Video Details

Video Title: What is the True Effect in the DGP?
Video Link: https://player.vimeo.com/video/379319240

Video Transcript

Student (off screen)
Dr. Ji, so we've talked about effect size, and that's what this is all about, right? This gives us the true effect in the population?

Dr. Ji
That's a good question!

Remember statistics is completely about going beyond the sample to know something about the DGP. So let's throw up our thumb by sex data.

[FACETED HISTOGRAM OF THUMB BY SEX IN FINGERS DATA FRAME APPEARS ON SCREEN. A BLUE VERTICAL LINE RUNS THROUGH THE MEAN OF EACH GROUP.]

Even though in this particular data set, we've fit a certain model. What we really want to know is what is the DGP that gave rise to that data,

[SPEAKER WRITES “DGP” WITH AN ARROW POINTING TO THE HISTOGRAMS]

what is the data generating process or the population that generated this data, that's what we really want to know about. So even though we fit a great model here, a complex model using sex. So here I'll write that model,

[WRITES “$Y_i = b_0 + b_1X_i + e_i$” ABOVE THE HISTOGRAMS]

y sub i equals b sub 0, plus b sub 1 x sub i, plus e sub i, right? Even though we have this great model that includes sex, and we could even find the best-fitting parameter estimates, for example, 58.23.

[WRITES “58.23” WITH AN ARROW POINTING TO “b_0”]

That's the mean of the females, right? And then we could also find that the mean difference right here,

[DRAWS A HORIZONTAL LINE ALONG THE DISTANCE FROM THE MEAN OF MALES TO THE MEAN OF FEMALES]

is 6.45. About 6.5 millimeters, right?

[WRITES “6.45” WITH AN ARROW POINTING TO “b_1”]

That's the mean difference between the mean of the females and the mean of the males. We don't actually know if those

[POINTS TO “b_0” AND “b_1”]

are actually the beta sub 1's in the DGP. Because remember in the DGP, we write the GLM as y sub i equals beta sub 0, plus beta sub 1 x sub i, plus epsilon sub i, right?

[DRAWS AN ARROW FROM “DGP” POINTING TO THE RIGHT AND WRITES “$Y_i = \beta_0 + \beta_0 X_i + \varepsilon_i$”]

What we need to keep in mind is, is this 6.45, beta sub 1, is that the mean difference in the population of all males and females and their thumb lengths, right? That's a different question. It could be 6.45, but could it be 6.41 or 7 or 2, right? It could be any of those numbers. Now there is a true mean difference out there in the world. We just don't know what it is. It's unknown to us, right? And this notation

[POINTS TO “$Y_i = \beta_0 + \beta_0 X_i + \varepsilon_i$”]

helps us keep in mind that this

[POINTS TO “$Y_i = b_0 + b_1 X_i + e_i$”]

is just the best-fitting parameter estimates from our data, but not necessarily the true mean difference in the population. So to answer your question, even though 6.45 is an effect size that we calculated from our sample, and perhaps that's the best we could do in guessing what the effect size in our DGP is. The true mean difference is something else, something that's unknown to us. Now another measure of effect size that we talked about is PRE. PRE is really interesting because it's not just about this complex model. It's actually a way of comparing two different models. We're comparing this more complex model with sex in it to the empty model, the grand mean of this data which was 60.1.

[A GREEN VERTICAL LINE THAT RUNS THROUGH BOTH HISTOGRAMS AT THE GRAND MEAN APPEARS ON SCREEN]

So let me write that down. Let me write down the GLM for just the empty model which is b sub 0 plus e sub i.

[WRITES “$Y_i = b_0 + e_i$” BELOW “$Y_i = b_0 + b_1 X_i + e_i$”]

So how much error has been reduced going from this simple model to this more complex model? And that PRE is 11 percent in the case of sex, 0.11, right?

[WRITES “PRE = .11” ON SCREEN]

And so in this particular set of data, going to this more complex model has reduced our error by 11 percent. Now, is that the case for the DGP, when we have this more complex model. If we compare it to the more simple model of beta sub 0 plus epsilon sub i,

[WRITES “$Y_i = \beta_0 + \varepsilon_i$” UNDERNEATH “$Y_i = \beta_0 + \beta_0 X_i + \varepsilon_i$”]

are we gonna have the same PRE in the DGP? It's not gonna be the case because this PRE is just for this set of data. These parameter estimates fit this data. Our beta sub 0 for our empty model and the beta sub 0 and beta sub 1 for a more complex model, they're gonna be different numbers. And the PRE is going to be different too. Given that we don't actually know what any of these things are,

[POINTS TO "$Y_i = \beta_0 + \beta_0 X_i + \varepsilon_i$"]

this is our best guess

[POINTS TO "$Y_i = b_0 + b_1 X_i + e_i$"]

for what the proportion reduction of error would be in the DGP. But again we don't know. So this is just the PRE we calculated from our sample. But could we have gotten a PRE of 0.1? Or maybe a PRE of 0.08 or 0.02 or even 0, right?

[WRITES "PRE = .10" AND "PRE = .08" ON SCREEN]

The DGP could have a true PRE that's different from our sample. And later what we're gonna do is figure out how to use this idea

[GESTURES TO"PRE = .10" AND "PRE = .08"]

to eventually evaluate and compare these two models more formally, so that we could actually figure out which of these two models [THE EMPTY VS THE COMPLEX MODEL] is best supported by our data?