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 (CLT) **states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution**. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.

   Central limit theorem **helps us to make inferences about the sample and population parameters and construct better machine learning models using them**. Moreover, the theorem can tell us whether a sample possibly belongs to a population by looking at the sampling distribution.

   Biologists use the central limit theorem **whenever they use data from a sample of organisms to draw conclusions about the overall population of organisms**. For example, a biologist may measure the height of 30 randomly selected plants and then use the sample mean height to estimate the population mean height.

2. What is sampling? How many sampling methods do you know?
   Ans:
   Sampling means **selecting the group that you will actually collect data from in your research**. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

Types of sampling: sampling methods

Samplingis of two types – probability sampling and non-probability sampling.

Probability sampling: Probability sampling is a sampling technique where a researcher sets a selection 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.

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.

There are four types of probability sampling techniques:

Simple random sampling: One of the best probability sampling techniques that helps in saving time and resources, is the Simple Random Sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.

Cluster sampling: Cluster sampling is a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback.

Systematic sampling: Researchers use the systematic sampling method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.

Stratified random sampling: Stratified random sampling is a method in which the researcher divides the population into smaller groups that don't overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.

Types of non-probability sampling with examples

Convenience sampling: This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as convenience sampling, because of the researcher's ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it's purely done based on proximity and not

representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

Judgmental or purposive sampling: Judgemental or purposive samples are formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience.

Snowball sampling: Snowball sampling is a sampling method that researchers apply when the subjects are difficult to trace. For example, it will be extremely challenging to survey shelterless people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed

Quota sampling: In Quota sampling, the selection of members in this sampling technique happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population.

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

Definition of Type I Error

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.

Definition of Type II Error

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

Key Differences Between Type I and Type II 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'.

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

Ans: 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.

5. What is correlation and covariance in statistics?
   Correlation: Correlation refers to the statistical relationship between two entities. In other words, it's how two variables move in relation to one another. Correlation can be used for various data sets, as well. In some cases, you might have predicted how things will correlate, while in others, the relationship will be a surprise to you. It's important to understand that correlation does not mean the relationship is causal.

- **Positive correlation:** A positive correlation would be 1. This means the two variables moved either up or down in the same direction together.
- **Negative correlation:** A negative correlation is -1. This means the two variables moved in opposite directions.
- **Zero or no correlation:** A correlation of zero means there is no relationship between the two variables. In other words, as one variable moves one way, the other moved in another unrelated direction.

Covariance: In statistics, covariance is **a measure of the relationship between two random variables**. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables.

covariance is measured in units. The units are computed by multiplying the units of the two variables. The variance can take any positive or negative values. The values are interpreted as follows:

Positive covariance: Indicates that two variables tend to move in the same direction.

- **Negative covariance**: Reveals that two variables tend to move in inverse directions.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.
   Ans: Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

Univariate Analysis

Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable. There are many different ways people use univariate analysis. The most common univariate analysis is checking the central tendency (mean, median and mode), the range, the maximum and minimum values, and standard deviation of a variable.

Common visual technique used for univariate analysis is a histogram, which is a frequency distribution graph. You could also use a box plot or violin plot to compare the spread of the variables and provides an insight into outliers. Using any of the above mentioned to compare the "*sepal_length*" in the iris dataset across species is only comparing one variable, therefore a Univariate analysis.

Bivariate Analysis

Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.

The most common visual technique for bivariate analysis is a scatter plot, where one variable is on the x-axis and the other on the y-axis. In addition to the scatter plot, regression plot and correlation coefficient are also frequently used to study the relationship of the variables. For example, continuing with the iris dataset, you can compare "*sepal_length*" vs "*sepal_width*" or "*sepal_length*" vs the "*petal_length*"to see if there is a relationship.

Multivariate Analysis

Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.

In terms of visualization, Seaborn library in Python allows for pairplots where it generates one large chart of selected variables against one another in a series of scatter plots and histograms depending on the type of variable, also known as scatter plot matrix. Again, in the series to come, I will provide the code and examples of this.

Depending on the dataset and the depth of analysis required, there are other techniques that you could deploy, such as Principal Component Analysis or logistic regression, linear regression, cluster analysis, etc. Again, in the series to come, I will provide the code and examples of this and dive deeper into PCA and its importance in data.

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

Ans: Sensitivity (equivalent to the True Positive Rate): **Proportion of positive cases that are well detected by the test**. In other words, the sensitivity measures how the test is effective when used on positive individuals.

Sensitivity – the proportion of people with the disease who tested positive compared to the number of all the people with the disease, regardless of their test result.

To calculate sensitivity, we'll need:

- Number of true positive cases (TP); and
- Number of false negative cases (FN).

And the following sensitivity equation:

```
Sensitivity = TP / (TP + FN)
```

TP + FN = Total number of people with the disease; and

TN + FP = Total number of people without the disease.

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

Ans: Hypothesis testing refers to **the predetermined formal procedures used by statisticians to determine whether hypotheses should be accepted or rejected**. The process of selecting hypotheses for a given probability distribution based on observable data is known as hypothesis testing.

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.

Null Hypothesis and Alternate Hypothesis

The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.

H0 is the symbol for it, and it is pronounced H-naught.

The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.

Problem Statement: The average height of students in a batch is 100 cm and the standard deviation is 15. However, Tedd believes that this has changed, so he decides to test the height of 75 random students in the batch. The average height of the sample comes out to be 105. Is there enough evidence to suggest that the average height has changed?

1. Specify the Null(H0) and Alternate(H1) hypothesis

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.

