WEBVTT

1 00:00:00.040 --> 00:00:00.873 Hi.

2 00:00:00.873 --> 00:00:01.706 Hi everybody.

3 00:00:01.706 --> 00:00:02.539 Students Hi.

4 00:00:02.539 --> 00:00:03.372 It's my pleasure today

5 00:00:03.372 --> 00:00:05.253 to introduce Professor Rebecca Andridge.

6 00:00:06.120 --> 00:00:09.920 Professor Andridge has a Bachelors' in Economics in Stanford

7 00:00:09.920 --> 00:00:13.290 and her Master's and PhD in Biostatistics

8 00:00:13.290 --> 00:00:14.890 from the University of Michigan.

9 00:00:15.744 --> 00:00:17.670 She an expert in group randomized trials

10 00:00:17.670 --> 00:00:19.440 and methods of missing data

11 00:00:19.440 --> 00:00:22.930 especially for that ever so tricky case that is not,

12 00:00:22.930 --> 00:00:25.700 or so where data is missing not at random.

13 00:00:25.700 --> 00:00:28.780 She's been faculty in Biostatistics in Ohio State University

14 00:00:28.780 --> 00:00:30.620 since 2009.

15 00:00:30.620 --> 00:00:32.210 She's an award-winning educator

16 00:00:32.210 --> 00:00:35.930 and a 2020 Fellow of the Americans Associates,

17 00:00:35.930 --> 00:00:38.290 and we're very honored to have a huge day.

18 00:00:38.290 --> 00:00:40.186 Let's welcome professor Andridge.

19 00:00:40.186 --> 00:00:43.470 (students clapping)

20 00:00:43.470 --> 00:00:45.860 Thank you for the very generous introduction.

21 00:00:45.860 --> 00:00:46.693 I have to tell you,

22 00:00:46.693 --> 00:00:50.800 it's so exciting to see a room full of students.

23 00:00:50.800 --> 00:00:52.440 I am currently teaching online class

24 00:00:52.440 --> 00:00:54.320 and the students don't all congregate in a room.

25 00:00:54.320 --> 00:00:56.883 So it's like been years since I've seen this.

26 00:00:57.830 --> 00:01:01.400 So I'm of course gonna share my slides.

27 00:01:01.400 --> 00:01:06.320 I want to warn everybody that I am working from home today.

28 00:01:06.320 --> 00:01:08.600 And while we will not be interrupted by my children

29 00:01:08.600 --> 00:01:10.580 we might be interrupted or I might be interrupted

30 00:01:10.580 --> 00:01:13.000 by the construction going on in my house,

31 00:01:13.000 --> 00:01:15.790 my cats or my fellow work at home husband.

32 00:01:15.790 --> 00:01:18.260 So I'm gonna try to keep the distractions to a minimum

33 00:01:18.260 --> 00:01:21.530 but that is the way of the world in 2020,

34 00:01:21.530 --> 00:01:23.700 in the pandemic life.

35 00:01:23.700 --> 00:01:25.880 So today I'm gonna be talking about some work

36 00:01:25.880 --> 00:01:26.960 I've done with some colleagues

37 00:01:26.960 --> 00:01:28.720 actually at the University of Michigan.

38 00:01:28.720 --> 00:01:31.090 Talking about selection bias

39 00:01:31.090 --> 00:01:34.373 in proportions estimated from non-probability samples.

40 00:01:35.690 --> 00:01:38.020 So I'm gonna start with some background and definitions

41 00:01:38.020 --> 00:01:40.460 and we'll start with kind of overview

42 00:01:40.460 --> 00:01:43.070 of what's the problem we're trying to address.

43 00:01:43.070 --> 00:01:45.120 So big data are everywhere, right?

44 00:01:45.120 --> 00:01:48.574 We all have heard that phrase being bandied about, big data.

45 00:01:48.574 --> 00:01:49.890 They're everywhere and they're cheap.

46 00:01:49.890 --> 00:01:53.360 You got Twitter data, internet search data, online surveys,

47 00:01:53.360 --> 00:01:56.280 things like predicting the flu using Instagram, right?

48 00:01:56.280 --> 00:01:59.170 All these massive sources of data.

49 00:01:59.170 --> 00:02:03.140 And these data often, I would say pretty much all the ways

50 00:02:03.140 --> 00:02:06.500 arise from what are called non-probability samples.

51 00:02:06.500 --> 00:02:08.320 So when we have a non-probability sample

52 00:02:08.320 --> 00:02:10.580 we can't use what are called design based methods

53 00:02:10.580 --> 00:02:11.413 for inference,

54 00:02:11.413 --> 00:02:13.880 you actually have to use model based approaches.

55 00:02:13.880 --> 00:02:16.310 So I'm not gonna assume that everybody knows

56 00:02:16.310 --> 00:02:17.750 all these words that I've found out here,

57 00:02:17.750 --> 00:02:20.393 so I'm gonna go into some definitions.

58 00:02:21.640 --> 00:02:25.120 So our goal is to develop an index of selection bias

59 00:02:25.120 --> 00:02:28.450 that lets us get at how bad the problem might be,

60 00:02:28.450 --> 00:02:32.200 how much bias might we have due to non-random selection

61 00:02:32.200 --> 00:02:33.173 into our sample?

62 00:02:34.380 --> 00:02:38.220 So a probability sample is a situation

63 00:02:38.220 --> 00:02:39.230 where you're collecting data

64 00:02:39.230 --> 00:02:41.020 where each unit in the population

65 00:02:41.020 --> 00:02:44.460 has a known positive probability of selection.

66 00:02:44.460 --> 00:02:47.330 And randomness is involved in the selection of which units

67 00:02:47.330 --> 00:02:48.970 come into the sample, right?

68 00:02:48.970 --> 00:02:52.940 So this is your stereotypical complex survey design

69 00:02:52.940 --> 00:02:54.670 or your sample survey.

70 00:02:54.670 --> 00:02:57.130 Large government sponsored surveys

71 00:02:57.130 --> 00:03:00.020 like the National Health and Nutrition Examination Survey,

72 00:03:00.020 --> 00:03:04.320 NHANES or NHIS or any number of large surveys

73 00:03:04.320 --> 00:03:05.760 that you've probably come across,

74 00:03:05.760 --> 00:03:09.000 you know, in application and your biostatistics courses.

75 00:03:09.000 --> 00:03:11.130 So for these large surveys

76 00:03:11.130 --> 00:03:13.560 we do what's called design-based inference.

77 00:03:13.560 --> 00:03:15.820 So that's where we rely on the design

78 00:03:15.820 --> 00:03:17.670 of the data collection mechanism

79 00:03:17.670 --> 00:03:19.770 in order for us to get unbiased estimates

80 00:03:19.770 --> 00:03:21.240 of population quantities,

81 00:03:21.240 --> 00:03:24.340 and we can do this without making any model assumptions.

82 00:03:24.340 --> 00:03:25.870 So we don't have to assume

83 00:03:25.870 --> 00:03:29.130 that let's say body mass index has a normal distribution.

