Mrs. Swati Amit Motugade

Statistics Assignment 4

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

Ans. The Central Limit Theorem is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

Importance:

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, as we will see in the next section.

The central limit theorem is important because it helps us to understand the behaviour of a population's average when samples are taken from it. It allows us to use standard statistical techniques to analyse data even if the population distribution is not normal.

The importance of the CLT stems from the fact that, in several real applications, a random variable is the sum of a large number of independent random variables. Thus, the CLT explains why the Gaussian probability distribution is observed so commonly in nature.

Q. 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 the 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. It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country's population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug's behaviour.

Sampling is of two types – probability sampling and non-probability sampling.

1. Probability sampling:

Probability sampling is a sampling technique in which researchers choose samples from a larger population using a method based on the theory of probability. This sampling method considers every member of the population and forms samples based on a fixed process.

For example, in a population of 1000 members, every member will have a 1/1000 chance of being selected to be a part of a sample. Probability sampling eliminates sampling bias in the population and gives all members a fair chance to be included in the sample.

There are four types of probability sampling techniques:

1. Simple Random Sampling
2. Cluster Sampling
3. Systematic Sampling
4. Stratified Random Sampling

2. non-probability sampling:

The non-probability method is a sampling method that involves a collection of feedback based on a researcher or statistician's sample selection capabilities and not on a fixed selection process. In most situations, the output of a survey conducted with a non-probable sample leads to skewed results, which may not represent the desired target population. But there are situations such as the preliminary stages of research or cost constraints for conducting research, where non-probability sampling will be much more useful than the other type.

Four types of non-probability sampling explain the purpose of this sampling method in a better manner:

1. **Convenience Sampling**

2. Judgmental or Purposive Sampling
3. Snowball Sampling
4. Quota Sampling

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

Ans. Type I Error:

A Type I error means rejecting the null hypothesis when it's actually true. It means concluding that results are statistically significant when, in reality, they came about purely by chance or because of unrelated factors.

Type II Error:

A type II error is a statistical term used within the context of hypothesis testing that describes the error that occurs when one fails to reject a null hypothesis that is actually false. A type II error produces a false negative, also known as an error of omission.

Difference Between Type I and II Error

Type I Error	Type II Error
Type 1 error 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.
Type I error is equivalent to a false positive.	Type II error is equivalent to a false negative.
It is a false rejection of a true hypothesis.	It is the false acceptance of an incorrect hypothesis.
Type I error is denoted by α.	Type II error is denoted by β.
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.
It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.
It is caused by luck or chance.	It is caused by smaller sample size or a less powerful test.
It happens when the acceptance levels are set too lenient.	It happens when the acceptance levels are set too stringent.
Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.

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

Ans. The normal distribution is probably the most important distribution in all of probability and statistics.

Many populations have distributions that can be fit very closely by an appropriate normal (or Gaussian, bell) curve.

Examples: height, weight, and other physical characteristics, scores on various tests, etc.

Definition: A continuous r.v. X is said to have a normal distribution with parameters μ and σ > 0 if the pdf of X is

$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$ where -∞<x<∞

The statement that X is normally distributed with parameters μ and σ^2 is often abbreviated X ~ N(μ, σ^2).

The normal distribution with parameter values μ = 0 and σ = 1 is called the standard normal distribution.

A r.v. with this distribution is called a standard normal random variable and is denoted by Z. Its pdf is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

For the normal distribution, the values of mean, mode and median are same. That is why this is also known as symmetric distribution and its skewness is 0.

Real Life Applications:

1. For practical purpose normal distribution is good enough to represent the distribution of continuous variable like height, weight, blood pressure etc.
2. Often used to approximate other distributions.
3. It has significant applications in Statistical quality control.
4. It can be use to describe the situation where very few individuals possess extreme values and more individuals are found near the average value.
5. Also used to approximate discrete distributions like Binomial distribution for large values of ‘n’.

Q.5. What is correlation and covariance in statistics?

Ans.

Covariance:

In statistics and probability theory, covariance deals with the joint variability of two random variables: x and y. Generally, it is treated as a statistical tool used to define the relationship between two variables.

Covariance is a measure of the relationship between two random variables and to what extent, they change together. In other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance:

If both the variables are moving in same direction that is both are increasing or both are decreasing then the covariance between them is known as positive covariance.

2. Negative Covariance:

If the variables are not moving in same direction i.e., if one is increasing and as a result of this other is decreasing or vice-versa then there is a negative covariance between these two variables.

Formula:

Let X and Y are two variables whose relationship is to be calculated then the formula for covariance between them is given by

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{n}$$

Correlation:

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.

The sample correlation coefficient, *r*, quantifies the strength of the relationship. Correlations are also tested for statistical significance.

Correlations are useful for describing simple relationships among data. For example, imagine that you are looking at a dataset of campsites in a mountain park. You want to know whether there is a relationship between the elevation of the campsite (how high up the mountain it is), and the average high temperature in the summer.

Formula:

Let X and Y be two variables then the correlation between these two is given by

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

The value of correlation lies between -1 and 1.

The negative value shows that there is a negative correlation. The positive value shows that there is a positive correlation. The 0 values means there is no correlation between the variables and the 1 values means the perfect correlation.

Q. 6. Differentiate between univariate, Bivariate and multivariate analysis.

Ans. **1. Univariate Analysis**:
This type of analysis consists of only one variable. The analysis of univariate data is the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

Heights in cm	167	162.3	154	168	162.5	163	159.4

2. Bivariate data –
This type of analysis involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables. Examples of bivariate data can be temperature and ice cream sales in summer season.

Temperature (in Celsius)	Ice-cream Sales
20	2000
25	2500
35	5000
43	7800

Suppose the temperature and ice cream sales are the two variables of a bivariate data. Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase. Thus, bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

3. Multivariate data –
When the analysis involves three or more variables, it is categorized under multivariate. Example of this type of analysis is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

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

Ans. The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as sensitive analysis. Its usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond's price.

It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system. All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.

Calculation of sensitivity analysis

Below are mentioned the steps used to conduct sensitivity analysis:

1. Firstly, the base case output is defined; say the NPV at a particular base case input value (V1) for which the sensitivity is to be measured. All the other inputs of the model are kept constant.
2. Then the value of the output at a new value of the input (V2) while keeping other inputs constant is calculated.
3. Find the percentage change in the output and the percentage change in the input.
4. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

This process of testing sensitivity for another input (say cash flows growth rate) while keeping the rest of inputs constant is repeated until the sensitivity figure for each of the inputs is obtained. The conclusion would be that the higher the sensitivity figure, the more sensitive the output is to any change in that input and vice versa.

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

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

few real-life examples of statistical hypothesis

- A teacher assumes that 60% of his college's students come from lower-middle-class families.
- A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

H0: H0 is generally known as null hypothesis. It basically defines the statement which states that there is no exact or actual relationship between the variables.

H1: H1 is generally known as Alternative hypothesis. It makes a statement that suggests or advises a potential result or an outcome that an investigator or the researcher may expect.

Two-tail test:

Two-tailed hypothesis tests are also known as nondirectional and two-sided tests because we can test for effects in both directions. When we perform a two-tailed test, we split the significance level percentage between both tails of the distribution. In the example below, I use an alpha of 5% and the distribution has two shaded regions of 2.5% (2 * 2.5% = 5%).

When a test statistic falls in either critical region, our sample data are sufficiently incompatible with the null hypothesis that we can reject it for the population.

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

- **Null**: The effect equals zero.
- **Alternative**: The effect does not equal zero.

The specifics of the hypotheses depend on the type of test we perform because we might be assessing means, proportions, or rates.

e.g., H0: $\mu=0$ v/s H1: $\mu\neq0$

Q. 9. What is quantitative data and qualitative data?

Ans.

1. Quantitative data: Quantitative data is the value of data in the form of counts or numbers where each data set has a unique numerical value. This data is any quantifiable information that researchers can use for mathematical calculations and statistical analysis to make real-life decisions based on these mathematical derivations.

e.g. Cost of laptops, weight in pounds, height in inches etc.

2.Qualitative Data: Qualitative data is the descriptive and conceptual findings collected through questionnaires, interviews, or observation. Analysing qualitative data allows us to explore ideas and further explain quantitative

results. While quantitative data collection retrieves numerical data (what, where, when), qualitative data, often presented as a narrative, collect the stories and experiences of individual patients and families (why, how)

e.g., result of exam pass or fail, marital status, blood groups etc.

Q. 10. How to calculate range and interquartile range?

Ans.

Range: Range is the difference between largest and smallest value of the data. To calculate the range, we need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution. It's used as a supplement to other measures, but it is rarely used as the sole measure of dispersion because it's sensitive to extreme values.

e.g., 25,21,17,31,15,9,35,26,14,23 is the given data then the range can be calculated as

Range = L – S where L is largest value and S is smallest value

Range = 35 – 9

Range = 26

Interquartile Range:

The interquartile range is the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q_1 known as the lower quartile, the second Quartile is denoted by Q_2 and the third Quartile is denoted by Q_3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

Interquartile range = Upper Quartile – Lower Quartile = $Q_3 - Q_1$

where Q_1 is the first quartile and Q_3 is the third quartile of the series.

Q. 11. What do you understand by bell curve distribution?

Ans. 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 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.

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.

Q. 12. Mention one method to find outliers.

Ans. Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on our statistical analysis and skew the results of any hypothesis tests.

It's important to carefully identify potential outliers in our dataset and deal with them in an appropriate manner for accurate results.

IQR method of removing outliers:

1. Sort the given data in ascending order.
2. Find the quartiles Q1, Q2 and Q3 of the data.
3. Calculate IQR by using formula IQR = Q3-Q1
4. Calculate the upper and lower boundaries by using the formulae

$$Upper\ boundary = Q3 + 1.5 * IQR$$

$$Lower\ Boundary = Q1 - 1.5 * IQR$$

The points below the lower boundary and above the upper boundary are considered as outliers.

Example: Suppose the following dataset is given and we have to find outliers for this data

35,24,31,53,37,29,22,28,64,41,25

Firstly, we have to arrange this data in ascending order

22,24,25,28,29,31,35,37,41,53,64

Let's find the Q1 and Q3 for this data

Since there are 11 observations in data i.e., n=11 means the value at 3rd place is Q1 and the value at 9th place Q3

i.e., Q1=25 and Q3=41

Let's find IQR

IQR = Q3-Q1 = 41-25 = 16

Now find the lower and upper boundaries

Lower Boundary = Q1-1.5*IQR = 25-1.5*16 = 25-24 = 1

Upper Boundary = Q3+1.5*IQR= 41+1.5*16 = 41+24 = 65

Which means that the points below 1 and the points above 65 are the outliers for above dataset.

Q. 13. What is p-value in hypothesis testing?

Ans. The P-value is known as the probability value. It is defined as the probability of getting a result that is either the same or more extreme than the actual observations. The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event. The P-value is used as an alternative to the rejection point to provide the least significance at which the null hypothesis would be rejected. If the P-value is small, then there is stronger evidence in favour of the alternative hypothesis.

If p-value>0.05, the result is not statistically significant and hence don't reject the null hypothesis.

If p-value<0.05, the result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis.

Generally, the level of statistical significance is often expressed in p-value and the range between 0 and 1. The smaller the p-value, the stronger the evidence and hence, the result should be statistically significant. Hence, the rejection of the null hypothesis is highly possible, as the p-value becomes smaller.

How to calculate p-value

The p-value is calculated using the sampling distribution of the test statistic under the null hypothesis, the sample data, and the type of test being done (lower-tailed test, upper-tailed test, or two-sided test). The p-value for: a lower-tailed test is specified by:

p-value = P (TS ts | H_0 is true) = cdf(ts)

Q. 14. What is the Binomial Probability Formula?

Ans. The binomial distribution formula helps to check the probability of getting "x" successes in "n" independent trials of a binomial experiment. To recall, the binomial distribution is a type of probability distribution in statistics that has two possible outcomes. In probability theory, the binomial distribution comes with two parameters n and p.

The probability distribution becomes a binomial probability distribution when it meets the following requirements.

The formula for the binomial probability distribution is as stated below:

$$P(x) = {}^{n}C_{x} \cdot p^{x} (1 - p)^{n-x}$$

Where, n = Total number of events

r (or) x = Total number of successful events.

p = Probability of success on a single trial.

${}^{n}C_{r} = [n! /r! (n-r)]!$

1 – p = Probability of failure.

Q. 15. Explain ANOVA and its applications.

Ans. An **ANOVA** ("Analysis of Variance") is a statistical technique that is used to determine whether or not there is a significant difference between the means of three or more independent groups. The two most common types of ANOVAs are the one-way ANOVA and two-way ANOVA.

A **One-Way ANOVA** is used to determine how one factor impacts a response variable. For example, we might want to know if three different studying techniques lead to different mean exam scores. To see if there is a statistically significant difference in mean exam scores, we can conduct a one-way ANOVA.

A **Two-Way ANOVA** is used to determine how two factors impact a response variable, and to determine whether or not there is an interaction between the two factors on the response variable. For example, we might want to know how gender and how different levels of exercise impact average weight loss. We would conduct a two-way ANOVA to find out.

It's also possible to conduct a three-way ANOVA, four-way ANOVA, etc. but these are much more uncommon and it can be difficult to interpret ANOVA results if too many factors are used.

Real Life Applications of ANOVA:

ANOVA is used in a wide variety of real-life situations, but the most common include:

- **Retail:** Store are often interested in understanding whether different types of promotions, store layouts, advertisement tactics, etc. lead to different sales. This is the exact type of analysis that ANOVA is built for.
- **Medical:** Researchers are often interested in whether or not different medications affect patients differently, which is why they often use one-way or two-way ANOVAs in these situations.
- **Environmental Sciences:** Researchers are often interested in understanding how different levels of factors affect plants and wildlife. Because of the nature of these types of analyses, ANOVAs are often used.