

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 Theroem 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 sample from that population will be roughly equal to the populationmean. The**Central Limit Theorem (CLT)** is a mainstay of**statistics**and **probability.** The theorem expresses that as the size of the sample expands, the distribution of the mean among multiple samples will be like a **Gaussian distribution.**

The CLT gives us a certain distribution over our estimations. We can utilize this to pose an inquiry about the probability of an estimate that we make.

1. **What is sampling? How many sampling methods do you know?**

**Sampling:**

Sampling is a process where researchers take predetermined number of observation from a larger population, or it is a technique to selecting individual member or subset of the population to make statistical inferences from them and estimate the characteristics of the whole population.

**Types of sampling method:**

1. **Probability sampling**
2. **Non Probability sampling**

**Probability Sampling:** The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of the population.

There are four parts of Probability Sampling:

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

**Non Probability sampling:** The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.

1. Convenience sampling
2. Consecutive sampling
3. Quota sampling
4. Snowball sampling
5. **What is the difference between type1 and typeII error?**

In [statistics](https://www.scribbr.com/?cat_ID=34372), a **Type I error** is a false positive conclusion, while a **Type II error** is a false negative conclusion.

Example: Type I vs Type II error

* **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.

1. **Differentiate between univariate ,Biavariate,and multivariate analysis.**

Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.

The simplest technique for analyzing quantitative data is called univariate analysis. In univariate analysis, there is just one reliable variable, as the name "Uni," which means "one," suggests. Inferences are made and the hypothesis is tested using it. The goal is to gather data, summaries and describe it, and look for patterns in it. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. The example of a univariate data can be height.

In Bivariate Analysis, there are two variables wherein the analysis is related to cause and the relationship between the two variables. This sort of data analysis examines relationships and causes, and it seeks to understand how the two variables are related. The temperature and ice cream sales during the summer is an examples of bivariate data analysis.

Multivariate data refers to data that has three or more variables. For instance, if an online marketer wanted to compare the popularity of four ads, they could analyses the click rates for men and women and then look at the correlations between the variables.

Plots including count plots, histograms, density curves, and distribution plots are used to visualise univariate analyses. Bar plots, scatter plots, joint plots, strip plots, and other types of plots can be used to visualise bivariate analyses. By adding hued data as an indication into the bivariate plots, multivariate analysis charts are created.

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

Sensitivity is the percentage of true positives (e.g. 90% sensitivity = 90% of people who have the target disease will test positive. The sensitivity of a test is the proportion of people who **test positive** among all those who actually have the disease. Mathematically, this can be stated as:

# Sensitivity=TP/ TP+FN

1. **What is correlation and covariance in statistics?**

Correlation is a statistical measure that measures the degree to which two or more random variables move sequentially. The variables are said to be correlated when, during the study of two variables, a comparable movement of one variable reciprocates the movement of the other variable in some manner. It is a statistical measure that indicates how strongly two variables are related. Correlation is limited to values between the range -1 and +1. The formula for correlation is:

# r= (n Σxy – Σx Σy) / √ (n Σx2 – ( Σx)2) (n Σy2 – ( Σy)2)

|  |  |
| --- | --- |
| 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 |

A change in one variable reflects a change in the other, which is referred to statistically as covariance. Covariance describes a systematic link between two random variables. A negative number for the covariance value indicates a negative association, whereas a positive value indicates a positive link. The covariance value can vary from -∞ to +∞. The relationship is more dependent the higher this value is. A positive figure for positive covariance indicates a direct relationship. An inverse link between the two variables is indicated by a negative value, which signifies negative covariance. Covariance is excellent at identifying the type of relationship, but it's terrible at determining its magnitude. The formula is :

# Cov(x,y)= Σ ((xi - x̄ ) (yi - ȳ )) / N

Where,

* xi = data value of x
* yi = data value of y
* x̄ = mean of x
* ȳ = mean of y
* N = number of data values.

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

[Hypothesis](https://www.merriam-webster.com/dictionary/Hypothesis) testing is a form of [statistical inference](https://www.britannica.com/science/inference-statistics) that uses data from a sample to draw conclusions about a population [parameter](https://www.britannica.com/topic/parameter) or a population [probability distribution](https://www.britannica.com/science/distribution-function).

A null hypothesis (H0) is a type of statistical hypothesis that proposes that no [statistical significance](https://www.investopedia.com/terms/s/statistical-significance.asp) exists in a set of given observations. [Hypothesis testing](https://www.investopedia.com/terms/h/hypothesistesting.asp) is used to assess the credibility of a hypothesis by using sample data. Sometimes referred to simply as the "null," it is represented as H0. Whatever information that is against the stated null hypothesis is captured in the alternative hypothesis (H1). The parameter or distribution is initially assumed in a preliminary manner. The null hypothesis, or H0, is what is meant by this assumption. Then, the opposite of what the null hypothesis claims is characterized as an alternative hypothesis (designated H1). Utilizing sample data to assess whether or not H0 can be rejected is a part of the hypothesis-testing technique. The alternative hypothesis, H1, is likely to be true if H0 is rejected, according to statistical analysis.

