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
   b) False
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
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
5. _______random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
8. 4. Normalized data are centered at_______and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

FLIP ROBO

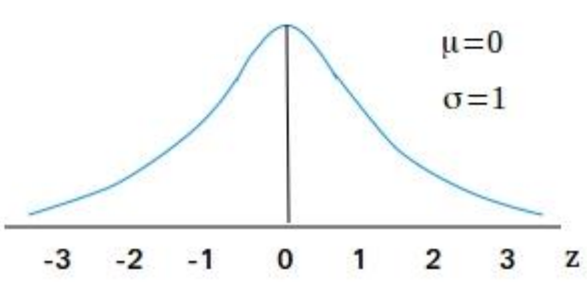
Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.

The normal distribution is often called the bell curve because the graph of its probability density looks like a bell.

A normal distribution is determined by two parameters the mean and the variance. A normal distribution with a mean of 0 and a standard deviation of 1 is called a standard normal distribution.


The standard normal distribution curve

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

When dealing with missing data, one can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Imputation Techniques to handle missing data - When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.

Instead of deletion, we have multiple solutions to impute the value of missing data. Depending upon, why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

1. **Mean, Median and Mode**

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, we can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data.

2. **Time-Series Specific Methods**

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

3. **Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)**

In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

4. **Linear Interpolation**

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

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

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you 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.

13. **Is mean imputation of missing data acceptable practice?**

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. Consider the following scenario: we have a table with age and fitness scores, and an eighty-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.

Second, 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.

14. **What is linear regression in statistics?**

Linear regression analysis is used to predict the value of a variable based on the value of another variable. 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.

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. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. **What are the various branches of statistics?**

The two main branches of statistics are Descriptive Statistics and Inferential Statistics. Both of these are employed in scientific analysis of data.

a) **Descriptive Statistics** - Deals with the presentation and collection of data. This is usually the first part of a statistical analysis. Descriptive statistics have two parts;

- **Central tendency** - measures specifically help statisticians evaluate the distribution center of values. These tendency measures are: Mean, Median, Mode
- **Variability** - measures 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.

b) **Inferential Statistics -** Involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. That is obtained by taking samples and testing how reliable they are.

- **Regression analysis:** It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear, and nonlinear. The most well-known models are simple linear and multiple linear.
- **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.

FLIP ROBO