STATISTICS WORKSHEET

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: a

2. Which of the following theorem states that the distribution of averages of iid variables,properly normalized, becomes that of standard normal as the sample size increases?

a) Central Mean Theorem

b)Centroid Limit Theorem

c)All of the mentioned

d)All of the mentioned

Ans: 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

Ans: a

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: b

5. ______________random variables are used to model rates.

a)Empirical

b)Binomial

c)Poisson

d)All of the mentioned

Ans: C

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

a)True

b)False

Ans: 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: 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: 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

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: Normal distribution is also called **Gaussian distribution** and the distribution curve is bell-shaped curve and the curve falls within data of standard deviation and mean and the curve with 95%of the data falls within the standard deviation and mean with 5% and in normal distribution mean = median = mode = 0.

In probability theory, a normal distribution is a type of continuous probability distribution for a real-valued random variable. The random variables following the normal distribution are those whose values can find any unknown value in a given range. the normal distribution doesn’t even bother about the range. The range can also extend to –∞ to + ∞ and still we can find a smooth curve.

Generally, the normal distribution has any positive standard deviation. We know that the mean helps to determine the line of symmetry of a graph, whereas the standard deviation helps to know how far the data are spread out.

Applications:

a)Marks scored on the test

b)Heights of different persons

c)Size of objects produced by the machine

d)Blood pressure and so on.

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

Ans: Missing data: The concept of missing data is implied in the name: it’s data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results. Missing data occurs when no data value is stored for the variable in an observation.Missing can be handled either with removal of the data i.e, deletion or dropping the data but this method is not always possible so **imputation method** is used to fill the missed data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

Before deciding which approach to employ to handle the missing data, we must understand why the data is missing:

a) Missing at Random (MAR): It means the data is missing relative to the observed data and it can be predicted based on the complete observed data.

b) Missing Completely at Random (MCAR) : In the MCAR situation, the data is missing across all observations regardless of the expected value or other variables.

c) Missing Not at Random (MNAR): The MNAR category applies when the missing data has a structure to it , means there appear to be reasons the data is missing.

The techniques mostly recomended are Multiple imputation , K-Nearest neighbors also few times median and mode can be used.

12. What is A/B testing?

Ans: A/B testing is also known as **bucket testing** or **Split testing** or also known as **2 Sample hypothesis testing** includes 2 variants A and B used in the statistical hypothesis testing. "Two-sample hypothesis tests" are appropriate for comparing the two samples where the samples are divided by the two control cases in the experiment.

A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

A/B testing allows individuals, teams and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses and to learn why certain elements of their experiences impact user behavior. In another way, they can be proven wrong—their opinion about the best experience for a given goal can be proven wrong through an A/B test.

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation (or mean substitution) replaces missing values of a certain variable by the mean of non-missing cases of that variable. Mean imputation is so simple but still risky in practice,it must be the last resort to be used because there are many other imputation techniques to give good accuracy and the mean imputation may not help us to find the relationships with stronger parameters and it does not preserve the relationships among the variables.

Pros and Cons of Mean Imputaton:

Pros: 1) Do not reduce your sample size
2) simple to understand and to apply
3) sample mean of your variable is not biased

Cons: 1) Bias in multivariate estimates such as correlation or regression coefficients
2) Standard errors and variance of imputed variables are biased

14. What is linear regression in statistics?

Ans: In statististics , linear regression is a linear approach for modelling the relationship scaled data and one or more dependent and independent variables. If it is with 1 feature then comes under "Simple linear regression" and if more than one then t\it comes under "Multiple linear regression". The dependent and independent variables show the linear relationship between slope and intercept.

Linear regression is one of the ways to perform predictive analysis. It is used to examine regression estimates.

a) To predict the outcome from the set of predictor variables

b) Which predictor variables have maximum influence on the

outcome variable?

The regression estimates explain the relationship between one dependent variable and one or more independent variables. The same is represented in the below equation.

The formula for linear Regression:

y = mx + c, Where **m** is the **Slope and c** is the **intercept.**

15. What are the various branches of statistics?

Ans: There are 2 branches of statistics they are: a) Descriptive statistics and
b) Inferential statististics.

Descriptive Statistics:

In this type of statistics, the data is summarised through the given observations. The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Descriptive statistics are also categorised into four different categories:

a)Measure of frequency
b)Measure of dispersion
c)Measure of central tendency
d)Measure of position

Inferential Statistics:

This type of statistics is used to interpret the meaning of Descriptive statistics. It is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population.

Example:

In a class, the collection of marks obtained by 50 students is the description of data.

Now when we take out the mean of the data, the result is the average of marks of 50 students.

If the average mark obtained by 50 students is 88 out of 100, then we can reach to a conclusion or give a judgment on the basis of the result.