84 00:03:29.130 --> 00:03:31.980 We literally don't have to specify distribution at all.

85 00:03:31.980 --> 00:03:34.540 It's all about the random selection into the sample

86 00:03:34.540 --> 00:03:35.850 that lets us get our estimates

87 00:03:35.850 --> 00:03:38.823 and be assured that we have unbiased estimates.

88 00:03:39.970 --> 00:03:42.590 So here's an example in case there are folks

89 00:03:42.590 --> 00:03:44.500 out in the audience who don't have experience

90 00:03:44.500 --> 00:03:47.600 with the sort of complex survey design or design features.

91 00:03:47.600 --> 00:03:49.240 So this is a really silly little example

92 00:03:49.240 --> 00:03:50.530 of a stratified sample.

93 00:03:50.530 --> 00:03:52.540 So here I have a population

94 00:03:52.540 --> 00:03:54.730 of two different types of animals.

95 00:03:54.730 --> 00:03:56.710 I have cats and I have dogs.

96 00:03:56.710 --> 00:04:00.023 And in this population I happen to have 12 cats and $8.

97 00:04:00.870 --> 00:04:02.590 And I have taken a sample.

98 00:04:02.590 --> 00:04:06.560 Stratified sample where I took two cats and two dogs.

99 00:04:06.560 --> 00:04:08.890 So in this design the selection probabilities

100 00:04:08.890 --> 00:04:10.890 are known for all of the units, right?

101 00:04:10.890 --> 00:04:13.980 Because I know that there's a two out of eight chance

102 00:04:13.980 --> 00:04:16.150 I pick a dog and a two out of 12 chance

103 00:04:16.150 --> 00:04:18.440 that I pick a cat, right?

104 00:04:18.440 --> 00:04:20.530 So the probability a cat is selected is 1/6

105 00:04:20.530 --> 00:04:23.300 then the probability of dog is selected is 1/4.

106 00:04:23.300 --> 00:04:25.550 Now, how do I estimate a proportion of interest?

107 00:04:25.550 --> 00:04:27.830 Let's say it's the proportion of orange animals

108 00:04:27.830 --> 00:04:28.730 in the population.

109 00:04:28.730 --> 00:04:30.100 Like here in my sample,

110 00:04:30.100 --> 00:04:32.270 I have one of four orange animals,

111 00:04:32.270 --> 00:04:34.390 but if I chose that as my estimator

112 00:04:34.390 --> 00:04:37.180 I'd be ignoring the fact that I know how I selected

113 00:04:37.180 --> 00:04:39.310 these animals into my sample.

114 00:04:39.310 --> 00:04:41.520 So what we do is we wait the sample units

115 00:04:41.520 --> 00:04:43.930 to produce design unbiased estimates, right?

116 00:04:43.930 --> 00:04:47.580 Because this one dog kinda counts

117 00:04:47.580 --> 00:04:49.570 differently than one cat, right?

118 00:04:49.570 --> 00:04:50.950 Because there were only eight dogs

119 00:04:50.950 --> 00:04:53.600 to begin with but there were 12 cats.

120 00:04:53.600 --> 00:04:56.590 So if I want to estimate the proportion of orange animals

121 00:04:56.590 --> 00:05:00.270 I would say this cat is a weight is six

122 00:05:00.270 --> 00:05:02.340 because there's two of them and 12 total.

123 00:05:02.340 --> 00:05:04.310 So 12 divided by two is six.

124 00:05:04.310 --> 00:05:06.280 So there's six in the numerator.

125 00:05:06.280 --> 00:05:08.210 And then the denominator is the sum of the weights

126 00:05:08.210 --> 00:05:09.570 of all the selected units,

127 00:05:09.570 --> 00:05:12.150 the cats are each six and the dogs are each four.

128 00:05:12.150 --> 00:05:14.740 So I actually get my estimate a proportion of 30%.

129 00:05:14.740 --> 00:05:16.550 So instead of 25%.

130 00:05:16.550 --> 00:05:17.920 So that kind of weighted estimator

131 00:05:17.920 --> 00:05:20.190 is what we do in probability sampling.

132 00:05:20.190 --> 00:05:22.310 And we don't have to say what the distribution

133 00:05:22.310 --> 00:05:24.160 of dogs or cats is in the sample

134 00:05:24.160 --> 00:05:25.940 or orangeness in the sample,

135 00:05:25.940 --> 00:05:28.623 we entirely rely on the selection mechanism.

136 00:05:29.870 --> 00:05:32.200 What ended up happening in the real world

137 00:05:32.200 --> 00:05:34.680 a lot of the time is we don't actually get to use

138 00:05:34.680 --> 00:05:36.230 those kinds of complex designs.

139 00:05:36.230 --> 00:05:37.580 And instead we collect data

140 00:05:37.580 --> 00:05:40.230 through what's called a non-probability sample.

141 00:05:40.230 --> 00:05:42.150 So in a non-probability sample,

142 00:05:42.150 --> 00:05:43.470 it's pretty easy to define.

143 00:05:43.470 --> 00:05:46.040 You cannot calculate the probability of selection

144 00:05:46.040 --> 00:05:47.170 into the sample, right?

145 00:05:47.170 --> 00:05:49.440 So we simply don't know what the mechanism

146 00:05:49.440 --> 00:05:52.720 was that made at unit enter our sample.

147 00:05:52.720 --> 00:05:55.020 I know there's the biostatistics students in the audience,

148 00:05:55.020 --> 00:05:57.290 and you've all probably done a lot of data analysis.

149 00:05:57.290 --> 00:05:59.680 And I would venture a guess that a lot of the times

150 00:05:59.680 --> 00:06:01.090 your application datasets

151 00:06:01.090 --> 00:06:02.540 are non-probability samples, right?

152 00:06:02.540 --> 00:06:05.090 A lot of the times there are convenience samples.

153 00:06:05.090 --> 00:06:06.960 I work a lot with biomedical researchers

154 00:06:06.960 --> 00:06:08.430 studying cancer patients.

155 00:06:08.430 --> 00:06:11.580 Well guess what, it's almost always a convenient sample

156 00:06:11.580 --> 00:06:12.850 of cancer patients, right?

157 00:06:12.850 --> 00:06:14.610 It's who will agree to be in the study?

158 00:06:14.610 --> 00:06:16.770 Who can I find to be in my study?

159 00:06:16.770 --> 00:06:18.610 Other types of non-probability samples

160 00:06:18.610 --> 00:06:21.950 include things like voluntary or self-selection sampling,

161 00:06:21.950 --> 00:06:23.690 quota sampling, that's a really old,

162 00:06:23.690 --> 00:06:27.850 old school method from polling back many years ago.

163 00:06:27.850 --> 00:06:30.040 Judgment sampling or snowball sampling.

164 00:06:30.040 --> 00:06:31.030 So there's a lot of different ways

165 00:06:31.030 --> 00:06:33.053 you can get non-probability samples.

