Study Design

Robin Donatello

Last Updated 2017-10-28 10:21:38

Contents

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider how data are collected so that they are reliable and help achieve the research goals.

Populations and samples

Recall that each research question is designed to make a statement or learn something about a target *population*. A *sample* represents a subset of the cases and is often a small fraction of the population.

Consider the following research question.

> What is the average mercury content in swordfish in the Atlantic Ocean?

The target population is all swordfish in the Atlantic Ocean, and each fish represents a case. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. For instance, 60 swordfish in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

For the following two questions, identify the target population, what a sample could consist of, and what represents an individual case.

1. Over the last 5 years, what is the average time to complete a degree for Chico State undergraduate students?
2. Does a new drug reduce the number of deaths in patients with severe heart disease?

Next consider the following possible responses to the three research questions (RQ):

- A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
- I met two students who took 10 years to graduate from Chico State, so it must take longer to graduate at Chico State than at other colleges.

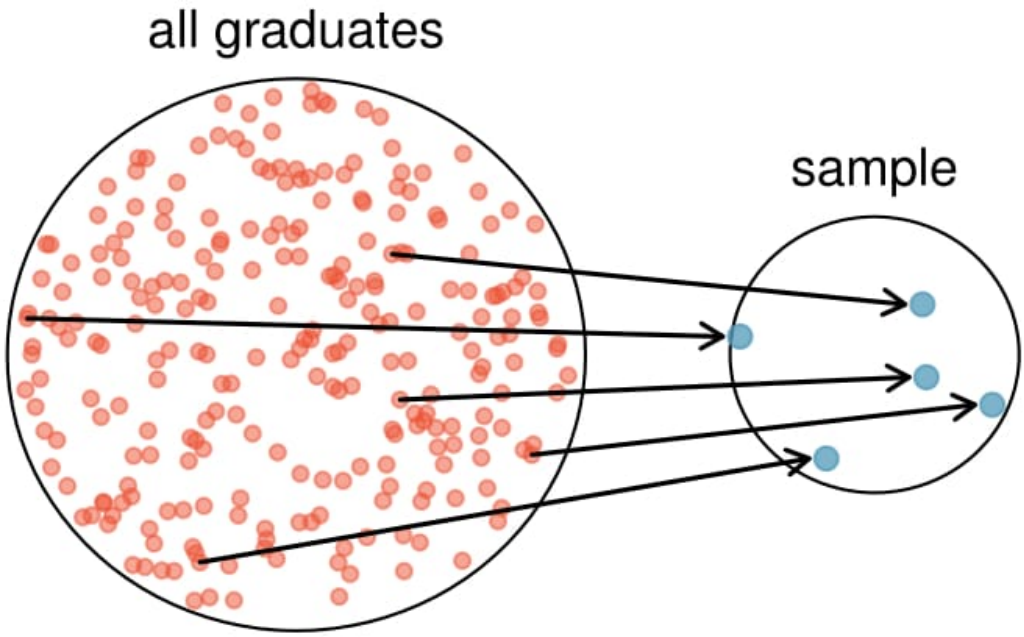

Figure 1:

- My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data. However, there are two problems.

1. The data only represent one or two cases.
2. It is unclear whether these cases are actually representative of the population.

Data collected in this haphazard fashion are called **anecdotal evidence**.

Such evidence may be true and verifiable, but it may only represent extraordinary cases.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 10 years to graduate than the 20 others who graduated in six years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

Sampling from a population

Lets try to estimate the time to graduation for Chico State undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population.

The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates.

Five graduates are randomly selected from the population to be included in the sample.

