Hi.

Hi everybody.

Students Hi.

It’s my pleasure today

to introduce Professor Rebecca Andridge.

Professor Andridge has a Bachelors’ in Economics in Stanford

and her Master’s and PhD in Biostatistics

from the University of Michigan.

She an expert in group randomized trials

and methods of missing data

especially for that ever so tricky case that is not,

or so where data is missing not at random.

She’s been faculty in Biostatistics in Ohio State University

since 2009.

She’s an award-winning educator

and a 2020 Fellow of the Americans Associates,

and we’re very honored to have a huge day.

Let’s welcome professor Andridge.

Thank you for the very generous introduction.

I have to tell you,

it’s so exciting to see a room full of students.

I am currently teaching online class

and the students don’t all congregate in a room.

So it’s like been years since I’ve seen this.

So I’m of course gonna share my slides.

I want to warn everybody that I am working from home today.
And while we will not be interrupted by my children, we might be interrupted or I might be interrupted by the construction going on in my house, my cats or my fellow work at home husband. So I’m gonna try to keep the distractions to a minimum but that is the way of the world in 2020, in the pandemic life. So today I’m gonna be talking about some work I’ve done with some colleagues actually at the University of Michigan. Talking about selection bias in proportions estimated from non-probability samples. So I’m gonna start with some background and definitions and we’ll start with kind of overview of what’s the problem we’re trying to address. So big data are everywhere, right? We all have heard that phrase being bandied about, big data. They’re everywhere and they’re cheap. You got Twitter data, internet search data, online surveys, all these massive sources of data. And these data often, I would say pretty much all the ways arise from what are called non-probability samples.
So when we have a non-probability sample we can’t use what are called design based methods for inference, you actually have to use model based approaches.

I’m not gonna assume that everybody knows all these words that I’ve found out here, so I’m gonna go into some definitions.

Our goal is to develop an index of selection bias that lets us get at how bad the problem might be, how much bias might we have due to non-random selection?

A probability sample is a situation where you’re collecting data where each unit in the population has a known positive probability of selection. Large government sponsored surveys like the National Health and Nutrition Examination Survey, NHANES or NHIS or any number of large surveys that you’ve probably come across, you know, in application and your biostatistics courses.
So for these large surveys we do what’s called design-based inference. So that’s where we rely on the design of the data collection mechanism in order for us to get unbiased estimates in order for us to get unbiased estimates of population quantities, and we can do this without making any model assumptions. So we don’t have to assume that let’s say body mass index has a normal distribution. We literally don’t have to specify distribution at all. It’s all about the random selection into the sample that lets us get our estimates and be assured that we have unbiased estimates. So here’s an example in case there are folks out in the audience who don’t have experience with the sort of complex survey design or design features. So this is a really silly little example of a stratified sample. So here I have a population of two different types of animals. And in this population I happen to have 12 cats and 8 dogs. And I have taken a sample. Stratified sample where I took two cats and two dogs. So in this design the selection probabilities
are known for all of the units, right?
Because I know that there's a two out of eight chance.
I pick a dog and a two out of 12 chance.
that I pick a cat, right?
So the probability a cat is selected is 1/6
then the probability of dog is selected is 1/4.
Now, how do I estimate a proportion of interest?
Let's say it's the proportion of orange animals in the population.
Like here in my sample,
I have one of four orange animals,
but if I chose that as my estimator
I'd be ignoring the fact that I know how I selected
these animals into my sample.
So what we do is we wait the sample units to produce design unbiased estimates, right?
Because this one dog kinda counts differently than one cat, right?
Because there were only eight dogs.
To begin with but there were 12 cats.
So if I want to estimate the proportion of orange animals
I would say this cat is a weight is six
because there's two of them and 12 total.
So 12 divided by two is six.
So there's six in the numerator.
And then the denominator is the sum of the weights
of all the selected units,
the cats are each six and the dogs are each four.

So I actually get my estimate a proportion of 30%.

So instead of 25%.

So that kind of weighted estimator is what we do in probability sampling.

And we don’t have to say what the distribution of dogs or cats is in the sample

we entirely rely on the selection mechanism.

What ended up happening in the real world a lot of the time is we don’t actually get to use those kinds of complex designs.

And instead we collect data through what’s called a non-probability sample.

So in a non-probability sample, it’s pretty easy to define.

You cannot calculate the probability of selection into the sample, right?

So we simply don’t know what the mechanism was that made at unit enter our sample.

I know there’s the biostatistics students in the audience,

and you’ve all probably done a lot of data analysis.

And I would venture a guess that a lot of the times your application datasets

are non-probability samples, right?
A lot of the times there are convenience samples. I work a lot with biomedical researchers studying cancer patients. Well guess what, it’s almost always a convenient sample of cancer patients, right? It’s who will agree to be in the study? Who can I find to be in my study? Other types of non-probability samples include things like voluntary or self-selection sampling, quota sampling, that’s a really old, old school method from polling back many years ago. Judgment sampling or snowball sampling. So there’s a lot of different ways you can get non-probability samples. So if we go back to the dog and cat example, if I didn’t know anything about how these animals got into my sample and I just saw the four of them, and one of them was orange, I guess, I’m gonna guess 25% of my population is orange. Right? I don’t have any other information. I can’t recreate the population. I like I could with the weighting. Where I knew how many cats in the population did each of my sampled cats represent and similarly for the dogs.
So of course our best guess looking at these data would just be 25%, right? One out of the four animals is orange. So when you think about a non-probability sample, how much faith do you put in that estimate, that proportion? Hard to say, right? It depends on what you believe about the population and how you selected this non-probability sample but you do not have the safety net of the probability sample that guaranteed you’re gonna get an unbiased estimate of repeated applications of the sampling. So I’ve already used the word selection bias a lot and sort of being assuming that, you know what I mean. So now I’m gonna come back to it and define it. So selection bias is bias arising when part of the target population is not in the sample population, right? So when there’s a mismatch between who got into your sample and who was supposed to get into your sample, right? Who’s the population? Or in a more general statistical kind of way, when some population units are sampled at a different rate.
It’s like you meant for there to be a certain selection probability for orange animals or for dogs but it didn’t actually end up that way. This will end up down the path of selection bias.

