

Correlation Design Essentials

Video Views: 1

Just let me know.

Um, so again, this is, this particular presentation is focused on quantitative correlation research design.

Um, and for those that don't know me, um, I'm Dr. Jim Rice.

I'm a, a faculty member in the College of Doctoral studies.

I teach primarily, um, the, in the business curriculum, um, in the DM programs.

And, uh, I am a dissertation chair.

I'm not actually a user research methodologist, but I am, I love this, um, design type.

It's, it's one that I've got a passion for.

And, uh, so I like to be able to share this, uh, share my experiences with doing research with this design when I can.

Um, so you will notice as we go through this, my, my approach tends to be, uh, describing how to apply this design in a dissertation process.

So I will tend to walk through how to integrate this into a dissertation proposal, this design.

So, um, it's not focused on the details of statistics or correlation design, but, but how to use it for, uh, uh, dissertation research.

Um, so I, I gotta start out by just defining a few terms here.

Uh, 'cause you will hear me use them almost interchangeably at times, and I want to make sure that I'm defining them upfront.

Um, this is a, um, uh, talk on correlation design, but you may also hear me use the term co variance.

And that's really what correlation designs are.

They're a study of the, uh, variation of variables, um, in tandem, so to speak.

So, uh, as one variable changes, another variable changes, and they cova meaning that there's a relationship and how they, uh, change.

Um, you'll also hear me occasionally use the term inferential statistics.

Boy, that's a mouthful, isn't it? Uh, inferential statistics means, um, using a statistical approach to try and predict an outcome.

That's the whole purpose of Covance, is to be able to look at one variable and have a reasonable expectation that I can com I can predict, uh, with a, within a level of tolerance, the value of another variable.

Um, so this is, you'll hear me use these terms.

Um, you will never hear me use the term cause.

And, and that's one thing that I, I always reinforce with students is that correlation is not a study in causation.

It is a study in correlation or co co variance.

Uh, it allows us to predict the outcome of one variable versus the other.

Um, and that's what makes this such a powerful, um, powerful tool because it allows us to step back and look at large populations of data and begin to look for patterns in them.

And that's what's really, that's what I love about this.

Um, quantitative research as opposed to qualitative research, is really looking at new insights and patterns in the, in the populations we're studying.

Um, qualitative research dives into those relationships and helps us understand, uh, understand those relationships more deeply.

But this is really one of the first steps we do, is take to look and see if there is a relationship, uh, that is significant and causes some predictability.

And so, when, when students are looking at a study and they're looking at trying to understand, is there a relationship between two variables, uh, or two attributes in a population or, uh, uh, patterns within that population, we're, we're gonna be looking at quantitative research, and probably the most common is correlation design.

We are gonna talk briefly about all of these steps.

These, these are the pieces and parts that are necessary to do a full correlation analysis.

Uh, we're gonna start by looking first at how do you, how do you state a problem and a purpose, uh, that's appropriate for correlation? How does that, uh, affect your selection of the methodology and design? And then how do you state and describe the variables, the research questions, and the hypotheses, uh, in an aligned way? These, these, this red row here always has to align very carefully, and it's one of the things that we watch very closely as we're reading a chapter one proposals, making sure that all the elements of a correlation design are aligned.

We're gonna talk briefly, uh, about, um, how to describe and sample a population.

What kind of techniques are valid, what, what sampling controls are to ensure that you've got, um, a representative sample.

I'm not gonna spend a lot of time on this because we have another webinar in the, uh, research hub that focuses specifically on population descriptions and sampling controls.

Uh, but I'm gonna fly over some of that material quickly, uh, just to reinforce how important that is, particularly in, uh, quantitative research.

Um, we'll talk a little bit about what you do after you get your IRB approval.

How do you recruit, uh, a sample? How do you, um, collect data for that sample? And then we're gonna spend most of our time in this hour talking about two very important, uh, aspects of correlation

research, which is the descriptive statistics, the often forgotten, uh, aspect of correlation design that's so important.

And the correlation statistic itself. What is it? How do we select it? Um, what are the dependencies that determine what the correlation statistic is that we're gonna use? And then we're gonna wrap up by how to analyze and how to interpret, uh, the information that you're getting from, um, a correlation statistical analysis.

Most students will use a tool like SaaS or SPSS to crunch the data, and they get all these numbers out.

And, uh, it's really important to understand what those numbers mean.

And so we're gonna talk a little bit about what that is.

Um, so we're gonna dive right in here.

Um, the key objectives, this is, this is again, why we do correlation research.

We're really looking to understand relationships in the data, and we're looking to see if we can predict an outcome or predict the value of a variable based, an independent variable based on a, or a base predict, a, an independent, a dependent variable based on an independent variable.

So we're, we're looking for those relationships that occur, and I'll give some examples of that in a minute.

Um, we have to make sure we understand how to, um, if we're testing something validly, are we, are we, for example, if we're looking for student outcomes based on a type of intervention or test, are we looking at, uh, those outcomes with, and getting that understanding of relationship, uh, that is valid and statistically significant? And we'll also use, uh, correlation designs in, uh, exploratory analysis.

Primarily what a correlation statistic is, is simply a data reduction.

We're taking a huge population of data, and we're re reducing it down to a small set of numbers, uh, that help us understand that data set better, and, uh, draw some conclusions from it.

And so these are the types of things that we're doing, uh, with correlation design.

And again, it's really exploring a population and looking for relationships that can help us, uh, understand the world around us a little bit better and predict how it'll behave.

I'm gonna, I'm gonna share an example of a problem statement, and I always map out problem statements with my students.

They've all, all problem statements have these same five components to them.

Uh, but you'll notice that there's two very important elements that are necessary for a correlation design.

One is we are looking for, uh, a problem that is seeking to understand if there's a relationship.

So, typically, a problem with statement will use that term relationship or co variance or, um, connection, or some way to tie together two variables that exist.

