Statistical Worksheet-4

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

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 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.

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

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed. There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

⦁ Random Sampling: Here, 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: It 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.

⦁ Convenience Sampling: Convenience sampling is a type of nonprobability sampling in which people are sampled simply because they are "convenient" sources of data for researchers. In probability sampling, each element in the population has a known nonzero chance of being selected through the use of a random selection procedure.

⦁ Cluster Sampling: It 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.

⦁ 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).

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

A type I error is the rejection of a true null hypothesis and also known as a "false positive" finding or conclusion, while a type II error is the non-rejection of a false null hypothesis, also known as a "false negative" finding or conclusion. Much of statistical theory revolves around the minimization of one or both of these errors, though the complete elimination of either is a statistical impossibility for non-deterministic algorithms. By selecting a low threshold (cut-off) value and modifying the alpha (p) level, the quality of the hypothesis test can be increased.

⦁ type 1 error is the rejection of a true null hypothesis as the result of a test procedure(false positive) , and is sometimes called an error of the first kind.

⦁ type 2 error is the failure to reject a false null hypothesis as the result of a test procedure. This sort of error is called a type II error (false negative) and is also referred to as an error of the second kind.

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

The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where 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. The normal distribution is often called the bell curve because the graph of its probability density looks like a bell.

5. What is correlation and covariance in statistics?

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 is a relationship between two variables in which both variables move in the same direction. A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. Understanding that relationship is useful because we can use the value of one variable to predict the value of the other variable.

Covariance provides insight into how two variables are related to one another. More precisely, covariance refers to the measure of how two random variables in a data set will change together. A positive covariance means that the two variables at hand are positively related, and they move in the same direction. Covariance is usually measured by analyzing standard deviations from the expected return or we can obtain by multiplying the correlation between the two variables by the standard deviation of each variable.

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

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn't deal with causes or relationships (unlike regression ) and it's major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. One example of a variable in univariate analysis might be "age". Another might be "height".

Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. Some of the examples are percentage table, scatter plot, etc.

Multivariate analysis is a set of statistical techniques used for analysis of data that contain more than one variable. Multivariate analysis refers to any statistical technique used to analyse more complex sets of data. This type of analysis is almost always performed with software (i.e. SPSS or SAS), as working with even the smallest of data sets can be overwhelming by hand.

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

Sensitivity is a statistical measure of the performance of a binary classification test that are widely used in medicine. Sensitivity (True Positive rate) measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition). Sensitivity refers to a test's ability to designate an individual with disease as positive. A highly sensitive test means that there are few false negative results, and thus fewer cases of disease are missed. The specificity of a test is its ability to designate an individual who does not have a disease as negative.

Sensitivity is calculated by the following formula:

Sensitivity=[a/(a+c)]×100Specificity=[d/(b+d)]×100Positive\_-predictive value(PPV)=[a/(a+b)]×100Negative predictive value(NPV)=[d/(c+d)]×100

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

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 in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.

Alternative Hypothesis, H1 is the hypothesis that we are interested in proving. Null hypothesis, H0 is the complement of the alternative hypothesis. This is the probability of falsely rejecting the null hypothesis.

Two-tailed hypothesis tests are also known as non-directional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution.

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. Here, we doesn’t know whether the average has gone up or down, but he believes that it has changed and is not 100 anymore.

H1: μ ≠100

9. What is quantitative data and qualitative data?

Quantitative data is the type of data whose value is measured in the form of numbers or counts, with a unique numerical value associated with each data set. Also known as numerical data, quantitative data further describes numeric variables. This data is any quantifiable information that can be used for mathematical calculations and statistical analysis, such that real-life decisions can be made based on these mathematical derivations.

Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods. Qualitative data is a type of data that describes information. It is investigative and also often open-ended, allowing respondents to fully express themselves. Also known as categorical data, this data type isn’t necessarily measured using numbers but rather categorized based on properties, attributes, labels, and other identifiers.

10. How to calculate range and interquartile range?

The range is the size of the smallest interval which contains all the data and provides an indication of statistical dispersion. It is measured in the same units as the data.The range is calculated by subtracting the lowest value from the highest value. Example: In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9. So the range is 9 − 3 = 6.

The interquartile range (IQR), also called the midspread, middle 50%, or H‑spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles. The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. Formula for calculating IQR is IQR = Q3 − Q1.

11. What do you understand by 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. The width of the bell curve is described by its standard deviation. 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 standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean. The mean, in turn, refers to the average of all data points in the data set or sequence and will be found at the highest point on the bell curve.

12. Mention one method to find outliers.

The most effective way to find all of the outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR.

13. What is p-value in hypothesis testing?

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. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis. 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. The p-value approach to hypothesis testing uses the calculated probability to determine whether there is evidence to reject the null hypothesis. The null hypothesis, also known as the conjecture, is the initial claim about a population (or data generating process). The alternative hypothesis states whether the population parameter differs from the value of the population parameter stated in the conjecture.

14. What is the Binomial Probability Formula?

Binomial probability refers to the probability of exactly 'x' successes on 'n' repeated trials in an experiment which has two possible outcomes. the Binomial Probability Formula is:

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.

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures,such as the "variation" among and between groups used to analyze the differences among means. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means. The analysis of variance can be used to describe otherwise complex relations among variables. Applications of ANOVA are as follows:

⦁ ANOVA checks the impact of one or more factors by comparing the means of different samples.

⦁ ANOVA can be used to prove/disprove if all the medication treatments were equally effective or not.

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