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) 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

   ANS: D) All of the above 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. 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 centered 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)

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

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

ANS: A normal distribution is the proper term for a probability 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.

A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. ... The graphs are commonly used in mathematics, statistics and corporate data analytics.

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

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

ANS: First, determine the pattern of your missing data.
There are three types of missing data:

1. Missing Completely at Random:
2. Missing at Random:
3. Missing Not at Random:
4.

And here are seven things we can do about that missing data:

1. Listwise Deletion: Delete all data from any participant with missing values. If your sample is large enough, then you likely can drop data without substantial loss of statistical power cipants.
2. Recover the Values: You can sometimes contact the participants and ask them to fill out the missing values. For in-person studies, we've found having an additional check for missing values before the participant leaves helps.
3. Imputation: It is replacing missing values with substitute values. The following methods use some form of imputation.
4. Educated Guessing: It sounds arbitrary and isn't your preferred course of action, but you can often infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all "4s", assume that the missing value is a 4.
5. Average Imputation: Use the average value of the responses from the other participants to fill in the missing value.
6. Common-Point Imputation: For a rating scale, using the middle point or most commonly chosen value.
7. Regression Substitution: You can use multiple-regression analysis to estimate a missing value. Multiple Imputation: The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation [pdf], software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions. It is one of a number of examples where computers continue to change the statistical landscape. Most statistical packages like SPSS come with a multiple-imputation feature. More on multiple imputation.

12. What is A/B testing?

1. 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.

2. A/B testing is a shorthand for a simple controlled experiment. in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

3. A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product. Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.

4. Today, A/B tests are being used also for conducting complex experiments on subjects such as network effects when users are offline, how online services affect user actions, and how users influence one another.[6] Many professions use the data from A/B tests. This includes data engineers, marketers, designers, software engineers, and entrepreneurs.[7] Many positions rely on the data from A/B tests, as they allow companies to understand growth, increase revenue, and optimize customer satisfaction.

13. Is mean imputation of missing data acceptable practice?
ANS:
No its not acceptable method. It has following drawbacks:

1. Mean imputation does not preserve the relationships among variables.
2. Mean Imputation Leads to An Underestimate of Standard Errors.
3. Mean Imputation Leads to An Underestimate of Standard Errors toward zero

14. What is linear regression in statistics?
1. Linear regression is a basic and commonly used type of predictive analysis.
2. 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?
3. 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.

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?
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.
1. Descriptive Statistics
    1. 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.

    2. 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.

Inferential Statistics

3. 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.

4. 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.