166 00:06:34.040 --> 00:06:36.800 So if we go back to the dog and cat example,

167 00:06:36.800 --> 00:06:38.970 if I didn't know anything about how these animals

168 00:06:38.970 --> 00:06:41.430 got into my sample and I just saw the four of them,

169 00:06:41.430 --> 00:06:43.210 and one of them was orange,

170 00:06:43.210 --> 00:06:48.210 I guess, I'm gonna guess 25% of my population is orange.

171 00:06:48.290 --> 00:06:49.123 Right?

172 00:06:49.123 --> 00:06:50.290 I don't have any other information

173 00:06:50.290 --> 00:06:52.500 I can't recreate the population

174 00:06:52.500 --> 00:06:54.090 like I could with the weighting.

175 00:06:54.090 --> 00:06:57.270 Where I knew how many cats in the population

176 00:06:57.270 --> 00:06:59.220 did each of my sampled cats represent

177 00:06:59.220 --> 00:07:00.790 and similarly for the dogs.

178 00:07:00.790 --> 00:07:02.830 So of course our best guess looking at these data

179 00:07:02.830 --> 00:07:04.610 would just be 25%, right?

180 00:07:04.610 --> 00:07:07.350 One out of the four animals is orange.

181 00:07:07.350 --> 00:07:10.410 So when you think about a non-probability sample,

182 00:07:10.410 --> 00:07:12.460 how much faith do you put in that estimate,

183 00:07:12.460 --> 00:07:13.403 that proportion?

184 00:07:14.640 --> 00:07:15.900 Hard to say, right?

185 00:07:15.900 --> 00:07:19.300 It depends on what you believe about the population

186 00:07:19.300 --> 00:07:22.530 and how you selected this non-probability sample

187 00:07:22.530 --> 00:07:25.620 but you do not have the safety net of the probability sample

188 00:07:25.620 --> 00:07:27.840 that guaranteed you're gonna get an unbiased estimate

189 00:07:27.840 --> 00:07:30.373 of repeated applications of the sampling.

190 00:07:31.810 --> 00:07:34.200 So I've already used the word selection bias

191 00:07:34.200 --> 00:07:36.920 a lot and sort of being assuming that, you know what I mean.

192 00:07:36.920 --> 00:07:39.580 So now I'm gonna come back to it and define it.

193 00:07:39.580 --> 00:07:42.420 So selection bias is bias arising

194 00:07:42.420 --> 00:07:44.800 when part of the target population

195 00:07:44.800 --> 00:07:46.950 is not in the sample population, right?

196 00:07:46.950 --> 00:07:49.390 So when there's a mismatch between who got into your sample

197 00:07:49.390 --> 00:07:51.250 and who was supposed to get into your sample, right?

198 00:07:51.250 --> 00:07:52.830 Who's the population?

199 00:07:52.830 --> 00:07:55.910 Or in a more general statistical kind of way,

200 00:07:55.910 --> 00:07:59.050 when some population units are sampled at a different rate

201 00:07:59.050 --> 00:08:00.100 than you meant.

202 00:08:00.100 --> 00:08:02.910 It's lik you meant for there to be a certain selection

203 00:08:02.910 --> 00:08:05.840 probability for orange animals or for dogs

204 00:08:05.840 --> 00:08:07.740 but it didn't actually end up that way.

205 00:08:07.740 --> 00:08:10.610 This will end up down the path of selection bias.

206 00:08:10.610 --> 00:08:13.090 And I will note that again, as you are biostats students

207 00:08:13.090 --> 00:08:15.080 you've probably had some epidemiology.

208 00:08:15.080 --> 00:08:17.490 And epidemiologists talk about selection bias as well.

209 00:08:17.490 --> 00:08:19.270 It's the same concept, right?

210 00:08:19.270 --> 00:08:21.810 That concept of who is ending up in your sample.

211 00:08:21.810 --> 00:08:24.383 And is there some sort of a bias in the mechanism?

212 00:08:25.610 --> 00:08:27.850 So selection bias is in fact the predominant

213 00:08:27.850 --> 00:08:30.270 concern with non-probability samples.

214 00:08:30.270 --> 00:08:32.410 In these non-probability samples,

215 00:08:32.410 --> 00:08:35.640 the units in the sample might be really different

216 00:08:35.640 --> 00:08:37.270 from the units not in the sample,

217 00:08:37.270 --> 00:08:39.570 but we can't tell how different they are.

218 00:08:39.570 --> 00:08:42.970 Whether we're talking about people, dogs, cats, hospitals,

219 00:08:42.970 --> 00:08:44.220 whatever we're talking about.

220 00:08:44.220 --> 00:08:47.260 However, these units got into my sample, I don't know.

221 00:08:47.260 --> 00:08:49.380 So I don't know if the people in my sample

222 00:08:49.380 --> 00:08:52.610 look like my population or not.

223 00:08:52.610 --> 00:08:54.560 And an important key thing to know

224 00:08:54.560 --> 00:08:56.520 is that probability samples

225 00:08:56.520 --> 00:08:59.120 when we have a low response rates, right?

226 00:08:59.120 --> 00:09:01.210 So when there are a lot of people not responding

227 00:09:01.210 --> 00:09:02.770 you're basically ending up back

228 00:09:02.770 --> 00:09:04.730 at a non-probability sample, right?

229 00:09:04.730 --> 00:09:06.660 Where we have this beautiful design,

230 00:09:06.660 --> 00:09:10.180 we know everybody's sampling weight, we draw a sample,

231 00:09:10.180 --> 00:09:13.510 oops, ut then only 30% of people respond to my sample.

232 00:09:13.510 --> 00:09:16.050 You're basically injecting that bias back in again.

233 00:09:16.050 --> 00:09:19.673 Sort of undoing the beauty of the probability sample.

234 00:09:20.920 --> 00:09:22.780 So when we think about a selection

235 00:09:22.780 --> 00:09:25.300 bias and selection into a sample,

236 00:09:25.300 --> 00:09:27.570 we can categorize them in two ways.

237 00:09:27.570 --> 00:09:30.400 And Dr. McDougal, actually,

238 00:09:30.400 --> 00:09:32.100 when he was giving you my brief little bio

239 00:09:32.100 --> 00:09:34.350 used the words that I'm sure you've used

240 00:09:34.350 --> 00:09:37.260 in your classes before like ignorable and non-ignorable.

241 00:09:37.260 --> 00:09:39.410 These are usually are more commonly applied

242 00:09:39.410 --> 00:09:40.660 to missingness, right?

243 00:09:40.660 --> 00:09:42.560 So ignorable missingness mechanisms

244 00:09:42.560 --> 00:09:45.210 and non-ignorable missingness mechanisms.

245 00:09:45.210 --> 00:09:47.640 Missing at random, missing completely at random

246 00:09:47.640 --> 00:09:49.900 or missing not at random, right?

247 00:09:49.900 --> 00:09:51.720 Same exact framework here.

248 00:09:51.720 --> 00:09:53.750 But instead of talking about missingness

249 00:09:53.750 --> 00:09:56.390 we're talking about selection into the sample.

250 00:09:56.390 --> 00:09:58.850 So when we have an ignorable selection mechanism,

251 00:09:58.850 --> 00:10:00.550 that means the probability of selection

252 00:10:00.550 --> 00:10:01.977 depends on things I observed.

