Math 140
Introductory Statistics

Professor B. Ábrego

Lecture 7

Sections 4.2, 4.3

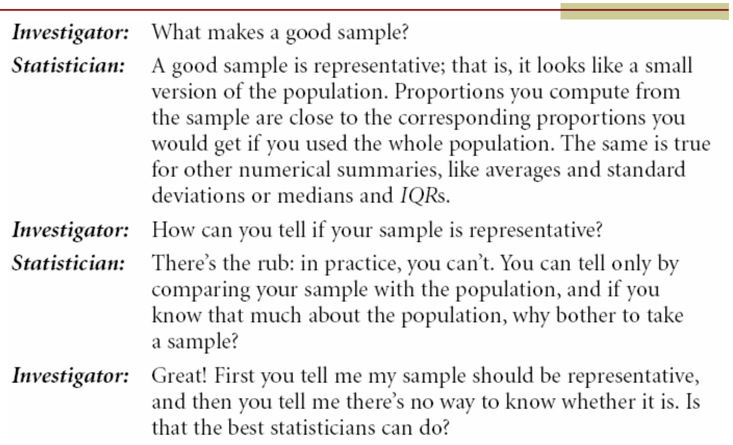
4.2 Why take samples and how not to.

- Advantages of taking samples
  - Sampling saves money and time.
  - Testing can be destructive.
  - Sampling can provide more information than can a cursory study of all items in the population.

Bias: A potential problem with sample data.

- A sampling method is biased if it tends to give samples in which some characteristic of the population is overrepresented or underrepresented.

Bias: Dialogue



Investigator: What makes a good sample?

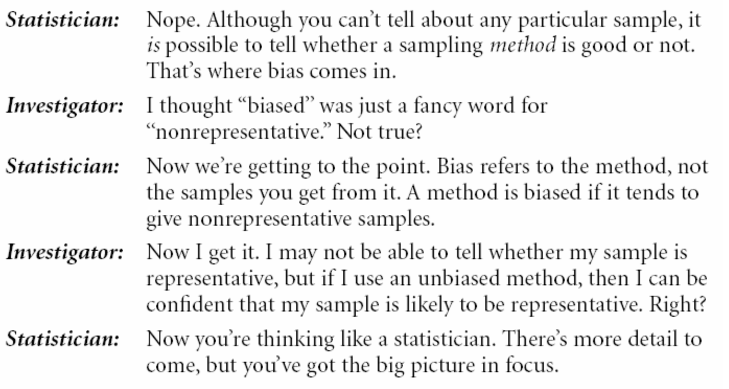
Statistician: A good sample is representative; that is, it looks like a small version of the population. Proportions you compute from the sample are close to the corresponding proportions you would get if you used the whole population. The same is true for other numerical summaries, like averages and standard deviations or medians and *IQR*s.

Investigator: How can you tell if your sample is representative?

Statistician: There's the rub: in practice, you can't. You can tell only by comparing your sample with the population, and if you know that much about the population, why bother to take a sample?

Investigator: Great! First you tell me my sample should be representative, and then you tell me there's no way to know whether it is. Is that the best statisticians can do?

Bias: Dialogue



Statistician: Nope. Although you can't tell about any particular sample, it *is* possible to tell whether a sampling *method* is good or not. That's where bias comes in.

Investigator: I thought "biased" was just a fancy word for "nonrepresentative." Not true?

Statistician: Now we're getting to the point. Bias refers to the method, not the samples you get from it. A method is biased if it tends to give nonrepresentative samples.

Investigator: Now I get it. I may not be able to tell whether my sample is representative, but if I use an unbiased method, then I can be confident that my sample is likely to be representative. Right?

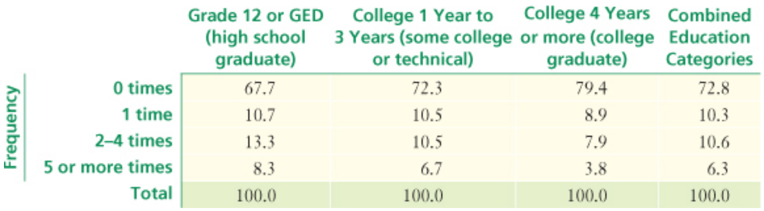
Statistician: Now you're thinking like a statistician. There's more detail to come, but you've got the big picture in focus.

Discussion: Binge Drinking

- The Behavioral Risk Factor Surveillance System (BRFSS) of the U.S. Centers for Disease Control (CDC) collects information on health risk behaviors, and many other health-related issues, for all 50 states, the District of Columbia, and U.S. territories by conducting large monthly surveys. Display 4.7 shows recent data on binge drinking as categorized by education level of the person responding. The numbers are the percentages for that column. Binge drinking is defined as drinking with the intention of becoming intoxicated.

Discussion: Binge Drinking

- a. Describe any patterns you see in the data.
- b. Discuss the possible effect of bias when questioning people on a sensitive question like binge drinking.