And you notice in this problem statement, there are two variables.

Um, I have shoelace material, which is a specific measurable variable, and I have an untie rate.

Um, and I'm in this particular research design, I'm looking to see if runner shoes have a relationship between their shoelace material and the rate that they become untied when they're running in a marathon.

So you can see this is a correlation study.

It gives me an opportunity to go and observe and collect data, uh, from runners, and then be able to look to see if there's a relationship between these two, um, variables.

And if that relationship exists, then as a researcher, I've now got an opportunity to, uh, delve further into that and see what I can do to, uh, improve our understanding and also potentially create some interventions that could, uh, uh, mitigate the issue.

So, again, I'm looking here for correlation study.

I'm making sure my problem statement is aligned.

And, and one of the things I look for are the variables and the, the, a key word that indicates a correlation design.

In this case, relationship purpose.

You're gonna see the same type of terminology used in the purpose statement.

Really, when you're looking for correlation research, you're looking for a relationship, uh, between variables.

So this purpose of this quantitative correlative study in the, um, correlational study in the, uh, example on the right is the relationship between transformational leadership and emotional intelligence.

Again, two variables to quantifiable variables.

And I'm looking to see if there's a relationship between the two.

How does one vary as the other varies? Um, so you can see that this is a very powerful tool.

We can use it for an awful lot of, uh, examinations of relationship.

I will tell you though, many students when they are, um, uh, ideating or, or figuring out what they want to do for their research, often forget how to just often forget to identify quantifiable variables.

Make sure that you're defining a variable and you understand where that data is coming from.

Um, you know, I've had students get all the way through chapter two and still haven't figured out where they're gonna get their data from.

And that really makes it difficult to do a correlation study.

If you don't, at the onset, have a vision for where you're gonna collect this data from, are you gonna get it from secondary sources? Are you gonna get it from a survey instrument that you administer? Where are you gonna get this data from? Are you gonna mix and match primary data and secondary data? There's a lot of different numeric, um, or quantifiable data sources to populate

variables that you really need to know where that data is coming from, even as you begin to propose and build this proposal for your research design.

So, never forget how important the variables are.

Um, so this, one of the things that's unique about quantitative research, particularly, uh, correlation studies, is that for every research question that you're gonna have, um, you're gonna also have hypotheses and null hypotheses.

So if I look for a, if i, if I look at a proposal that has a, uh, proposed correlation design, one of the things I go look for is the problem statement.

I look at the purpose statement, but then I go look at the research questions to see what the hypotheses are.

Um, the hypotheses should directly relate to the research question.

And a primary hypothesis and null hypothesis should be the mirror of each other.

The primary hypothesis is what we're seeking to, um, confirm.

And a null hypothesis is what we're, we're seeking to reject.

Um, and we may or may not reject it.

Um, it's perfectly okay in a research study, by the way, to look for a correlation, study a correlation, and find out that there isn't one there that's a perfectly legitimate outcome.

And I, I actually had a student one time call me crying because she had found the null hypothesis to be confirmed, and, uh, couldn't reject the null hypothesis.

She thought she'd failed. Um, I said, no, not at all.

That's a powerful finding.

Um, let's, let's, let's, and analyze that and figure it out.

So, just recognize that a correlation study is seeking to, to accept or reject, um, one or the other of these hypotheses.

That's the purpose of the correlation assessment.

And the intention there is to provide insight into the research question that we posed.

Let's talk about variables for a minute.

There's two different ways to look at your variables.

And, and again, I will, I will honestly tell you this is probably one of the hardest things that Stu students struggle with, is being able to articulate their variables.

Um, so in a correlation study, you were looking for a relationship between an independent variable and a dependent variable.

Those are always, those are always present.

Um, but there's also, uh, potentially gonna be control variables, moderating variables, mediator variables, uh, for example, a moderating variable.

Um, good example, if, if I'm looking for a relationship between, um, the, the temperature outside and the amount of, um, time that a student spend studying there, there might actually be a relationship there.

It might be a strong one, might be a weak one, might be a moderate one, uh, but there's actually maybe an in, uh, Indian, a moderating variable that says the windows are open or not in the house.

So it's always important when we're looking at studying relationships to step back and take a look to see if the, the variables that we are examining are, uh, complete, and that they tell the whole, the full story.

Um, so always explore your variables and make an attempt anyway in your proposal to define, um, whether variables are control variables, moderating variables, mediating, mediating variables.

You have to, it is a requirement that you have a dependent and independent variable, but you will get, you'll produce more insight into your study.

If you can identify potential, uh, control moderating and mediating variables.

It will speak to the strength and also the repeatability of your study.

So, uh, again, if there's, uh, a lot of moderating variables or mediating variables, um, it may challenge the repeatability of your study.

So, step back, look at your variable data, make sure that you understand what type of variables that you're, you're positing that you wanna examine the relationships with.

This is the, this is the really important one.

Um, this is almost always forgotten by students.

Um, and you, uh, it's important to think about what type of variables you have, what's the nature of the variables that you're using.

There are different kinds of variables.

Um, variables can be ratio, they can be interval, they can be ordinal, they can be nominal, discreet, continuous.

Um, I put a number of examples of, of what each of these variable types look like.

Uh, so you can assess when you're putting variables into your study.

You can be, you are in a position to be able to assert what types of variables are, why is this important? Um, for starters, uh, when you choose a correlation statistic, there are correlation statistics, whether it's a Pearson or a Spearman or roe or, you know, there's a variety of different correlation statistics, and we'll talk about those later.

Um, most of those have a dependency on the type of variables that are included in the, in the dataset.

Um, so if you, for example, attempt to do a, uh, a piercing correlation on nominal variables, you'll get an answer.

I guarantee you SPSS or or SaaS will kick out an answer for you.

It'll be wrong, it'll be inaccurate.

Um, but, and that's why it's important to make sure you are, you understand what your variables are before you go pick your statistic.

Um, and if you change your variable, you have to go back and reexamine your statistic.

It is, okay. I have a, a number of students that are doing multiple correlations in a study.

They may have three or four variables.

They wanna look for relationships between.

And, uh, they may be using different correlation statistics, um, when they, uh, do that analysis because they're comparing different types of, they're, they're correlating different types of variables.

So, um, you know, as you're putting your proposal together and you're describing your population, uh, and the, the variables within your, your study, um, I'm sorry, the purpose section, when you're describing the variables in your study, I, I always encourage you to identify this, not just that they're a variable, but the type of variable they are.

'cause that is very informative to the person reading your, your proposal.

Um, again, I, I said this at the beginning, one of the most common problems is students will throw out ideas for variables without any idea where the data's gonna come from.

Um, make sure literally before you get past the population section in chapter one, you know where your variables are coming from.

You're either gonna do a survey, uh, and collect the data as primary data, or you've got access to data sources.

Uh, and you can see the schemas in those databases and see specifically what the variables are that are available.

You wanna know where your data's coming from before you get too far in this proposal.

Um, I've, I've had students get all the way through 7 41 and realize they can't actually get the data.

It doesn't exist anywhere.

And, uh, they have to rewind and start over again.

That's a long way to go before you, uh, uh, before that person made the effort to figure out where they were gonna get their data from, figure it out early, you know, have a good, good, a good vision for where your data's gonna come from, and be very specific about it.

So, we've talked about the, uh, uh, problem, purpose, methodology variables.

Um, I'm gonna talk briefly about how to describe a population for a correlation study.

And we'll talk about sampling techniques and sampling controls.

Uh, an often overlooked portion of the proposal, uh, that we often have to go back and, and put in, uh, before we can get into IRB.

So, let's talk about the, the population section.

This is, this is a common chapter one, uh, section.

You're gonna have a population and sample section that's comprised of a, uh, paragraph or two on the population.

And a paragraph or two on the sampling technique, um, when you're describing your population, is really important to make sure, you know, um, what constrained your population.

So, for example, if you were studying, um, business executives and manufacturing organizations, but you've only really got access to those in Michigan and Illinois, make sure you define the, the limita, the limit, the limited popul, because that understanding that delimited population is critical to your power analysis that you're gonna do later.

This is the end value you see all over the place in the statistics books.

This is the number of people that exist in your population.

Um, be careful to avoid trying to do a, uh, a power analysis on the entire population where you're gonna wind up having to do a much larger sample than you would otherwise want to do.

So, make sure you've got a good delimited population identified.

Uh, it has to still be substantive.

Um, it has to be a, a substantial population to study, uh, so that the findings can be generalized.

But make sure that it, you, you've delimited it properly, and that you specifically know how many people are in that population.

It's very, very important to know.

Uh, and it doesn't have to be a perfect number.

I mean, you don't have to know that it's 64 people.

Um, saying it's between 60 and 70 people is fine.

Uh, but making sure that you've got a good idea, uh, at least within a reasonable order magnitude of how, um, how large that population is.

Um, what are the limitations, estimate the size.

And I, and I say estimate the size, because sometimes it's impossible to know the exact size.

Don't get too daunted by that.

Um, it's okay to step back and say, there's, I can go look up and see there's, uh, 700 companies in this region, and on average, they're gonna have five HR resources, and therefore, I'm gonna have x number of members of my population.

As long as you articulate how you come to that estimate and what it's based on, and you've got some support for that, then you can use that number for your, um, uh, your population size.

Now, this next part is really important, um, demographic information about your population.

And this is gonna go directly to the size of your, uh, sample and whether or not your sample is representative of the population.

Uh, so when you are sampling a population, you wanna make sure you know something about that population.

Uh, what's the ratio of men and women? What's the, uh, cultural mix? What's the educational mix? Um, what's their average height? I mean, any kind of demographic information is fine.

And for a dis as a dissertation chair, I'm not that picky about what you use, as long as you have good demographic information within that describes your population and provides the reader with some insight, uh, about that population, education levels, et cetera.

Why is it important? Uh, it's important because when you sample, you want to have your sample roughly represent the ratios of the population.

So if your population is, is, uh, 40% men and 60% women, your sample should be roughly 40% men and 60% women.

You could, if without the demographic, without the sampling controls in place, you cannot assert as a researcher that your study is examining a sample of the population that represent or sample of the population that represents the population.

Um, you could wind up with a st a statistically significant sample.

We'll talk about that in a minute.

But, um, you can, you can achieve that minimum sample size that's required for statistical significance, um, and have it all women, in which case that population does not.

That sample does not rep represent the population.

So you may have to continue inviting people, you might have to continue sampling until you get the right ratios within your sample to be able to assert that it's representative.

This is true of, uh, qualitative research, but it's also true of, um, quantitative research, probably more so in quantitative research because of the size of the populations you're dealing with.

Um, so again, we, we frequently talk about the minimum sample size that, and this is why in the proposal in chapter one, before you actually start writing chapter four and five, I always encourage students to describe the minimum sample size.

The minimum sample size for quantitative research is going to be what's called a statistically significant sample.

You're gonna do a, a power analysis, uh, we'll talk about that in just a second.

You're gonna do a power analysis, which will give you the number of participants you have to have in order to be able to accept or reject the null hypothesis.

We call it the minimum sample, because you, again, you may have to continue sampling to get a representative sample.

Um, so in quantitative research, we're still, we're trying to get both a statistically significant sample and a representative sample.

Both of those criteria very important. Okay.

Uh, this just talks about this a little bit more.

Um, researchers commonly have to collect data above their minimum sampling levels.

Again, the minimum is where we reach statistical significance and the, um, actual sample sizes where we've got a representative population, uh, sample.

We're representing the demographics of the population, however you've described the demographics.