$H_0 : \mu = 100$

Alternate hypothesis (H1): The alternate hypothesis is always what is being **claimed**. “In our case, Tedd believes(**Claims**) that the actual value has changed”. He doesn’t know whether the average has gone up or down, but he believes that it has changed and is not 100 anymore.

$H_1: \mu \neq 100$

9. What is quantitative data and qualitative data?

Ans: 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.

Qualitative data **describes qualities or characteristics**. It is collected using questionnaires, interviews, or observation, and frequently appears in narrative form. For example, it could be notes taken during a focus group on the quality of the food at Cafe Mac, or responses from an open-ended questionnaire.

10. How to calculate range and interquartile range?

Ans:

The range is calculated by subtracting the lowest value from the highest value.

Calculate the range

The formula to calculate the range is:

$$R = H - L$$

- R = range
- H = highest value
- L = lowest value

The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), **first find the median (middle value) of the lower and upper half of the data**. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1

he formula for finding the interquartile range takes the third quartile value and subtracts the first quartile value. **IQR = Q3 – Q1**. Equivalently, the interquartile range is the region between the 75th and 25th percentile (75 – 25 = 50% of the data).

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.

- 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.

Bell curves (normal distributions) are used commonly in statistics, including in analyzing economic and financial data.

Understanding a Bell Curve

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.

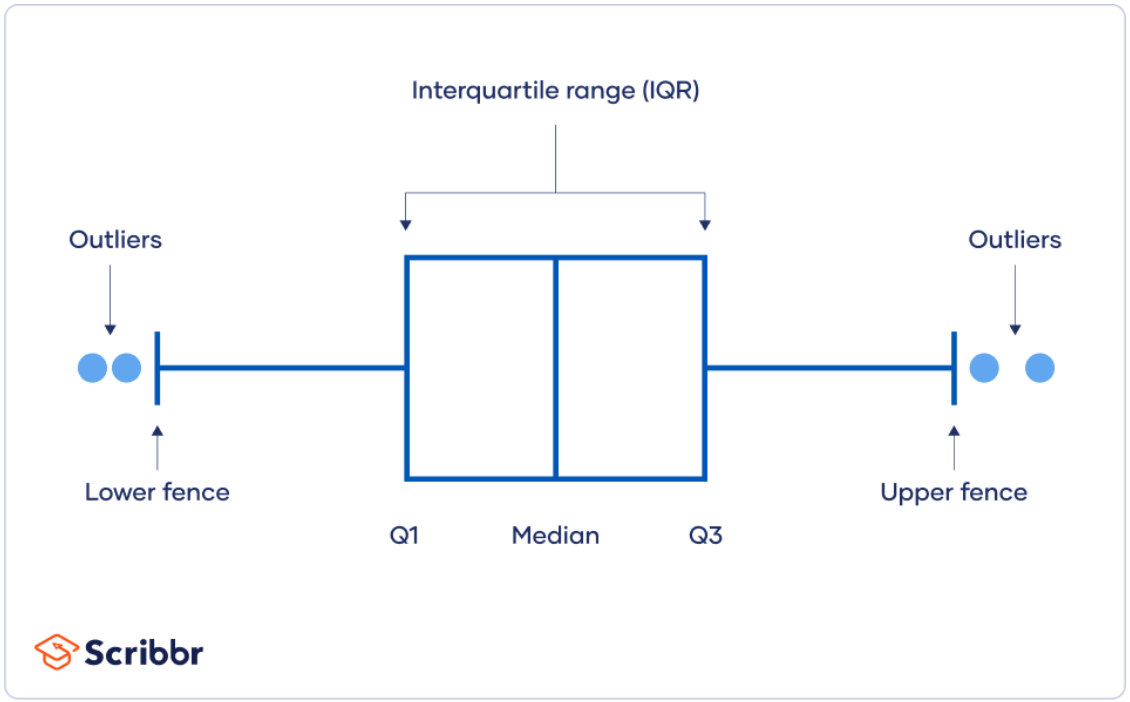
12. Mention one method to find outliers.

There are four ways to identify outliers:

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

Using the interquartile range

The **interquartile range** (IQR) tells you the range of the middle half of your dataset. You can use the IQR to create “fences” around your data and then define outliers as any values that fall outside those fences.



This method is helpful if you have a few values on the extreme ends of your dataset, but you aren’t sure whether any of them might count as outliers.

Interquartile range method

1. Sort your data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate your IQR = Q3 – Q1
4. Calculate your upper fence = Q3 + (1.5 * IQR)
5. Calculate your lower fence = Q1 – (1.5 * IQR)
6. Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

13. What is p-value in hypothesis testing?

Ans: The p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis

The smaller the *p*-value, the more likely you are to reject the null hypothesis.

14. What is the Binomial Probability Formula?

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

Binomial Distribution Formula	
Binomial Distribution	$P(x) = {}^{n}C_{x} \cdot p^{x} (1 - p)^{n-x}$
Or,	$P(r) = [n!/r!(n-r)!] \cdot p^{r} (1 - p)^{n-r}$

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.

15. . Explain ANOVA and it's applications

Ans: Analysis of variance, or ANOVA, is **a statistical method that separates observed variance data into different components to use for additional tests**. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables..

Applications of Anova:

. This is particularly applied in experiment otherwise difficult to implement such as in the case of clinical trails. compare the heights of plants with or with out galls

Compare birth weights of deer in different geographical regions.

Compare responses of patients to real medications vs placebo.

Compare attention spans of undergraduate students in different programs at PC