

Key concepts for working with data

Author: Nicholas G Reich

Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US

Foundational concepts for data

Unit of analysis

- observational level at which data is collected and/or analyzed.

Sampling Frame

- “a list of the items or people forming a population from which a sample is taken” (Oxford Languages via Google).

Tidy data

- a set of principles about how to store data in a standardized way, for easier interoperability across data wrangling, visualization, and modeling tools (hence, the ‘tidyverse’!)

Unit of analysis or observation

Conceptually: the smallest unit at which data is collected and able to be analyzed. Note: this may be determined by design or convenience (what data happens to be available).

In practice: The unit that defines a row of your data frame.

For example: If you have one observation for every state and every week, those together define your “unit of observation”, or what makes every row in your dataset unique. You might call it a “state-week”.

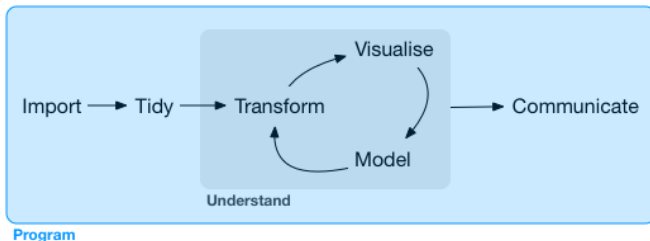
Sampling frames and populations

Three features dictate what conclusions can be drawn.

- **Population:** “set of all possible units which might have been included” (Daniel Kaplan)
- **Sampling frame:** “a list of the items or people forming a population from which a sample is taken” (Oxford/Google)
- **Sample:** “a selection of cases from the population” (Daniel Kaplan)

For the Sad Songs analysis:

Tidy data is a promise for downstream tools



Central principles

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table. (For more complicated analyses that may have several different “units of analysis”).

Breakout rooms

Revisit “Sad songs” and SAT analyses in the context of these concepts.