Statistical Learning

Introduction to Statistics

Outline

1. **Why Statistics**
2. **Statistical Methods**
3. **Types of Statistics - Descriptive and Inferential Statistics**
4. **Data Sources and Types of Datasets**
5. **Attributes of Datasets**

Classification

- *Classification* techniques helps in segmenting the customers into appropriate groups based on key characteristics.

- For example, using *appropriate statistical model*, an organization could easily segment the customers into Long Term Customers, Medium Term Customers, and Brand Switchers.

- Another application in this context is classifying customers into "Buyers and Non-Buyers."

- Classification helps professionals understand the customer behavior and position their products and brands using appropriate strategies.

Statistics - Methods

Pattern Recognition

- "A picture is worth thousand words" and it reveals hidden pattern in the data that could be leveraged by retail professionals. Pattern recognition techniques include *Histogram*, *Box Plot*, *Scatter Plot* and other *Visual Analytics*.

- For example, histogram drawn for income of a particular class of customers may reveal a symmetrical bell curve pattern or may be left or right skewed.

- Relationship between age and expenditure could be captured using a scatter plot.

- *Box Plot* enables identification of outliers (extreme points) apart from providing the distribution pattern.

Statistics - Methods

Association

- *Association* Analysis helps in determining which of the items go together. Association rules include a set of analytics that focuses on discovering relationships that exist among specific objects.

- In this context, market basket analysis refers to an association rule that generates the probability for an outcome.

- For example, market basket analysis may lead to a finding that if customers buy coffee, there is a 40% probability that they also buy bread.

- Association rules can be adapted by organizations to store lay cross selling among other items, discount and sales promotion decisions.

Statistics - Methods

Predictive Modeling

- Both customer segmentation as well as identifying and targeting most profitable customers can be facilitated by predictive models.

- *Regression* can be used for predicting the amount of expenditure on a particular product based on input variables income, age, and gender.

$$y = a + bx$$

- Organizations can leverage on other advanced models that comprise *Logistic Regression*, and *Neural Networks* for predicting a target variable as well as classifying and predicting into which group the consumer belongs to.

- For example, these models can classify and predict buyers and non-buyers, and defaulters and non-defaulters on credit card loan.

Classical Definition of Statistics

" By Statistics, we mean methods specially adopted to the elucidation of quantitative data affected to a marked extent by multiplicity of causes".

Yule and Kendal

It is interesting to see what *Thomas Davenport* means by Business Analytics and note the similarities and dissimilarities between the two.

"Business Analytics (BA) can be defined as the broad use of data and quantitative analysis for decision making within organizations".

Types of Statistics

Descriptive Statistics is concerned with Data Summarization, Graphs/Charts, and Tables

Inferential Statistics is a method used to talk about a Population Parameter from a Sample.

Population, Parameter, Sample, Statistic

A **Population** is the universe of possible data for a specified object. Example: People who have visited or will visit a website.   *not observed*

A **Parameter** is a numerical value associated with a population. Example: The average amount of time people spend on a website.   *not observed*
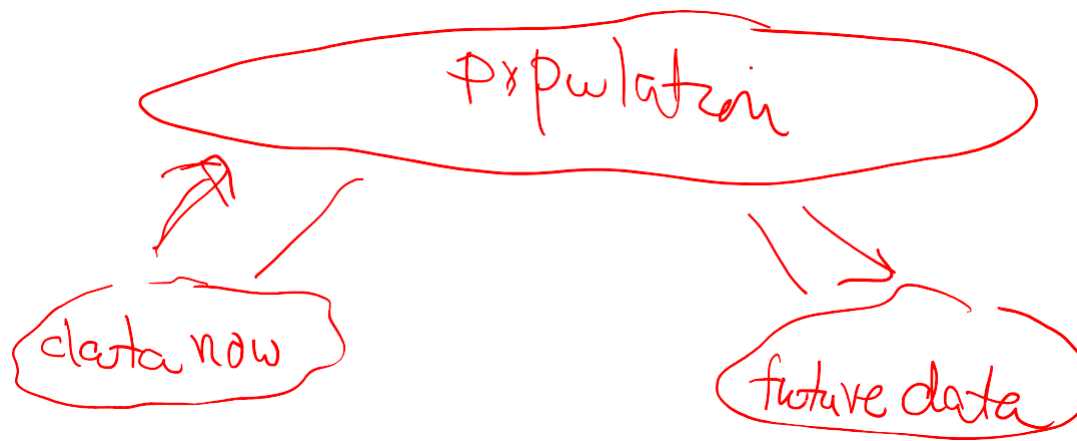
A **Sample** is a selection of observations from a population. Example: People (or IP addresses) who visited a website on a specific day.   *observed*

A **Statistic** is a numerical value associated with an observed sample. Example: The average amount of time people spent on a website on a specific day.   *observed*

Population

data now   *future data*

Data Sources

Primary Data are collected by the organization itself for a particular purpose. The benefits of primary data are that they fit the needs exactly, are up to date, and reliable.

Secondary Data are collected by other organizations or for other purposes. Any data, which are not collected by the organization for the specified purpose, are secondary data. These may be published by other organizations, available from research studies, published by the government, web, social media and so on.

Types of Data

Qualitative Data are nonnumeric in nature and can't be measured. Examples are gender, religion, and place of birth.

Quantitative Data are numerical in nature and can be measured. Examples are balance in your savings bank account, and number of members in your family.

Quantitative data can be classified into discrete type or continuous type. **Discrete type** can take only certain values, and there are discontinuities between values, such as the number of rooms in a hotel, which cannot be in fraction. **Continuous type** can take any value within a specific interval, such as the production quantity of a particular type of paper (measured in kilograms).

Types of Data Sets

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- **Graph and network**
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- **Ordered**
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- **Spatial, image and multimedia:**
  - Spatial data: maps
  - Image data
  - Video data

term

document

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

transaction ID