Um, yeah, this is just going into more detail about that.

Uh, describe the population. What's the limitation? How's the size of the limited population and the demographics? This is starting to talk about the sampling.

And again, whatever sampling technique you use to collect your data, if it's secondary data, typically is gonna be a large representative, uh, data set for the population.

But if it's primary data survey data, you're gonna be collecting data from population members until you at least get to the minimum, uh, sample size.

And then you begin applying the sampling technique or sampling criteria to ensure you get a representative sample.

That representative sample those sampling cri, those sampling, uh, criteria are particularly important with some of the non-random sampling methods.

And we'll talk about briefly in a minute.

Um, I think I beat this horse quite a bit.

Um, again, minimum sample size is always done with the power analysis.

You could do this in SurveyMonkey with the, the sample calculator.

You can do this with G Power.

There are hundreds of sample size calculators out there.

They're all going ask for the same basic information.

They're gonna ask for your population size.

And remember, that's a delimited population.

They're gonna ask for the confidence interval.

This is the interval or that part of the bell curve that you're trying to hit to have, uh, the ability to accept or reject the null hypothesis.

For most business research, um, most, um, social research, were gonna be in the 90%, 92.5%, 95% confidence intervals.

Um, work with your chair, work with your UM to find out what they consider acceptable, uh, for your particular field and for your particular population.

Uh, some populations are just very hard to sample, and it may be acceptable to get to an 80% confidence interval.

That's really a conversation between you and your committee.

Um, the other variable that you need to be concerned about is the margin of error.

This is your, um, what are, there's two different kinds of errors, but the, the margin of error is our confidence in our confidence interval, if you will.

Most of the, most of the time, our, uh, research is gonna use about a 5% or 10% margin of error, uh, for the calculation.

So the combination of the population size, guided by the confidence level, and the margin of error will give you the minimum sample size that you should be targeting.

Now, again, this is for your proposal.

So again, this is your minimum sample size.

You're gonna do sampling controls that are above this.

Once you've collected your data and done your correlation analysis, you're gonna come back and do your power analysis again.

Um, uh, normally you'll do that with a Z statistic and look, look up your P values in a table somewhere, or, or just have SPSS, uh, do that.

What it's gonna do is this, is that power analysis is gonna tell you, can you, with confidence accept or reject the, uh, the null hypothesis.

The reason you do that with the actual sample size is we wanna know what the real confidence level is, and the real margin of error is of your study based on your sample size that you collected.

Um, it will often deviate.

It will almost always deviate from your proposal.

Um, and it's important for you to make sure that you meet or exceed, um, the proposed confidence interval and margin error.

If you don't achieve that confidence level or margin of error, you have to explain why.

Um, and that'll have to be, uh, part of your, uh, chapter four, chapter five, write up.

Um, and again, if you don't meet those thresholds of your proposal, work with your chair on it.

I've had many students not able to collect their minimum samples, uh, and we just go, you know, they're shooting for 95 and they actually get a 90% confidence in a role.

That's fine. Um, I generally will.

I, I will generally let students in, in business research go down into the 85, 80% actual intervals if they just can't get to the right population.

We get below that. We have questions about how do we get a larger sample rather than, uh, stopping the data collection.

Uh, so this power analysis is important both in the proposal, but it's also very important.

Um, doing, getting your p values at the end is important to understand how significant your findings are.

Um, yeah, this is not intended to be a statistics class A, so I'm just, I'm not gonna skip over some of these things.

Uh, again, this is all about when do you stop sampling.

You stop sampling when you've got at least the minimum and you satisfied the sampling criteria.

Um, I'm gonna skip, oh, oh, sampling methods.

There's a lot of different sampling methods.

Um, for, you know, the, the gold standard for quantitative research is random sampling.

Random sampling is almost impossible for most dissertation research.

You have to have complete access to the entire population, and then you have to pull the data from it randomly.

Um, and that's complicated to do, and, and not easily available.

Most students don't have complete and uniform access to their entire population.

And so they'll use one of these other sampling methods.

Uh, purpose of sampling, snowball sampling, convenience sampling, random sampling, case study sampling, stratified sampling, maximum variation sampling, theoretical sampling.

I have seen every one of these used in correlation research.

It's really a description of how you are inviting people to participate in your study.

Each of these have strengths.

Each of these have weaknesses, um, mostly as it relates to the bias that you're just injecting into your study.

Um, this is part of the reason, um, the sampling criteria are so important.

The sampling criteria are a control against sampling bias.

Um, if you are, you are ensuring that you're camp collecting a sample that represents your population, you're less likely to fall into the trap that you might fall into.

For example, snowball sampling, where you go pick one person who selects their friend, who they know agrees with them, who selects their friend, who knows they agree with them, and you wind up with a non-representative sample that's got a bias finding.

Um, sampling controls help us mitigate the, the risk of sampling bias within our studies.

And so, make sure you've got sampling controls in here.

And, um, think about how you can sample your proposed populations in a way that minimizes the potential bias or skewing of your data.

Uh, that's, that's a big part of the description in the sampling section in chapter one.

What is your sampling technique that you're using? And why did you choose that and justify that selection based on the ability to minimize, uh, sampling bias? Okay, boy, did I go through the blue stuff fast? So, okay.

So let's, let's dive into, um, descriptive statistics.

So there's two green boxes down here.

This is kind of the heart and soul of correlation research.

Uh, there's two types of statistics that we do within, uh, correlation research.

One is called descriptive statistics.

Never ever skip descriptive statistics.

I say, way too many people skip it.

Um, and then once you've completed your descriptive statistics, you can then move into correlation.

I'm gonna talk about why the descriptive statistics are so important.

Uh, the reality, and, and I'll, I'll, I'll give you the cliff notes.

You can't choose a correlation statistic without doing the descriptive statistics.

