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

Answer: 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

Answer: 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

Answer: b) Modelling 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

5. \_\_\_\_\_\_ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer : c) Poisson

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

a) True

b) False

Answer: 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

Answer: b) Hypothesis

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

Answer: 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

Answer: 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?

Answer:

The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let’s say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

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

Answer:

Types of Missing Data

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

1. Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.
2. Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.
3. Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

**Common Methods**

1. Mean or Median Imputation

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

There may not be enough observations with non-missing data to produce a reliable analysis

In predictive analytics, missing data can prevent the predictions for those observations which have missing data

2. Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

12. What is A/B testing?

Answer:

A/B testing:

A/B testing is also called split testing or bucket testing—compares the performance of two versions of content to see which one appeals more to visitors/viewers. It tests a control (A) version against a variant (B) version to measure which one is most successful based on your key metrics. As a digital marketing practitioner does either B2B marketing or B2C marketing.

With the data we collected from the activity of users of our website, we can compare the efficacy of the two designs A and B. Simply comparing mean values wouldn’t be very meaningful, as we would fail to assess the statistical significance of our observations. It is indeed fundamental to determine how likely it is that the observed discrepancy between the two samples originates from chance.

In order to do that, we will use a two-sample hypothesis test. Our null hypothesis H0 is that the two designs A and B have the same efficacy, i.e. that they produce an equivalent click-through rate, or average revenue per user, etc. The statistical significance is then measured by the p-value, i.e. the probability of observing a discrepancy between our samples at least as strong as the one that we actually observed.

13. Is mean imputation of missing data acceptable practice?

14. What is linear regression in statistics?

Answer:

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 = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

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?

Answer:

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

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

**Inferential statistics:** As the name suggests, 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.