STATISTICS WORKSHEET

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

Answer : The central limit theorem states that if you 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 means will be approximately normally distributed.

The theorem states that the sampling distribution of the mean approaches a normal distribution as the size of the sample increases, regardless of the shape of the original population distribution.

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 .

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

The CLT works from the center out. That implies on the off chance that you are presuming close to the center, for example, that around two-thirds of future totals will fall inside one standard deviation of the mean, you can be secure even with little samples.

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.

In statistics, sampling is a method when researchers determine a representative segment of a larger population that is then used to conduct a study. Sampling generally comes in two forms — probability sampling and non-probability sampling.

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are:

- Probability Sampling
- Non-probability Sampling

The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of the population.

The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.

3. What is the difference between type 1 and type II error?

Answer : Type -1 Error (Error of the first kind):

It is also known as a false-positive.

It occurs if the researcher rejects a correct null hypothesis in the population.

i.e., incorrect rejection of the null hypothesis.

Measured by alpha (significance level).

If the significance level is fixed at 5%,

It means there are about five chances of type – 1 error out of 100.

Cause of Type – 1 Error

The significance level is decided before testing the hypothesis

Sample size is not considered

This may occur due to chance

It can be reduced by decreasing the level of significance.

Type-2 Error(Error of second kind):

It is also known as a false negative.

It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis.

Measured by beta (the power of test).

The probability of committing a type -2 error is calculated by 1 – beta (the power of test).

Cause of Type-2 Error:

A statistical test is not powerful enough.

It is caused by a smaller sample size.

It may hide the significance level of the items being tested.

It can be reduced by increasing the level of significance.

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

Answer : 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.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena.

The normal distribution is often called the bell curve because the graph of its probability density looks like a bell. It is also known as called Gaussian distribution.

A normal distribution **comes with a perfectly symmetrical shape**. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve.

A normal distribution has a probability distribution that is centered around the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the center. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution.

Properties of a normal distribution:

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

5. What is correlation and covariance in statistics?

Answer: 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.

A positive correlation means that both variables change in the same direction. A negative correlation means that the variables change in opposite directions.

Correlation is what you are doing when you compare two sets of measurements (each set is called a variable). If you were to measure everyone's height and weight, you could then compare heights and weights and see if they have any relationship to each other.

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

Covariance can have both positive and negative values. Based on this, it has two types:

- Positive Covariance
- Negative Covariance

Covariance formula is a statistical formula, used to evaluate the relationship between two variables. It is one of the statistical measurements to know the relationship between the variance between the two variables. Let us say X and Y are any two variables, whose relationship has to be calculated. Thus the covariance of these two variables is denoted by Cov(X,Y).

If cov(X, Y) is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.

If cov(X, Y) is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.

If cov(X, Y) is zero, then we can say that there is no relation between two variables.

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

Answer: Univariate data –
This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

Bivariate data –
This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.Example of bivariate data can be temperature and ice cream sales in summer season.

Multivariate data –
When the data involves **three or more variables**, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved.Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA).

Univariate analysis looks at one variable, Bivariate analysis looks at two variables and their relationship. Multivariate analysis looks at more than two variables and their relationship.

Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable. There are many different ways people use univariate analysis.

Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other.

Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.

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

Answer: 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.

It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system. All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.

Sensitivity analysis is a financial model that determines how target variables are affected based on changes in other variables known as input variables. It is a way to predict the outcome of a decision given a certain range of variables.

Sensitivity analysis is used **to identify how much variations in the input values for a given variable impact the results for a mathematical model**. Sensitivity analysis can identify the best data to be collected for analyses to evaluate a project's return on investment (ROI).

Mathematically, this can be stated as:

Sensitivity=TP /TP+FN

Where

True Positive(TP) = The number of cases correctly identified as patient

False Negative(FN) = The number of cases incorrectly identified as healthy

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

Answer: Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

The test provides evidence concerning the plausibility of the hypothesis, given the data.

Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

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.

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. H0 : μ = 100.

Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.

9. What is quantitative data and qualitative data?

Answer: Qualitative data is information that cannot be counted, measured or easily expressed using numbers. It is collected from text, audio and images and shared through data visualization tools, such as word clouds, concept maps, graph databases, timelines and infographics.

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.

Quantitative data is fixed and Universal while qualitative data is subjective and dynamic. For example, if something weighs 20 kilograms, that can be considered an objective fact. However, two people may have very different qualitative accounts of how they experience a particular event.

Since quantitative data is defined as the value of data in the form of counts or numbers, each data set has a numerical value associated with it.

Examples of Quantitative Data:

- Weight in pounds.
- Length in inches.
- Distance in miles.
- Number of days in a year.
- A heatmap of a web page.

Examples of Qualitative Data

- Observation Notes.
- Semi-structured interviews. ...
- Open-ended survey.
- Participant diaries or journals.

10. How to calculate range and interquartile range?

Answer : In statistics, the range is the spread of your data from the lowest to the highest value in the distribution. It is a commonly used measure of variability.

Along with measures of central tendency, measures of variability give you descriptive statistics for summarizing your data set.

The range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution.

The formula to calculate the range is:

$$R = H - L$$

R = range

H = highest value

L = lowest value

The range is the easiest measure of variability to calculate. To find the range, follow these steps:

- Order all values in your data set from low to high.
- Subtract the lowest value from the highest value.

InterQuartile Range:

In descriptive statistics, the interquartile range tells you the spread of the middle half of your distribution.

Quartiles segment any distribution that's ordered from low to high into four equal parts. The interquartile range (IQR) contains the second and third quartiles, or the middle half of your data set.

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.

The interquartile range formula is the first quartile subtracted from the third quartile:

IQR = Q3 – Q1.

11. What do you understand by bell curve distribution ?

Answer : 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.

- 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.

Answer: Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests.

It’s important to carefully identify potential outliers in your dataset and deal with them in an appropriate manner for accurate results.

Sorting method:

You can sort quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.

This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

180 156 9 176 163 1827 166 171

You sort the values from low to high and scan for extreme values.

9 156 163 166 171 176 180 1872

13. What is p-value in hypothesis testing?

Answer: The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis 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.

- A p-value is a statistical measurement used to validate a hypothesis against observed data.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.
- A p-value of 0.05 or lower is generally considered statistically significant.
- P-value can serve as an alternative to—or in addition to—preselected confidence levels for hypothesis testing.

P-values are usually found using p-value tables or spreadsheets/statistical software. These calculations are based on the assumed or known probability distribution of the specific statistic tested. P-values are calculated from the deviation between the observed value and a chosen reference value, given the probability distribution of the statistic, with a greater difference between the two values corresponding to a lower p-value.

Mathematically, the p-value is calculated using integral calculus from the area under the probability distribution curve for all values of statistics that are at least as far from the reference value as the observed value is, relative to the total area under the probability distribution curve.

14. What is the Binomial Probability Formula?

Answer: The binomial distribution is a commonly used discrete distribution in statistics. The normal distribution as opposed to a binomial distribution is a continuous distribution. The binomial distribution represents the probability for 'x' successes of an experiment in 'n' trials, given a success probability 'p' for each trial at the experiment.

The binomial distribution formula is for any random variable X, given by;

$P(x:n,p) = {}^{n}Cxx\ p^{x}\,(1-p)^{n-x}$ **Or** $P(x:n,p) = {}^{n}C_{x}\ p^{x}\ (q)^{n-x}$

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 ${}^{n}C_{x} = n!/x!(n-x)!$. Hence, $P(x:n,p) = n!/[x!(n-x)!].p^{x}.(q)^{n-x}$

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.

15. Explain ANOVA and it's applications.

Answer: 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.

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

Applications:

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer >> 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. It is employed with subjects, test groups, between groups and within groups.

A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.