First, you have proposed that you're gonna do a Pearson or a Spearman or whatever, based on your variable data.

But those, um, those statistics themselves, those correlation statistics also make assumptions about the population distributions that you can't affirm until you've actually collected your data and done the descriptive statistics on them.

So, if you jump to the correlation analysis without doing the descriptive statistics, um, you very well could have wasted all that time doing correlation analysis.

'cause it doesn't fit the data. Um, so, so the question was, is the order descriptive before correlation? Always, always, always, always.

Um, you can't, and we'll, we'll talk about that more when we talk about the correlation statistics.

So, and why you'll, and you'll, you'll begin to see why it's important we do it this way.

So let's talk about the descriptive statistics.

Um, so once you've put your proposal together, once you have, um, uh, gotten through chapter three, articulated how you're gonna do your, uh, data collection and your analysis, what's your process for analysis? Uh, what's the framework that you can apply to their findings? So you can draw conclusions.

You're gonna get all those first three chapters done.

Now you're gonna do the fun part.

You're gonna go collect data.

Um, and again, the data could be secondary data.

It could be primary data.

Um, but, uh, before you do any correlation, you have to be able to describe the characteristics of the data.

Um, so, uh, the descriptive statistics, uh, include some very common, very straightforward things that we always have to include.

We have to be able to articulate the central tendencies of the data.

I'm gonna talk about that more in a minute.

Uh, we're talking about, um, using these descriptive statistics to identify some patterns in the data.

And we'll, we'll talk about that more in a little bit too.

But is my data normally distributed? Is it skewed one way or the other? Is my, is my normal curve very flat or very spiky? Um, we have to understand these things so that when we're looking at our correlation, that we're comparing that data with like patterns.

Um, particularly if you're doing a piercing correlation, um, uh, where you're, you're assuming normality and you're assuming a, a uniform normality.

Uh, other, other correlation statistics are, are much more forgiving in, uh, the, the patterns of the data.

So, we'll, we'll talk about that more in a minute.

And you also have to assess your assumptions.

Um, you may have made assumptions that the data would be coming in a normal distribution.

You have to, you have to demonstrate that you have to show that your data actually does have a normal distribution, that it does, um, have a reasonable pattern for correlation.

Otherwise, you're not gonna be able to determine a linear relationship.

You may actually determine, and this was true in my dissertation, which is what spent me so long to finish my analysis, is I had a, uh, one data variable that was beautifully normally distributed, and another one was massively skewed.

So I had to go through transforms and figure, what does that mean? 'cause I'm now doing a correlation that isn't linear.

I'm doing a non-linear correlation.

Um, that's a little more advanced.

Don't worry about that right now, but make sure that your data's got roughly the same pattern in both variables. We'll show that in a minute.

Dr. Weiss, there's a question. Yeah.

Uh, clarify primary versus secondary data.

Yeah. Yeah. Okay. So, primary data, primary data, uh, very simply put is data that you collect.

Uh, so if you do a survey, um, and you send it out to a population, you are collecting the data that's considered primary data.

It was not collected by somebody else.

Secondary data basically means it was collected by somebody else.

So if you are pulling data from him or from HHS or you're, you're gaining access to, um, your, your business', uh, CRM environment, you're pulling data from that.

Those are called secondary data sources.

Um, you're gonna have to have, um, permission to use for those.

There's a lot of additional steps in the IRB process for that, um, for secondary data.

But the, the real difference is primary data is data you collect.

Secondary data is data that you collect from somebody who collected it, which is why it's called secondary data.

Um, regardless of whether it's primary or secondary, you can use, uh, correlation designs that are secondary, secondary, primary, secondary, primary, primary.

You can combine and, uh, compare and contrast different types of data.

But you always have to be able to articulate what, whether it's primary or secondary in your proposal, um, so that it can be so that you can assert, particularly if it's secondary data, you can assert that you have an unusual or unique treatment of the data in your analysis.

Um, that is more than just simply pulling data from a database and running statistics on it.

So, um, okay.

So again, data cleaning.

So, um, almost any data collection is gonna have outliers in it, meaning data, uh, a data variable or data point that is wide, wildly out of context with all the risk of the data.

Um, these are called outliers.

And part of the data cleaning process is going through and identifying outliers.

And in many cases, we will reject outliers.

You have to report that you've rejected an outlier and how many outliers you rejected.

But, um, it's common for you to reject outliers to get a more accurate correlation assessment.

This is an example, a, a, a scatter plot where I've plotted the, um, dependent variable and the independent variable.

You can see the vast majority of them have a nice linear correlation.

This is up into the right, but there's two data points out here out in the margins.

And, uh, those could be somebody that just wasn't paying attention when they were filling out a survey, or they misinterpreted what the max versus bin value was, uh, on the survey instrument.

And doing an analysis of an identification of outliers is an important first step you have to do in data cleansing.

So you take a look at your data, many students will just do a scatter plot and look for major outliers, but you can actually calculate this by looking at the difference between the median and mean.

And if there's a significant difference between the median and mean, it typically means you've got substantial outliers.

They're gonna influence your, your data findings, or your data isn't normally distributed, one or the other.

Um, but, uh, this is why the different, um, measures of central tenancy are important.

You want to be able to compare them to see if there's significant differences in them, because significant differences in them typically mean that you've got obstacles to linearity.

And you may have to do some data cleansing.

You may have to do some data transforms.

Uh, but the very first step you do with your data is do this analysis of mean media and mode, and look for substantial differences between them before you ever do any correlation.

'cause if they, if you've got significant differences in these measures of central tenancy, um, you're gonna wind up with a, a false finding in your correlation, uh, analysis.

Two other measures, uh, that are really important to do.

Um, most correlation designs assume that there's a normal distribution in the data.

And I think everybody knows that normal distribution is this nice clean bell curve.

The reality is most data sets aren't perfect bell curves.

