Solved Statistics Worksheet

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

Ans. The central limit theorem 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. The CLT also performs a significant part in statistical inference. It depicts precisely how much an increase in sample size diminishes sampling error, which tells us about the precision or margin of error for estimates of statistics, for example, percentages, from samples.

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

Ans. Sampling means selecting a group (a sample) from a population from which we will collect data for our research. Sampling is an important aspect of a research study as the results of the study majorly depend on the sampling technique used.

There are two types of sampling methods:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. Probability sampling involves the following methods-
    - Simple Random Sampling- In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population. To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.
    - Systematic Sampling- Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals. If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample.

    - Stratified Sampling- Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample. To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).
    - Cluster Sampling- Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups. If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called multistage sampling.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data. Probability sampling involves the following methods-
    - Convenience Sampling- A convenience sample simply includes the individuals who happen to be most accessible to the researcher. This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can’t produce generalizable results.
    - Voluntary Response Sampling- Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.
    - Purposive Sampling- This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research. It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion.
    - Snowball Sampling- If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The

number of people you have access to “snowballs” as you get in contact with more people.

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

Ans.

BASIS OF DIFFERENCE	TYPE I ERROR	TYPE II ERROR
Meaning	Type I error refers to non-acceptance of hypothesis which ought to be accepted.	Type II error is the acceptance of hypothesis which ought to be rejected.
Equivalent to	False positive	False negative
What is it?	It is incorrect rejection of true null hypothesis.	It is incorrect acceptance of false null hypothesis.
Represents	A false hit	A miss
Probability of committing error	Equals the level of significance	Equals the power of test.
Indicated by	Greek letter 'α'	Greek letter 'β'

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

Ans. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In a normal distribution, the mean, mode and median are all the same. Moreover, these values all represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation. In graphical form, the normal distribution appears as a "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.

5. What is correlation and covariance in statistics?

Ans. Correlation is a statistical measure that indicates how strongly two variables are related. Correlation is limited to values between the range -1 and +1. A change in scale has no effect on correlation. It is a unit-free measure.

Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. The value of covariance lies in the range of -∞ and +∞. A change in scale has effects on correlation. It is not a unit-free measure.

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

Ans.

Univariate Analysis	Biavariate Analysis	Multivariate Analysis
Univariate Analysis deals with data consisting of only one variable.	Biavariate Analysis deals with data involving two different variables.	Multivariate Analysis deals with data involving three or more variables
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.	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	It is similar to bivariate but contains more than one dependent variable.
The example of a univariate data can be height.	Example of bivariate data can be temperature and ice cream sales in summer season.	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.

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

Ans. The sensitivity of a test is its ability to determine the patient cases correctly. To estimate it, we should calculate the proportion of true positive in patient cases.

Mathematically, this can be stated as:

Sensitivity=TP/(TP+FN)

Where, TP is True Positive (the number of cases correctly identified as patient)

FN is False Negative (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?

Ans. Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.

The null hypothesis, H0, is a statistical proposition stating that there is no significant difference between a hypothesized value of a population parameter and its value estimated from a sample drawn from that population. The alternative hypothesis, H1 or Ha, is a statistical proposition stating that there is a significant difference between a hypothesized value of a population parameter and its estimated value

In two-tail test we consider Null Hypothesis H0 as the statement being tested. Here we claim about μ or historical value of μ, whereby the given null hypothesis is μ is equal to k, where k is a value of the mean given μ is the population mean. Similarly we consider Alternative Hypothesis H1 as the statement one will adopt in the situation in which evidence(data) is strong so H0 is rejected. Such hypothesis testing comes into picture when sample mean may be different from the population mean. In two tailed test μ is not equal to k, i.e., μ is different from the value stated in H0

9. What is quantitative data and qualitative data?

Ans. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often).

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code. Qualitative data are data about categorical variables (e.g. what type).

10. How to calculate range and interquartile range?

Ans. 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 gives a better idea of the dispersion of data. To calculate these two measures, one needs 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 (Q3) and lower(Q1) quartiles.

11. What do you understand by bell curve distribution ?

Ans. 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.

12. Mention one method to find outliers.

Ans. A method to detect outliers is the z score method. Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean. To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, the formula for that process is the following:

$$Z=(X - \mu)/ \sigma$$

The further away an observation's Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers are Z-scores of +/-3 or further from zero. In a population that follows the normal distribution, Z-score values more extreme than +/- 3 have a probability of 0.0027 (2 * 0.00135), which is about 1 in 370 observations. However, if the data does not follow the normal distribution, this approach might not be accurate.

13. What is p-value in hypothesis testing?

Ans. The *p*-value, or probability value, explains how likely it is the data could have occurred under the null hypothesis. It does this by calculating the likelihood of the test statistic, which is the number calculated by a statistical test using the data. The *p*-value tells how often one would expect to see a test statistic as extreme or more extreme than the one calculated by the statistical test if the null hypothesis of that test was true. The *p*-value gets smaller as the test statistic calculated from the data gets further away from the range of test statistics predicted by the null hypothesis. The *p*-value is a proportion: if the *p*-value is 0.05, that means that 5% of the time one would see a test statistic at least as extreme as the one found if the null hypothesis was true.

14. What is the Binomial Probability Formula?

Ans. 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-

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

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)

15.Explain ANOVA and it’s applications.

Ans. 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. ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1. The formula of ANOVA is-

$$F=MST/MSE$$

Where, F is ANOVA coefficient, MST is mean sum of squares due to treatments and MSE is mean sum of errors due to error.

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. If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

There are two main types of ANOVA: one-way (or unidirectional) and two-way. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. With a two-way ANOVA, there are two independents.

Applications- The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. A two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.