Statistics–WORKSHEET-1

Assignment

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

2. Which of the following theorem states that the distribution of averages of id 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

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modelling event/time data b) Modelling bounded count data

c) Modelling contingency tables d) All of the mentioned

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

5. random variables are used to model rates.

a) Empirical b) Binomial

c) Poisson d) All of the mentioned

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

a) True 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

8. Normalized data are centred at and have units equal to standard deviations of the original data.

a) 0 b) 5

c) 1 d) 10

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

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 is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

Common Properties of the Normal Distribution

- They're all symmetric. The normal distribution cannot model skewed distributions.
- The mean, median, and mode are all equal.
- Half of the population is less than the mean and half is greater than the mean.
- The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean.

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

Ans: The most common way of dealing with missing data is to remove all rows with missing data if there are not too many rows with missing data.

If more than 50-60% of rows of a specific column are missing data, it is common to remove the column. The main problem with removing missing data thus, is that it could introduce substantial bias.

1. Imputation of data is also a common technique used to deal with missing data where the data is substituted with the best guess.
    1. Imputation with mean: Missing data is replaced by the mean of the column. This is a commonly used technique. However, this might not be appropriate if the data is not unimodal (for example suppose we fill missing value of weights, the mean of weights for males might be different from females and this might not be a unimodal distribution).
    2. Imputation with median: Missing data is replaced by the median of the column. A median is better than the mean when there are outliers, but once again, if the data is multi-model with multiple clusters, median might not work.
    3. Imputation with Mode: Missing data is replaced with mode of the column. This also leads to similar problems as the above two methods.
    4. Imputation with linear regression: With real valued data, this is another common technique. The missing value is replaced by performing linear regression based on the other feature values. This overcomes the problems with the above simpler forms of imputation.

12. What is A/B testing?

Ans: A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. 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.

A/B test is the shorthand for a simple controlled experiment. As the name implies, two versions (A and B) of a single variable are compared, which are identical except for one variation that might affect a user's behaviour. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, this becomes more complex.

A/B tests are useful for understanding user engagement and satisfaction of online features, such as 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.

Today, A/B tests are being used to run more complex experiments, such as network effects when users are offline, how online services affect user actions, and how users influence one another. Many jobs use the data from A/B tests. This includes, data engineers, marketers, designers, software engineers, and entrepreneurs.

13. Is mean imputation of missing data acceptable practice?

Ans: It is considered bad practice in general to adopt because

- 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
- If the data is sorted mean cannot be accepted as the value might be increasing or decreasing so can't put the mean value everywhere it's better to use B-fill or F-fill method.

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.

Three major uses for regression analysis are

(1) determining the strength of predictors

(2) forecasting an effect

(3) trend forecasting.

15. What are the various branches of statistics?

Ans: If we consider the branches of statistics, there are two branches in it. They are

- **Descriptive statistics**
- **Inferential statistics**

Descriptive-statistics:

It organizes raw data into meaningful information. A house hold articles manufacturing company would like to know what people feel about their products. For that purpose, the company forms a team of people and tries to collect information from the public. The team of people formed by the company is trying to collect data from the public directly. The data which is being collected directly from the public will always not be meaning full. Hence, the data which is being collected directly from the public has to be converted in to meaningful information. This is the work being done in this particular branch “descriptive-statistics". That is, it focuses on collecting, summarizing and presenting set of data.

For example, Industrial statistics, population statistics, trade statistics etc.,

Inferential-statistics:

It analyses sample data to draw conclusion about population. It analyses sample data to draw conclusion about population. Marketing research team of a company wants to know how far the people need a particular product manufactured by the company. There are one hundred thousand populations in a particular city. It is bit difficult to go and ask all one hundred thousand people, due to time consumption and other factors. Hence, it takes a sample of 1000 people to draw conclusion for the whole population. That is making general statement from the study of particular cases or any treatment of data, which leads to prediction or inference concerning a larger group of data.

For example, we want to have an idea about percentage of illiterates in a country. We take a sample from a population and the proportion of illiterates in the sample. That sample with the help of probability enables us to find the proportion to the original population.