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

b) False

Ans. (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

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans. (a) Central Limit Theorem

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

Ans. (b) Modeling bounded count data

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

Ans. (d) All of the mentioned

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans. (c) Poisson

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

a) True

b) False

Ans. (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

Ans. (b) Hypothesis

8. 4. Normalized data are centred at______and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

Ans. (a) 0

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

Ans. (c) Outliers cannot conform to the regression relationship

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

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

Ans. Normal distribution is the most common distribution function for independent, randomly generated variables. It describes how the values of a variable are distributed. It is a continuous probability distribution that is symmetrical around its mean. Most of the values cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. It is also called as Gaussian distribution and the bell curve.

The normal distribution fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

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

- They are 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 to determine the proportion of values that fall within certain distances from the mean.

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

Ans. Imputation techniques depend upon the type of missing data cases. In case of Missing at Random (MAR) and Missing Completely at Random (MCAR), it is safe to remove the data with missing values depending upon their occurrences. While in the case of Missing not at Random (MNAR), removing observations with missing values can produce a bias in the model. So, we have to be really careful before removing observations.

Some of the recommended imputation techniques to handle missing data are-

(1). Complete Case Analysis (CCA): It directly removes the rows that have missing data, i.e., we consider only those rows where we have complete data, i.e. the data is not missing. This method is also popularly known as “Listwise deletion”. It is easy to implement and no data manipulation is required. However, deleted data can be informative and can lead to deletion of a large part of the data and the production model will not know what to do with the missing data. It is good for Mixed, Numerical, and Categorical data.

(2). Arbitrary Value Imputation: This technique can handle both the Numerical and Categorical variables. In this, 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. The missing data is imputed with an arbitrary value that is not part of the dataset or Mean/Median/Mode of data. It is easy to implement and retains the importance of "missing values" if it exists. But it can distort original variable distribution and arbitrary values can create outliers.

(3). Frequent Category Imputation**:** 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. Its implementation is easy and we can obtain a complete dataset in very little time. **However,** the higher the percentage of missing values, the higher will be the distortion. Also, it may lead to over-representation of a particular category and can distort original variable distribution.

Q12. What is A/B testing?

Ans. A/B testing is basically statistical hypothesis testing, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics. It is a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

The A/B testing process can be simplified as follows:

1. You start the A/B testing process by making a claim (hypothesis).
2. You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.
3. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

To put this in more practical terms, a prediction is made that Page Variation #B will perform better than Page Variation #A. Then, data sets from both pages are observed and compared to determine if Page Variation #B is a statistically significant improvement over Page Variation #A.

Q13. Is mean imputation of missing data acceptable practice?

Ans. It is a bad practice in general. If just estimating means: mean imputation preserves the mean of the observed data. It leads to an underestimate of the standard deviation. Also, it distorts relationships between variables by "pulling" estimates of the correlation toward zero.

Q14. What is linear regression in statistics?

Ans. Linear regression is a basic and commonly used type of predictive analysis. 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.

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. 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.

The three major uses for regression analysis are-

(1) determining the strength of predictors,

(2) forecasting an effect, and

(3) trend forecasting.

Q15. What are the various branches of statistics?

Ans. There are two main branches of statistics-

(1). Descriptive statistics and

(2). Inferential statistics.

Both of these are employed in scientific analysis of data.

(1) Descriptive Statistics: Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It describes the important characteristics/ properties of the data using the measures the central tendency like mean, median, mode and the measures of dispersion like range, standard deviation, variance etc. Descriptive statistics provides us the tools to define our data in a more understandable and appropriate way and data can be summarized and represented in an accurate way using charts, tables and graphs. 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. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

(2) Inferential Statistics: Inferential statistics 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. It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc. 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.