Statistics Worksheet

1.Ans -a

2.Ans-d

3.Ans-b

4.Ans-d

5.Ans-c

6.Ans-b

7.Ans -b

8.Ans-a

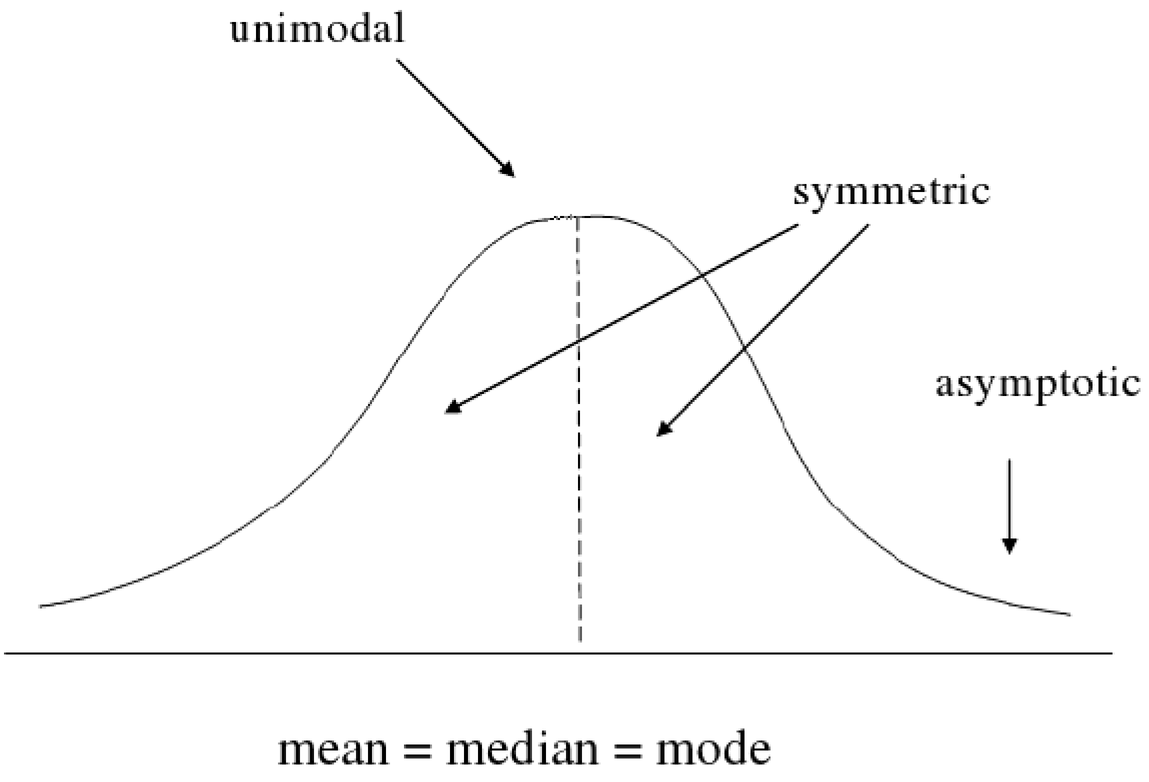
9.Ans -c

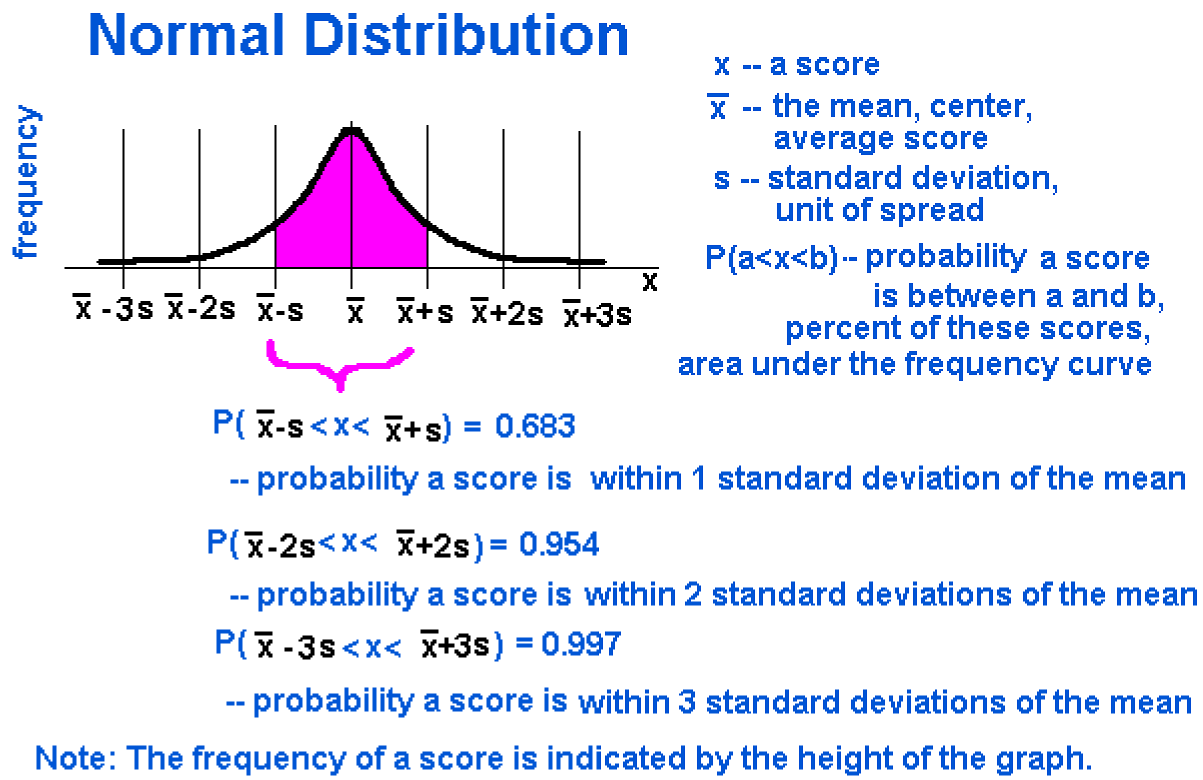
10.Ans -Normal Distribution: -

Normal distribution, also known as the Gaussian distribution, is **a probability distribution that is symmetric about the mean**, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

Characteristics of Normal Distribution :-

Normal distributions are **symmetric, unimodal, and asymptotic, and the mean, median, and mode are all equal**. A normal distribution is perfectly symmetrical around its center. That is, the right side of the center is a mirror image of the left side. There is also only one mode, or peak, in a normal distribution. Normal distributions are continuous and have tails that are asymptotic, which means that they approach but never touch the x-axis. The **center of a normal distribution** is located at its peak, and 50% of the data lies above the mean, while 50% lies below. It follows that the mean, median, and mode are all equal in a normal distribution.





x -- a score

$\bar{x}$ -- the mean, center, average score

s -- standard deviation, unit of spread

P(a<x<b) -- probability a score is between a and b, percent of these scores, area under the frequency curve

$P(\bar{x}-s < x < \bar{x}+s) = 0.683$

-- probability a score is within 1 standard deviation of the mean

$P(\bar{x}-2s < x < \bar{x}+2s) = 0.954$

-- probability a score is within 2 standard deviations of the mean

$P(\bar{x}-3s < x < \bar{x}+3s) = 0.997$

-- probability a score is within 3 standard deviations of the mean

Note: The frequency of a score is indicated by the height of the graph.

11. Ans -Imputation techniques to handle missing data:

- Numerical Values :- a) Mean/Median Imputation c)End of tail imputation

  b)Arbitary Value Imputation d) Mode imputation

- Categorical Variable : - a) Frequency Category imputation

  b) Adding a missing category

- Mixed Variable(Both) : -a) Complete Case Analysis
  b) Adding a missing indicator
  c) Random sample imputation

Complete Case Analysis(CCA):-

This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. This method is also popularly known as "Listwise deletion".

