Worksheet – 1

Statistics

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

a) True b) False

Answer – a) True

Question – 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

Question – 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

Question – 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

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

a) Empirical b) Binomial

c) Poisson d) All of the mentioned

Answer - c) Poisson

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

a) True b) False

Answer – b) False

Question 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 - b) Hypothesis

Question 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 – a) 0

Question 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

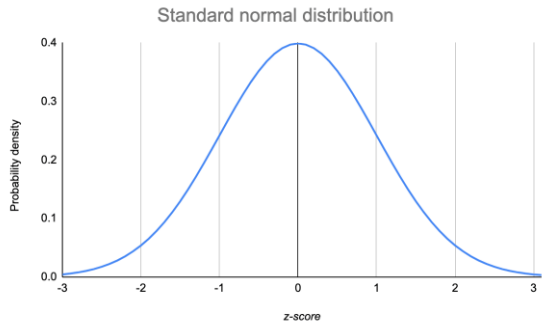
c) Outliers cannot conform to the regression relationship d) None of the mentioned

Answer - c) Outliers cannot conform to the regression relationship

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

Answer- A normal distribution is a type of continuous probability distribution in which data points cluster towards the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.

Graphically, a normal distribution is a bell curve because of its flared shape. The precise shape can vary according to the distribution of the values within the population. The population is the entire set of data points that are part of the distribution.



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

Answer - Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people.It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution

Assume the value from a new person who was not included in the sample.To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables.To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10.Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

Cold deck imputation

A value picked deliberately from an individual with similar values on other variables.In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

Regression imputation

The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Stochastic regression imputation

The predicted value of a regression plus a random residual value.This has all of the benefits of regression imputation plus the random component's benefits.The majority of multiple imputation is based on stochastic regression imputation.

Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time.Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

Question 12- . What is A/B testing?

Answer- A/B testing refers to the experiments where two or more variations of the same webpage are compared against each other by displaying them to real-time visitors to determine which one performs better for a given goal.

A/B testing is basically statistical hypothesis testing, or in other words statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

IN statistics hypothesis breaks down into:

- Null hypothesis
- Alternative hypothesis

The null hypothesis states the default position to be tested or the situation as it is (assumed to be) now, i.e the status quo.

The alternative hypothesis challenges the status quo (the null hypothesis) as is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that you're a/B test will prove to be true.

Question 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 follow 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 the people between the ages of q5 and 80 the eighty-year old will appear to have a significant greater fitness level than he actually does.

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

Question 14-. What is linear regression in statistics?

Answer- 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.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

Question 15-What are the various branches of statistics?

Answer - The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.