And I will note that again, as you are biostats students you’ve probably had some epidemiology. And epidemiologists talk about selection bias as well. It’s the same concept, right? That concept of who is ending up in your sample.

And is there some sort of a bias in the mechanism? So selection bias is in fact the predominant concern with non-probability samples. In these non-probability samples, the units in the sample might be really different from the units not in the sample, but we can’t tell how different they are. Whether we’re talking about people, dogs, cats, hospitals, whatever we’re talking about.

However, these units got into my sample, I don’t know. So I don’t know if the people in my sample look like my population or not. And an important key thing to know is that probability samples when we have a low response rates, right?
So when there are a lot of people not responding, you’re basically ending up back at a non-probability sample, right? Where we have this beautiful design, we know everybody’s sampling weight, we draw a sample, oops, ut then only 30% of people respond to my sample. You’re basically injecting that bias back in again. Sort of undoing the beauty of the probability sample. So when we think about a selection bias and selection into a sample, we can categorize them in two ways. And Dr. McDougal, actually, when he was giving you my brief little bio used the words that I’m sure you’ve used in your classes before like ignorable and non-ignorable. These are usually are more commonly applied to missingness, right? So ignorable missingness mechanisms and non-ignorable missingness mechanisms. Missing at random, missing completely at random. or missing not at random, right? Same exact framework here. But instead of talking about missingness, we’re talking about selection into the sample. So when we have an ignorable selection mechanism,
That means the probability of selection depends on things I observed. Right, it depends on the observed characteristics. When I have a non-negotiable selection mechanism now that probability of selection depends on observed characteristics. Again, this is not really a new concept if you understood about missing data, just apply to selection into the sample. So in a probability sample we might have different probabilities of selection for different types of units like for cats versus dogs. But we know exactly how they differ, right? It’s because I designed my survey based on his characteristic of dog versus cat and I know exactly the status of dog versus cat for my entire population in order to do that selection. So I absolutely can estimate the proportion of orange, animals unbiasedly in the sense of taking repeated stratified samples and estimating that proportion. I hadn’t guaranteed that I’m gonna get an unbiased estimate, right? So this selection mechanism is definitely not non-ignorable, right?
This is definitely an ignorable selection mechanism in the sense that it only depends on observed characteristics. But if my four animals had just come from, I don’t know where? Convenience. Well now why did they end up in my sample? It could depend on something that we didn’t observe. What breed of dog it was? The age of the dog, the color of the dog. It could have been pretty much anything, right? That’s the problem with the convenient sample. You don’t know why those units often self-selected to be into your sample. So now I’m gonna head into the kind of ugly statistical notation portion of this stock. We’ll start with estimated proportions. So we’ll use Y as our binary indicator for the outcome, okay? But here I’m gonna talk about Y more generally as all the survey data. So we’ll start with Y as all the survey data, then we’re gonna narrow it down to Y as the binary indicator? So we can partition our survey data into the data for the units we got in the sample.
and the data for units that are not in the sample.

I so selected into the sample versus not selected into the sample. But for everybody I have Z, I have some fully observed what are often called design variables.

So this is where we are using information that we know about an entire population to select our sample in the world of probability sampling.

And then S is the selection indicator.

So these three variables have a joint distribution.

And most of the time, what we care about is Y given Z.

Right, we’re interested in estimating some outcome characteristic conditional on some other characteristic, right?

Average weight for dogs, average weight for cats, right?

Y given Z.

But Y given Z is only part of the issue, there’s also a selection mechanism, right?

So there’s also this function of how do you predict selection S with Y and Z.

And I’m using this additional Greek letter ψ here to denote additional variables that might be involved, right?
'Cause selection could depend on more than just Y and Z.

It could depend on something outside of that set of variables.

So when we have probability sampling, we have what's called an extremely ignorable selection mechanism, which means selection can depend on Z, like when we stratified on animal type but it cannot depend on Y.

Either the selected units Y or the excluded units Y doesn't depend on either. Kind of vaguely like the MCAR of selection mechanisms. It doesn't depend on Y at all. Observed or unobserved. But it can depend on Z.

So that makes it different than MCAR. Including a unit into the sample is independent of those survey outcomes Y and also any unobserved variables, right?

That phi here, that phi goes away. So selection only depends on Z. If I'm interested in this inference target I can ignore the selection mechanism. This is kind of parallels that idea in the missingness, in the missing data literature, right? If I have an ignorable missingness mechanism I can ignore that part of it.

I don't have to worry about modeling
the probability that a unit is selected.

But the bad news in our non-probability sampling,

very, very arguably true

that you could have non ignorable selection, right?

It's easy to make an argument for why the people

who ended up into your sample, your convenient sample are different than the people

who don’t enter your sample.

Think about some of these big data examples.

Think about Twitter data.

Well, I mean, you know,

the people who use Twitter are different

than the people who don’t use Twitter, right?

In lots of different ways.

So if you’re going to think about drawing

some kind of inference about the population,

you can’t just ignore that selection mechanism.

You need to think about how do they enter

into your Twitter sample

and how might they be different than the people

who did not enter into your Twitter sample.

So when we’re thinking about the selection mechanism

basically nothing goes away, right?

We can’t ignore this selection mechanism.

But we have to think

about it when we want to make inference,
380 00:14:55.930 --> 00:14:58.590 even when our inference is about Y given Z, right?
381 00:14:58.590 --> 00:14:59.900 Even when we don’t actually care
382 00:14:59.900 --> 00:15:01.970 about the selection mechanism.
383 00:15:01.970 --> 00:15:03.970 So the problem with probability samples
384 00:15:03.970 --> 00:15:07.350 is that it’s often very, very hard to model S
385 00:15:07.350 --> 00:15:09.790 or we don’t really have a good set of data
386 00:15:09.790 --> 00:15:11.290 with which to model the probability
387 00:15:11.290 --> 00:15:13.500 someone ended up in your sample.
388 00:15:13.500 --> 00:15:17.050 And that’s basically what you have to do to
generalize
389 00:15:17.050 --> 00:15:18.690 to the population, right?
390 00:15:18.690 --> 00:15:21.370 There’s methods that exist for non-probability
samples
391 00:15:21.370 --> 00:15:23.790 require you to do something along the lines
392 00:15:23.790 --> 00:15:25.750 of finding another dataset
393 00:15:25.750 --> 00:15:27.190 that has similar characteristics
394 00:15:27.190 --> 00:15:29.860 and model the probability of being in the
probability
395 00:15:29.860 --> 00:15:31.090 sample, right?
396 00:15:31.090 --> 00:15:33.540 So that’s doable in many situations
397 00:15:33.540 --> 00:15:35.490 but what we’re looking for is a method
398 00:15:35.490 --> 00:15:37.040 that doesn’t require you to do that
399 00:15:37.040 --> 00:15:40.030 but instead says, let’s do a sensitivity analysis.
400 00:15:40.030 --> 00:15:43.140 Let’s say, how big of a problem
401 00:15:43.140 --> 00:15:46.100 might selection bias be if we ignored
402 00:15:46.100 --> 00:15:47.250 the selection mechanism, right?
403 00:15:47.250 --> 00:15:49.240 If we just sort of took our sample on faith
404 00:15:49.240 --> 00:15:51.970 as if it were an SRS from the population.
405 00:15:51.970 --> 00:15:53.530 How wrong would we be
406 00:15:53.530 --> 00:15:57.173 depending on how bad our selection bias prob-
lem is?
407 00:15:58.570 --> 00:16:00.220 So there has been previous work done
408 00:16:00.220 --> 00:16:03.140 in this area, in surveys often.
409 00:16:03.140 --> 00:16:05.560 Try to think about how confident
410 00:16:05.560 --> 00:16:07.890 are we that we can generalize to the population
411 00:16:07.890 --> 00:16:10.320 even when we’re doing a probability sample.
412 00:16:10.320 --> 00:16:13.620 So there’s work on thinking about the representativeness
413 00:16:13.620 --> 00:16:14.510 of a sample.
414 00:16:14.510 --> 00:16:18.290 So that’s again, the generalizability to the population.
415 00:16:18.290 --> 00:16:20.710 So there’s something called an R-indicator,
416 00:16:20.710 --> 00:16:24.870 which is a function of response probabilities
417 00:16:24.870 --> 00:16:25.980 or propensities,
418 00:16:25.980 --> 00:16:27.870 but it doesn’t involve the survey variables.
419 00:16:27.870 --> 00:16:31.810 So it’s literally comparing the probability of response
420 00:16:31.810 --> 00:16:34.330 to a survey for different demographic,
421 00:16:34.330 --> 00:16:36.850 across different demographic characteristics, for example.
422 00:16:36.850 --> 00:16:37.683 Right.
423 00:16:37.683 --> 00:16:40.030 And seeing who is more likely to respond then who else?
424 00:16:40.030 --> 00:16:41.470 And if there are those differences
425 00:16:41.470 --> 00:16:43.143 then adjustments need to be made.
426 00:16:44.180 --> 00:16:46.500 There’s also something called the H1 indicator,
427 00:16:46.500 --> 00:16:49.430 which does bring Y into the equation
428 00:16:49.430 --> 00:16:51.910 but it assumes ignorable selection.
429 00:16:51.910 --> 00:16:53.600 So it’s going to assume that the Y
430 00:16:53.600 --> 00:16:55.583 excluded gets dropped out.
431 00:16:57.690 --> 00:16:59.470 The selection mechanism is only depends
432 00:16:59.470 --> 00:17:02.830 on things that you observe, so you can ignore it, right?
433 00:17:02.830 --> 00:17:04.490 So it’s ignorable.
So that’s not what we’re interested in.
'Cause we’re really worried in the non probability space
that we can’t ignore the selection mechanism.
And there isn’t relatively new indicator
called that they called the SMUB, S-M-U-B.
That is an index that actually extends
this idea of selection bias
to allow for non ignorable selection.
So it lets you say, well, what would my point estimate
be for a mean if selection were in fact ignorable,
and now let’s go to the other extreme,
suppose selection only depends on Y.
And I’m trying to estimate average weight
and whether or not you entered my sample
is entirely dependent on your weight.
That’s really not ignorable.
So that SMUB, this estimator is really close
to what we want but we want it for proportions.
especially because in survey work and in large datasets,
we very often have categorical data
or very, very often binary data.
If you think about if you’ve ever participated
in an online survey or filled out those kinds of things
very often, right, You’re checking a box.
It’s multiple choice, select all that apply.

It’s lots and lots of binary data floating around out there.

And I’ll show you a couple of examples.

So that was a lot of kind of me talking at you about the framework.

Now, let me bring this down to a solid example application.

So I’m going to use the national survey of family growth as a fake population.

So I want you to pretend that I have a population.

of 19,800 people, right?

It happens to be that I pulled it from the national survey of family growth, that’s not really important that that was the source.

I’ve got this population of about 20,000 people.

But let’s pretend we’re doing a study and I was only able to select into my sample smartphone users.

Because I did some kind of a survey that was on their,

you had to take it on your phone.

So if you did not have a smartphone you could not be selected into my sample.

In this particular case, in this fake population, it’s a very high selection fraction.

So about 80% of my population is in my sample.

That in and of itself is very unusual, right?
A non-probability sample is usually very, very small compared to the full population let’s say of the United States if that’s who we’re trying to generalize to. But for the purposes of illustration it helps to have a pretty high selection fraction. And we’ll assume that the outcome we’re interested in is whether or not the individual has ever been married.

So this is person level data, right? Ever been married. we wanna estimate it by gender, and I will note that the NSFG only calculate or only captures gender as a binary variable. This is a very long standing survey, been going on since the seventies. We know our understanding of gender as a construct has grown a lot since the seventies but this survey, and in fact many governmental surveys still treat gender as a binary variable. So that’s our limitation here but I just want to acknowledge that. So in this particular case, we know the true selection bias, right? Because I actually have all roughly 20,000 people so that therefore I can calculate what’s the truth,
and then I can use my smartphone sample and say, "Well, how much bias is there?"

So it turns out that in the full sample 46.8% of the females have never been married. And 56.6% of the males had never been married.

But if I use my selected sample of smartphone users I'm getting a, well, very close, but slightly smaller estimate for females. 46.6% never married. And for males it's like about a percentage point lower than the truth, 55.5%. So not a huge amount of bias here. My smartphone users are not all that non-representative with respect to whether or not they've ever been married.

When we have binary data, an important point of reference is what happens if we assume everybody not in my sample is a one, right? What if everybody not in my sample was never married?

So when we have binary data, everybody not in my sample is a one, right? What if everybody not in my sample was never married?

When you fill in all zeros or fill in old bonds for the missing values or the values
for those non-selected folks. So we can bound the bias. So the bias of this estimate of 46.6 or 46.6% has to be by definition between negative 0.098 and positive 0.085. Because those are the two ends of putting all zeros or all ones for the people who are not in my sample. So this is unlike a continuous variable, right? Where we can’t actually put a finite bound on the bias. We can with a proportion, right? So this is why, for example, if any of you ever work on smoking cessation studies they’re getting the Manski bounds. Okay. So the question is, can we do better than the Manski bounds? Because these are actually pretty wide bounds, relative to the size of the true bias, and these are very wide. And imagine a survey where we didn’t have 80% selected. Well, then the Manski bounds are gonna be useless, right?
plug in, all zeros plug in all ones,
you’re gonna get these insane estimates
that are nowhere close to what you observed.
So going back to the statistical notation,
this is where I said we’re going to use Y
in a slightly different way.
Now, Y, and now forward is the binary variable of interest.
So in this case, in this NSFG example
it was never married.
We have a bunch of auxiliary variables that we observed
for everybody in the selected sample;
age, race, education, et cetera,
and I’m gonna call those Z.
Assume also that we have summary statistics
on Z for the selected cases.
So I don’t observe Z for everybody, right?
All my non-smartphone users,
I don’t know for each one of them, what is their gender?
What is their age? What is their race?
But I don’t actually observe that.
But I observed some kinda summary statistic.
But a mean vector and a covariance matrix of Z.
So I have some source of what does my population
look like at an aggregate level?
And in practice, this would come from something
like census data or in a very large probability sample,
something where we would be pretty confident
This is reflective of the population.
Will note that if we have data for the population
and not the non-selected,
then we can kinda do subtraction, right?
We can take the data for the population
and aggregate and go backwards
to figure out what it would be for the non-selected
by effectively backing out the selected cases.
And similarly another problem
is that we don’t have the variance.
We could just assume it’s what we observe
in the selected cases.
So how are we gonna use this in order
to estimate of selection bias,
what we’re gonna come up
with this measure of unadjusted bias for proportions
called the MUBP.
So the MUBP is an extension of the SMUB
that was for means, for continuous variables
to binary outcomes, right?
To proportions.
High-level, it’s based on pattern-mixture models.
It requires you to make explicit assumptions
about the distribution of the selection mechanism,
and it provides you a sensitivity analysis,
I don’t know what that distribution is, and you’re gonna get a range of bias. So that’s that idea of how wrong might we be? So we’re trying to just tighten those bounds compared to the Manski bounce. Where we don’t wanna have to rely on plug in all zeros, plug in all ones, we wanna shrink that interval to give us something a little bit more meaningful. So the basic idea behind how this works before I show you the formulas is we can measure the degree of selection bias in Z, right? Because we observed Z for our selected sample, and we observed at an aggregate for the population. So I can see, for example, that if in my selected sample, I have 55% females but in the population it’s 50% females. Well, I can see that bias. Right, I can do that comparison. So absolutely I can tell you how much selection bias there is for all of my auxiliary variables. So if my outcome Y is related to my Zs then knowing something about the selection bias in Z tells me something about the selection bias in Y.
It doesn’t tell me exactly the selection bias in Y, but it gives me some information in the selection bias in Y.

So in the extreme imagine if your Zs in your selected sample in aggregate looked exactly like the population. Well, then you’d be pretty confident, right? That there’s not an enormous amount of selection bias assuming that Y was related to the Z.

So we’re gonna use pattern-mixture models to explicitly model that distribution of S, right? And we’re especially gonna focus on the case when selection depends on Y.

It depends on our binary outcome of interest. So again, Y is that binary variable interest, we only have it for the selected sample. In the NSFG example it’s whether the woman or man has ever been married.

We have Z variables available for the selected cases in micro data and an aggregate for the non-selected sample, a demographic characteristics like age, education, marital status, et cetera.

And the way that we’re gonna go about doing this is we’re gonna try to get back to the idea of normality, because then as you all know, when everything’s normal
It’s great, right?

It’s easy to work with the normal distribution.

So the way we can do that with a binary variable is we can think about latent variables.

So we’re going to think about a latent variable called U.

That is an underlying, unobserved latent variables.

So unobserved for everybody, including our selected sample.

And it’s basically thresholded.

And when U crosses zero, well, then Y goes from zero to one.

So I’m sure many, all of you have seen probit regression,

or this is what happens and this is how probit regression is justified,

via latent variables.

So we’re going to take our Zs that we have for the selected cases,

and essentially reduce the dimensionality.

We’re gonna take the Zs,

run a probate regression of Y on Z in the selected cases,

and pull out the linear predictor and I’m gonna call that X.

That is my proxy for Y or my Y hat, right?

It’s just the predicted value from the regression.
And I can get that for every single observation in my selected sample, of course, right? Just plug in each individual’s Z values and get out their Y hat. That’s my proxy value. And it’s called the proxy because it’s the prediction, right? It’s our sort of best guess at Y based on this model. So I can get it for every observation in my selected sample, but very importantly I can also get it on average for the non-selective sample. So I have all my beta hats for my probit regression, and I’m gonna plug in Z-bar. And I’m going to plug in the average value of my Zs. And that’s going to give me the average value of X for the non-selected cases. I don’t have an actual observed value for all those non-selective cases but I have the average, right? So I could think about comparing the average Z value in the aggregate, in the non-selected cases to that average Z among my selected cases. And that is of course exactly where we’re gonna get those index from. So I have my selection indicator S, so in the smartphone example,
716 00:28:33.000 --> 00:28:35.080 that’s S equals one for the smartphone users
717 00:28:35.080 --> 00:28:37.230 and S equals zero for the non-smartphone users
718 00:28:37.230 --> 00:28:38.670 who weren’t in my sample.
719 00:28:38.670 --> 00:28:40.150 And importantly, I’m going to allow
720 00:28:40.150 --> 00:28:42.750 there to be some other covariates V
721 00:28:42.750 --> 00:28:46.010 floating around in here that are independent of Y and X
722 00:28:46.010 --> 00:28:48.220 but could be related to selection.
723 00:28:48.220 --> 00:28:49.113 Okay.
724 00:28:49.113 --> 00:28:51.110 So it could be related to how you got into my sample
725 00:28:51.110 --> 00:28:53.310 but importantly, not related to the outcome.
726 00:28:54.870 --> 00:28:58.550 So diving into the math here, the equations,
727 00:28:58.550 --> 00:29:01.890 we’re gonna assume a proxy pattern-mixture model for U,
728 00:29:01.890 --> 00:29:04.510 the latent variable underlying Y
729 00:29:04.510 --> 00:29:07.883 and X given the selection indicator.
730 00:29:07.883 --> 00:29:11.110 So what a pattern-mixture model does is it says
731 00:29:11.110 --> 00:29:13.530 there’s a totally separate distribution
732 00:29:13.530 --> 00:29:16.400 or joint distribution of Y and X for the selected units
733 00:29:16.400 --> 00:29:17.770 and the non-selected units.
734 00:29:17.770 --> 00:29:21.010 Notice that all my mus, all my sigmas, my rho,
735 00:29:21.010 --> 00:29:23.420 they’ve all got a superscript of j, right?
736 00:29:23.420 --> 00:29:26.810 So that’s whether your S equals zero or S equals one.
737 00:29:26.810 --> 00:29:31.240 So two totally different bi-variate normal distributions
738 00:29:31.240 --> 00:29:32.690 before Y and X,
739 00:29:32.690 --> 00:29:35.000 depending on if you’re selected or non-selected.
740 00:29:35.000 --> 00:29:36.650 And then we have a marginal distribution
just Bernoulli, for the selection indicator. However, I'm sure you all immediately are thinking, "Well, that's great, but I don't have any information to estimate some of these parameters for the non-selected cases." Clearly, for the selected cases, right? S equals one. I can estimate all of these things. But I can't estimate them for the non-selected sample because I might observe X-bar but I don't observe anything having to do with you. 'Cause I have no Y information. So in order to identify this model and be able to come up with estimates for all of these parameters, we have to make an assumption about the selection mechanism. So we assume that the probability of selection into my sample is a function of U. So we're allowing it to be not ignorable. Remember that's underlying Y and X, that proxy which is a function of Z. So that's observed and V, those other variables. And in particular, we're assuming that it's this funny looking form of combination of X and U.
That depends on this sensitivity parameter $\phi$.

So $\phi$ it’s one minus $\phi$ times $X$ and $\phi$ times $U$.

So that’s essentially weighting the contributions of those two pieces.

How much of selection is dependent on the thing that I observe or the proxy builds off the auxiliary variables and how much of it is depending on the underlying latent $U$ related to $Y$.

Okay.

That’s sort of a technical detail.

We’re rescaling the proxy.

So it has the same variance as $U$.

Very unimportant mathematical detail.

So we have this joint distribution that is conditional on selection status.

And in addition to, we need that one assumption to identify things.

We also have the latent variable problem.

So latent variables do not have separately identifiable mean and variance, right?

So that’s just...

Outside of the scope of this talk.

That’s just a fact, right?
So without loss of generality, we're gonna set the variance of the latent variable for the selected sample equal to one. So it's just the scale of the latent variable. So what we actually care about is a function of you, right? It's the probability Y equals one marginally in my entire population. And so the probability Y equals one, is a probability U is greater than zero. That's that relationship. And so it's a weighted average of the proportion in the selected sample and the proportion in the non-selected sample, right? These are just... If U has this normal distribution this is how we get down to the probability U equals zero. Like those are those two pieces. So the key parameter that governs how this MUBP works is a correlation, right? It's the strength of the relationship between Y and your covariates. How good of a model do you have for Y, right? So remember we think back to that example of what if I had no biases Z. Or if Y wasn't related to Z, well, then who cares that there is no bias in Z.
But we want there to be a strong relationship between Z and Y so that we can kind of infer from Z to Y.

So that correlation in this latent variable framework is called the biserial correlation of the binary X and the continuous. I mean, sorry, the binary Y and the continuous X, right?

There’s lots of different flavors of correlation, biserial is the name for this one that’s a binary Y and a continuous X when we’re thinking about the latent variable framework.

Importantly, you can estimate this in the selected sample, right? So I can estimate the correlation between you and X among the selected sample.

I can’t for the non-selected sample, of course, but I can for the selected sample. So the non-identifiable parameters of that pattern-mixture model, here they are. Like the mean for the latent variable, the variance for the latent variable and that correlation for the non-selected sample are in fact identified when we make this assumption on the selection mechanism.

So let’s think about some concrete scenarios. What if phi was zero?
If \( \phi \) is zero, we look up here at this part of the formula, well, then \( \phi \) drops out it. So therefore selection only depends on \( X \) and those extra variables \( V \) that don’t really matter because \( V \) isn’t related to \( X \) or \( Y \). This is an ignorable selection mechanism, okay.

If on the other hand \( \phi \) is one, well, then it entirely depends on \( U \). \( X \) doesn’t matter at all. This is your worst, worst, worst case scenario, right?

Where whether or not you’re in my sample only depends on \( U \) and therefore only depends on the value of \( Y \). And so this is extremely not ignorable selection.

And of course the truth is likely to lie somewhere in between, right?

Some sort of non-ignorable mechanism, a \( \phi \) between zero and one, so that \( U \) matters but it’s not the only thing that matters.

Right, that \( X \) matters as well. Okay.

So this is a kind of moderate, non-ignorable selection. That’s most likely the closest to reality with these non-probability samples. So for a specified value of \( \phi \).
So we pick a value for our sensitivity parameter. There's no information in the data about it. We just pick it and we can actually estimate the mean of Y and compare that to the selected sample proportion. So we take this select a sample proportion, subtract what we get as the truth for that particular value of phi, and that's our measure of bias, right? So this second piece that's being subtracted here depends on phi. Right, it depends on what your value of your selected parameter is, or selection for your sensitivity parameter is. So in a nutshell, pick a selection mechanism by specifying specifying phi, estimate the overall proportion, and then subtract to get your measure of bias. And again, we don't know whether we're getting the right answer because it’s depending on the sensitivity parameter but it’s at least going to allow us to bound the problem. So the formula is quite messy, but it gives some insight into how this index works. So this measure of bias is the selected sample mean minus that estimator, right? This is the overall mean of Y based on those latent variables.
And what gets plugged in here importantly for the mean and the variance for the non-selected cases depends on a component that I’ve got colored blue here, and a component that I’ve got color red. So if we look at the red piece this is the comparison of the proxy mean for the unselected and the selected cases. This is that bias in Z, the selection bias in Z, and it’s just been standardized by its estimated variance, right? So that’s how much selection bias was present in Z via X, right. Via using it to predict in the appropriate regression. Similarly, down here, how different is the variance of the selected and unselected cases for X. How much bias, selection bias is there in estimating the variance? So we’re going to use that difference and scale the observed mean, right? There’s that observed the estimated mean of U in the selected sample and how much it’s gonna shift by it depends on the selection, the sensitivity parameter phi, and also that by serial correlation.
So this is why the by serial correlation is so important.

It is gonna dominate how much of the bias in X we’re going to transfer over into Y.

So if phi were zero, so if we wanna assume that it is an ignorable selection mechanism, then this thing in blue here, think about plugging zero here, zero here, zero everywhere, is just gonna reduce down to that correlation.

So we’re gonna shift the mean of U for the non-selective cases based on the correlation times that difference in X.

Whereas if we have phi equals one, this thing in blue turns into one over the correlation.

So here is where thinking about the magnitude of the correlation helps.

If the correlation is really big, right? If the correlation is 0.8, 0.9, something really large than phi and...

I mean, sorry, then rho and one over rho are very close, right?

0.8 and 1/0.8 are pretty close.

So if we’re thinking about bounding this between phi, equals zero and equals one, our interval is gonna be relatively small.

But if the correlation is small, the correlation were 0.2, oh, oh, right?
We’re gonna get a really big interval because that correlation, we’re gonna shift with the factor of multiplied by 0.2 but then one over 0.2.

That’s gonna be a really big shift in that mean of the latent variable U and therefore the mean of Y.

So how do we get these estimates? We have two possibilities. We can use what we call modified maximum likelihood estimation.

It’s not true. Maximum likelihood because we estimate the biserial correlation with something called a two step method, right?

So instead of doing a full, maximum likelihood, we kind of take this cheat in which we set that mean of X for the selected cases equal to what we observe, and then conditional not to estimate the by serial correlation.

We kind of take this cheat in which we set that mean of X for the selected cases equal to what we observe, and then conditional not to estimate the by serial correlation.

And as a sensitivity analysis we would plug in zero one and maybe 0.5 in the middle as the values sensitivity parameter.

Alternatively, and we feel is a much more attractive approach is to be Bayesian about this.

So in this MML estimation, we are implicitly assuming that we know the betas.
from that probate regression.
That we’re essentially treating $X$ like we know it.
But we don’t know $X$, right?
That probate regression, those parameters have error associated with them.
Right?
And you can imagine that the bigger your selected sample,
the more precisely estimating those betas, that’s not being reflected
at all in the modified maximum likelihood.
So instead we can be Bayesian.
Put non-informative priors on all the identified parameters.
That’s gonna let those, the error in those betas be propagated.
And so we’ll incorporate that uncertainty.
And we can actually, additionally put a prior on $\phi$, right?
So we could just say
let’s have it be uniform across zero one.
Right?
So we can see what does it look like if we in totality,
if we assume that $\phi$ is somewhere evenly distributed
across that interval.
We’ve done other things as well.
We’ve taken like discreet priors.
Oh, let’s put a point mass on 0.5 and one
or other different, right?
You can do whatever you want for that prior.
So let’s go back to the example
see what it looks like.
If we have the proportion ever married for females
on the left and males on the right,
the true bias is the black dot.
And so the black is the true bias.
The little tiny diamond is the MUBP for 0.5.
An so that’s plugging in that average value.
Some selection mechanism that depends on why some,
where we are actually coming pretty close.
That happens to be, that’s pretty close.
And the intervals in green
are the modified maximum likelihood intervals
from phi equals zero to phi equals one,
and the Bayesian intervals are longer, right?
Naturally.
We’re incorporating the uncertainty.
Essentially these MUBP,
modified maximum likely intervals are too short.
And we admit that these are too short.
If we plug in all zeros and all ones
for that small proportion of my NSFG population
that we aren’t selected into the sample,
we get huge bounds relative to our indicator.
Right?
So remember when I showed you that slide, that bounded,
we know the bias has to be between these two values.
That’s what’s going on here.
That’s what these two values are.
But using the information in Z,
we’re able to much, much narrow
or make an estimate on where our selection bias is.
So we got much tighter bounds.
An important fact here
is that we have pretty good predictors.
Our correlation, the biserial correlation
is about 0.7 or 0.8.
So these things are pretty correlated
with whether you’ve been married, age, education, right?
Those things are pretty correlated.
Another variable in the NSFG is income.
So we can think about an indicator for having low income.
Well, as it turns out those variables
we have on everybody; age, education, gender,
those things are not actually that good of predictors,
of low income, very low correlation.
So our index reflects that.
Or you get much, Y, your intervals.
Sort of closer to the Manski bounds.
And in fact, it’s exactly returning one of those bounds.
The filling in all zeros bound is returned by this index. So that’s actually an attractive feature. Right? We’re sort of bounded at the worst possible case. on one end of the bias. but we are still capturing the truth. The Manski bounds are basically useless, right in this particular case. So that’s a toy example. Just gonna quickly show you a real example, and I’m actually gonna to skip over the incentive experiment, which well, very, very interesting is there’s a lot to talk about, and I’d rather jump straight to the presidential polls. So there’s very much in the news now, and over the past several years, this idea of failure of political polling and this recent high profile failure of pre-election polls in the US. So polls are probability samples but they have very, very, very low response rates. I don’t know how much you know about how they’re done, but they’re very, very low response rate. But think about what we’re getting at in a poll, a binary variable, are you going to vote for Donald Trump?
Yes or no?

