FLIP ROBO WORKSHEET

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

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

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

4. Point out the correct statement.
   a) The exponent of 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

5. ______ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
ANS:--C

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False
ANS:--B

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

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

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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

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. For a normal distribution, 68% of the observation are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

The normal distribution model is motivated by the Central Limit Theorem. 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 (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

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

Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programmed will make the decision for you.

Your application will remove things in a list wise sequence most of the time. Depending on why and how much data is gone, list wise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values.

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 utilizing 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 imputations are based on stochastic regression imputation.

Interpolation and extrapolation

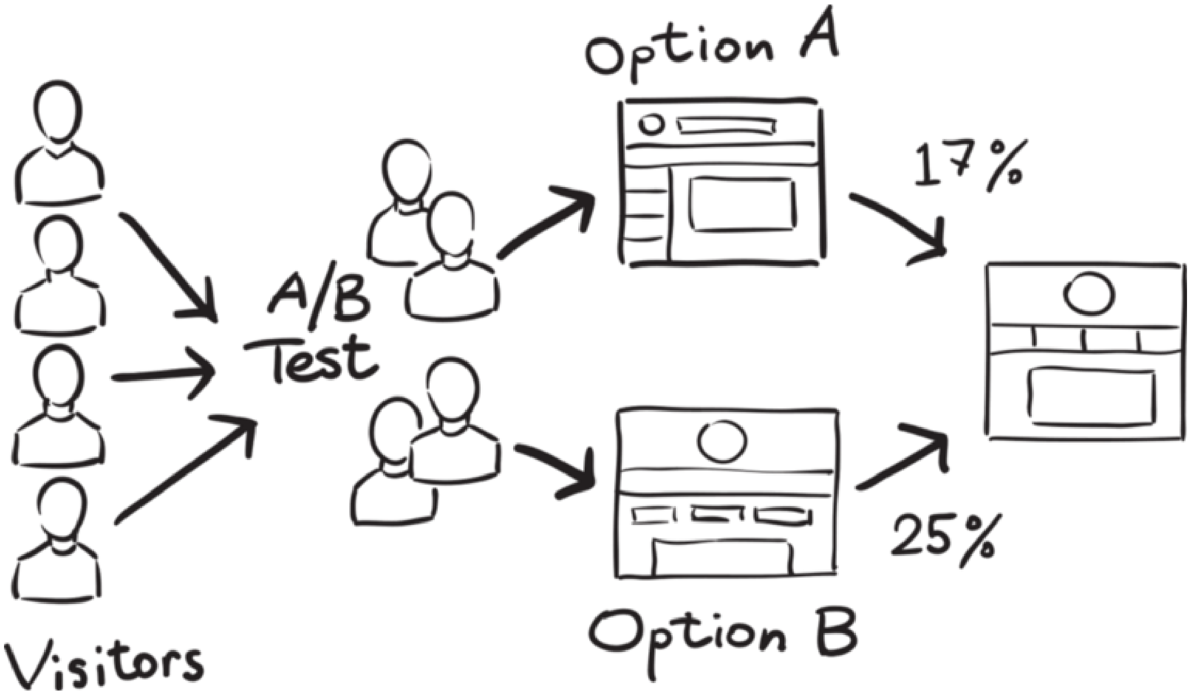
An estimate based on other observations made by the same person. It generally only works with data that is collected over time. Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating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

Q12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



Source

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

14. What is linear regression in statistics?

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 regressed. The independent variables can be called exogenous variables, predictor variables, or repressors.

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

Linear regression is **a kind of statistical analysis that attempts to show a relationship between two variables**. Linear regression looks at various data points and plots a trend line

15. What are the various branches of statistics?

There are three real branches of statistics: **data collection, descriptive statistics and inferential statistics**

Descriptive statistics:

This is a branch of statistics which deals with methods of collection of data, its presentation and organization in various forms, such as distribution tables, graphs (e.g., give, Lorenz curves, etc.), diagrams (e.g., pie charts) and finding measures of central tendency and measures of dispersion or spread which are used in the description of data. Managers, CEOs. Etc. make use of descriptive statistics in presenting their annual reports, financial accounts and bank statements.

Descriptive statistics is used to present the data in an understandable way, so that a meaningful description can be made.

Inferential or predictive statistics:

This is a branch of statistics which deals with techniques used for analysis of data, making estimates that lead to predictions and drawing conclusions or inferences from limited information taken on sample basis and testing the reliability of the estimates or predictions.

Inferential statistics is used to make comparisons or predictions about a larger group, known as population, using information gathered about a small part of that population called a sample.