STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

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 .

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

Sampling, in simple terms, means selecting a group (a sample) from a population from which we will collect data for our research. Sampling is an important aspect of a research study as the results of the study majorly depend on the sampling technique used. So, in order to get accurate results or the results that can estimate the population well, the sampling technique should be chosen wisely.

Sampling Methods

- Simple Random Sampling (SRS):

Suppose we have a population of 20 people and we need to get a sample of 7 people from this population. For the sake of understanding, let us number these people. Now, we will randomly choose 7 numbers between 1 and 20 and the people against those numbers will be a part of our sample. If the person against the chosen number is already in our sample, we will just skip that number and choose another number.

- Stratified Sampling:

Let us take the same sample as above. Let us say we want a sample of size 9 this time. Let us arrange these people in different groups and let these groups be based on the color of the clothes these people are wearing.

3. What is the difference between type1 and type II error?

BASIS FOR COMPARISON	TYPE I ERROR	TYPE II ERROR
Meaning	Type I error refers to non-acceptance of hypothesis which ought to be accepted.	Type II error is the acceptance of hypothesis which ought to be rejected.
Equivalent to	False positive	False negative
What is it?	It is incorrect rejection of true null hypothesis.	It is incorrect acceptance of false null hypothesis.

BASIS FOR COMPARISON	TYPE I ERROR	TYPE II ERROR
Represents	A false hit	A miss
Probability of committing error	Equals the level of significance.	Equals the power of test.
Indicated by	Greek letter 'α'	Greek letter 'β'

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

The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Most people recognize its familiar bell-shaped curve in statistical reports.

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. For example, the Student's t, Cauchy, and logistic distributions are symmetric.

5. What is correlation and covariance in statistics?

Basis	Covariance	Correlation
Meaning	Covariance is an indicator of how two random variables are dependent on each other. A higher number denotes higher dependency.	Correlation indicates how strongly these two variables are related, provided other conditions are constant. The maximum value is +1, representing a perfect dependent relationship.
Relationship	We can deduct correlation from a covariance.	Correlation provides a measure of covariance on a standard scale. It is deduced by dividing the calculated covariance by standard deviation.
Values	The value of covariance lies in the range of -∞ and +∞.	Correlation is limited to values between the range -1 and +1.
Scalability	Covariance is affected.	Correlation is not affected by a change in scales or multiplication by a constant.

Basis	Covariance	Correlation
Units	Covariance has a definite unit as deduced by the multiplication of two numbers and their units.	Correlation is a unitless absolute number between -1 and +1, including decimal values.

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

Univariate Data Analysis:

Univariate data is used for the simplest form of analysis. It is the type of data in which analysis arem de only based on one variable. For example, there are sixty students in class VII. If the variablemarks obtained in math were the subject, then in that case analysis will be based on the number ofsubject fall into defined categories of marks.

Bivariate Data Analysis:

Bivariate data is used for little complex analysis than as compared with univariate data.

Bivariatedata is the data in which analysis are based on two variables per observation simultaneous y.

Multivariate Data Analysis:

Multivariate data is the data in which analysis are based on more than two variables perobservation Usually multivariate data is used for explanatory purposes.

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

Sensitivity Analysis is a tool used in financial modeling to analyze how the different values of a set o independent variables affect a specific dependent variable under certain specific conditions. In gene al, sensitivity analysis is used in a wide range of fields, ranging from biology and geography to econom s and engineering.

It is especially useful in the study and analysis of a “Black Box Process” where the output is an opa ue function of several inputs. An opaque function or process is one which, for some reason, can’t be st died and analyzed. For example, climate models in geography are usually very complex. As a result, the exact relationship between the inputs and outputs are not well understood.

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

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a popul ion parameter to the test. It is used to estimate the relationship between 2 statistical variables.

H_0	H_1
Equal (=)	Not equal to (≠) or greater than (>) or less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

9. What is quantitative data and qualitative data?

	Qualitative	Quantitative
Conceptual	Concerned with understanding human behaviour from the informant's perspective	Concerned with discovering facts about social phenomena
	Assumes a dynamic and negotiated reality	Assumes a fixed and measurable reality
Methodological	Data are collected through participant observation and interviews	Data are collected through measuring things
	Data are analysed by themes from descriptions by informants	Data are analysed through numerical comparisons and statistical inferences
	Data are reported in the language of the informant	Data are reported through statistical analyses

Source: Adapted from Minchiello *et al.* (1990, p. 5)

10. How to calculate range and interquartile range?

How to Find Interquartile Range

The interquartile range IQR is the range in values from the first quartile Q_1 to the third quartile Q_3. Find the IQR by subtracting Q_1 from Q_3.

- $IQR = Q_3 - Q_1$

How to Find the Minimum

The minimum is the smallest value in a sample data set.

Ordering a data set from lowest to highest value, $x_1 \le x_2 \le x_3 \le \ldots \le x_n$, the minimum is the smallest value x_1. The formula for minimum is:

Min=x1=min(xi)ni=1Min=x1=min(xi)i=1n

How to Find the Maximum

The maximum is the largest value in a sample data set.

Ordering a data set from lowest to highest value, $x_1 \le x_2 \le x_3 \le \ldots \le x_n$, the maximum is the largest value x_n. The formula for maximum is:

Max=xn=max(xi)ni=1Max=xn=max(xi)i=1n

How to Find the Range of a Set of Data

The range of a data set is the difference between the minimum and maximum. To find the range, calculate x_n minus x_1.

R=xn−x1

11. What do you understand by bell curve distribution ?

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.

The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).

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.

Unfortunately, there are no strict statistical rules for definitively identifying outliers. Finding outliers depends on subject-area knowledge and an understanding of the data collection process. While there is no solid mathematical definition, there are guidelines and statistical tests you can use to find outlier candidates.

13. What is p-value in hypothesis testing?

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That's pretty tiny. On the other hand, a large p-value of .9(90%) means your results have a 90% probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important ("significant") your results.

14. What is the Binomial Probability Formula?

A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes (the prefix "bi" means two, or twice). For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

15. Explain 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.