Statistics worksheet 1

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

Answer:- A) True

Q.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?

Answer:- A) Central Limit Theorem

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

Answer:- B) Modeling bounded count data

Q.4- Point out the correct statement.

Answer:- D) All of the mentioned

Q.5- _____ random variables are used to model rates.

Answer:- C) Poisson

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

Answer:- B) False

Q.7- Which of the following testing is concerned with making decisions using data?

Answer:- B) Hypothesis

Q.8- Normalized data are centered at_____and have units equal to standard deviations of the original data.

Answer:- A) 0

Q.9- Which of the following statement is incorrect with respect to outliers?

Answer:- C) Outliers cannot conform to the regression relationship

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

Answer:- A normal distribution is a type of probability distribution that describes how likely it is to observe a certain value of a continuous variable.

It has a bell-shaped curve that is symmetric around the mean, which is the average value of the variable.

The standard deviation is a measure of how spread out the values are from the mean.

The normal distribution is useful for modelling many natural phenomena, such as heights, weights, IQ scores, and errors.

The normal distribution is also called the Gaussian distribution. After the mathematician Carl Friedrich Gauss who discovered many of its properties.

The normal distribution is completely determined by its mean and standard deviation, and it has the same shape regardless of the units of measurement.

*Some properties of normal distribution are:-

- Normal Distribution curve is symmetric about the mean.

- Normal Distribution is unimodal in nature. i.e., it has single peak value.

- Normal Distribution curve is always bell-shaped.

- Mean, Mode and Median for normal distribution is always same.

- Normal distribution follows the empirical rule.

The normal distribution has some important properties, such as the 68-95-99.7 rule, which states that about 68%, 95% and 99.7% of the values are within one, two and three standard deviations from the mean.

The normal distribution is often used as an approximation for other distributions, such as the binomial, poisson and exponential distributions, when certain conditions are met.

The normal distribution is also the basis for many statistical tests and methods, such as hypothesis testing, confidence intervals, and linear regression.

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

Answer:- Missing data is a common problem in real-world datasets that can affect the quality and performance of machine learning models.

There are different ways to handle missing data, depending on the type and cause of the missingness, the amount and pattern of the missing values, and the goal of the analysis.

One way to handle missing data is to drop the rows and columns that have missing values.

This is a simple and fast method, but it can result in a loss of information and reducing the sample size. This method is only recommended when the missing data is completely at random (MCAR) and the proportion of the missing values is very small.

Another way to handle missing data is to impute the missing values, that is, to fill in the missing values with some substitude values.

This can preserve the size and structure of the data, but it can also introduce bias and uncertainty.

There are many imputation techniques available, such as:

- Imputing with a constant number, such as zero, the mean, or the median of the column. This is simple and easy method, but it can distort the distribution and variance of the data and ignore the relationship between variables.
- Imputing with a random value, such as a value drawn from the same distribution as the observed values. This can preserve the distribution and variance of the data, but it can also introduce noise and randomness.
- Imputing with a value based on other variables, such as using a regression model, a nearest neighbor method, or a machine learning algorithm. This can capture the relationship between variables and produce more realistic values, but it can also be computationally expensive and prone to overfitting.
- Imputing with multiple values, such as using a multiple imputation method that generates several imputed datasets and then combines them using some rules. This can account for the uncertainty and variability of the imputation process, but it can also be complex and time consuming.

The choice of imputation techniques depends on the characteristics and objectives of the data and the analysis.

There is no single best method that works for all cases.

*Some of the factors that can influence the choice of imputation technique are:

- The type of data, such as numeric, categorical, date-time or mixed. Different types of data may require different imputation methods or transformations.

- The mechanism of missingness, such as MCAR, MAR, or NMAR. Different mechanisms of missingness may imply different assumptions and implications for the imputation process.

- The pattern of missingness, such as univariate, multivariate, monotone or arbitrary. Different patterns of missingness may affect the complexity and feasibility of the imputation methods.

- The amount of missingness, such as the percentage of missing values, that number of missing values per row or column, and the number of variables with missing values. Different amount of missingness may affect the accuracy and reliability of the imputation methods.

- The goal of the analysis, such as descriptive, inferential, predictive or exploratory. Different goals of the analysis may require different levels od precision and validity of the imputed data.

Q.12- What is A/B testing?

Answer:- A/B testing is a method of comparing two versions of something, such as a website, an app, or a machine learning model, to see which one performs better.