253 00:10:01.977 --> 00:10:05.170 Right, it depends on the observed characteristics.

254 00:10:05.170 --> 00:10:07.700 When I have a non-negotiable selection mechanism

255 00:10:07.700 --> 00:10:09.514 now that probability of selection depends

256 00:10:09.514 --> 00:10:11.820 on observed characteristics.

257 00:10:11.820 --> 00:10:13.790 Again, this is not really a new concept

258 00:10:13.790 --> 00:10:15.310 if you understanded about missing data,

259 00:10:15.310 --> 00:10:18.453 just apply to selection into the sample.

260 00:10:19.670 --> 00:10:21.560 So in a probability sample

261 00:10:21.560 --> 00:10:24.060 we might have different probabilities of selection

262 00:10:24.060 --> 00:10:27.760 for different types of units like for cats versus for dogs.

263 00:10:27.760 --> 00:10:30.670 But we know exactly how they differ, right?

264 00:10:30.670 --> 00:10:32.890 It's because I designed my survey

265 00:10:32.890 --> 00:10:35.720 based on his characteristic of dog versus cat

266 00:10:35.720 --> 00:10:38.110 and I know exactly the status of dog versus cat

267 00:10:38.110 --> 00:10:41.690 for my entire population in order to do that selection.

268 00:10:41.690 --> 00:10:45.320 So I absolutely can estimate the proportion of orange,

269 00:10:45.320 --> 00:10:49.390 animals unbiasedly in the sense of taking repeated

270 00:10:49.390 --> 00:10:51.910 stratified samples and estimating that proportion.

271 00:10:51.910 --> 00:10:54.360 I hadn't guaranteed that I'm gonna get an unbiased

272 00:10:54.360 --> 00:10:55.430 estimate, right?

273 00:10:55.430 --> 00:10:57.300 So this selection mechanism

274 00:10:57.300 --> 00:10:59.760 is definitely not non-ignorable, right?

275 00:10:59.760 --> 00:11:01.980 This is definitely an ignorable selection mechanism

276 00:11:01.980 --> 00:11:03.540 in the sense that it only depends

277 00:11:03.540 --> 00:11:05.800 on observed characteristics.

278 00:11:05.800 --> 00:11:09.200 But if my four animals had just come from,

279 00:11:09.200 --> 00:11:10.033 I don’t know where?

280 00:11:10.033 --> 00:11:11.030 Convenience.

281 00:11:11.030 --> 00:11:13.830 Well now why did they end up in my sample?

282 00:11:13.830 --> 00:11:16.110 It could depend on something that we didn’t observe.

283 00:11:16.110 --> 00:11:17.670 What breed of dog it was?

284 00:11:17.670 --> 00:11:20.080 The age of the dog, the color of the dog.

285 00:11:20.080 --> 00:11:22.340 It could have been pretty much anything, right?

286 00:11:22.340 --> 00:11:24.180 That’s the problem with the convenient sample.

287 00:11:24.180 --> 00:11:25.410 You don’t know why those units

288 00:11:25.410 --> 00:11:28.303 often self-selected to be into your sample.

289 00:11:29.350 --> 00:11:32.050 So now I’m gonna head into the kind of ugly statistical

290 00:11:32.050 --> 00:11:34.750 notation portion of this stock.

291 00:11:34.750 --> 00:11:36.720 So we’ll start with estimated proportions.

292 00:11:36.720 --> 00:11:40.658 So we’ll use Y as our binary indicator

293 00:11:40.658 --> 00:11:42.860 for the outcome, okay?

294 00:11:42.860 --> 00:11:45.310 But here I’m gonna talk about Y

295 00:11:45.310 --> 00:11:48.670 more generally as all the survey data.

296 00:11:48.670 --> 00:11:50.110 So we’ll start with Y as all the survey data,

297 00:11:50.110 --> 00:11:51.210 then we’re gonna narrow it down to Y

298 00:11:51.210 --> 00:11:52.940 as the binary indicator?

299 00:11:52.940 --> 00:11:56.740 So we can partition our survey data into the data

300 00:11:56.740 --> 00:11:58.197 for the units we got in the sample

301 00:11:58.197 --> 00:12:01.020 and the data for units that are not in the sample.

302 00:12:01.020 --> 00:12:02.700 I so selected into the sample versus

303 00:12:02.700 --> 00:12:04.640 not selected into the sample.

304 00:12:04.640 --> 00:12:07.180 But for everybody I have Z,

305 00:12:07.180 --> 00:12:08.740 I have some fully observed

306 00:12:08.740 --> 00:12:11.310 what are often called design variables.

307 00:12:11.310 --> 00:12:13.960 So this is where we are using information

308 00:12:13.960 --> 00:12:16.140 that we know about an entire population

309 00:12:16.140 --> 00:12:19.520 to select our sample in the world of probability sampling.

310 00:12:19.520 --> 00:12:21.653 And then S is the selection indicator.

311 00:12:22.520 --> 00:12:25.840 So these three variables have a joint distribution.

312 00:12:25.840 --> 00:12:27.070 And most of the time,

313 00:12:27.070 --> 00:12:29.940 what we care about is Y given Z.

314 00:12:29.940 --> 00:12:31.950 Right, we're interested in estimating

315 00:12:31.950 --> 00:12:34.120 some outcome characteristic

316 00:12:34.120 --> 00:12:36.890 conditional on some other characteristic, right?

317 00:12:36.890 --> 00:12:40.360 Average weight for dogs, average weight for cats, right?

318 00:12:40.360 --> 00:12:42.010 Y given Z.

319 00:12:42.010 --> 00:12:45.440 But Y given Z is only part of the issue,

320 00:12:45.440 --> 00:12:47.750 there's also a selection mechanism, right?

321 00:12:47.750 --> 00:12:49.120 So there's also this function

322 00:12:49.120 --> 00:12:53.320 of how do you predict selection S with Y and Z.

323 00:12:53.320 --> 00:12:56.210 And I'm using this additional Greek letter psi here

324 00:12:56.210 --> 00:12:58.230 to denote additional variables

325 00:12:58.230 --> 00:12:59.830 that might be involved, right?

326 00:12:59.830 --> 00:13:02.540 ’Cause selection could depend on more than just Y and Z.

327 00:13:02.540 --> 00:13:04.230 It could depend on something outside

328 00:13:04.230 --> 00:13:05.593 of that set of variables.

329 00:13:06.670 --> 00:13:08.230 So when we have probability sampling,

330 00:13:08.230 --> 00:13:09.140 we have what’s called

331 00:13:09.140 --> 00:13:12.270 an extremely ignorable selection mechanism,

332 00:13:12.270 --> 00:13:14.320 which means selection can depend on Z,

333 00:13:14.320 --> 00:13:16.440 like when we stratified on animal type

334 00:13:16.440 --> 00:13:18.470 but it cannot depend on Y.

335 00:13:18.470 --> 00:13:21.960 Either the selected units Y or the excluded units Y

336 00:13:21.960 --> 00:13:23.830 doesn’t depend on either.

