The Podcast *Quantitude*

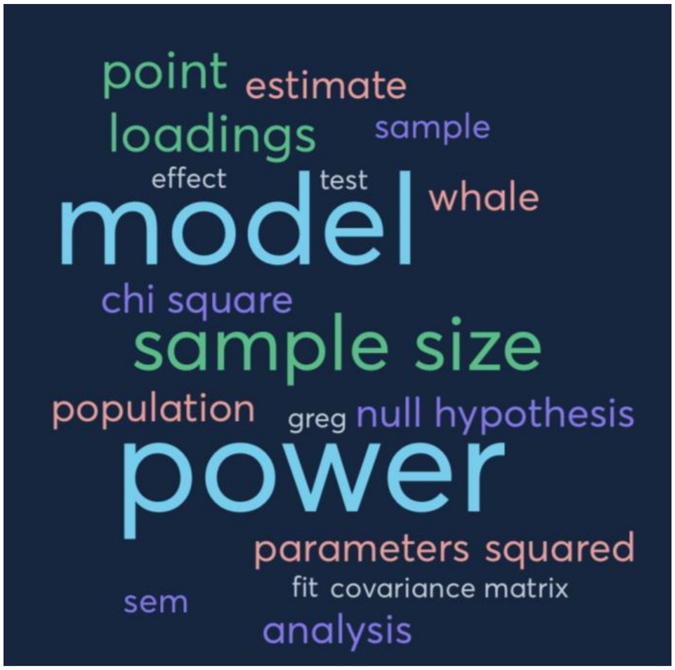
Greg Hancock & Patrick Curran

Season 4, Episode 13

*S4E13 Model-Based Power Analysis...The Power of *What**

Published Tuesday January 10, 2023 • 54:01

SUMMARY KEYWORDS
power, model, sample size, loadings, point, whale, population, R-squared, estimate, parameters, chi square, analysis, null hypothesis, SEM, sample, covariance matrix, test, effect, fit



Patrick 00:05
Welcome, my name is Patrick Curran and along with my dark Kherson Campbell lightning friend Greg Hancock, we make up quantity. We're a podcast dedicated to all things quantitative ranging from the irrelevant to the completely irrelevant. In this week's episode, Greg and I revisit a topic we addressed in our second ever episode statistical power. Here we continue our discussion by attempting to clarify the power of what and we explore ways of obtaining meaningful power estimates using the structural equation modeling framework. Along the way, we also discuss tearing arms off German dentists, booby prizes, Dr. Strangelove, making it look like an accident. Shrug emojis, the whale petting machine, baseball and war. Where's Waldo? Whale holes, the big R squared, throwing reviewers against the wall. DIY power. In fairness to me, egg plants. And screw you guys. I'm going home. We hope you enjoy this week's episode.

Greg 01:08
I'm looking at you to see how you're doing. But I can't really tell you're very stoic.

Patrick 01:12
Oh, dude, I am falling apart at the seams. What's going on? You know, from when I came up to visit you over the summer, were you surprised to me by saying I'm going to take you to this place, you're going to love it. And I was foolish enough to say where we going and you said you'll see when we get in the parking lot. After we drive. 45 minutes, we drive 45 minutes. We pull into the parking lot. And it's one of those parachute places where they blow a fan and you ride on the air. I have two immediate thoughts. One is this was so thoughtful of you. And the second is you're going to tear my arm out of its socket because I have a severed rotator cuff tendon which I had failed to tell you about. I am finally sucking it up and then a couple of weeks I'm going to have reconstructive surgery on the shoulder as a booby prize a couple of days after I scheduled the surgery, I had to have an emergency root canal.

Greg 02:13
Oh boy,

Patrick 02:14
I don't want you to picture something. All right, you and I are producers on a movie. We say we've got a couple of minutes scene, but I need you to send a brilliant endodontic surgeon who has a German accent and seems to way overly enjoy her job. And central casting says, Okay, I got it. This wouldn't be the person who I went to about 45 minutes into it. She starts talking to herself. It's all done through a microscope. So she has this big microscope that goes off a mirror as she's working on my tooth. I can't do the German accent. Maybe I can say some lines. And you can say, okay, so I want you to say you're the perfect patient, you don't move at all, you just leave me alone with my microscope,

Greg 03:05
you as a perfect patient. You just leave me alone with my microscope to do this work that I'm doing on your tools right now.

Patrick 03:15
That was pretty much it. Okay, I have never had so much confidence in a health care provider. She was amazing. But oh my god, she just scared the living crap out of me. And at one point, she stops now I'm going to try Okay, ready, I'm gonna buy there she stops and she says too complex. You must come back. She throws in a temporary filling. And I have to go back next week. That was my Oh, you're about to have shoulder surgery. Why we have something for you. Wait, don't answer yet. Dr. Strangelove, is going to do a dental procedure on you.

Insert 03:54
The whole point of the Doomsday machine is lost if you keep it a secret, right? I didn't tell them.

Patrick 04:01
So yes, thank you for asking.

Greg 04:04
I mean, at some point, we're just going to have to have you put down

Patrick 04:06
Almost verbatim that's what one of my kids said. And the other one said make it look like an accident or else we won't get the insurance money. But what it did make me think about and it's what we're going to segue into today's conversation, you can start something and then at some point, you say, This is too complex, and you just stop. We did that with our power episode. Ooh, taken out of context, I might have advocated for using emojis for power analysis, one of which was a poop emoji. And somehow an eggplant came into play as well. Came off the rails a little bit. But we in some ways were misconstrued because I have heard multiple people say, Well, I know Greg and Patrick are anti power. Yeah, right. Right. And that is patently untrue. Uh, I am not anti power, you are not anti power. But what the punch line of that episode was that came up multiple times, because we're really bad at editing. And if we'd

redid it now, it only would have come up once is the power of what Yes. And I expressed frustration in being part of grant reviews where there is a multiple group bivariate latent curve model with structured residuals. And the reviewer says, The applicant must demonstrate adequate power. And I don't even know what that means, because it is too complex to complex.

Greg 05:39
Let me see if I can pull all this together, you are likening power analysis to a root canal and that we didn't finish the job. The first time around, we got into power analysis, but was like episode two.

Patrick 05:52
I know I just I have repressed the first two years.

Greg 05:57
I would characterize that as what I might call a curse the darkness episode, we were grousing about power analysis, power analysis. First of all is hard. It is complex. And to try to distill it down into one thing is just foolish. I think we talked about the complexities of it a lot. But we didn't light the candle, because at the end of the day, that root canal needs to be finished or you're still going to have a problem. So I think what we could do to pair with that maybe some suggestions for actually how to conduct a power analysis, how to think about it, because we have to do it. And to be crystal clear. You and I support doing power analysis.

Patrick 06:34
You mean there's more to academia than cursing the darkness? Dude, I've been in the game for like 30 years and nobody has ever talked to me about lighting a candle. What kind of boat is that?

Greg 06:46
Let's put on some acoustic guitar Patrick and light some candles. And that probably is going to require us to do a speed recap of power analysis. In a nutshell,

Patrick 07:02
I think that this is going to involve the whale petting machine.

Greg 07:07
If anybody gets that reference, you've either taken a class with Patrick and God help you. Or you listen to one of our very, very early episodes. And again, I repeat that, yeah, either way, you're kind of screwed. All right, tell us about your whale.

Patrick 07:21
Every example I ever use either uses baseball or war. It turns out that those work really well as examples. Unless you have no interest in baseball, or war. The backstory briefly is I talked about type one and type two error. And I talked about an enemy submarine and you're in a submarine. And you have to know whether the enemy sub is out there. You don't want to miss it if it's there, because it puts you in danger. But you don't want to give away your position by responding in a way to something that's not there. Some students very good naturedly said Is there any example that you can use that doesn't

involve one person trying to kill another person? I came up with the whale petting machine, which is you are underwater and you're not looking for an enemy submarine. But you're looking for a whale. Everybody knows whales like to be petted. You don't want to miss a whale that goes by without petting it because it's going to make it sad. Now in the prior conversation, you told us what a sad whale was. So go ahead, or was that me? I forgot you did.

Greg 08:28
You. Alright, let's hear what remind us Patrick. What a sad whale sounds like that sounded better before you needed a root canal. Oh, you

Patrick 08:38
want me to make the whales out? Actually, I was just living my two. So you don't want to miss a whale? That's really there. But what's super important is you don't want to extend the Petit machine if a whale is not there, because it scares away the other whales. That's power.

Greg 08:57
Okay, it's crystal clear everybody to clarify that a little bit. Do you have any other language you could use?

Patrick 09:04
Or Starburst? Or Starburst or whale man God now we're gonna hear from Tove about that one, aren't we? Spear horned bears here.

Tove 09:14
Hi, this Tove Larsen faculty member in applied linguistics and quantitative Swedish consultant. I am happy to report that neither listener in Sweden was offended. Right? Hendrick nodded.

Patrick 09:25
Okay, here's the deal, folks. This is going to be a whirlwind tour through a frequentist perspective on no hypothesis testing. Right. Levy pay attention. All right, maybe you'll learn something out of it.

Roy Levy 09:37
Hey, this is where we live in Arizona State University. When you start thinking about statistical power, a Bayesian perspective can really be

Patrick 09:44
Nope. Null hypothesis testing is like Where's Waldo. Okay, we have two conditions in the population that exist. Imagine that we have a treatment group and a control group and you put 1000 hours of blood sweat and tears into your dissertation. You do an intervention with kids in schools to try to improve reading and you want to know, was your reading intervention successful? Did it improve reading to a greater extent than you would expect the by chance alone? The null hypothesis now we're going to super stressed this in the population is the two population means are equal. Okay? There is no difference in the population. The alternative hypothesis is we're just a roomful of cowards. And we say, well, I'm going to die on the hill of the null hypothesis that they're equal. But if the null doesn't hold, I'm

going to say the two population means are not equal. Ooh, yeah, very good. See, so either they are equal, or they are not equal. So that's the null hypothesis and the alternative. Now, that's what exists in the population that we believe to be there. But that we don't have access to we want to make a probabilistic inference about that. So now think about the decision that we're going to make you do your 1000 hours of blood, sweat and tears, you find the treatment group had higher reading skills than the control group did. So now you have to make a decision? Are the group means different from what you would expect by chance alone? Or are they not picture a little two by two contingency table? The columns are the population, the no holds, or the Nolde does not hold? And the rose is, what did you decide? Did you accept the null? Or did you reject the null? And that's the whale petting machine? Is there a whale there? Is there not a whale there? Did you extend the petting machine? Or did you not extend the petting machine? Do you understand? Now, Greg, I could not have been more clear with the whale petting machine. I'm not sure I needed all this other crap.

