**STATISTICS WORKSHEET-4**

**Q1to Q15 are descriptive types. Answer in brief.**

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

ANS: The CLT 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. 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.

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.

Sampling Methods are:

**Random Sampling:** With [random sampling](https://www.investopedia.com/terms/s/simple-random-sample.asp), every item within a population has an equal probability of being chosen.

### Judgement Sampling: Auditor judgement may be used to select the sample from the full population. An auditor may only be concerned about transactions of a material nature.

### Block Sampling: Block sampling takes a consecutive series of items within the population to use as the sample.

### [Systematic sampling](https://www.investopedia.com/terms/s/systematic-sampling.asp): begins at a random starting point within the population and uses a fixed, periodic interval to select items for a sample.

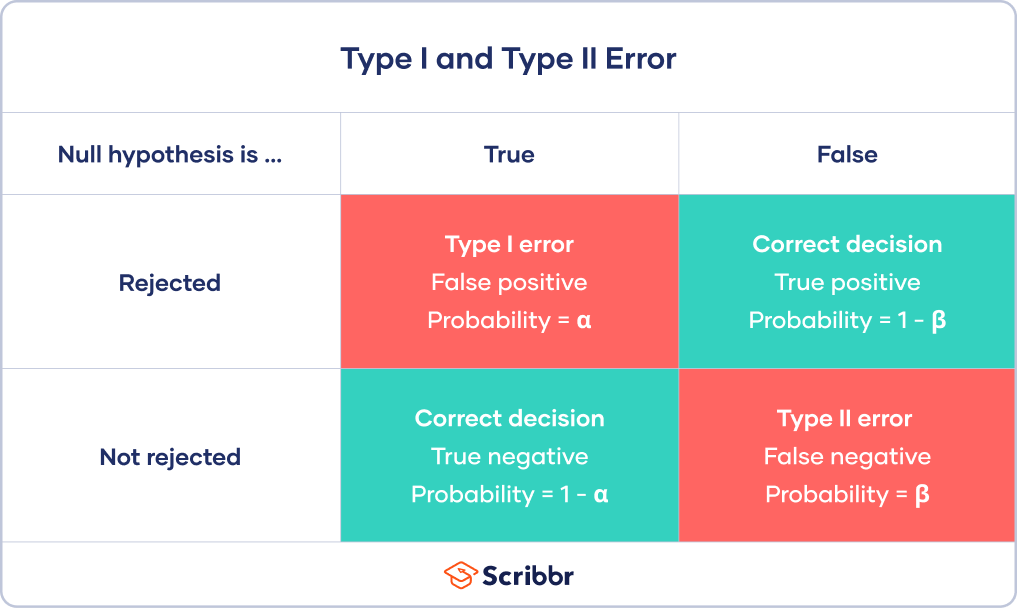
3. What is the difference between type1 and typeII error?

ANS: 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.

Making a statistical decision always involves uncertainties, so the risks of making these errors are unavoidable in [hypothesis testing](https://www.scribbr.com/statistics/hypothesis-testing/).

The probability of making a Type I error is the [significance level](https://www.scribbr.com/statistics/statistical-significance/#significance-level), or alpha (α), while the probability of making a Type II error is beta (β). These risks can be minimized through careful planning in your study design.

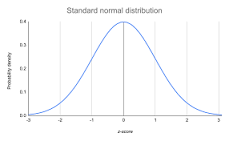
Type I error (false positive): the test result says you have coronavirus, but you actually don’t.

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?

ANS: 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.

Example: Using the empirical rule in a normal distribution You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150.



5. What is correlation and covariance in statistics?

ANS: Covariance is a measure to indicate the extent to which two random variables change in tandem. Correlation is a measure used to represent how strongly two random variables are related to each other. Covariance is nothing but a measure of correlation. Correlation refers to the scaled form of covariance.

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.

1. Univariate data –  
This type of data consists of only one variable. 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 patterns that exist within it. The example of a univariate data can be height.

2. Bivariate data –  
This type of data 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.Example of bivariate data can be temperature and ice cream sales in summer season.

3. Multivariate data –  
When the data involves three or more variables, it is categorized under multivariate. 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 and women and relationships between variables can then be examined.

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

ANS: Sensitivity analysis works on the simple principle: **Change the model and observe the behavior.**

The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

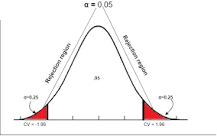
This process of testing sensitivity for another input (say cash flows growth rate) while keeping the rest of inputs constant is repeated until the sensitivity figure for each of the inputs is obtained. The conclusion would be that the higher the sensitivity figure, the more sensitive the output is to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

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

ANS: In a jury trial the hypotheses are: **H0: defendant is innocent; • H1: defendant is guilty**. H0 (innocent) is rejected if H1 (guilty) is supported by evidence beyond “reasonable doubt.” Failure to reject H0 (prove guilty) does not imply innocence, only that the evidence is insufficient to reject it.

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. Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed.

9. What is quantitative data and qualitative data?

ANS:  Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

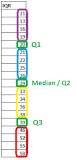
Quantitative data is data expressing a certain quantity, amount or range. Usually, there are measurement units associated with the data, e.g. metres, in the case of the height of a person. It makes sense to set boundary limits to such data, and it is also meaningful to apply arithmetic operations to the data.

10. How to calculate range and interquartile range?

ANS: 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.

The formula for finding the interquartile range takes the third quartile value and subtracts the first quartile value. 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 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.

Look at the symmetrical shape of a bell curve. The center should be where the largest portion of scores would fall. The smallest areas to the far left and right would be where the very lowest and very highest scores would fall. Read across the curve from left to right.

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 your [statistical analyses](https://www.scribbr.com/category/statistics/) and skew the results of any [hypothesis tests](https://www.scribbr.com/statistics/hypothesis-testing/).

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

Outliers are values at the extreme ends of a dataset.

Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry, equipment malfunctions, or other [measurement errors](https://www.scribbr.com/methodology/random-vs-systematic-error/).

An outlier isn’t always a form of dirty or incorrect data, so you have to be careful with them in [data cleansing](https://www.scribbr.com/methodology/data-cleansing/). What you should do with an outlier depends on its most likely cause.

**Sorting method**

You can sort [quantitative variables](https://www.scribbr.com/methodology/types-of-variables/#quantitative-variables) from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.

This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

Example: Sorting methodYour dataset for a pilot experiment consists of 8 values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 180 | 156 | 9 | 176 | 163 | 1827 | 166 | 171 |

You sort the values from low to high and scan for extreme values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 9 | 156 | 163 | 166 | 171 | 176 | 180 | 1872 |

**Using visualizations**

You can use software to visualize your data with a box plot, or a box-and-whisker plot, so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the [range](https://www.scribbr.com/statistics/range/)), the [median](https://www.scribbr.com/statistics/median/), and the interquartile range for your data.

Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

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

In the P-Value approach to hypothesis testing, a calculated probability is used to decide if there’s evidence to reject the null hypothesis, also known as the conjecture. The conjecture is the initial claim about a data population, while the alternative hypothesis ascertains if the observed population parameter differs from the population parameter value according to the conjecture.

Effectively, the significance level is declared in advance to determine how small the P-value needs to be such that the null hypothesis is rejected.  The levels of significance vary from one researcher to another; so it can get difficult for readers to compare results from two different tests. That is when P-value makes things easier.

14. What is the Binomial Probability Formula?

ANS: Binomial [probability](https://www.varsitytutors.com/hotmath/hotmath_help/topics/probability.html)refers to the probability of exactly xx successes on nn repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment).

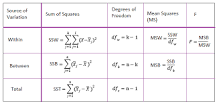
If the probability of success on an individual trial is pp , then the binomial probability is nCx⋅px⋅(1−p)n−xnCx⋅px⋅(1−p)n−x .

Here nCxnCx indicates the number of different [combinations](https://www.varsitytutors.com/hotmath/hotmath_help/topics/combinations.html)of xx objects selected from a set of nn objects. Some textbooks use the notation (nx)(nx) instead of nCxnCx .

Note that if pp is the probability of success of a single trial, then (1−p)(1−p) is the probability of failure of a single trial

15. Explain ANOVA and it’s applications.

ANS: Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.



The Anova test is performed by comparing two types of variation, the variation between the sample means, as well as the variation within each of the samples. The below mentioned formula represents one-way Anova test statistics: Alternatively, F = MST/MSE. MST = SST/ p-1.