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

a) True

b) False

Answer: 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

Answer: 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

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.

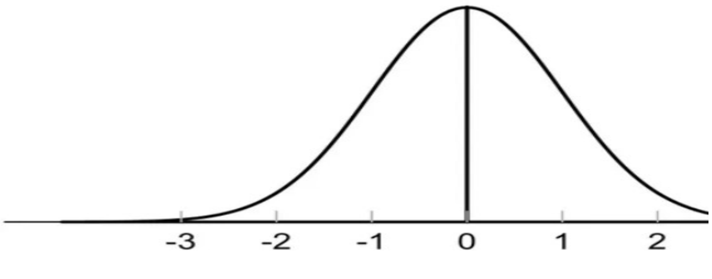
10. What do you understand by the term Normal Distribution?

Answer: The **Normal Distribution**, also called the **Gaussian distribution**, is the most significant continuous probability distribution. Sometimes it is also called a bell curve. A large number of random variables are either nearly or exactly represented by the normal distribution, in every physical science and economics.

A probability function that specifies how the values of a variable are distributed is called the normal distribution. It is symmetric since most of the observations assemble around the central peak of the curve. The probabilities for values of the distribution are distant from the mean narrow off evenly in both directions.

There are two main parameters of normal distribution in statistics namely mean and standard deviation. The location and scale parameters of the normal distribution can be estimated using these two parameters.

Normal Distribution Graph



Normal Distribution Formula

The probability density function of normal or Gaussian distribution is given by;

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

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

Important Properties of Normal Distribution

Some of the important properties of the normal distribution are listed below:

- In a normal distribution, the mean, median and mode are equal.(i.e., Mean = Median= Mode).
- The total area under the curve should be equal to 1.
- The normally distributed curve should be symmetric at the centre.
- There should be exactly half of the values are to the right of the centre and exactly half of the values are to the left of the centre.
- The normal distribution should be defined by the mean and standard deviation.
- The normal distribution curve must have only one peak. (i.e., Unimodal)
- The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.

Applications of Normal Distribution

The normal distributions are closely associated with many things such as:

- Marks scored on the test
- Heights of different persons
- Size of objects produced by the machine
- Blood pressure and so on.

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

Answer: When dealing with missing data, most commonly used two primary methods to solve the error: **imputation or the removal of data**.

The **imputation** method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to **remove data**. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Before deciding which approach to employ, we must understand why the data is missing.

Missing at Random (MAR)

Missing at Random means the data is missing relative to the observed data. It is not related to the specific missing values. The data is not missing across all observations but only within sub-samples of the data It is not known if the data should be there; instead, it is missing given the observed data. The missing data can be predicted based on the complete observed data.

Missing Completely at Random (MCAR)

In the MCAR situation, the data is missing across all observations regardless of the expected value or other variables. Data scientists can compare two sets of data, one with missing observations and one without. Using a t-test, if there is no difference between the two data sets, the data is characterized as MCAR.

Data may be missing due to test design, failure in the observations or failure in recording observations. This type of data is seen as MCAR because the reasons for its absence are external and not related to the value of the observation.

It is typically safe to remove MCAR data because the results will be unbiased. The test may not be as powerful, but the results will be reliable.

Missing Not at Random (MNAR)

The MNAR category applies when the missing data has a structure to it.

Deletion

There are two primary methods for deleting data when dealing with missing data**: listwise and dropping variables.**

Listwise

In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, it may be the most efficient method to eliminate those cases from the analysis. However, in most cases, the data are not missing completely at random (MCAR). Deleting the instances with missing observations can result in biased parameters and estimates and reduce the statistical power of the analysis.

Pairwise

Pairwise deletion assumes data are missing completely at random (MCAR), but all the cases with data, even those with missing data, are used in the analysis. Pairwise deletion allows data scientists to use more of the data. However, the resulting statistics may vary because they are based on different data sets. The results may be impossible to duplicate with a complete set of data.

Dropping Variables

If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

Dropping complete columns

If a column holds a lot of missing values, say more than 80%, and the feature is not meaningful, that time we can drop the entire column.

Imputation

When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.

Instead of deletion, we have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data

Imputation techniques:

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem. Some Imputation techniques can be broadly classified as follows:

Complete Case Analysis (CCA)

Imputation with a constant value

Imputation using the statistics (mean, median, mode)

K-Nearest Neighbor Imputation

Complete Case Analysis (CCA):-

This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing.

Imputation with constant value:

It replaces the missing values with either zero or any constant value.

Imputation using Statistics:

The syntax is the same as imputation with constant only the Simple Imputer strategy will change. It can be "**Mean" or "Median" or "Most_Frequent".**

"Mean" *will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.*

"Median" *will replace missing values using the median in each column. It is preferred if data is numeric and skewed.*

"Most_frequent" *will replace missing values using the most_frequent in each column. It is preferred if data is a string(object) or numeric.*

Before using any strategy, the foremost step is to check the type of data and distribution of features(if numeric)

K_Nearest Neighbor Imputation:

The KNN algorithm helps to impute missing data by finding the closest neighbors using the Euclidean distance metric to the observation with missing data and imputing them based on the non-missing values in the neighbors.

12. What is A/B testing?

Answer: A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing**, A** refers to **'control'** or the original testing variable, whereas **B** refers to **'variation'** or a new version of the original testing variable.

The version that moves your business metric(s) in the positive direction is known as the **'winner**.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

The metrics for conversion are unique to each website. For instance, in the case of ecommerce, it may be the sale of the products. Meanwhile, for B2B, it may be the generation of qualified leads.

A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc. If you're not A/B testing your website, you're surely losing out on a lot of potential business revenue

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 and not acceptable since it ignores feature correlation. Mean imputation reduces the variance of the imputed variables. Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.

Mean imputation does not preserve relationships between variables such as correlations. 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. This method **can lead into severely biased estimates even if data are** Missing Completely at Random **(MCAR)**

14. What is linear regression in statistics?

Answer: Linear regression in statistics 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.

A regression is **a statistical technique that relates a dependent variable to one or more independent (explanatory) variables**. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.

Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line

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.

15. What are the various branches of statistics?

Answer: Statistics is a method of interpreting, analyzing and summarizing the data. Hence, the types of statistics are categorized based on these features: Descriptive and inferential statistics. Based on the representation of data such as using pie charts, bar graphs, or tables, we analyze and interpret it.

Statistics is the application of Mathematics, which was basically considered as the science of the different types of stats. For example, the collection and interpretation of data about a nation like its economy and population, military, literacy, etc.

In terms of mathematical analysis, the statistics include linear algebra, stochastic study, differential equation and measure-theoretic probability theory.

Statistics have majorly categorized into two Branches:

1. **Descriptive statistics**
2. **Inferential statistics**

Descriptive Statistics

In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, 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.

Descriptive statistics are also categorized into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

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.

Measure of position describes the percentile and quartile ranks.

Inferential Statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed and summarized 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.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.