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 is A

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 is A

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 is B

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 is D

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
Answer is C

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

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned
Answer is B

8. 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 is A

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 is 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?
Answer
1. Normal distribution is also known as Gaussian Distribution. The Normal Distribution is a continuous probability distribution that is described by the probability density function.
2. In graph form, normal distribution will appear as a bell curve. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve
3. The normal distribution is the most common type of distribution assumed in statistical analyses. However Real life data rarely follows a perfect normal distribution.
4. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.
5. The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution.
6. The skewness measures the symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set has a skewness less than zero, or negative skewness, then the left tail of the distribution is longer than the right tail. positive skewness implies that the right tail of the distribution is longer than the left.
7. The kurtosis statistic measures the thickness of the tail ends of a distribution in relation to the tails of the normal distribution.
8. The normal distribution model is motivated by the Central Limit Theorem.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer
1. Missing data is a huge problem for data analysis because it distorts findings. It's difficult to see that some entries are missing values and they must be addressed. There are three types of missing data they are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern, Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data and Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.
2. Ways to eliminate missing data and imputation techniques: Deletion method to eliminate missing data, Regression analysis to systematically eliminate missing data, Data imputation techniques can also be used like Mean imputation, Substitution, Hot deck imputation, Cold deck imputation etc.

12. What is A/B testing?
Answer
1. A/B testing is comparison experiment and used to determine which type variation performs better out of two or more versions provided in a controlled environment. Technically we can say that A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.
2. A/B testing demonstrates the efficacy of potential changes, enabling data-driven decisions and ensuring positive impacts.
3. Benefits of A/B testing: Improved user engagement, Improved content, Reduced bounce rates, Increased conversation rates, Higher conversion values, Ease of analysis, Quick results, Reduced cart abandonment, Increased sales.
4. A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.
5. A/B testing doesn't work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

13. Is mean imputation of missing data acceptable practice?
Answer
Mean Imputation: Mean imputation is a very basic type of imputation technique. It is the only tested function that takes no advantage of the time series characteristics or relationship between the variables. Mean imputation reduces the variance of the imputed variables, it shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.

It does not preserve relationships between variables such as correlations. It is 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. Distorts relationships between variables by "pulling" estimates of the correlation toward zero

14. What is linear regression in statistics?

Answer

Linear regression is a linear approach for modelling relationship between two variables by fitting a linear equation with one dependent and one independent variable. It is the process of finding a line that best fits the data points available on the plot which can be used to predict output values for inputs.

15. What are the various branches of statistics?

Answer

There are two branches of statistics they are:-

i) Descriptive statistics

ii) Inferential statistics

Descriptive statistics:- Descriptive statistics summarizes or describes the characteristics of a data set. Descriptive statistics consists of two basic categories of measures: measures of central tendency and measures of variability (or spread).

Measures of central tendency describe the center of a data set.

Measures of variability or spread describe the dispersion of data within the set.

Inferential statistics:- Inferential statistics is a statistical method that deduces from a small but representative sample the characteristics of a bigger population. In other words, it allows the researcher to make assumptions about a wider group, using a smaller portion of that group as a guideline.