337 00:13:23.830 --> 00:13:27.340 Kind of vaguely like the MCAR of selection mechanisms.

338 00:13:27.340 --> 00:13:29.340 It doesn’t depend on Y at all.

339 00:13:29.340 --> 00:13:30.520 Observed or unobserved.

340 00:13:30.520 --> 00:13:31.460 But it can depend on Z.

341 00:13:31.460 --> 00:13:33.680 So that makes it different than MCAR.

342 00:13:33.680 --> 00:13:35.800 So including a unit into the sample

343 00:13:35.800 --> 00:13:38.930 is independent of those survey outcomes Y

344 00:13:38.930 --> 00:13:41.110 and also any unobserved variables, right?

345 00:13:41.110 --> 00:13:43.720 That phi here, that phi goes away.

346 00:13:43.720 --> 00:13:46.310 So selection only depends on Z.

347 00:13:46.310 --> 00:13:49.170 So if I’m interested in this inference target

348 00:13:49.170 --> 00:13:51.490 I can ignore the selection mechanism.

349 00:13:51.490 --> 00:13:54.060 So this is kind of parallels that idea

350 00:13:54.060 --> 00:13:56.320 in the missingness, in the missing data literature, right?

351 00:13:56.320 --> 00:13:58.520 If I have an ignorable missingness mechanism

352 00:13:58.520 --> 00:14:00.350 I can ignore that part of it.

353 00:14:00.350 --> 00:14:01.870 I don’t have to worry about modeling

354 00:14:01.870 --> 00:14:03.870 the probability that a unit is selected.

355 00:14:05.010 --> 00:14:08.230 But the bad news in our non-probability sampling,

356 00:14:08.230 --> 00:14:10.610 very, very arguably true

357 00:14:10.610 --> 00:14:13.010 that you could have non ignorable selection, right?

358 00:14:13.010 --> 00:14:16.030 It's easy to make an argument for why the people

359 00:14:16.030 --> 00:14:17.500 who ended up into your sample,

360 00:14:17.500 --> 00:14:20.210 your convenient sample are different than the people

361 00:14:20.210 --> 00:14:22.130 who don't enter your sample.

362 00:14:22.130 --> 00:14:24.030 Think about some of these big data examples.

363 00:14:24.030 --> 00:14:25.610 Think about Twitter data.

364 00:14:25.610 --> 00:14:26.840 Well, I mean, you know,

365 00:14:26.840 --> 00:14:28.730 the people who use Twitter are different

366 00:14:28.730 --> 00:14:30.720 than the people who don't use Twitter, right?

367 00:14:30.720 --> 00:14:32.400 In lots of different ways.

368 00:14:32.400 --> 00:14:33.670 So if you're going to think about drawing

369 00:14:33.670 --> 00:14:35.940 some kind of inference about the population,

370 00:14:35.940 --> 00:14:39.100 you can't just ignore that selection mechanism.

371 00:14:39.100 --> 00:14:40.770 You need to think about how do they enter

372 00:14:40.770 --> 00:14:42.210 into your Twitter sample

373 00:14:42.210 --> 00:14:44.150 and how might they be different than the people

374 00:14:44.150 --> 00:14:47.040 who did not enter into your Twitter sample.

375 00:14:47.040 --> 00:14:49.280 So when we're thinking about the selection mechanism

376 00:14:49.280 --> 00:14:50.860 basically nothing goes away, right?

377 00:14:50.860 --> 00:14:53.297 We can't ignore this selection mechanism.

378 00:14:53.297 --> 00:14:54.570 But we have to think

379 00:14:54.570 --> 00:14:55.930 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 problem 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.

434 00:17:04.490 --> 00:17:05.820 So that’s not what we’re interested in.

435 00:17:05.820 --> 00:17:08.870 ’Cause we’re really worried in the non probability space

436 00:17:08.870 --> 00:17:11.603 that we can’t ignore the selection mechanism.

437 00:17:12.670 --> 00:17:14.840 And there isn’t relatively new indicator

438 00:17:14.840 --> 00:17:17.773 called that they called the SMUB, S-M-U-B.

439 00:17:18.710 --> 00:17:21.140 That is an index that actually extends

440 00:17:21.140 --> 00:17:22.840 this idea of selection bias

441 00:17:22.840 --> 00:17:25.410 to allow for non ignorable selection.

442 00:17:25.410 --> 00:17:28.760 So it lets you say, well, what would my point estimate

443 00:17:28.760 --> 00:17:32.880 be for a mean if selection were in fact ignorable,

444 00:17:32.880 --> 00:17:34.500 and now let’s go to the other extreme,

445 00:17:34.500 --> 00:17:37.080 suppose selection only depends on Y.

446 00:17:37.080 --> 00:17:39.050 And I’m trying to estimate average weight

447 00:17:39.050 --> 00:17:40.490 and whether or not you entered my sample

448 00:17:40.490 --> 00:17:42.740 is entirely dependent on your weight.

449 00:17:42.740 --> 00:17:44.680 That’s really not ignorable.

450 00:17:44.680 --> 00:17:47.080 And then it kinda bounds the potential magnitude

451 00:17:47.080 --> 00:17:48.083 for the problem.

452 00:17:48.930 --> 00:17:51.870 So that SMUB, this estimator is really close

453 00:17:51.870 --> 00:17:54.720 to what we want but we want it for proportions.

454 00:17:54.720 --> 00:17:59.720 especially because in survey work and in large datasets,

455 00:18:00.390 --> 00:18:02.630 we very often have categorical data

456 00:18:02.630 --> 00:18:05.160 or very, very often binary data.

457 00:18:05.160 --> 00:18:06.860 If you think about if you’ve ever participated

458 00:18:06.860 --> 00:18:09.710 in an online survey or filled out those kinds of things

459 00:18:09.710 --> 00:18:11.230 very often, right, You’re checking a box.

460 00:18:11.230 --> 00:18:13.200 It's multiple choice, select all that apply.

461 00:18:13.200 --> 00:18:16.540 It's lots and lots of binary data floating around out there.

462 00:18:16.540 --> 00:18:19.200 And I'll show you a couple of examples.

463 00:18:19.200 --> 00:18:21.780 So that was a lot of kind of me talking

464 00:18:21.780 --> 00:18:23.460 at you about the framework.

465 00:18:23.460 --> 00:18:27.250 Now, let me bring this down to a solid example application.

466 00:18:27.250 --> 00:18:29.650 So I'm going to use the national survey

467 00:18:29.650 --> 00:18:32.370 of family growth as a fake population.

468 00:18:32.370 --> 00:18:35.600 So I want you to pretend that I have a population

469 00:18:35.600 --> 00:18:37.880 of 19,800 people, right?

470 00:18:37.880 --> 00:18:40.440 It happens to be that I pulled it from the national survey

471 00:18:40.440 --> 00:18:41.273 of family growth,

472 00:18:41.273 --> 00:18:43.150 that's not really important that that was the source.

473 00:18:43.150 --> 00:18:46.310 I've got this population of about 20,000 people.

474 00:18:46.310 --> 00:18:48.240 But let's pretend we're doing a study

475 00:18:48.240 --> 00:18:49.890 and I was only able to select

476 00:18:49.890 --> 00:18:51.890 into my sample smartphone users.

477 00:18:51.890 --> 00:18:54.430 Because I did some kind of a survey that was on their,

478 00:18:54.430 --> 00:18:55.750 you had to take it on your phone.

479 00:18:55.750 --> 00:18:57.170 So if you did not have a smartphone

480 00:18:57.170 --> 00:19:00.050 you could not be selected into my sample.

481 00:19:00.050 --> 00:19:02.740 In this particular case, in this fake population,

482 00:19:02.740 --> 00:19:04.490 it's a very high selection fraction.

483 00:19:04.490 --> 00:19:07.260 So about 80% of my population is in my sample.

484 00:19:07.260 --> 00:19:10.620 That in and of itself is very unusual, right?

485 00:19:10.620 --> 00:19:12.540 A non-probability sample is usually very,

486 00:19:12.540 --> 00:19:15.370 very small compared to the full population

487 00:19:15.370 --> 00:19:16.580 let's say of the United States

488 00:19:16.580 --> 00:19:18.220 if that's who we're trying to generalize to.

489 00:19:18.220 --> 00:19:19.640 But for the purposes of illustration

490 00:19:19.640 --> 00:19:22.330 it helps to have a pretty high selection fraction.

491 00:19:22.330 --> 00:19:24.280 And we'll assume that the outcome we're interested

492 00:19:24.280 --> 00:19:27.930 in is whether or not the individual has ever been married.

493 00:19:27.930 --> 00:19:29.390 So this is person level data, right?

494 00:19:29.390 --> 00:19:30.600 Ever been married.

495 00:19:30.600 --> 00:19:32.410 And it is...

496 00:19:32.410 --> 00:19:33.980 we wanna estimate it by gender,

497 00:19:33.980 --> 00:19:36.400 and I will note that the NSFG only calculate

498 00:19:36.400 --> 00:19:39.000 or only captures gender as a binary variable.

499 00:19:39.000 --> 00:19:40.930 This is a very long standing survey,

500 00:19:40.930 --> 00:19:42.430 been going on since the seventies.

501 00:19:42.430 --> 00:19:44.800 We know our understanding of gender as a construct

502 00:19:44.800 --> 00:19:46.590 has grown a lot since the seventies

503 00:19:46.590 --> 00:19:48.320 but this survey, and in fact

504 00:19:48.320 --> 00:19:50.840 many governmental surveys still treat gender

505 00:19:50.840 --> 00:19:51.930 as a binary variable.

506 00:19:51.930 --> 00:19:53.840 So that's our limitation here

507 00:19:53.840 --> 00:19:56.330 but I just want to acknowledge that.

508 00:19:56.330 --> 00:19:57.980 So in this particular case,

509 00:19:57.980 --> 00:19:59.960 we know the true selection bias, right?

510 00:19:59.960 --> 00:20:03.580 Because I actually have all roughly 20,000 people

511 00:20:03.580 --> 00:20:05.990 so that therefore I can calculate what's the truth,

512 00:20:05.990 --> 00:20:08.287 and then I can use my smartphone sample and say,

513 00:20:08.287 --> 00:20:10.630 ”Well, how much bias is there?”

514 00:20:10.630 --> 00:20:12.930 So it turns out that in the full sample

515 00:20:12.930 --> 00:20:16.320 46.8% of the females have never been married.

516 00:20:16.320 --> 00:20:19.830 And 56.6% of the males had never been married.

517 00:20:19.830 --> 00:20:22.890 But if I use my selected sample of smartphone users

518 00:20:22.890 --> 00:20:24.880 I’m getting a, well, very close,

519 00:20:24.880 --> 00:20:27.710 but slightly smaller estimate for females.

520 00:20:27.710 --> 00:20:30.170 46.6% never married.

521 00:20:30.170 --> 00:20:31.990 And for males it’s like about a percentage

522 00:20:31.990 --> 00:20:35.290 point lower than the truth, 55.5%.

523 00:20:35.290 --> 00:20:37.610 So not a huge amount of bias here.

524 00:20:37.610 --> 00:20:41.070 My smartphone users are not all that non-representative

525 00:20:41.070 --> 00:20:42.920 with respect to the entire sample,

526 00:20:42.920 --> 00:20:44.390 at least with respect to whether

527 00:20:44.390 --> 00:20:46.810 or not they’ve ever been married.

528 00:20:46.810 --> 00:20:48.670 So when we have binary data,

529 00:20:48.670 --> 00:20:52.790 an important point of reference is what happens if we assume

530 00:20:52.790 --> 00:20:55.410 everybody not in my sample is a one, right?

531 00:20:55.410 --> 00:20:58.030 What if everybody not in my sample was never married

532 00:20:58.030 --> 00:21:00.660 or everyone not in my sample

533 00:21:00.660 --> 00:21:02.730 is a no to never married, right?

534 00:21:02.730 --> 00:21:05.260 So like has, has ever been married?

535 00:21:05.260 --> 00:21:07.410 And these are what’s called the Manski bounds.

536 00:21:07.410 --> 00:21:10.140 When you fill in all zeros or fill in old bonds

537 00:21:10.140 --> 00:21:12.167 for the missing values or the values

538 00:21:12.167 --> 00:21:14.080 for those non-selected folks.

539 00:21:14.080 --> 00:21:15.490 So we can bound the bias.

540 00:21:15.490 --> 00:21:20.490 So the bias of this estimate of 46.6 or 46.6%

541 00:21:20.770 --> 00:21:22.680 has to be by definition

542 00:21:22.680 --> 00:21:25.910 between negative 0.098 and positive 0.085.

543 00:21:25.910 --> 00:21:28.850 Because those are the two ends of putting all zeros

544 00:21:28.850 --> 00:21:32.090 or all ones for the people who are not in my sample.

545 00:21:32.090 --> 00:21:34.610 So this is unlike a continuous variable, right?

546 00:21:34.610 --> 00:21:37.810 Where we can't actually put a finite bound on the bias.

547 00:21:37.810 --> 00:21:39.670 We can with a proportion, right?

548 00:21:39.670 --> 00:21:42.140 So this is why, for example,

549 00:21:42.140 --> 00:21:45.010 if any of you ever work on smoking cessation studies

550 00:21:45.010 --> 00:21:46.850 often they do sensitivity analysis.

551 00:21:46.850 --> 00:21:49.710 People who drop out assume they're all smoking, right?

552 00:21:49.710 --> 00:21:51.400 Or assume they're all not smoking.

553 00:21:51.400 --> 00:21:53.180 They're not calling it that

554 00:21:53.180 --> 00:21:56.240 but they're getting the Manski bounds.

555 00:21:56.240 --> 00:21:57.200 Okay.

556 00:21:57.200 --> 00:22:00.080 So the question is, can we do better than the Manski bounds?

557 00:22:00.080 --> 00:22:02.360 Because these are actually pretty wide bounds,

558 00:22:02.360 --> 00:22:04.100 relative to the size of the true bias,

559 00:22:04.100 --> 00:22:06.170 and these are very wide.

560 00:22:06.170 --> 00:22:10.190 And imagine a survey where we didn't have 80% selected.

561 00:22:10.190 --> 00:22:11.870 What if we had 10% selected?

562 00:22:11.870 --> 00:22:13.990 Well, then the Manski bounds are gonna be useless, right?

563 00:22:13.990 --> 00:22:15.670 plug in, all zeros plug in all ones,

564 00:22:15.670 --> 00:22:17.420 you're gonna get these insane estimates

565 00:22:17.420 --> 00:22:19.620 that are nowhere close to what you observed.

566 00:22:20.800 --> 00:22:22.920 So going back to the statistical notation,

567 00:22:22.920 --> 00:22:24.400 this is where I said we're going to use Y

568 00:22:24.400 --> 00:22:25.550 in a slightly different way.

569 00:22:25.550 --> 00:22:30.070 Now, Y, and now forward is the binary variable of interest.

570 00:22:30.070 --> 00:22:32.680 So in this case, in this NSFG example

571 00:22:32.680 --> 00:22:34.003 it was never married.

572 00:22:34.900 --> 00:22:38.490 We have a bunch of auxiliary variables that we observed

573 00:22:38.490 --> 00:22:41.180 for everybody in the selected sample;

574 00:22:41.180 --> 00:22:43.310 age, race, education, et cetera,

575 00:22:43.310 --> 00:22:44.843 and I'm gonna call those Z.

576 00:22:47.560 --> 00:22:50.640 Assume also that we have summary statistics

577 00:22:50.640 --> 00:22:52.950 on Z for the selected cases.

578 00:22:52.950 --> 00:22:55.460 So I don't observe Z for everybody, right?

579 00:22:55.460 --> 00:22:56.950 All my non-smartphone users,

580 00:22:56.950 --> 00:22:59.670 I don't know for each one of them, what is their gender?

581 00:22:59.670 --> 00:23:01.650 What is their age? What is their race?

582 00:23:01.650 --> 00:23:03.310 But I don't actually observe that.

583 00:23:03.310 --> 00:23:05.610 But I observed some kinda summary statistic.

584 00:23:05.610 --> 00:23:09.150 But a mean vector and a covariance matrix of Z.

585 00:23:09.150 --> 00:23:12.240 So I have some source of what does my population

586 00:23:12.240 --> 00:23:14.300 look like at an aggregate level?

587 00:23:14.300 --> 00:23:16.120 And in practice, this would come from something

588 00:23:16.120 --> 00:23:19.510 like census data or in a very large probability sample,

589 00:23:19.510 --> 00:23:21.020 something where we would be pretty confident

590 00:23:21.020 --> 00:23:23.440 This is reflective of the population.

591 00:23:23.440 --> 00:23:27.000 Will note that if we have data for the population

592 00:23:27.000 --> 00:23:28.510 and not the non-selected,

593 00:23:28.510 --> 00:23:30.180 then we can kinda do subtraction, right?

594 00:23:30.180 --> 00:23:32.460 We can take the data for the population

595 00:23:32.460 --> 00:23:34.630 and aggregate and go backwards

596 00:23:34.630 --> 00:23:36.320 to figure out what it would be for the non-selected

597 00:23:36.320 --> 00:23:40.090 by effectively backing out the selected cases.

598 00:23:40.090 --> 00:23:41.590 And similarly another problem

599 00:23:41.590 --> 00:23:42.530 is that we don't have the variance.

600 00:23:42.530 --> 00:23:44.040 We could just assume it's what we observe

601 00:23:44.040 --> 00:23:45.140 in the selected cases.

602 00:23:46.450 --> 00:23:48.490 So how are we gonna use this in order

603 00:23:48.490 --> 00:23:52.410 to estimate of selection bias,

604 00:23:52.410 --> 00:23:53.243 what we're gonna come up

605 00:23:53.243 --> 00:23:56.210 with this measure of unadjusted bias for proportions

606 00:23:56.210 --> 00:23:57.823 called the MUBP.

607 00:23:58.760 --> 00:24:01.940 So the MUBP is an extension of the SMUB

608 00:24:01.940 --> 00:24:04.470 that was for means, for continuous variables

609 00:24:04.470 --> 00:24:06.030 to binary outcomes, right?

610 00:24:06.030 --> 00:24:07.470 To proportions.

611 00:24:07.470 --> 00:24:10.380 High-level, it's based on pattern-mixture models.

612 00:24:10.380 --> 00:24:12.700 It requires you to make explicit assumptions

613 00:24:12.700 --> 00:24:15.470 about the distribution of the selection mechanism,

614 00:24:15.470 --> 00:24:17.730 and it provides you a sensitivity analysis,

615 00:24:17.730 --> 00:24:20.010 basically make different assumptions on S,

616 00:24:20.010 --> 00:24:21.910 I don't know what that distribution is,

617 00:24:21.910 --> 00:24:24.240 and you're gonna get a range of bias.

618 00:24:24.240 --> 00:24:27.950 So that's that idea of how wrong might we be?

619 00:24:27.950 --> 00:24:29.990 So we're trying to just tighten those bounds

620 00:24:29.990 --> 00:24:30.910 compared to the Manski bounce.

621 00:24:30.910 --> 00:24:33.480 Where we don't wanna have to rely on plug in all zeros,

622 00:24:33.480 --> 00:24:34.550 plug in all ones,

623 00:24:34.550 --> 00:24:35.750 we wanna shrink that interval

624 00:24:35.750 --> 00:24:38.420 to give us something a little bit more meaningful.

625 00:24:38.420 --> 00:24:40.910 So the basic idea behind how this works

626 00:24:40.910 --> 00:24:44.160 before I show you the formulas is we can measure

627 00:24:44.160 --> 00:24:47.480 the degree of selection bias in Z, right?

628 00:24:47.480 --> 00:24:50.390 Because we observed Z for our selected sample,

629 00:24:50.390 --> 00:24:53.170 and we observed at an aggregate for the population.

630 00:24:53.170 --> 00:24:56.370 So I can see, for example, that if in my selected sample,

631 00:24:56.370 --> 00:25:00.970 I have 55% females but in the population it's 50% females.

632 00:25:00.970 --> 00:25:02.590 Well, I can see that bias.

633 00:25:02.590 --> 00:25:04.330 Right, I can do that comparison.

634 00:25:04.330 --> 00:25:08.360 So absolutely I can tell you how much selection bias

635 00:25:08.360 --> 00:25:11.380 there is for all of my auxiliary variables.

636 00:25:11.380 --> 00:25:15.670 So if my outcome Y is related to my Zs

637 00:25:15.670 --> 00:25:18.550 then knowing something about the selection bias in Z

638 00:25:18.550 --> 00:25:21.970 tells me something about the selection bias in Y.

639 00:25:21.970 --> 00:25:24.700 It doesn't tell me exactly the selection bias in Y

640 00:25:24.700 --> 00:25:28.380 but it gives me some information in the selection bias in Y.

