# STATISTICS WORKSHEET-4 Answers

Ans.1 :-

**central limit theorem:-**

- 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.

**why is it important:-**

- The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases.

- This is useful, as the research never knows which mean in the sampling distribution is the same as the population mean, but by selecting many random samples from a population the sample means will cluster together, allowing the research to make a very good estimate of the population mean.

- Thus, as the sample size (N) increases the sampling error will decrease.

Ans.2 :-

**sampling:-**

- Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population.

- When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample.

- The sample is the group of individuals who will actually participate in the research.

- There are two types of sampling methods:

1. Probability sampling: 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.

2. Non-probability sampling: In non-probability 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.

Ans.3 :-

**difference between type I and type II error :-**

- Type I error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.

  Type II error is the error that occurs when the null hypothesis is accepted when it is not true.

- Type I error is equivalent to false positive.

  Type II error is equivalent to a false negative.

- Type I error is a false rejection of a true hypothesis.

  Type II error is the false acceptance of an incorrect hypothesis.

- Type I error is denoted by α.

  Type II error is denoted by β.

- The probability of type I error is equal to the level of significance.

  The probability of type II error is equal to one minus the power of the test.

- Type I error can be reduced by decreasing the level of significance.

  Type II error can be reduced by increasing the level of significance.

- Type I error is caused by luck or chance.

  Type II error is caused by a smaller sample size or a less powerful test.

- Type I error is similar to a false hit.

  Type II error is similar to a miss.

- Type I error is associated with rejecting the null hypothesis.

  Type II error is associated with rejecting the alternative hypothesis.

- Type I error happens when the acceptance levels are set too lenient.

  Type II error happens when the acceptance levels are set too stringent.

Ans.4 :-

**Normal distribution :-**

- 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.

- The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions.

- Extreme values in both tails of the distribution are similarly unlikely.

- While the normal distribution is symmetrical, not all symmetrical distributions are normal.

- 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.

- Characteristics that are the sum of many independent processes frequently follow normal distributions.

- For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

Ans.5 :-

**Correlation:-**

- In statistics, correlation is a measure that determines the degree to which two or more random variables move in sequence.

- When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated.

- The formula for correlation is:

$$\rho_{xy} = Correlation\ (x, y) = \frac{cov(x, y)}{\sqrt{var(x)}\sqrt{var(y)}.}$$

where,

var(X) = standard deviation of X

var(Y) = standard deviation of Y

- Positive correlation occurs when two variables move in the same direction. When variables move in the opposite direction, they are said to be negatively correlated.

- Correlation is of three types:

  1. Simple Correlation: In simple correlation, a single number expresses the degree to which two variables are related.
  2. Partial Correlation: When one variable's effects are removed, the correlation between two variables is revealed in partial correlation.
  3. Multiple correlation: A statistical technique that uses two or more variables to predict the value of one variable.

**Covariance:-**

- Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

- The covariance value can range from -∞ to +∞, with a negative value indicating a negative relationship and a positive value indicating a positive relationship.

- The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number.

- A negative number, on the other hand, denotes negative covariance, which indicates an inverse relationship between the two variables.

- Covariance is great for defining the type of relationship, but it's terrible for interpreting the magnitude.

- Let Σ(X) and Σ(Y) be the expected values of the variables, the covariance formula can be represented as:

$$\mathrm{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Where,

- xi = data value of x
- yi = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

Ans.6 :-

**Differentiate between Univariate, Biavariate, and Multivariate analysis :-**

**Univariate:-**

- 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.

- The example of a univariate data can be height.

**Biavariate:-**

- 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:-**

- 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).

Ans.7 :-

**Sensitivity:-**

- Sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions.

- In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty.

- This technique is used within specific boundaries that depend on one or more input variables.

- Sensitivity analysis is used in the business world and in the field of economics.

- It is commonly used by financial analysts and economists and is also known as a what-if analysis.

**Calculate sensitivity:-**

1. Firstly, the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model are kept constant.

2. Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated.

3. Find the percentage change in the output and the percentage change in the input.

4. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

Ans.8 :-

**Hypothesis testing :-**

- Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.
- It is used to estimate the relationship between 2 statistical variables.
- 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.
- Let's discuss few examples of statistical hypothesis from real-life –
    1. A teacher assumes that 60% of his college's students come from lower-middle-class families.
    2. A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

**H0 and H1:-**

**H0:-**

- The null hypothesis is a concise mathematical statement that is used to indicate that there is no difference between two possibilities.
- In other words, there is no difference between certain characteristics of data.
- This hypothesis assumes that the outcomes of an experiment are based on chance alone.
- It is denoted as H0.

- Hypothesis testing is used to conclude if the null hypothesis can be rejected or not.

- Suppose an experiment is conducted to check if girls are shorter than boys at the age of 5.

- The null hypothesis will say that they are the same height.

**H1:-**

- The alternative hypothesis is an alternative to the null hypothesis.

- It is used to show that the observations of an experiment are due to some real effect.

- It indicates that there is a statistical significance between two possible outcomes and can be denoted as H1 or Ha.

- For the above-mentioned example, the alternative hypothesis would be that girls are shorter than boys at the age of 5.

**two-tail test:-**

- A two-tailed test, in statistics, is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

- It is used in null-hypothesis testing and testing for statistical significance.

- If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.

Ans.9 :-

**quantitative data:-**

- Quantitative data refers to any information that can be quantified.
- If it can be counted or measured, and given a numerical value, it's quantitative data.
- Quantitative data can tell you "how many," "how much," or "how often".
- for example, how many people attended last week's webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking?
-

**qualitative data:-**

- qualitative data cannot be measured or counted.
- It's descriptive, expressed in terms of language rather than numerical values.
- Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labelling a particular ice cream flavour as vanilla.

Ans.10 :-

**How to calculate range and interquartile range:-**

- 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 formula to calculate interquartile range is :- IQR = Q3 – Q1

  Where,

  - IQR = interquartile range
  - Q3 = 3$^{rd}$ quartile or 75$^{th}$ percentile
  - Q1 = 1$^{st}$ quartile or 25$^{th}$ percentile

Ans.11 :-

**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 consists of a symmetrical bell-shaped curve.
- In a bell curve, the centre contains the greatest number of a value and, therefore, it is the highest point on the arc of the line.
- This point is referred to the mean, but in simple terms, it is the highest number of occurrences of an element (in statistical terms, the mode).
- 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.

Ans.12 :-

**Find outlier using interquartile range:-**

- The interquartile range (IQR) tells you the range of the middle half of your dataset.
- You can use the IQR to create “fences” around your data and then define outliers as any values that fall outside those fences.
- This method is helpful if you have a few values on the extreme ends of your dataset, but you aren’t sure whether any of them might count as outliers.

- Interquartile range method

1. Sort your data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate your IQR = Q3 – Q1
4. Calculate your upper fence = Q3 + (1.5 * IQR)
5. Calculate your lower fence = Q1 – (1.5 * IQR)
6. Use your fences to highlight any outliers, all values that fall outside your fences.

- Your outliers are any values greater than your upper fence or less than your lower fence.

Ans.13 :-

**p-value:-**

- In statistics, 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.

- P-value is often used to promote credibility for studies or reports by government agencies.

- 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.

Ans.14 :-

**Binomial Probability Formula:-**

- 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)

Ans.15 :-

**ANOVA and it's applications?:-**

- 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.

- The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method.

- ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests.