6.1 Describe, Predict, or Explain

6.1.1 Describe, Predict, or Explain: Overview

This element addresses the following learning objective of this course:

LO4: Justify an analytic approach that informs decision making.

Let's imagine you work for a petroleum company. On the one hand, if you want to describe production, you might work on an observational study that uses time series data to describe historical patterns. On the other hand, if you want to figure out if we can increase production via the use of different preventative maintenance schedules, you might design the experiment to randomly assess this new strategy in some wells.

These approaches are not mutually exclusive. You could design a project to describe the historical conditions. You Realize that production is down, and then you want to explore how this new maintenance schedule can help eliminate costly shutdowns when equipment breaks. You test this hypothesis that a new preventative maintenance schedule can reduce the number of shutdowns and find out this new schedule works.

And then prediction comes in. Based on your experiment, you're pretty confident that this new maintenance schedule will save money. But you don't know which rigs to fix first. So you employ various methodologies to predict which rigs will fail first. Now, remember, you may not always have this kind of nice, nested type of descriptive, explanatory, and predictive research. These approaches can exist in isolation.

In closing, I want to remind you about the qualitative and quantitative approaches that can inform these various types of inference. For example, you may want to talk to mechanics and engineers in the field to better inform your project.

6.1.2 Descriptive Research

This element addresses the following learning objective of this course:

LO4: Justify an analytic approach that informs decision making.

Both qualitative and quantitative approaches can help us answer the "what's going on" question, which is the focus of descriptive research. Let's say we work for the San Francisco Department of Public Health. And you are tasked with identifying the population of residents that are unhoused.

Who are they? If you put your quantitative hat on, you might want to improve the way that the city measures the prevalence of homelessness in a community. You may use survey methods to better understand the demographics of this population.

If you put your qualitative hat on, you may want to talk to those who are homeless to better understand who they are.The demographic information one gathers from interviews will probably better capture people's stories about how they became homeless compared to structure and surveys. Now, the goal isn't to pit one approach against another, but rather the goal is to recognize that each approach can answer a different aspect of a question.

Through your discussions, you might realize that you need to use some kind of block randomization when you design your experiments because there are distinct types of wells that we as data scientists back at headquarters didn't know about. We were just going to lump all the wells together. And that's why you need to talk to people.

6.1.3 When Is Prediction Enough?

When Is Prediction Enough?
This element addresses the following learning objective of this course:

- LO4: Justify an analytic approach that informs decision making.

When might we focus exclusively on prediction? Theory and hypothesis testing may not apply to some of the projects you work on. Again, sometimes all you want to do is predict. Let's go through some examples.

Maybe we care about how to predict what video, song, or advertisement to recommend to a user. Maybe we care about how to predict whether or not a credit card transaction is fraudulent. Maybe we care about how to predict whether or not online content is appropriate. Oftentimes, we don't care about why the prediction is correct. We just care about accuracy and reliability.

6.1.4 When Is It Necessary to Understand the Causal Mechanism?

This element addresses the following learning objective of this course:

- LO4: Justify an analytic approach that informs decision making.

[? Comparative ?] predictive approaches-- when do we need to know why something is going on? When do we need to engage in explanatory research? Maybe when the stakes are higher, we should understand the mechanism.Perhaps it's important to understand the mechanism when we want to recommend the best medical treatment for a patient that's really sick. Perhaps it's important to understand the mechanism when we want to recommend whether or not someone gets a bank loan.

Or maybe the mechanism doesn't matter if we could predict with great accuracy and reliability. As long as we're getting it quote unquote "right" often enough, maybe we're OK with the fact that we don't know the mechanism. Or Maybe we should understand the causal mechanism if the consequences of getting it wrong are significant.

Here, we're talking about false positives and false negatives. If we recommend a wrong movie for me to watch next on my streaming service, who cares? It's probably OK. But if a business will fail because we didn't extend a line of credit to them, that's a different situation. If we are providing recommendations about the likelihood that someone in the criminal justice system will reoffend when they're released, and we get that wrong, that can have a real impact on the conditions of release. It may be more important to understand the mechanism, simply put, when the consequences of getting it wrong are grave.

We also may need to understand the mechanism when we want to change behavior. Maybe we need to understand the mechanism so we know which lever to pull. Or finally, maybe we need to be able to explain when our client demands an explanation. But just be aware here-- here's a small word of caution. Be aware that a client may pushback about prediction versus explanation, based on how much they agree with what you find. They might say, nah,we don't agree with this because we don't understand the mechanism. And your reply might be, but we didn't designthis to understand the mechanism. Or they may say, we love this, even though we don't understand the mechanism because we like what we saw.

