**STATISTICS WORKSHEET-4**

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

What is central limit theorem and why is it important?

Answer- **Statement of Central limit Theorem:**

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?

Answer- Sampling means **selecting the group that you will actually collect data from in your research**. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

 There are two primary types of sampling methods that you can use in your research:

* [**Probability sampling**](https://www.scribbr.com/methodology/probability-sampling/) involves random selection, allowing you to make strong statistical inferences about the whole group.
* [**Non-probability sampling**](https://www.scribbr.com/methodology/non-probability-sampling/) involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

3- What is the difference between type1 and typeII error?

## Answer- Key Differences Between Type I and Type II Error

The points given below are substantial so far as the differences between type I and type II error is concerned:

1. Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
2. Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.
3. When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.
4. Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.
5. The probability of committing type I error is the sample as the level of significance. Conversely, the likelihood of committing type II error is same as the power of the test.
6. Greek letter ‘α’ indicates type I error. Unlike, type II error which is denoted by Greek letter ‘β’.

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

Answer- 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.

The normal distribution model is important in statistics and is key to the [Central Limit Theorem](https://www.investopedia.com/terms/c/central_limit_theorem.asp) (CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled.

5- What is correlation and covariance in statistics?

Answer- Statistical analysis is an excellent method for mathematically exploring the relationships between different variables. Two important statistical methods for analyzing data include **covariance** and **correlation**. In general terms, covariance describes the difference between two variables, while correlation looks at the relationship between two variables.

Covariance and standard deviation are easy to be confused, as are correlation and standard deviation. However, each term - covariance, correlation, and standard deviation- represents distinct statistical concepts. Covariance specifically examines whether a change in one variable produces a change in a second variable. While covariance shows how two variables are related to each other, it does not show the strength of the relationship between the two variables. This is where correlation is important. Correlation looks at the strength of the relationship between two variables. To calculate the correlation between two variable, the covariance value is divided by the standard deviation of both variables. **Standard deviation** is a measure of the amount of variation that is found in a data set.

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

**Univarate 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.

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.

**Bivarate 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
* Generalized Procrustean Analysis
* MANOVA
* Multidimensional Scaling
* Multiple Regression Analysis
* Partial Least Square Regression
* Principal Component Analysis / Regression / PARAFAC
* Redundancy Analysis.

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

Answer - *Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions. It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result.*

The formula for sensitivity analysis is basically a financial model in excel where the analyst is required to identify the key variables for the output formula and then assess the output based on different combinations of the independent variables.

Mathematically, the dependent output formula is represented as,

**Z = X2 + Y2**

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

Answer- Hypothesis testing is a tool for making statistical inferences about the population data. It is an analysis tool that tests assumptions and determines how likely something is within a given standard of accuracy. Hypothesis testing provides a way to verify whether the results of an experiment are valid.

A null hypothesis and an alternative hypothesis are set up before performing the hypothesis testing. This helps to arrive at a conclusion regarding the sample obtained from the population. In this article, we will learn more about hypothesis testing, its types, steps to perform the testing, and associated examples.

In hypothesis testing there are two mutually exclusive hypotheses; **the Null Hypothesis (H0) and the Alternative Hypothesis (H1)**. One of these is the claim to be tested and based on the sampling results (which infers a similar measurement in the population), the claim will either be supported or not.

Two-tailed Test **H0 : µ = k H1 : µ 6= k** P-value = 2P(z > |t|) If P-value  α, we reject H0. If P-value > α, we do not reject H0. Note: For each formula to find z-scores, if you can assume that x has a normal distribution, then any sample size n will work. If you cannot assume this, use a sample size n 30.

9- What is quantitative data and qualitative data?

Answer – Qualitative- Qualitative data is non-statistical and is typically unstructured or semi-structured. This data isn’t necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers.

Qualitative data can be used to ask the question “why.” It is investigative and is often open-ended until further research is conducted. Generating this data from qualitative research is used for theorizations, interpretations, developing hypotheses, and initial understandings.

Quantitative - Contrary to qualitative data, quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined. This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended. It can be used to ask the questions “how much” or “how many,” followed by conclusive information.