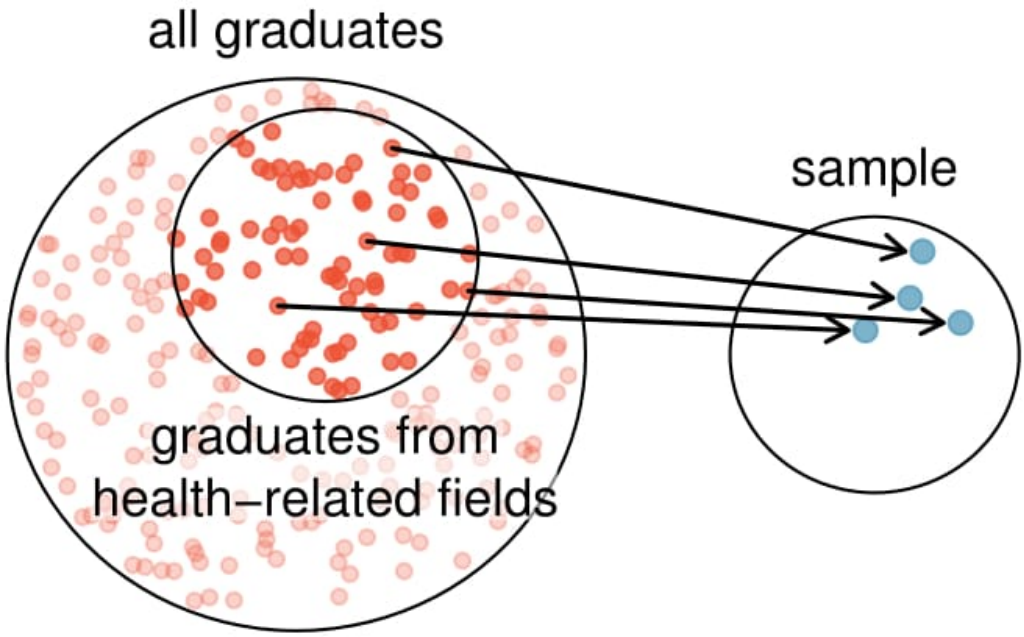

Figure 2:

Representation and Bias

Suppose we ask a student who happens to be majoring in nutrition to select several graduates for this study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces **bias** into a sample.

> A recent review of 74 studies of antidepressant agents found 38 studies with positive results and 36 studies with negative or questionable results. All but 1 of the 38 positive studies were published. Of the remaining 36, 22 were not published, and 11 were published in such a way as to convey a positive outcome.

Describe how this selective reporting can have adverse consequences on health care.

Simple Random Sampling

Sampling randomly helps resolve this problem. The most basic random sample is called a **simple random sample**, and which is equivalent to using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often. The act of taking a simple random sample helps minimize bias, however, bias can crop up in other ways.

Even when people are picked at random, e.g. for surveys, caution must be exercised if the **non-response** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are *representative* of the entire population. This **non-response bias** can skew results.

Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

Convienience Sample

The next type of sampling to discuss is called a **convenience sample**. This is where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

> We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?

Volunteer Sample (Opt-In)

In a volunteer sample, individuals choose to be included. In general, volunteer samples tend to be comprised of individuals who have a particularly strong opinion about an issue and are looking for an opportunity to voice it.

- A student posts a music-lovers' survey to campus announcements, asking people to vote for their favorite type of music.
- The Student Evaluations of Teaching (SET) that you have to fill out each semester for each teacher.

Such a sample is almost guaranteed to be biased.

Whether the responses obtained from such a sample are over- or under-estimated, and to what extent, cannot be determined. As a result, data obtained from a voluntary response sample is quite useless when you think about the "Big Picture", since the sampled individuals only provide information about themselves, and we cannot generalize to any larger group at all.

However, a volunteer sample is not so problematic for a study conducted for the purpose of comparing several treatments.

In some cases volunteer samples are the only ethical way to obtain a sample. In medical studies, for example, in which new treatments are tested, subjects must choose to participate by signing a consent form that highlights the potential risks and benefits.

Clustered Sampling

Cluster sampling is used when our population is naturally divided into groups (which we call clusters).

- All the students in a university are divided into majors
- all the nurses in a certain city are divided into hospitals
- all registered voters are divided into precincts (election districts).

In cluster sampling, we take a random sample of clusters, and use all the individuals within the selected clusters as our sample.

> In order to get a sample of high school seniors from a certain city, you choose 3 high schools at random from among all the high schools in that city and use all the high school seniors in the three selected high schools as your sample.

Ask your professors for email rosters of all the students in your classes. Randomly sample some addresses and email those students with your question about musical preference.

Here is a case where the sampling frame (list of potential individuals to be sampled) does not match the population of interest.

- The population of interest consists of all students at the university, whereas the sampling frame consists of only your classmates.
- There may be bias arising because of this discrepancy.

  Students with similar majors will tend to take the same classes as you, and their musical preferences may also be somewhat different from those of the general population of students.

It is always best to have the sampling frame match the population as closely as possible.

Identifying problems with sampling strategies

A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Three research strategies for collecting data are described below. In each, describe any potential problems and bias you might expect.

1. He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
2. He gives out the survey only to his friends, and makes sure each one of them fills out the survey.
3. He posts a link to an online survey on his Facebook wall and asks his friends to fill out the survey.

