STATISTICS WORKSHEET – 4

Q1to Q15 are descriptive types. Answer in brief.

1. **What is the central limit theorem and why is it important?**
   According to Central Limit Theorem, for sufficiently large samples with size greater than 30, the shape of the sampling distribution will become more and more like a normal distribution, irrespective of the shape of the parent population. This theorem explains the relationship between the population distribution and sampling distribution. It highlights the fact that if there are large enough set of samples then the sampling distribution of mean approaches normal distribution.

   The significance of the central limit theorem lies in the fact that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample.

2. **What is sampling? How many sampling methods do you know?**
   When we conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research. To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. There are two types of sampling methods:

   - **Probability sampling –** It means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce a result that is representative of the whole population, Probability sampling is the most valid choice.

   - **Non-probability sampling –** In this technique, individuals are selected based on non-random criteria, and not every individual has a chance to be selected. This type of sampling is easier and cheaper to access, but it has a higher risk of sampling bias. This sampling does not aim to test a hypothesis about the broad population but to develop an initial understanding of a small or under-researched population.

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

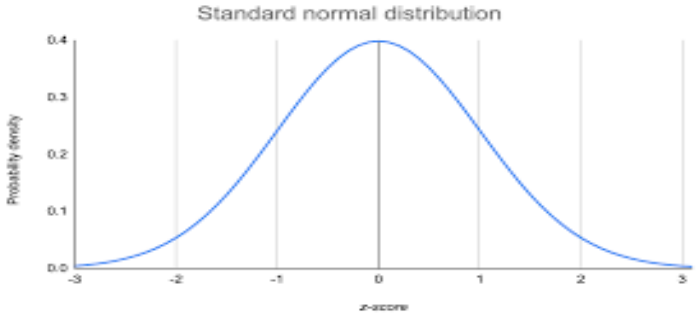
A Type I error (Alpha)(false-positive) means rejecting the null hypothesis when it's actually true.

A Type II error (Beta)(false-negative) means not rejecting the null hypothesis when it's actually false.

Example: Type I error (false positive): the test result says you have coronavirus, but you actually don't. Type II error (false negative): the test result says you don't have coronavirus, but you actually do.

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

A normal distribution is also called Gaussian distribution. In this distribution, the mean, median, and mode are all equal to one another. In the normal distribution, the mean is zero, the standard deviation is 1 and it has zero skewness. It is visually depicted as the "bell curve".



5. **What are 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.

<u>Covariance</u>:

- 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.

Formula –

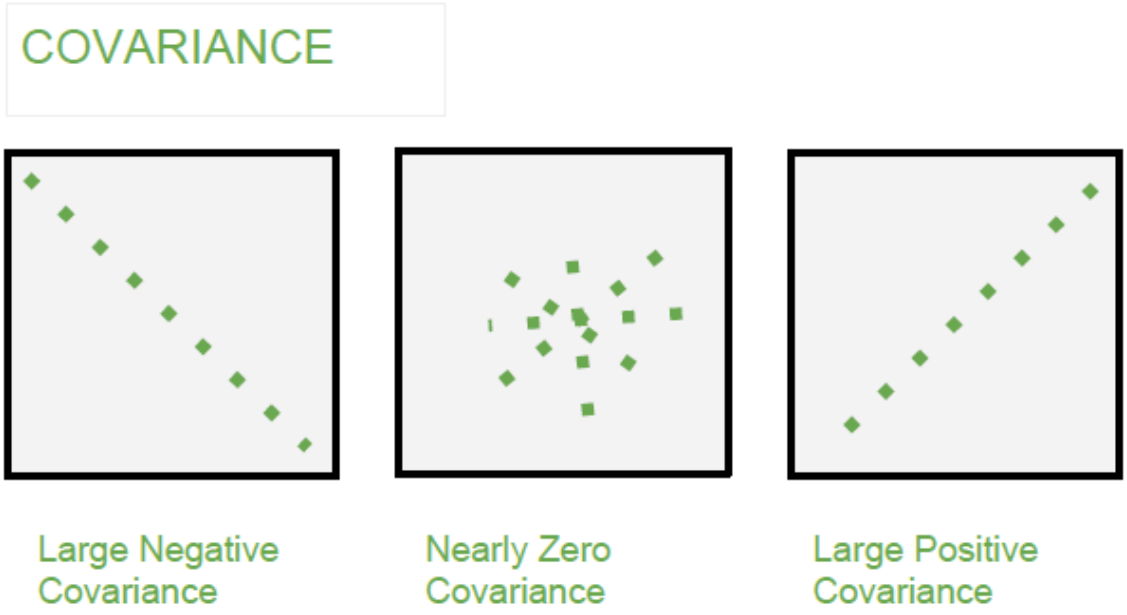
For Population:

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n}$$

For Sample:

$$Covari(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n-1}$$

Example –



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.

Formula –

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

Here,
x' and y' = mean of given sample set

n = total no of sample
xi and yi = individual sample of set

Example –

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

   Univariate analysis: Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. An example of univariate analysis may be height.

