Lecture 1

D&P: 1.3

I. Things we do with data:

A. Description

B. Inference

D&P: 2.2, 2.7

II. Collecting Data via Surveys and Experiments:

A. Surveys: want to learn about the value of a variable associated with units in a population

1. Some examples …

2. Techniques for collecting data:
   a. Ask everyone: census, election
      i. Expensive and/or difficult
      ii. If you manage to do this, subsequent analysis does not require any techniques from inferential statistics.
   b. Ask a portion, and hope that results are similar to what you would get from whole group.
      i. Procedure is called a survey,
      ii. Resulting collection of info is called a sample,
      iii. If you select people systematically they may systematically be
Lecture 1

different from everyone else

3. Requirements for sample to be representative of population:
   a. Simplest kind of representativeness is called \textit{simple random sampling}
   b. Each person has the same chance of being asked
   c. People are selected independently
   d. cf., take separate simple random samples from people in the various groups and pool the results: \textit{stratified random sampling}:
      i. This way we are sure that we get reasonable numbers of people from each demographic group,
      ii. if we expect responses to vary with some demographic variable.

4. Pitfalls that leading to systematically wrong answers: \textit{biases}
   a. If people’s chances of being asked are tied to the answer they are likely to give, results will be systematically wrong: \textit{selection bias} or \textit{under-coverage}.
   b. If people’s chances of responding are tied to the answer they are likely to give, you are likely to get an answer that is systematically wrong: \textit{response bias}
B. Experiments:

1. Answer a question about a hypothetical population

2. For example, "How much longer generally do certain cardiac patients live if they get a bypass operation"?

3. Different from survey kind of question:
   a. We are interested not just in an existing group of patients, but also for all patients in the future.
   b. Population wouldn’t exist except for experimenter’s intervention.

4. An example:
   a. Pediatricians studying ways to avoid lead poisoning.
   b. Does cleaning behavior affect blood lead levels in infants.
   c. Study:
      i. Rochester, NY area, early 1995
      ii. All new-borns in high-risk parts of city were recruited
      iii. Half of parents were selected to receive instruction and supplies for cleaning; rest were left as is.
      iv. All head blood lead measured at beginning and end of study.
5. Experiments are *controlled*:

   a. We want to compare group getting treatment to those not getting treatment.

   b. We want those treated to be exactly like those not treated except for treatment

      i. (so, for ex., we wouldn’t use old data or data from other cities as basis for comparison).

   c. Progress should be compared to that of a group, undergoing the passage of time but not the treatment.

   d. The group receiving the experimental intervention is typically called the *treatment group*, and the other group is called the *control group*.

      i. *randomize*: Assign participants at random into treatment and control groups (using, for example, shuffled cards).

      ii. *extraneous factors* (variables that are not fundamentally related to the treatment but which could influence the outcome) are usually close to evenly distributed.

        • rental status
        • whether renovation had been done
Lecture 1

iii. Randomizing within groups defined by extraneous factors is called **blocking**.

D&P: 2.5

6. Experiments should be **blinded** if possible:

a. Since a patient’s status on treatment or control might influence their response, while on study they shouldn’t know which group they are in: **single-blinded**.

i. In order to avoid having some patients infer from not being given medication that they are assigned to a control group, control patients are given what appears to be a study medication, but is actually inactive: **placebo**

b. Since a patient’s status on treatment or control might influence the assessment of their response, while on study the person assessing response shouldn’t know which group they are in: This plus single-blinding makes them **double-blinded**.

D&P: 1.4A

III. Taxonomy of Data:

A. Classify each **variable**

1. Definition:
Lecture 1

a. a single quantity, perhaps out of many, assessed in a study.

b. measured on a number of different individuals

B. Categorizing the variable:

1. Qualitative or Categorical:

   a. Data represents a classification rather than a count or measurement.

b. Examples:

   i. Race
   
   ii. Gender
   
   iii. Treatment group

   c. Ordinal: A qualitative variable whose categories have a natural ordering;

   d. Numerical or Quantitative: Data represents a count or measurement.

   i. Often too many values to usefully represent in a table.

   ii. Classified according to how many potential values there are:
      
      • Discrete if we can make a list of the possible values before seeing the data:
      
      ▶ eg., number of colds in a year, or weight rounded to nearest
Lecture 2

pound.

- *Continuous* otherwise: eg, exact weight.