Study Design

There are two primary types of data collection: **observational** studies and **experiments**.

Observational Study

The researcher simply monitors and collects data on things as they are. There is no manipulation of the study by the researcher.

- The Youth Risk Behavior Surveillance System (YRBSS) monitors six types of health-risk behaviors that contribute to the leading causes of death and disability among youth and adults.
- Measurements are taken on post-spawn carcasses of Chinook Salmon in Butte Creek to assess the annual population health.
- We collect a representative sample from the population of smokers who are just now trying to quit by using a nationwide telephone survey of 1,000 individuals. We ask them how they are trying to quit and classify it into one of four groups: 1) Drugs that alleviate nicotine addiction; 2) Therapy that trains smokers to quit; 3) A combination of drugs and therapy; or 4) Neither form of intervention (quitting "cold turkey"). One year later, we contact the same 1,000 individuals and determine whether they succeeded in quitting.

In general, observational studies can provide evidence of a naturally occurring association between variables, but **they cannot by themselves show a causal connection.**

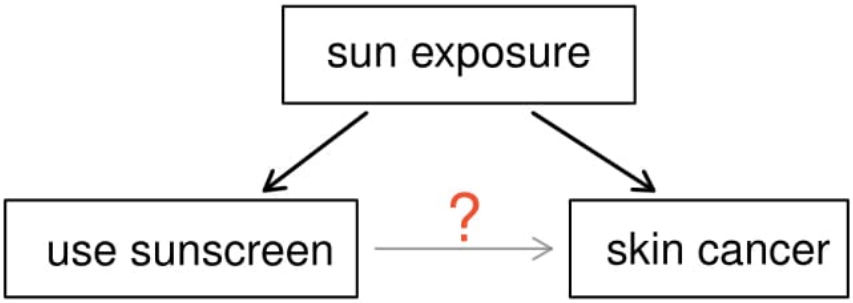

Figure 3:

Experiment

In a controlled experiment, the researcher controls the value of the explanatory variable for each unit. In other words, the researcher controls which subjects go into which treatment groups.

- We may suspect that diet and exercise reduce mortality in heart attack patients over the following year. Researchers collect a sample of individuals who have recently experienced a heart attack and split them into groups. The individuals in each group are assigned a treatment, one group per level of the explanatory variable. Patients are followed over a year and record the number of deaths per group.
- To study the effect of tar contained in cigarettes researchers (Wynder 1953) painted tobacco tar on the back of some mice but not others, and observed if the painted mice had cancer at a higher rate than those not exposed to the tar.
- We collect a representative sample from the population of smokers who are just now trying to quit by using a nationwide telephone survey of 1,000 individuals. We divide the sample into 4 groups of 250 and assign each group to use one of the four methods to quit. One year later, we contact the same 1,000 individuals and determine whose attempts at quitting succeeded while using our designated method.

The value of such control is that **cause-and-effect relationships can be established** between the response and explanatory variable.

Confounding and Lurking variables

> Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen AND more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.

Sun exposure is what is called a **confounding variable**,(a.k.a **lurking variable**, **confounder** or a **confounding factor**.) which is a variable that is correlated with both the explanatory and response variables.

We could control for the lurking varible of sun exposure by collecting information on the individuals and categorizing their amount of sun exposure as *high* and *low* exposure, then we could run a stratified model to see if the relationship between sunscreen use and skin cancer is the same between each exposure group.