6.2 Predict vs. Explain

Spend five minutes on the following discussion prompt.
Think of your industry, or think of a domain you care about.

- Write down the domain
- What is a situation where prediction is sufficient? Why?
- With the same topic in mind, when is it important to understand the causal mechanism? Why?

6.3 Look at Both Successes and Failures

In my view, the killer app for data scientists when it comes to changing people's minds and getting stuff done is having a really tight research design. I think it's the single best way to save time and effort in answering the question that you need to answer. And it, kind of like a force multiplier, it just multiplies your effectiveness across the board. People used to say there's gold out there and then there are hills. And that's true of most data sets. But just like in mining, the difference between a really successful mining company and a bankrupt mining company is how they design their system for getting the gold out of the hills. It sounds obvious, and we all know it. But it is amazing how often really smart, really capable people just screw this up and don't pay enough attention to the research design at the front of the project. Give you an example. Literally tens of times in my career, I've seen really fantastic graduate students go out to the field to collect data. And sometimes, they go to places that are actually pretty hard or dangerous to get to. So one example, I had a student who went to Bratislava for a year. It's not necessarily the best place to spend February and March. And she came home with boxes and boxes of data. And then, guess what she had? Boxes and boxes of data. A couple months later, she called me and said, oh my god, what am I going to do with all this data? Well, actually, in a really significant way, it wasn't data. She didn't know what her research design was. She didn't know why she was collecting that data. And you know what? She had to go back to Bratislava for another year. And actually, as we all know, even when data becomes cheaper and easier to collect, it's still a distraction, unless it's organized for a specific purpose. So now, I want to tell you another story, which is in some ways even more distressing. I had a client, an important client, at a major foundation that wanted to run a competition about urban resilience. They were concerned with trying to make cities more resilient to natural disasters. And so, what did they ask me to do? They said, look, we want to design a competition. And we want that competition to generate the best ideas that could possibly out there about urban resilience. And so in order to help us kind of figure out how to design that, could you go out there and look at five or six recent competitions that got a lot of press attention and tell me what are the key design criteria for us to use? Because competitions are all the rage these days in generating new ideas. So what did I do? Well, I tried to explain to them the problem in their implied research design. By the way, they didn't like hearing this, as any client wouldn't. But I was able to convince them, look, here's one problem. Your sample size is too small. Five or six recent competitions, that's not enough to really know what works. So could we go for a bigger sample size? And they agreed, yeah. OK. So that sounds like a good idea. Let's get a bigger sample size. Then, I pressed them further and said we got a nasty kind of selection bias going on here. So you've asked me to look at the competitions that got a lot of press attention, but maybe those aren't the ones we should be looking at. Maybe they got press attention, not because they were good competitions or generated good ideas, but because the people who were running them had hired a really excellent public relations firm. So maybe that's not the right set for us to look at. The data set's biased. And these are not statistics geeks, but they got that one too. But here's where we just ran into a wall-- with the core research design issue. Look, I tried to explain to them, you cannot possibly determine the causes of success by looking only at successful cases. It's like asking what explains high market capitalization by comparing Exxon to Apple. They might share some traits, but unless we look also at Yahoo, we couldn't possibly know if those are the traits that actually matter. Every person who has taken basic statistics knows this. Every scientist knows this. But some of the people that you'll be working for and some of the folks to whom you'll have to explain your findings just will not be able to grasp this. And despite the fact that I spent literally an hour on the phone with this client, they just couldn't grasp that you couldn't determine the causes of success by looking only at successful cases. And as you can imagine, my blood pressure was going up, my heart rate was going up, and I was sweating. But at the end of the day, it was a core research design issue. And there was absolutely nothing I could do about it. So here's the thing. We know proper inference, proper inference from data is not always intuitive. We know that the brain resists training in the proper rules of inference. That's why statistics courses are so hard. And we're just not born with those statistics in our head. And even great statisticians sometimes make the same mistakes when they're out in regular life. Even very smart people sometimes make very silly mistakes. What's the classic example of this? Look at any stock market technical analysis filled with this kind of thing. And gazillions of dollars depend on it. Let me show you my favorite example in this lovely slide of the Dow Jones Industrial Average-- or, excuse me. Actually, this is the S&P 500. Well, look at the data. Every time the S&P 500 breaks 1,500, if you look at the score of the Super Bowl that year and you add the numbers together, you will get precisely the percentage by which the S&P 500 will then fall within about a year and a half. So if you're worried right now, you've got to be upset about what happened in the 2012 season Super Bowl. In early 2013, the Ravens defeated the 49ers by 34 to 31. That predicts a 65% fall in the S&P 500. There's the data. You tell me if people aren't going to bet on that. You know they will.

6.4 Variables

6.4.1 Conceptualize

This element addresses the following learning objectives of this course:

- LO3: Assess and select specific data and the data collection methods that best fit a specific outcome or need.

- LO4: Justify an analytic approach that informs decision making.

Imagine you work for the National Department of Health and your working research question is what are the effects of isolation on health? This is a pretty clear research question, but let's do a little bit of conceptualization.Conceptualization is when we refine abstract concepts, when we make imprecise concepts more specific. It's like moving from a really high level of abstraction to something that's a little bit more concrete.

Let's go through an example. Your question is what are the effects of isolation on health? Well, what type of isolation are we talking about? Are we talking about isolation in hospitals to prohibit the transmission of diseases? Are we talking about the low frequency of social interaction with others? And what type of health are we talking about? Are We talking about isolation in hospitals, in which case maybe we're talking about physical health? But if we're talking about social interactions with others, maybe we're talking about mental health, or maybe we're talking about both.

And so here we've gone through a process of refinement. In this example, kind of either through internal dialogue orvia a chat with your colleagues, you've come to realize that you want to focus on social isolation and the effects of mental health. And so your revised question is what are the effects of social isolation on mental health?

6.4.2 Operationalize

This element addresses the following learning objectives of this course:

- LO3: Assess and select specific data and the data collection methods that best fit a specific outcome or need.
- LO4: Justify an analytic approach that informs decision making.

Imagine you want to study the effects of social isolation on mental health. Let's operationalize. Operationalization is how we move from concept to measure. Here we want to look at the effect of social isolation on mental health. How Can we measure social isolation? Is it a dichotomous measure? There is social contact or there is not? Is it the number of social interactions? Is it the quality?

Can we simply ask respondents whether or not they feel socially isolated? Can we really just ask them that straight up? Maybe there's an established series of questions that capture the degree of social isolation without having to directly ask the respondent if they feel isolated.

This nondirect series of questions might be a preferred approach because A, maybe respondents don't know themselves if they feel socially isolated, or B, there might be some social stigma of directly saying they do feel isolated.

What about mental health? How can we measure that? Do we want to focus on whether or not people feel lonely?Whether or not they're depressed? What metric do we want to focus on? Do we want to focus on feelings? Or what about the number of interactions with medical professionals?

For example, if a subject currently under the care-- excuse me. For example, is a subject currently under the care of mental health professional? That could be a relatively clean, dichotomous measure. But the individual may feel like they can't answer a question about mental health services honestly because of social stigma.

Or unless we ask really specifically why they're seeking medical services, we might unintentionally conflate seeking a therapist with feeling lonely or depressed. There may be many other reasons individuals seek guidance from therapists.

Or if we only focus on whether or not individuals seek medical attention, not only have we set the bar pretty high for what we categorize as an indicator for the presence of mental health concerns, but we may underestimate the social isolation because there are likely individuals who have similar symptoms, but chose not to seek medical help.

As you can see, this can get pretty complicated pretty fast. And that's OK. The idea is to think early and to have these tough conversations about measurement. You want to have these at the design phase of the project and not once you've started collecting the data, and you realize, oh, man, this is not really what we want to measure. Now,depending on the task, you may want to consult with experts on definitions and on how to operationalize the ideas.

Let me step back real quick before I close this video. Let's talk about how this project on social isolation and mental health might benefit from both qualitative and quantitative approaches.

First, we might do an unstructured interview with public and mental health experts to make sure that we're thinking about this question in the right way i.e. conceptualization. Then we might consult with those same experts in the future to think about how to measure those concepts, operationalization.

Second, if we want to survey study participants, we might have both open-ended and close-ended questions. And we could apply quantitative methods to the close-ended questions. And we could apply either or qualitative or quantitative methods to the open-ended questions.

Third and finally, we may want to use this survey as a focal point for our study. We may want to conduct in-depth interviews, though, with a small number of respondents to better understand the lives of people in the study.

6.5 Conceptualize Company Success

Spend five minutes on the following prompt:

- What are three or four ways to conceptualize "company success?"

Do not worry about how to measure it yet, just think at the conceptual level. In the next question, you will be prompted to think about measurement.

