22S:30/105
Statistical Methods and Computing

Introduction to Types of Studies

Lecture 7
February 10, 2006

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

Association does not by itself imply causation.

Confounding

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

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 indivduals 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.

Example

- I wish to get an idea as to how well undergrad students in 22S:30 like the textbook. To do this, I want to administer a lengthy interview and I have time to do only 3. Therefore, I want to draw a simple random sample of size 3 from the population of 24 undergrad students in the class.

- Begin by giving each student a unique numeric identifier.

1. Derek A
2. Kara
3. Courtney
4. Karen
5. Cory
6. Catherine
7. Katie H
8. Ryan
9. Jenna
10. Peter
11. Anne
12. Todd
13. Anthony
14. Katie McE
15. Kimbra
16. Phil

17. Derek N
18. Tuyet
19. Ben
20. Mitchell
21. Nicole
22. Cristina
23. Joanna
24. Jessica

- Use Table B in your book to find the first 3 of these identifiers that appear.

Table of random digits

- Each entry in the table is equally likely to be any of the 10 digits from 0 to 9 inclusive.
- The entries are "independent" of each other; i.e., knowledge of what digits are in one part of the table gives no information about the digits in any other part.

Using SAS to draw a simple random sample

```
options linesize = 79 ;

data students ;
input name $9. ;
datalines ;
Derek A
Kara
Courtney
Karen
Cory
Catherine
Katie H
Ryan
Jenna
Peter
Anne
Todd
Anthony
Katie McE
Kimbra
Phil
Derek N
```



```
Tuyet
Ben
Mitchell
Nicole
Cristina
Joanna
Jessica
;

proc print  data = students ;
run ;
```


Output

```
                 Obs    Name

                   1    Derek A
                   2    Kara
                   3    Courtney
                   4    Karen
                   5    Cory
                   6    Catherine
                   7    Katie H
                   8    Ryan
                   9    Jenna
                  10    Peter
                  11    Anne
                  12    Todd
                  13    Anthony
                  14    Katie McE
                  15    Kimbra
                  16    Phil
                  17    Derek N
                  18    Tuyet
                  19    Ben
                  20    Mitchell
                  21    Nicole
```



```
                  22    Cristina
                  23    Joanna
                  24    Jessica
```

Proc plan

```
proc plan seed = 72950 ;
factors a = 3 of 24 ;
run ;
```

```
                 The PLAN Procedure

     Factor       Select       Levels      Order

     a                 3           24     Random


                      ----a---

                       1 24  7
```


Using the same seed will reproduce exactly the same "random" choice!

```
proc plan seed = 72950 ;
factors a = 3 of 24 ;
run ;
```

```
                 The PLAN Procedure

     Factor       Select       Levels      Order

     a                 3           24     Random


                      ----a---

                       1 24  7
```


Using a different seed will produce a different set of choices.

```
proc plan seed = 32542 ;
factors a = 3 of 24 ;
run ;
```

```
Procedure PLAN


     Factor       Select       Levels      Order

     a                 3           24     Random


                      ----a---

                       2 16  4
```


Drawing from a larger population

```
proc plan seed = 241 ;
factors a = 100 of 1000 ;
run ;
```

```
Procedure PLAN

Factor    Select  Levels   Order
------    ------  ------  -------
a            100    1000   Random

      a
----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+

  576  792  359  517  110  598  859  144    9   52  462  262  673  202  648
  630  705  286  412  597  868  488  621  240  674  651  923  298  419  865
  550  120  441  921  139  644  269  861  775  529  168  939   50  281   57
  119  944  692  265  432  470  311  585   69  329  143  562  974  996  904
  901  767  507  819  844  518  264  822  897  271  820  239  435  341  442
  497  773  687  449   41  424   24  326  863  178  752  423  233  834  358
  864  481  362  584   28  479  594  235  337  175
```

Other statistical sampling designs

- Statistical sampling is based on *chance*.
- A **probability sample** gives each member of the population of interest a *known* chance of being selected.
- **stratified random sampling**
  - procedure
    - ∗ first divide the population into *strata* – groups of similar individuals
    - ∗ draw a simple random sample from each stratum
    - ∗ combine the SRSs to form the full sample
  - ensures that each stratum is represented in the overall sample

  - Example: survey of class opinions on the textbook
    - ∗ I might divide the class into men and women and take a SRS within each gender
- Probability sampling methods other than SRSs require more complicated statistical analysis than do SRSs.
  - But meaningful results **can** be obtained because we know what population was actually sampled and exactly how it was done.
  - This contrasts with voluntary response samples, convenience samples, and judgment samples.

Other possible sources of bias in surveys

- Undercoverage
  - The list of individual items from which a sample is chosen is called the *sampling frame*
  - Some segments of the population of interest are likely to be missed even with careful sampling methods because they are not included in the sampling frame
    - ∗ Example: telephone surveys systematically miss the 6% of American households without phones.

- Nonresponse
  - Some members of the chosen sample cannot be contacted or refuse to answer.
  - This biases the results of the survey if the members who do not respond are different from the general population.
  - Example: in surveys that include questions about household income, families with unusually low or unusually high incomes are less likely to answer that question than are families with moderate income.

- Response bias
  - Respondents may lie, especially about sensitive subjects.
  - Attributes or behavior of interviewers can make this more likely.

  - Example: In a survey concerning roles of family members, a father might tend to respond differently to the question

    > "How many hours per week do you spend caring for your children on average?"

    depending on the gender of the interviewer.

- Bias due to wording of questions
  - leading questions
  - confusing questions
  - questions involving undefined terms
  - Example: Do you eat 5 servings of fruits and vegetables per day?