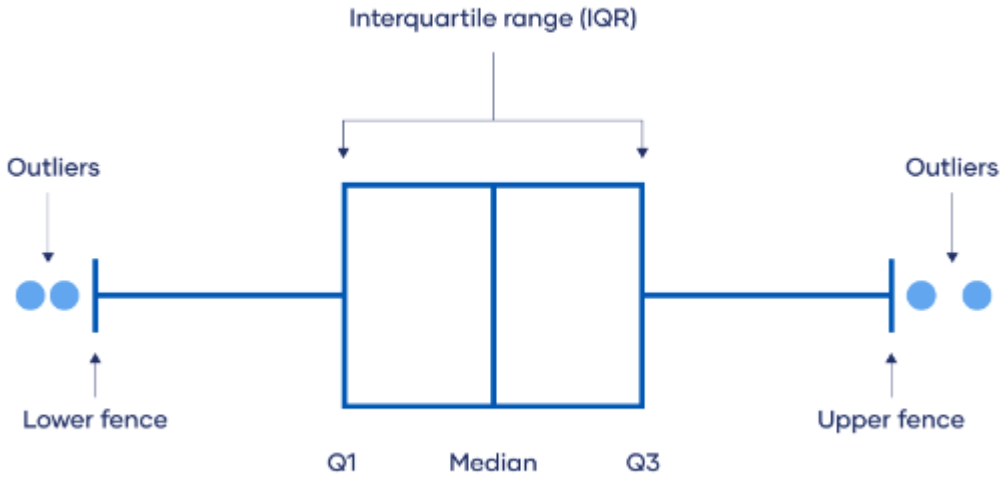
**11. What do you understand by bell curve distribution?**

**Ans: -** A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.




**12. Mention one method to find outliers.**

**Ans: -** Statistical outlier detection involves **applying statistical tests or procedures to identify extreme values**. You can convert extreme data points into z scores that tell you how many standard deviations away they are from the mean. If a value has a high enough or low enough z score, it can be considered an outlier.



**13. What is p-value in hypothesis testing?**

**Ans: -** The p-value is defined as **the probability of obtaining the result at least as extreme as the observed result of a statistical hypothesis test, assuming that the null hypothesis is true**.

| P-value | Decision |
|---|---|
| P-value > 0.05 | The result is not statistically significant and hence don't reject the null hypothesis. |
| P-value < 0.05 | The result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis. |
| P-value < 0.01 | The result is highly statistically significant, and thus rejects the null hypothesis in favour of the alternative hypothesis. |

The p value is **a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true**. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

14. What is the Binomial Probability Formula?

**Ans:-** Binomial probability refers to **the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes** (commonly called a binomial experiment).

| **Binomial Distribution Formula** | |
|---|---|
| Binomial Distribution | $P(x) = {}^{n}C_{x} \cdot p^{x} (1 - p)^{n-x}$ |
| Or, | $P(r) = [n!/r!(n-r)!] \cdot p^{r} (1 - p)^{n-r}$ |

Where,

- n = Total number of events
- r (or) x = Total number of successful events.
- p = Probability of success on a single trial.
- ${}^{n}C_{r} = [n!/r!(n-r)]!$
- 1 – p = Probability of failure.

We know that the binomial probability distribution is

$$P(r) = {}^{n}C^{r} \cdot p^{r} (1 - p)^{n-r}.$$

**15. Explain ANOVA and it's applications.**

**Ans: -** ANOVA is **helpful for testing three or more variables**. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.

**ANOVA assumes that the data is normally distributed**. The ANOVA also assumes homogeneity of variance, which means that the variance among the groups should be approximately equal. ANOVA also assumes that the observations are independent of each other.