

STATISTICS WORKSHEET-4

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

1. What is central limit theorem and why is it important?
2. What is sampling? How many sampling methods do you know?
3. What is the difference between type1 and typeII error?
4. What do you understand by the term Normal distribution?
5. What is correlation and covariance in statistics?
6. Differentiate between univariate ,Biavariate,and multivariate analysis.
7. What do you understand by sensitivity and how would you calculate it?
8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
9. What is quantitative data and qualitative data?
10. How to calculate range and interquartile range?
11. What do you understand by bell curve distribution ?
12. Mention one method to find outliers.
13. What is p-value in hypothesis testing?
14. What is the Binomial Probability Formula?
15. Explain ANOVA and it’s applications.
16. 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.

This is useful since the researcher never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from a population, the sample means will cluster together, allowing the researcher to obtain a very accurate estimate of the population mean.

1. When one conducts 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, one has to carefully decide how to select a sample that is representative of the group as a whole. There are two types of sampling methods:

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

1. A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population
2. A normal distribution is **an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme**.
3. Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable. In statistics, correlation is a measure that determines the degree to which two or more random variables move in sequence. When an equivalent movement of another variable
4. WORKSHEET

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

1. The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as **sensitive analysis**. It’s usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond’s price. Sensitivity is calculated by **dividing the number of true-positive results by the total number of positives** (which include false positives)

Sensitivity=[a/(a+c)]×100Specificity=[d/(b+d)]×100Positive predictive value(PPV)=[a/(a+b)]×100Negative predictive value(NPV)=[d/(c+d)]×100.

1. In hypothesis testing there are two mutually exclusive hypotheses; **the Null Hypothesis (H0) and the Alternative Hypothesis (H1)**. One of these is the claim to be tested and based on the sampling results (which infers a similar measurement in the population), the claim will either be supported or not.
2. 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 colors or emotions
3. To calculate the range, you need to find the largest observed value of a variable (the maximum) and subtract the smallest observed value (the minimum). 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 and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, you 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 in increasing 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. The semi-interquartile range is half the interquartile range.

1. The term bell curve is used to describe the mathematical concept called **normal distribution**, sometimes referred to as Gaussian distribution. 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.
2. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR.. An outlier is defined as being any point of data that lies over 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3)in a data set.  
   High = (Q3) + 1.5 IQR  
   Low = (Q1) – 1.5 IQR
3. 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
4. Binomial probability refers to the probability of exactly x successes on n 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 p , then the binomial probability is **nCx⋅px⋅(1−p)n−x** .
5. 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.