Greg 11:56
I'm sad having to listen to that explanation. But go ahead. Yeah.

Patrick 12:00
We actually have in that two by two table to correct decisions. And we have two incorrect decisions or errors. And if there are two of them, I got a hankering to call them a type one error, or a type two error. That's where those terms come from. All right, well, what are the correct decisions? The correct is you conclude there is not an effect when there really is not an effect, or you conclude there is an effect when there really is an effect. But what are those two errors, and this is what we pay a lot of attention to in classes and in work is an error of the first kind type one is what our standard p value is focused around, which is what is the probability you're going to say there really is an effect when there is not a false positive, you extend the whale petting machine, and there's not a whale there, and you scare all the other whales away? Alright, what does it mean for your dissertation? What you concluded the reading intervention was successful, but it wasn't Yeah, type two error is you say that there is not an effect when there really is an effect. And that makes the whale sad, because there is a whale and it wants to be petted, and you're not going to pet it. And whales are highly ruminative. And it's going to go back to its little whale hole that it lives in on the ocean floor. And it's just going to ruminate for the rest of the day.

Greg 13:23
Where did you go to school

Patrick 13:25
Colorado?

Roy Levy 13:27
Not a lot of whale holes,

Patrick 13:30
what we focus a lot on is type one error, what is the probability that you're going to reject the null if the null is actually true? Alright, so you say there's an effect when there's not. What power is, though is

what is the probability we're going to reject the null if the null is false? What that means is, what's the probability that you're going to find an effect? If an effect really exists? That's power?

Greg 13:59
Well, it's sort of power in the following way, right? When you set up that dichotomy of the null hypothesis is true, or the null hypothesis is false. That all hypothesis being true is one case. The null hypothesis being false is many, many, many cases, it's all the cases where mu one and mu two are not equal to each other. And when we talk about power, it's not the null hypothesis being false it is the null hypothesis being false in a specific way to a specific degree that seems to get lost sometimes in the way we talk about things. So power has to do with the probability of rejecting the null hypothesis when not only is it false, but it is false to a very specific

Patrick 14:41
degree power is easiest to think about and to learn and to get our head around in this two group kind of scenario. We have two population means are they equal or not? We have two sample means are they sufficiently different that we would not attribute those differences to chance alone? If For the null hypothesis was true, we don't do the field a service because what we do is we teach power in these hyper contrived situations, you have to group means you have a correlation, you have a multiple r squared. And I feel like that's the framework that we get for power. All right now all of you as you're driving or mowing the lawn or cooking dinner, or whatever you're doing right now, think about the work that you're doing right now, not hypothetically, not sometime in the future that you're doing right now, how many of you are doing your entire research project comparing two means? Nobody? Nobody out there

Greg 15:36
sit on a grant panel? When is the last time that you saw a two sample T test? When is the last time you even saw a multiple regression? And the answer is never. But those are the things for which we are at best trained to think about power. But even if I take a multiple regression, it is almost always the power associated with the omnibus R squared associated with the whole multiple regression model. How many subjects do you need to be able to have enough power to get a statistically significant R squared to be able to proclaim that there is some nonzero population multiple correlation coefficient? That's all fine and good. But my research question is not about the big R squared. My research question is almost never about the big R squared. My research question is almost always about the individual predictors, right? To what extent is this helping me to understand why above and beyond these other things that I'm controlling for? And already, we're asking you to think about power in things that transcend the level that you're typically formally trained to do? You and

Patrick 16:38
I have talked ad nauseam about the structural equation model about the generality of that about how a lot of things that we do t test to know Vancouver regression, CFA, all of these fit into that framework, not everything, but a whole lot of things. And it turns out, there is 30 freakin years, maybe 40 years of work on power analysis within the SEM, but you don't see that a lot. And I think what we should do is pivot into the year though candle

Insert 17:10
light a candle, Patrick? Can the burpees light a candle?

Patrick 17:16
Let's revisit some of these classic concepts. But through the perspective of the SEM, because what I was grousing about on that earlier episode is a grant reviewer says the applicant did not adequately demonstrate sufficient power to conduct the analyses proposed here, and I throw him against the wall, not the reviewer. One time, the judge told me those records there. We can't talk about that. The reviews I throw against the wall. And I say the power of what? Well, what we're going to talk about within this SEM framework is it's going to involve Coors Light Green Day in my garage, I know

Greg 18:01
Yeah.

Patrick 18:10
You're gonna go out in the garage, and you're gonna build the damn model to get the damn power effect that is specific to your hypotheses. So we can reach inside these really complicated models, and say, This is what the power is, we are going to use the SEM framework as a calculator to get the power estimate for exactly what we want, not what G power gives you. Not what colons table says, which are brilliant, totally. And those are hugely advantageous, unless you're doing any of the work that we all are doing.