Um, they, but they, they generally will approximate them, but it's not uncommon.

In fact, it's very common to have a data set that is skewed, meaning that the, the, the peak of the, um, peak of the bell curve is skewed to one side or the other has a positive skew.

If it's to the left, it's a negative skew if it's to the right.

And if you have data that is substantially skewed, then if you were to do a linear, if you tried to do a, a comparison of that data with a, uh, a normal distribution, you'd find that you wouldn't find a good correlation, even though there probably is.

It would be typically you would do a transform of this data, um, uh, a log transformation or an exported transformation to normalize the data, you know, shift that SKU back over.

You're not invalidating the data when you do a transform.

What you're doing is asserting that that relationship is no longer linear.

It might be a logarithmic relationship, it might be an exponential relationship.

So by normalizing your skewed data, um, you can continue to do your core, your linear correlation analysis.

But because it was transformed, you now make a different statement about the relationships between the relationship between the two variables.

It's no longer linear. It may be, uh, logarithmic or exponential or some other, um, relationship depending on the transform you used Dr. Rice. There's a question.

It, um, it states, is it best to narrow our topic to help the sampling for data? Um, that's narrowing a topic is something we always do with doctorals, because everybody wants to boil the ocean.

Um, uh, the, the most important thing you do in, um, correlation analysis is I clearly identify the two variables you're comparing.

So the, if those two variables help you describe your topic, then you've done a great job.

If it takes three variables to describe your topic, that's okay.

Now you're doing multi correlation analysis, that's fine.

The broader your topic, the more variables, the more complex the correlation analysis is.

I, I encourage students to try and find a significant relationship that they wanna study and focus on that.

Um, if that means narrowing the topic, that's means narrowing the topic.

But sometimes it's just being very thoughtful about the type of variables.

You remember, we talked about the different types of variables, independent dependent, uh, but also the mediating variables, et cetera.

So, uh, you may want to have five variables with three of 'em mediating or moderating, um, that you're doing an analysis with that you're, you're complicating the design.

It's a more complex statistic, but it's all the same basic principles.

It's just kind of a rinse and repeat with, uh, additional variables.

So I I, for a dissertation, I would encourage students to try and get down to two variables and, and focus on a, on a correlation, and used that multi correlation analysis as a great postdoc study.

So, so Another question has jumped in the chat.

How big are the sample sizes? Um, they're asking for a range.

Uh, the, the sample size is entirely dependent on the size of your population.

Mm-Hmm. So the, the power analysis that you do will determine the size of the minimum size of your population.

So if you want to study the entire, the population of North America, um, you're gonna find that when you do a power, once you've done a

lot of power analysis, you start to see patterns in them.

And there's what we call a, a terminus, um, value.

Your populations can keep getting larger, but to get a representative sample, you start to, to terminate around 350 members of your sample.

I, um, I just completed a study for all parents in North America, and I had 360 people in my population to get the, um, minimum sample size, but also to get the, um, uh, the representative nature of the genders mixed into there.

So I had to over sample to get the enough fathers in that study.

Um, so I have seen samples as small as, you know, 60, 70.

Um, it, it depends a lot on the size of your population, but it also depends on the access you've got to the population.

Some populations are just very hard to study.

And so you may set a confidence interval at 80 or 85%, um, to get good insight into the data, but not as strong, uh, a finding, not as significant a finding, but it's the only population sample you can get to because the population's very hard to reach.

Uh, so it's a conversation you would have between you and your methodologist to make sure that you've got, you clearly understand the delimited population, you've agreed to the confidence interval and error rate, and you do the power analysis to get the minimum sample size.

So it, there is no standard range.

It's a very analytic process to get to the, uh, minimum sample size that answer the question.

Um, I believe it. Did we have an, okay, great.

But then there's a question about, um, bias and sample size.

I think you might have addressed it earlier, but can sample size cause bias and data Sample size does not bias the data.

It does. Sample size does determine what the significance of your findings.

So you may have a stronger ability to accept or reject the null hypothesis based on the si the, the, uh, by, based on the size of your sample, larger sample sizes will give you more significant findings.

Smaller sample sizes will give you less significant si findings.

And if your confidence interval drops below, drops too low, then it's not a study that you can, uh, base any real, uh, mitigations on.

So, um, it's rare to see a, uh, population minimum size go below an 80%, uh, confidence interval with a 10% error rate.

You start to get below that, it's almost a random number generator, So thank you.

Uh, okay. Um, so we talked about skewing.

Uh, so, uh, bell curves that are skewed, uh, are actually measured by a ANA or by a number called skewness.

That really is a number.

Um, so you can, when you're in SPSS, when you're in, uh, SaaS, uh, or any of the other statistics packages, when you're analyzing your data set, your, your population sample, uh, you will, you can kick out a number that call skewness.

Uh, it varies from minus one to zero to one.

Uh, it really depends on how far left or how far right it is.

Uh, the closer your skewness is to zero, the better.

Um, kurtosis is the other one.

Uh, kurtosis means that your bell curve is either really spiky, really tall and skinny, or really, really flat.

So if you have a kurtosis of zero, it's just a flat line.

Um, what you're looking for, uh, in kurtosis is having roughly a, a, a standard normal distribution.

And again, you may do a transform transformation to the data to get the, the flatness of the, uh, bell curve to be more normal.

Uh, to be closer to a normal, you may lift up and exaggerate the medians points, or you may flatten the median points.

And again, it, it is a nonlinear distribution if you do that, or nonlinear, uh, uh, correlation analysis if you do that.

Um, I would s most of our data sets, we don't do a lot of kurtosis adjustments, but it's pretty common for us to do s skewness adjustments.

Uh, we have a lot of data sets, particularly at like student testing, uh, where students will do really well on the test or do really poorly on the tests and your data sets get skewed, uh, and you don't have a good normal distribution in your findings.