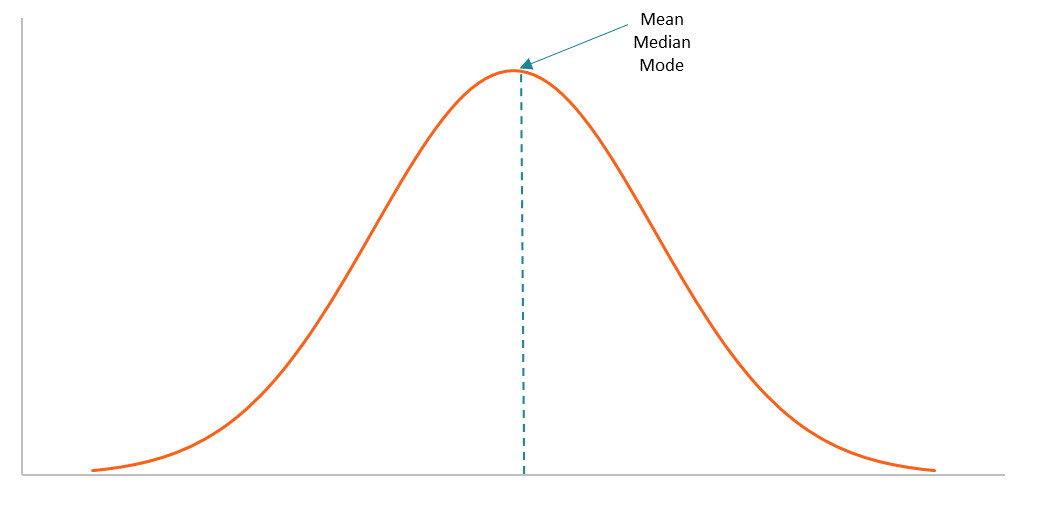
1. - How to calculate range and interquartile range?

Answer - To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). 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.

The interquartile range and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, you need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

11- What do you understand by bell curve distribution ?

Answer- A bell curve is the informal name of a graph that depicts a normal probability distribution. The term obtained its name due to the bell-shaped curve of the normal probability distribution graph.



However, the term is not quite correct because the normal probability distribution is not the only probability distribution whose graph shows a bell-shaped curve. For example, the graphs of the Cauchy and logistic distributions also demonstrate a bell-shaped curve.

12- Mention one method to find outliers.

Answer- **Standard Deviation:**

In statistics, If a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and **about 99.7%**lie within three standard deviations.

Therefore, if you have any data point that is more than 3 times the standard deviation, then those points are very likely to be anomalous or outliers.

The output of this code is a list of values above 80 and below -40. Notice that the dataset I am passing is a one-dimensional dataset. Now, let’s explore more advanced methods for multi-dimensional datasets.

13- What is p-value in hypothesis testing?

Answer- In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical [hypothesis test](https://www.investopedia.com/terms/h/hypothesistesting.asp), assuming that the [null hypothesis](https://www.investopedia.com/terms/n/null_hypothesis.asp) is correct. The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

P-value is often used to promote credibility for studies or reports by government agencies. For example, the United States Census Bureau stipulates any analysis with a p-value greater than 0.10 must be accompanied by a statement that the difference is not statistically different from zero. The Census Bureau also has standards in place stipulating which p-values are acceptable for various publications.

14- What is the Binomial Probability Formula?

Answer- The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a single outcome such as success/failure is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process. Consider an experiment where each time a question is asked for a yes/no with a series of n experiments. Then in the binomial probability distribution, the boolean-valued outcome the success/yes/true/one is represented with probability p and the failure/no/false/zero with probability q (q = 1 − p). In a single experiment when n = 1, the binomial distribution is called a Bernoulli distribution.

**What Is the Binomial Distribution Formula?**

The binomial distribution formula is for any random variable X, given by;  P(x:n,p) = nCxx px(1-p)n-x **Or** P(x:n,p) = nCx px (q)n-x

where,

* n = the number of experiments
* x = 0, 1, 2, 3, 4, …
* p = Probability of success in a single experiment
* q = Probability of failure in a single experiment (= 1 – p)

The binomial distribution formula is also written in the form of n-Bernoulli trials, where nCx = n!/x!(n-x)!. Hence, P(x:n,p) = n!/[x!(n-x)!].px.(q)n-x

15-Explain ANOVA and it’s applications.

Answer- An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.

Another Key part of ANOVA is that it splits the independent variable into 2 or more groups. For example, one or more groups might be expected to influences the dependent variable while the other group is used as a control group, and is not expected to influence the dependent variable.

There are two main types of ANOVA viz. one–way ANOVA and two–way ANOVA.

* **One Way ANOVA –** It is also known as one factor ANOVA. Here, we are using one criterion variable (or called as a factor) and analyze the difference between more than two sample groups. Suppose in glass industry, we want to compare the variation of three batches (glass) for their average weight (factor).
* **Two Way ANOVA –**Here, we are using two independent variables (factors) and analyze the difference between more than two sample groups. Similarly, we want to compare the variation of three batches of glass w.r.t weight and hardness (two factors)
* **Null Hypothesis**– It is a type of hypothesis which is presumed to be true i.e. no statistical significance exists in a set of given observations.
* **Alternative Hypothesis –** It is alternative or contradictory to the null hypothesis. There is a significant difference in the process.
* **Level of significance (α) –** It represents the probability of making the wrong decision when the null hypothesis is true.
* **p-value –** It is used in hypothesis testing for making a decision whether to accept or reject the null hypothesis.