# STATISTICS WORKSHEET 1

Name - AMIT MATREJA
Internship 21
Datatrained batch - 1834

**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

Ans 1. 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 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

Ans 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

Ans 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

Ans 5. c) Poisson

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

Ans 6. a) True

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 7. 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 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

Ans 9. c) Outliers cannot confirm to the regression relationship.

Ans 10. The term Normal Distribution mean that the data is evenly distributed. In this Distribution all three are equal i.e. Mean = Median = Mode and they are aligned to the center of the curve which makes the evenly balanced or evenly distributed to the left and right.

# STATISTICS WORKSHEET 1

Normal distribution is most commonly and widely used of all the distributions. The normal distribution is also identifiable by its bell shaped design, sometimes referred to as a bell curve.

Ans 11.Missing data is most common problem which exist in almost all the large datasets.There are many different imputation techniques to handle the missing data which also depends on what kind of Data we have.The easiest and could be riskiest way is deletion of missing data, If missing data is very small in size or something which is avoidable or not useful in our observation we can simply delete that missing that particular data and carry on with our analysis.

If it is unavoidable or among the relevant part of data then I would recommend the simplest or common method of Mean or Median Imputation. Through this technique we can use the mean or median of the non missing data.this technique is very useful where we have important but very small amount of missing data.

Ans 12. A/B testing is the randomized testing. It is a way to compare two different type of variable to check which performs better in the given situation.A/B testing is most prominent and commonly used statistical tool. This technique is majorly used to analyze large datasets by picking out sample from it.

Ans 13. Mean imputation of missing data is not acceptable as mean imputation ignores feature correlation and decreases the variance of data. As reduced variance the model is also going to be less accurate.

Ans 14. Linear regression is most important and widely used statistical technique.using Linear regression we try to model relationship between two variables, one independent variable and one dependent variable.it is a straight line that minimizes discrepancies.linear regression models are relatively simple and also provide an easy mathematical formula that can generate predictions easily.

Ans 15. There are two branches of Statistics:-

1. Descriptive Statistics- if we are able to describe any data without using any Statistical tools then it is called descriptive statistics.it helps in the analysis of large datasets by shrinking them to a manageable summary. So that decision can be easily made.

# STATISTICS WORKSHEET 1

2. Inferential Statistics- If the data is too big to explain to reach to any conclusion, we drawn out some sample data from different places and by analyzing that sample data we try to make conclusion or result out of it.eg. Exit polls