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

Why Statistics is So Important?

Three significant events triggered the current meteoric growth in the use of analytical decisions making and Statistics is central to all of them.

Event1

- Technological developments, the Revolution of the Internet and social networks, and data generated from mobile phones and other electronic devices produce a large amount of data from which insights will have to be sifted.
- The discovery of patterns and trends from these data for organizations will pave the way for improving profitability, understanding customer expectations, and appropriately pricing their products so that they can gain a competitive advantage in the marketplace.

Why Statistics is So Important?

Event 2

- Advances in enormous computing power to effectively process and analyze massive amounts of data
- Sophisticated and faster algorithms for solving problems
- Data Visualization for Business Intelligence and Artificial Intelligence

Why Statistics is So Important?

Event 3

- Large data storage capability
- Parallel computing and cloud computing coupled with better computer hardware have enabled businesses and other organizations to solve large scale problems faster than ever before without sacrificing

Big Data

Big data

- A set of data that cannot be managed, processed or analyzed with traditional software/algorithms within a reasonable amount of time.
- Big data revolves around

  Volume Velocity Variety Value Veracity

Walmart handles over one million purchase transactions per hour.

Facebook processes more than 250 million picture uploads per day.

Statistics - Methods

Classification

- Classification techniques help in segmenting the customers into appropriate groups based on key characteristics.
- For example, using an 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 a thousand words” and it reveals a 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, a histogram drawn for income of a particular class of customers may reveal a symmetrical bell curve pattern or maybe left or right-skewed.
- Relationship between age and expenditure could be captured using a scatter plot.
- Box Plot enables the 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 the 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.
- Organizations can leverage 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.
- For example, these models can classify and predict buyers and non-buyers, and defaulters and non-defaulters on credit card loans.

Classical Definition of Statistics

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

A Parameter is a numerical value associated with a population. Example: The average amount of time people spend on a website.

A Sample is a selection of observations from a population. Example: People (or IP addresses) who visited a website on a specific day.

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.

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

	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

Data Objects

- Data sets are made up of data objects.
- A data object represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called samples, examples, instances, data points, objects, tuples.
- Data objects are described by attributes.
- Database rows -> data objects; columns ->attributes.

Attributes

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
    - *E.g., customer _ID, name, address*
- Types:
    - Nominal
    - Binary
    - Ordinal
    - Numeric: quantitative
        - Interval-scaled
        - Ratio-scaled

Attribute Types

- Nominal: categories, states, or “names of things”
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- Binary
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes are equally important
    - E.g., gender
- Asymmetric binary: outcomes not equally important.
  - E.g., medical test (positive vs. negative)
  - Convention: assign 1 to the most important outcome (e.g., HIV positive)
- Ordinal
  - Values have a meaningful order (ranking) but the magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- Interval
  - Measured on a scale of equal-sized units
  - Values have an order
    - E.g., the temperature in C˚or F˚, calendar dates
  - No true zero-point
- Ratio
  - Inherent zero-point
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚)
  - E.g., the temperature in Kelvin, length, counts, monetary quantities