Data Processing and Analysis

Edited by: Dr. Mamunur Rashid

Learning Objectives

- Understand the definition and purpose of data processing
- Define key concepts in data analysis

References

- Data processing, Naseem Ahmed Khan.
- Analyzing and interpreting data, Matt Calvert.
- Exploratory Data Analysis, anonymous.
- Inferential Data Analysis, anonymous.

Outline

- Data Processing
- Data Analysis
  - Descriptive Statistics
  - Inferential Statistics

Data

- Data

The word data is derived from Latin language. It is plural of Datum (But Data is usually used as a singular term.) Datum (singular) – Data (plural). Data is any collection of facts of figures. The data is the raw material to be processed by a computer.

Example

Names of students, marks obtained in the examination, designation of employees, addresses, quantity, rate, sales figures or anything that is input to the computer is data. Even pictures, photographs, drawings, charts and maps can be treated as data. Computer processes the data and produces the output or result

Types of Data

Mainly Data is divided into two types:

1. Numeric Data
2. Character Data

1. Numeric Data

The data which is represented in the form of numbers is known as Numeric Data. This includes 0-9 digits, a decimal point (.), +, /, – sign and the letters "E" or "D".

2. Character Data

Character data falls into two groups.

i. String Data
ii. Graphical Data

Types of Data

- *String Data*

  String data consists of the sequence of characters. Characters may be English alphabets, numbers or space. The space, which separates two words, is also a character. The string data is further divided into two types.

  a. Alphabetic Data
  b. Alphanumeric Data

- *Graphical Data*

  It is possible that pictures, charts and maps can be treated as data. The scanner is normally used to enter this type of data. The common use of this data is found in the National Identity Card.

Information

A collection of data which conveys some meaningful idea is information. It may provide answers to questions like who, which, when, why, what, and how.

or

The raw input is data and it has no significance when it exists in that form. When data is collated or organized into something meaningful, it gains significance. This meaningful organization is information

or

Observations and recordings are done to obtain data, while analysis is done to obtain information

Data Processing

Any operation or set of operations performed upon data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration to convert it into useful information.

- Real time processing
- Batch processing

Real Time Data Processing

In a real time processing, there is a continual input, process and output of data. Data has to be processed in a small stipulated time period (real time), otherwise it will create problems for the system.

For example, when a bank customer withdraws a sum of money from his or her account it is vital that the transaction be processed and the account balance updated as soon as possible, allowing both the bank and customer to keep track of funds.

Batch Data Processing

In a batch processing group of transactions collected over a period of time is collected, entered, processed and then the batch results are produced. Batch processing requires separate programs for input, process and output. It is an efficient way of processing high volume of data.

For example: Payroll system, Examination system and billing system.

Data Analysis

Data analysis and interpretation

- Think about analysis EARLY
- Start with a plan
- Code, enter, clean
- Analyze
- Interpret
- Reflect
  - What did we learn?
  - What conclusions can we draw?
  - What are our recommendations?
  - What are the limitations of our analysis?

Analyzing and Interpreting Quantitative Data

- Quantitative Data is

  Presented in a numerical format

  Collected in a standardized manner

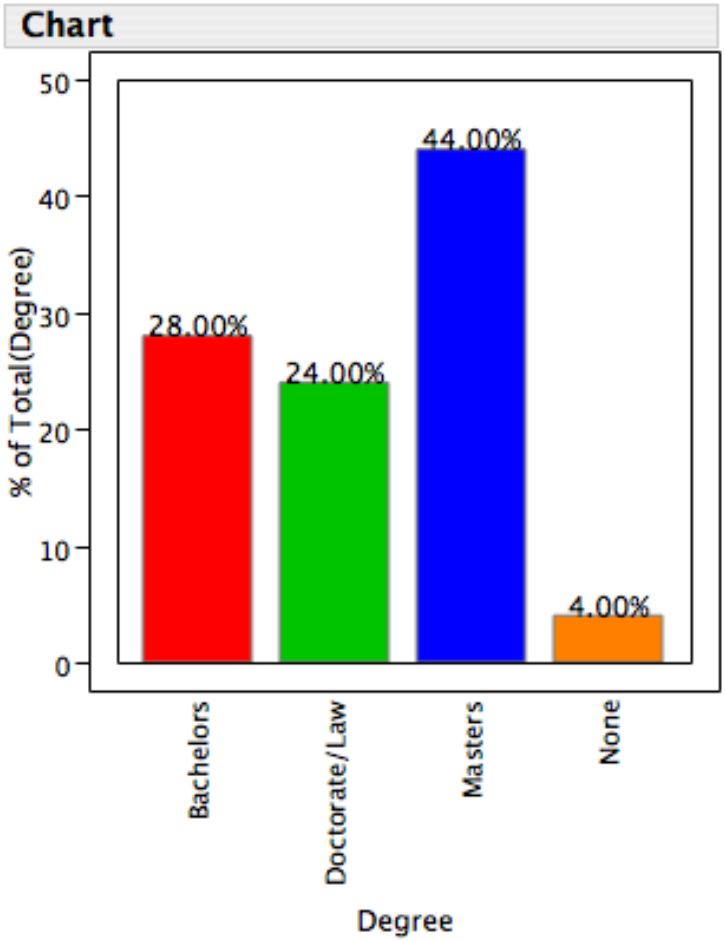
  e.g. surveys, closed-ended interviews, tests, measurements, etc.

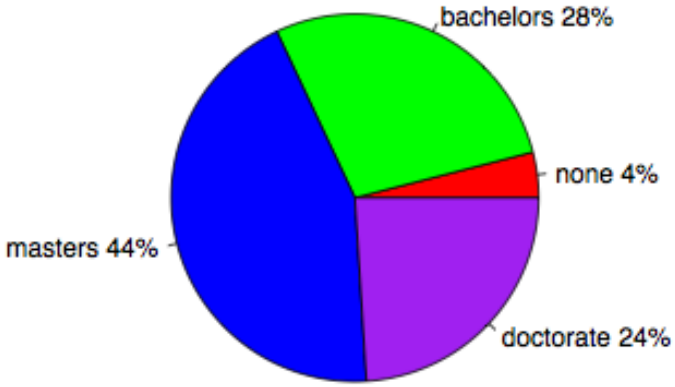
  Analyzed using statistical techniques

Common descriptive statistics

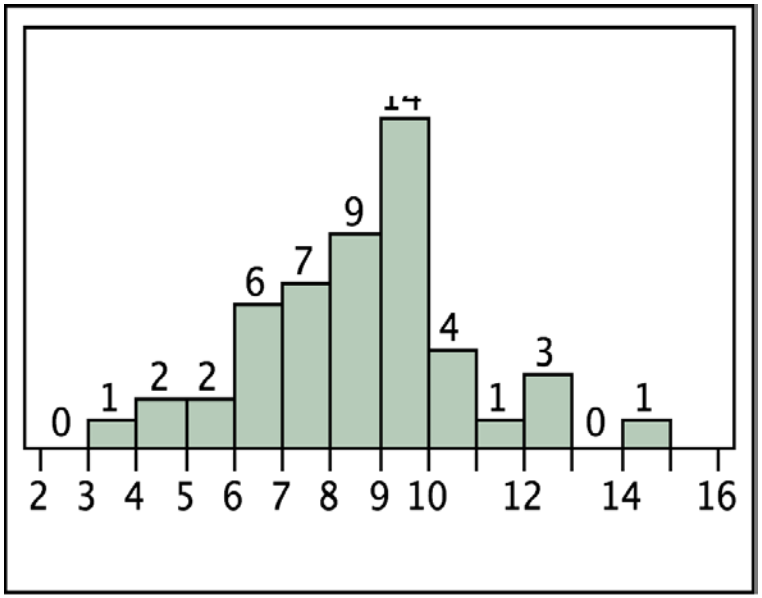
- Count (frequencies)
- Percentage
- Mean
- Mode
- Median
- Range
- Standard deviation
- Variance
- Ranking

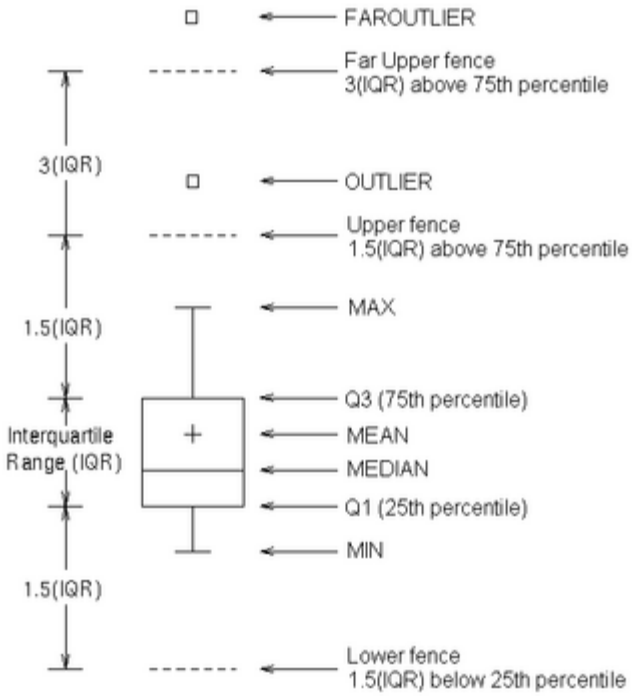
Common descriptive statistics





Common descriptive statistics





Inferential Statistics

- Techniques that allow us to study samples and then make generalizations about the population. Inferential statistics are a very crucial part of scientific research in that these techniques are used to test hypotheses.

Uses for Inferential Statistics

- Statistics for determining differences between experimental and control groups in experimental research
- Statistics used in descriptive research when comparisons are made between different groups
- These statistics enable the researcher to evaluate the effects of an independent variable on a dependent variable

Hypothesis Testing

- The purpose of the statistical test is to evaluate the null hypothesis (H_0) at a specified level of significance (e.g., $p < .05$)
  - In other words, do the treatment effects differ significantly so that these differences would be attributable to chance occurrence less than 5 times in 100?

t-tests

- Characteristics of t-tests
  - requires interval or ratio level scores
  - used to compare two mean scores
  - easy to compute
  - pretty good small sample statistic

Hypothesis Testing Errors

- Hypothesis testing decisions are made without direct knowledge of the true circumstance in the population. As a result, the researcher's decision may or may not be correct
- Type I Error
- Type II Error

Hypothesis Truth Table

DECISION	NULL HYPOTHESIS: TRUE	NULL HYPOTHESIS: FALSE
ACCEPT	CORRECT DECISION	TYPE II ERROR
REJECT	TYPE I ERROR	CORRECT DECISION

ANOVA - Analysis of Variance

- A commonly used family of statistical tests that may be considered a logical extension of the t-test
  - requires interval or ratio level scores
  - used for comparing 2 or more mean scores
  - maintains designated alpha level as compared to experimentwise inflation of alpha level with multiple t-tests
  - may also test more than 1 independent variable as well as interaction effect

Multivariate Tests

- Univariate statistic –
  - used in situations where each participant contributed one score to the data analysis, or in the case of a repeated measures design, one score per cell
- Multivariate statistic –
  - used in situations where each participant contributes multiple scores

Multiple Analysis of Variance

- MANOVA
- Analogous to ANOVA except that there are multiple dependent variables
- Represents a type of multivariate test

Prediction and Regression Analysis

- Correlational technique
- Simple prediction
  - Predicting an unknown score (Y) based on a single predictor variable (X)
  - $Y' = bX + c$
- Multiple prediction
  - Involves more than one predictor variable
  - $Y' = b_1X_1 + b_2X_2 + c$

Discussing limitations

Written reports:

- Be explicit about your limitations

Oral reports:

- Be prepared to discuss limitations
- Be honest about limitations
- Know the claims you cannot make
  - Do not claim causation without a true experimental design
  - Do not generalize to the population without random sample and quality administration (e.g., <60% response rate on a survey)