FLIP ROBO

STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

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

   Answer-

   In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a "bell curve") as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

   The central limit theorem is useful **when analyzing large data sets** because it allows one to assume that the sampling distribution of the mean will be normally-distributed in most cases. This allows for easier statistical analysis and inference.

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

   Answer-

   Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling

   Types of Sampling Methods are-

   1. **Probability sampling:** Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.

   2. **Non-probability sampling:** In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

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

   Answer-

   In statistics, **type I** error is defined as an error that occurs when the sample results cause the rejection of the null hypothesis, in spite of the fact that it is true. In simple terms, the error of agreeing to the alternative hypothesis, when the results can be ascribed to chance.

   When on the basis of data, the null hypothesis is accepted, when it is actually false, then this kind of error is known as Type II Error. It arises when the researcher fails to deny the false null hypothesis. It is denoted by Greek letter 'beta (β)' and often known as beta error.

The points given below are substantial so far as the differences between type I and type II error is concerned:

1. Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
2. Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.
3. When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.
4. Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.
5. The probability of committing type I error is the sample as the level of significance. Conversely, the likelihood of committing type II error is same as the power of the test.
6. Greek letter 'α' indicates type I error. Unlike, type II error which is denoted by Greek letter 'β'.

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

Answer-

The Normal Distribution, also called the Gaussian Distribution, is the most significant continuous probability distribution for independent, randomly generated variables. Sometimes it is also called a bell curve.A large number of random variables are either nearly or exactlyrepresented by the normal distribution, in every physical science andeconomics. Furthermore, it can be used to approximate other probability distributions, therefore supporting the usage of the word 'normal 'as in about the one, mostly used.

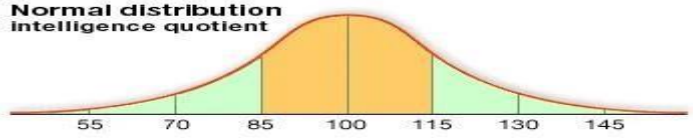
Normal Distribution Formula:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

The graph of the normal distribution is characterized by two parameters: the mean, or average, which is the maximum of the graphand about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from themean.



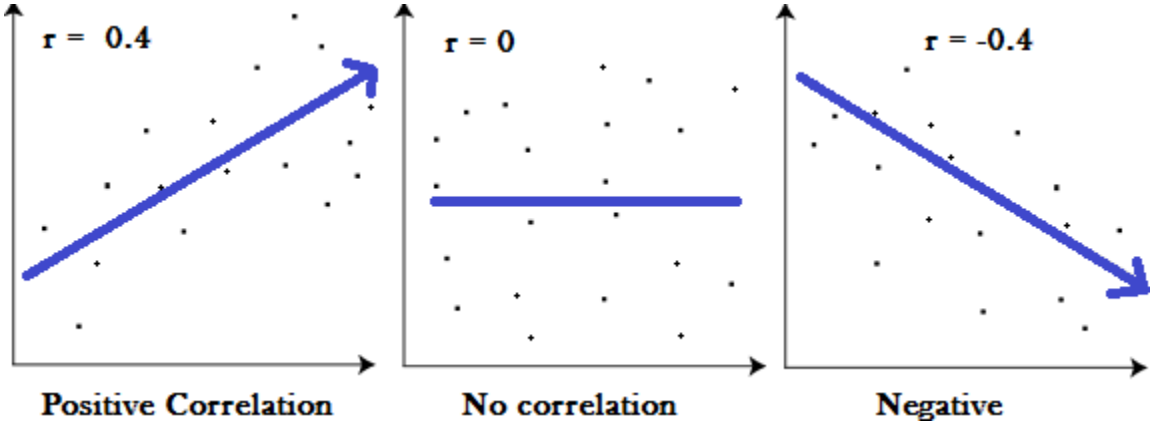
Q-5) What is correlation and covariance in statistics?

Answer-

Correlation is a statistical measure that expresses the extent to whichtwo variables are linearly related. It's a common tool for describing simple relationships without making a statement about cause and effect. It is used to test relationships between quantitative variables or categorical variables. In other words, it's a measure of how things are related. The study of how variables are correlated is called correlation analysis.

A correlation coefficient is a way to put a value to the relationship. Correlation coefficients have a value of between -1 and 1. A "0" meansthere is no relationship between the variables at all, while -1 or 1 means that there is a perfect negative or positive correlation.

