STATISTICS WORKSHEET - 4

1. Q1to Q15 are descriptive types. Answer in brief:

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

➢ The theorem states that as the size of the sample increases, the distribution of the mean across multiple samples will approximate a Gaussian distribution. It is important in applications of statistics and in the understanding of nature.- It confirms that the normal distribution is essential to nature. This builds confidence that we, math people, can understand and explain nature. We have other indications that the normal distribution is natural(e.g. that the normal distribution is a Maximum entropy probability distribution) but CLT is the main reason that we think so highly of the normal distribution. Before simulation based methods, such as Bootstrapping (statistics) and permutation tests, were widespread, a CLT was the only tool available when confidence intervals and p values should be found in many situations. Today you can choose either simulation or a CLT based result(though sometimes, as in GWAS, simulation would take too long). Hence, the cruciality of the CLTs is not as big anymore in applications. They are still being used a lot because CLT results are easier to use than simulations and often just as good as simulations.

    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, but it may include simple random sampling or systematic sampling.

There are Five types of Sampling –

1)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.

2) Systematic sampling is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every k th 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.

3)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.

4)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.

5)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 type I and type II error?

➢

Type I	Type II
• Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesi when it is true. • 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 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?

➢ 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.

5. What is correlation and covariance in statistics?

➢ Covariance and Correlation are two mathematical concepts which are commonly used in the field of probability and statistics. Both concepts describe the relationship between two variables.

Covariance –

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. 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.

3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.

Correlation-

1. It show whether and how strongly pairs of variables are related to each other.
2. 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.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

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

➢ 1) Univariate statistics summarize only one variable at a time.
2) Bivariate statistics compare two variables.
3) Multivariate statistics compare more than two variables.

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

➢ 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 model's overall uncertainty. This technique is used within specific boundaries that depend on one or more input variables.
We can calculate It with :- Sensitivity: $A/(A+C) \times 100$.

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 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. The word "population" will be used for both of these cases in the following descriptions.

null (H0) and alternative (H1) hypothesis, hypothesis testing is the technique of analyzing sample data to make one of the following two decisions: We have enough evidence to reject Ho in favor of H1. We don’t have enough evidence to reject Ho in favor of H1.

In Two Tail test H0 And H1 are : 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: $\mu = 100$

Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed H1: $\mu \neq 100$

9. What is quantitative data and qualitative data?

➢ Quantitative data is information about quantities, and therefore numbers, and qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language. Qualitative data is defined as non-numerical data, such as text, video,

photographs or audio recordings. This type of data can be collected using diary accounts or in-depth interviews, and analyzed using grounded theory or thematic analysis.

10. How to calculate range and interquartile range?

- Range - The Range is the difference between the lowest and highest values. We can find the interquartile range or IQR in four simple steps:
  1. Order the data from least to greatest
  2. Find the median
  3. Calculate the median of both the lower and upper half of the data
  4. The IQR is the difference between the upper and lower medians

11. What do you understand by bell curve distribution ?

- 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. 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?

- Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.
  We can Use Z-scores to Detect Outliers

13. What is p-value in hypothesis testing?

- The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested. P is also described in terms of rejecting H0 when it is actually true

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 (commonly called a binomial experiment).
  If the probability of success on an individual trial is p , then the binomial probability is
  nCx px (1−p)n−xnCx px (1−p)n−x · · · · .
  Here nCxnCx indicates the number of different combinations of x objects selected from a set of n objects. Some textbooks use the notation (nx)(nx0) instead of nCxnCx. Note that if p is the probability of success of a single trial, then (1−p)(1−p) is the probability of failure of a single trial.

15. Explain ANOVA and it's applications.

- 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. We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not