Name- Akanksha B Amarnani
Batch- 1841

STATISTICS WORKSHEET 1

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False
Ans- 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- 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- 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- 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- c) Poisson

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

b) False
Ans- b) False

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned
Ans- 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
Ans- 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- c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
Ans- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In normal distribution mean is zero and standard deviation is one. In graph, normal distribution is illustrated as a bell shaped curve.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans- Missing data can be handled in 2 ways-

a) Deleting the Missing values- It is one of the quick and dirty techniques one can use to deal with missing values; however not recommended as it may cause loss of vital data in some scenarios. If the missing value is of the type Missing Not At Random (MNAR) (Missing values depend on the unobserved data), then it should not be deleted.If the missing value is of type Missing At Random (MAR) or Missing Completely At Random (MCAR) then it can be deleted. There are 2 ways one can delete the missing values- Deleting the entire row (If a row has many missing values then you can choose to drop the entire row) or Deleting the entire column (If a certain column has many missing values then you can choose to drop the entire column.)

b) Imputing the Missing Value- There are different ways of replacing the missing values-

- Replacing with Arbitrary Value- If you can make an educated guess about the missing value then you can replace it with some arbitrary value using the following code.
- Replacing with Mean- This is the most common method of imputing missing values of numeric columns. If there are outliers then the mean will not be appropriate. In such cases, outliers need to be treated first.
- Replacing with Mode- Mode is the most frequently occurring value. It is used in the case of categorical features.
- Replacing with Median- Median is the middlemost value. It's better to use the median value for imputation in the case of outliers.
- Replacing with Previous Value (Forward Fill) - In some cases, imputing the values with the previous value instead of mean, mode or median is more appropriate. This is called forward fill. It is mostly used in time series data.
- Replacing with Next Value (Backward Fill) - In backward fill, the missing value is imputed using the next value.
- Interpolation- Missing values can also be imputed using interpolation. Pandas interpolate method can be used to replace the missing values with different interpolation methods like 'polynomial', 'linear', 'quadratic'. Default method is 'linear'.

12. What is A/B testing?

Ans- A/B testing is an analytical method for making decisions that estimates population parameters based on sample statistics. It 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. For instance, let's say you own a company and want to increase the sales of your product. Here, you may 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 your product, while the sample refers to the number of customers that participated in the test.

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. Mean imputation despite of preserving the mean of the observed data, being easy and restoring the original size of the dataset, is not advisable due to the following reasons-

- Mean imputation does not preserve the relationships among variables. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution. This causes it's ill acceptance as an imputation technique

- Mean Imputation Leads to an Underestimate of Standard Errors

14. What is linear regression in statistics?

Ans- Linear regression models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent and dependent variables, respectively. When there is one independent variable (IV), the procedure is known as simple linear regression. When there are more IVs, statisticians refer to it as multiple regression.

15. What are the various branches of statistics?

Ans- The branches of statistics are as follows-

a) Descriptive Statistics- Descriptive statistics focuses on collecting, summarizing and presenting a set of data. The presentation of data can be in two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages, dispersion and so on).
b) Inferential Statistics- Inferential statistics analyses sample data to draw conclusions about a population. It involves sampling data and infers the result to describe the entire population.