Week 2

Variables

Michael W. Kearney, PhD

School of Journalism
Informatics Institute
University of Missouri

2018-01-25

Outline

Recap

Sampling

Variables

Independent and dependent variables

Hypotheses/research questions

Research designs

Scientific method

- Try to be “objective”
- Theories should be falsifiable
- Research should be reproducible
- Knowledge is cumulative and provisional

Outline

Recap

Sampling

Variables

Independent and dependent variables

Hypotheses/research questions

Research designs

Random sample

- Technically, it should be a “uniform” random sample
- Everyone in a sampling frame (all possible data points) has an equal probability of being selected
- Best method for making inferences but expensive and difficult to do

Quasi-random sample

- Samples that use mathematical rules but lack access to the full population
- One common technique is **probability matching**, which is when you match the demographics in a sample with the desired population

Snowball sample

- When you use one to recruit more into the sample
- Not very representative, but useful for getting access to niche or hard-to-reach groups

Convenience sample

- A sample that is not random but selected largely due to convenience (low cost, ease of access, etc.)

Why care about sampling?

- The sampling method and study design has a direct effect on one's ability to make **inferences** from data
- **Inferences** are conclusions drawn from a sample and applied to a population

Outline

Recap

Sampling

Variables

Independent and dependent variables

Hypotheses/research questions

Research designs

Variable

- A **constant** is a fixed value that never changes.
    - e.g., pi, the number 1, etc.
- A **variable** is a value that differs across observations.
    - can often be thought of as features or characteristics

Variable values

- **Values** are measurements (observations) on a given variable
    - e.g., Tracy's height (variable) is `6'6''` (value)
    - e.g., Avery's final race position (variable) is `1` (value)
    - e.g., Cory's skill level in chess (variable) is `master` (value)
    - e.g., Rory's hometown (variable) is `Kansas City` (value)
- Different levels of measurement enable different levels of analysis

Levels of measurement

- **Nominal**: values represent different categories [or named things]
- **Ordinal**: values represent meaningful sequence
- **Interval**: values represent meaningful sequence using equi-distant intervals
- **Ratio**: values represent real numbers

Nominal

Values represent different categories [or named things]

- Can be used to operationalize anything
- This is often done using *dummy codes*

Ordinal

Values represent meaningful sequence

- The order people finish in a race
- The distance from 1st to 2nd can vary wildly with the distance from 2nd to 3rd

Interval

Values represent meaningful sequence using equi-distant intervals

- Likert-type items, e.g., *I always make my bed in the morning: Strongly Agree… Strongly Disagree* and other survey items that measure a range of feelings/attitudes using numbers
- This is why the visual representation of numbers on a survey is often important

Ratio

Values represent real numbers

- Numbers correspond to some non-arbitrary meaning.
- A true 0 (zero) exists

Temperature

Do the statements below add up?

- The temperature today is 100 degrees Fahrenheit
- The temperature yesterday was 50 degrees Fahrenheit
- Today is twice as hot as yesterday

Celsius

Do the statements below add up?

- The temperature today is 38 degrees Celsius
- The temperature yesterday was 10 degrees Celsius
- Today is twice as hot as yesterday

Kelvin

Do the statements below add up?

- The temperature today is 311 degrees Celsius
- The temperature yesterday was 283 degrees Kelvin
- Today is twice as hot as yesterday

Actual temperature ratios

Fahrenheit

```
100/50
## [1] 2
```

Celcius

```
38/10
## [1] 3.8
```

Kelvin

```
311/283
## [1] 1.09894
```

Categorical/character

- I will tend to refer to nominal variables as categorical variables
- Categorical variables are represented in R as `character` and `factor` variables

```
x <- c('a', 'b', 'c')
class(x)
## [1] "character"

x <- as.factor(x)
class(x)
## [1] "factor"
```

Integer

- In R, we will refer to **discrete** numbers as `integer` values.

```
## R assumes numbers are continuous (numeric)
x <- c(1, 2, 3)
class(x)
## [1] "numeric"

## use "L" after whole numbers to indicate integer
x <- c(1L, 2L, 3L)
class(x)
## [1] "integer"
```

Numeric

- In R, we will refer to **continuous** interval and ratio variables as `numeric` values.
- Values are continuous if they don't only exist as discrete units

```
x <- c(1.25, -3.5, 4)
class(x)
## [1] "numeric"
```

Defining variables

Conceptual definition

- A description of a variable's **theoretical** meaning.

