Chapter 13:

Experiments and Observational Studies

Observational Studies

An ***observational study*** takes place when researchers don't assign choices they simply observe them:

For instance, a study trying to find a connection between students who play an instrument and academic performance. Instead of assigning some students to learn an instrument the researchers simply ***observed*** student who did and did not play an instrument and recorded their grades.

- This is also an example of a ***retrospective study*** because researchers first identified subjects who studied music and then collected data on their past grades.

Problems?

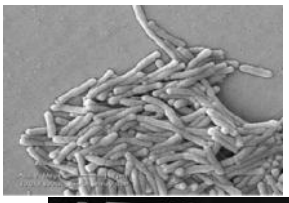
- Can you **conclude**, even if the study showed a connection, that playing a musical instrument improves grades?

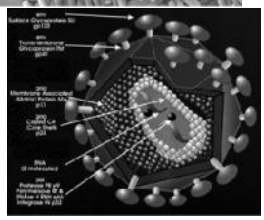
Some Lurking Variables:

- Students who play an instrument might have better work habits to start with, making them more successful in both.
- Music students may have more parental support (who paid for the instrument and lessons?)
- Maybe smarter kids just play musical instruments and it isn't the instrument playing that causes anything at all.

Good Uses

- Observational studies are valuable for **discovering trends and possible relationships**
- Observational studies can help discover variables related to rare outcomes, such as a specific disease when the study is also retrospective (The likely causes of both legionnaires' disease and HIV were initially identified from such retrospective studies)





Prospective Study

- Identifying subjects **in advance** and collecting data as events unfold is a ***prospective study.***
- For example, back to our musical students, the study might have started by selecting young students who have not begun music lessons. Then track their academic performance over several years, comparing those who later chose to study music with those who do not.

An Example! ☺ (With Dead Puppies)* ☹

- In earl
  usual
  develo
  died. I
  know
  obser
  invest
- Suppo
  to play
  cause
  your c
  prosp

**Statistics is sometimes bleak ☹*

Randomized, Comparative Experiments

- Is it ever possible to find evidence of a cause and effect relationship?
- Yes!
- Experiments!
- Say we take a group of third graders and randomly assign half to music lessons and forbid the other half to do so. Then we could compare their grades several years later.
- This is the kind of study design we are talking about when we talk about an ***experiment***.

Experiments

- Experiments require a ***random assignment*** of subjects to treatments.
- "Does taking vitamin C reduce the chance of getting a cold?
- Does working with computers improve performance in statistics class?
- Is this drug a safe and effective treatment for that disease?

Experiments

- An experiment is a study design that allows us to prove a cause-and-effect relationship.
- Experiments study the relationship between two or more variables.
- An experiment must identify at least one explanatory variable, called a ***factor***, to manipulate and at least one ***response variable*** to measure.
- The experimenter actively and deliberately manipulates the factors to control the details of the possible treatments and assigns the subjects to those treatments at random.
- An **experiment**:
  - ***Manipulates*** factor levels to create treatments.
  - ***Randomly*** assigns subjects to these treatment levels.
  - ***Compares*** the responses of the subject groups across treatment levels.

Experiment Terms

- Humans who are experimented on are commonly called ***subjects*** or ***participants***.
- Other individuals (rats, days, petri dishes of bacteria) are commonly referred to by the more generic term ***experimental unit.***
- The specific values that the experimenter chooses for a factor are called the ***levels*** of the factor.

Example

- We want to perform a sleep deprivation experiment to see whether this has any effect on test performance.
- The **factor** is the number of hours of sleep our participants (experimental units) receive. The **levels** might be 4, 6, or 8 hours of sleep.
- The **response variable** is the performance on some standard test we give each of our subjects.

Factors, Levels

- Often there are several factors at a variety of levels.
- Our subjects might also be assigned to a treadmill for 0 or 30 minutes
- The combination of specific levels from all the factors that an experimental unit receives is known as its ***treatment.***
- Our subjects could have any one of **six different treatments** – three sleep levels, each at two exercise levels.

Remember Our Example? ☹

- In 2007, deaths of a large number of pets were found to be linked to contamination from some brands of pet food. The manufacturer now claims that the food is safe but before it can be released, it must be tested.

