Data Analysis Course

Basics & Terminology(Version-1)

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

What is “Statistics”?

- ***Statistics*** is the science of data that involves:
  - ***Collecting***
  - ***Classifying***
  - ***Summarizing***
  - ***Organizing and***
  - ***Interpretation***

Of numerical information.

- *Examples:*
  - Cricket batting averages
  - Stock price
  - Climatology data such as rainfall amounts, average temperatures
  - Marketing information
  - Gambling?

Key Terms

- **What is Data?**
  - facts or information that is relevant or appropriate to a decision maker
- **Population?**
  - the totality of objects under consideration
- **Sample?**
  - a portion of the population that is selected for analysis
- **Parameter?**
  - a summary measure (e.g., mean) that is computed to describe a characteristic of the population
- **Statistic?**
  - a summary measure (e.g., mean) that is computed to describe a characteristic of the sample

Variables

- Traits or characteristics that can change values from case to case.
- Examples:
  - Age
  - Gender
  - Income
  - Social class

Types Of Variables

- In causal relationships:

  CAUSE → EFFECT

  independent variable → dependent variable
- **Independent variable:** is a variable that can be controlled or manipulated.
- **Dependent variable:** is a variable that cannot be controlled or manipulated. Its values are predicted from the independent variable.
- **Discrete** variables are measured in units that cannot be subdivided. Example: Number of children
- **Continuous** variables are measured in a unit that can be subdivided infinitely. Example: Height

Lab

- Print product sales data
- What are cause variables, what are effect variables
- Identify the continuous & discrete variables
- What is the population
- Filter data and pick a sample
- Calculate a parameter (Mean of the population)
- Calculate a statistic
- How close is the statistics to parameter? Is it a good estimate?
- **Self study:** Randomly pick 10 samples, calculate mean for each sample. Find the mean of the means & see whether it is a good estimate of the population mean

Descriptive Statistics

- Gives us the overall picture about data
- Presents data in the form of tables, charts and graphs
- Includes summary data
- Avoids inferences
- Examples:
  - Measures of central location
    - **Mean, median, mode and midrange**
  - Measures of Variation
    - **Variance, Standard Deviation, z-scores**

Details later

Lab

- Download product sales data
- Run proc means to print the descriptive statistics
- Run proc univariate to print the descriptive statistics
- Identify Measures of central location
- Identify Measures of variation

Inferential Statistics

- Take decision on overall population using a sample
- “Sampled” data are incomplete but can still be representative of the population
- Permits the making of generalizations (inferences) about the data
- *Probability theory* is a major tool used to analyze sampled data

-Details later

Predictive Modeling

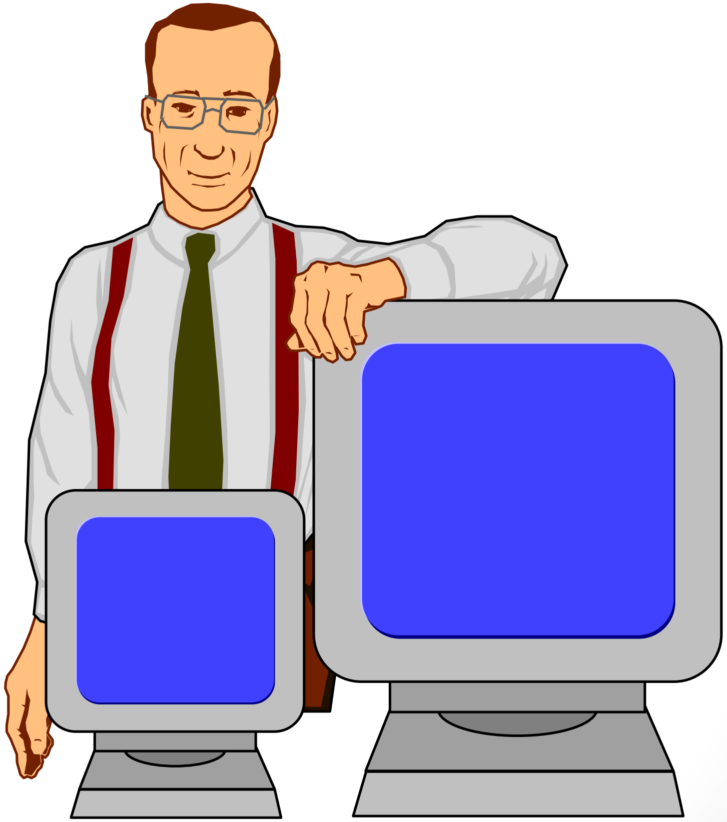
- The science of predicting future outcomes based on historical events.
- ***Model Building: "Developing set of equations or mathematical formulation to forecast future behaviors based on current or historical data."***
- Regression, logistic Regression, time series analysis etc.,

-Details later

Statistical Computer Packages

Typical Software

- SAS
- R
- SPSS
- MINITAB
- Excel



Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

+91 9886 768879