Wow, did we just kind of fly through sampling and, and or fly through, uh, descriptive statistics? So again, descriptive statistics are all about looking hard at your data.

Before you do correlation, you need to make sure that the, the profile of the data set that you have matches the expectations of the correlation statistic you choose.

Uh, so let's talk a little bit about correlation statistics.

So when choosing a correlation statistic, you have to understand, and these are all essential attributes to choosing a correlation statistic.

You don't get to just pick a random one.

The, uh, the correlation statistic is, is driven by the nature of your data.

You have to understand your variables.

What's the variable type? Are they continuous ordinal, di economical? Um, check your assumptions.

If you assumed that you are, you were gonna get, uh, normal data distribution, a bell curve, uh, because you were thinking that you want to use a piercing correlation assessment.

You have to, you have to test that.

You have to make sure your data is actually in a normal distribution before you can apply Pearson.

Um, and that's just one example.

Um, and also consider the research question, what are you trying to accomplish? Uh, because different correlation statistics will tell you different things about the correlation.

So you may choose to use more than one correlation statistic depending on the attributes of your data, but you can't just choose a random set.

It has to match the, uh, the profile of your data set, and it has to match the data types or the variable types that you've got in your, uh, data.

So let's talk about some different correlation statistics.

Um, Pearson, if you, if you're a person that wants to use the Pearson correlation coefficient, and I, and I, I, frankly, I just like him as a researcher, so that's wonderful.

But, um, your data set has to be continuous variables.

So if you have nominal data, or if you have non-continuous data, Pearson correlation would give you a false or false, uh, finding you, it is also assuming that the variables are normally distributed or close to normal.

So if, if you do your descriptive statistics and you find your bell curve is really skewed to one side, you're gonna have to do a transform to bring it back into a normal curve before you're gonna apply Pearson to it.

And Pearson is also very sensitive to outliers.

So if you haven't scrubbed your data, then uh, you're gonna get, uh, an incorrect correlation coefficient.

If I ran Pearson on that, uh, scatter plot that I showed you earlier, had those two outliers, I guarantee you that as linear as that relationship looked, uh, Pearson would not report a linear relationship.

Um, so you're gonna have to make sure you do you scrub your data.

Pearson's very sensitive.

Um, it's a very simple correlation to run.

Um, but again, the variable types, the nature of the data distribution and sensitivity to the outliers have to be considered.

Spearman's, on the other hand, is a non-parametric measure, meaning you don't have to have continuous variables, you can have ranked variables.

Um, it's, it's really more suitable to ordinal non nominal, uh, distributed data.

So skew data is fine, um, uh, data with high, um, not only high skewness, but uh, high skewness is fine and it's less sensitive to outliers, but, um, it's not gonna give you as, um, significant of finding because it's, it's, it's going to lean more

towards a correlation or, or a strength of relationship rather than correlation.

Uh, Kendall's tau, uh, I've actually used this more lately.

It's, it's actually a great statistic.

It's for non-parametric data with rank ranked variables.

So if you are looking at, um, uh, the ability of somebody to succeed based on their education level, you know, a variable that's got discrete values, um, Kendall's Tau is a, a powerful tool.

Um, it's really robust.

Um, uh, it's insensitive to outliers in ranked variables.

You don't have a lot of outliers, but, um, it, it's very insensitive to those.

And so, and it doesn't require normality.

Uh, so this is, this is one that I've been, I've been playing with actually on even, uh, uh, places where I've used Spearman's historically, just to check the Spearman correlation to see is, you know, those two seem to be a good pair.

Uh, they, they do a good job of supporting each other because tau is so insensitive to outliers.

Um, it will give a better, um, answer For nons scrub data, you still need to go back and figure out why did I have all those outliers? Uh, go analyze your data, but it's kind of cool.

Um, fi coefficient, cre, I'll let you read through the deck and look at the different attributes, but always make sure that your data variable types, your data distribution and, and profile and the, the nature of the data, the size of the populations, et cetera, are consistent with the correlation coefficient that you choose to run.

All of these are great coefficients.

They all have, uh, equal value in describing the relationship between variables.

Um, but they are accurate or not accurate depending on the profile of the data that, um, they're expecting to see.

So if you feed a data that is garbage from its perspective, you're gonna get a garbage out response.

So make sure you've got good data in that it matches the assumptions that the, uh, coefficient has.

Um, so getting wrapped up here on the end.

So all correlation coefficients are gonna return a value between minus one and one.

Um, basically you're looking for whether or not, um, the correlation is positive or negative, meaning that it's a, and I'll give you visuals for that in just a minute.

Or the closer to zero, it is the, the weaker the correlation is.

So on the right hand side, these are kind of the norms, uh, in most of the research.

If you get a, uh, a correlation coefficient that is, uh, above a 0.7 that or below a minus 0.7, that's considered a strong correlation.

Um, medium correlations tend to be between 0.3 and 0.7 weak correlations below 0.3, but I will tell you there's some fuzziness around that.

Depending on your industry, depending on the type of field you're in, you may find, uh, some that require a 0.8 or 0.9 correlation to be considered a strong correlation.

Um, so this is where going and looking at, and remember in chapter two, you're going and looking at, uh, other studies that are, are in your field and using your type of design.

What you're looking for is how do they define the strength of these relationships so that you can support what you're defining as a strong relationship, medium relationship, or weak relationship in your findings.

You wanna be consistent with your field if you want your study to be accepted.

Um, direction of correlation.

This really just says whether the correlation is positive, meaning that as one variable grows, the other variable grows.

Um, the, a negative correlation is where one variable grows and the other one declines.

Uh, so, uh, an inverse relationship.

So the upper row there, the first three, there are examples of positive correlations.

And you can see as the r value gets higher or closer to one, that the, um, the, the data points that scattered plaque gets closer and closer to a line.

The, the further out those scattered, those, uh, those, uh, that scatterplot goes from that linear relationship, the weaker the correlation.