Greg 18:47
Those are great for stuff you don't do. Yeah, yeah. All right. So to draw on the analogy, from regression, when we talk about fit in regression, and how successful a particular regression model is, as a whole, the R squared is often our go to write a big R squared, we go yay, good for us. I don't know which predictors are doing it. But yeah, good for us. And a small r squared, and we are sad. And so the power for regression is often treated as power of this whole rather than power of the different parts inside. Within structural equation modeling. We also assess the model as a whole, but we don't use an r squared, we use Fit indices fit indices, like a comparative fit index, or a root mean square error of approximation or Hakka model chi square, right? There are many, many different ways of assessing fit of a structural equation model. And there's a lot of question about how one should use those if one should use those. Now, if I just asked a question about power at this omnibus level, one way to think about it is how many subjects would I need to be able to say this is a good model? Right? That seems like a reasonable question because we all want good models. But the problem with that generally speaking, the chi square is a funny index in the sense that bigger values of the chi square test to you that your model is doing worse, and smaller values of your chi square tell you that your model is doing better. So if someone used a chi square as a characterization of fit of a model as a whole, and said, Now, how can I use that in some sort of a priori power analysis, the answer I would have at the end of the day is just get a really small sample, right, because if you get a really small sample, you will tend to get a really small chi square. And so you will tend to say you've got a great model. And that should bother you on some level, deeply. There's sort of a logic problem that we have here with the chi square. And instead, what we do is we kind of flip the logic from the chi square, the index that is used in part of this process Most commonly, and it absolutely doesn't have to be the one that's used is the root mean

square error of approximation. The root mean square error of approximation has a known distribution. And here's the logic of power analysis at the omnibus level using the root mean square error of approximation. Let's imagine that you set for yourself some standard that you consider to be the barrier between good and evil, right, bad fit, good fit truth.

Insert 21:05
What it means is Old Testament fire and brimstone coming down from the sky, human sacrifice dogs and cats living together enough, I get the point.

Greg 21:14
What do you use? Patrick? What do you think of is the cutoff for RMSEA. As

Patrick 21:17
I wrote a paper that shows there is no universal cut point, the impact on the field was zero. So I'm just going to cross my arms and sit here grumbling,

Greg 21:29
someone might say, oh, we'd like to use point o five as the cutoff for the root mean square error of approximation, or we'd like to use point o six. But the thing is, you have to pick some value in this world. And what Patrick said about maybe there isn't a good value, I think, is part of the challenge with this. But imagine you did lock in, let's say you locked in and said point o five. So what you do is you imagine what your true models RMSEA is, how on earth would I know that true? RMSEA for my model, the first thing I say is exactly. But then the second thing I say is that's not really different from other power analyses, where you have to suppose something about what's true in the population. So we're not really being held to a different standard here. It's just a different numerical description. And so imagine that you say, I believe that my model is perfect in the population, I say, okay, great. It's not but okay, good for you. If you were to take samples of size, let's say 50. And do studies over and over and over, you will get a confidence interval around your samples, root mean square error of approximation. And when you have a smaller sample size, that confidence interval would tend to be much wider. And when you have a bigger sample size, that confidence interval would come in and be tighter. The operational question is, what sample size would you need? So that let's say 80% of the time your confidence interval is entirely below, wherever you have set that threshold, like an RMSEA of point oh, five. So if you say I want 80% power? The answer is what sample size would you need to shrink your RMSEA confidence interval down so that 80% of the time it is under that threshold? And essentially, you are rejecting badness of fit in favor of good fit. That's the logic of it. There's the realistic part where you say, but whose model really has perfect fit whose model has an RMSEA of zero and truth? And the answer is probably nobody. And so I say well inject a little bit of badness. But imagine that you're RMSEA. And truth is point O two little bit of badness if it kind of realistic, maybe and now the question is what sample size would you need to shrink your confidence intervals down so that 80% of the time they are below whatever threshold you've set for yourself, like point oh five? Well, it's going to take a bigger sample size now. Because truth isn't sitting down at zero truth is sitting up at point O two. And so oh my gosh, I'm gonna need a bigger sample size so that I'm under that point oh, five 80% of the time. That's what the point oh two, what if your model fit is point oh, four, right. And RMSEA of point oh four, in truth characterizes the fit of your model in the population. Which by the way, I might be pretty darn happy

with right, we can question whether or not the RMSEA is the be all end all characterization of fit. But if someone told me, frankly, at an RMSE of point oh four, I'd be pretty happy. Well imagine that's the truth in the population. What sample size are you going to need so that your confidence intervals around your samples? RMSEA are below that threshold of point o five 80% of the time, and the answer is probably you're going to need a pretty darn big sample size. Because your truth of point of four is standing right next to that threshold of point oh five, you're gonna need a big old sample size to get a tight confidence interval around that RMSEA. This is the logic of power analysis or sample size determination when you're using a fit index like the RMSEA. But there are a lot of challenges associated with this.

Patrick 24:43
We'll put this on the show notes but McCallum Brown and Sugawara have a very important paper in sight methods on all of this in 1997. What they did and what Greg did here in the telling of the story, we're actually turning two knobs at once in what we Think about. The first one is the exact fit versus close fit, right. So the chi square in our usual way is the null hypothesis is sigma equals sigma theta. That is the population covariance matrix is equal to the model implied covariance matrix. And that is the equivalent of mu one equals mu two. Now, as Greg described, and I won't reiterate, we come to a different conclusion. On that note, the t test is saying your intervention was not effective. And you have to bring empirical data to demonstrate that it was the Nolan the SEM is saying your model is correctly specified. And you have to bring empirical information to demonstrate that it is not Karl Popper is over there, mixing drinks at this point shaking his jaw, let's head out that little intellectual Judo that we did. But what we're doing is we're thinking about one, we've got this omnibus test of power for the model as a whole. And to do we do close fit or exact fit. All right, and what McCallum argues and just as Greg said, rarely, if ever, does a model fit exactly in the population. And so we're going to be, what are we going to give, we're gonna give a little bit of oh two, we're gonna give a little bit of both three. But let's go back to the R squared and regression situation, where we said, oh, you have some power to detect a given r squared. But that's for the model as a whole, you could have all of your predictors be significant, you could have one predictor be significant. Indeed, you could align things in a way where none of your predictors are significant, but you still have an r square to point to Yep. And you kind of sigh and say, well, crap, we are evaluating the SEM as a holistic entity. If we have complete data, we have a sample covariance matrix, we have a model implied covariance matrix that the model says this holds given the structure you gave me, we subtract those two, and we get a matrix of residuals. What we're doing in the test of exact fit is, are those residuals taken jointly larger than we would expect by chance alone under the null hypothesis? Well, we could have one bloody honkin, residual, and all the rest are near zero, or we could just have drips and drabs, right? The Irish loadings. Remember Oh, Curran, O'Shaughnessy? Oh, 703, right. If you're doing a CFA, we don't believe why to only loads on 801 and done more than any other factors. It may have a cross loading of O one or O two or three, but we're fixing those to be zero. Well, that omnibus test is just like the r squared, all we can say is somewhere within the confines of the model, we have a 78% chance of detecting and misspecification. I look at that and think, okay, in one way, super exciting. In another way, that's not what we're really interested in. We're interested in a treatment effect. We're interested in a non linearity, we're interested in the covariance between two growth factors, those omnibus tests don't give us that and the other poke in the eye is to do the entire RMSEA omnibus test, as you already alluded to, is we have to pick a value that we believe is indicative of close fit and RMSEA votes, you are oh three or Oh 405, there's

been a fair amount of work that shows we can't do that, because it varies over degrees of freedom and model complexity and determinacy and all of these things. So enter stage left one of my heroes in the field, Albert Satorra. I think Albert Satorra is one of the most important contributors to SEM. If you use robust maximum likelihood, you can thank Albert, if you use adjusted chi squared, a robust chi square you can think Albert, he has made unbelievable contributions to the field. One is he said, Wait a minute. under the null hypothesis, our test statistic follows a central chi square under the alternative, which has some misspecification. In it, it follows a noncentral chi square that Nan centrality parameter represents the degree of misfit from that misspecification damned if I can't turn that into a power estimate. And that's the Satorra-Saris method,

Greg 29:28
right? If we think about pivoting, then from power, and sample size for the model as a whole to power and sample size for the parameters, where our hypotheses are actually residing within a model, he laid the groundwork for all of that

Patrick 29:44
it's very similar to what we do when we compare nested models within the sample. But what we're doing using the Satorra-Saris method is we're doing that at the level of the population. Remember that in a likelihood ratio test for nested models, we have Model A ie that has some parameterization. We have Model B, where we can impose restrictions on Model A to get to model B. So they're nested, and they're each going to have their own chi square, we take the difference between them, and we can test is there a significant difference in model fit? Well, what Albert did is said, look, go up to the whiteboard and draw out whatever model you have. Now we're starting to talk about the power of what draw a growth model, draw a path model, draw CFA go nuts, man, draw whatever you want on that board, declare that to be your population model. Now, this is a little tricky, because you're gonna have to give me every parameter value. Yeah, your loadings are point seven, your regression coefficient is point five, your correlations are point three. Right now, this isn't no different than what we do in other kinds of things. When we say I have an effect size of point two, you're implicitly saying these are the values in the population. It's just we're up at the whiteboard now. And now it's power analysis on spring break, you got to write in communality estimates, factor loadings, all of these things, you're gonna get a model implied covariance matrix and mean vector that correspond to that model. Now, for your sample planning, go up and say, I'm going to remove these three parameters. And I'm going to get the covariance matrix and mean vector that that model implies. Now keep in mind, we're still at the level of the population, we don't have any sample data. That's right. But we have the covariance matrix of your population model, we have the covariance matrix of A misspecified model that's defined by a very specific Miss specification, and he showed how you can use the difference in those chi squares to calculate what is the probability you would detect that Miss specification, given the other characteristics of your model, it's

Greg 31:52
freaking brilliant, is what it is, you can do some version of it mathematically without running models, you can do it with software by running models, and just setting parameter values, it translates perfectly into the non centrality stuff that Patrick and I talked about previously. Because if you specify a proper population model, and then you take out a parameter and run the analysis on that population, you will

get an estimate of the non centrality associated with that misspecification given a particular sample size, and given the context of the rest of the parameters of the model that translates just boom, right into power into sample size stuff. And he laid the groundwork for everything really,

