STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
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

3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modelling bounded count data

4. Point out the correct statement.
d) All of the mentioned

5. _____ random variables are used to model rates.
c) Poisson

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

7. 1. Which of the following testing is concerned with making decisions using data?
b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0

9. Which of the following statement is incorrect with respect to outliers?
c) Outliers cannot conform to the regression relationship

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

10. What do you understand by the term Normal Distribution?
Ans- In normal distribution the data is scattered in such a way that it forms a bel shaped curve. The data is most concentrated in the middle and goes on reducing at the edges of the curve.
--There is no skew and data is uniformly distributed on each side of the median
-- Normal distribution is also called as Gaussian distributions or bell curves because of their shape.
--The normal distribution curves have certain properties which are as follow:

- The mean, median and mode are exactly the same.
- Distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- Distribution can be described by two values: the mean and the standard deviation.
- Mean is the location parameter while the standard deviation is the scale parameter.
- Mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans- Missing data can be either removed or replaced/substituted.
-- Removing data can be used to reduce the bias, when the missing data is occurring at random
--Replacing missing values with substituted data is called as imputation. It is useful when the percentage of the missing data is low. Following are few techniques where data can be replaced:

- Educated Guessing – It is an arbitrary method. For eg: if most vales are 4, then we can substitute the missing values with 4.
- Common-point imputation- use the most commonly chosen value or common point.
- Average Imputation – using the average value of the responses.

- Regression substitution- use multiple-regression analysis to estimate a missing value
- Multiple imputation- It is the most effective method.

12. What is A/B testing?

- A/B tests, also known as split tests which allows you to compare 2 versions of data, say data A and data B to learn which is more effective or efficient.
- -The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes. Eg: If you put 2 seeds in 2 cups of dirt and put one on the fridge with not much sunlight and the other by the window with sunlight, you'll see different results. This kind of experimental setup is A/B testing.

13. Is mean imputation of missing data acceptable practice?

Ans- Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

--It can be used as one of the techniques, but it is not a good solution and it has its drawbacks and the results will not be accurate, as it does not preserve the relationships among variables and leads to low standard errors.

14. What is linear regression in statistics?

Ans- Liner regression is the type of predictive analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line. There are two types of linear regression - Simple and Multiple.

15. What are the various branches of statistics?

Ans- There are 2 branches of statistics.

- Descriptive
- Inferential

--Descriptive statistics- If data can be described without any statistical tools, then it is called descriptive statistics. ex, marks in class, height of student.

--Inferential statistics- If data is too big then then inferential statistics is used. Few samples from different data are taken and average is calculated. The average is then applicable to all the data from where we have selected our samples.