- **Assumptions:-**
  - Data is Missing At Random(MAR).
  - Missing data is completely removed from the table.
- **Advantages:-**
  - Easy to implement.
  - No Data manipulation required.
- **Limitations:-**
  - Deleted data can be informative.
  - Can lead to the deletion of a large part of the data.
  - Can create a bias in the dataset, if a large amount of a particular type of variable is deleted from it.
  - The production model will not know what to do with Missing data.
- **When to Use:-**
  - Data is MAR(Missing At Random).
  - Good for Mixed, Numerical, and Categorical data.
  - Missing data is not more than 5% – 6% of the dataset.
  - Data doesn't contain much information and will not bias the dataset.

2. Arbitrary Value Imputation

This is an important technique used in Imputation as it can handle both the Numerical and Categorical variables. This technique states that we group the missing values in a column and assign them to a new value that is far away from the range of that column. Mostly we use values like 99999999 or -9999999 or "Missing" or "Not defined" for numerical & categorical variables.

- **Assumptions:-**
  - Data is not Missing At Random.
  - The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data.
- **Advantages:-**
  - Easy to implement.
  - We can use it in production.
  - It retains the importance of "missing values" if it exists.
- **Disadvantages:-**
  - Can distort original variable distribution.
  - Arbitrary values can create outliers.
  - Extra caution required in selecting the Arbitrary value.
- **When to Use:-**
  - When data is not MAR(Missing At Random).

- Suitable for All.

3. <u>Frequent Category Imputation</u>

This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column. This technique is also referred to as **Mode Imputation.**

- **Assumptions:-**
  - Data is missing at random.
  - There is a high probability that the missing data looks like the majority of the data.
- **Advantages:-**
  - Implementation is easy.
  - We can obtain a complete dataset in very little time.
  - We can use this technique in the production model.
- **Disadvantages:-**
  - The higher the percentage of missing values, the higher will be the distortion.
  - May lead to over-representation of a particular category.
  - Can distort original variable distribution.
- **When to Use:-**
  - Data is Missing at Random(MAR)
  - Missing data is not more than 5% – 6% of the dataset.

12 Ans – A/B Testing :-

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say a company want to increase the sales of product. Here, either we can use random experiments, or we can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, we may divide the products into two parts – A and B. Here A will remain unchanged while we make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

1. **Make a Hypothesis**

In hypothesis testing, we have to make two hypotheses i.e Null hypothesis and the alternative hypothesis. Let's have a look at both.

1. Null hypothesis or H_0:

   The **null hypothesis** is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is **no** difference between the control and variant groups. It states the default position to be tested or the situation as it is now, i.e. the status quo. Here our H_0 is " there is no difference in the conversion rate in customers receiving newsletter A and B".

2. **Alternative Hypothesis or** H_0**:**

The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

2. Create Control Group and Test Group :

For this experiment, select we randomly 1000 customers – 500 each for our Control group and Test group.

Random sampling is important in hypothesis testing because it eliminates sampling bias, and **it's important to eliminate bias because we want the results of your A/B test to be representative of the entire population rather than the sample itself.**

3. Conduct the A/B Test and Collect the Data for the specific testing time period.

13 Ans- The process of replacing null values in a data collection with the data's mean is known as **mean imputation.** Mean imputation is **typically considered terrible practice** since it ignores feature correlation. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14 Ans- ***Linear regression*** quantifies the relationship between one or more *predictor variable(s)* and one *outcome variable.* Linear regression is commonly used for predictive analysis and modelling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as *multiple regression*, *multivariate regression*, *ordinary least squares (OLS)*, and *regression*.

Example :- **Simple Linear Regression** : - For any company we want to predict sales(outcome variable) ,depending on advertisement expenditure (Predictor variable).

Multiple Linear Regression: For example if we want to predict salary(outcome variable) ,based on experience, technology ,location(metro cities) etc.Here exp,technology and location all are predictor variables.

15 Ans – Branches of Statistics :-

Descriptive Statistics :- Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential Statistics:- ,Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.