6.6 Operationalize

Spend five minutes on the following question:

Go back to the previous question and review the ways your peers conceptualized "company success." Choose one peer's response and reply to their post with variables you could use to measure company success. Think of two ways to measure the same concept. Use the following framework:

Concept 1 => Measure A

Concept 1 => Measure B

Example:

Revenue => Quarterly earnings

Revenue => Change in quarterly earnings from last quarter

6.7 What Can We Measure?

6.7.1 What Concepts Do We Capture?

This element addresses the following learning objectives of this course:

- LO3: Assess and select specific data and the data collection methods that best fit a specific outcome or need.
- LO4: Justify an analytic approach that informs decision making.

What concepts can we capture? How do we measure things that are difficult to measure? Or rather how does measurement inform our research design? The goal is to encourage you to spend time on conceptualization, especially if you want to measure concepts that are frankly tough to measure.

Let's say you care about measuring absolute wealth. We could ask people close-ended questions on annual income, income plus savings,income plus savings and stock, real estate, whether you own or rent, et cetera. And we could reasonably expect our respondents to know what these questions mean.

If instead you care to measure perceived wealth, this concept is a little bit more difficult to measure. The goal of this idea is to capture one'sbelief about how they see themselves compared to their peers. But here we may not be able to ask close-ended questions similar to those that we asked to measure absolute wealth. We can ask, hey, relative to your peers in your neighborhood, do you perceive your wealth to be greater than your peers?

But compared to when we ask someone about income, i.e. how much money do you make, it's unclear if respondents will understand the intent of this question that wants to capture relative wealth. It might be tough to place your relative wealth compared to your peers. So maybe you could take a more quantitative approach to get at absolute wealth. And maybe you use a more qualitative approach to get at perceived wealth.

So a separate but related concept is non-response and response bias. I think it's particularly relevant to this example that we just explored.Now, the length of a survey may affect one's willingness to answer all the questions. You might get drop-off because people simply get tired of your survey.

Or if you ask sensitive questions, you might get a non-response because the respondent may intentionally skip the question because they feel there's a social norm that affects their ability to answer the question honestly. If you ask about wealth, someone may not want to answer.

And even if they do, you might observe response bias because the respondent may not answer truthfully because they are self-conscious about discussing salary. In some cultures, it's not appropriate to boast about how much money you make, or you might be embarrassed about how little money you make.

While we may not have to worry about our machines-- let's say we only work with machines-- we may not have to worry about them not reporting a metric or trying to deceive us. But if you work with machines, we still have to think about conceptualization. There are many ways to define improved efficiency.

So in summary, what we care about are concepts. And we have to think about whether or not what we measure captures those concepts we care about. So define concepts as early as you can. Enumerate out all the ways you can measure the concepts. And be honest about what you can and can't capture. Think early so you could design the most appropriate study.

6.8 Relationship Between Variables

6.8.1 Think Early About How Variables Are Related

This element addresses the following learning objectives of this course:

- LO3: Assess and select specific data and the data collection methods that best fit a specific outcome or need.
- LO4: Justify an analytic approach that informs decision making.

Let's bring together a few concepts. So when we theorize and conceptualize, we're forced to think about how concepts are related. When we operationalize, we translate those concepts into measurement. And when we articulate a hypothesis, we think about in a very concrete way how the variables are related.

Now, up to this point, I've presented these as three distinct steps. But really, our theory about how concepts are related will inform our expectations about how measurables or variables are related and vice versa. So let's go through an example of how we put all these steps together.

Let's say we want to explore a potential causal relationship between two concepts that are highly correlated--customer churn and decreased engagement with our platform. Our theory is that decreased engagement causes churn. So here's your working story. You observe a decrease in customer engagement before a user leaves.

Here's the challenge, though. Is their decreased engagement with the product the reason they leave? Or does the customer already know that they're leaving the platform for whatever reason, and they reduce their engagement because they know they're on their way out? Which is it? This is a pretty tough problem. So let's walk through loud whether or not an experiment would be appropriate here. It may or may not be.

So when we use the experimental method, just as a refresher, we randomly assign subjects to either a treatment or control condition. If we do the randomization correctly, we can say pretty confidently that any difference in outcome between the treatment and the control group is due to the intervention. This method is most appropriate to determine causal relationships. But depending on the context, it's not always available.

In the context described above where I'm describing disengagement and user churn, it's not exactly clear how you could set up an experiment. You can't randomly assign some users to disengage with the platform to see if differences in the assigned level of engagement have a differential impact on churn. This is just not realistic, and you can't force people how to feel.