Are you gonna vote for Joe Biden?

Yes or no?

These binary variables.

We want to estimate proportions.

That’s what political polls aimed to do.

Pre-election polls.

So we have these political polls with these failures.

So we’re thinking, maybe it’s a selection bias problem.

And that there is some of this people

are entering into this poll differentially,

depending on who they’re going to vote for.

So think of it this way,

and I’m gonna use Trump as the example

’cause we’re going to estimate,

I’m gonna try to estimate

the proportion of people who will vote

for Former President Trump in the 2020 election.

So, might Trump supporters

just inherently be less likely to answer the call, right?

To answer that poll or to refuse to answer the question

even conditional demographic characteristics, right?

So two people who otherwise look the same

with respect to those Z variables, age, race, education,

the one who’s the Trump supporter, someone might argue,
you might be more suspicious of the government and the polls, and not want to answer and not come into this poll, not be selected. As it would be depending on why. So the MUBP could be used to try to adjust poll estimates. Say, well, there’s your estimate from the poll but what if selection were not ignorable? How different would our estimate of the proportion voting for Trump? So in this example, our proportion of interest is the percent of people who are gonna vote for Trump. The sample that we used are publicly available data from seven different pre-election polls conducted in seven different states by ABC in 2020. And the way these polls work is it’s a random digit dialing survey. So that’s literally randomly dialing phone numbers. Many of whom get throughout ’cause their business, et cetera, very, very low response rates, 10% or lower. Very, very low response rates to these kinds of polls. They do, however, try to do some weighting. So it’s not as if they just take that sample and say, there we go let’s estimate the proportion for Trump.
We do waiting adjustments and they use what’s called inter proportional fitting or raking to get the distribution of key variables in the sample to look like the population. So they use census margins for, again, it’s gender as binary, unfortunately, age, education, race, ethnicity, and party identification. Because we’re doing this after the election we know the truth. We have access to the true official election outcomes in each state. So I know the actual proportion of why. My population is likely voters, and that’s who we’re trying to target with these pre-election polls. You wanna know what’s the estimated proportion would vote for Trump among the likely voters. So the tricky thing is that population is hard to come by summary statistics. Likely voters, right? It’s easy to get summary statistics from all people in the US or all people of voting age in the US but not likely voters. So here Y is an indicator for voting for Trump. Z is auxiliary variable in the ABC poll.
So all those variables I mentioned before gender, age, et cetera. We actually have very strong predictors of why basically because of these political ideation, party identification variables, right? Not surprisingly the people who identify as Democrats, very unlikely to be voting for Trump. The data set that we found that can give us population level estimates of the mean of Z for the non-selected sample is a dataset from AP/NORC. It’s called their VoteCast Data. And they conduct these large surveys and provide an indicator of likely voter. So we can basically use this dataset to describe the demographic characteristics of likely voters, instead of just all people who are 18 and older in the US. The subtle issue is of course, these AP VoteCast data are not without error, but we’re going to pretend that they are without error. And that’s like a whole other papers. How do we handle the fact that my population data have error? So we’re gonna use the unweighted ABC poll data as the selected sample and estimate the MUBP
1183 00:46:25.530 --> 00:46:27.270 with the Bayesian approach with phi
1184 00:46:27.270 --> 00:46:29.270 from the uniform distribution.
1185 00:46:29.270 --> 00:46:32.280 The poll selection fraction is very, very, very
1186 00:46:32.280 --> 00:46:34.030 Right, these polls in each state
1187 00:46:34.030 --> 00:46:36.050 have about a thousand people in them
1188 00:46:36.050 --> 00:46:38.060 but we’ve got millions of voters in each state.
1189 00:46:38.060 --> 00:46:40.040 So the selection fraction is very, very, very
1190 00:46:40.040 --> 00:46:42.090 total opposite of the smartphone example.
1191 00:46:42.980 --> 00:46:45.760 So we’ll just jump straight into the answer,
1192 00:46:45.760 --> 00:46:46.593 did it work?
1193 00:46:46.593 --> 00:46:48.090 Right, this is really exciting.
1194 00:46:48.090 --> 00:46:51.820 So the red circle is the true proportion,
1195 00:46:51.820 --> 00:46:53.410 oh, sorry, the true bias,
1196 00:46:53.410 --> 00:46:54.720 this should say bias down here.
1197 00:46:54.720 --> 00:46:55.600 In each of the states.
1198 00:46:55.600 --> 00:46:56.540 So these are the seven states
1199 00:46:56.540 --> 00:46:59.270 we looked at Arizona, Florida, Michigan,
1200 00:46:59.270 --> 00:47:01.550 Minnesota, North Carolina, Pennsylvania, and Wiscon-
1201 00:47:01.550 --> 00:47:05.960 sin.
1202 00:47:05.960 --> 00:47:08.140 So this horizontal line here at zero that’s no
1203 00:47:08.140 --> 00:47:09.490 bias, right?
1204 00:47:09.490 --> 00:47:12.920 And we can see then in Arizona where sort
1205 00:47:12.920 --> 00:47:14.060 of underestimated
1206 00:47:14.060 --> 00:47:16.277 we’ve got underestimated the support for
1208 00:47:19.140 --> 00:47:22.250 And so that was really the failure was the
1209 00:47:22.250 --> 00:47:24.250 underestimation
Notice that our Bayesian bounds cover the true bias everywhere except in Pennsylvania and Wisconsin. And so Wisconsin had an enormous bias, or they way under called the support for Trump in Wisconsin by 10 percentage points. Huge problem. So we're not getting there but notice that zero is not in our interval. So our bounds are suggesting that there was a negative bias from the poll. So even though we didn’t capture the truth, we’ve at least crossed the threshold saying very likely that you are under calling the support for Trump. The true proportion who voted for Trump are now these red triangles and then the black circles are the point estimates from three different methods of estimation, of obtaining an estimate. Unweighted from the poll weighted estimate from the poll.
1235 00:48:24.720 --> 00:48:27.820 and the adjusted by our measure of selection bias,
1236 00:48:27.820 --> 00:48:30.340 the non-ignorable selection bias is the last one.
1238 00:48:32.330 --> 00:48:34.850 So we can see that in some cases
1239 00:48:34.850 --> 00:48:39.140 our adjustment and the polls are pretty similar, right?
1240 00:48:39.140 --> 00:48:40.700 But look at, for example, Wisconsin,
1241 00:48:40.700 --> 00:48:42.080 all the way over here on the right.
1242 00:48:42.080 --> 00:48:43.887 So again, remember I said, we didn’t cover the truth
1243 00:48:43.887 --> 00:48:45.700 and we didn’t cover the true bias
1244 00:48:45.700 --> 00:48:48.650 but our indicator is the only one, right?
1245 00:48:48.650 --> 00:48:52.020 That’s got that much higher shift up towards Trump.
1246 00:48:52.020 --> 00:48:53.430 So this is us saying, well,
1247 00:48:53.430 --> 00:48:57.190 if there were an underlying selection mechanism
1248 00:48:57.190 --> 00:48:58.980 saying that Trump supporters
1249 00:48:58.980 --> 00:49:02.860 were inherently less likely to enter this poll,
1250 00:49:02.860 --> 00:49:03.900 this is what would happen.
1251 00:49:03.900 --> 00:49:07.330 Or this is what your estimated support for Trump would be.
1252 00:49:07.330 --> 00:49:08.830 It’s shifted up.
1253 00:49:08.830 --> 00:49:10.780 We’ve got a similar sort of success story
1254 00:49:10.780 --> 00:49:12.270 I’ll say in Minnesota,
1255 00:49:12.270 --> 00:49:15.650 we’re both of the ABC estimators did not cover the truth
1256 00:49:15.650 --> 00:49:18.000 in these pre-election polls but ours did, right.
1257 00:49:18.000 --> 00:49:20.660 We were able to sort of shift up and say,
1258 00:49:20.660 --> 00:49:22.440 look, if there were selection bias
1259 00:49:22.440 --> 00:49:24.660 that depended on whether or not you supported Trump
we would captured that.

So the important idea here is, you know, before the election, we wouldn’t have these red triangles.

But it’s important to be able to see that this is saying you’re under calling the support for Trump if there were a non-negligible selection, right?

So it’s that idea of a sensitivity analysis? How bad would we be doing?

And what we would say is in Minnesota and Wisconsin we’d be very worried about under calling the support for Trump.

I’ll summarize.

The MUBP is a sensitivity analysis tool to assess the potential for non-ignorable selection bias.

If we have a phi equals zero, an ignorable selection, we can adjust that away via weighting or some other method, right?

So if it’s not ignorable, I mean, if it is ignorable we can ignore the selection mechanism.

On the other extreme if phi is one, totally not ignorable, selection is only depending on that outcome we’re trying to measure.

Somewhere in between we’ve got the 0.5. That if you really needed a point estimate
of the bias, that would be 0.5.

And in fact, that’s what this black dot is.

That’s the adjustment at 0.5 for our adjusted estimator.

This MUBP is tailored to binary outcomes, and it is an improvement over the normal base SMUB.

I didn’t show you the results from simulations that basically show

if you use the normal method on a binary outcome

you get these huge bounds.

You go outside of the Manski bounds, right?

'Cause it’s not properly bounded between zero and one,

or your proportion isn’t properly bounded.

And importantly, our measure only requires summary statistics for Z,

for the population or for the non-selected sample.

So I don’t have to have a whole separate data set

where I have everybody who didn’t get selected

into my sample,

I just need to know the average of these co-variants, right.

I just needs to know Z-bar in order to get my average

proxy for the non-selected.

With weak information,

so if my model is poor then my Manski bounds
are gonna be what’s returned. So that’s a good feature of this index.

Is that it is naturally bound unlike the normal model version. And we have done additional work to move beyond just estimating means and proportions.

We have indices of selection bias for regression coefficients. So instead of wanting to know the mean of $Y$ or the proportion with $Y$ equals one, what if you wanted to do a regression of $Y$ on some covariates? So we have a paper out in the animals of applied statistics that extends those two regression coefficients. So I believe I’m pretty much right on the time I was supposed to end, so I’ll say Thank you everyone.

And I’m happy to take questions. I’ll put on my references of my meeny, miny fonts, yes.

Robert Does anybody have any questions? From the room?

So. Dr. Rebecca Let me stop my share. Student Hey.

I have a very basic one,

mostly more of curiosity (indistinct)
Sure, sure.

What is it that caused the...

We know after the fact that in your example that there was the direction of the bias, but why is it that it only shifted in the Trump direction?

Why?

You don’t know in advance if something is more likely or less likely?

Okay.

So excellent question.

So that is effectively, the direction of the shift is going to match...

The direction of the shift in the mean of Y, when the proportion is going to match

the shift in X, right?

If what you get as your mean for your proxy,

for the non-selected sample is bigger than for your selected sample

then your proportion is gonna get shifted in that direction?

Right.

It’s only ever going to shift it to match the bias in X.

Right?

And so then, which way that shifts Y

depends on what the relationship

is between the covariates Z and X in the probate regression.

But it will always shift it in a particular direction.
I will notice that I fully admit, our index actually shifted the wrong direction in one particular case. Right? So actually in Florida, we actually shifted down when we shouldn’t. Right. So here’s the way to estimate and we’re shifting down, but actually the truth is higher. So we’re not always getting it right we’re getting it right when that X is shifting in the correct direction. Right? So it isn’t true that we always... It’s true that it always shifts the direction of X, but it’s not a hundred percent true that X always shifts in the exact same way as Y. Just most of the time. There was evidence of underestimating the Trump support, and that was in fact reflected in that probate regression, right in that relationship. The people who replied to the poll were older, they were higher educated, right? And so those older, higher educated people in aggregate were less likely to vote for Trump. So that’s why we ended up under calling the support.
for Trump when we don’t account for that potential non-ignorable selection bias.

Good question though.

Robert Go it, Thank you.

Any other questions (indistinct)

Anybody?

I know I talk fast and that was a lot of stuff so you know, like get it.

Alright.

Well, Andridge, Thank you again.

And.

(students clapping)

Thank you.

Thank you for having me.

Robert Yeah.