- In an experiment to test whether the food is now safe for dogs to eat what would be the treatments and what would be the **response variable**?

SEE PART 1 FOR BETTER PPT

Our Example…

- The **treatments** would be ordinary-size portions of two dog foods:

  the new one from the company (the test food)

  and

  a food you knew to be safe

- The response variable would be a veterinarians assessment of the health of the test animals

Another Example…

Driving Study

Explanatory = ***hours of sleep & amount of alcohol***

Treatments: **normal sleep, no alcohol**

normal sleep, alcohol

reduced sleep, no alcohol

reduced sleep, alcohol

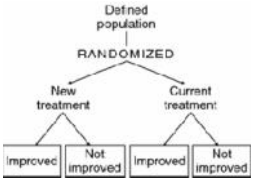
Response = ***Driving Impairment***

Drinking and Driving

Four Principles of Experimental Design

- **Control**
  We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups.

- **Randomize**
  Randomization allows us to equalize the effects of unknown or uncontrollable sources of variation. It is only by randomly assigning experimental units to treatments at random that we are able to draw conclusions from an experiment.



Four Principles of Experimental Design

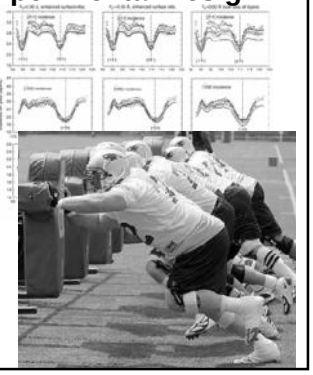
- **Replicate**
  We should apply each treatment to a number of subjects
  The outcome of an experiment on a single subject is an anecdote, not data.

  Replication of an entire experiment with the controlled sources of variation at different levels is an essential step in science: Your sleep deprivation experiment should be replicable in another part of the world or country with people of different ages and at different times of the year.

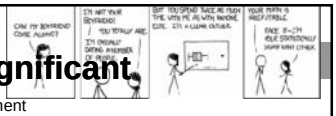
Four Principles of Experimental Design

- **Block**
  Sometimes attributes of the experimental units that we are not studying and that we can't control may nevertheless affect the outcomes of an experiment. If we group similar individuals together and then randomize within each of these **blocks** we can remove much of the variability due to the difference among the blocks.

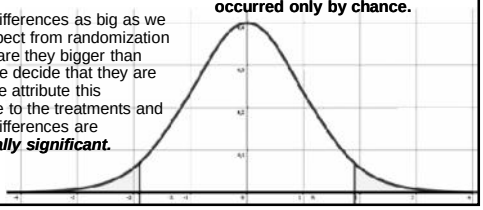
  Blocking is an important compromise between randomization and control but unlike the first three principles blocking is **not *required* for experimental design**.



Statistically Significant

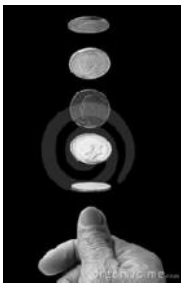


- If the difference among treatment groups are big enough, we will attribute the differences to the treatments.
- How can we decide whether the differences are big enough?
- Are the differences as big as we might expect from randomization alone or are they bigger than that? If we decide that they are bigger, we attribute this difference to the treatments and say the differences are ***statistically significant.***
- We'll talk more about statistical significance later on. For now, the important point is that **a difference is statistically significant if we don't believe that it's likely to have occurred only by chance.**



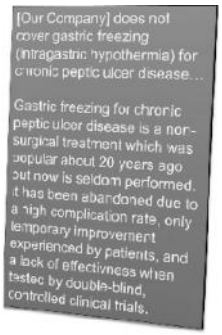
Example

- Say you flip a coin 100 times
- You expect, *on average,* 50 heads
- You get 54 heads out of 100 flips. This doesn't seem very surprising.
- What if you got 94 heads? Seems very out of the ordinary.
- What about 74?



