**Week 6 transcript--Inference, Variables, and Sampling (updated Sept 2022)**

# 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?

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.

## 6.4.3 Levels of Measurement

This element addresses the following learning objective 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 try not to go too far down the measurement and variable rabbit hole. But it's important to be aware of different levels of measurement. So let's go through an example.

Let's say we care about partisanship in the United States. In the conceptualisation phase, you might think, OK, do I care about partisanship or party ID, which is a team or brand of US politics that people identify with and that politicians also identify with. Or do I care more about ideology that is a broader belief system about whether or not there should be more or less government involvement on issues such as equity or morality?

And after you kind of conceptualize a little while, you're like, OK, I definitely care about partisanship. I want to better understand why people choose to self identify with particular brand or team of politics. And so I'll operationalize partisanship, that is, move from concept to measure in the following way.

Our variable is party affiliation, the attributes or the characteristics of the variable are Republican, Independent, or Democrat. And the variables can take on a value of one for Republican, two for Independent, and three for Democrat. In our data, we might have a row for each survey respondent, and then a column of party affiliation that has either a value of one, two, or three that stands for Republican, Independent, or Democrat.

And what's next? Do we do stats on this, or what's next? To that, I would say, hold your horses. We've got to think about level of measurement first. We've got to think about the different levels of measurements that we know what type of information we could infer from the variables in our data set. And we need to know what type of analysis we could do on a given set of variables.

Here, I follow the textbook I learned research design from in grad school by Trochum. And we'll include a link in the syllabus under Optional Readings if you want to read a little bit more about these levels of measurement that I'm going to introduce.

So there's four levels of measurement. The first level of measurement is nominal, also known as categorical. The numerical value of a variable attribute does not mean anything. In the previous example, we thought of one for Republican, two for Independent, three for Democrats. That number is really only a placeholder for the attribute of the variable.

Said differently, we could easily change it so that three is Republican and one is Democrat, and we haven't changed the concepts at all. The numbers themselves don't convey information.

Nominal variables also represent concepts that can't be ordered. There's no order here. Let's say that one is Republican and three is Democrat. It doesn't mean that three, or Democrat, is more than or higher than one, a Republican. You know, absent any kind of philosophical or normative argument, to rank Republicans, Independents, and Democrats doesn't make much sense.

The second level of measurement is ordinal. Here, there is a rank order to the attributes. But importantly the distance between the attributes is not meaningful. For example, let's say we want to measure the impact of education level on partisanship.

Here, our dependent variable, or outcome of interest, is partisan affiliation. And our independent variable is education. Educational attainment could range in this particular example from zero to five. Zero did not graduate high school, one graduated high school, two, some college, three, four-year college degree, four, master's level degree, and five, doctoral degree.

So unlike nominal, where higher numbers and party affiliation doesn't mean more-- here, higher numbers do mean more education. But importantly, the distance between the numbers is not the same for each level. For example, to move from zero to one-- did not graduate high school to graduated high school-- is not a commensurate jump. There's not the same as jumping from three to four, graduated from four-year college degree program to master's level degree. We can order ordinal variables, but we can't directly interpret the distance between the attributes.

The third level of measurement is interval. Here, the distance between attributes do have meaning, or the intervals do have meaning. OK. Let's switch gears a little bit as far as a substantive example. We still care about politics, but we want to better understand turnout.

The outcome variable or the dependent variable is in-person turnout, that is whether or not someone decides to vote on election day in person. And let's imagine you're interested in turnout in an area of the country where elections happen in the winter, or when elections happen when it's cold.

One's decision to turn out in person might be a function of the weather. If it's cold as hell outside, you may not want to go outside and stand in line to vote. Here, if we're looking at temperature, the distance between the attributes does have meaning. And a plus one degree in temperature means the same thing whether we're moving from 20 degrees Fahrenheit to 21 degrees Fahrenheit, or if we're moving from 80 degrees Fahrenheit to 81 degrees Fahrenheit. Either way, it's an additional degree.

So you can say something-- you are able to say something like an X increase in temperature would lead to a Z increase or decrease in turnout. But you cannot say when it's twice as hot outside, you get a Z percent increase in turnout. So even though the number 80 is twice as large as 40, it doesn't necessarily make sense to say it's twice as hot. There's no absolute zero in temperature when we measure in Fahrenheit. And that's why, because it's not ratio-- and we'll get to that in just a second-- we can't say the number 80 is twice as large-- excuse me-- that it is twice as hot outside when we move from 40 to 80.

Now folks may quibble about zero Kelvins. But we're not talking about that right now. In the Fahrenheit, there's no absolute zero.

OK. The fourth and final level of measurement is ratio. So as you can see, each of these levels of measurement build upon each other. And the fourth level of measurement, as I mentioned already, is ratio. So here there is an absolute zero that has meaning, and you could construct meaningful ratios. So let's stick with the turnout model.

Let's say you care about how income impacts turnout. So income or salary can range from zero to a lot. And in the previous example about temperature and the interval level of measurement, we were not able to say that 80 degrees is twice as hot as 40 degrees Fahrenheit. But here you can say that an annual salary of $80,000 a year is twice as much as an annual salary as $40,000 a year. This is because there is an absolute zero.

And zero itself is also meaningful. If you have an annual salary of zero, that means you have zero dollars. If instead, you experience 0 degrees of Fahrenheit, that doesn't mean that you have zero heat.

So let's step back. What does this all mean and why does it matter? We want to understand levels of measurement so that we can best understand what information is contained in the measurement, in the variables, and also so that we know what the appropriate methodology is to analyze the information.

## 6.4.4 Use Caution if Combine Variables

This element addresses the following learning objective 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 understand one's level of enthusiasm for a product. Imagine you have a seven-point scale that is derived from the following question-- "how satisfied are you with the product?" The seven-point scale ranges from extremely dissatisfied on the left, moderately dissatisfied, slightly dissatisfied, neutral, slightly satisfied, moderately satisfied, and then extremely satisfied on the far right. So on the far left, we have extremely dissatisfied. On the far right, we have extremely satisfied, and then all of these different categories in the middle.

So as an aside, this is an ordinary-- excuse me-- an ordinal level of measurement because there is a meaningful order to this. Respondents who say that they are moderately satisfied are more satisfied than those who say they are slightly satisfied. But we can't say that the distance between slightly satisfied and moderately satisfied is the same as the distance between moderately and extremely satisfied. So this is an ordinal level of measurement.

And let's say that you have another seven-point scale. How likely is it that you would recommend this product to a friend or colleague? It's a seven-point scale-- again, very unlikely on the left to very likely on the right.

So you think to yourself, well, these are both on seven point scales. And I generally want to measure how excited people are, how happy they are, how stoked they are about the product. So we asked each respondent both of these questions. So maybe we could just combine these to get us a 14-point scale To that, I would say, hold your horses. So here's a couple of ways to think about it.

So one is, do you think how satisfied you are with the product is the same as how likely you would recommend it to a friend? Maybe. Maybe not. So let's think about why that might be problematic.

So I might be willing to work with something that's subpar because I already know how it works. And I know that there's transaction costs of learning a new skill set. But I may not want to recommend a subpar product to a friend.

Or maybe flip that on its head. Maybe this is a cool product and I like it, but I wouldn't recommend it to a friend because I know they already use something different. And they are already pretty happy with it. Or they're in a different industry, or something. You could come up with a bunch of different reasons why these two questions are fundamentally different.

And so let's not get too ahead of ourselves, but maybe we already did. And let's imagine that you want to do some more advanced-- something a little bit more advanced. And you want to create an index of happiness or satisfaction.

And so you create an index with these questions, or maybe other questions. And you get a value, and you're like, OK. That makes sense. Maybe these variables are highly correlated, and it makes sense to combine them. But what does the index even mean?

What does it measure? What's the underlying concept we care about? And does this index capture that?

So let's step back. Why does this matter? Well, sometimes, many of the variables we care about are highly correlated. And for reasons you'll discuss in other classes, it may not make sense to include all of them in our model separately. And so you might be tempted to combine them.

But the punchline is if you want to combine variables in however way you do that, think about the assumptions you make about the individual variables or the individual metrics, and then think about the assumptions you make when you combine them. We've got to be precise. We've got to be careful. We've got to be deliberate and systematic. That's what makes us data scientists.

If we start combining things in a way that's not mindful, we might advance our project. We might find some results. But then we might struggle to derive insight from the results. You might find yourself saying, OK. But what does this even mean?

So one of the major themes in the class is to slow down our thinking, to be deliberate, to be systematic, so that when we are in the final phases of the project, we can have more confidence in the insight that we do derive.

# 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:

Employee satisfaction => average employee tenure

Employee satisfaction => employee responses to job satisfaction
survey

# 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 Am I Thinking About This Correctly?

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 potatoes. Tier 1 is the highest quality and Tier 3 is the lowest quality.

The current business problem is that potatoes do not always get sorted appropriately. In particular, too often, Tier 2 potatoes get mistakenly graded as Tier 3 potatoes. The company is leaving money on the table since they can charge customers more for Tier 2 potatoes than for Tier 3 potatoes.