Patrick 32:33
because it's embedded within the SEM. And all of our traditional friends from high school are members of the SEM, right, we can use this approach for a t test or an ANOVA. And uncover a multiple regression or a multiple regression with an interaction. If you want to know what is the power to detect a three way interaction in a multiple regression, you can do it using this framework, but we can then generalize it to all the other things that we can do. And so way back in the day, Muthen and I have a paper back in 97, in Psych methods, where we use this method in a multiple group latent growth curve model to detect a treatment effect. And I wrote a little do loop, I'm not talking like some massive things, I wrote a little doom loop in SAS, I just went through sample size one at a time, up, up, up, up, up up from one to 1000. And I plotted out power curves, you can build power curves across what ever samples you want, whatever effect sizes you want. And this is the Coors Light Green Day garage. You can't say, well, Cohen, ADA doesn't have a table for a bivariate growth model, where my hypothesis is about the correlation between the two growth factors go into the garage, write a couple of do loops, I'm not exaggerating, because we're not doing Monte Carlo simulation. This is all analytic. I just referenced a noncentral chi square given a degree of freedom sample size and non centrality parameter and plotted the function, and that's your power. So when the grant reviewer says, Well, you need to give power for your bivariate growth model. You say here is the power that I would achieve with this sample size to detect a covariance between the growth factors this large or larger, there you go, man, there's your power.

Greg 34:35
I remember reading your 1997 paper pretty close to 1997. Honestly, I remember also thinking Dang, that's kind of hard to do. You know, if I were an applied person, am I literally going to write a do loop to be able to grind through this? I certainly could. And I appreciate you know, when you say yeah, it only took a few lines to be able to do but for some people that might still be a big ask a lot of the Torah And Soros things can be done with the loops that you talked about where you try different sample sizes, it can be done using the mathematics of power analysis or and this, I think is where things have started to go toward, we can hand everything over to simulation techniques. When you do the loops that you are talking about, or when you do the stuff that's the Torah and Soros have come up with the mathematics for, you actually have to specify the entire variance covariance matrix. And if I do that, what am I assuming I'm assuming a lot about the actual data that give rise to that I'm assuming that all the cases or there might be making some distributional assumptions, all of that it's a bit of a leap, maybe to think that you're holding the population covariance matrix in your hand, even if we grant that you got all the relations in the model, right? There are still issues about the data that we're kind of sweeping away,

Patrick 35:51
you were exactly right. This is an asymptotic method based on a mean vector and a covariance matrix. That means you have an infinite sample size, you have continuity, independence, normality, linearity, complete data, stop me when I've hit something that might be in your own study. So what we do is we say, well, wait a minute, all of us either have been in a class or teaching a class where we talk about if

you were to sample and fit your model an infinite number of times, and gather all the parameters together, we're gonna say, oh, wait a minute, I actually could do that and see what proportion I would say was significant. And that's power.

Greg 36:33
So let's give an example. What I want to make clear from the start is that we could pick any kind of model as an example from something that is as simple as the model equivalent of a two sample T test to a multiple regression with three predictors to a latent growth curve model with structured residuals. So what we're about to talk about generalizes, to all of those kinds of things. But let's still pick a pretty simple example, let me pick an example where I have a latent variable path model. And in my latent variable path model, I have three factors, I am just going to call them for simplicity, F one, F two, and F three, the model that I'm interested in is a latent path model where there's a path from f1 to f2, and a path from F two to F three. And then I also have a direct path from f1 to f3. Right? So simple structural model. And then I have indicators for all of those factors, we could just say three indicators for each of those. And imagine that this is someone's model, what they want to do. And when I say that, what I mean is what they want to be able to do sample size planning for so then the question is, what are the steps that someone would go through to be able to do sample size planning for a model of this type. So now someone has come into my office, and they have said, this is the study, I am planning, this is what it looks like, I want to know about the indirect effect of f1 on f3. And maybe the direct effect that it has as well. So this is where my theory is I say, great. So we need to plan your sample size, here are the things that we need to do to start the first thing now that we have it drawn up on the whiteboard, is I need you to go up there, I need you to tell me what you think the numerical values are of those structural relations that you're going to have to detect

Patrick 38:17
things get a little uncomfortable, but no more so than with Satorra-Saris,

Greg 38:22
exactly. Right. So no matter whether you're doing the mathematical version of this, that Satori and Sarris laid out, or, frankly, you're doing the power analysis for an ANOVA. And I asked you to specify the effect sizes for the differences associated with the means. Power of what you have to tell me the what, in order for us to talk about power. So you say, Okay, well, what metrics should I write it in on the board? So let's just use a standardized metric and say, Okay, I think this path here is a point five, I think this path here is a point three, and I think this path over here is a point six, is that okay? And you might have answered that based on your theoretical knowledge, you might have answered that based on sort of the smallest value that you would be interested in being able to detect and even that is the conversation that we have. So are we ready to do this? Well, there's some other things that we're going to have to do, because there's a measurement model associated with each of these factors. And so now I'm going to ask you, what are the loadings associated with each of these factors? You're like, what I have to know the loadings of these,

Patrick 39:20
you got to know how many items to put on the dam for how

Greg 39:23
many items do you put on there? That's right, three, five, how many you got? And honestly, sometimes people don't even know that when they enter into the conversation. But let's assume they've nailed it down and we'll go with the three items per factor only because they've said that's what we have, as past research told you what these loadings ought to be. Is there reliability information on any of those individual indicators that might give us some sense of how strongly it would load on a common factor? Okay, then by God, we might use that or is there some value that you think is sort of on the weaker side, but you're being cautious and Okay, great. Let's try these things. At least let's put those in as placeholders for now but I Have to get numbers on those loadings, do we have all the moving parts? Patrick,

