WORKSHEET

STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

b) Modelling bounded count data

4. Point out the correct statement.

d) All of the mentioned

5. _____ random variables are used to model rates.

c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

b) False

7. Which of the following testing is concerned with making decisions using data?

b) Hypothesis

8. **Normalized data are centered at_____and have units equal to standard deviations of the original data.**

a) 0

9. **Which of the following statement is incorrect with respect to outliers?**

c) Outliers cannot conform to the regression relationship

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

Ans: The normal distribution is a form presenting data by arranging the probability distribution of each value in the data. Most values remain around the mean value making the arrangement symmetric. We use various functions in numpy library to mathematically calculate the values for a normal distribution. Histograms are created over which we plot the probability distribution curve.

Some excellent properties of a normal distribution:

- The mean, mode, and median are all equal.
- The total area under the curve is equal to 1.
- The curve is symmetric around the mean.

Empirical rule tells us that:

- 68% of the data falls within one standard deviation of the mean.
- 95% of the data falls within two standard deviations of the mean.
- 99.7% of the data falls within three standard deviations of the mean.

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

Data can have missing values for a number of reasons such as observations that were not recorded and data corruption. Handling missing data is important as many machine learning algorithms do not support data with missing values.There are machine learning algorithms that are robust with missing data. Some examples include:

- **kNN (k-Nearest Neighbor)**
- **Naïve Bayes**

Techniques to handle Missing Data

*** Find Missing Values**

Find how many missing values there are per column by running:

```
data.isnull().sum()
```

Mark Missing Values

Display the general statistical data for a dataset by running:

Drop Missing Values

The easiest way to handle missing values in Python is to get rid of the rows or columns where there is missing information.

Impute Missing Values

Imputation is a method of filling missing values with numbers using a specific strategy. Some options to consider for imputation are:

- A mean, median, or mode value from that column.
- A distinct value, such as 0 or -1.
- A randomly selected value from the existing set.
- Values estimated using a predictive model.

The Pandas DataFrame module provides a method to fill NaN values using various strategies. For example, to replace all NaN values with 0:

```
data.fillna(0)
```

12. What is A/B testing?

Ans: 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.

When Should We Use A/B Testing?

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not. A/B testing doesn't work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

13. Is mean imputation of missing data acceptable practice?

Ans: Bad practice in general. If just estimating means: mean imputation preserves the mean of the observed data. Leads to an underestimate of the standard deviation. Distorts relationships between variables by "pulling" estimates of the correlation.

14. What is linear regression in statistics?

Ans: linear regression is a statistical method that tries to show a relationship between variables. It looks at different data points and plots a trend line. A simple example of linear regression is finding that the cost of repairing a piece of machinery increases with time. Linear regression tries to model the relationship between two variables by applying a linear equation to the observed data. A linear regression line can be represented using the equation of a straight line:

$$y = mx + b$$

In this simple linear regression equation:

- **y** is the estimated dependant variable (or the output)
- **m** is the regression coefficient (or the slope)
- **x** is the independent variable (or the input)
- **b** is the constant (or the y-intercept)

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

15. What are the various branches of statistics?

Ans: Branches Of Statistics

Descriptive Statistics

Descriptive statistics is the first part of statistics that deals with the collection of data. People think it is too easy, but it is not that easy. The statisticians need to be aware of the design and experiments. They also need to select the correct focus group and keep away from biases. On the contrary, Descriptive statistics are used to do various kinds of analysis on different studies.

Descriptive statistics have two parts;

- Central tendency measures
- Variability measures

Measures of Central Tendency

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:mean.mediun,mode

Measures of Variability

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

Inferential Statistics

Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Different types of inferential statistics include:

- **Regression analysis:** It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables.
- **Analysis of variance (ANOVA):** ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the relationship between the dependent and independent variables.
- **Analysis of covariance (ANCOVA):** It is used to test categorical variables' main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The control variables are known as covariates.
- **Statistical significance (t-test):** It is used to determine a significant difference between the means of two groups related to particular features. A t-test studies the t-statistic, the t-distribution values, and the degree of freedom to learn the statistical significance.
- **Correlation analysis:** It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.