

## 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

ANS. 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. 10. Usually replacing the standard error by its estimated value does change the CLT.

- a) True
- b) False

Ans. B) False

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. 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. 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

**Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.**

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

Ans. the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. This is also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. Its just bell-shaped curve in statistical reports.

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

Ans. There are some of techniques to treat missing value.

>Ignore the records with missing values : Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

>Substitute a value such as mean : When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this method could cause bias distribution and variance. That's where the following imputation methods come in.

> Predict missing values : Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

> Predict missing values - Multiple Imputation : Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

12. What is A/B testing?

Ans. A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product. Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.

13. Is mean imputation of missing data acceptable practice?

Ans. Bad practice in general, If just estimating means: mean imputation preserves the mean of the observed data, Leads to an underestimate of the standard deviation Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14. What is linear regression in statistics?

Ans. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable.

The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

15. What are the various branches of statistics?

Ans. Statistics have majorly categorised into two types:

A. Descriptive Statistics

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

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.

B. Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.