Econ 300/QAC 201: Quantitative Methods in Economics/Applied Data Analysis

1st Class **6/7/10**

"It is easy to lie with statistics, but it is easier to lie without them."--Frederick Mosteller

--go through syllabus
--hand out first homework for Econ 300 students
--introductory remarks on course
    --most important lessons of the course are in the first three chapters
    Don't be fooled by statistics! Learn to use and interpret them correctly
    --are data collected in a reliable manner?
    --think about how experimental outcomes are influenced by the way the experiment is set up
        --truly random?
        --truly independent?
    --relation, or correlation, does not mean causation
    --we are constantly faced with the need to make decisions under uncertainty
        --comes up in economics in both micro and macro contexts;
        --need to work out sound methods to avoid using bad heuristics/techniques
    --will often ask a short question/pose a puzzle to begin class; encourage discussion in class
    --have a "quote of the day" feature in class

--Chapter 1
Importance of random sampling: trying to avoid bias in the sample
[cartoon handout]

    Types of bias:
    --response bias (phrasing of questions may be leading) [question-wording handout]
    --nonresponse bias (e.g., asking people to answer a mail survey)
    --large-household bias--pollsters often sample whoever answers the door or the phone, so people in small households are more likely to be chosen
    --selection bias--occurs whenever a survey is designed so that certain people have a lower

chance of being chosen

example: how to pick people to survey randomly in order to predict an election outcome?

consider bias sources in the following methods:

--telephone numbers

--on a city street

--from a list of registered voters

in practice, multistage sampling more efficient than simple random sampling (e.g., sample cities/counties, then blocks, then people)

rule of thumb: larger samples generally better than smaller, but can still be biased in important ways

note this also assumes that people don't respond differently when they know they are being polled than when they aren't (I'll call this Heisenberg bias)

for example, the 1947 film "Magic Town" depicts a Midwestern town named Grandview whose citizens' opinions always statistically match those of the country. Jimmy Stewart plays a pollster who uses the town as a shortcut for measuring national public opinion. But when the citizens learn what is happening, they feel obligated to make the most informed choices possible. They arrange for their own surveys, providing library reference materials at every polling booth.

Basic rule for constructing a confidence interval:

$$\pi = P \pm \text{sampling allowance}$$

For simple random sampling, we can state with approximately 95% confidence that:

$$\pi = P \pm 1.96 \sqrt{\frac{P(1-P)}{n}}$$

where π and P are the population and sample proportions respectively and n is the sample size

notice the error gets smaller as n increases; notice also that P(1-P) is maximized if P=.5 and the error then gets smaller the more skewed the sample is in either direction

95% very commonly used as degree of confidence; also in economics tend to see 99% confidence and occasionally 90% confidence

Example of polling:

Results from the final Gallup poll before the 2000 Presidential election
[source: http://www.gallup.com/Election2000/trends.asp; no longer loaded]
Nov. 4-5 Gallup poll by phone of (about) 1448 "likely" voters
Bush polled 47%, Gore 45%, Nader 4%, fringe or undecided 4%

calculate confidence interval for percentage voting for Bush (π):

$$\pi = .47 \pm 1.96 \sqrt{\frac{.47 * .53}{1448}}$$

$\pi = .47 \pm .03$; another way to write this is $.44 \leq \pi \leq .50$

actual results: Bush 47.87% (48%), Gore 48.38% (48%), Nader 2.74% (3%), other 1%
[source: http://www.fec.gov/pubrec/2000presgeresults.htm]

Final Gallup poll before the 2008 election
Oct. 31-Nov. 2 of 2472 likely voters
Obama polled 53%, McCain 42%, 5% fringe or undecided
[source: http://www.gallup.com/poll/111703/Final-Presidential-Estimate-Obama-55-McCain-44.aspx]

calculate confidence interval for percentage voting for Obama (π):

$$\pi = .53 \pm 1.96 \sqrt{\frac{.53 * .47}{2472}}$$

$\pi = .53 \pm .02$; another way to write this is $.51 \leq \pi \leq .55$

results from the November 2008 election
Obama 52.93% (53%), McCain 45.65% (46%), Nader 0.56%, other 0.86%
[source: http://www.fec.gov/pubrec/fe2008/tables2008.pdf]

[note you can get all recent federal election official results at
http://www.fec.gov/pubrec/electionresults.shtml]

The example of polling is an example of statistical inference

Deduction vs. Induction:

population known → sample? This is deduction (probability)
--predicting sample composition

sample known → population? This is induction (statistical inference)
--constructing confidence intervals
--testing hypotheses

Why not just survey the entire population? i.e., the Census attempts to do this (ask how it fails)

1. limited resources (e.g. decide on a purchase after spending a short amount of time shopping around due to time constraints)
2. scarcity--sometimes only a small sample is available--in other words, the whole population is not appropriate for testing the hypothesis
   (e.g. identical twins who have been raised apart to test heredity vs. environment hypotheses)
3. destructive testing (can think of this as a case of the Heisenberg principle--observation alters outcome) (e.g. blood testing--all or a sample?)
4. sampling may be more accurate--we think it will reduce measurement errors (e.g. running the Census vs. the CPS)--spend more money per observation to make sure it is collected carefully

Importance of randomization in trying to avoid bias in an experiment

--randomly assign subjects to treatment and control groups
--keep both subjects and observers/evaluators unaware of the assignments, if possible
evaluator only in the dark is called a blind experiment
evaluator and subject in the dark is a double-blind experiment
(note. this is not standardized terminology. In journal refereeing, a blind journal is one in which the referees know who wrote the papers but the writers do not know who the referees are)

thereby groups are initially equal and continue to be treated equally

Still need to consider original source of subjects (e.g. in case of bonding experiments, what socioeconomic group were the mothers drawn from?) to consider whether results are applicable to the entire population, or only to a subpopulation

Problem for some disciplines: randomized experiments not possible, not practical, or both; must rely on observational data instead

--not possible--often can't do random assignment (e.g., reassigning sex/gender status to a group of professors to see how sex affects salaries)

--not practical--can't afford to do random assignment (e.g., randomly assigning people to live in the city or country to observe effect on crime rate)

--ethical issues (think about use of medical treatments)

Still, there are times when it could be done and isn't

--e.g. random assignment of preschoolers to different educational regimes

And there are techniques for reducing bias in observational studies

multiple regression analysis--use this statistical technique to emulate the holding of confounding factors constant

So why is randomization a good thing?--Because while it is uncertainty, it is manageable uncertainty

Before next class, please read chapter 1 to review material covered this time, read chapter 2.

organize workshop dates and times; posted on course website