Quick Review

- At one time a method called "gastric freezing" was used to treat people with peptic ulcers.
- An inflatable bladder was inserted down the esophagus and into the stomach, and then a cold liquid was pumped into the bladder.
- Now you can find the following notice on the Internet site of a major insurance company:



Quick Review

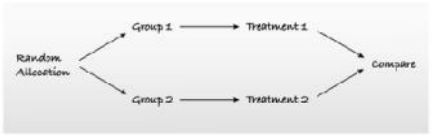
- What did that "controlled clinical trial" (experiment) probably look like? (Don't worry about the double-blind part, getting there)
- A) What was the factor in the experiment?
- **The factor was type of treatment for peptic ulcers**
- B) What was the response variable?
- **The response variable could be a measure of relief from gastric ulcer pain or an evaluation by a physician on the state of the disease.**
- C) What were the treatments?
- **Treatments would be gastric freezing and some alternative control treatment.**
- D) How did researchers decide which subjects would receive which treatments?
- **Treatments should be assigned randomly.**
- E) Were the results statistically significant?
- **No. The website reports "lack of effectiveness" indicating that no large differences in patient healing were noted.**

Experiments and Samples

- Sample surveys try to estimate population parameters, so the sample needs to be as representative of the population as possible.
- By contrast, experiments try to assess the effects of treatments.
- We want a sample to exhibit the diversity and variability of the population, but for an experiment the more homogenous the subjects the more easily we will spot differences in the effects of the treatments.

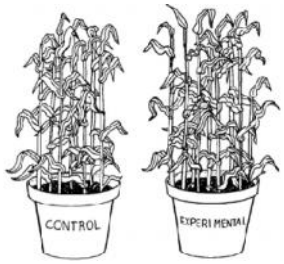
Diagram of an Experiment

- It's often helpful to diagram the procedure of an experiment.
- The following diagram emphasizes the random allocation of subjects to treatment groups, the separate treatments applied to these groups, and the ultimate comparison of results:



Control Treatments

- Example:
  Suppose you wanted to test a $300 piece of software designed to shorten download times. You could try it on several files and record download times but you probably want to *compare* the speed with what would happen *without* the software installed.
- Such a baseline measurement is called a ***control*** treatment and experimental units to whom it is applied are called a ***control group.***



Blinding

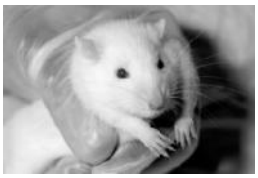
- Humans are notoriously susceptible to errors in judgment.
- All of us. ⊗
- When we know what treatment was assigned, it is difficult not to let that knowledge influence our assessment of the response, even when we try to be careful.

Blinding

- ***Blinding*** participants to a treatment is the process of intentionally disguising which treatment is which.
- This is why Coke and Pepsi do a "blind" taste test – to avoid brand loyalty.

Blinding

- But it isn't just participants who should be blind!
- Experimenters themselves often subconsciously behave in different ways that favor what they believe.
- Even lab technicians may treat plants or test animals differently if, for example, they expect them to die.



Blinding

- There are two main classes of individuals who can affect the outcome of the experiment:

  - **Those who could influence the results** (the subjects, treatment administrators, or technicians)

  - **Those who evaluate the results** (judges, treating physicians, etc.)

- When all the individuals in either one of these classes are blinded, an experiment is said to be ***single-blind***. When everyone in ***both*** classes is blinded, we call the experiment ***double-blind.***

More With Puppies

- In our experiment to see if the new pet food is now safe, we're feeding one group of dogs the new food and another group a food we know to be safe.
- Our response variable is the health of the animals as assessed by a veterinarian.
- Questions: Should the vet be blinded? Why or why not? How would you do this?
  Can this experiment be double-blind? Would that mean that the test animals wouldn't know what they are eating?

More With Puppies

- Whenever the response variable involves judgment it is a good idea to blind the evaluator to the treatments.
- The veterinarian should not be told which dogs ate which food.
- There is a need for **double-blinding**. In this case, the workers who care for and feed the animals should not be aware of which dogs are receiving which food. We'll need to make the "safe" food look as much like the "test" food as possible.