Let's think about a similar example where observational methods can help inform a subsequent experimental intervention. Let's say you work for an online company that relies on its users to create content. Many of your content creators you find are leaving the platform. And your hunch is that many of them are burning out. The business problem is that you need productive content producers to help ensure that customers return to your site.

So you step back and think, OK, can I predict when a content creator will leave? And if I can predict content creator churn in stage one of my research, then maybe in stage two, I could develop different experimental interventions to try to keep the content creators on my platform.

Now, without this first stage of prediction, your intervention to avoid churn may be ill timed. So in this particular example, stage one would predict churn, and stage two, design an experiment to test the effectiveness of different interventions on churn. And so the punchline here is to spend time early in the design process to think about how variables are related.

6.8.2 Convince the Audience

This element addresses the following learning objective of this course:

- LO4: Justify an analytic approach that informs decision making.

Once you've convinced yourself that the variables you will measure in fact capture the concepts you care about, then you have to convince your audience. The argument you could put forward is as follows. This is what our measure captures, and this is how we know that. That second step is critical.

You can help convince others of the validity of your measures with the convergent and predictive validity framework.You demonstrate convergent validity when you show that your measure is highly correlated to a similar measure that we already accept as valid.

For example, let's say we work in agriculture. And we have a new metric to identify the quality of the crop yield. To Demonstrate convergent validity, we would show that this measure is highly correlated with existing measures oncrop quality. You demonstrate predictive validity when you show that the measure can predict something we think valid measures should be able to predict.

Let's go with the same example. We know that existing crop metrics can predict how much customers will pay for the crops. If we can also demonstrate that the new measure can also predict crop prices, then we've demonstrated predictive validity. The one-line takeaway is that you can't just say this is what our measure captures. You need to show it.

6.9 Two-Stage Design

Spend five minutes on the following prompt:

Imagine you work for a major producer of mini potatoes in California. The company has three tiers of potato quality. The current business problem is that potatoes do not always get sorted appropriately. In particular, too often, Tier 2 potatoes get graded as Tier 3. The company is leaving money on the table since they can charge more for Tier 2 potatoes than Tier 3.

The company knows that you recently enrolled in a data science program. They are convinced that there is a two-stage solution to this problem, but need help thinking through this. Please respond to your manager's email below.

TO: You
FROM: boss
RE: Two-stage approach
Hi,
I'm following up on our in-person discussion. I'm convinced there is a two-stage approach to our sorting problem. I sketched what I think are the two stages below. Can you send back a couple sentences on your thoughts? Am I even close?
Stage one: Identify when sorting issues occur the most.
Stage two: Implement different interventions to try to minimize sorting issues.
Best,
Your Awesome Boss

6.10 Sampling

6.10.1 Population vs. Sample

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

The population in a study is the broad group of entities. It could be people or non sentient beings that you want to make general conclusions about. The sample is the subset of units that you will include in your study. In many cases,it's simply not feasible to work with the entire population. This can be because it's prohibitively expensive to do that. It would take too long. You can't, or because you have specific reasons not to use the entire population.

For example, let's say you want to test a new intervention. You may not want to include the entire population of users in your study in case it goes poorly. Or if the intervention is relatively invasive, you may want to preserve some of the population members for future studies because it's possible that exposure to one intervention may influence future work.

We'll talk about sampling strategies in another video. But the punch line is we must be mindful of how each sampling approach does or does not allow us to generalize to the population.

6.10.2 Sampling Frame

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

The sampling frame is the list of units from which you will sample from. Ideally, this is a list that includes the entire population. But oftentimes, you won't have a list that includes everyone in the population of interest.

For example let's say you want to reach out to Berkeley graduates to see what their post-graduation salary is. You may not want to burn all your goodwill and email or call every graduate. So instead, you decide to sample 10%. At this point, you're pretty confident that you're good to go. You already have the emails and phone numbers.

But then you decide you want to text graduates instead because you have a hunch that their response rate will be a little bit higher. Then you determine which phone numbers are cell phones, and now you have your sampling frame.So you're sampling frame was reduced from all graduates, to those with phone numbers, and then finally graduates who had valid cell phone numbers. Now you could sample from that last list.

Let's go through another example of how to create a sampling frame. Let's say you work for a power company, and you want to experiment on a new monitoring strategy of power transformers. Your population is all transformers. Nowyour sampling frame could be the entire population, or maybe you only want to focus on a subset of non critical transformers.

