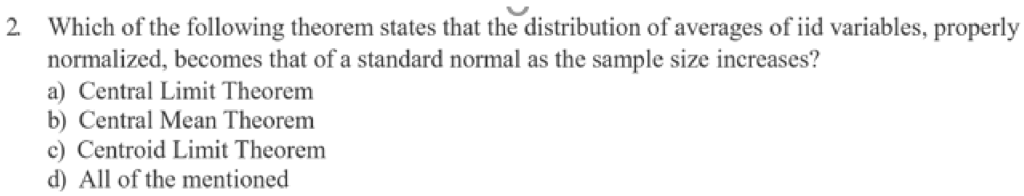
STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False

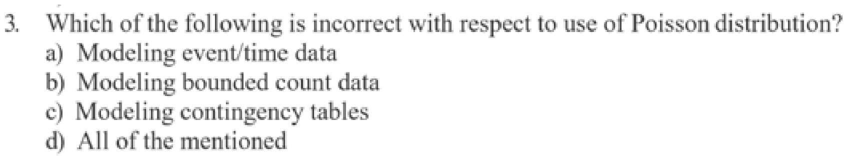
Answer: True

It takes on a 1 if an experiment with probability p resulted in success and a 0 otherwise.



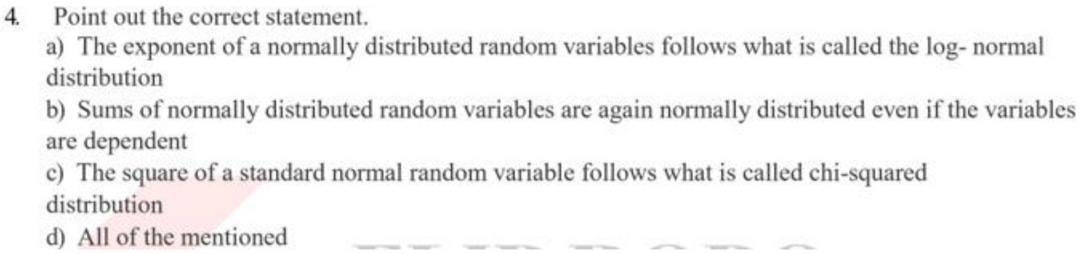
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned

Answer: a) Central Limit Theorem



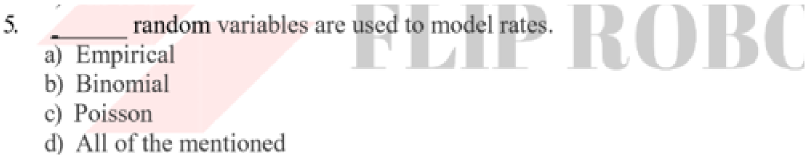
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned

Answer: b) Modelling bounded count data



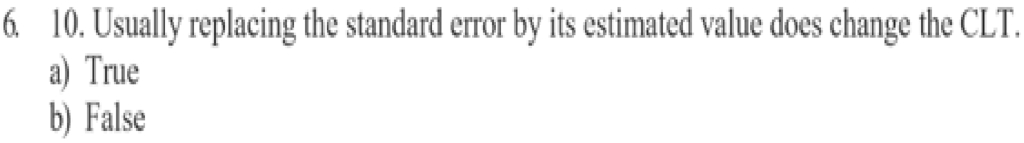
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned

Answer: d) All of the mentioned



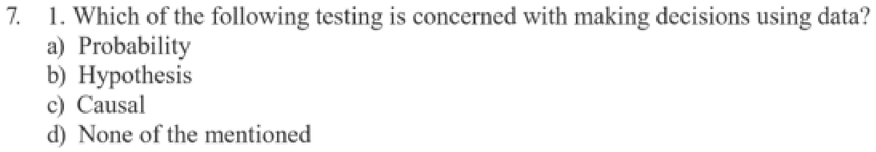
5. ______ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

Answer: c) Poisson



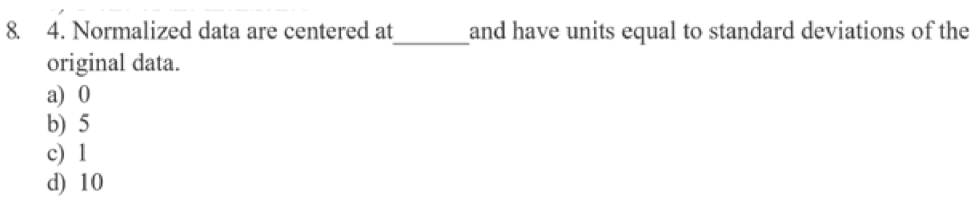
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

Answer: **b)False**



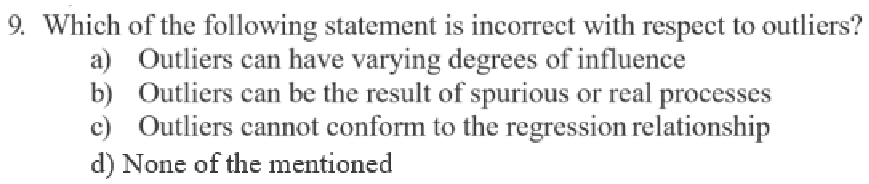
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

Answer: **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 can confirm to the regression relationship

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

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**.

If something is said to follow the normal distribution, it means in the simplest terms that most of the data lie around the average. An easy example is the distribution of test grades in schools. Most people will score around the average, with a few high scores and a few low scores. This means that most people get C's, while only a few get A's and F's.

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

Zero Replacement: Here, you replace the missing value with zero irrespective of everything.

Min or Max Replacement: Replace the missing value with the minimum or maximum value of a feature.

Mean/ Median/ Mode Replacement: Replace missing value with mean or median or most frequent feature value.

Also there are more others ways too to handle missing values ,one can do the comparison and analysis on the basis of other columns in the data and then can fill the missing values.

12. What is A/B testing?

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.For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is **typically considered terrible practice** since it ignores feature correlation.

14. What is linear regression in statistics?

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?

The two main branches of statistics are descriptive statistics and inferential statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of 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 avoiding biases that are so easy to creep into the experiment.

Inferential Statistics

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.