Digital Natives: Data Literacy

March 14th 2018
Cynthia Kremer
Science Outreach Librarian
ckremer@ben.edu
Statistics 101

- Population/Universe
- Case
- Sample
- Dependent Variable vs. Independent Variable
- Column Variable vs. Row Variable
- Control Variable
- Frequency
- Median
- Top coded
- Recode
- Longitudinal vs. Cross sectional

(Bordelon, 2017)
Interpreting Documentation – Key Terms

• **Description**: Summary or abstract of the dataset; often includes investigator(s), subject terms, geographic coverage, time period, funding bodies and any access restrictions. Best place to get a good overview of the dataset.

• **Codebook**: Explains the data file structure and layout. Provides information about how responses are coded, the column width/location, questions and any skip patterns used during interviewing.
Interpreting Documentation – Key Terms

• **Universe**: Information about the coverage of the survey. For public opinion polls, the *universe* may include the age, location, and means of access (ex: telephone or in-person interview) to respondents.

• **Sample Size**: Number of units surveyed for the study.

• **Unit of Analysis**: Family, Individual. Business, State, Zip code?

(Bordelon, 2017)
Interpreting Documentation – Key Terms

• **dependent variable**: question that is influenced by some external factor; the effect of the independent variable. *For example*, if you are trying to determine what influences an individual's attendance at performing arts events, then attendance at performing arts events would be a dependent variable.

• **independent variable**: causal factor, often demographic; responses to this variable may influence the responses to other variables. *For example*, if a researcher hypothesizes that income affects the likelihood of purchasing a luxury automobile, then income would be an independent variable.

(Bordelon, 2017)
Interpreting Documentation – Key Terms

• **column variable**: the question in the analysis that is considered to be an independent variable. Responses will appear as columns in the final results table.

• **row variable**: the primary question of interest in the analysis; the dependent variable. Responses will appear as rows in the final results table.

• **control variable**: question added to an analysis in order to determine whether the observed relationship between the independent and dependent variables might be affected by an additional influencing factor

(Bordelon, 2017)
Interpreting Documentation – Key Terms

• **frequency (or frequency distribution)**: table of results showing how many responses were given in each category for a specific variable
• **median**: middle-most score; a measure of central tendency that can be computed for ordinal or interval level variables
• **recode**: to create new boundaries for the categories of a variable being analyzed
• **top coded**: values above a certain level are not provided above an upper limit to protect anonymity (very common for income)

(Bordelon, 2017)
Interpreting Documentation – Key Terms

• **cross-sectional study:** observational study that involves the analysis of data collected from a population, or a representative subset, at one specific point in time. Examples: National Health Interview Survey; General Social Survey; Survey of Consumer Finance

• **longitudinal study:** observational research method in which data is gathered for the same subjects repeatedly over a period of time. Sometimes called a panel study. Examples: National Longitudinal Survey of Youth 1997; Panel Study of Income Dynamics; National Education Longitudinal Study 1988

(Bordelon, 2017)
Interpreting Documentation – Key Terms

• **Methodology**: Describes the theory & methods incorporated into the development or the “how” of the survey. May include information on sampling, weighting, how the data were collected, etc.

• **Version history**: Provides information about any previous editions of the survey. Very important if researchers are trying to replicate results. Version histories are often updated when data are added/removed from a survey or there are changes to the data documentation (ex: methodology).

(Bordelon, 2017)
Handout: Interpreting documentation

- Using the Survey of Public Participation in the Arts 1982-2012 Combined File as an example
- [http://www.icpsr.umich.edu/icpsrweb/NADAC/studies/35596](http://www.icpsr.umich.edu/icpsrweb/NADAC/studies/35596)
- **Documentation Exercise**
  - What is the scope of the study?
  - Longitudinal?
  - Are there any missing years?
  - What is the sample size?
  - What is the lowest level of geography? Are estimates provided for all places at that level of geography?
  - Is it a stand alone survey?
  - Who is/is not included?
    - Can one segment by: Age?
    - Marital Status?
    - Income? Actual, range, top coded?
    - Gender?
    - Race?
    - Citizenship status?

- What is a variable/question you found you did not expect to find?

(Bordelon, 2017)
Summary Statistics

Using SDA to create tables – Survey of Public Participation in the Arts 1982-2012 Combined File

1. [http://www.icpsr.umich.edu/icpsrweb/NADAC/studies/35596](http://www.icpsr.umich.edu/icpsrweb/NADAC/studies/35596)
2. Choose SDA Online Analysis

(Bordelon, 2017)
Summary Statistics

• We are looking for reading by income level for 2012.
• Under analysis choose “Run Frequency or Crosstabulations”

(Bordelon, 2017)
Summary Statistics

• Under Demographics, choose INCOME and click on Col.
• Go to Reading and choose BOOKS which represents read any books in the last 12 months. Click on Row.
• Under Wave, choose YEAR and click on CTRL.
• Click on Filter.
• In the Filter box, follow YEAR with 2012.
• Click on Run the Table

(Bordelon, 2017)
Summary Statistics

SDA Frequencies/Crosstabulation Program
Help: General / Recoding Variables

REQUIRED Variable names to specify
Row: BOOKS

OPTIONAL Variable names to specify
Column: INCOME
Control: YEAR
Selection Filter(s): YEAR(2012) Example: age(18-50)
Weight: WEIGHT - SPPA CROSS SECTIONAL WEIGHT TO PRODUCE ESTIMATES FOR O

(Bordelon, 2017)
### Summary Statistics

**Survey of Public Participation in the Arts 1982-2012 Combined File [United States]**

Dec 19, 2016 (Mon 09:25 AM EST)

#### Statistics for YEAR = 2012

<table>
<thead>
<tr>
<th>BOOKS</th>
<th>INCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1: YES</strong></td>
<td></td>
</tr>
<tr>
<td>2,973,301</td>
<td>6,332,009</td>
</tr>
<tr>
<td>1,915,963</td>
<td>4,402,944</td>
</tr>
<tr>
<td>1,933,564</td>
<td>4,724,949</td>
</tr>
<tr>
<td>2,649,923</td>
<td>7,046,108</td>
</tr>
<tr>
<td>2,743,114</td>
<td>6,748,891</td>
</tr>
<tr>
<td>4,864,834</td>
<td>10,496,133</td>
</tr>
<tr>
<td>5,623,116</td>
<td>13,229,480</td>
</tr>
<tr>
<td>6,336,256</td>
<td>14,300,477</td>
</tr>
<tr>
<td>7,213,590</td>
<td>14,639,324</td>
</tr>
<tr>
<td>8,029,219</td>
<td>12,080,730</td>
</tr>
<tr>
<td>10,011,917</td>
<td>20,249,892</td>
</tr>
<tr>
<td>11,019,407</td>
<td>20,067,444</td>
</tr>
<tr>
<td>12,033,694</td>
<td>21,674,002</td>
</tr>
<tr>
<td>13,637,522</td>
<td>20,213,940</td>
</tr>
</tbody>
</table>

| **2: NO** |             |
| 3,556,707 | 6,332,009 |
| 2,486,900 | 4,402,944 |
| 2,704,368 | 4,724,949 |
| 4,200,867 | 7,046,108 |
| 4,655,779 | 6,748,891 |
| 5,634,349 | 10,496,133 |
| 7,606,367 | 13,229,480 |
| 7,124,211 | 14,300,477 |
| 7,625,733 | 14,639,324 |
| 6,051,528 | 20,249,892 |
| 8,648,036 | 20,067,444 |
| 11,019,407 | 21,674,002 |
| 12,033,694 | 20,213,940 |

#### Color coding:
- Z: Smaller than expected
- X: Larger than expected

*(Bordelon, 2017)*
Data Visualization

Data visualization can:

• provide clear understanding of patterns in data
• detect hidden structures in data
• condense information explain complex data to an external audience
• R and Tableau have been popular among a wide range of users

(Womack, 2017)
Anscombe’s Quartet

- For example, see Anscombe’s quartet (image source: http://commons.wikimedia.org/wiki/File:Anscombe%27s_quartet_3.svg):
From Barchart to Dot Plot

- The Cleveland dot plot
- use to compare labeled quantities, ordered lists

(Womack, 2017)
Barchart vs Dot Plot

(Womack, 2017)
Visualizing Distributions of Data

Box and Whiskers Plot

• illustrate quantiles and outliers. There is also a Tufte version.

Violin plot

• Blends density information with box and whiskers style (in an artistic manner)

(Womack, 2017)
Box Plot v. Violin Plot

(Womack, 2017)
Visualizing Categorical Data

Beyond the pie chart

• The mosaic plot allows multiple categories to be displayed on the same graph, but can be complicated to interpret.

• The spineplot is a variant of the mosaic plot, plotting proportions in 2 dimensions.

(Womack, 2017)
Pie Chart v. Mosaic Plot

(Womack, 2017)
Maps and Glyphs

• Maps are obviously an important and widespread way of presenting data.
• Choropleth maps, in which shading indicates data levels
• See also Interactive Maps in R and 5 kinds of Interactive maps in Plot.ly for further exploration
• Glyphs present iconic representations of data elements.
• Weather maps often use glyphs.
• As an R example, consider Chernoff faces from the aplpack package.

(Womack, 2017)
Choropleth Map v. Chernoff Faces

(Womack, 2017)
The Face of Crime in the United States

Violent Crime
- HEIGHT OF FACE: Murder
- HEIGHT OF MOUTH: Aggravated Assault
- SHAPE OF FACE: Robbery
- WIDTH OF FACE: Forcible Rape

Property Crime
- WIDTH OF EYES: Motor Vehicle Theft
- HEIGHT OF EYES: Larceny-Theft
- WIDTH OF MOUTH: Burglary

United States
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Interactive DataViz - Principles

• Why aren’t all of our graphs interactive?
• Brushing is used to select data points and track them through various analyses.
• Drilling down, zooming, and subsetting are also interactive techniques.
• Data displays can be linked so that a selection in one panel modifies the output displayed in another panel.
• Interactivity is especially useful for data exploration, studying multidimensional relationships.

(Womack, 2017)
In many contexts, visualizing the relationships between data elements is made easier by viewing related data interactively.

Making this easy are googleVis and other “Vis” packages, e.g. `bdvis` for biodiversity or `rainfreq`.

A Library example - comparing selected ARL Statistics for publicCIC universities

Web services are making interactivity accessible to all -> Google Fusion Tables

(Womack, 2017)
Interactive Power

• Population pyramids are one example where
  
  *interactivity + animation = insight.*

• Populationpyramid.net - for all countries, basic animation

• The German Population Pyramid from Destatis is even more interactive

• Doing it in R is possible with these instructions (Part 1) and (Part 2)

(Womack, 2017)
References


Slides by R. Womack are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.