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: 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: b) 5

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

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 Distributions is a probability distribution that is symmetric about mean, showing that data mean is more frequent in occurrence that data far from mean.
- Normal distribution is most widely known and used of all distribution.
- It is easy to work with mathematically, method developed using normal theory work quite well even when the distribution is not normal.
- In graphical form, the normal distribution appears as a “bell curve”.
- In normal distribution mean is zero and standard deviation is one. It has zero skew and a kurtosis is 3.
- Normal distribution are symmetrical, but not all symmetrical distributions are normal. Normal distributions occur when where a dividing line produces two mirror images.
- Normal distribution is key to central limit theorem.
- Formula for normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

σ = Standard Deviation
μ = Mean
***x* = Value of variable**
***f(x)* = Probability function**

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

ANS: We often encounter missing values while we are trying to analyze and understand our data. There will be missing values because the data might be corrupted or some collection error. Missing values can cause bias and can affect the efficiency of how the model performs. Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step.
The following are some of the most prevalent methods:

- Mean imputation
  Calculate the mean of the observed values for that variable for all non-missing people.It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks.
- Substitution
  Assume the value from a new person who was not included in the sample.To put it another way, pick a new subject and employ their worth instead.
- Hot deck imputation
  A value picked at random from a sample member who has comparable values on other variables.
  One benefit is that you are limited to just feasible values.
  Another factor is the random element, which introduces some variation.
- Cold deck imputation
  A value picked deliberately from an individual with similar values on other variables.In most aspects, this is comparable to Hot Deck, but without the random variance.
- Regression imputation
  The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.
- Interpolation and extrapolation
  An estimate based on other observations made by the same person. It generally only works with data that is collected over time.Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

12. What is A/B testing?

ANS: A/B testing is a user experience research methodology that allows testers to experiment by comparing two versions of a single variable. Also known as “split-run” testing, this method provides measurable feedback on the performance of the digital products we create. Ambiguity, hunches, and guesswork can be discarded to allow data-driven decisions.

Experiments are run to achieve a single desired outcome using two distinct methods or variations, which attempt to achieve the outcome. Users interact with both the A and B versions of the product, and measurements are made to assess which version comes closest to delivering the best outcome.

Once a clear winner has emerged from these first two choices, it is common to retest more times with a new A or B version introduced to assess if the previous winner is still the best choice.

A/B testing is regularly used across all digital platforms to optimize conversion rates, dwell-time, and engagement with apps and websites.

Here are a few examples where you may wish to A/B test and how A/B testing has been successfully used to improve performance:

- A/B testing for email campaigns
- A/B testing for ad copy
- A/B testing your opt-in forms
- A/B testing for button and user interface design
- A/B testing for images

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.

This technique isn't a good idea.

Drawbacks of mean imputation**:**

1. Mean imputation reduces the variance of the imputed variables.
2. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
3. Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

ANS: Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?
These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.
The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

- Types of linear regression:
  Simple linear regression
  Multiple linear regression
  Logistic regression
  Ordinal Regression

Multinomial regression
Discriminant Analysis

15. What are the various branches of statistics?

ANS: Statistics is the main branch of mathematics. Used to perform different operations, i.e., Data collection, organization, analysis, and so on.
There are two branches of statistics:

- **Descriptive:**
  Descriptive statistics is the first part of statistics that deals with the collection of data. Descriptive statistics are used to do various kinds of analysis on different studies. Descriptive statistics has two parts:
  - Central tendency measures
  - Variability measures

Measures of central tendency:
Central tendency measures specifically help statisticians evaluate the distribution centre of values. These tendency measures are:
Mean
Mean is a conventional method used to describe the central tendency. calculate the average of values, count all values, and then divide them with the number of available values.
Median
It is the result that is in the middle of a set of values.
Mode
The mode is the frequently occurring value in the given data set.
Measures of variability:
The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

- **Inferential Statistics:**
  Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.
  these techniques are used for data analysis, drafting, and making conclusions from limited information.
  Inferential Statistics includes:

1. Regression analysis
2. Analysis of variance (ANOVA)
3. Analysis of covariance (ANCOVA)
4. Statistical significance (t-test)
5. Correlation analysis

THE END