The graph of Correlation is:



Correlation is of three types:

- **Simple Correlation:** In simple correlation, a single number expresses the degree to which two variables are related.
- **Partial Correlation:** When one variable's effects are removed, the correlation between two variables is revealed in partialcorrelation.
- **Multiple correlation:** A statistical technique that uses two or more variables to predict the value of one variable.

Correlation coefficient Formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Where,

n: Quantity of Information

Σx: Total of the First Variable Value

Σy: Total of the Second Variable Value

Σxy: Sum of the Product of & Second Value

Σx2: Sum of the Squares of the First Value

Σy2: Sum of the Squares of the Second Value

Covariance:

Covariance is a statistical tool that is used to determine the relationship between the movements of two random variables. Whentwo stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

Covariance is different from the correlation coefficient, a measure of the strength of a correlative relationship. It is a significant tool in modern portfolio theory used to ascertain what securities to put in a portfolio. Risk and volatility can be reduced in a portfolio by pairing assets that have a negative covariance.

Types of Covariance:

- **Positive Covariance:** If the covariance for any two variables is positive, that means, both the variables move in the same direction.
- **Negative Covariance:** If the covariance for any two variables is negative, that means, both the variables move in the opposite direction.

Formula:

- Covariance formula for population:

$$Cov(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n}$$

- Covariance Formula for a sample:

$$Cov(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Where,

- X_i is the values of the X-variable
- Y_i is the values of the Y-variable
- $\overline{X}$ is the mean of the X-variable
- $\overline{Y}$ is the mean of the Y-variable
- n is the number of data points

Q-6)Differentiate between univariate ,Bivariate,and multivariate analysis

Answer-

- **Univariate:** It summarize only one variable at a time. The analysis of univariate data is thus 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 patternsthat exist within it.

The example of a univariate data can be height.

- **Bivariate:** It compare two variables. The analysis of this type ofdata deals with causes and relationships and the analysis is doneto find out the relationship among the two variables.

Example of bivariate data can be temperature and ice cream sales in summer season.

- **Multivariate:** It compare more than two variables. It is similarto bivariate but contains more than one dependent variable. Theways to perform analysis on this data depends on the goals to beachieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

Example of this type of data 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 andwomen and relationships between variables can then be examined.

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

Answer-

Sensitivity analysis determines how different values ofan independent variable affect a particular dependent variable undera given set of assumptions.

This model is also referred to as a what-if or simulation analysis. It canbe used to help make predictions in the share prices of publicly tradedcompanies or how interest rates affect bond prices. It allows for forecasting using historical, true data.

While sensitivity analysis determines how variables impact a single event, scenario analysis is more useful to determine many different outcomes for more broad situations.

Calculate Sensitivity Analysis:

Sensitivity analysis is often performed in analysis software, and Excel has built in functions to help perform the analysis. In general, sensitivity analysis is calculated by leveraging formulas that referencedifferent input cells. For example, a company may perform NPV analysis using a discount rate of 6%. Sensitivity analysis can be performed by analyzing scenarios of 5%, 8%, and 10% discount rates as well by simply maintaining the formula but referencing the different variable values.

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

Answer- Hypothesis Testing is a type of statistical analysis in which we put ourassumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables. Example of statistical hypothesis is:

A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effectivefor diabetic patients.
Types of Hypothesis Testing: Two types of hypothesis testing are:

I. **Null Hypothesis:** It is denoted by symbol H0. The Null Hypothesis is the assumption that the event will not occur. A nullhypothesis has no bearing on the study's outcome unless it is rejected.
II. **Alternate Hypothesis:** It is denoted by symbol H1. The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis followsthe rejection of the null hypothesis.

Two-Tailed Hypothesis Testing:

In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the criticaldistribution area is two-sided.

Example 1:
Suppose,
H0: mean = 50 and
H1: mean ≠ 50 (the mean can be greater than or less than 50 butnot equal to)

Example 2:
The average height of students in a batch is 100 cm and the standard deviation is 15. However, one teacher believes that this has changed, so he/she decides to test the height of 75 random students in the batch. The average height of the sample comes outto be 105. The steps for performing hypothesis testing are:

- Specify the Null(H0) and Alternate(H1) hypothesis
- Choose the level of Significance(α)
- Find Critical Values
- Find the test statistic
- Draw conclusion

➢ **Specify the Null(H0) and Alternate(H1) hypothesis for Two- Tailed Hypothesis Testing:**

- Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In this example, 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. In this example, teacher believes (Claims) that the actual value has changed. He/she doesn't know whether the average has gone up ordown, but he/she believes that it has changed and is not 100 anymore.