It is also called split testing or randomized controlled trial.

A/B testing is based on statistics and probability, and it can help measure the impact of a change on a certain outcome, such as click-through rate, conversion rate, or accuracy.

A/B testing can also help optimize a machine learning model by testing different hyperparameters, features or algorithms.

*To conduct an A/B test, you need to:

- Define a clear and measurable goal, such as increasing sales, reducing errors, or improving user satisfaction.

- Split your population into two groups: A(control) and B(treatment).

The groups should be randomly assigned and have similar characteristics.

-Run the test for a sufficient amount of time and collect data on the outcome of interest for both groups.

- Analyze the data and compare the result of groups A and B.

Use a statistical test, such as a t-test, a z-test, or a chi-square test, to determine if the difference between the groups is statistically significant.

A/B testing is a common technique in statistics and machine learning, and it can help you improve your products, services, or models.

*The Process of A/B testing:-

The steps for carrying out an insightful A/B test are listed below. After selecting your variable, you should:

- Build on your hypothesis (what do you expect the result should become?)
- Based on the chosen criteria, create a “Control” group and a “Challenger” group.
- Divide your sample groups at random into subgroups of the same size.
- Decide on the sample size (if applicable to your test)
- Specify what constitudes a statistically significant outcome.
- Make sure that each campaign is only having one test running at a time. (Doing many tests at once risks compromising results and invalidating your test).

Q.13- Is mean imputation of missing data acceptable practice?

Answer:- Mean imputation is a technique that replaces missing values of a variable with the mean of the non-missing values.

It is simple and easy method, but it has many drawbacks, some of the disadvantages are:

- It reduces the variability and standard deviation of the imputed variable, making the data less representative of the true population.
- It biases the estimates of the correlations and regression coefficients involving the imputed variable, as it creates the artificial zeros between the imputed values and other variables.
- It does not account for the mechanism of missingness, which may be related to the variable itself or other variables. This can lead to biases estimates of the mean and other statistics of the imputed variable.

Therefore, mean imputation is not an acceptable practice in most cases, as it can distort the analysis and lead to incorrect conclusions.

There are more advanced methods of imputation, such as multiple imputation or predictive mean matching, that can handle missing data more appropriately and preserve the properties of the data.

Q.14- What is linear regression in statistics?

Answer:- Linear regression is a way of finding the best straight line that describes the relationship between two variables.

For example: if you want to know how the height of a person affects their weights, you can use linear regression to draw a line that shows the average weight for each height.

The line can also help you to predict the weight of a person based on their height, or vice versa.

Linear regression uses a formula like this:

Y = C + b*x,

Where Y is the variable you want to predict (weight), x is the variable you use to make prediction (height), C is a constant (the point where the line crosses the y-axis), and b is the slope of the line (how much y changes when x changes by one unit).

Linear regression is based on some assumptions, such as the data being normally distributed, the relationship being linear, and the errors being independent and equal.

If these assumptions are not met, you may need to use a different type of regression, such as logistic or nonlinear regression.

Linear regression is a common and useful tool in statistics, as it can help you to understand how variables are related, estimate the strength of the relationship, and make predictions based on the data.

Q.15- What are the various branches of statistics?

Answer:- The two main branches of statistics are:

1) Descriptive statistics
2) Inferential statistics

Statistics is the branch of mathematics that deals with collecting , organizing, analyzing, and interpreting data.

[1] Descriptive statistics summarizes and displays the data using measures of central tendency, measures of variability, graphs, tables and charts.

It helps us to understand the main features and patterns of the data.

*Descriptive statistics have two parts:

- Central tendency measures

- Variability measures

Example: The average score of the college students in the math test.

The average age of the people who voted for the winning candidate in the last election.

The average length of the statistics book.

[2] Inferential statistics uses the data to make generalizations and predictions about a larger population based on a sample.

It helps us to test hypotheses, estimate parameters and draw conclusions.

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

Besides, a statistician uses these techniques for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

*Different types of inferential statistics include:

- Regression analysis

- Analysis of variance (ANOVA)

- Analysis of covariance (ANCOVA)

- Statistical significance (t-test)

- correlation analysis

Example: Suppose you want to get an idea about the percentage of the people who love shopping at H&M .

We take the sample of the population and find the proportions of individuals who love the H&M Brand. With the assistance of probability, this sample proportion allows us to make a few assumptions about the population proportion. This study belongs to Inferential statistics.