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
Answer : **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
Answer : **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
Answer : **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
Answer : **D) All of the mentioned**

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
Answer : **C) Poisson**

6. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False
Answer: **B) False**

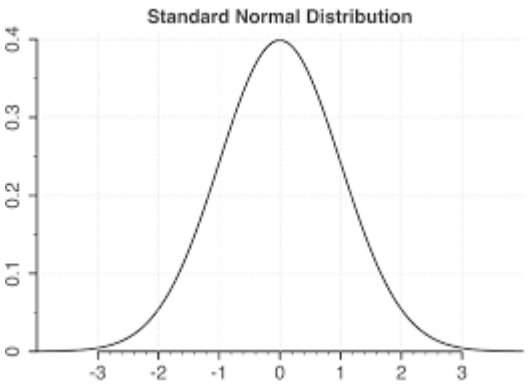
7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned
Answer : **B) Hypothesis**

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
Answer : **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
Answer: **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?
Answer: The normal distribution, also known as the Gaussian distribution, is a symmetric probability distribution centred on the mean, indicating that data around the mean occur more frequently than data far from it. The normal distribution will show as a bell curve on a graph.



11. How do you handle missing data? What imputation techniques do you recommend?
Answer: When dealing with missing data, we have two options for resolving the problem:

1. Imputation: For missing data, the imputation method generates credible predictions. When the percentage of missing data is low, it's the most beneficial. If the percentage of missing data is too high, the results will be lacking of natural variation, which will make it difficult to build an effective model.
2. Data removal: The other option is to remove data. To eliminate bias when dealing with data that is missing at random, relevant data can be erased. If there aren't enough observations to make a reliable analysis, removing data may not be the best solution. Observation of specific events or factors may be essential in some circumstances.

Most recommended method of Imputation technique is Imputation Using Datawig (Deep Learning):

- When compared to other approaches, this one is quite accurate.
- It has various routines for dealing with categorical data (Feature Encoder).
- It works with both CPUs and GPUs.

12. What is A/B testing?

Answer: A/B testing is a type of randomised control trial. It's a method of comparing two versions of a variable in a controlled environment to see which performs better. A/B testing also known as bucket testing or split-run testing, it contains statistical hypothesis testing, also known as "two-sample hypothesis testing."

13. Is mean imputation of missing data acceptable practice?

Answer: Mean Imputation has two major flaws.

1. The relations between variables are not preserved by mean imputation.
2. Mean Imputation Leads Standard Errors to be Undervalued.

14. What is linear regression in statistics?

Answer : The most fundamental and widely used type of predictive analysis is linear regression. The goal of regression is to look at two things:

(1) Is it possible to forecast an outcome (dependent) variable using a set of predictor variables?

(2) Which variables in particular are significant predictors of the outcome variable, and how do they affect the outcome variable (as indicated by the size and sign of the beta estimates)?

These regression estimations are used to illustrate how one dependent variable interacts with one or more independent variables. The simplest expression of the regression equation with one dependent and one independent variable is **y = c + b*x,** where y represents the estimated dependent variable score, c represents the constant, b represents the regression coefficient, and x represents the independent variable score.

Types of Linear Regression

- Simple linear regression: 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
- Multiple linear regression: 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)
- Logistic regression: 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
- Ordinal regression: 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)
- Multinomial regression: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)
- Discriminant analysis: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

15. What are the various branches of statistics?

Answer: Descriptive statistics and inferential statistics are the two types of statistics.

Descriptive Statistics:

Descriptive statistics is the initial step in statistical analysis, and it deals with data gathering and presentation. In scientific terms, descriptive statistics are concise explanatory coefficients used by statisticians to summarise a set of data. A data set might represent a sample of a population or the complete population in general. There are two types of descriptive statistics.

Central tendency measures :

- Mean

- Mode
- Median

Variability measures: quartiles, range, variance and standard deviation

Inferential Statics:

Inferential statistics are methods that allow analysts to draw conclusions, assessments, or predictions about a population based on the information acquired from a sample. Inferential statistics often talks in probability terms by using descriptive statistics. Statistical techniques are primarily used by statisticians to analyse data, make estimates, and draw conclusions from limited information obtained through sampling, as well as to test the accuracy of the estimates.

The different types of calculation of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis