Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 3
January 6, 2022

Describing Data

Terminologies

- Population, sample, observations - The units on which we measure data - such as persons, cars, animals are called observations. The collection of all units is called population which is denoted by Ω. A selection of observations $\omega_1, \ldots, \omega_n \in \Omega$ is called a sample.

Describing Data

Terminologies

- Population, sample, observations - The units on which we measure data - such as persons, cars, animals are called observations. The collection of all units is called population which is denoted by Ω. A selection of observations $\omega_1, \ldots, \omega_n \in \Omega$ is called a sample.
- Variable (random/statistical) - qualitative and quantitative variables
  - Discrete variables
  - Continuous variables (measured rather than counted)

Describing Data

Terminologies

- Population, sample, observations - The units on which we measure data - such as persons, cars, animals are called observations. The collection of all units is called population which is denoted by Ω. A selection of observations $\omega_1, \ldots, \omega_n \in \Omega$ is called a sample.
- Variable (random/statistical) - qualitative and quantitative variables
  - Discrete variables
  - Continuous variables (measured rather than counted)
- Scales - different variables contain different amounts of information

Describing Data

Terminologies

- Population, sample, observations - The units on which we measure data - such as persons, cars, animals are called observations. The collection of all units is called population which is denoted by Ω. A selection of observations $\omega_1, \ldots, \omega_n \in \Omega$ is called a sample.
- Variable (random/statistical) - qualitative and quantitative variables
  - Discrete variables
  - Continuous variables (measured rather than counted)
- Scales - different variables contain different amounts of information
  - Nominal scale - values can not be ordered

Describing Data

Terminologies

- Population, sample, observations - The units on which we measure data - such as persons, cars, animals are called observations. The collection of all units is called population which is denoted by Ω. A selection of observations $\omega_1, \ldots, \omega_n \in \Omega$ is called a sample.
- Variable (random/statistical) - qualitative and quantitative variables
  - Discrete variables
  - Continuous variables (measured rather than counted)
- Scales - different variables contain different amounts of information
  - Nominal scale - values can not be ordered
  - Ordinal scale - can be ordered but cannot be interpreted/compared in a meaningful/numerical way

Describing Data

Terminologies

- Population, sample, observations - The units on which we measure data - such as persons, cars, animals are called observations. The collection of all units is called population which is denoted by Ω. A selection of observations $\omega_1, \ldots, \omega_n \in \Omega$ is called a sample.
- Variable (random/statistical) - qualitative and quantitative variables
  - Discrete variables
  - Continuous variables (measured rather than counted)
- Scales - different variables contain different amounts of information
  - Nominal scale - values can not be ordered
  - Ordinal scale - can be ordered but cannot be interpreted/compared in a meaningful/numerical way
  - Continuous scale - can be ordered but can be interpreted/compared in a meaningful/numerical way

