Where do data come from?

- Observational studies
  - Interpretation of results subject to effects of **confounding**
  - **Confounding variable** is one that affects the response variable and is related to the explanatory variable
  - Can often establish an association, but generally can’t conclude cause and effect

- Experiments
  - Observe effects on subjects after the application of some **treatment**
  - Might want to compare a treatment versus a control or multiple treatments
  - Key elements in experimental design are
    - Control for effects of variables
    - Use replication
    - Use randomization
**Where do data come from?**

- Controlling for effects of variables
  - blinding
  - placebo effect
  - double dummy
  - blocking
- Replication and sample size
  - need sufficiently large enough samples to be able to distinguish between a true effect and natural variability
  - experimental results should be reproducible

**Where do data come from?**

- Using randomization
  Expect all components of the population to be approximately proportionately represented
  - Random sample: each individual has an equal chance of being selected
  - Simple random sample: each sample of the same size n has the same chance of being chosen
  - Probability sample: each member has a known chance of being selected

**Where do data come from?**

- Other sampling techniques
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling
  - Convenience sampling
- Good design combines elements
  - Randomized block design

**Where do data come from?**

- Surveys
  - a type of observational study
  - phone, mail, email, web-based, in person
  - some additional issues
    - Wording of questions can introduce bias (deliberate or unintentional)
      - “Do you agree…?”
    - Ordering of questions (planting ideas)
    - Convenience samples/Self-selected samples
    - Desire of respondents to please
    - Confidentiality concerns may influence responses
    - Non-response bias
**Where do data come from?**

Fundamental Rule:
Data must be representative of the population with regards to the question(s) of interest
-- regardless of how data were collected
-- randomization important
  Helps to ensure that all components of the population will be approximately proportionately represented

**Sampling error**
Difference between the sample result and the true population result due to chance sample fluctuations.

**Nonsampling error**
Errors due to sample data that are incorrectly collected, recorded, or analyzed

**Things to be aware of**

• Already mentioned
  -- Sample sizes
  -- Loaded questions
  -- Order of questions
  -- Nonresponse
  -- Association versus causation

• Some other potential issues
  -- Graphs used to exaggerate or understate (scaling of axes)
  -- Pictographs
  -- Percentages (misleading or unclear)
  -- Missing data
    • At random meaning unrelated to values
    • Special reasons
  -- Self-interest studies
  -- Precise numbers ➔ accuracy