Lecture 6

Nancy Pfenning Stats 1000

Designing Observational Studies

The main advantage of an experiment is that it gives researchers as much control as possible over the explanatory variable of interest. One disadvantage is a possible lack of **ecological validity** if the setting is too unrealistic.

Example

> The article **When your hair's a real mess, your self-esteem is much less** reports: "A Yale University study of the psychology of bad hair days found that people's self-esteem goes awry when their hair is out of place. They feel less smart, less capable, more embarrassed and less sociable." The study is described as follows: "For the study, researchers questioned 60 men and 60 women ages 17 to 30, most of them Yale students...The people were divided into three groups. One group was questioned about times in their lives when they had bad hair. The second group was told to think about bad product packaging, like leaky containers, to get them in a negative mindset. The third group was not asked to think about anything negative. All three groups then underwent basic psychological tests of self-esteem and self-judgment. The people who pondered their bad hair days showed lower self-esteem than those who thought about something else." Is thinking about bad hair in this artificial setting truly comparable to experiencing bad hair in real life? Interesting to note is the fact that the study was paid for by Procter & Gamble, which makes shampoo.

We must also be aware that there are many "treatments"—such as income or smoking—which are impractical or unethical for researchers to impose. Thus, observational studies are a very commonly used alternative for gathering statistical data.

Example

> In 1939, an experiment was carried out on 22 orphans in Iowa: one group of 11 was given positive speech therapy, and the other 11 were induced to stutter by constant badgering on the part of their speech therapist. Out of those 11, 8 became chronic stutterers. One of them, interviewed over sixty years later, said, "It's affected me right now. I don't like to read out loud because I'm afraid of making a mistake. I don't like talking to people because of saying the wrong word." The therapist herself came to deeply regret her role in this experiment, which eventually led to a theory that helped thousands of children overcome stuttering. Most people today would agree that the price of this knowledge was simply too high for the unwitting victims.

Thus, many problems must be researched via an observational study, rather than an experiment. There are many approaches that may be taken in such a study, and many pitfalls to be avoided.

Parents putting more time into kids is the headline of an article published in 2001. "The University of Michigan research showed that children spent four to six more hours a week with their parents in 1997 than they did in 1981...The gains recorded were significant: In 1997 children ages 3 to 12 spent about 31 hours a week with their mothers, a gain of six hours over 1981, and 23 hours a week with their fathers, a gain of four hours... 'A lot of popular culture has been saying that we're spending less time with our kids and that it's bad for our children, and it turns out we're spending more time with them,' said study co-author John F. Sandberg."

First, notice that we are mainly interested in one categorical explanatory variable (whether the year was 1981 or 1997) and one quantitative response variable (amount of time spent with parents). How do you think the data were gathered for this study? A **retrospective** study would ask particpants to recall how much time their parents spent with them almost twenty years ago: relying on people's memories for something like this would surely produce inaccurate results. Better to use a **prospective** study, carried forward in

time. In fact, the study "compared two nationally representative samples of children who recorded their minute-by-minute activities over two days..." But is the prospective study really flawless? The study was "based on time diaries completed by the children, with parental help if needed." Can you think of any influences which could prevent these samples from being truly representative? Keeping in mind Sandberg's comment, when would parents be more inclined to bias their schedules, or reported schedules, towards extra time with their children: in 1981 or in 1997?

Another article is titled **Light drinking in pregnancy has long-term effects on kids**. It's easy to see why this would have to have been conducted as an observational study rather than an experiment. Women were recruited when they were four months pregnant, and it was recorded whether they had up to 1.5 drinks per week during their first trimester; researchers also recorded second trimester drinking. Thus, the study was both retrospective and prospective. The article focuses on comparing values of a quantitative response variable (weight) for two categorical groups (children of light vs. non-drinkers), but the research included moderate and heavy drinkers, too. Children of moms who were light drinkers weighed an average of 2.5 pounds less at age 14 than children of non-drinkers.

Pitfalls in Experiments and Observational Studies

Your textbook lists these potential difficulties: confounding variables; extending results inappropriately; interacting variables; placebo, Hawthorne, and experimenter effects; ecological validity; relying on memory or secondhand sources.

Example

> **Study finds cell phones related to accidents** states that researchers drew their conclusions from interviewing 699 drivers in Toronto who acknowledged being owners of cellular phones when they came to a Toronto police collision reporting center. Their cellular phone bills were checked for whether they were using the phones either at or just before the time their accidents occurred.
>
> Was this an experiment or an observational study? An observational study, because researchers can't very well impose cell phone conversations on drivers.
>
> Did the study rely on people's memories about cell phone use? No.
>
> Why did the researchers only study cell phone *owners*? Because whether or not someone owns a cell phone could impact the likelihood of accidents.
>
> Are there any other possible lurking variables? "The researchers said there was no evidence of a cause-and-effect link between use of the phones and collisions, but a numerical risk making the chances more likely on the days they used the phones than on days they didn't...For example, emotional stress may lead to both increased use of a cellular telephone and decreased driving ability." This may be so, but in the end our common sense tells us that use of cell phones can, and does, to some extent cause accidents.

Example

> **Net surfing can put you in a funk, CMU finds**: this 1998 experiment (see text page 6) concluded that internet use leads to higher levels of depression and loneliness. Some of the study's detractors felt that the subjects were bound to become more depressed anyway, because they were all living in Pittsburgh! It's always a good idea to decide if it's appropriate to extend your results from the particular sample to a larger population.

Sometimes the explanatory variable affects the response only when a certain trigger is present. This additional variable, whose presence or absence impacts the relationship, is called an **interacting variable**.

Example

> **Starchy foods tied to cancer** states, "A diet high in starchy foods like potatoes, rice and white bread appears to increase the risk of pancreatic cancer in women who are already overweight."

It was important for reporters to mention this interacting variable (weight) so that women of normal weight would not feel compelled to reduce their starch intake.

We have already mentioned that observational studies have the advantage of observing variables' values as they naturally occur. Often, participants in an experiment behave differently than they normally would, a phenomenon called the **Hawthorne effect**.

Example

Researchers at UPMC wanted to compare the effectiveness of ordinary and vegetarian low-fat diets for overweight women. Participants recruited for such a study may modify their eating habits in general, just because they are being monitored by the researchers. In order to take the Hawthorne effect into account, researchers let half of the subjects choose which of the diets they'd be more inclined to adopt naturally, and divided them into regular and vegetarian sub-groups by request. The other half of the subjects were given no choice; they were assigned to regular and vegetarian sub-groups randomly by the researchers. Comparing results for those who did and did not choose their diet preference enabled researchers to assess to some extent the effect of being told what to do by researchers, as opposed to making one's own choices as one would in an observational study. Interestingly, the subjects who were permitted to choose their own type of diet were less successful in terms of weight loss. Diets do tend to be only temporarily successful, and this is perhaps due to the fact that people adhere only as long as their food intake is being strictly enforced by an outsider.

Besides the risk of subjects behaving differently, there is also the possibility that researchers would elicit or record responses in a biased way, called the **experimenter effect**. Double-blinding is the simplest way to avoid this pitfall.

Example

Comparative experiments have been conducted to establish whether or not calcium supplements can lower blood pressure. If a researcher is using a hand blood pressure monitor, assessing the numbers is not entirely cut-and-dried. For this reason, whoever measures the subjects' blood pressures should *not* know whether they are in the calcium or the control group.

Consider if any of these pitfalls apply to our opening examples from Lecture 1. The earnings example may involve incorrect data due to relying on students' imperfect memories about what they earned. In the Larry Flynt example, he was guilty of extending results inappropriately. In the family size vs. IQ example, a confounding variable (heredity) was involved. A confounding variable may have entered into a sun-time vs. sunscreen observational study, but probably not in the experiment. If SAT scores of students who did and did not take a prep course were being compared for individuals in an observational study, then confounding variables could easily enter in. If an experiment with a control group *not* receiving preparatory help were carried out, then those receiving preparatory help may end up doing better just as a result of a boost in confidence (a type of placebo effect).

Example

An article entitled **Family dinners benefit teens** reports: "Eating dinner together as a family is one way of keeping teen-agers well adjusted and out of trouble, a new study shows. Adjusted teens—who were less likely to take drugs or be depressed and were more motivated at school and had better peer relationships—ate with their families an average of five days a week; nonadjusted teens ate with their families only three days a week, according to the study presented at last week's American Psychological Association's convention in Chicago. Family mealtimes, researchers said, appear to plan an important role in helping teens deal with adolescence."

This was an observational study because no treatment was imposed by researchers (nobody obliged some teenagers to have dinner with their families a certain number of days per week).

The explanatory variable is how often a teen ate dinner with the family, and seems to be treated as categorical (averaging five vs. averaging three). The response is being well- or non-adjusted, also apparently categorical, although quantitative levels of depression, motivation, etc. may have been recorded before ultimately classifying a teen as belonging in one group or the other.

Results could easily have come about because of confounding variables, such as how dedicated parents were in general, and how healthy the parent-child relationship was. Imagine taking a well-adjusted teen in a happy family environment and sending the teen out to eat away from home two additional days a week, while holding all other circumstances intact. Is this really likely to cause the teen to become nonadjusted? In discussions of this article, some students have suggested that forcing a non-adjusted teen to eat dinner with the family more often could conceivably make matters go from bad to worse.

There is even a possibility that causation works in the opposite direction from what the researchers suggest: perhaps being well-adjusted simply causes a teenager to choose to eat with the family more often.

Exercise: Find an article or report about an observational study. Tell what the variables of interest are, whether they are quantitative or categorical, which is explanatory and response (if there are two variables). Are there any potential confounding variables that should have been controlled for? Are there any other pitfalls of concern?

Lecture 7

Chapter 4: Sampling Surveys and How to Ask Questions

Sample surveys are an extremely common method of gathering information and opinions: a subgroup of a large population is questioned on a set of topics. Because no treatment is imposed by researchers, a sample survey is a type of observational study, not an experiment.

As long as the sample has been selected at random, the **sample proportion** with a given response tends to be surprisingly close to the unknown **population proportion**. The **margin of error** measures how far they tend to be from one another: 95% of the time, sample proportion will come within one margin of error of population proportion. By Chapter 10 we will have developed the theory necessary to calculate the precise margin of error in a given situation. For now, we can use the following conservative estimate: margin of error for proportions is roughly one over square root of sample size, or $\frac{1}{\sqrt{n}}$ where n denotes the sample size. Often the proportion and accompanying margin of error are reported as percentages. If the proportion is p, then the percentage is 100% times p. Instead of making a statement about our sample proportion as it relates to unknown population proportion, our real goal is to make a statement about the unknown population proportion as it relates to sample proportion. Probability theory developed in future chapters will justify our making statements like the following:

We are approximately 95% confident that population proportion falls between sample proportion minus $\frac{1}{\sqrt{n}}$ and sample proportion plus $\frac{1}{\sqrt{n}}$.

Example

664 teenagers who reported having sex for the first time between 1999 and 2000 were asked where this first encounter took place; 56% said it was at their own or their partner's home. Assuming those 664 constitute a random sample, what can we say about the location of all teens' first sexual encounter?

$$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{664}} = .0388 \approx 4\%$$

We are approximately 95% confident that the proportion of all teens having their first sexual encounter at their or their partner's home is between 56%-4% and 56%+4%, that is, between 52% and 60%.

Example

40% (that is, 400) of 1000 husbands and wives surveyed admitted that they kept secrets from their spouses; the most common secret, admitted by 48% of those 40%, that is, 19%, was not telling their spouses about the real price of something they bought. (In case you're curious, this percentage held about equally for men and women.) Since $\frac{1}{\sqrt{1000}} = \frac{1}{32} = .03$ or about 3%, we are about 95% confident that between 16% and 22% of all married people have kept a secret about the price of something they bought.

Example

Of 5685 respondents in a survey, 4948 confessed to singing in their cars on a regular basis. Since $\frac{4948}{5685} = 80\%$, and since $\frac{1}{\sqrt{5685}} = .013$ or about 1%, we are 95% sure that the percentage of all people who sing in their cars is between approximately 79% and 81%. [To get a margin of error of exactly 1%, we'd need to sample 10,000 people.]

These examples should impress on you what a powerful tool random sampling is: it lets us, with a fairly high degree of confidence, come within just a few percentage points of the real percentage that holds for the entire population, whose size may be huge. Pollsters typically sample about 1000 people, which results in a margin of error of about 3%. There are many advantages to taking a sample rather than a **census** of the whole population, which is often impractical or even impossible, costly in terms of time and money, and still subject to bias because of undercoverage of certain groups.

Note: Estimating population proportion is appropriate when one categorical variable is involved. Such was the case for the three examples above: location of first sexual encounter being in the home or not, keeping a secret from one's spouse about a purchase or not, singing in the car or not.

The following example can be carried out in class and the results examined for bias in the selection process:

Example

Each student picks three states at random from the following, exactly as they are shown here:

Alabama	Hawaii	Massachusetts
Alaska	Idaho	Michigan
Arizona	Illinois	Minnesota
Arkansas	Indiana	Mississippi
California	Iowa	Missouri
Colorado	Kansas	Montana
Connecticut	Kentucky	Nebraska
Delaware	Louisiana	Nevada
Florida	Maine	New Hampshire
Georgia	Maryland	New Jersey

Bias, the systematic tendency to over- or under-estimate, can creep into a survey in a variety of ways. **Selection bias** results from taking a non-representative sample. **Nonresponse bias** occurs when too many individuals selected cannot be reached or choose not to respond. Such individuals may be different in an important way, so that the few who *do* respond are not truly representative of the population. **Response bias** represents failure of the respondents to provide honest or accurate responses, a pitfall which can be avoided to some extent by careful wording of questions and by ensuring anonymity.

Example

Suppose I want to survey a random sample of 6 from 80 class members to get their opinion about our textbook. How can this be accomplished?

1. I could ask for 6 volunteers from the class. This would result in selection bias, favoring people with strong positive or negative feelings. **Self-selected** or **volunteer** samples, as in call-in or internet polls, are practically guaranteed to be biased, often quite heavily.
2. I could ask the next 6 students who come in to my office hours. This is called a **convenience**, or **haphazard** sample—an excellent way to predispose a survey to bias!
3. I could assign each student in the room a number from 1 to whatever, then use a row of random digits (details to follow) to select 6 at random. The problem here is that the **sampling frame**—all students attending—does not match the **population**—all students enrolled. Absent students might tend to feel negatively about the course in general, including the textbook; or on the other hand maybe they don't attend because they feel the book is good enough to teach them all they need to know!
4. I could take a random sample from my roster of students and mail them a questionnaire. What if the questionnaires accidentally get mailed to students' family homes, instead of to where they reside during the school year, and only 2 students are reached? What if all 6 students are reached, but only 2 choose to respond? Can we assume those 2 opinions to be representative? No; non-respondents may differ from respondents in important ways, and if there are too many of them, our conclusions may be biased.
5. I could take a random sample of 6 students from all those enrolled, set up meetings with them, and interview them about the textbook. This approach may make students feel some pressure to respond favorably and not disappoint me, again leading to bias.

The way to avoid selection bias in a survey is to let chance govern our selection of a sample by using a **probability sampling plan**. The simplest probability sampling plan, already mentioned in Chapter 3, is a **simple random sample** (sampled at random and without replacement).

Example

Again, suppose I want to survey a random sample of 6 from 80 class members to get their opinion about our textbook. In case a computer isn't handy, a **table of random digits** can be used to select a simple random sample. In Table 4.1 or any random digit table, all digits 0-9 are equally likely and each digit is independent of the others.

1. Give each individual a number label with as few digits as possible. [List class members' names alphabetically and assign them numbers 01-80; each label has 2 digits.]
2. Starting anywhere in Table 4.1, read the digits off by ones, or twos, or threes,... according to the number of digits in our labels. [Arbitrarily, let's start at row 5. We use two digits at a time—if we'd been selecting a sample from 8000 students, we'd use four digits at a time.]
3. As we read off each set of digits, we determine if it is a label for one of our units. [For this example, anywhere from 01 to 80.] If so, then that individual is selected. Repeat numbers are skipped over [which means our sampling is done *without* replacement], as are numbers which are not useable as labels.

   98 | 87 | 9 3 | **40** | **72** | **04** | **18** | 9 3 | **16** | 72 | **33** | 35 | 7 5 | 31 | 91 | ...

   Students labeled 40, 72, 4, 18, 16, and 33 constitute our simple random sample, and will be interviewed.

Example

Use row 7 to pick 5 letters at random from the alphabet. You should get P, X, A, N, B.

Other, more complicated, probability sampling plans include:

1. stratified random sample: take separate random samples from similar groups (strata) within the population (eg. regions of the U.S.)

2. cluster sampling: small similar groups (clusters) within the population are sampled at random, but *all* units in each cluster are sampled (eg. sample *all* people living on each of randomly selected streets in Pittsburgh).

3. telephone surveys: using a phone book would reduce the sampling frame to people with listed numbers; instead, most telephone surveys use random digit dialing.

4. multistage sampling: stratify in stages (eg. divide the U.S. into regions NE, SE, NW, etc.; within each region, stratify by urban, suburban, or rural; then take random samples of communities and street blocks; finally, sample clusters of people on the same block).

A **systematic sampling plan**, such as choosing every 10th name on my alphabetical roster of students, does not utilize randomness, but in some situations this would not prevent the sample from being representative of the population. Such a plan is at times a reasonable alternative to a probability sampling plan.