Patrick 40:03
we've still done nothing different than we did with Satori. Oh, yeah, we are building the model we believe to exist in the population. So nothing is different yet, you're still at the whiteboard, you've still drawn your model, you've still put in your hypothesized values. So this is just a repeat of what we've done so far. But now here's where we're going to take a right

Greg 40:25
turn. And this is where it gets cool, right? I can put this into m plus, I can use a variety of our packages that you can do this, once you communicate your model, this will now become a simulation. And what I mean by that is that once you specify the nature of the population, if I say something to the program, like let's take samples of size 100, what the software can do with some assumptions is imagine reaching into the population drawing out a sample of size 100, running your model, computing estimates for all of the parameters, the loadings, everything, but you are particularly interested in tracking the behavior of those three key structural parameters from F one, F two to F three, we do that for the first sample of size 100. The second sample, the third sample, we do this some reasonably large number of times, I don't know, let's just say 10,000 times. And what I will get out of that will be an empirical sampling distribution of each of those structural parameters that I care about, as well as all of the other parameters that we have. So now the statistical question is, how many of those would be statistically significant? If I take the path from f1 to f2? How many of those reached statistical significance? And if the answer is 40%, then I'm like, Oh, heck, that's an estimate of 40% power. That's not what I want at all right? I want something more. So that tells me that, at least with respect to that parameter, I'm going to have to increase my sample size. But while I'm at it, I would check how many of the tests of the path from F to F three were statistically significant, how many of the path from f1 to f3. So I'm getting these empirically generated values for power. That gives me a sense of whether or not I need to turn the sample size knob up, or the sample size knob down. And what I learned pretty quickly is that there is going to be a weakest link in this whole model, that's going to be the one that holds everything up, right, there's going to be one parameter, the other parameters are going to be like, Hey, we're okay, we've got enough power. And you've got this other parameter that's holding everybody else up where you have to keep turning up the sample size knob to get that slowest one across the finish line. But that's essentially the setup for things under whatever assumed distributional conditions we have. Currently, I'll just say normal under the assumption of complete data and all of that this is essentially how it works, where we keep tweaking that sample size knob until we get all the parameter

estimates that we care about to have statistical significance of some threshold power level of something that we're happy with, let's say point eight, oh,

Patrick 42:50
you may say cheese sounds an awful lot like what you get out of Satorra-Saris and you would be right, if you meet assumptions, if you have a large enough sample size, and you have continuity, and you have normality. And you do what Greg just described in the Monte Carlo, and you do it 10,000 times. And you compute what is the mean number of parameters that I rejected at a P of oh five that is going to converge on the Satorra-Saris analytic method, they are one in the same way are they separate is when we start to violate those assumptions. So you're doing a longitudinal growth model in Satorra-Saris? Do you got a covariance matrix in the population, a covariance matrix under some misspecification in the population? Well, in this simulation, you shrug we do a lot of shrugging, and you say, well, let's punch out 20% of the observations in any given sample under an Mar assumption, let's make them a little skewed. Let's introduce a touch of dependence, let's put in a couple of Irish loadings. And now let's do it 10,000 times and see what proportion we would reject. Because now we have what's sometimes called an empirical estimate of power, where Satorra-Saris has an analytical or asymptotic estimate of power. The empirical is you literally say, of my 10,000 regression coefficients estimated between f1 and f2. Under all of those characteristics I described 68% had a p value less than point oh five, your empirical power estimate is point six, eight, if you do the study, as you described in your study corresponds to the characteristics as you generated the data. You have a point six, eight chance of finding an effect if an effect really exists. Boom, in and out, nobody gets hurt. There is your power of what

Greg 44:50
sort of sorta sorta kinda sorta I'll tell you my kind of sorta here and that is under those specific To conditions, that degree of missingness, that type of missingness, with those particular loadings with Saturn in Jupiter's fourth house. Under those specific conditions, you have a beautiful, beautiful estimate of power,

Patrick 45:27
in fairness to me,

Greg 45:29
which is what's important.

Patrick 45:30
I said, if your sample corresponds to the characteristic that you defined in your model,

Greg 45:37
yeah. Oh, that's such a pregnant if isn't it at this point, right? I'm of two minds. One, mine is like, boss screw this whole thing. I'm outta here. Right? All right, that does it.

Insert 45:49
I'm going home.

