

**STATISTICS WORKSHEET-4**

**Q1to Q15 are descriptive types. Answer in brief.**

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

The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases. This fact holds especially true for sample sizes over 30. Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ . Central Limit theorem states thats if n = no of sample size is large enough, it follows normal distribution. This is very important as you can compute many statistical properties which can be applied to many problems as it would be computed for a normal distribution. Once it holds true and you know underlying distribution is normal distribution, you compute mean and variance and so other statistical properties which would otherwise be very difficult if underlying distribution is unknown and you cannot make assumptions.

1. What is sampling? How many sampling methods do you know?

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in [market research](https://www.questionpro.com/blog/what-is-market-research/) so that they do not need to research the entire population to collect actionable insights. It is also a time-convenient and a cost-effective method and hence forms the basis of any [research design](https://www.questionpro.com/blog/research-design/). Sampling techniques can be used in a research survey software for optimum derivation.

Types of sampling: sampling methods

* + Probability sampling: [Probability sampling](https://www.questionpro.com/blog/probability-sampling/) is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.
  + Non-probability sampling: In [non-probability](https://www.questionpro.com/blog/non-probability-sampling/) sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

1. What is the difference between type1 and typeII error?



In statistical hypothesis testing, a type I error is the rejection of a 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 non-rejection of a false null hypothesis (also known as a "false negative" finding or conclusion; example: "a guilty person is not convicted").Much of statistical theory revolves around the minimization of one or both of these errors, though the complete elimination of either is a statistical impossibility for non-deterministic algorithms. By selecting a low threshold (cut-off) value and modifying the alpha (p) level, the quality of the hypothesis test can be increased. The knowledge of Type I errors and Type II errors is widely used in medical science, biometrics and computer science.Intuitively, type I errors can be thought of as errors of commission, i.e. the researcher unluckily concludes that something is the fact. For instance, consider a study where researchers compare a drug with a placebo. If the patients who are given the drug get better than the patients given the placebo by chance, it may appear that the drug is effective, but in fact the conclusion is incorrect. In reverse, type II errors as errors of omission. In the example above, if the patients who got the drug did not get better at a higher rate than the ones who got the placebo, but this was a random fluke, that would be a type II error. The consequence of a type II error depends on the size and direction of the missed determination and the circumstances. An expensive cure for one in a million patients may be inconsequential even if true.

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

Normal distribution, also known as the Gaussian distribution, is a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a [bell curve](https://www.investopedia.com/terms/b/bell-curve.asp).

* A normal distribution is the proper term for a probability bell curve.
* In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
* Normal distributions are symmetrical, but not all symmetrical distributions are normal.
* In reality, most pricing distributions are not perfectly normal.

1. What is correlation and covariance in statistics?

We know that variance measures the spread of a random variable, so Covariance measures how two random random variables vary together. Unlike Variance, which is non-negative, Covariance can be negative or positive (or zero, of course). A positive value of Covariance means that two random variables tend to vary in the same direction, a negative value means that they vary in opposite directions, and a 0 means that they don’t vary together. If two random variables are independent, their Covariance is 0, which makes sense because they don’t affect each other and thus don’t vary together (this relation doesn’t necessarily hold in the opposite direction,). For two random variables X and Y , you can define the Covariance Cov(X,Y) as:

Cov(X,Y)=E((X−E(X))(X−E(Y)))

Correlation is the Covariance divided by the standard deviations of the two random variables. Of course, you could solve for Covariance in terms of the Correlation; we would just have the Correlation times the product of the Standard Deviations of the two random variables. Consider the Correlation of a random variable with a constant. We know, by definition, that a constant has 0 variance  our mathematical definition is as follows for random variables XX and YY:

ρ=Corr(X,Y)=Cov(X,Y)/σ(x)σ(y)

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

Univariate analysis is the simplest form of data analysis where the data being analyzed 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.

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 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

Generalized Procrustean Analysis

MANOVA

Multidimensional Scaling

Multiple Regression Analysis

Partial Least Square Regression

Principal Component Analysis / Regression / PARAFAC

Redundancy Analysis.

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

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. There are different methods to carry out the sensitivity analysis:

* Modeling and simulation techniques
* Scenario management tools through Microsoft excel

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

The major purpose of hypothesis testing is to choose between two competing hypotheses about the value of a population parameter. For example, one hypothesis might claim that the wages of men and women are equal, while the alternative might claim that men make more than women. The hypothesis actually to be tested is usually given the symbol H0, and is commonly referred to as the null hypothesis. As is explained more below, the null hypothesis is assumed to be true unless there is strong evidence to the contrary – similar to how a person is assumed to be innocent until proven guilty. The other hypothesis, which is assumed to be true when the null hypothesis is false, is referred to as the alternative hypothesis, and is often symbolized by HA or H1. Both the null and alternative hypothesis should be stated before any statistical test of significance is conducted. In other words, you technically are not supposed to do the data analysis first and then decide on the hypotheses afterwards.

1. What is quantitative data and qualitative data?

Quantitative data is defined as the value of data in the form of counts or numbers where each data-set has an unique numerical value associated with it. This data is any quantifiable information that can be used for mathematical calculations and statistical analysis, such that real-life decisions can be made based on these mathematical derivations. Quantitative data is used to answer questions such as “How many?”, “How often?”, “How much?”. This data can be verified and can also be conveniently evaluated using mathematical techniques.

Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting [focus groups](https://www.questionpro.com/blog/focus-group/), and similar methods. Qualitative data in statistics is also known as categorical data – data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon

1. How to calculate range and interquartile range?

Range = highest-lowest

Interquartile Range Formula for a given set of data can be expressed as:

IQR = Q3 - Q1

where,

IQR = Interquartile range

Q1 = First Quartile

Q3 = Third Quartile

1. 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](https://www.investopedia.com/terms/n/normaldistribution.asp) consists of a symmetrical bell-shaped curve.

* 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.

1. Mention one method to find outliers.

An outlier is defined as being any point of data that lies over 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3)in a data set.

High = (Q3) + 1.5 IQR

Low = (Q1) – 1.5 IQR

1. What is p-value in hypothesis testing?

In statistical hypothesis testing, the p-value or probability value is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical summary (such as the absolute value of the sample mean difference between two compared groups) would be greater than or equal to the actual observed results*.*

1. What is the Binomial Probability Formula?

The Binomial Probability distribution of exactly x successes from n number of trials is given by the below formula-

P (X) = nCx px qn – x

Where,

n = Total number of trials

x = Total number of successful trials

p = probability of success in a single trial

q = probability of failure in a single trial = 1-p

15. Explain ANOVA and it’s applications

A common approach to figure out a reliable treatment method would be to analyse the days it took the patients to be cured. We can use a statistical technique which can compare these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples on the basis of their means, is called ANOVA.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

WORKSHEET