Ch. 4 Designing Studies

Section 4-2: Experiments
4-2 Experiments

Scenario: Last year ECRCHS offered an after school ACT prep class that students could volunteer to take. 44 students took the course and then took the ACT. The average ACT score for this group was 23.2. The average ACT score for all students who did not take the prep class was 21.4.

Can you conclude that taking the prep course will cause a student’s ACT score to increase?  no
What are some lurking variables here? Motivation, intelligence, money, test taking ability
How could we show causation? Carefully design an experiment.

lurking variable – a variable that may influence the response variable

Sample surveys aim to gather info about a population without disturbing the population, which is a type of observational study.

In this section, you’ll be learning about statistical designs for experiments, a different way to produce data.
Definitions:

- Observing individuals and measuring variables of interest but without attempting to influence the responses is called **an observational study**.

**GOALS:**
1) to describe/compare groups or situations
2) to examine relationships

- Deliberately imposing a treatment on individuals and then measuring their responses is called **an experiment**. The advantage of an experiment over an observational study is that it allows you to show **causation**.

**GOAL:** to determine whether a treatment causes a change in the response
Should women take hormones such as estrogen after menopause, when natural production of these hormones ends? In 1992, several major medical organizations said “Yes.” Women who took hormones seemed to reduce their risk of a heart attack by 35% to 50%. The risks of taking hormones appeared small compared with the benefits.

The evidence in favor of hormone replacement came from a number of *observational studies* that compared women who were taking hormones with others who were not. But the women who chose to take hormones were richer and better educated and saw doctors more often than women who didn’t take hormones. Because the women who took hormones did many other things to maintain their health, it isn’t surprising that they had fewer heart attacks.
To get convincing data on the link between hormone replacement and heart attacks, we should do an experiment.

Experiments don’t let women decide what to do. They assign women to either hormone replacement or to dummy pills that look and taste the same as the hormone pills. The assignment is done by a coin toss, so that all kinds of women are equally likely to get either treatment.

By 2002, several experiments with women of different ages agreed that hormone replacement does not reduce the risk of heart attacks. The National Institutes of Health, after reviewing the evidence, concluded that the first studies were wrong. Taking hormones after menopause quickly fell out of favor.
A variable that is not among the explanatory or response variables in a study but that may influence the response variable is called a **lurking variable**.

Lurking variables often lead to **confounding**.

Let’s consider a **lurking variable** from the observational studies of hormone replacement: number of doctor visits per year. The women who chose to take hormones visited their doctors more often than the women who didn’t take hormones.

Did the women in the hormone group have fewer heart attacks because they got better health care or because they took hormones? We can’t be sure. A situation like this, in which the effects of two variables on a response variable cannot be separated from each other, is called **confounding**.
Definitions:

- Observing individuals and measuring variables of interest but without attempting to influence the responses is called an **observational study**.

- Deliberately imposing a treatment on individuals and then measuring their responses is called an **experiment**. The advantage of an experiment over an observational study is that it allows you to show **causation**.

- A variable that is not among the explanatory or response variables in a study but that may influence the response variable is called a **lurking variable**.

- Lurking variables often lead to **confounding**.

○ Picture:

```
\[\text{ACT Prep Class} \rightarrow \text{Correlation} \rightarrow \text{ACT Scores} \rightarrow \text{motivation, money, intelligence, test taking, ability}\]
```

*BEWARE OF LURKING VARIABLES!*
Lurking variables are most common in observational studies. They are much less common in designed experiments because we randomize to avoid such things. Lurking variables will show up only if we fail to randomize completely or correctly.

Confounding, however, can show up in an experiment either through poor design or just because there is no reasonable way to avoid it. Things happen that are not under our control, but might be confounded with the factors, making it impossible to tell whether our treatment or the external change was responsible for the response.

The Language of Experiments

- A specific condition applied to individuals in an experiment is called the treatment.
- The smallest collection of individuals to which treatments are applied are called experimental units. Change on your paper
- When the units are human beings, they are called subjects.
- The explanatory variables in an experiment are often called factors. Each factor may have several different levels.
1) An experiment because a treatment (brightness of screen) was imposed on the laptops.

2) An observational study. Students were not assigned to a particular number of meals to eat with their family per week.

3) Explanatory variable: the number of meals per week eaten with their family. Response variable: probably their GPA.

4) This is an observational study and there may well be lurking variables that are actually influencing the response variable. For instance, families that eat more meals together may also be families where the parents show more interest in their children’s education and therefore help them to do better in school.
The Randomized Comparative Experiment. Most well-designed experiments compare 2 or more treatments.

Scenario: ECHRHS has decided to offer an ACT prep class again this year. It will be offered in two different formats: online or classroom teacher. The counselors want to know which teaching method will yield higher ACT scores so they have allowed me to set up an experiment. 50 students have signed up to take some form of the ACT prep class.

We have to think about how to put the students into the two formats. What do you think would happen if we let the students choose? More motivated students will probably do online, which would cause bias results.

Therefore, we need to do random assignment.

For Completely Randomized Design: Can use “hat” method, random number generator, or table of random digits.

Does using chance to assign treatments guarantee a completely randomized design? No, if students flipped a coin, there probably wouldn’t be 25 in each group, so making them even would be “forced”, which isn’t random.
Discuss an appropriate experimental design and outline it below.

For random assignment, start by labeling students 01-50, alphabetically. Then use a table of random digits by looking at 2 digits at a time and ignoring repeats, numbers greater than 50, and 00. The first 25 students who come up in the table of random digits are in Group 1 and receive the ACT class with a teacher. The remaining 25 students get an online class. Then compare ACT scores. On AP Exam, diagram isn’t enough to get full credit. Must describe how treatments are assigned to the experimental units and state what will be measured or compared.
No, multiple treatments serve as control.

Primary purpose of a **control group** is to provide a baseline for comparing the effects of treatments.

Definitions:

- In an experiment, we must always use **random assignment** to assign experimental units to treatments. This means using some sort of chance process.
- When the treatments are assigned to all the experimental units completely by chance, we have a **completely randomized design**.
- A **control** group provides a baseline for comparing the effects of all the treatments. **Multiple treatments** often serve as a control for the experiment and a separate control group is not necessary.
1) Using an alphabetical list of the students, assign each student a number between 01 and 29. Pick a line of the random number table and read off two-digit numbers until you have 15 numbers between 01 and 29, ignoring repeats. These students belong in the treatment group, where students will meet in small groups. The other students will view the videos alone. Then compare the evaluations from each group.

2) The purpose of the control group is to have a group with which to compare the treatment group. Presumably, the students have been evaluating their own performances by themselves. If you incorporate such a group into your experiment, you can evaluate whether the group work is actually better.
Three Principles of Experimental Design

1. **Control** for lurking variables that might affect the response. Use a **comparative** design and ensure that the only systematic difference between the groups is the treatment administered.

2. **Random Assignment**: Use impersonal chance to assign experimental units to treatments. This helps create roughly equivalent groups of experimental units by balancing the effect of lurking variables that aren’t controlled on the treatment group.

3. **Replication**: Use enough experimental units in each group so that any differences in the effects of the treatments can be distinguished from chance differences between the groups.

Experiments: What Can Go Wrong?  [https://youtu.be/z03FQGlGgo0](https://youtu.be/z03FQGlGgo0)

Good experiments require careful attention to details to ensure that all subjects are treated identically, except for the actual treatments being compared.

If some subjects in a medical experiment take a pill each day and a control takes no pill, the subjects are not treated identically. How can we fix this?
Some experiments give a placebo (fake treatment) to a control group. That helps prevent confounding due to the placebo effect, in which some patients get better because they expect the treatment to work even though they have received an inactive treatment.

Since the placebo effect is so strong, it’s foolish to tell patients they’re receiving a placebo. Bias can occur.

In an experiment, you do not want the experimental units to know which treatment (or placebo) they have been given. This means the experiment is single blind.

Experimenters shouldn’t know who receives a placebo as well, since they can interact and diagnose differently between placebo and treatment patients.

In addition, you do not want the experimenters to know which treatment (or placebo) is being given to the experimental units. This means the experiment is double blind.
1) No. It is possible that women who “thought” they were getting an ultrasound would have different reactions to pregnancy than those who knew that they hadn’t received an ultrasound.

2) No. While the people weighing the babies at birth may not have known whether that particular mother had an ultrasound or not, the mothers did know whether they had had an ultrasound or not. This means that the mothers may have affected the outcome since they knew whether they had received the treatment or not.

3) An improved design would be one in which all mothers are treated as if they had an ultrasound, but for some mothers the ultrasound machine just wasn’t turned on (a fact that would not be obvious to the women). This means that the ultrasound would have to be done in such a way that the women could not see the screen.
Inference for Experiments

Scenario: Recall our experiment of online ACT prep class vs. classroom instruction ACT prep class. The results are in.

Average ACT score (online class): 23.2. Average ACT score (classroom teacher): 21.0

Can we therefore conclude that the online ACT prep class caused better scores? _maybe_

These results could’ve happened by chance.

An observed effect that is so large that it would rarely occur by chance is called _statistically significant_.

Remember, strong association _does not_ imply causation.

However, _statistically significant_ association from a well-designed experiment _does_ imply causation.

- In an experiment, researchers usually hope to see a difference in the responses so _large_ that it is unlikely to happen due to chance.

- If we observe statistically significant differences, we have good evidence that the treatments actually _caused_ these differences.
**Block Design**

Completely randomized designs are the simplest statistical designs for experiments.

But just as with sampling, there are times when the simplest method (SRS) doesn’t yield the most precise results.

**Blocks** are another form of control.
New scenario: The counselors at ECRCHS hypothesize that the online vs. classroom results are being greatly affected by the grade of students that were put into each treatment group. They know that Juniors generally score better on the ACT than Sophomores. They think that the average ACT score for the online classes is much higher because there may have been more Juniors in this class than in the classroom teacher prep course.

How could we argue against this? **We used random assignment.**

Suppose the counselors won’t let go of their hypothesis. Design a new experiment.

30 juniors, 20 sophomores

- **30 juniors** → Random Assignment → Group 1 (15) → Treatment 1 (teacher) → Compare ACT scores
- **50 students**
- **20 sophomores** → Random Assignment → Group 2 (10) → Treatment 2 (online) → Compare ACT scores
In the previous experimental design, the population was split into two groups (Juniors and Sophomores) called **blocks**. A block is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.

When the random assignment of experimental units to treatments is carried out separately within each block, we have a **randomized block design**.

When using a randomized block design, separate conclusions can be made for each **block**.

**Don’t confuse randomized block design with a stratified random sample!!**

When selecting a sample for a **survey**, use a **stratified random sample**.

When setting up an **experiment**, use a **randomized block design**.

**Be careful!** Don’t mix the language of experiments and sample surveys/observational studies when answering questions.
Matched Pairs Design  A common type of randomized block design for comparing two treatments.

Back to the ACT prep class. Let’s look only at the Juniors. The counselors are now worried that a student’s GPA is certainly going to affect their ACT score and they want to be sure that the different GPAs are being evenly distributed into the two treatment groups.

How could we be sure the GPAs are evenly distributed? **Make a ranked list of students by GPA, pairs students with similar GPAs, randomly assign one to Group 1 and the other to Group 2.**

This is called a _matched pairs design_. Essentially, we are creating blocks of size **2**.

To compare scores at the end, **take the difference** of the scores and compare. For example: *Online - Teacher.* A lot of positive values would mean that the online course was better.

More often, each “pair” in a matched pairs design consists of just **1** experimental unit that receives **2** treatments.

The order of the treatments can influence the response, so we **randomize** the order in which the treatments are given.
4-3: Scope of Inference

What type of inference that can be made from a particular study depends on the design of the study.

What can you conclude in each of the following situations?

(1) The U.S. Census Bureau carries out a monthly Current Population Survey of about 60,000 households using Table D to select individuals. They find that 12% of the sample is unemployed. Because random sampling is used, we can infer 12% of the population is unemployed.

Conclusion: ________

(2) Scientists performed an experiment that randomly assigned 21,000 volunteer subjects to one of two treatments: sleep deprivation for one night or unrestricted sleep. Two days later, the unrestricted sleep group scored significantly higher on a performance test than the sleep deprived group. Because of random assignment of treatments, we can infer more sleep causes increased scores.

Conclusion: ________
Inferences about Populations

If the individuals were randomly selected, then we can make inference about the **population**. Think observational study.

Inferences about Cause and Effect

If the individuals were randomly assigned to treatment groups, then we can make inference about **cause and effect**. Think experiment.

When discussing the scope of inference, always be sure to use **context**.

**IP** – inference about the population
**ICE** – inference about cause and effect

<table>
<thead>
<tr>
<th>Were individuals randomly assigned to groups?</th>
<th>Were individuals randomly selected?</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>NO</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
104) **Frozen batteries** Will storing batteries in a freezer make them last longer? To find out, a company that produces batteries takes a random sample of 100 AA batteries from its warehouse. The company statistician randomly assigns 50 batteries to be stored in the freezer and the other 50 to be stored at room temperature for 3 years. At the end of that time period, each battery’s charge is tested. *Result:* Batteries stored in the freezer had a higher average charge, and the difference between the groups was statistically significant. What conclusion can we draw from this study? Explain.

Since this study involved random assignment to the treatments, we can infer cause and effect. Therefore, we can conclude that storing batteries in the freezer leads to a higher average charge for batteries produced by this company. Also, since the batteries were randomly chosen, we can generalize to the whole population of batteries.
Attend church, live longer? One of the better studies of the effect of regular attendance at religious services gathered data from a random sample of 3617 adults. The researchers then measured lots of variables, not just the explanatory variable (religious activities) and the response variable (length of life). A news article said: “Churchgoers were more likely to be nonsmokers, physically active, and at their right weight. But even after health behaviors were taken into account, those not attending religious services regularly still were about 25% more likely to have died.” What conclusion can we draw from this study? Explain.

Since this study involved a random sample, we can make an inference about the population. It appears that those who attend religious services regularly have a lower risk of dying younger. However, we cannot infer cause and effect. We do not know that attending religious services is the reason for this lower risk.
Studying frustration

A psychologist wants to study the effects of failure and frustration on the relationships among members of a work team. She forms a team of students, brings them to the psychology lab, and has them play a game that requires teamwork. The game is rigged so that they lose regularly. The psychologist observes the students through a one-way window and notes the changes in their behavior during an evening of game playing. Can the psychologist generalize the results of her study to a team of employees that spends months developing a new product that never works right and is finally abandoned by their company? Explain.

No. She has “put together” a team of students. This suggests that there was no randomization involved. Besides that, students are likely to be in a different place in their lives than employees who are on the job for at least several months and likely much longer. Also, the disappointment associated with losing games during an evening is not likely to be equivalent to the disappointment felt after months of hard work.