Statistics– WORKSHEET

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
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
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
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-
5. _ random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   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
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
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

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

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?

13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

10 answer: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mea are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

11 answer: A slightly better approach towards handling missing data is Imputation. Imputation means to replace or fill the missing data with some value. There are lo of ways to impute the data. As you can see the above code imputes the building area column values with the mean values of that column.

There are many ways one could handle missing data - and there is single answer to this problem. Some specific ways of handling this with imputation are :
Imputation with mean : Missing data is replaced by mean of the column
Imputation with median : Missing data is replaced by mean of the column
Imputation with Mode: Missing data is replaced with mode of the column
Imputation with linear regression : With real valued data, this is a common technique. The missing value is replaced by performing linear regression based on the oth feature values.

12 answer: 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.
In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Nov on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

13 answer: Bad practice in general
If just estimating means: mean imputation preserves the mean of the observed data
Leads to an underestimate of the standard deviation
Distorts relationships between variables by "pulling" estimates of the correlation toward zero

14 answer: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be explanatory variable, and the other is considered to be a dependent variable.

15 answer: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