STATISTICS WORKSHEET-4¶

Q1to Q15 are descriptive types.

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

ANS> 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.

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

ANS> Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population. Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights.

There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

Random sampling is analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occuring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling. You can generate random numbers using the TI82 calculator.

Systematic sampling is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every kth element is taken. This is similar to lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.

Convenience sampling is very easy to do, but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first people the surveyor runs into.

Cluster sampling is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

Stratified sampling also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

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

ANS> Type 1 error:-

Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.

STATISTICS WORKSHEET-4¶

Type 1 error is caused when the hypothesis that should have been accepted is rejected.

Type I error is denoted by α (alpha) known as an error, also called the level of significance of the test.

This type of error is a false negative error where the null hypothesis is rejected based on some error during the testing.

The null hypothesis is set to state that there is no relationship between two variables and the cause-effect relationship between two variables, if present, is caused by chance.

Type 1 error occurs when the null hypothesis is rejected even when there is no relationship between the variables. As a result of this error, the researcher might end up believing that the hypothesis works even when it doesn't.

Type 2 error:-

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

In simple words, Type II error means accepting the hypothesis when it should not have been accepted.

The type II error results in a false negative result.

In other words, type II is the error of failing to accept an alternative hypothesis when the researcher doesn't have adequate power.

The Type II error is denoted by β (beta) and is also termed as the beta error.

The null hypothesis is set to state that there is no relationship between two variables and the cause-effect relationship between two variables, if present, is caused by chance.

Type II error occurs when the null hypothesis is acceptable considering that the relationship between the variables is because of chance or luck, and even when there is a relationship between the variables.

As a result of this error, the researcher might end up believing that the hypothesis doesn't work even when it should.

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 graph form, normal distribution will appear as a bell curve.

The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

5. What is correlation and covariance in statistics?

ANS> Correlation:-

Correlation is a statistical measure that indicates how strongly two variables are related.

STATISTICS WORKSHEET-4¶

It show whether and how strongly pairs of variables are related to each other.

Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.

In this variable are indirectly related to each other.

It gives the direction and strength of relationship between variables.

Covariance:-

It is the relationship between a pair of random variables where change in one variable causes change in another variable.

It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

It is used for the linear relationship between variables.

It gives the direction of relationship between variables.

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

ANS>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 our goals.

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

ANS> A 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

STATISTICS WORKSHEET-4¶

model's overall uncertainty. This technique is used within specific boundaries that depend on one or more input variables.

Sensitivity is calculated as :- A/(A+C) × 100 where, A = True positives C = False negatives

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

ANS> Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

H0 is null hypothesis.

H1 is alternate hypothesis, the hypothesis we are interested in proving.

In a two-tailed test, the generic null(H0) and alternative hypotheses(H1) are the following:

Null(H0): The effect equals zero.

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

9. What is quantitative data and qualitative data?

ANS> Quantitative data:-

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.

Qualitative data:-

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.

10. How to calculate range and interquartile range?

ANS> Range:-

The Range is the difference between the lowest and highest values. Example: In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9. So the range is 9 − 3 = 6.

Interquartile rage:-

We can find the interquartile range or IQR in four simple steps:

Order the data from least to greatest

STATISTICS WORKSHEET-4¶

Find the median

Calculate the median of both the lower and upper half of the data

The IQR is the difference between the upper and lower medians

11. What do you understand by bell curve distribution ?

ANS> The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean.

12. Mention one method to find outliers.

ANS> Using Z-scores to Detect Outliers:-

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.

13. What is p-value in hypothesis testing?

ANS> 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 is used 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.

14. What is the Binomial Probability Formula?

ANS> The Binomial Probability distribution of exactly x successes from n number of trials is given by the below formula-

P (X) = nCx px qn – x Where,

n = Total number of trials

x = Total number of successful trials

p = probability of success in a single trial

q = probability of failure in a single trial = 1-p

15. Explain ANOVA and it’s applications.

ANS> A common approach to figure out a reliable treatment method would be to analyse the days it took the patients to be cured. We can use a statistical technique which can compare

STATISTICS WORKSHEET-4¶

these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples on the basis of their means, is called ANOVA.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

One real-life application of analysis of variance is the recommendation of a fertilizer against others for the improvement of a crop yield.

Anova can be in different fields of sciences, i.e. all the problems of testing more than three groups.