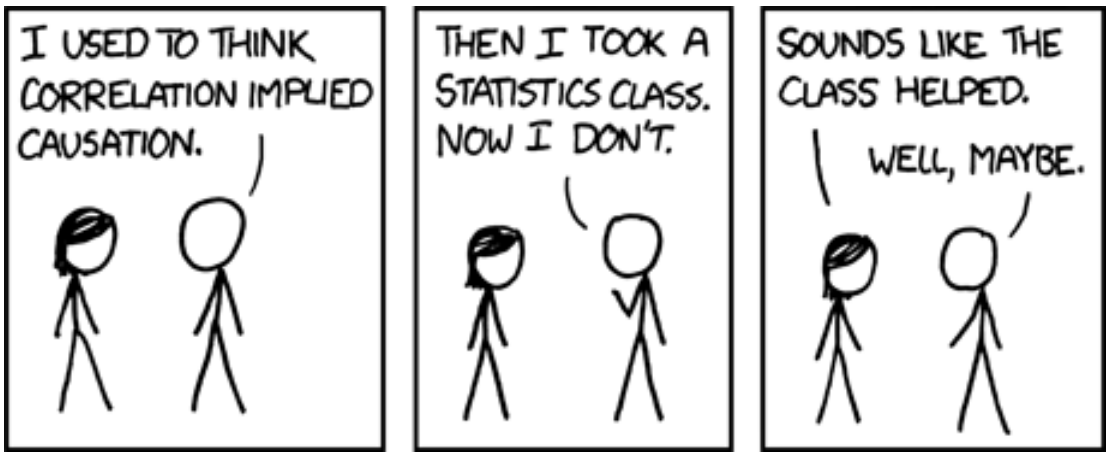

Figure 4:

- But then what if race/ethnicity tends to drive how much sun exposure they have? Or sunscreen use? It is easy to imagine that those with very light skin color will either tend to use suncreen more, or stay out of the sun more than those with a darker skin color

While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured. It is because of the existence of a virtually unlimited number of potential lurking variables that we can never be 100% certain of a claim of causation based on an observational study.

Confounding is a major threat to the validity of inferences made about statistical associations. In the case of a confounding variable, the observed association with the response variable should be attributed to the confounder rather than the explanatory variable. We test for confounders by including these additional in our statistical models that may explain the association of interest.

> In other words, we want to demonstrate that our association of interest is significant even after controlling for potential confounders.

We do so by building a **Multivariable Model**.

Types of multivariable relationships

How can I keep them all separate?

- The exposure works through the *mediator*
- The *confounder* affects both exposure and outcome
- The *moderator* changes the relationship between the exposure and the outcome.
- Everything else is a *covariate*

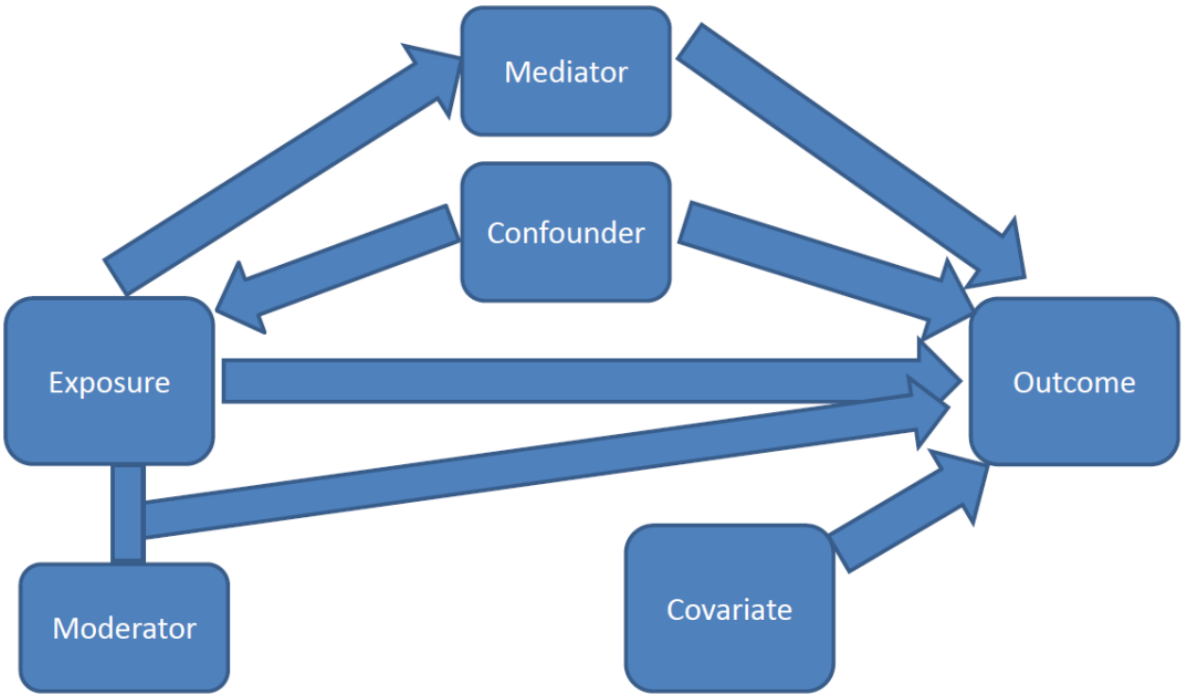
Mediator
Confounder
Exposure
Outcome
Moderator
Covariate

Figure 5: