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 1 : 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

Answer 2 : (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 3 : (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

Answer 4 : (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 5 : (c) Poisson

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

Answer 6 : (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 7 : (b) Hypothesis

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 8 : (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

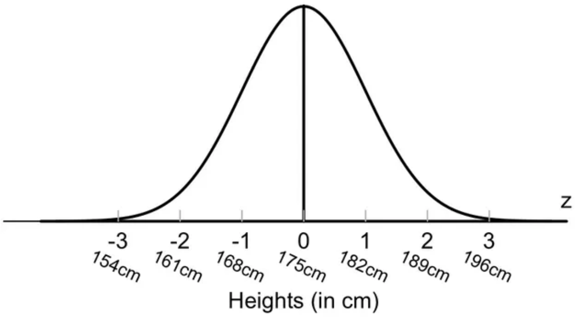
Answer 9 : (c) Outliers cannot conform to the regression relationship

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer 10 : If a distribution has standard deviation between -3 to +3 then it is called Normal Distribution. The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the "bell curve."

An example of Normal Distribution is a distribution of heights of a population.



most people conform to the mean height whereas taller and shorter people exist, but with decreasing frequency in the population. All (almost) the data falls between -3 to +3 standard deviation.

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

Answer 11 : Missing data can be ignored or deleted but may or may not be a good idea. So it is preferable to use Imputation.

In Imputation missing values are replaced with estimated or mean value. Most of the time **Mean imputation** is used. In this mean value is used to replace missing value. It has some drawbacks hence Some other type of Imputation are also used like:

Hot deck imputation : A value picked at random from a sample member who has comparable values on other variables.

Cold deck imputation : A value picked deliberately from an individual with similar values on other variables

Regression imputation : We use regression technique to find the predicted value and replace it with missing value.

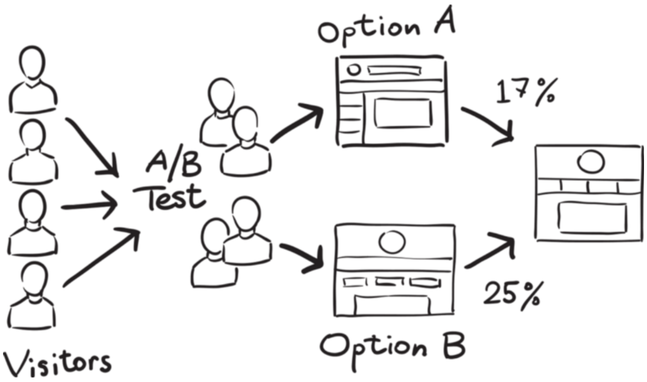
I recommend all the above imputation methods specially Mean Imputation and Regression Imputation and can be used to best fit the prediction on case to case basis.

12. What is A/B testing?

Answer 12 : 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.

Suppose we want to increase sales of our product. A/B testing is one of the most prominent and widely used statistical tools to decide so.

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



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying our product, while the sample refers to the number of customers that participated in the test.

In A/B Testing one set of Hypothesis are made and based on the experiment data any one from Null Hypothesis or Alternative Hypothesis is selected.

13. Is mean imputation of missing data acceptable practice?

Answer 13 : The process of replacing null values in a data collection with the data’s mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Hence Mean Imputation can’t be accepted in all cases. Based on scenario different imputation may be used

14. What is linear regression in statistics?

Answer 14 : 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).

Basic equation of Linear Regression is as under :

$$\mathbf{Y=a+bx+e}$$

Here y is the target variable for which the value need to be predicted.

a is intercept which specifies from which point the regression line starts.

b is coefficient which specifies the amount what gets increased in the value for every unit increased in x or input variable.

x is input variable for which value of y is to be predicted.

e is the error which comes in the prediction.

Error is the calculated value which may be either positive or negative and the same value is added or deduction from the result of the equation to get best prediction.

The common method used to get the value of the error is **Least Squares Error (Method).**

Following is the formula of Least Squares Error :

$$S = \sum_{i=1}^{n} {d_i}^2$$

$$S = \sum_{i=1}^{n} [y_i - f_{x_i}]^2$$

$$S = {d_1}^2 + {d_2}^2 + {d_3}^2 + \cdots + {d_n}^2$$

It is basically the sum of squares of all the differences of actual and predicted values of a particular data set.

15. What are the various branches of statistics?

Answer 15 : Statistics is a study of presentation, analysis, collection, interpretation and organization of data. There are two main branches of statistics:

1. Inferential Statistic.
2. Descriptive Statistic.

Inferential Statistics: Inferential statistics is used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics: Descriptive statistics are used to get a brief summary of data. The summary of data can be represented in numerical or graphical form.

Both the statistics are very useful in making many business decisions.

Submitted by : Amit Puri Goswami
Batch No. : 1843
Project: Internship 30