The company knows that you recently enrolled in a data science program. They need help thinking through this problem. Please respond to your manager's email below.

TO: You

FROM: Cool Manager

RE: Sorting Issue

Hi,

I'm convinced there is a two-stage approach to our sorting problem. Can you send back a couple sentences on your thoughts? Am I even close? It’s ok to reply back with more questions if you I’m thinking about this the wrong way.

Stage one: Identify when the sorting issues occur.

Stage two: Implement some kind of intervention to try to minimize sorting issues.

Variables: I’m convinced some of the variables we need to think about are time of the day or shift / volume of potatoes at a given time / which sorting line, etc.

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.

# 6.11 Necessary Features of an Experiment

# 6.11.1 What is an Experiment?

This element addresses the following learning objective of this course:

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

## Overview

To compliment what is discussed in Creswell and Creswell, we introduce the experiments framework that is the backbone of Data Science w241: Experiments and Causal Inference

Imagine you want to investigate the effectiveness of a vaccine and you want to make a causal claim about its impact. The **dependent variable** (or outcome of interest) is whether or not the participant contracts the disease of interest. The **independent variable** is whether or not the participant received the treatment vaccine or the placebo vaccine (also known as the control condition).

You recognize that if you conduct an **observational study** and allow participants to self-select if they receive the treatment vaccine or the placebo, there is the risk that **confounding factors** impact your inference. (See page 50 in Creswell & Creswell 5th ed. for more on confounds.)

For example, perhaps only the most health-conscious participants choose to receive the treatment, and those that are less concerned about health choose the placebo. In this case, you may incorrectly infer that the vaccine really works! You see that on average those that received the treatment had better health outcomes than those received the placebo. But you cannot be sure that this difference is **caused** by the vaccine. You cannot separate the potential benefits of a generally healthy lifestyle from the impact of the vaccine.

Here, lifestyle is the confounding factor because it may impact both one's willingness to receive the treatment (independent variable) and one's baseline risk of contracting the disease (dependent variable).

This becomes an even more complicated problem because the relationship could be the opposite–perhaps the most health-conscious participants are more likely to choose the placebo. Furthermore, your research team may not have **any**theory about how baseline health impacts the effectiveness of a vaccine. Finally, what happens if you have **no idea** what confounds are most relevant? Is it baseline health? Occupation? Number of romantic partners? Something else?

The experimental method helps reduce this confounding risk. In particular, if you do the **randomization** correctly, you can convincingly say that any difference you observe in the average outcome of those who received the treatment, compared to the average outcome of those who received the placebo, was **caused by** the vaccine.

To convince yourself and your audience that you did the randomization correctly, you would show a balance table that demonstrates that the treatment and control groups are virtually identical on relevant observable characteristics for your study (e.g., income, race, etc). If you do the randomization correctly, you can also convincingly say that the groups are virtually identical on non-observable characteristics.

### The necessary characteristics to call this an experiment are outlined below.

**1) Intervention:** Your team must have an intervention. That is, you must have a treatment that you caused. This contrasts with a “natural experiment” or “quasi-experiment” where one focuses on an event that was not assigned by the researcher (e.g., natural disaster).

- **Example:** You conduct a medical trial and randomly assign half of the subjects in your study to receive the vaccine (the treatment). The other half received a placebo syringe that contains a saline solution (the control).

**2) Random assignment to treatment:** You must assign at random whether or not a subject receives the treatment or control. Said differently, subjects have the same probability of being assigned to the treatment or control conditions.

- **Example:** You flip a coin to determine whether each subject is assigned to the treatment condition. If the coin comes up heads, that subject is assigned to the treatment and receives the vaccine; if it comes up tails, that subject is assigned to the control and receives the placebo.

**3) Excludability:** Only your treatment causes the treatment effect, and not some other feature of the experiment.

- **Example:** To meet this excludability criteria, anyone who is assigned to receive the treatment does receive any additional benefits. This criterion would not be met if, in addition to being assigned the vaccine, subjects in the treatment group were allowed access to exercise gyms where unvaccinated control subjects were barred. In this case, you cannot separate the impact of being vaccinated from the impact of working out at the gym.

**4) Non-interference:** There is no spill over or other interaction between subjects. One subject’s assignment to treatment or control does not affect any other subjects.

- **Example:** Even if you only vaccinate those in the treatment condition, this could reduce the prevalence of the illness in the entire population. Said differently, this may reduce the chance that a control subject gets infected. This non-interference criterion would not hold if a control subject’s illness outcomes changed because

of the effects of the vaccine on a treatment subject.
- Interference may occur in multiple scenarios: subjects in the treatment impact those in the control (the scenario described above); subjects in the control impact those in the treatment; subjects in the control impact others in the control; subjects in the treatment impact others in the treatment.

# 6.12 Creswell and Creswell Textbook

## 6.12.1 Chapter 8: Quantitative Methods

This will help you digest the required Creswell and Creswell reading. Please use these ideas as a starting point.

This element addresses the following learning objective of this course:

- LO2: Design and apply research questions.
- LO3: Assess and select data and the data collection methods that best fit a specific outcome or need.
- LO4: Justify an analytic approach that informs decision-making.

This chapter focuses on quantitative methods. Let's talk a little bit about sampling first. I'll focus on population, sampling frame, and then the sample.

So in many quantitative studies, we care to generalize to a broader population. And so the population is the broad group of folks that we want to generalize to. So let's say we want to generalize to cell phone users. The sampling frame is the list from which we sample. So let me step back.

The population is cell phone users. The sampling frame is the list of cell phone users we either have internally or we get from a vendor, because we have to sample from something. We can't directly sample from the population.

Now sometimes, the sampling frame is the same thing as the population, especially if we're in the kind of online space or we're working in the kind of data native companies. But oftentimes, the sampling frame is a subset of the population.

So again, we have, in this scenario, the population is cell phone users. The sampling frame is the list of cell phone users we got from a vendor or that we have internally. And the sample is who is in our study. These are the folks that were selected into our study.

And the whole reason we care about sampling and-- that we care about doing sampling correctly is because we want to be able to generalize back to the population. Let's talk briefly about power analysis.

And here, you know, it can get a little bit more complicated. It is more complicated than I will describe it today, and you will cover this in other classes.

But the high-level idea is to think about how many observations you may need in your study to have a chance of detecting something. The idea here, in general, is that if the anticipated effect size is small, you're going to need more observations. If the anticipated effect size is large, you'll need fewer observations. And the idea here is that if the effect size is small, to be able to distinguish the signal from the noise, you'll need more observations. And again, if the effect size is really large, you won't need as many observations to separate that signal from noise.

And power analysis is a little bit of art, a little bit of science, because it's based on the anticipated effect size. And so that's kind of where that art comes into it. And so you're going to base this anticipated effect size on existing literature, on best practices. But especially if you're in a domain that's maybe relatively undefined or you're kind of trailblazing, you might not have as much kind of insight on this front.

This chapter also talks about variables, which we've mentioned in different videos, but I'll also talk about it here. Again, the dependent variable is your outcome of interest. And the independent variable is the variable we think influences the outcome of interest.

The dependent variable again is the outcome you care to explain. And the independent variable is the variable you think influences, or in some case, causes the dependent variable. So let's go through an example.

Let's imagine we work for University athletics. And our dependent variable, our outcome of interest, is athletic performance. There's different ways to operationalize that, depending on the sport. Let's say we're working for the football team. And we're talking about American football. We could think about it as the number of touchdowns, yards run, the number of key plays or sacks, or something like that.

Let's say that our key independent variable in the study is hydration. That is, we want to look at the impact of hydration on athletic performance. And let's say we want to run an experiment. So we-- hold on.

So I'm going to use this description of a hypothetical experiment to talk about internal and external validity. So again, the outcome of interest is athletic performance. And we're going to vary in a kind of experimental way the hydration regimen. So let's talk about internal validity and external validity.

So in the study, we may want to focus on two things. That is, say something convincing, and then be able to apply that convincing insight beyond our sample to the broader population. So let's talk about internal validity first.

So in the words of Creswell and Creswell-- and this is on page 169 in the fifth edition of the text-- quote, "internal validity threats are experimental procedures, treatment, or experiences of the participants that threaten the researcher's ability to draw correct inferences from the data about the population in an experiment," end quote. So what does this mean?

This means that we have to make sure that the way we design the experiment and the way the subjects engage with our study do not limit our ability to draw valid inferences. For example, take diffusion of treatment, one of the internal validity threats that is covered in the text.

Let's again imagine you work for the University sport's team and you want to assess how the hydration regimen or an athletic mix that we put into an athlete's water influences performance. And so we randomly assign different kind of hydration regimens to students or to athletes-- to student athletes.

The diffusion here is that athletes may share water bottles. People in the control condition may have consumed water or this kind of hydration, magic potion from individuals that were in the treatment condition, and vise versa. So you're going to get people that were assigned to the control condition. They got treated.

