

Statistics Worksheet 4

QUES 1: What is central limit theorem and why is it important?

Answer. **The Central Limit Theorem** is a statistical theory that states that if we 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 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 is important because it performs a significant part in statistical inference. It depicts precisely how much an increase in sample size diminishes sampling error. Sample sizes equal to or greater than 30 are often considered sufficient for the Central Limit Theorem to hold.

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

Answer. Sampling is the action or process of taking samples of something for analysis. It is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. There are two primary types of sampling methods that you can use in your research:

1. **Probability Sampling**
2. **Non-Probability Sampling**

Probability Sampling is a sampling technique in which samples from a larger population are chosen using a method based on the theory of probability.

Non-probability sampling is a sampling technique in which the researcher selects samples based on the researcher's subjective judgment rather than random selection.

Probability Sampling Methods: They are as follows:

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Cluster Sampling

Non-Probability Sampling Methods: They are as follows:

1. **Convenience sampling**
2. **Quota Sampling**
3. **Purposive Sampling**
4. **Referral Sampling**

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

Answer. **Type – 1** error is known as false positive, i.e., when we reject the correct null hypothesis.

Type -2 error is also known as a false negative, i.e., when we fail to reject the false null hypothesis.

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

Answer. A **Normal Distribution** is the bell-shaped frequency distribution curve of a continuous random variable. The Normal Distribution, also called the Gaussian distribution. 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.

QUES 5. What is correlation and covariance in statistics?

Answer. **Correlation** is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

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.

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

Answer. **Univariate** statistics summarize only one variable at a time.

Bivariate statistics compare two variables.

Multivariate statistics compare more than two variables.

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

Answer. **Sensitivity** is the percentage of true positives (e.g. 90% sensitivity = 90% of people who have the target disease will test positive).

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

Answer. 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.

H0: The null hypothesis is generally denoted as **H0**. It states the exact opposite of what an investigator or an experimenter predicts or expects. It basically defines the statement which states that there is no exact or actual relationship between the variables.

H1: The alternative hypothesis, **H1**. It is a statistical proposition stating that there is a significant difference between a hypothesized value of a population parameter and its estimated value. When the null hypothesis is tested, a decision is either correct or incorrect.

In a two-tailed test, the generic null and alternative hypotheses are the following:

Null(H0): The effect equals zero. The null hypothesis here is what currently stated to be true about the population.

Alternative(H1): The effect does not equal zero.

The alternate hypothesis is always what is being claimed.

QUES 9. What is quantitative data and qualitative data?

Answer. **Quantitative data** is data expressing a certain quantity, amount or range. Usually, there are measurement units associated with the data, e.g. meters, 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. Example. Weight in Pounds, Length in inches, Distance in Miles.

Qualitative data is the descriptive and conceptual findings collected through questionnaires, interviews, or observation. Analyzing qualitative data allows us to explore ideas and further explain quantitative results. Example Photographs, Audio recordings, Video Recordings.

QUES 10. How to calculate range and interquartile range?

Answer. **The range** is calculated by subtracting the lowest value from the highest value.

Range= Highest Value- Lowest Value.

The interquartile range is the distance from the largest quartile to the smallest quartile.

IQR=Q3-Q1. To find the interquartile range (IQR):

First find the median (middle value) of the lower and upper half of the data.

QUES 11. What do you understand by bell curve distribution?

Answer. A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.

A normal distribution curve typically follows Bell-Shaped Curve, hence the name. The peak is always in the middle and the curve is always symmetrical. When distribution is normal Mean, Mode, Median are all the same.

QUES 12. Mention one method to find outliers.

Answer. Sorting Method.

QUES 13. What is p-value in hypothesis testing?

Answer. 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 less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

QUES 14. What is the Binomial Probability Formula?

Answer. **Binomial probability** refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p, then the binomial probability is $nCx \cdot p^x \cdot (1-p)^{n-x}$.

QUES 15. Explain ANOVA and it's applications.

Answer. **Analysis of Variance (ANOVA)** is a technique which is used to compare the means of multiple samples. Whether there is a significant difference between the mean of 2 samples, can be evaluated using z-test or t-test but in case of more than 2 samples, t-test cannot be applied as it accumulates the error and it will be more difficult as the number of sample will increase (for example: for 4 samples — 12 t-test will have to be performed). The ANOVA technique enables us to perform this simultaneous test.