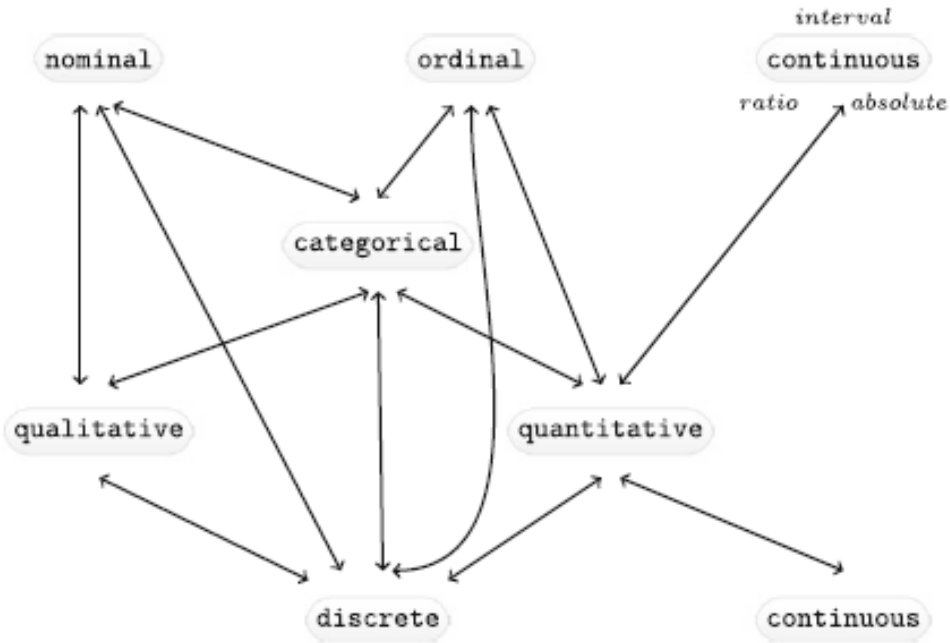
Describing Data

Terminologies

- Population, sample, observations - The units on which we measure data - such as persons, cars, animals are called observations. The collection of all units is called population which is denoted by Ω. A selection of observations $\omega_1, \ldots, \omega_n \in \Omega$ is called a sample.
- Variable (random/statistical) - qualitative and quantitative variables
  - Discrete variables
  - Continuous variables (measured rather than counted)
- Scales - different variables contain different amounts of information
  - Nominal scale - values can not be ordered
  - Ordinal scale - can be ordered but cannot be interpreted/compared in a meaningful/numerical way
  - Continuous scale - can be ordered but can be interpreted/compared in a meaningful/numerical way
- Grouped Data (categorical variable) - instead of the original value, one may only know the category or group the value belongs to

Describing Data

Summary of variable classifications



Units of data

- smallest unit of storage, a 0 or 1

Units of data

- smallest unit of storage, a 0 or 1
- anything that can store two states - now "transistors", used to be vacuum tubes
- with n bits, can store 2^n patterns - so one byte can store 256 patterns

Units of data

- smallest unit of storage, a 0 or 1
- anything that can store two states - now "transistors", used to be vacuum tubes
- with n bits, can store 2^n patterns - so one byte can store 256 patterns
- eight bits = one byte

Units of data

- smallest unit of storage, a 0 or 1
- anything that can store two states - now "transistors", used to be vacuum tubes
- with n bits, can store 2^n patterns - so one byte can store 256 patterns
- eight bits = one byte

The encoding paradigm: Here

Units of data

Multi-byte units

unit	abbreviation	total bytes	nearest decimal equivalent
kilobyte	KB	$1,024^1$	1000^1
megabyte	MB	$1,024^2$	1000^2
gigabyte	GB	$1,024^3$	1000^3
terabyte	TB	$1,024^4$	1000^4
petabyte	PB	$1,024^5$	1000^5
exabyte	EB	$1,024^6$	1000^6
zettabyte	ZB	$1,024^7$	1000^7
yottabyte	YB	$1,024^8$	1000^8

Units of data

Multi-byte units

unit	abbreviation	total bytes	nearest decimal equivalent
kilobyte	KB	$1,024^1$	1000^1
megabyte	MB	$1,024^2$	1000^2
gigabyte	GB	$1,024^3$	1000^3
terabyte	TB	$1,024^4$	1000^4
petabyte	PB	$1,024^5$	1000^5
exabyte	EB	$1,024^6$	1000^6
zettabyte	ZB	$1,024^7$	1000^7
yottabyte	YB	$1,024^8$	1000^8

Observation this is why 1GB is greater than 1 billion bytes

Describing Data

Data Collection - Primary and secondary data - gather data on a sample and draw conclusions about the population of interest

- Survey - collecting data by asking questions

Describing Data

Data Collection - Primary and secondary data - gather data on a sample and draw conclusions about the population of interest

- ▷ Survey - collecting data by asking questions
- ▷ Experiment - generated by the researcher with full control over one or many variables of interest

Describing Data

Data Collection - Primary and secondary data - gather data on a sample and draw conclusions about the population of interest

- Survey - collecting data by asking questions
- Experiment - generated by the researcher with full control over one or many variables of interest
- Observational - without a survey or conducting an experiment

