**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

The Bernoulli distribution is a discrete probability distribution in which the random variable can take only two possible values 0 or 1,

where 1 is assigned in case of success or occurrence (of the desired event) and 0 on failure or non-occurrence.

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

Poisson distribution is used for modeling unbounded 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

Many random variables, properly normalized, limit to a normal distribution.

5. \_\_\_\_\_\_ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

ans: c

Poisson distribution is used to model counts.

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

a) True

b) False

ans: b

Usually, replacing the standard error by its estimated value doesn't change the CLT.

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

ans: b

The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative 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

Ans: a

In statistics and applications of statistics, normalization can have a range of meanings.

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 can conform to the regression relationship.

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

Ans: A normal distribution resembles an asymmetric arrangement of most of the values around the mean, such that the curve so formed

looks like a bell. It has two key parameters: the mean (µ) and the standard deviation (σ). This probability method plays a crucial

role in asset return calculation and risk management strategy decisions.

The possible outcomes of the function are given in terms of whole real numbers lying between -∞ to +∞. The tails of the bell curve extend on both sides of the chart (+/-) without limits.

Approximately 68% of all observations fall within +/- one standard deviation(σ).

About 95% of all observations fall within +/- two standard deviations (σ).

Nearly 99.7% of all observations fall within +/- three standard deviations (σ).

=>Skewness refers to symmetry. If skewness is 0, the data is perfectly symmetrical. If the normal distribution is uneven with a skewness greater than zero or positive skewness, then its

right tail will be more prolonged than the left. Similarly, for negative skewness, the left tail will be longer than the right tail. Negative skewness means skewness is less than zero.

=>Kurtosis is a measure of peakiness. If the kurtosis is 3, the probability data is neither too peaked nor too thin at tails. If the kurtosis is more than three, then the data curve is

heightened with fatter tails. Alternatively, if the kurtosis is less than three, then the represented data has thin tails with the peak point lower than the normal distribution.

For a normal distribution, the kurtosis is 3.

\*Characteristics of Normal Distribution:

Normal Distribution has the following characteristics that distinguish it from the other forms of probability representations:

->Empirical Rule: In a normal distribution, 68% of the observations are confined within -/+ one standard deviation, 95% of the values fall within -/+ two standard deviations,

and almost 99.7% of values are confined to -/+ three standard deviations.

->Bell-shaped Curve: Most of the values lie at the center, and fewer values lie at the tail extremities. This results in a bell-shaped curve.

Mean and Standard Deviation: This data representation is shaped by mean and standard deviation.

->Equal Central Tendencies: The mean, median, and mode of this data are equal.

->Symmetric: The normal distribution curve is centrally symmetric. Therefore, half of the values are to the left of the center, and the remaining values appear on the right.

->Skewness and Kurtosis: Skewness is the symmetry. The skewness for a normal distribution is zero. Kurtosis studies the tail of the represented data. For a normal distribution, the kurtosis is 3.

->Total Area = 1: The total value of the standard deviation, i.e., the complete area of the curve under this probability function, is one. Also, the entire mean is zero.

Normal Distribution Curve:

The curve takes the shape of a bell due to the symmetrical arrangement of the values that are concentrated towards the central tendency. At the same time, the tail consists of an insignificant number of values.

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

Ans: Missing data is a sad fact when it comes to data analytics. We cannot avoid situations like these entirely because there are several remedial steps we need to take to make sure it

doesn’t adversely affect the analytics process. While these methods are helpful, they are not foolproof because they are contentious, meaning, their effectiveness depends heavily on

circumstances. The best option available to data scientists is to work with powerful, processing tools that can make the data capturing and analysis process significantly easier.

It is the best way to handle missing data.

->Use deletion methods to eliminate missing data:

The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion.

It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without

significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical

for large datasets. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organizations can fill out the gaps manually. Pairwise deletion is the

process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries were

variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

->Use regression analysis to systematically eliminate data:

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic

regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is

that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

->Data scientists can use data imputation techniques:

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data

entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the

data scientists utilize the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilizing this method is the

three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

12. What is A/B testing?

Ans: A/B testing allows any organization to be more data-driven and strategic about their digital communications. It removes the guesswork from decision making and lets the data decide the path forward.

Instead of spending valuable meeting time debating what color the button should be (who doesn’t love that, right?), split testing helps facilitate the conversation to focus more on the data, rather

than opinion or emotion.

->A/B testing allows individuals, teams, and companies to make careful changes to their user experiences while collecting data on the results. This allows them to construct hypotheses and to learn why

certain elements of their experiences impact user behavior. In another way, they can be proven wrong—their opinion about the best experience for a given goal can be proven wrong through an A/B test.

->When running an A/B test on a webpage, traffic is usually split between some users who will see the control, or the original experience (for example, the blue button), and those who will see the variation,

or the test experience (the yellow button). Unlike qualitative testing or research where users tell us what they will do, during an A/B test, data are collected on what the users do when choosing

between the control and the variation.

->There are two foundational principles that everyone should understand about experimentation: random selection and statistical significance.