Greg 45:53
Seriously, because even if we just take the model that we started with, how do I know the value of those paths? Well, you don't but you make an educated guess, how do I know the value of those loadings? Or don't I just make an educated guess? How do I know what the distributions are? Like? I don't know it's making. It's an this is such a Jenga tower of assumptions that you might feel like what the heck is the point of all of this, so there's a part of me that would lead me to throw my hands up. But the other side of this is that it sets the stage so beautifully, for being able to do the equivalent of a sensitivity analysis, think about all the different knobs that are being turned here. One is just the values that you put in structurally, the point five, the point three, the point four, or whatever the numbers are that we put in there, you can try? What would it be like if those were different, you can try? Well, what about the loadings? Well, you can try it if the loading values are different. What if there are some Irish cross loadings? Like Patrick said, yeah, you can put in some, you can put in a few more, you can twiddle the knobs on the magnitude of those. What about missingness? Yeah, that's right. You can try different levels of missingness. What about non normality? Yeah, you can try different levels of non normality. You can sort of ask yourself, What is the worst perfect storm that I can imagine of conditions? And maybe the answer is, you've got some wicked non normality. And you've got some missingness in a particular pattern, and you've got some low loadings, and you've got some cross loadings, and you've got some weak structural pads, you're like, Bring it on, right, you define the worst case scenario, you run the simulation, you turn the crank, and you power your study for that, then you are ready for just about anything. So I would say this framework gives us the ability to understand how robust our estimates are, and hope for the best, but plan for the worst. When I'm writing up that grant proposal, I don't just say, I looked at the power for the root mean square error of approximation, because first of all, it doesn't correspond to any of my hypotheses, I can say I did a sample size planning endeavor where we essentially defined the worst case scenario. And this is what we got here. And if you can get your reviewers to buy into your conception of that worst case scenario, and that you have a sample size to overcome that, then you are in phenomenal shape. What I

Patrick 48:13
love to see are these ranges as you talk about sensitivity analysis is our best faith effort is we have a power of point six, eight, to detect this very specific effect. If things really shift sideways on us the worst possible situation, we're going to get as point six one. And if things break our way, point eight, three, give kind of a low medium high estimate. And that conveys a huge amount of information to the reviewer. But you know, we always talk to the reviewer, screw the reviewer, it's to you, you have an ethical responsibility to build a study that has a fighting chance to detect an effect if an effect really exists. And this is how you go about doing it. Yeah, dude, we're getting long in the tooth. So let's walk to the exit.

Greg 49:01
In that first power episode, episode two, season one, we talked about the challenges associated with power analysis just as an endeavor, that it's very difficult to characterize power in some single number, let alone with any degree of precision. And that was, as we said before, the curse the darkness of power analysis episode, and in this episode, what we tried to do is we tried to pair that with, you gotta do the power analysis, you still got to do the sample size planning. So how do you do it and you and I

both lean toward the structural equation modeling framework, not because it's power in structural equation modeling because it's power in darn near anything you want to do not literally everything, but darn near anything that you want to do. And it has a mathematical side to it that you could do if you want and it also has what I think is incredibly versatile, this simulation based side and I think that for people out there who are doing honest sample size planning, this is a wonderful go to framework for them to use To lay things out,

Patrick 50:01
hopefully, we may have got you thinking a little bit more about the power of what don't let your advisor don't let an article reviewer, don't let a grant reviewer get away by saying, Well, what is the power of your model? Yeah, turn it back on them and to say, Well, what do you mean? Do you mean the power of an omnibus exact fit? Do you mean the power of an omnibus close fit? Do you mean a one degree of freedom? misspecification? Do you mean a multiple degree of freedom? misspecification? Do you mean under asymptotic assumptions, the power of what and with the simulation methodology, whether you use canned stuff that exists in is very good, but also being able to program this for any model you want? If you have some weird three level cross classified model, and you want to know what is the power to detect a random effect, you can write a Monte Carlo simulation that you generate 10,000 random effect estimates and just count up how many are significant. You can, in principle, find the empirical power of any parameter in any model that you can draw on the board. And

Greg 51:16
I think as scientists, this is really what we ought to be doing making this full, honest attempt at trying to get at the power of the things that we care about. Right? These are my research questions. This is how I'm going to get at those research questions. Here are the sample sizes I need to get at those research questions in this particular context.

Patrick 51:34
And you know what, I'm going to double down on what I advocated in the initial episode here because I still feel this way about the precision of power estimates you really nicely described well, what loadings, what residual variances, what there's so many moving parts and subjective decisions. I still think we should have a smiley face emoji, a shrug, emoji, and OOP emoji you for power. But now we fix those emojis to the power of what now it's not just I have a shrug emoji for my entire model. I have somewhere between a shrug emoji and a smiley face emoji to detect my treatment effect and a bivariate growth model over five time periods. With 22% attrition and six ordinal level indicators. I still advocate the emojis

Greg 52:30
I feel an eggplant coming on.

Patrick 52:31
That is for your work on.

Greg 52:34
All right, thanks very much. Tootles. Everybody, take care. Thanks so much for joining us. Don't forget to tell your friends to subscribe to us on Apple podcasts, Spotify, or wherever they go. When life is so overwhelming. They'd rather listen to a podcast about statistics and root canals. You can also follow us on Twitter where we are at quantity pod and visit our website quantity pod.org where you can leave us a message find organized playlists and shownotes. Listen to past episodes and other fun stuff. And finally, you can get cool quantity merch like shirts, mugs, stickers and spiral notebooks from Red bubble.com where all proceeds from non bootleg authorized merch go to donorschoose.org to help support low income schools. You've been listening to quantitative the podcast equivalent of that weird second cousin who stays just a bit too long at holiday gatherings. Today's episode has been sponsored by the bar chart. When the rest of the world hears those words. They think of the thing that tells them where to go next on their fun Saturday night pub crawl with friends. When you hear a bar chart, you think of a tool to visualize distributions of data. Please tell me you see the problem here, and by Fisher's LSD, way more powerful than two keys magic mushrooms. And finally by stepwise variable selection methods and regression, maybe not the best idea but still way better than whale holes. This is most definitely not NPR