Statistics 467/567: How to Create a Data Set and Formulate Research Questions

July 6, 2020

Due: Monday, July 13, 2020 (can send link via e-mail or upload to Sakai Drop Box)

1. What is the topic of your data?

2. Are the data numeric - i.e. quantitative (numerical) or narrative (ethnographic) - qualitative? What recoding may be required to convert from string variable (i.e. name such as male/female) to numeric?

3. What research questions do you want to answer from your data? (Apply the same questions when writing the abstract.)

   ➢ Who
   ➢ What
   ➢ Where
   ➢ When
   ➢ Why
   ➢ How

4. Decide on sample size. For example, if US States, then use all 50. If large data set, such as population data, use a random number table to select subjects, or choose every tenth (or fifteenth) case to form a smaller data set (pandect).

5. Which variables will you use to test your hypotheses (three (3) total hypotheses)?

6. What is the unit of analysis? Define your universe.

   • Human Subjects or Cases (i.e. women --with breast cancer -- restricted to females only)
   • Cities (U.S. only or world)
   • U.S. States
   • Countries (all 7 continents or Europe only??)
   • Cars (class and size--comparison)
   • Year/Month (trend analyses or compare 2 different years) (note: month may have problem with autocorrelation).
   • Companies (Fortune 500 or Dow Jones)
   • Sports Teams (US NFL, MLB, NHL, FIFA, etc.)

7. What are the measures – variables-predictors - available in the data set to answer your question(s)?

Examine the code book (i.e. description of the measures in the data set. Define the variables by using univariate analyses to assess frequency, range (minimum and maximum), plus measures of central tendency and dispersion.

8. From where is the sample drawn? What is the data source? (country? state? county? Government website or school? University repository. Or hospital? Clinic? Corporate website, etc.).

9. When were the data collected? (what year?).

10. Why is this research important? (theoretical basis?). Why is this problem worth studying? (See existing literature.). Select one or two seminal papers as a basis for main hypothesis. Goal is to replicate and extend.

11. How are the phenomenon (the variables) related? How are the associations measured--via survey or samples from the field? What are the variables? Define?

12. Are the data primary (collected by you) or secondary (collected by someone else or organization)?

13. What type of methods will you apply to these data based on the structure of the variables?

Abstract n = 250 word abstract, even if in proposal form only, using this basic structure:

- ❖ Introduction
- ❖ Methods
- ❖ Results
- ❖ Discussion