STATISTICS WORKSHEET-1

1) A)True
2) A)Central Limit theorem
3) B)Modeling bounded count data
4) D)All of the mentioned
5) C)poisson
6) B)False
7) B)Hypothesis
8) A)0
9) D)None of the mentioned

Q10 to Q15 Subjective Answers

Q10) What do you understand by the term Normal Distribution?

Ans:

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.

The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

Common Properties for All Forms of the Normal Distribution:

Despite the different shapes, all forms of the normal distribution have the following characteristic properties.

They’re all symmetric bell curves. The Gaussian distribution cannot model skewed distributions.

The mean, median, and mode are all equal.

Half of the population is less than the mean and half is greater than the mean.

The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean. More on this below!

While the normal distribution is essential in statistics, it is just one of many probability distributions, and it does not fit all populations. To learn how to determine whether the normal distribution provides the best fit to your sample data, read my posts about How to Identify the Distribution of Your Data Assessing Normality: Histograms vs. Normal Probability Plots.

The uniform distribution also models symmetric, continuous data, but all equal-sized ranges in this distribution have the same probability, which differs from the normal distribution.

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

Ans:

Q12) What is A/B testing?

Ans:

A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

Q13) Is mean imputation of missing data acceptable practice?

Ans:

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.

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.

Q14) What is linear regression in statistics?

Ans:

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Types of Linear Regression

Simple linear regression
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

Logistic regression
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis

1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

Q15) What are the various branches of statistics?

Ans:

Descriptive Statistics and Inferential Statistics

Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

Submitted By: Netra Dalvi