H1: $\mu \neq 100$

If the alternate hypothesis is written with a ≠ sign that means that we are going to perform a two-tailed test because chances,are it could be more than 100 or less than 100 which makes it two-tailed as an alternate hypothesis is always written with a ≠ or < or > sign.

Q-9) What is quantitative data and qualitative data?

Answer-

Quantitative data is anything that can be counted or measured; it refers to numerical data. Qualitative data is descriptive, referring to things that can be observed but not measured—such as colours or emotions.

Quantitative data: It refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it's quantitative data. Quantitative data can tell us "How many,""how much," or "how often"—for example, how many people attended last week's webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking?

Qualitative data: It cannot be measured or counted. It's descriptive, expressed in terms of language rather than numerical values.

Researchers will often turn to qualitative data to answer "Why?" or "How?" questions. For example, if our quantitative data tells us that acertain website visitor abandoned their shopping cart three times in one week, we'd probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps we want to know how a user feels about a particular product;again, qualitative data can provide such insights. In this case, we're not just looking at numbers; we're asking the user to tell us, using language, why they did something or how they feel.

Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labelling a particular ice cream flavour as vanilla.

Q-10) How to calculate range and interquartile range?

Answer-

To calculate the **range,** we need to find the maximumvalue of a variable and subtract the minimum value. 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.

The interquartile range (IQR) give a better idea of the dispersion of data. To calculate it, we need to know the values of the lower and upper quartiles. The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged inincreasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles.

Formula: IQR = Q3 - Q1

Q-11) What do you understand by bell curve distribution ?

Answer-

A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side. Bell curves are visual representations of normal distribution, also called Gaussian distribution.

A normal distribution curve, when graphed out, typically follows a bell-shaped curve, hence the name. While the precise shape can vary according to the distribution of the population, the peak is always in the middle and the curve is always symmetrical.

Bell curves are useful for quickly visualizing a data set's mean, mode and median because when the distribution is normal, the mean,median and mode are all the same.

The long tail refers to the part of the bell curve that stretches out in either direction. If the diagram above represents a population under study, the fat area under the bell curve is where most of the population falls.

Q-12) Mention one method to find outliers.

Answer-

Outliers are data points that are far from other datapoints. 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.

There are four ways to identify/detect outliers:

- Sorting method.
- Data visualization method.
- Statistical tests (z scores)
- Interquartile range method.

Z-scores:

Z-scores can quantify the unusualness of an observation when our data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls.

For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies itis two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.

To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation.

Formula of Z-score:

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation

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

Answer-

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 tounderstand 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. That's pretty tiny. On the otherhand, 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 our results.

Q-14) What is the Binomial Probability Formula?

Answer-

The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a singleoutcome such as success/failure is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process. Consider an experiment where each time a question is asked for a yes/no with a series of n experiments. Then in the binomial probability distribution, the Boolean-valued outcome the success/yes/true/one is represented with probability p and the failure/no/false/zero with probability q (q = 1 − p). In a single experiment when n = 1, the binomial distribution is called a Bernoulli distribution.

Formula for binomial distribution:

$$P(X) = {}_nC_x\, p^x (1-p)^{n-x}$$

Where,

- n = the number of experiments
- x = 0, 1, 2, 3, 4, …
- p = Probability of success in a single experiment
- q = Probability of failure in a single experiment (= 1 – p)

Q-15) Explain ANOVA and it's applications

Answer-

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.

Analysis of variance, or ANOVA, is a statistical method that separatesobserved variance data into different components to use foradditional tests.

Types of ANOVA:

- **One Way ANOVA**– It is also known as one factor ANOVA. Here, we are using one criterion variable (or called as a factor) and analyze the difference between more than two sample groups. Suppose in glass industry, we want to compare the variation of three batches (glass) for their average weight (factor).

- **Two Way ANOVA**– Here, we are using two independent variables (factors) and analyze the difference between more than two sample groups. Similarly, we want to compare the variation of three batches of glass w.r.t weight and hardness (two factors).

The Formula for ANOVA is:

$$F = MST/MSE$$

where:

F=ANOVA coefficient

MST=Mean sum of squares due to treatmentMSE=Mean sum of squares due to error