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 1. 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

Answer 2. 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 3. 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

Answer 4. 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 5. c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

Answer 6. 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

Answer 7. 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 8. 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 9. Outliers cannot conform to the regression relationship

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 10. 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 graphical form, the normal distribution appears as a "bell curve".

In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.Normal distributions are symmetrical, but not all symmetrical distributions are normal. Many naturally-occurring phenomena tend to approximate the normal distribution. In finance, most pricing distributions are not, however, perfectly normal.

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

Answer 11.

You use the model to fill in the missing value of that variable. This technique is utilized for the MAR and MCAR categories when the features in the dataset are dependent on one another. For example using a linear regression model.

There is no single method to handle missing values. Before applying any methods, it is necessary to understand the type of missing values, then check the datatype and skewness of the missing column, and then decide which method is best for a particular problem.

i will recommend K_Nearest Neighbor Imputation. The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbors.

12. What is A/B testing?

Answer 12. A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

13. Is mean imputation of missing data acceptable practice?

Answer 13. 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.

14. What is linear regression in statistics?

Answer 14. 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?

Answer 15. The two main branches of statistics are descriptive statistics and inferential statistics.
Descriptive statistics is the part of statistics that deals with resenting the data we have. This can take two basic forms –presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on).
Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking ‘What is this data telling us, and what should we do? ’For example, a council might be considering altering the speed limit on a main road, after a number of accidents. They might do this by surveying the speeds of cars (data collection) and then arrive at a conclusion as to whether the speed limit needs to be lowered (if, 6 Introductory statistics for example, a number of cars are driving too fast). Note, though, that this may not be the case; everyone might be driving at a perfectly acceptable speed, and the accidents are down to something other than speed (a blind spot or a pothole, for example). This is inferential statistics: take the data you have and make an ‘inference’ or ‘conclusion’ from it. We shall see much more of this later when we discuss things such as hypothesis testing, where we test to see whether the data supports a belief that we have.