Placebos

- Often, simply applying *any* treatment can induce improvement.
- Some of the improvement seen with a treatment – even an effective treatment – can be due simply to the act of treating.
- To separate these two effects, we can use a control treatment that mimics the treatment itself.
- A "fake" treatment that looks just like the treatments being tested is called a ***placebo.***

Placebos

- Especially when psychological attitude can affect the results, control group subjects treated with a placebo may show an improvement.
- It's not unusual for 20% or more of subjects given a placebo treatment to report reduction in pain, improved movement, or greater alertness, or even demonstrate improved health or performance.

Placebo Effect

- The ***placebo effect*** highlights both the importance of effective blinding and the importance of comparing treatments with a control.
- You should use placebo controls as an almost essential tool for blinding whenever possible.
- The best experiments should be: Randomized, Double-Blind, Comparative, Placebo-Controlled

Does Ginkgo Biloba Improve Memory?

- Researchers investigated the purported memory-enhancing effect of ginkgo biloba tree extract (P. R. Solomon, F. Adams, A. Silver, J. Zimmer, R. De Veaux, "Ginkgo for Memory Enhancement. A Randomized Control Trial.")
- In a randomized, comparative, double-blind placebo-controlled study, they administered treatments to 230 elderly community members.

Ginkgo Biloba

- Treatments to 230 community members.
- One group received Ginkoba according to manufacturers instructions. The other received a similar-looking placebo. Thirteen different tests of memory were administered before and after the treatment.
- The placebo group showed greater improvement on 7 of the tests, the treatment group on the other 6. **None showed any significant differences**.

Blocking

- When groups of experimental units are similar, it's often a good idea to gather them together into ***blocks***. By blocking, we isolate the variability attributable to the differences between blocks, so that we can see the differences caused by the treatments more clearly.

Blocking

- For example, say we want to use 18 tomato plants for a fertilizer experiment but the garden store had only 12 plants left.
- So we drive down to a different store and buy 6 more plants.
- We worry that the tomato plants from the two stores are different somehow and in fact they don't even appear similar.

Blocking

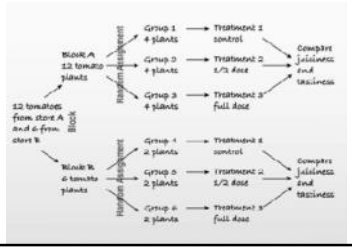
- How do we design an experiment so that the differences between the stores don't mess up our attempts to see differences among fertilizer levels?
- Here we would define the plants from each store to be a block.
- The randomization is introduced when we randomly assign treatments within each block.
- To isolate the store effect, we block on store by assigning the plants from each store to treatments at random.

Blocking

- We now have six treatment groups, three for each block.
- Within each block we will randomly assign the same number of plants to each of the three treatments.
- The experiment is still fair because each treatment is still applied (at random) to the same number of plants and to the same proportion from each store: 4 from store A and 2 from store B

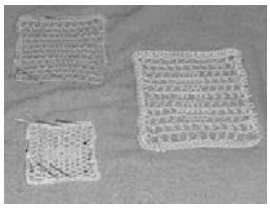
Blocking

- Because the randomization occurs only within the blocks (plants from one store cannot be assigned to treatment groups for the other) we call this a ***randomized block design***.
- In effect, we conduct two parallel experiments, one for tomatoes from each store, and then combine the results.



Blocking

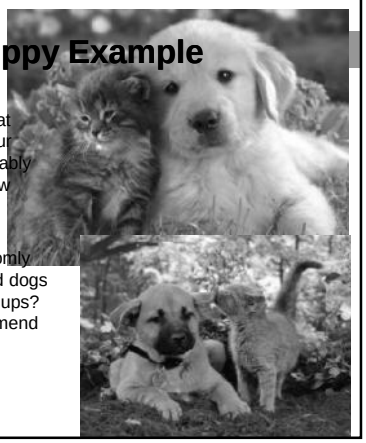
- Blocking is the same idea for experiments as stratifying is for sampling. Both methods group together subjects that are similar and randomize within those groups as a way to remove unwanted variation.
- We use blocks to reduce variability so we can see the effects of the factors; we are not usually interested in studying the effects of the blocks themselves.



Matching

- In a retrospective or prospective study, subjects are sometimes paired because they are similar in ways *not* under study.
- ***Matching*** subjects in this way can reduce variation in much the same way as blocking.
- For example, a retrospective study of music education and grades might match each student who studies an instrument with someone of the same sex who is similar in family income but didn't study an instrument.
- When we compare grades of music students with those of non-music students, the matching would reduce the variation due to income and sex differences.

Blocking: Puppy Example



- In 2007, pet food contamination put cats at risk, as well as dogs. Our experiment should probably test the safety of the new food on both animals.
- Why shouldn't we randomly assign a mix of cats and dogs to the two treatment groups? What would you recommend instead?

Puppy Blocking

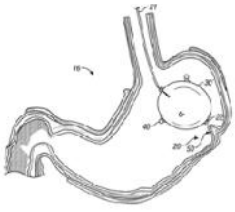
- Dogs and cats might respond differently to the foods and that variability could obscure my results.
- Blocking by species can remove that superfluous variation. I'd randomize cats to the two treatments (test food and safe food) separately from the dogs. I'd measure their responses separately and look at the results afterward.

Checking In

- Recall the experiment about gastric freezing, an old method for treating peptic ulcers that we talked about last class.
- A major insurance company doesn't cover this treatment because "double-blind, controlled clinical trials" failed to demonstrate that gastric freezing was effective.
- What does it mean that the experiment was double-blind?
- **Neither the patients who received the treatment nor the doctor who evaluated them afterward knew what treatment they had received.**
- Why would you recommend a placebo control?
- **The placebo is needed to accomplish blinding. The best alternative would be using body-temperature liquid rather than the freezing liquid.**

Checking In

- Suppose that researchers suspected that the effectiveness of the gastric freezing treatment might depend on whether a patient had recently developed the peptic ulcer or had been suffering the condition for a long time. How might researchers have designed the experiment?
- **The researchers should block the subjects by the length of time they had had the ulcer, then randomly assign subjects in each block the freezing and placebo groups.**



Adding More Factors

- There are two kinds of gardeners in this world:
- Those who make sure their plants are never dry and water constantly
- And those who let Mother Nature take its course
- The makers of OptiGro want to ensure their product will work on a wide variety of watering conditions.

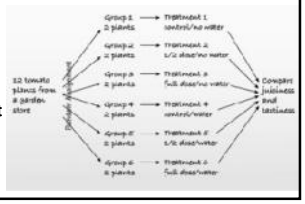
Adding More Factors

- Can we study a second factor at the same time and still learn as much about fertilizer?
- We now have two factors: fertilizer at three levels and irrigation at two levels.

	No Fertilizer	Half Fertilizer	Full Fertilizer
No Added Water	1	2	3
Daily Watering	4	5	6

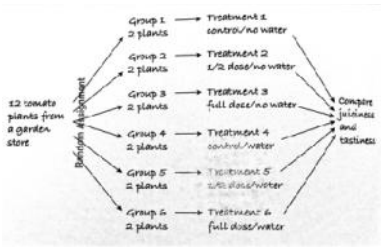
Adding More Factors

- With the original 12 plants the experiment now assigns 2 plants to each of these six treatments at random.
- The experiment is a **completely randomized two-factor experiment** because any plant could end up assigned at random to any of the six treatments (and we have two factors)
- **With two factors we can account for more of the variation**. That lets us see the underlying patterns more clearly.



Adding Factors

- Experiments with more than one factor are both more efficient and provide more information than one-at-a-time experiments.



Confounding

- Professor Stephen Ceci of Cornell University performed an experiment to investigate the effect of a teacher's classroom style on student evaluations.
- He taught a class in developmental psychology during two successive terms to a total of 472 students in two very similar classes.
- He kept everything about his teaching identical (same text, same syllabus, same office hours, etc.) and modified only his style in class.
- During the fall term he maintained a subdued demeanor.
- During the spring term he used expansive gestures and lectured with more enthusiasm – varying his vocal pitch and using more hand gestures.
- He administered a standard student evaluation form at the end of each term.