And in this case-- yeah. And then also people that maybe were assigned to the treatment condition, where they should have got the magic power-- or the magic powder-- might drink from people's water bottles who were in the control conditions. So they're not getting the level of treatment intended, and in this case, might dilute the effect that we see.

So that's internal validity. Second is external validity. In the words of Creswell and Creswell, again, from the fifth edition page 171, "external validity threats arise when experiments draw incorrect inferences from the sample data to other persons, other settings, and past or future situations," end quote. So what this means is that if we are-- that we need to pay attention to external validity threats. Otherwise, it impacts our ability to generalize our findings out of sample.

So for example, let's say we do this kind of hydration experiment during a super hot summer. Maybe the findings don't generalize well to other summers or fall sports because of the difference that the drink powder might make only when it's really hot outside. So the kind of condition that makes this experiment different and maybe because we-- and the reason that we may not be able to generalize is because we did it in a really hot summer.

This doesn't kind of discount the fact that the finding was internally valid, assuming that we figured out how to make athletes only drink out of their water bottles. But we just might have to be a little bit extra careful about making decisions about fall sports based on a study in summer.

So I think those are the main ideas in Creswell and Creswell. Thanks.

## 6.12.2 Chapter 9: Qualitative Methods

This will help you digest the required Creswell and Creswell reading. Please use these ideas as a starting point.

This element addresses the following learning objective of this course:

- LO2: Design and apply research questions.
- LO3: Assess and select data and the data collection methods that best fit a specific outcome or need.
- LO4: Justify an analytic approach that informs decision-making.

This chapter focuses on qualitative methods. And the two main concepts I want to talk about are validity and reliability. And I also want to talk about how one's, kind of, backgrounds and lived experiences can influence our research. So let's talk about the validity first.

Validity in the quantitative sense that we talk about is kind of are we capturing the concepts we want to capture. Do our metrics capture the concepts we intended to capture? Are they kind of accurate? And I think that approach to validity applies here as well. Creswell really pushes us to think about validity in the qualitative sense in a slightly different way.

Creswell and Creswell talk about trustworthiness, authenticity, credibility. And so I think it's important to remember that validity could mean different things depending on kind of our approach. And I think it's important to do some thinking beforehand before we start a project, to think about the ideas we want to measure, the ideas we want to capture, and how will we know if we've in fact captured those concepts in a way that's valid, that's trustworthy, that's authentic, that's credible. And the chapter goes over a couple of ways to kind of think about that.

Let's think about reliability-- that is, is our approach consistent across researchers and contexts or projects? Is it consistent and stable? If another researcher followed the exact same procedures, would they come to the same conclusion?

So let's imagine we've developed semi-structured interviews. And we develop a codebook that instructs the various research assistants to identify major themes in the interviews live. Maybe we don't want to record the interviews in this context because of the sensitivity of the topic and maybe, as a result, we're not able to use automated text analysis.

And so if we were doing this kind of coding live, is the codebook written in a way that two research assistants would come up with the same coding? If not, how could we write it differently or how can we train differently? And again, the codebook would be kind of detailed description of when to categorize, let's say, certain words and phrases as one thing, and when to categorize other words and phrases as another thing. So it's basically the set of rules that determines our coding.

The text also provides, I think, a really good substantive example of the application of qualitative procedures. There's a case study of the ethnography of the president of a four-year college. And the researcher explicitly calls out their prior experience working with college administrations. They say, quote, "due to previous experiences working closely with the new college president, I bring certain biases to this study. Although every effort will be made to ensure objectivity, these biases may shape the way I view and understand the data I collect and the way I interpret my experiences," end quote.

This is not that different from the way that we view-- that we interact in our day to day. We have prior experiences that influence our perception of research and the questions we ask. And I think it's important to keep that in mind. In this particular example, this particular researcher, she was upfront about how her prior experiences might influence or bias, maybe, in a way that's undesirable.

But in the text, she also talks about how-- because she worked with administrators in the past. She has unique knowledge. And she might be able to, in her observations, derive knowledge and information from certain experiences that a outsider wouldn't notice. That is, she might identify important nuance that someone without those experiences would overlook.

Another way to kind of think about this is to think about it in the sense that we maybe can't be fully objective. We can't-- we're not robots. We come to the table with our previous kind of professional and personal experiences. So maybe we can't be fully objective, but I think we can be transparent. And we could be honest about how our prior experiences influence our analysis.

Now depending on the domain, this declaration of our prior experiences and our prior sets of beliefs may or may not be client facing. But I think, at least internally, I would urge you to think about how your prior education, your prior professional experience, and your prior life experiences influence how you see a problem and influence how you conduct analysis.