22S:30/105 Statistical Methods and Computing

Introduction to Types of Studies

Lecture 7
February 15, 2008

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

Experiments and observational studies

- In an *experiment*, the investigator studies the effect of varying some factor that he/she controls.
- In an *observational study*, the investigator merely observes and records information on the subjects but does not manipulate any factors.
- It is very difficult to establish *causation* between one variable and another.
  - especially difficult based on observational studies

Koch's postulates

- In 1890 the German microbiologist Robert Koch attempted to develop criteria for establishing whether a particular microorganism *causes* a particular disease
- not considered completely satisfactory today
- "... first, the organism is always found with the disease, in accord with the lesions and clinical stage observed; second, the organism is not found with any other disease; third, the organism, isolated from one who has the disease and cultured through several generations, reproduces the disease in a susceptible experimental animal. Even where an infectious disease cannot be transmitted to animals, the 'regular' and 'exclusive' presence of the organism proves a causal relationship."

More formal criteria for judging whether an observed association is causal

- strength of the association
- dose-response relationship
- consistency of the association
  - Is the association observed in one study observed in other study populations, in studies using different methods, etc.
- temporally correct association
- specificity of the association
  - the alleged effect is rarely if ever observed without the alleged cause
- plausibility

Example: Female literacy and infant mortality

Confounding

Two variables (explanatory or lurking) are **confounded** when their effects on a response variable cannot be separated.

Association does not by itself imply causation.

Populations and samples

- A **population** is the *entire set* of items about which we might wish to draw conclusions.
  - Example: I wish to find out the average income of families of current UI undergrads.
  - Example: A political pollster would like to know the Presidential preference of every registered voter in South Carolina.
  - Some populations we would like to study are hypothetical.
    - Example: all pregnant women who are infected with the HIV virus now and in the future
- A **sample** is the subset of the population that we can actually study (on which we can measure values of variables).

- How a sample is drawn from a population affects how valid it is to apply conclusions based on the sample to the population.
- The **sample design** is the method used to choose the sample from the population.

Bias

- The results of a study are **biased** if they are subject to systematic error.
  - i.e., there is something about the way the study is carried out such that, if we did many studies in this way, on average we'd get the wrong conclusions!
- One source of bias is if the sample is not *representative* of the entire population.
- The design of a study is **biased** if it systematically favors certain outcomes.

Kinds of sample designs

- simple random sample (SRS)
  - a sample of size n individuals chosen in such a way that every set of n individuals in the population has an equal chance to be the sample
  - the ideal
  - biased or unbiased?
- voluntary response sample
  - consists of people who choose themselves by responding to a general appeal
  - biased or unbiased?
- convenience sample
  - consists of subjects who are easy to get
  - biased or unbiased?

- judgment sample
  - consists of subjects chosen by an expert to be representative of the population
  - biased or unbiased?

How simple random samples are drawn

- each member of the population is uniquely identified in some way
  - example: the population of interest is UI students; each has a unique ID number
- intuitive idea: the identifiers are put in a hat and drawn at random
- usually actually done by a computer
- can be done manually using a table of random digits
  - first assign a unique numeric label to each member of the population
  - use table of digits to select labels at random.