Operational definition

- A description of a variable's **observable** meaning.

Example

- Let's walk through an example of defining variable(s) and thinking about it in more practical terms

Class size

Define the construct *class size*

- **Class size** refers to the…
  - Physical dimensions of the classroom
  - The number of students
  - The mass of the students

Comparing classes sizes

Let's say we want to compare two observed class sizes

- **CLASS 1**: 30' x 30' room with 5 students weighing 1500 total lbs
- **CLASS 2**: 20' x 20' room with 10 students weighing 1250 total lbs

Defining “class size”

- **CLASS 1**: 30’ x 30’ room with 5 students weighing 1500 total lbs
- **CLASS 2**: 20’ x 20’ room with 10 students weighing 1250 total lbs
- If we define class size **physical space**, class 1 is biggest
- If we define class size as **number of students**, class 2 is biggest
- If we define class size as **mass of students**, class 1 is biggest

Is any definition right or wrong?

Practice

- Let’s practice conceptually and operationally defining variables

Media exposure

- Conceptual definition:
- Operational definition:

Views on climate change

- Conceptual definition:
- Operational definition:

Coverage of environmental issues

- Conceptual definition:
- Operational definition:

Accuracy in reporting

- Conceptual definition:
- Operational definition:

Fake news

- Conceptual definition:
- Operational definition:

Social media use

- Conceptual definition:
- Operational definition:

Political ideology

- Conceptual definition:
- Operational definition:

Perceived role of journalists

- Conceptual definition:
- Operational definition:

Outline

Recap

Sampling

Variables

Independent and dependent variables

Hypotheses/research questions

Research designs

Classifying variables

- Variables can be classified according to their status in a theoretical and/or statistical model
- There are several different labels applied toward this end, but for the most part they all do the same thing
    - Variables that represent the *cause*
    - Variables that represent the *effect*

Terminology

- The cause is the **independent** variable
- The effect is the **dependent** variable
- There's also exogeneous and endogenous, predictors and outcomes, and probably some other synonyms

Cause isn’t always cause

- Although we use the cause/effect heuristic when modeling the relationship between two variables, we usually acknowledge multiple, reinforcing, or some underlying “root” causes
- As long as we all know that, the cause/effect heuristic is still helpful for organizing our theoretical and statistical models
    - Bad reporting of scientific research happens when we take cause/effect terminology too literally

Practice

- Let’s practice identifying independent and dependent variables

Independent or dependent?

- Sun
- Surface temperature

Independent or dependent?

- Reading
- Knowledge

Independent or dependent?

- Public perceptions of an event
- Reporting information about an event

Independent or dependent?

- Breakfast
- Health

Independent or dependent?

- Exercise
- Health

Independent or dependent?

- Social media
- Face-to-face interaction

Outline

Recap

Sampling

Variables

Independent and dependent variables

Hypotheses/research questions

Research designs

Relationships between variables

- Research questions ask about the relationship between two or more variables
    - These questions can guess direction (e.g., positive vs negative) but I think that kind of defeats the purpose
- Hypotheses predict the relationship between two or more variables
    - Hypotheses can either predict **whether** there is a relationship or what **direction** the relationship is in

Hypotheses allow for more definitive inferences

RQs/Hs

- Use **research questions** when there isn't a lot of research in an area and the theory isn't clear about what exactly to expect
- Otherwise, use **hypotheses**

Outline

Recap

Sampling

Variables

Independent and dependent variables

Hypotheses/research questions

Research designs

Overview of study designs

- Surveys
- Experiments
- Content/text analysis

Next week

- Survey research with Dr. Ben Warner from Communication Department
- Make sure you've completed your IRB training

Table of Contents

Recap

Sampling

Variables

Independent and dependent variables

Hypotheses/research questions

Research designs