Statistics Worksheet

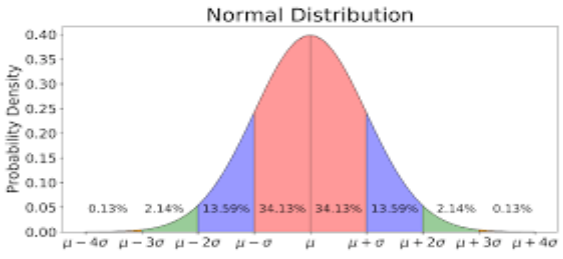
Q. Multiple choice question

1. **A**(True)
2. **A** (Central limit theorem)
3. **B** (Modelling bounded count data)
4. **D** (All above mentioned)
5. **C** (Poisson)
6. **B** (False)
7. **B** (Null hypothesis)
8. **A** (0)
9. **C** (Outliers cannot conform to the regression relationship)

Q. Briefly explain the following questions.

Q 10. what do you understand by term normal distribution?

1.The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables.
2.Most people recognize its familiar bell-shaped curve in statistical reports.
3.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.
4. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal.
5.As with any probability distribution, the normal distribution describes how the values of a variable are distributed.
6.It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena.
7.Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.



Q11. How do you handle missing data? What imputation techniques do you recommend?

1.Some common ways of handling missing values are Deletions and Imputations.

2.Deletions of Missing Values

3.Deleting data may be a crucial thing in Machine learning as a result of we tend to find ourselves losing data observations, trends, and patterns from one feature to another.

3.sometimes there is a lack of a lot of data observations and that data is not heavily influenced by target variables that we used for the analysis, which means we can choose the Data Deletion Techniques.

Two types of Deletions:

1.Listwise Deletions

2.Pairwise Deletions

a. It is recommended that these deletion techniques only be used when the data set contains fewer missing values.

b. Deleting Columns with Missing Values

Imputation of Missing Values: Imputation is that the method of substituting missing data with substituted values.

Two types of Imputations are majorly categorized

- General
- Time-Series

A. General Data: General data is mainly imputed by mean, mode, median, Linear Regression, Logistic Regression, Multiple Imputations, and constants.
Further General data is divided into two types Continuous and Categorical.

Imputation techniques:

1.Imputation Using (Mean/Median) Values:

This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

2.Imputation Using (Most Frequent) or (Zero/Constant) Values:

Most Frequent is another statistical strategy to impute missing values . It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column.

3.Dropping rows with null values:

The easiest and quickest approach to a missing data problem is dropping the offending entries. This is an acceptable solution if we are confident that the missing data in the dataset is missing at random, and if the number of data points we have access to is sufficiently high that dropping some of them will not cause us to lose generalizability in the models we build

Q12. What is A/B testing?

1.A/B testing is a basic randomized control experiment.

2. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

3. A/B testing is one of the most prominent and widely used statistical tools.

4.For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods.

5. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics.

6. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

Q13. Is mean imputation of missing data acceptable practice?

1.The process of replacing null values in a data collection with the data's mean is known as mean imputation.
2.Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.
3.Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.
4. By imputing the mean, you are able to keep your sample size up to the full sample size.

Q14. What is linear regression in statistics?

1. Linear regression analysis is used to predict the value of a variable based on the value of another variable.
2. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
3. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.
4. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.
5. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data.

6. Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions.
7. Linear regression can be applied to various areas in business and academic study.
8. You'll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business.
9. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

Q15. What are the various branches of statistics?

1.Descriptive Statistics:

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis.

2. 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.

3.Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time.

4.The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

2.Inferential Statistics:

1.Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics.

2. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

3.Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics.

4. 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.