   Bivariate analysis: Bivariate analysis is used to find out if there is a relationship between two different variables. An example of bivariate analysis, may be a plot of calorie intake versus weight.

   Multivariate analysis: Multivariate analysis is the analysis of three or more variables.
   An example of multivariate 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.

7. **What do you understand by sensitivity and how would you calculate it?**
   Sensitivity (or Recall) aims to calculate the proportion of actual positive that was identified incorrectly. It can be calculated as True Positive or predictions that are actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

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

   Hypothesis testing is a statistical method that is used in making a statistical decision using experimental data. In other words, Hypothesis testing is basically an assumption that we make about the population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

   H0 - Null Hypothesis (Null hypothesis is a general given statement)

   H1 – Alternative Hypothesis (Alternative hypothesis is used in hypothesizing that is contrary to the null hypothesis).

   The two-sample t-test (Two tail test) compares the mean of two independent groups in order to determine the mean of two different variables are identical or not.

   Here,

   H0 – Two variables are independent.

   H1 – Two variables are dependent.

9. **What are quantitative data and qualitative data?**

   Quantitative data: Quantitative data can be counted, measured, and expressed using numbers. It gives the hard facts.
   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?, etc.

   Qualitative data: Qualitative data is descriptive and has categories and it is used to gain an understanding of human behavior, intentions, attitudes, experience, etc., based on the observation and the interpretation of the people.
   For example, if our quantitative data tells us that a certain website visitor abandoned their shopping cart three times in one week, we would probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps we would want to know how a user feels about a particular product; again, qualitative data can provide such

insights. In this case, we are not just looking at numbers; we are asking the user to tell us, using language, why they did something or how they feel.

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

    To calculate the range, you need to find the largest observed value of a variable (maximum) and subtract the smallest observed value (minimum).to calculate the range we need maximum value and minimum value.

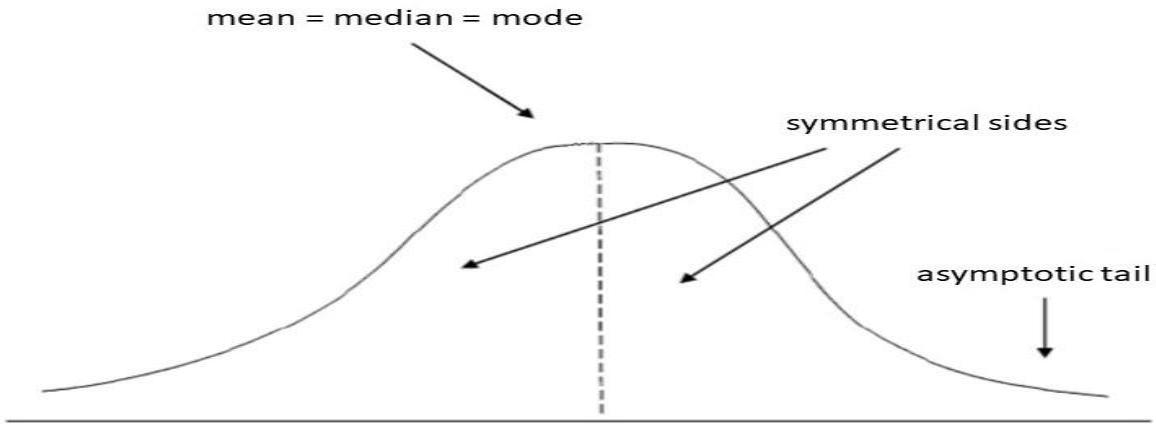
    Range(r) = Max – Min

    To calculate the interquartile range we need the 25th percentile value and the 75th percentile value.

    IQR = Q3 – Q1
    Q1 = 25th percentile value
    Q3 = 75th percentile value

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

    A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean.



12. **Mention one method to find outliers.**

Z-score is one of the methods to find the outliers. Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. Usually z-score =3 is considered a cut-off value to set the limit. Therefore, any z-score greater than +3 or less than -3 is considered an outlier.

Formula for calculating Z-Score:

$$z = (X - \mu) / \sigma$$

where,
z = Z-Score,
X = The value of the element,
μ = The population mean, and
σ = The population standard deviation

13. What is p-value in hypothesis testing?

The p-value or calculated probability is the probability of finding the observed/extreme results when the null hypothesis of a given study is true.
If,

- P-value > 0.05 – Null hypothesis is correct and accepted and alternative hypothesis is rejected.
- P-value < 0.05 – Null hypothesis is rejected and the alternative hypothesis is accepted.

14. What is the Binomial Probability Formula?

The binomial distribution is a commonly used discrete distribution in statistics. The normal distribution as opposed to a binomial distribution is a continuous distribution. The binomial distribution represents the probability for 'x' successes of an experiment in 'n' trials, given a success probability 'p' for each trial at the experiment.

The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a single outcome 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.

Binomial Distribution Formula:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
q = 1 - p = the probability of getting a failure in one trial

15. Explain ANOVA and it's applications.

The Anova test allows a comparison of two or more than two groups at the same time. It helps to determine whether a relationship exists between them or not. Anova is used to compare the difference of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found. Anova compares the amount of variation between groups with the amount of variation within the group.