The takeaway here is that if you ultimately want to generalize back to the population-- which is often what we want todo-- you're sampling frame should look pretty close to the population. If it's not, be explicit about that, and be very clear about how it's different. And describe how you think any differences between the way your population looksand the way you're sampling frame looks will influence what conclusions you can draw.

6.10.3 Selection Effect: How Observations End up in a Sample

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

How do observations from your sampling frame end up in your sample? Is there anything systematic about how an observation makes its way into your sample? If selection into the sample is at random, then you're pretty much OK.But still keep in mind a higher level concern of how observations end up in the sampling frame.

Now let's think about a situation where people enter your sample in a process that is not at random. Imagine you're engaged with the following research question. What is the impact of private school education on the caliber of college one attends?

We know that we need to gather information from students that go to both public and private schools because we know can't determine the effect of private school education by only looking at students who go to private schools. So our sampling frame is students who attend private schools and students who attend public schools.

Now, is there any systematic process that determines what school one attends? Or is it completely at random? If it's completely at random, then any difference we see in performance between students who attend private and public schools can be attributed to the difference in the type of education.

But we know, as is the case with many social processes, that what school you attend is far from random. We know family resources help determine what kind of school one attends. Private school tuition isn't cheap.

If we want to determine the impact of the type of school children attend, we need to be pretty clever about what other variables we should collect to account for confounds. For example, we probably need to collect parental income because existing theory suggests that socioeconomic status influences both one's propensity to go to private school and one's propensity to attend a prestigious college.

6.10.4 Selection Effect: Response Rate

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

Imagine that we've already identified a sampling frame. And we start contacting users. What determines whether or not someone replies to you? Again, if it's at random, you're probably good, but that's very unlikely. It's problematic if the response rate is correlated with the characteristic you care about.

Let's imagine you want to better understand future salaries of Cal graduates. You send out surveys. You're relatively happy with the response rate. And you begin some preliminary analysis. You see that the average salary was higher than you expected. You know that Cal Bears are awesome, but you didn't expect such high salaries. So what could be happening here?

It could be that grads, in fact, make a lot of money. It could also be that individuals over-report how much money they make. Or it could also be that individuals who don't make that much money either didn't reply to the survey at all or skipped that question. One way to help ensure that you're getting responses that are representative of your population is to offer incentives for people's participation and follow up multiple times to encourage them to participate.

Now let's think of a response rate in a different example in a slightly different context. Let's imagine you're working with power transformers. We love power transformers in this class. You implement a new change in preventative maintenance. You receive a log of transformer performance. Your preliminary analysis suggests that transformers that received the new preventive maintenance schedule are performing better.

You're pretty stoked at this point. But you dive deeper, and you realize you're not getting any data from many of the transformers that didn't get the new maintenance. In this case, your estimation of the effect of this new approach might be off.

6.10.5 Overview of Sampling Strategies

This element addresses the following learning objectives of this course:

- LO2: Design and apply research questions.
- LO4: Justify an analytic approach that informs decision making.

You'll spend more time in other classes on sampling, but it's worth knowing the broad categories of sampling earlyon. Broad categories are probability sampling and nonprobability sampling. Probability sampling relies on probability theory, usually random selection, to determine what person or element is selected into the sample. If your sample and design are built correctly, you can generalize to the population.

Some examples of probability sampling include random selection, where each element has an equal probability of selection into the sample, or stratified random sampling, where you want to ensure specific representation of subgroups in the sample that otherwise may not happen by simple random selection.

Nonprobability sampling is a sampling method that does not rely on probability theory. Therefore, to the extent that a researcher wants to generalize to the population, it will require argumentation that does not rely on probability theory.

One example is snowball sampling. It's particularly useful to capture difficult-to-reach populations. In this method of sampling, you start with an individual, and then you ask them to recommend other people to interact with.

This particular form of sampling and nonprobability sampling in general may be appropriate if you don't have the sampling frame to sample from. Imagine you want to talk to people who are unhoused or undocumented. In those cases, it's unlikely that you have a list you could sample from. Now, just keep in mind that the composition of the sample can be heavily influenced by the starting point in snowball sampling.

Another example is quota sampling where you select individuals based on some characteristics to reflect the distribution of characteristics in the population you care about. Let's imagine you're in an elementary school setting.And you want to interview some of the top, middle, and low-performing schools-- excuse me-- low-performing students. During each class's weekly visit to the library, you ask the teacher about the low, middle, and high-performing students, and then you interview them.

The headline takeaway is that we should be mindful of how the sampling approach does or does not allow us to generalize to the population.