Describing Data

Creating a data set - the data is stored in a data matrix (= data set) with p columns and n rows

Describing Data

Creating a data set - the data is stored in a data matrix (= data set) with p columns and n rows

$$
\begin{array}{c} \omega \\ 1 \\ 2 \\ \vdots \\ n \end{array}
\begin{array}{c}
\begin{array}{cccc} \textit{Variable}\,1 & \textit{Variable}\,2 & \dots & \textit{Variable}\,\mathrm{p} \end{array} \\
\begin{bmatrix}
x_{11} & x_{12} & \dots & x_{1p} \\
x_{21} & x_{22} & \dots & x_{2p} \\
\vdots & \vdots & \vdots & \vdots \\
x_{n1} & x_{n2} & \dots & x_{np}
\end{bmatrix}
\end{array}
$$

Describing Data

Creating a data set - the data is stored in a data matrix (= data set) with p columns and n rows

$$\begin{array}{c} \omega \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} \textit{Variable}\,1 & \textit{Variable}\,2 & \dots & \textit{Variable}\,\mathrm{p} \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \end{array}$$

A statistical convention: Variables are are aften called *features*- the columns

Dataset manipulation

- A dataset is a "rectangular" formatted table of data in which all the values of the same variable must be in a single column
- A dataset is not:
  - the results of tabulating a dataset
  - any set of summary statistics on a dataset
  - a series of relational tables
- Many of the datasets we use have been artificially reshaped in order to fulfill this criterion of rectangularity

Describing Data

- The first column is usually an identifier or index column, where each unit/row is given a unique name or ID

[1] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet. The R Software: Fundamentals of Programming and Statistical Analysis. Springer, New York, 2014.

Describing Data

- The first column is usually an identifier or index column, where each unit/row is given a unique name or ID
- Certain columns (features) can correspond to the design of the experiment, specifying, for example, to which experimental group the unit belongs

[1] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet. The R Software: Fundamentals of Programming and Statistical Analysis. Springer, New York, 2014.

Describing Data

- The first column is usually an identifier or index column, where each unit/row is given a unique name or ID
- Certain columns (features) can correspond to the design of the experiment, specifying, for example, to which experimental group the unit belongs
- Other columns represent the observed measurements of the experiment. Usually, these measurements exhibit variability; that is, they would change if the experiment were to be repeated

[1] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet. The R Software: Fundamentals of Programming and Statistical Analysis. Springer, New York, 2014.

Describing Data

- The first column is usually an identifier or index column, where each unit/row is given a unique name or ID
- Certain columns (features) can correspond to the design of the experiment, specifying, for example, to which experimental group the unit belongs
- Other columns represent the observed measurements of the experiment. Usually, these measurements exhibit variability; that is, they would change if the experiment were to be repeated

A well-known repository of data sets is the Machine Learning Repository maintained by the University of California at Irvine (UCI), found at HERE

[1] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet. The R Software: Fundamentals of Programming and Statistical Analysis. Springer, New York, 2014.

Describing Data

- The first column is usually an identifier or index column, where each unit/row is given a unique name or ID
- Certain columns (features) can correspond to the design of the experiment, specifying, for example, to which experimental group the unit belongs
- Other columns represent the observed measurements of the experiment. Usually, these measurements exhibit variability; that is, they would change if the experiment were to be repeated

A well-known repository of data sets is the Machine Learning Repository maintained by the University of California at Irvine (UCI), found at HERE

- Iris.data
- abalone.data
- nutrition_elderly.xls[1]

[1] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet. The R Software: Fundamentals of Programming and Statistical Analysis. Springer, New York, 2014.

Describing Data

Tidy data - Hadley Wickham's three rules for "tidy" datasets

- Each variable must have its own column

Describing Data

Tidy data - Hadley Wickham's three rules for "tidy" datasets

- Each variable must have its own column
- Each observation must have its own row

Describing Data

Tidy data - Hadley Wickham's three rules for "tidy" datasets

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell

Describing Data

Tidy data - Hadley Wickham's three rules for "tidy" datasets

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell

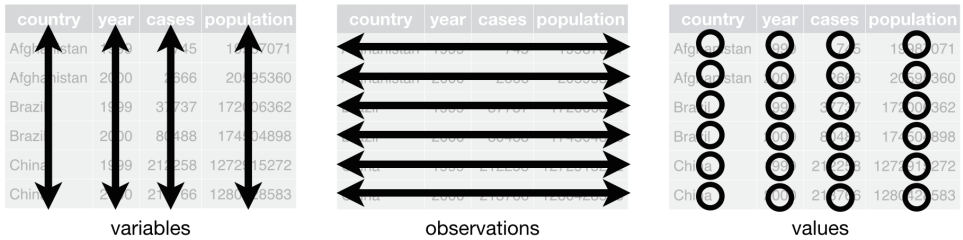
Read Tidy data

Describing Data

Tidy data - Hadley Wickham's three rules for "tidy" datasets

- ▷ Each variable must have its own column
- ▷ Each observation must have its own row
- ▷ Each value must have its own cell

Read Tidy data



Question Change the xls file into CSV file

Summary statistics

Let $x = [x_1, \ldots, x_n]^T$ be a column vector with n numbers.

Summary statistics

Let $x = [x_1, \ldots, x_n]^T$ be a column vector with n numbers.

- ▷ *sample mean:*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Summary statistics

Let $x = [x_1, \ldots, x_n]^T$ be a column vector with n numbers.

- ▷ *sample mean:*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- ▷ *sample quantile:* The p-sample quantile $(0 < p < 1)$ of x is a value x_p such that at least a fraction p of the data is less than or equal to x_p and at least a fraction $1 - p$ of the data is greater than or equal to x_p

Summary statistics

Let $x = [x_1, \ldots, x_n]^T$ be a column vector with n numbers.

- *sample mean:*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- *sample quantile:* The p-sample quantile $(0 < p < 1)$ of x is a value x_p such that at least a fraction p of the data is less than or equal to x_p and at least a fraction $1 - p$ of the data is greater than or equal to x_p
- The *sample median* is the sample 0.5-quantile

Summary statistics

Let $x = [x_1, \ldots, x_n]^T$ be a column vector with n numbers.

- *sample mean:*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- *sample quantile:* The p-sample quantile $(0 < p < 1)$ of x is a value x_p such that at least a fraction p of the data is less than or equal to x_p and at least a fraction $1 - p$ of the data is greater than or equal to x_p
- The *sample median* is the sample 0.5-quantile
- The p-sample quantile is also called the $100 \times p$ percentile. The 25, 50, and 75 sample percentiles are called the *first, second, and third quartiles* of the data.

Summary statistics

dispersion (spread) of the data is measured by

- ▷ *sample range*

$$\max_i x_i - \min_i x_i$$

- ▷ *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Summary statistics

dispersion (spread) of the data is measured by

- *sample range*

$$\max_i x_i - \min_i x_i$$

- *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- *sample standard deviation* $s = \sqrt{s^2}$

Summary statistics

dispersion (spread) of the data is measured by

- ▷ *sample range*

$$\max_i x_i - \min_i x_i$$

- ▷ *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- ▷ *sample standard deviation* $s = \sqrt{s^2}$

Question Calculate the above measures for a feature of the nutrition data

Visualizing data

- Plotting Qualitative Variables - Barplot
- Plotting Quantitative Variables -
  (a) Boxplot - a graphical representation of the five-number summary of the data consisting of the minimum, maximum, and the first, second, and third quartiles
  (b) Histogram - a graphical representation of the distribution of a quantitative feature
  (c) Empirical Cumulative Distribution Function - denoted by F_n, is a step function which jumps an amount k/n at observation values, where k is the number of tied observations at that value. For observations $x_1, \ldots, x_n$,

$$F_n(x) = \frac{\text{number of } x_i \leq x}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\}$$

where $\mathbb{1}$ denotes the indicator function