STATISTICS WORKSHEET-4

ANSWERS:

1. 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. 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 -

- 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.

Basis for comparison	Type I Error	Type II Error
Definition	Type 1 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.
Also termed	Type I error is equivalent to a false positive.	Type II error is equivalent to a false negative.
Meaning	It is a false rejection of a true hypothesis.	It is the false acceptance of an incorrect hypothesis.
Symbol	Type I error is denoted by α.	Type II error is denoted by β.
Probability	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.
Reduced	It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.
Cause	It is caused by luck or chance.	It is caused by smaller sample size or a less powerful test.
What is it?	Type I error is similar to a	Type II error is similar to a

	false hit.	miss.
Hypothesis	Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.
When does it happen?	It happens when the acceptance levels are set too lenient.	It happens when the acceptance levels are set too stringent.

4. 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. 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.
- 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.

Correlation –

- 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.

6. Univariate statistics summarize only one variable at a time, Bivariate statistics compare two variables and Multivariate statistics compare more than two variables.

7. 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.

Sensitivity: $A/(A+C) \times 100$.

8. 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. 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. Range - The Range is the difference between the lowest and highest values. We can find the interquartile range or IQR in four simple steps:

- Order the data from least to greatest
- 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. 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. Outliers are data points that are far from other data points. In other words, they are 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. 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. 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. 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.