So 0.8 has dots further off the line, 0.5 has lotts further off the line, and another R zero, uh, or frankly anything below it, like 0.1 is gonna be considered no correlation.

And then you can see where the line goes down.

Those are negative correlations.

So the, the correlation coefficient has a negative number in it.

This tells you a lot about the relationship.

So if I had a positive relationship between, uh, shoelace materials and um, uh, and untie rates, then I could, I could positively say that by changing the shoelace material, I can predict that I will have more or less, uh, untied shoes during a marathon.

Uh, so these correlations can have, uh, really strong implications, uh, when we step back and look at how to practically apply them.

If I'm a marathon, um, organizer and I see that I've got a much lower untie rate with a shoelaces material, I'm gonna give out shoe laces with the right material in it to all of my runners so that I can pay less

for my run, my race insurance, uh, 'cause I'll have fewer injuries from untied shoes.

So these kind of studies can really have, uh, powerful and profound implications.

Um, great studies to do, um, significance.

Now remember I said, once you're done with your correlation analysis, you need to go back and say, this is how many people were actually in my study.

This is where I go back and say, okay, I see that I have a relationship here.

I have, I am asserting or disprove, I'm asserting or not asserting my accepting or not accepting my hypothesis and null hypothesis.

But how confident am I in that? And in order to have a high confidence that I can reject the null hypothesis, I have to make sure that I test the significance based on my sample size.

And if my significance, my p value is lower than my confidence interval, then I can reject that null hypothesis and I can assert that my, my hypothesis has been accepted.

Um, this is really how you can assert that you can accept or reject an all hypothesis.

You have to both do the correlation and you have to test the significance of your correlation based on your sample size.

Um, so don't forget this step.

Too many people forget this step.

They go see the correlation, say it's a strong correlation, and they go, oh, I've accepted my hypothesis.

And nope, can't do that.

You gotta tell me what your significance is so that I know what confidence level I have in, uh, your decision to reject the null hypothesis.

So, um, people that have been doing statistics forever do this all the time.

It's, it's basically doing a normalized, uh, uh, a z do, basically doing a Z test and looking at the Z value in a, in a statistics table, or if you're using SaaS or SBSS, it just kicks out the number for you.

You're really looking for a p value that's lower than your confidence interval.

And if that exists, that means you've had a large enough sample to truly accept or reject the null hypothesis.

That's really what it's saying. Did I collect enough data to accept or reject my null hypothesis? So, um, there's a really fun tool.

Um, people will always struggle with that P value, and there's a fun little toy out there.

It's a website that describes this, accept or reject all hypotheses using, uh, Santa Claus and cookies under the tree.

Uh, it just kind of fun to do, uh, and I encourage people to just go out and give it a try.

You know, I, I put the web, I put the link up here, it's on a, on a site called towards data science that I like to haunt.

Um, but, uh, it's just a fun way and a, and a kind of a practical way to go look at why we do this, um, significance test after we've done the, uh, correlation analysis again.

So I am, I am over my hour, uh, I'm almost exactly an hour.

'cause we started about five minutes after.

But this is the, these are all the topics we touched on.

Oh my goodness. You did you actually think we get through all these in an hour? Um, I just need to say this is, has nothing to do with, uh, teaching anybody how to do statistics.

This is just all the elements of a correlation study.

Know your variables, make sure your problem, purpose, uh, design, uh, questions, hypotheses are all aligned.

You've done your, uh, it, it is very sensitive to understanding your population and your sampling controls.

'cause you have to get to a significant sample.

You have to do your descriptive statistics before you do your correlation statistics.

Make sure you're using the right correlation statistic, and then understand how to interpret the data.

Wow. I like how'd that I, I, I ran a couple minutes over, but did I, did I survive? Dr. Smith? I think you more than I think you conquered.

I think you did More than survive.

So I, um, thank you Dr. Rice for that amazing, uh, presentation and that survey of correlational research and that primer in statistics, even though it wasn't meant to be a primer in statistics.

I, I know that that was helpful for students.

Was helpful for me. Um, are there any questions quickly? I know we're over time, but there's a whole bunch of accolades in the, in the chat.

Dr. Rice. Hi. Hi. I miss you. Yeah.

Hi, Dr. Rice. This was wonderful.

Um, because I am at that time for four and five, so you were so perfect for me.

But my question is, um, are the notes you just showed us available? Yes. So I've, I've sent them to Dr. Smith Mm-Hmm.

She will wrap these up with the recording and she'll post 'em out on the research hub and they'll be a, you can go look at all the past webinars that the research hub has put on, including this one very shortly. Yeah.

'cause I'm right up to choosing that Pearson or One of the others. One of them. So you're so apropos for my needs. Thank You'll be

Oh, oh, I'm so sorry. It'll be Quick. No, I was just gonna say glad I could help. So Yes, it'll be quickly up on Research methodology group.

Mm-Hmm. It takes a couple of more days for it to be posted in the research hub, but it'll be posted on research methodology group probably tonight because I, since the recording and I can access the information that you shared, Dr. Rice.

Okay. So you'll have, you'll have it quickly.

Yeah. And the PowerPoint is in the files tab Mm-Hmm.

Of the correlation channel on the research, uh, center.

So, uh, you can actually get it from there as well.

Thank you. Thank you. Any other questions? Lots of thank yous.

This is, this is sadly an underused design.

I'm really excited to see so many people, uh, engaging with it.

'cause it's, it's really powerful.

It's really not hard to do once you've gone through it once.

It's fairly straightforward.

Um, I would say it takes a little longer to get your proposal done, but typically your data collection and analysis get done very quickly.

Mm-Hmm. Um, so it's a, it's a great tool for a dissertation.

Okay. Thank you. Have a good evening.

Goodnight. Good night. I'm stop the recording right now.

Okay. Here we go.