Statistics –Worksheet -1

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

Answer : 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 : 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 : 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 : All of the mentioned

5. __________random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer : Poisson

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

a) True

b) False

Answer : 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 : 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 : 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 : Outliers cannot conform to the regression relationship

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

Answer : Normal distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The normal distribution is the proper term for a probability bell curve. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew. Normal distributions are symmetrical, but not all symmetrical distributions are normal.

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation.

The normal distribution model is important in statistics and is key to the Central Limit Theorem(CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled.

The normal distribution is one type of symmetrical distribution. Symmetrical distributions occur when where a dividing line produces two mirror images. Not all symmetrical distributions are normal, since some data could appear as two humps or a series of hills in addition to the bell curve that indicates a normal distribution. First, its mean (average), median(midpoint), and mode(most frequent observation) are all equal to one another. Moreover, these values represent the peak, or highest point, of the distribution. The distribution then falls symmetrically around the mean, the width of which is defined by the standard deviation.

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

Answer : Missing data is handled by some advanced imputational techniques or by using some Imputer Function. The imputer is an estimator function used to fill the missing values in datasets. For numerical values , it uses mean , median and constants . For Categorical values , it uses most frequently used and constant value.

We can handle missing values using some Imputers such as :

- **Simple Imputer**
- **KNN Imputer**
- **Iterative Imputer**

Simple Imputer : Simple Imputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset**.** It replaces the NaN values with a specified placeholder.

KNN Imputer: Each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors found in the training set. Two samples are close if the features that neither is missing are close.It is an imputation technique for completing missing values using k-Nearest Neighbors.

Iterative Imputer: This method treats other columns(which doesn't have nulls as feature , train them , and treat null column as label. Finally it will predict tha NaN data and impute . It's just like Regression , here null column is Label.

12. What is A/B testing?

Answer: A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.

To understand what A/B testing is about, let’s consider two alternative designs: A and B. Visitors of a website are randomly served with one of the two. Then, data about their activity is collected by web analytics. Given this data, one can apply statistical tests to determine whether one of the two designs has better efficacy.

Now, different kinds of metrics can be used to measure a website efficacy. With **discrete metrics**, also called **binomial metrics**, only the two values **0** and **1** are possible. The following are examples of popular discrete metrics.

- Click-through rate — if a user is shown an advertisement, do they click on it?

- Conversion rate — if a user is shown an advertisement, do they convert into customers?

- Bounce rate — if a user is visits a website, is the following visited page on the same website?

With continuous metrics, also called non-binomial metrics,, the metric may take continuous values that are not limited to a set two discrete states.

13. Is mean imputation of missing data acceptable practice?

Answer: The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Mean imputation (MI) is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean.

14. What is linear regression in statistics?

Answer : 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.

linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Understanding linear regression is important because it provides a scientific calculation for identifying and predicting future outcomes. The ability to find predictions and evaluate them can help provide benefits to many businesses and individuals, like optimized operations and detailed research materials.

Real life example of Linear Regression:

Linear regressions can be used in business to evaluate trends and make estimates or forecasts. For example, if a company's sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months.

15. What are the various branches of statistics?

Answer: Statistics, the science of collecting, analyzing, presenting, and interpreting data. Governmental needs for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics. Currently the need to turn the large amounts of data available in many applied fields into useful information has stimulated both theoretical and practical developments in statistics.

Branches of Statistics:

The two major areas of statistics are known as **descriptive statistics, which describes the properties of sample and population data, and inferential statistics**, **which uses those properties to test hypotheses and draw conclusions. Descriptive statistics include mean (average), variance, skewness.**

Descriptive Statistics : Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables.

Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.

Measures of central tendency describe the center of the data set (mean, median, mode).

Measures of variability describe the dispersion of the data set (variance, standard deviation).

Measures of frequency distribution describe the occurrence of data within the data set (count).

Inferential Statistics : Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data. Apart from inferential statistics, descriptive statistics forms another branch of statistics. Inferential statistics help to draw conclusions about the population while descriptive statistics summarizes the features of the data set.

There are two main types of inferential statistics - hypothesis testing and regression analysis. The samples chosen in inferential statistics need to be representative of the entire population. In this article, we will learn more about inferential statistics, its types, examples, and see the important formulas.

Inferential statistics helps to develop a good understanding of the population data by analyzing the samples obtained from it. It helps in making generalizations about the population by using various analytical tests and tools. In order to pick out random samples that will represent the population accurately many sampling techniques are used. Some of the important methods are simple random sampling, stratified sampling, cluster sampling, and systematic sampling techniques.

Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. The goal of inferential statistics is to make generalizations about a population. In inferential statistics, a statistic is taken from the sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).