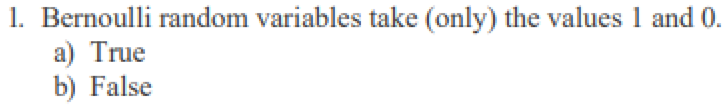
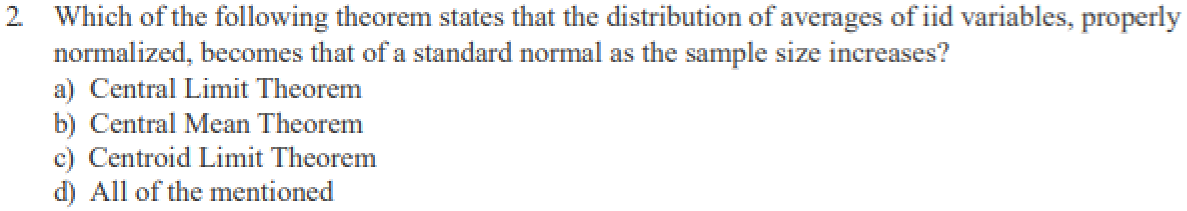
STATISTICS



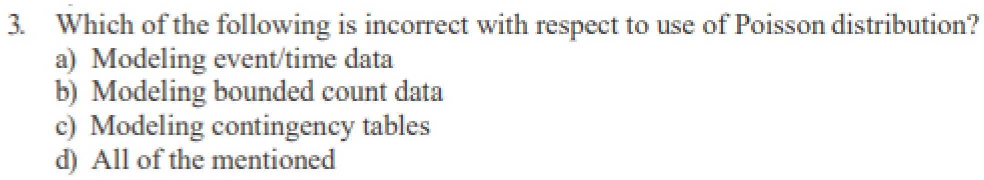
1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False

Ans. a) True



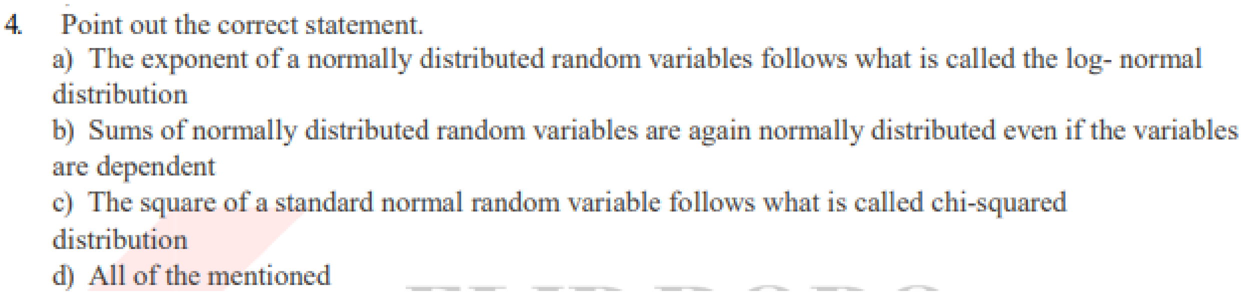
2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned

Ans. a) Central Limit Theorem



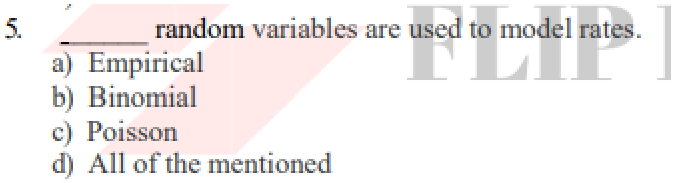
3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned

Ans. b) Modeling bounded count data



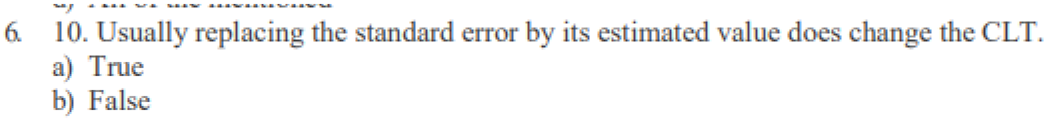
4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal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

Ans. d) All of the mentioned



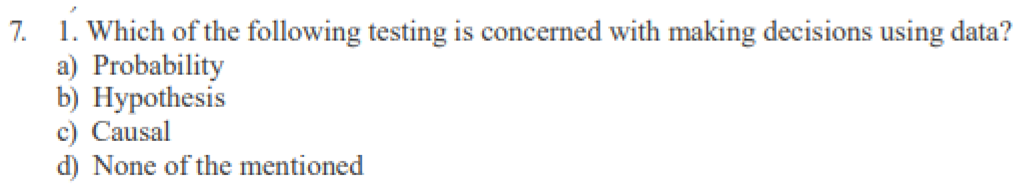
5. ______ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

Ans. c) Poisson



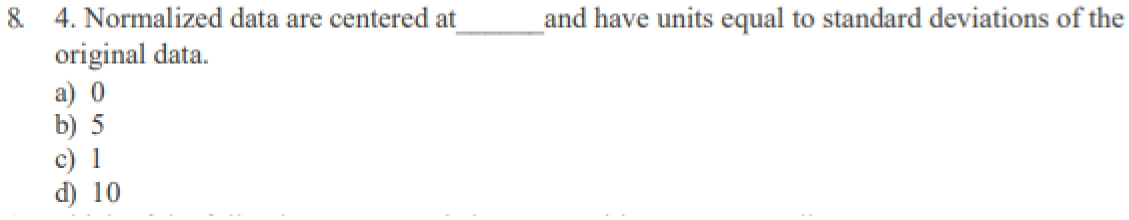
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

Ans. b) False



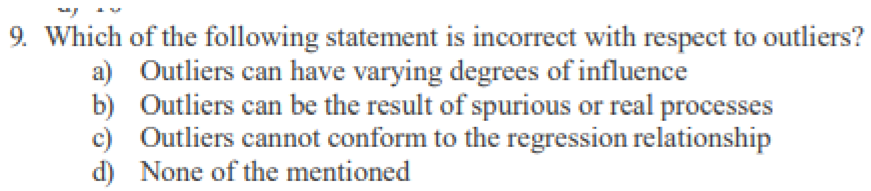
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned

Ans. b) Hypothesis



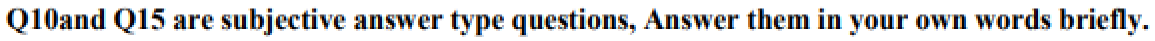
8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10

Ans. a) 0

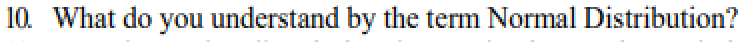


9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

Ans. c) 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?

Ans.

According to my understanding normal distribution means the data set which are having very similar values or range of value is very small and fewer outliers on high and low ends of the data range.

When we have data sets of Normal Distribution and we plot them on graph we could see below image and if we observe below points in that graph then we can say that this is normal distribution.

1. Bell Shaped image plotted
2. Mean, median and mode are nearby or same value.
3. Less spread of data in pattern of -1,0,1

These are some points considering we can decide if data is normal distribution or not.

One real life example which I can give is Hight of male persons. Average Hight of male person is nearby 5 to 6.2 feet always if we collect thousands of samples and plot them on diagram we can see that data are between above mentioned range and bell shape curved got plotted.

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

Ans.

We have multiple methods available to handle missing data. However, we should choose best fit method according to our data set and our requirement. Every time using single method would not be best practice and if we do then we can not able to get actual output from data set.

1. **Dropping the data points;**

   We should use this method when we have good knowledge of related domain and we are aware of that if we drop some rows are column then there will be not impact on desired out put result. Drawback of this method is if we drop important information then result can be not fruitful.

2. **Mean imputation**

   This is most common method where we replace missing data with mean of data set. We should not use this method blindly for large data set because they affect outliers.

3. **Regression/Classification Imputing**

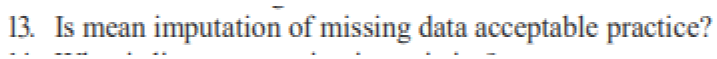
This is also of the best fit method to handle missing data. In this method we train an ML model to predict missing values. Most used algorithm in this method is KNN because it takes distance between two points .

12. What is A/B testing?

Ans.

In Simple word AB testing means way of measuring performance between two variables. This is very popular testing which has been implemented by business to make changes in their business strategies.

If we talk about practical use then suppose there is one business, they just developed one product which is updated version of old product and business wanted to know if this product will be accepted by customers or not. Business will do AB testing to find out that and to start AB testing they should start first with hypothesis testing.



13. Is mean imputation of missing data acceptable practice?

Ans.

You can fill in missing values with the mean of the variable over the time period of observation. *Pros*: Easy to compute and understand. Decent option if you know your variables to be distributed normally. *Cons*: You if your data has a trend (if the rolling-mean is increasing over time) your added values may make your charting look odd. Also, this is not acceptable if your variables have an odd distribution that makes the mean value meaningless.

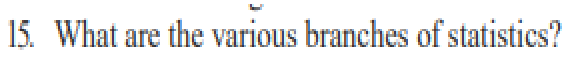
So, my personnel view is it is acceptable but not in every case. One should have good understanding of data and domain part so that he can decide if Mean imputation should be used or not.

14. What is linear regression in statistics?

Ans.

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Also called simple regression or ordinary least squares (OLS), linear regression is the most common form of this technique. Linear regression establishes the linear relationship between two variables based on a line of best fit. Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero. Non-linear regression models also exist, but are far more complex.



15. What are the various branches of statistics?

Basically, there are two main branches of statistic.

Descriptive Statistics-

Descriptive statistics are broken down into **measures of central tendency and measures of variability (spread)**. Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Inferential statistics-

Inferential statistics is concerned with inferences and predictions about populations based on sample data taken from the populations in question.

The main indices of inferential statistics are:

(1) Binomial Theorem

(2) Hypothesis Testing

(3) Normal Distributions

(4) Linear Regression

(5) Central Limit Theorem

These are some of the more important topics that will consume much of the analytical time spent in inferential statistics.