Confounding

- **The students in the fall term rated him only as an average teacher.**
- **Those in the spring term rated him as an excellent teacher, praising his knowledge and accessibility, and even the quality of the textbook.**
- On the question, "How much did you learn in the course?" the average response changed from 2.93 to 4.05 on a 5-point scale.

Confounding

- Now! How much of the difference he observed was due to his difference in manner, and how much might have been due to the season of the year?
- The fall term in Ithaca, NY (home to Cornell) starts off colorful and pleasantly warm but ends cold and bleak.
- The spring term starts out bitter and snowy but ends with blooming flowers and singing birds.
- Might students overall happiness have been affected by the season and reflected in their evaluations?

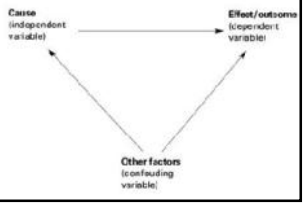
Unfortunately, there is no way to tell.

Confounding

- Nothing in the data enables us to tease apart the two effects, because all the students who experienced the expansive manner did so during the spring.
- When the levels of one factor are associated with the levels of another factor, we say these two factors are ***confounded***.

A ***confounding variable*** is …

- related to the explanatory variable (something that will exist more commonly if the explanatory variable does)
- affects the response variable



Puppppppppppies

- Would it be a bad design to feed the test food to some dogs and the safe food to some cats?
- Yes! This would create confounding. We would not be able to tell whether the difference in animal's health was attributable to the food they had eaten or to differences in how the two species responded.

A Two-Factor Example

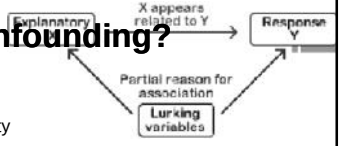
- Confounding can also arise from a badly designed multifactor experiment.
- A credit card bank wanted to test the sensitivity of the market to two factors: the annual fee charged for a card and the annual percentage rate charged.
- The bank selected 100,000 people at random from a mailing list.
- It sent out 50,000 offers with a low rate and no fee and 50,000 with a higher rate and a 50$ annual fee.

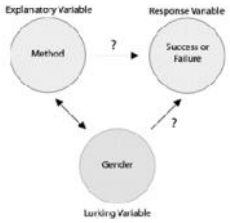
A Two-Factor Example

- What happened?
- No surprise…
- People signed up for the lower rate/no fee card at almost twice the rate of the other offer.
- But how much of the change was due to the rate and how much was due to the fee?
- There is simply no way to know.
- What should they have done?

Lurking or Confounding?

- Confounding and lurking variables are alike in that they interfere with our ability to interpret our analyses simply.
- However, there are important differences.
- A **lurking variable creates an association between two other variables** that tempts us to think that one might cause the other.



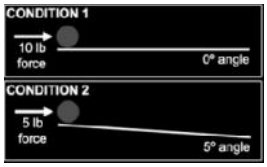


Lurking or Confounding?

- Lurking variables can come up when a lurking variable influences both the explanatory and response variable.
- Our example was that countries with more TV sets per capita tend to have longer life expectancies.
- We should not conclude that it is TVs "causing" longer life.
- We suspect instead that a generally higher standard of living may mean that people can afford more TVs and get better health care, too.

Lurking v. Confounding

- A lurking variable is usually thought of as a variable associated with both *y* and *x* that makes it appear that *x* may be causing *y*.
- A **confounding variable is associated in a non-causal way with a factor and affects the response**. Because of the confounding we find that we can't tell whether any effect we see was caused by our factor or by the confounding variable – or even by both working together.



Confounding

- Confounding can arise in experiments when some other variable associated with a factor has an effect on the response variable.
- In a designed experiment, the experimenter *assigns* treatments (at random) to subjects rather than just observing them.
- A confounding variable cannot be thought of as causing that assignment.
- It is worth noting that the role of blinding in an experiment is to combat possible sources of confounding. There is a risk that knowledge about the treatments can lead the subjects or those interacting with them to behave differently or could influence judgments made by the people evaluating responses.
- That means we do not know whether the treatments really do produce different results or if we are being fooled by those confounding influences.