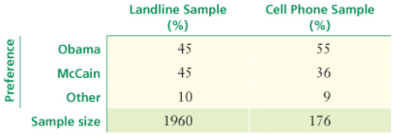
		Grade 12 or GED (high school graduate)	College 1 Year to 3 Years (some college or technical)	College 4 Years or more (college graduate)	Combined Education Categories
Frequency	0 times	67.7	72.3	79.4	72.8
	1 time	10.7	10.5	8.9	10.3
	2–4 times	13.3	10.5	7.9	10.6
	5 or more times	8.3	6.7	3.8	6.3
	Total	100.0	100.0	100.0	100.0

Discussion: Bias in Election Polling

- In the historic presidential race of 2008 between Barack Obama and John McCain, more preference polls were conducted than ever before in attempts to find trends in voting patterns that would shed light on predicting the winner.
- Nearly all of these surveys were conducted by telephone, and the traditional method of telephone polling is to use landlines. But many people now rely mainly on their cell phones, and that fact raises interesting questions about the effect of cell phones on election polling.
- To provide data on the issue, the Pew Research Center for the People & the Press conducted preelection polls with both cell phone and landline samples between the primaries and the election. Some of their data are summarized in the display below.

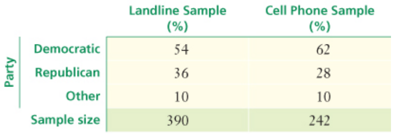
Discussion: Bias in Election Polling

- a. Discuss the differences in the results of the voting preferences for Obama and McCain in the landline survey versus the cell phone survey, as well as possible reasons for the differences.
- b. Discuss the differences in party affiliation between the landline and cell phone surveys for the "under age 30" group and possible reasons for these differences.



Preference	Landline Sample (%)	Cell Phone Sample (%)
Obama	45	55
McCain	45	36
Other	10	9
Sample size	1960	176

Registered voters, September 9–14, 2008.

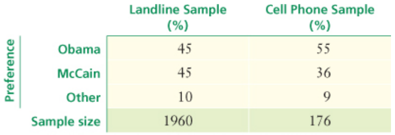


Party	Landline Sample (%)	Cell Phone Sample (%)
Democratic	54	62
Republican	36	28
Other	10	10
Sample size	390	242

Registered voters under age 30, August and September, 2008.

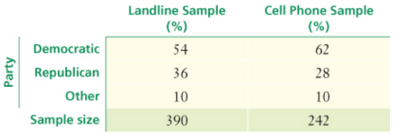
Discussion: Bias in Election Polling

- c. Discuss the potential for bias to affect a presidential preference poll if cell phones are not adequately represented in the sample.
- d. So, why not just add cell phone numbers to the sample? Two main reasons are that cell phone calls are more expensive (costing the respondent by the minute) and the federal Telephone Consumer Protection Act (TCPA) bans unsolicited calls to a cell phone using automated dialing devices.
- Discuss how these issues might relate to bias in the results of telephone surveys in general.



Preference	Landline Sample (%)	Cell Phone Sample (%)
Obama	45	55
McCain	45	36
Other	10	9
Sample size	1960	176

Registered voters, September 9–14, 2008.



Party	Landline Sample (%)	Cell Phone Sample (%)
Democratic	54	62
Republican	36	28
Other	10	10
Sample size	390	242

Registered voters under age 30, August and September, 2008.

Sample Bias

- Size bias: Larger units are more likely to be included.
- Voluntary Response Bias: Those who care about the issue respond.
- Convenience Sample Bias: Units are chosen because of convenience.
- Judgment Sample Bias: Units are choses according to the judgment of someone (expert)

- An Unbiased Sample Method requires that **all** units in the population have a chance of being in the sample.
- A sampling frame is the list of units you use to create the sample. "bad frame, bad sample".

Sample Bias Discussion

- D4. Identify the type of sampling method used in each of these surveys. Would you expect the estimate of the parameter to be too high or too low?
  - a. You use your statistics class to estimate the percentage of students in your school who study at least 2 hours a night.
  - b. You send a survey to all people who have graduated from your school in the past 10 years. You use the mean annual income of those who reply to estimate the mean annual income of all graduates of your school in the past 10 years.
  - c. A study was designed to estimate how long people live after being diagnosed with dementia. The researchers took a random sample of the people with dementia who were alive on a given day. The date the person had been diagnosed was recorded, and after the person died the date of death was recorded.

Sample Bias Discussion

- D5. You want to know the percentage of voters who favor state funding for bilingual education. Your population of interest is the set of people likely to vote in the next election. You use as your frame the phone book listing of residential telephone numbers. How well do you think the frame represents the population? Are there important groups of individuals who belong to the population but not to the frame? To the frame but not to the population? If you think bias is likely, identify what kind of bias and how it might arise.