-> In order to have confidence in the results, users who are shown the variation should be representative of the targeted user base – for example, all users should be people in the market to buy a pair of

boots. In most cases, the number of users is split evenly between control and variation. This is what we mean by random selection, and this is usually employed in testing to avoid any bias. Note, however,

that sometimes only a small portion of users are selected to see the test variation to minimize risk.

->Statistical significance is a measure of the probability of an outcome – whether it is accurate or simply due to luck or random chance. For example, if an analyst says that the test result of 5% increase

in conversion rate has a statistical significance of 90% confidence, it means that you can be 90% sure that the test results can be trusted.

=>A/B testing process:

The following is an A/B testing framework you can use to start running tests:

Collect data: Your analytics will often provide insight into where you can begin optimizing. It helps to begin with high traffic areas of your site or app to allow you to gather data faster. Look for pages

with low conversion rates or high drop-off rates that can be improved.

Identify goals: Your conversion goals are the metrics that you are using to determine whether the variation is more successful than the original version. Goals can be anything from clicking a button

or link to product purchases and e-mail signups.

Generate hypothesis: Once you've identified a goal you can begin generating A/B testing ideas and hypotheses for why you think they will be better than the current version. Once you have a list of ideas,

prioritize them in terms of expected impact and difficulty of implementation.

Create variations: Using your A/B testing software (like Optimizely), make the desired changes to an element of your website or mobile app experience. This might be changing the color of a button, swapping

the order of elements on the page, hiding navigation elements, or something entirely custom. Many leading A/B testing tools have a visual editor that will make these changes easy. Make sure to QA your

experiment to make sure it works as expected.

Run experiment: Kick off your experiment and wait for visitors to participate! At this point, visitors to your site or app will be randomly assigned to either the control or variation of your experience.

Their interaction with each experience is measured, counted, and compared to determine how each performs.

Analyze results: Once your experiment is complete, it's time to analyze the results. Your A/B testing software will present the data from the experiment and show you the difference between how the two

versions of your page performed and whether there is a statistically significant difference.

If your variation is a winner, congratulations! See if you can apply learnings from the experiment on other pages of your site and continue iterating on the experiment to improve your results. If your

experiment generates a negative result or no result, don't worry. Use the experiment as a learning experience and generate new hypothesis that you can test.

13. Is mean imputation of missing data acceptable practice?

Ans: Mean imputation is generally bad practice because it doesn’t take into account feature correlation. It is acceptable when the missing value proportion is not large enough. But, when the missing values

are large enough and you impute them with the mean, the standard errors will be lesser than what they actually would have been.

->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

In summary: There are a few advantages, but many serious drawbacks. On top of that, we can also benefit from the advantages with more advanced imputation methods (e.g., predictive mean matching or stochastic

regression imputation). To make it short, there is basically no excuse for using mean imputation.

14. What is linear regression in statistics?

Ans: In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

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.

->Linear regression shows the linear relationship between two variables.

15. What are the various branches of statistics?

Ans: Statistics is a form of mathematical analysis that uses quantitative models to give a set of experimental data or studies of real life. Statistics examine the methodology for collecting, reviewing, analyzing,

and making data conclusions. Some statistical measures include the following:

Mean: It is an important concept in mathematics and statistics. The mean is an average and the most common value in the collection of numbers.

Regression analysis: It is a powerful statistical method. It allows us to examine the relationship between two or more variables of interest.

Skewness: In statistics, skewness is a degree of asymmetry that is observed in a probability distribution. Distributions can display right (positive) skewness or left (negative) skewness to differing degrees.

A normal distribution (bell curve) presents zero skewness.

Kurtosis: It is a measure of the combined weight of a distribution’s tails relative to the center of the distribution.

Variance: It estimates the variability from the mean or average.

Analysis of variance: The method of statistics that separates the variance data into several components used for additional tests.

->Branches of Statistics:

DESCRIPTIVE statistics: describe and summarize data. Examples include numerical measures, like averages and correlation. Standard deviation is another descriptive statistic.

Descriptive statistics explain only the population you are studying. Scientists cannot use the information to generalize other groups. There are two types of descriptive statistics: measures of spread and measures

of central tendency.

=>Descriptive statistics have two parts.

Central tendency measures

Variability measures

->Measures of Central Tendency:

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

MEAN: Mean is a conventional method used to describe the central tendency. Typically, 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. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.

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.

Inference statistics often speak in terms of probability by using descriptive statistics.

=>Different types of inferential statistics include:

REGRESSION ANALYSIS: It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear,

and nonlinear. The most well-known models are simple linear and multiple linear.

ANALYSIS OF variance (ANOVA): ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the

relationship between the dependent and independent variables.

ANALYSIS OF COVARIANCE (ANCOVA): It is used to test categorical variables’ main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The

control variables are known as covariates.

STATISTICAL SIGNIFICANCE(t-test): It is used to determine a significant difference between the means of two groups related to features. A t-test studies the t-statistic, the t-distribution values, and

the degree of freedom to learn the statistical significance.

CORRELATION ANALYSIS: It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.