Preparing Data for Analysis

MATH 456

1/25/2016

Types of variables

- **Continuous:** Things that can be measured
  - Height, weight, density
- **Discrete / Integer:** Can only hold whole number values
  - counts of things
- **Categorical:** Non-numeric characteristics
  - Gender, ethnicity
- **Binary:** Two categories only
  - Gender
- **Indicator:** A binary variable that is yes/no depending on a specified criteria
  - Typically holds values of 1/0 or TRUE/FALSE
  - Female, on medicare

Stevens's classification of variables

- **Nominal:** A.k.a categorical. Each observation belongs to one of several distinct categories. No inherent ordering.
  - Gender (M/F), Ethnicity (White/AfAm/Hispanic)
- **Ordinal:** Categorical variables with an inherent ordering.
  - First(1), second(2), third(3), Strongly agree(5), agree(4), neutral(3), disagree(2), strongly disagree(1)
- **Interval:** Differences between successive values are always the same.
  - Temperature, calendar dates
- **Ratio:** Interval variables with a natural zero point
  - Height, weight, density, time duration

Q: Why is Temperature not ratio?

Data Types found in computer data sets

- **Numerical:** 3.14159
- **Integer:** 0, 1, 2, 3, 4, 5
- **Character / String:** “Hello World”, “Blue”
- **Logical:** TRUE/FALSE

Describing relationships between variables

Multiple names for the same concept

- Response / Outcome / Dependent variable
- Covariate / Explanatory / Predictor / Independent variable

The direction of the relationship is situation dependent.

- We could use weight as an outcome variable with height, sex, age and diet as predictors
- We could use blood pressure as the outcome with weight, sex, age and diet as predictors

Practice - Vocabulary

1. Classify the following types of data by using Stevens's measurement system: decibels of noise level, father's occupation, parts per million of an impurity in water, density of a piece of bone, rating of a wine by one judge, net profit of a firm, and score on an aptitude test.
2. Pose two possible research questions from the CORD study (Lung function data) and decide on the appropriate dependent and independent variables.
3. Give an example of nominal, ordinal, interval, and ratio variables from a field of application you are familiar with.

Tidy Data structures

```
head(iris[c(1,2,3,5)])
```

```
##   Sepal.Length Sepal.Width Petal.Length Species
## 1          5.1         3.5          1.4  setosa
## 2          4.9         3.0          1.4  setosa
## 3          4.7         3.2          1.3  setosa
## 4          4.6         3.1          1.5  setosa
## 5          5.0         3.6          1.4  setosa
## 6          5.4         3.9          1.7  setosa
```

- One row per observation/case/individual
- One column per characteristic
- Tidy data / Rectangular data / Structured data