Response Bias

- Non-Response Bias: You get no data or not enough data. e.g. 80% of people contacted refuse to answer a Survey (or answer incorrectly)

I'M FILLING OUT A READER SURVEY FOR *CHEWING* MAGAZINE.
SEE, THEY ASKED HOW MUCH MONEY I SPEND ON GUM EACH WEEK, SO I WROTE, "$500." FOR MY AGE, I PUT "43," AND WHEN THEY ASKED WHAT MY FAVORITE FLAVOR IS, I WROTE "GARLIC/CURRY."
CHEWING
THIS MAGAZINE SHOULD HAVE SOME AMUSING ADS SOON.
I LOVE MESSING WITH DATA.
© 1995 Watterson/Dist. by Universal Press Syndicate
8-23 WATTERSON

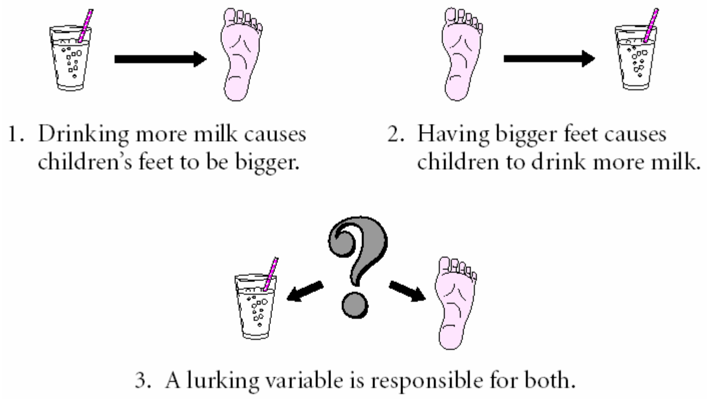
Response Bias

- Non-Response Bias: You get no data or not enough data. e.g. 80% of people contacted refuse to answer a Survey
- Questionnaire Bias: Arises from the way the questions are asked.

  Example.
  - I would be disappointed if Congress cut its funding for public television.
  - Cuts in funding for public television are justified as part of an overall effort to reduce federal spending.

4.3 Experiments and Inference about Cause

- Cause and Effect



1. Drinking more milk causes children's feet to be bigger.
2. Having bigger feet causes children to drink more milk.
3. A lurking variable is responsible for both.

4.3 Experiments and Inference about Cause

- Cause and Effect
- Lurking Variable:

  A variable in the background that could explain a pattern between the variables investigated.
- How to establish cause and effect?

Answer: Conduct an experiment.

Experiments

- Goal: To establish cause and effect by comparing two or more ***conditions*** (called treatments) using an ***outcome variable*** (called the response).
- To be a real experiment, the subjects must be randomly assigned to their treatments. To make this distinction sometimes we call these Randomized Experiments.

Example: Kelly's Hamsters

- Assumptions
  - Golden Hamsters hibernate.
  - Hamsters rely on the amount of daylight to trigger hibernation.
  - An animal's capacity to transmit nerve impulses depends in part on an enzyme called Na^+K^+ ATP-ase.
- Question: If you reduce the amount of light a hamster gets, from 16 hours to 8 hours per day, what happens to the concentration of Na^+K^+ ATP-ase.

Example: Kelly's Hamsters

- Subjects: Eight golden hamsters.
- Treatments: Raised in long days (16 hours) or short days (8 hours) of daylight.
- Random Assignment of Treatments: Kelly randomly assigns four of the hamsters to short days, and four to long days.
- Replication: Each treatment was given to four hamsters.
- Response Variable: Enzyme concentration.

Kelly’s Hamsters (Results)

- Results

Enzyme concentrations in milligrams per 100 milliliters.

Short Days	12.500	11.625	18.275	13.225
Long Days	6.625	10.375	9.900	8.800

Kelly’s defense of her design

Kelly: I claim that the observed difference in enzyme concentrations between the two groups of hamsters is due to the difference in daylight.

Skeptic: Wait a minute. As you can see, the concentration varies from one hamster to another. Some just naturally have higher concentrations. If you happened to assign all the high-enzyme hamsters to the group that got short days, you’d get results like the ones you got.

Kelly: I agree, and I was concerned about that possibility. In fact, that’s precisely why I assigned day lengths to hamsters by using random numbers. The random assignment makes it **extremely unlikely** that all the high-enzyme hamsters would get assigned to the same group. If you have the time, I can show you how to compute the probability.

Skeptic: (*Hastily*) That’s OK for now. I’ll take your word for it. But maybe you can catch me in Chapter 6.

Discussion

- **D18**. Plot Kelly's results in a dot plot, using different symbols for hamsters raised in short days and those raised in long days. Do you think the evidence supports a conclusion that the number of daylight hours causes a difference in enzyme concentration?

- **D19**. Kelly has shown that hamsters raised in less daylight have higher hormone concentration than hamsters raised with more daylight. In order for Kelly to show that less daylight *causes* an increase in the hormone concentration, she must convince us that there is no other explanation. Has she done that?