STATISTICS

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

The central limit theorem states that if we have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample mean is asymptotically normal. We can calculate the mean of the sample means for the random samples we choose from the population:
μX¯=μ
As well as the standard deviation of sample means:
σX¯=σn
According to the central limit theorem, the form of the sampling distribution will approach normalcy as the sample size is sufficiently large (usually n>30). regardless of the population distribution.
Importance of Central Limit Theorem:
This is useful since the researcher never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from a population, the sample means will cluster together, allowing the researcher to obtain a very accurate estimate of the population mean.

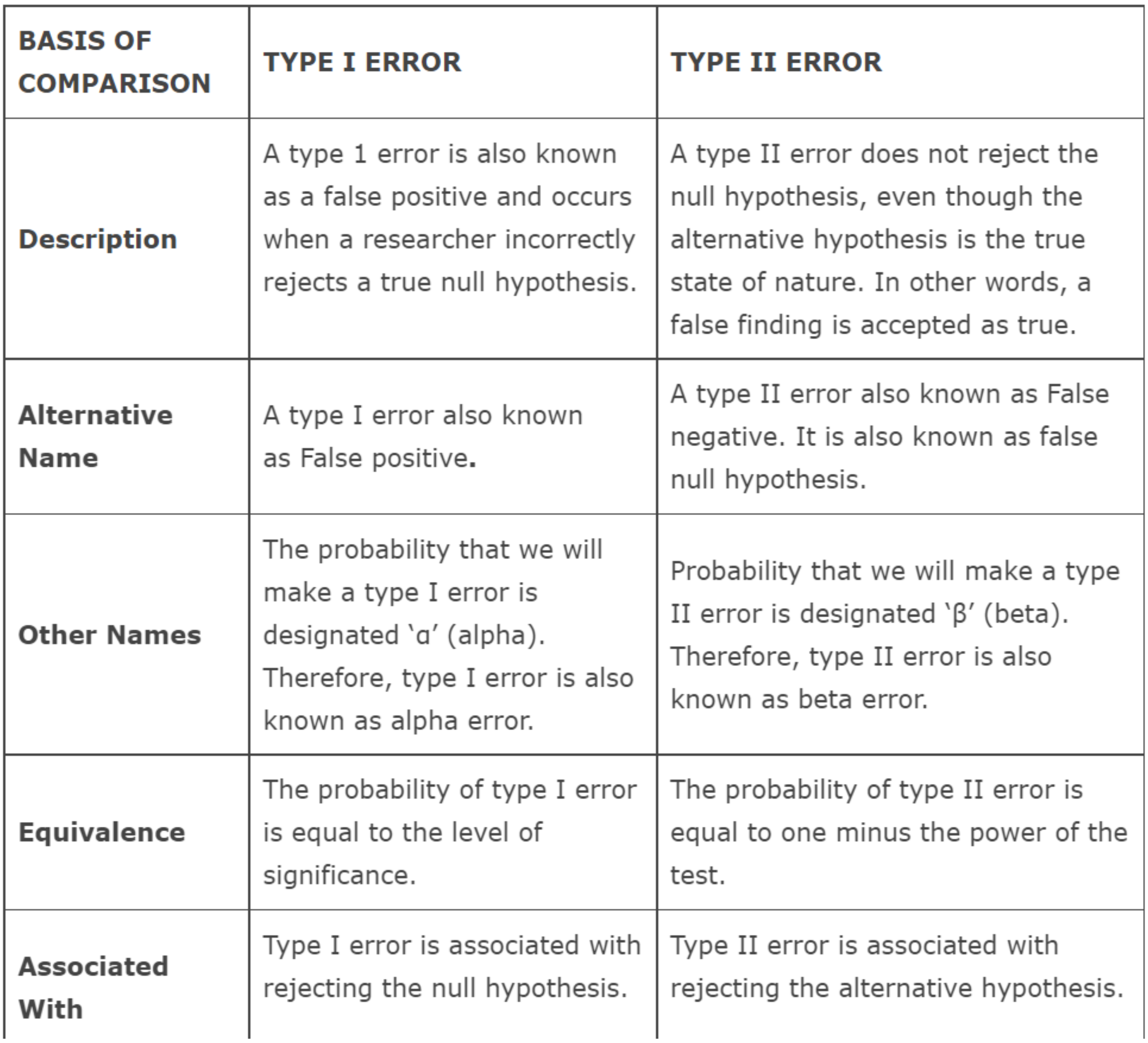
2. What is sampling? How many sampling methods do you know? The sampling method or sampling technique is the process of studying the population by gathering information and analysing that data. It is the basis of the data where the sample space is enormous. There are several different sampling techniques available, and they can be subdivided into two groups. All these methods of sampling may involve specifically targeting hard or approach to reach groups.

Types of Sampling Method: There are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are:

Probability Sampling: It involves random selection, allowing us to make strong statistical inferences about the whole group.

Non-probability Sampling: It involves non-random selection based on convenience or other criteria, allowing us to easily collect data.

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



BASIS OF COMPARISON	TYPE I ERROR	TYPE II ERROR
Description	A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis.	A type II error does not reject the null hypothesis, even though the alternative hypothesis is the true state of nature. In other words, a false finding is accepted as true.
Alternative Name	A type I error also known as False positive**.**	A type II error also known as False negative. It is also known as false null hypothesis.
Other Names	The probability that we will make a type I error is designated 'α' (alpha). Therefore, type I error is also known as alpha error.	Probability that we will make a type II error is designated 'β' (beta). Therefore, type II error is also known as beta error.
Equivalence	The probability of type I error is equal to the level of significance.	The probability of type II error is equal to one minus the power of the test.
Associated With	Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.

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

The Normal Distribution is the most significant continuous probability distribution for independent, randomly generated variables. It is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics.

Normal Distribution Formula:

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

Where

μ = Mean

σ = Standard deviation

x = Normal random variable

5. What is correlation and covariance in statistics?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It’s a common tool for describing simple relationships without making a statement about cause and effect. It is used to test relationships between quantitative variables or categorical variables. In other words, it’s a measure of how things are related. The study of how variables are correlated is called correlation analysis. A correlation coefficient is a way to put a value to the relationship. Correlation coefficients have a value of between -1 and 1. A “0” means there is no relationship between the variables at all, while -1 or 1 means that there is a perfect negative or positive correlation

6. Differentiate between univariate ,Biavariate,and multivariate analysis.

Univariate analysis

Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn’t deal with a causes or effect relationships then a Univariate analysis technique is used.

Bivariate analysis

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique.

Multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

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

Hypothesis Testing is a type of statistical analysis in which we put our assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

Types of Hypothesis Testing

Null Hypothesis: It is denoted by symbol H0. The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.

Alternate Hypothesis: It is denoted by symbol H1. The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis.

Two-Tailed Hypothesis Testing: In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two-sided.

9. What is quantitative data and qualitative data?

Quantitative data is anything that can be counted or measured; it refers to numerical data.

Qualitative data is descriptive, referring to things that can be observed but not measured—such as colours or emotions.

10. How to calculate range and interquartile range?

To calculate the range, we need to find the maximum value of a variable and subtract the minimum value. The range only takes into account these two values and ignore the data points between the two extremities of the distribution. It's used as a supplement to other measures, but it is rarely used as the sole measure of dispersion because it’s sensitive to extreme values

Formula: IQR = Q3 - Q1

11. What do you understand by bell curve distribution ?

A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode, and median in this case), while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

KEY TAKEAWAYS

- A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell.
- The top of the curve shows the mean, mode, and median of the data collected.
- Its standard deviation depicts the bell curve's relative width around the mean.
- Bell curves (normal distributions) are used commonly in statistics, including in analyzing economic and financial data.

12. Mention one method to find outliers.

Outliers are data points that are far from other data points. In other words, they’re unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

Calculate the interquartile range

The interquartile range (IQR) measures the dispersion of the data points between the first and third quartile marks. The general rule for using it to calculate outliers is that a data point is an outlier if it is over 1.5 times the IQR below the first quartile or 1.5 times the IQR above the third quartile.

To calculate the IQR, you need to know the percentile of the first and third quartile. The median of the upper half of the data set is the percentile for the third quartile, and the median of the lower half of the data set is the percentile for the first quartile.

To find the IQR, you subtract the first quartile from the third quartile:

IQR = Q3− Q1

where:

Q3 = the third quartile = the median of the upper half of the data set

Q1 = the first quartile = the median of the lower half of the data set

You can then use the IQR to find any outliers in your data set. The equations to calculate low or high outliers via the IQR range are:

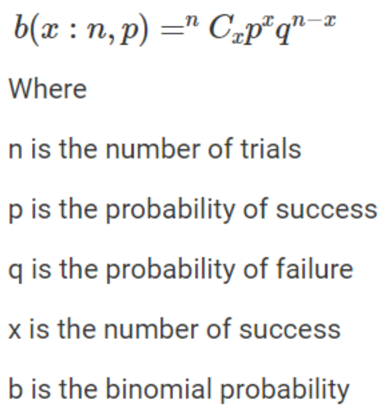
High outlier ≥ Q3 + (1.5 x IQR)

Low outlier ≤ Q1 − (1.5 x IQR)

13. What is p-value in hypothesis testing?

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis. P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage.

14. What is the Binomial Probability Formula?



$$b(x : n, p) = ^n C_x p^x q^{n-x}$$

Where

n is the number of trials

p is the probability of success

q is the probability of failure

x is the number of success

b is the binomial probability

15. Explain ANOVA and it’s applications.

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The Formula for ANOVA is: F= MST/MSE

where: F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to erro