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
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.

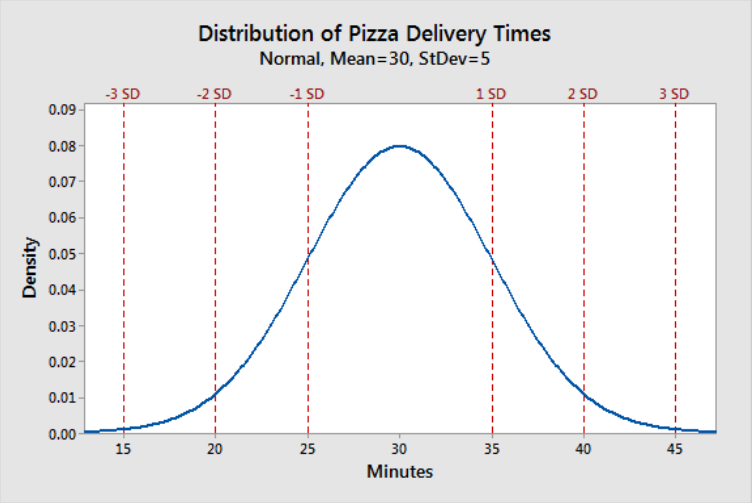
10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

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

Normal distribution, also called Gaussian distribution, the most common distribution function for independent, randomly generated variables. Its familiar bell-shaped curve is ubiquitous in statistical reports, from survey analysis and quality control to resource allocation.

The normal distribution is a probability distribution that describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques.

The graph of the normal distribution is characterized by two parameters: the mean, or average, which is the maximum of the graph and about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from the mean. A small standard deviation produces a steep graph, whereas a large standard deviation (again compared with the mean) produces a flat graph.



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

There are several ways to handle missing values in the given data:-

- ✓ Dropping the values
- ✓ Deleting the observation (not always recommended).
- ✓ Replacing value with the mean, median and mode of the observation.
- ✓ Predicting value with regression
- ✓ Finding appropriate value with clustering

A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations.

12. What is A/B testing?

A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

An AB test is an example of **statistical hypothesis testing**, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

To put this in more practical terms, a prediction is made that Page Variation B will perform better than Page Variation A. Then, data sets from both pages are observed and compared to determine if Page Variation B is a statistically significant improvement over Page Variation A.

This process is an example of statistical hypothesis testing.

13. Is mean imputation of missing data acceptable practice?

No, Mean imputation of missing data not acceptable practice

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations

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. In simple words, linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line
The Linear Regression Equation:-
The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

15. What are the various branches of statistics?

Statistics is the study and manipulation of data, including ways to gather, review, analyze, and draw conclusions from data. Statistics can be used to make better-informed business and investing decisions

The two major areas of statistics are descriptive and inferential statistics.

.