STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
   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?
   Answer: a) Central Limit Theorem
3. Which of the following is incorrect with respect to use of Poisson distribution?
   Answer: b) Modeling bounded count data
4. Point out the correct statement.
   Answer: d) All of the mentioned
5. _____ random variables are used to model rates.
   Answer: c) Poisson
6. Usually replacing the standard error by its estimated value does change the CLT.
   Answer: b) False
7. Which of the following testing is concerned with making decisions using data?
   Answer: b) Hypothesis
8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   Answer: a) 0
9. Which of the following statement is incorrect with respect to outliers?
   Answer: c) Outliers cannot conform to the regression relationship
10. What do you understand by the term Normal Distribution?

Answer:

Normal Distribution is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes.

The formula for the calculation can be represented as

$$\mathbf{X \sim N (\mu, \alpha)}$$

Where

- N= no of observations
- μ= mean of the observations
- α= standard deviation

In most of the cases, the observations do not reveal much in its raw form. So it is essential to standardize the observations to be able to compare that. It is done with the help of the **z-score formula**. It is required to calculate the Z-score for an observation.
The equation for Z Score Calculation for the normal distribution is represented as follows,

$$\mathbf{Z=(X- \mu)/ \alpha}$$

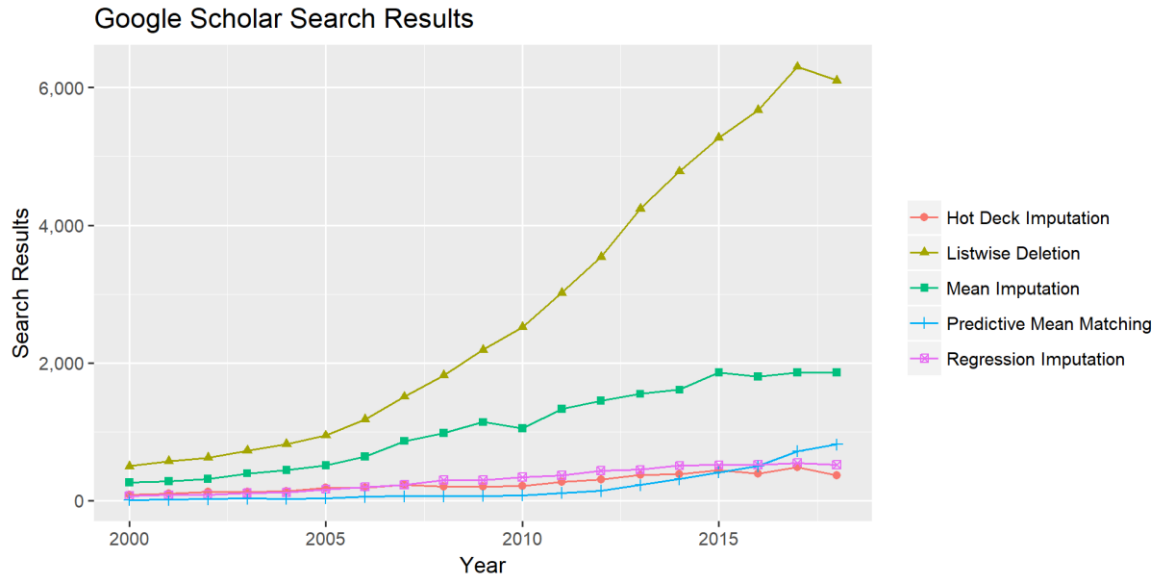
11. How do you handle missing data? What imputation techniques do you recommend?

Answer:

Missing Values – Statistical Analysis & Handling of Incomplete Data. Missing Data Definition: Missing data (or missing values) appear when no value is available in one or more variables of an individual. Missing data can occur due to several reasons, e.g. interviewer mistakes, anonymization purposes, or survey filters.

In order to bring some clarity into the field of missing data treatment, I'm going to investigate in this article, which imputation methods are used by other statisticians and data scientists.

More precisely, I'm going to investigate the popularity of the following **five imputation methods**:



Imputation is the process of replacing missing values with substituted data. It is done as a preprocessing step.

In missing data research literature, these three methods are highly respected for their ability to **improve data quality** (Learn more: regression imputation; predictive mean matching; hot deck imputation).

Regression imputation and hot deck imputation seem to have increased their popularity until 2013. Afterwards, however, both methods **converge at approximately 500** Google Scholar search results per year. In contrast, the popularity of predictive mean matching imputation is pretty low until 2010 (no surprise, the method is quite new), but afterwards its **popularity increases quickly**.

12. What is A/B testing?

Answer :

- A/B testing (also known as split testing or bucket testing) is **a method of comparing two versions of a webpage or app against each other to determine which one performs better**.
- It is **one of the most popular controlled experiments used to optimize web marketing strategies**. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B.
    - Two-sample hypothesis testing
    - Randomized experiments with two variants: A and B
    - A: control; B: variation
    - User-experience design: identify changes to web pages that increase clicks on a banner
    - Current website: control; NULL hypothesis
    - New version: variation; alternative hypothesis
- Now, different kinds of metrics can be used to measure a website efficacy. With **discrete metrics**, also called **binomial metrics**, only the two values **0** and **1** are possible. The following are examples of popular discrete metrics.
    - Click-through rate — if a user is shown an advertisement, do they click on it
    - Conversion rate — if a user is shown an advertisement, do they convert into customers
    - Bounce rate — if a user is visiting a website, is the following visited page on the same website
- With **continuous metrics**, also called **non-binomial metrics**,, the metric may take continuous values that are not limited to a set two discrete states. The following are examples of popular continuous metrics.
    - Average revenue per user — how much revenue does a user generate in a month?
    - Average session duration — for how long does a user stay on a website in a session?
    - Average order value — what is the total value of the order of a user?

13. Is mean imputation of missing data acceptable practice?

Answer:

- Bad practice in general
- 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

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 regress and. 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
(3) Trend forecasting

15. What are the various branches of statistics?

Answer:

- Summary statistics. These are statistics that summarize the data using a single number.
- Graphs. Graphs help us visualize data. Common types of graphs used to visualize data include boxplots, histograms, stem-and-leaf plots,and scatterplots.
- Tables. Tables can help us understand how data is distributed...

Descriptive statistics have two parts:

- Central tendency measures
- Variability measures

To help understand the analyzed data, the tendency measures and variability measures use tables, general discussions, and charts.

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.

Formula of Mean

m= Sum of the terms/numbers of terms

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.

Formula of Median

To solve the median, there are two formulas;

- **When n is odd,**

(n+1 / 2)th observation

- **When n is even,**

median=(n/2)th + (n/2+ 1)th observation / 2

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.