A two-tailed test in statistics determines if a sample is more than or less than a specific range of values by using a two-sided critical area of a distribution. It is used in testing the null hypothesis and determining statistical significance.

1. **What is quantitative data and qualitative data?**

Non-statistical qualitative data is usually unstructured or semi-structured. Hard numbers are not always employed to quantify this data in order to create graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers. In other hand, Quantitative data, as contrast to qualitative data, is statistical in nature and often structured, making it more rigid and defined. This data type is better suited for data analysis because it is measured using numbers and values.

A statistical and numerical analysis of numerical and statistical data (numbers and statistics) is quantitative research. On the other hand, open-ended and non-numerical data (concepts, descriptions, meanings, words, and more) are the focus of qualitative research.

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

The [range](https://www.scribbr.com/statistics/range/) gives you the spread of the whole data set. The range is calculated by subtracting the lowest value from the highest value.

The interquartile range in descriptive statistics describes the spread of the middle half of the distribution. Any distribution that is sorted from low to high is divided into four equal portions using quartiles. The second and third quartiles, or the center half of the data set, are contained in the interquartile range.

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

A bell curve is a form of graph that is used to show how a collection of selected values are distributed; it often has a peak at the center that tends to be normal, with low and high extremes tapering out rather symmetrically on either side. The normal distribution, commonly known as the Gaussian distribution, is visually represented as bell curves. When graphed out, a normal distribution curve often has a bell-shaped appearance, hence the name. The peak is always in the center and the curve is always symmetrical, although the specific shape may change depending on the population distribution.

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

Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your [statistical analyses](https://www.scribbr.com/category/statistics/) and skew the results of any [hypothesis tests](https://www.scribbr.com/statistics/hypothesis-testing/). There are 4 different ways to identify the outliers. They are: [Sorting method,](https://www.scribbr.com/statistics/outliers/#sorting) [Data visualization method](https://www.scribbr.com/statistics/outliers/#visualizations), [Statistical tests (z scores)](https://www.scribbr.com/statistics/outliers/#statistical) and Interquartile range method. Let’s discuss [**Statistical tests (z scores)**](https://www.scribbr.com/statistics/outliers/#statistical) **method**.

Applying statistical tests or techniques to find extreme values is the process of statistical outlier detection. Extreme data points can be transformed into z scores that indicate how far they deviate from the mean. A value can be classified as an outlier if its z score is sufficiently high or low. Generally speaking, values with a z score of larger than 3 or lower than -3 are considered as outliers.

1. **What is p-value in hypothesis testing?**

In [statistical hypothesis testing,](https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics) P-Value or probability value can be defined as the measure of the probability that a real-valued test statistic is at least as extreme as the value actually obtained. P-value shows how likely it is that the set of observations could have occurred under the null hypothesis. P-Values are used in statistical hypothesis testing to determine whether to reject the null hypothesis. The smaller the p-value, the stronger the likelihood that you should reject the null hypothesis.

1. **What is the Binomial Probability Formula?**

The binomial distribution formula aids in determining the likelihood that "x" successes will occur in "n" separate trials of a binomial experiment. Recall that there are two possible outcomes for the statistics probability distribution known as the binomial distribution. The binomial distribution in probability theory has two parameters, n and p. The formula for the binomial probability distribution is as stated below:

# P(X) = (n! / (n-X)! X!) \* (p)^X \* (q)^n-X

* + n = Total number of events
  + X = Total number of successful events.
  + p = Probability of success on a single trial.
  + q = 1 – p = Probability of failure.

1. **Explain ANOVA and it’s applications.**

By dividing the observed aggregate variability within a data set into systematic factors and random factors, the Analysis of variance (ANOVA) is a statistical analysis method. The random factors have no statistical impact on the presented data set, whereas the systematic factors do. The ANOVA test is used by analysts to ascertain how independent factors in a regression analysis affect the dependent variable. The formula is :

## F= MST/MSE

F=ANOVA coefficient

MST=Mean sum of squares due to treatment MSE=Mean sum of squares due to error

## Applications of ANOVA:

* + ANOVA (Analysis of Variance) is used when we have more than two sample groups and determine whether there are any statistically significant differences between the means of two or more independent sample groups. Suppose in the Manufacturing Process, we want to compare and check which are the most reliable procedures, materials, etc. We can use the ANOVA test to compare different suppliers and select the best available.
  + Understanding the impact of different catalysts on chemical reaction rates
  + Comparing the gas mileage of different vehicles, or the same vehicle under different fuel types, or road types.

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

A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution. A Gaussian distribution or probability bell curve are other names for the normal distribution. Because it is symmetric around the mean, it shows that values close to the mean happen more frequently than those distant from the mean. For a normally distributed feature 68.26% of the data lies in the 1st standard deviation, 95.44% of the data lies in the 2nd standard deviation area and 99.73% of data lies within 3 standard deviation of the feature.

1. .
2. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
3. What is quantitative data and qualitative data?

How to calculate range and interquartile range?

1. What do you understand by bell curve distribution ?
2. Mention one method to find outliers.
3. What is p-value in hypothesis testing?

