What are the computational and data sciences?

Data overview

Definitions

Variable

A quantity, quality, or property that you can measure.

Definitions

Variable

A quantity, quality, or property that you can measure.

Value

The state of a variable when you measure it. The value of a variable may change from measurement to measurement.

Definitions

Variable

A quantity, quality, or property that you can measure.

Value

The state of a variable when you measure it. The value of a variable may change from measurement to measurement.

Observation

A set of measurements made under similar conditions (you usually make all of the measurements in an observation at the same time and on the same object). An observation contains several values, each associated with a different variable.

Definitions

Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other

Definitions

Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other

$$\text{explanatory variable} \xrightarrow{\text{might affect}} \text{response variable}$$

Definitions

Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other

$$\text{explanatory variable} \xrightarrow{\text{might affect}} \text{response variable}$$

Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

Definitions

Tabular data (rectangular data)

A set of values, each associated with a variable and an observation.

Definitions

Tabular data (rectangular data)

A set of values, each associated with a variable and an observation.

Stu.	**sex**	**sleep**	$\cdots$	**dread**
1	male	5	$\cdots$	3
2	female	7	$\cdots$	2
3	female	5.5	$\cdots$	4
4	female	7	$\cdots$	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
21	male	6	$\cdots$	3

Data collected on students in a data science class on a variety of variables

Kinds of data

Numerical

Data that is a number, either an *integer* (whole numbers) or a *float* (real numbers). This kind of data is collected from device sensors, through counting and polling, outputs of computational simulations, etc.

Kinds of data

Numerical

Data that is a number, either an *integer* (whole numbers) or a *float* (real numbers). This kind of data is collected from device sensors, through counting and polling, outputs of computational simulations, etc.

Categorical

Groups observations into a set. Categories can be in text form (*strings* or *characters*), for example brand names for a certain kind of product, or numerical, for example labeling city districts by numbers.

Kinds of data

Numerical

Data that is a number, either an *integer* (whole numbers) or a *float* (real numbers). This kind of data is collected from device sensors, through counting and polling, outputs of computational simulations, etc.

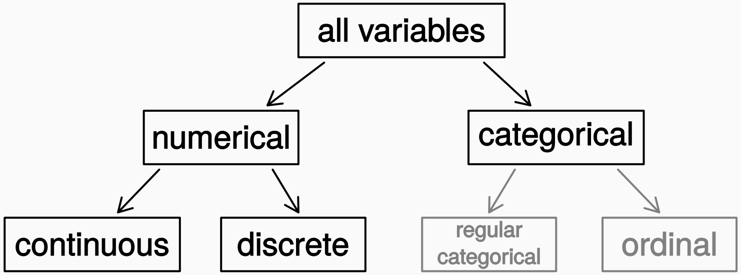
Categorical

Groups observations into a set. Categories can be in text form (*strings* or *characters*), for example brand names for a certain kind of product, or numerical, for example labeling city districts by numbers.

Textual

Plain text that is too varied to be treated as a category. Some examples can be full names, the text of a literary work, tweets, etc.

Kinds of data



Example: types of variables

Stu.	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

Example: types of variables

Stu.	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical

Example: types of variables

Stu.	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous

Example: types of variables

Stu.	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous
- *bedtime:* categorical, ordinal

Example: types of variables

Stu.	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous
- *bedtime:* categorical, ordinal
- *countries:* numerical, discrete

Example: types of variables

Stu.	sex	sleep	bedtime	countries	dread
1	male	5	12 – 2	13	3
2	female	7	10 – 12	7	2
3	female	5.5	12 – 2	1	4
4	female	7	12 – 2		2
5	female	3	12 – 2	1	3
6	female	3	12 – 2	9	4

- *sex:* categorical
- *sleep:* numerical, continuous
- *bedtime:* categorical, ordinal
- *countries:* numerical, discrete
- *dread:* categorical, ordinal (or numerical)

Modes of data collection

There are two main modes of data collection that affect the strength of a researcher's conclusions.

Modes of data collection

There are two main modes of data collection that affect the strength of a researcher's conclusions.

- **Observational/field study**: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe".

Modes of data collection

There are two main modes of data collection that affect the strength of a researcher's conclusions.

- **Observational/field study**: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe".
- **Experiment**: Researchers systematically control variables in order to establish causal connections

Modes of data collection

There are two main modes of data collection that affect the strength of a researcher's conclusions.

- **Observational/field study**: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe".
- **Experiment**: Researchers systematically control variables in order to establish causal connections
    - Careful tuning of one parameter of an experimental appartus, changing a single chemical component, altering one nutrient in an organism's diet, etc.

Modes of data collection

There are two main modes of data collection that affect the strength of a researcher's conclusions.

- **Observational/field study**: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe".
- **Experiment**: Researchers systematically control variables in order to establish causal connections
    - Careful tuning of one parameter of an experimental appartus, changing a single chemical component, altering one nutrient in an organism's diet, etc.
    - Blind studies: randomly assign subjects to treatments. Becomes double blind if experimental observers are also randomly assigned.

How do we obtain data?

Manual measurements

- Compared to a baseline: ruler, scale, stopwatch
- Record-keeping: counting, behaviorial notes, ledgers, timelines, relationships
- Self-reporting: surveys and interviews

How do we obtain data?

Manual measurements

- Compared to a baseline: ruler, scale, stopwatch
- Record-keeping: counting, behaviorial notes, ledgers, timelines, relationships
- Self-reporting: surveys and interviews

Sensor measurements

- Electrical, temperature, mechanical, chemical, electromagnetic, navigation, cameras/light, pressure, etc.
- A lot of these are in a cell phone!
- Benefits: automation, precision, access to properties that manual methods cannot measure

How do we obtain data?

Digital artifacts

- Internet: server logs, social network activity, web search, online transactions, data transmissions, etc.
- Digital text corpus: digital books, articles, government documents, email, messaging, etc.
- Databases: scientific, social, government, business, etc.

Credits

License

Acknowledgments

Content adapted from:

- The chapter 1 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license
- Chapter 2 from *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton
- The Lecture 7 - Sensors and Scientific Measurements by John Wallin