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

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

Subjective questions are answered in the next page.

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

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

Answer-

In simpler terms normal distribution represents the behavior of most of the situations in the universe. It is a bell shaped curve, where the value of mean, median, mode is equal at the centre. It is a symmetric. It indicates bulk of data with average value. Exactly half of the value are to the left of centre and half to the right of centre. The total area under the bell shaped curve is 1. In the curve 68% data lies in 1st standard deviation. 95% lies till 2nd std. deviation. 99.7% data lies till 3rd std. deviation. Remaining (.3%) lies outside the curve and are called outliers. This is called 68,95,99.7 emperical rule.

In statistical term the formula is.

$$Z= \frac{x-\mu}{\sigma}$$

where x= value

μ= mean

σ= Standard deviation.

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

Answer-

While working with datasets and models we get various datasets where value goes missing and to handle such missing data we have different techniques.

- Deleting the record - This is the easiest method to handle missing data, in this method we delete the whole set of data of that particular row and move ahead to train the machine with remaining data. But the limitation here is it can be enforced on a dataset where there is huge record else it will effect in the prediction of model.
- Creating a separate model - in this model, what we do is we take missing data from data sets as testing dataset and the dataset where data is not missing as training dataset and based on training whatever output the model generates gets filled into the missing dataset. Its drawback is that it consumes much time and is efficient on dataset with smaller records.
- Statistical methods - Here based on the requirement of dataset we compute mean or median or mode with the remaining value present in dataset and fill the outcome in missing dataset. The problem with this model is it doesn't take into the correlation between data.
- Linear Regression – in this technique dataset with complete data are used to generate the regression equation whereas the dataset of missing values are dependent variables and best predictor value are used as independent variable. The equation is then used to predict the missing value in an iterative way, values of missing data are inserted and with the help of dependent variable values are predicted. Steps are repeated until there is little difference between the predicted value from one step to next. This process is someway near to good predictability of model.

12. What is A/B testing?

Answer-

A-B testing is a testing tool for predicting outcomes through known information which we don't know. It is a basic random experiment, where we compare two variables or product to find out which one is better in a controlled environment. For example, a packing company/logistic company pack two products. One A, which is packed without any change, while product B is packed with some innovative significant change. Now the customers who used both packed products gives the review about the packing, which one was better. And based on the review which company got, it makes decision about further packing.

13. Is mean imputation of missing data acceptable practice?

Answer-

To some model mean imputation work, but model where data is co-related with other datasets this techniques fails badly. In this model, we calculate the mean of the data present and fill it into the missing data. Now let's take an example of a model where age, gender and efficiency data set is present and we know a man is more efficient than a women of same age naturally or a middle aged man is more efficient than older ones. If a data is missing in efficiency dataset, mean imputation just calculates the mean from the remaining values and puts the value of outcome in place of missing value, since this value is just average value the outcome of efficiency of old women can be greater than that of a middle aged man. So mean imputation dosen't take into account that efficiency is correlated to age and gender. It also reduces variance of the data set.

14. What is linear regression in statistics?

Answer-

A linear regression is a approximation of a relationship between two or more variables. In this regression with the help of sample data, we design a model that works on that particular sample and then makes prediction for whole group of data. For that we have a dependent variable(y) and a independent variable(x), also we have a slope(m) and intercept(b). This algorithm is used for prediction and forecast by plotting the best fit line, which touches or passes near to most of data in a graph. we also have a "least square error" method which help in plotting best fit line. In this method if a data is away from most of data we find the difference between the original data and predicted data and then it gets plotted on the line.

Linear Regression is of two types.

1. Simple = $y=b+mx$.
2. Multiple= $y=\alpha 0+\alpha 1x1+\alpha 2x2+\alpha 3x2+\ldots\ldots\alpha nxn$.

15. What are the various branches of statistics?

Answer-

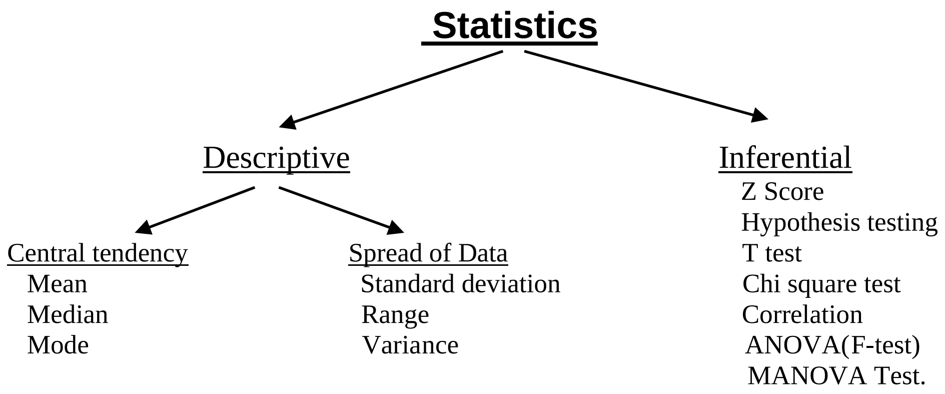
Statistics is the science of collecting, organizing, analyzing, presenting and implementing data to assist in making more effective decisions. Statistics is categorized in two branches.

1. Descriptive statistics
2. Inferential statistics

- Descriptive statistics is the representation of data in tabular, numerical and graphical form.it is used in non-experimental research. The data which we see in newspapers or magazines are examples of descriptive statistics.
- Inferential statistics studies a sample of data and tries to interpret the meaning of data collected as sample in descriptive data.it uses various techniques and testing methods for the estimation of data or come to a meaningful conclusion like probability, hypothesis testing etc.

Further Descriptive is sub divided in two parts central tendency and spread of data.

Central tendency consists of mean, median and mode. While spread of data consists of Standard deviation, Range and variance.


Statistics
Descriptive
Inferential
Central tendency
Mean
Median
Mode
Spread of Data
Standard deviation
Range
Variance
Z Score
Hypothesis testing
T test
Chi square test
Correlation
ANOVA(F-test)
MANOVA Test.