641 00:25:28.380 --> 00:25:31.850 So in the extreme imagine if your Zs

642 00:25:31.850 --> 00:25:33.340 in your selected sample

643 00:25:33.340 --> 00:25:36.210 in aggregate looked exactly like the population.

644 00:25:36.210 --> 00:25:39.600 Well, then you'd be pretty confident, right?

645 00:25:39.600 --> 00:25:41.850 That there's not an enormous amount of selection bias

646 00:25:41.850 --> 00:25:44.623 in Y assuming that Y was related to the Z.

647 00:25:46.290 --> 00:25:48.020 So we're gonna use pattern-mixture models

648 00:25:48.020 --> 00:25:51.770 to explicitly model that distribution of S, right?

649 00:25:51.770 --> 00:25:53.960 And we're especially gonna focus on the case

650 00:25:53.960 --> 00:25:55.930 when selection depends on Y.

651 00:25:55.930 --> 00:25:59.483 It depends on our binary outcome of interest.

652 00:26:00.320 --> 00:26:02.880 So again, Y is that binary variable interest,

653 00:26:02.880 --> 00:26:05.380 we only have it for the selected sample.

654 00:26:05.380 --> 00:26:08.420 In the NSFG example it's whether the woman or man

655 00:26:08.420 --> 00:26:09.740 has ever been married.

656 00:26:09.740 --> 00:26:12.970 We have Z variables available for the selected cases

657 00:26:12.970 --> 00:26:16.280 in micro data and an aggregate for the non-selected sample,

658 00:26:16.280 --> 00:26:17.590 a demographic characteristics

659 00:26:17.590 --> 00:26:20.713 like age, education, marital status, et cetera.

660 00:26:21.740 --> 00:26:23.610 And the way that we're gonna go

661 00:26:23.610 --> 00:26:24.920 about doing this is we're gonna try

662 00:26:24.920 --> 00:26:27.230 to get back to the idea of normality,

663 00:26:27.230 --> 00:26:30.330 because then as you all know, when everything's normal

664 00:26:30.330 --> 00:26:31.680 it's great, right?

665 00:26:31.680 --> 00:26:34.210 It's easy to work with the normal distribution.

666 00:26:34.210 --> 00:26:36.720 So the way we can do that with a binary variable

667 00:26:36.720 --> 00:26:39.330 is we can think about latent variables.

668 00:26:39.330 --> 00:26:42.150 So we're going to think about a latent variable called U.

669 00:26:42.150 --> 00:26:44.840 That is an underlying, unobserved latent variables.

670 00:26:44.840 --> 00:26:48.040 So unobserved for everybody, including our selected sample.

671 00:26:48.040 --> 00:26:49.950 And it's basically thresholded.

672 00:26:49.950 --> 00:26:54.460 And when U crosses zero, well, then Y goes from zero to one.

673 00:26:54.460 --> 00:26:57.940 So I'm sure many, all of you have seen probit regression,

674 00:26:57.940 --> 00:26:59.250 or this is what happens

675 00:26:59.250 --> 00:27:01.360 and this is how probit regression is justified,

676 00:27:01.360 --> 00:27:02.583 via latent variables.

677 00:27:03.540 --> 00:27:05.920 So we're going to take our Zs

678 00:27:05.920 --> 00:27:08.220 that we have for the selected cases,

679 00:27:08.220 --> 00:27:11.030 and essentially reduce the dimensionality.

680 00:27:11.030 --> 00:27:12.680 We're gonna take the Zs,

681 00:27:12.680 --> 00:27:17.080 run a probate regression of Y on Z in the selected cases,

682 00:27:17.080 --> 00:27:18.890 and pull out the linear predictor

683 00:27:18.890 --> 00:27:20.320 from the regression, right?

684 00:27:20.320 --> 00:27:22.430 The X beta, right?

685 00:27:22.430 --> 00:27:24.050 Sorry, Z beta.

686 00:27:24.050 --> 00:27:25.460 And I'm gonna call that X.

687 00:27:25.460 --> 00:27:29.580 That is my proxy for Y or my Y hat, right?

688 00:27:29.580 --> 00:27:31.560 It's just the predicted value from the regression.

689 00:27:31.560 --> 00:27:34.660 And I can get that for every single observation

690 00:27:34.660 --> 00:27:36.770 in my selected sample, of course, right?

691 00:27:36.770 --> 00:27:39.120 Just plug in each individual's Z values

692 00:27:39.120 --> 00:27:40.390 and get out their Y hat.

693 00:27:40.390 --> 00:27:42.240 That's my proxy value.

694 00:27:42.240 --> 00:27:43.540 And it's called the proxy

695 00:27:43.540 --> 00:27:45.060 because it's the prediction, right?

696 00:27:45.060 --> 00:27:46.820 It's our sort of best guess at Y

697 00:27:46.820 --> 00:27:47.903 based on this model.

698 00:27:48.760 --> 00:27:52.000 So I can get it for every observation in my selected sample,

699 00:27:52.000 --> 00:27:55.720 but very importantly I can also get it on average

700 00:27:55.720 --> 00:27:57.480 for the non-selective sample.

701 00:27:57.480 --> 00:28:01.130 So I have all my beta hats for my probit regression,

702 00:28:01.130 --> 00:28:03.050 and I'm gonna plug in Z-bar.

703 00:28:03.050 --> 00:28:05.880 And I'm going to plug in the average value of my Zs.

704 00:28:05.880 --> 00:28:08.160 And that's going to give me the average value

705 00:28:08.160 --> 00:28:10.890 of X for the non-selected cases.

706 00:28:10.890 --> 00:28:12.930 I don't have an actual observed value

707 00:28:12.930 --> 00:28:14.580 for all those non-selective cases

708 00:28:14.580 --> 00:28:16.390 but I have the average, right?

709 00:28:16.390 --> 00:28:19.240 So I could think about comparing the average Z value

710 00:28:19.240 --> 00:28:22.170 in the aggregate, in the non-selected cases

711 00:28:22.170 --> 00:28:24.180 to that average Z among my selected cases.

712 00:28:24.180 --> 00:28:25.540 And that is of course

713 00:28:25.540 --> 00:28:27.890 exactly where we're gonna get those index from.

714 00:28:28.970 --> 00:28:31.100 So I have my selection indicator S,

715 00:28:31.100 --> 00:28:33.000 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

741 00:29:36.650 --> 00:29:39.123 just Bernoulli, for the selection indicator.

742 00:29:40.070 --> 00:29:43.367 However, I'm sure you all immediately are thinking,

743 00:29:43.367 --> 00:29:44.627 "Well, that's great,

744 00:29:44.627 --> 00:29:47.187 "but I don't have any information to estimate

745 00:29:47.187 --> 00:29:50.830 "some of these parameters for the non-selected cases."

746 00:29:50.830 --> 00:29:52.970 Clearly, for the selected cases, right?

747 00:29:52.970 --> 00:29:53.803 S equals one.

748 00:29:53.803 --> 00:29:55.220 I can estimate all of these things.

749 00:29:55.220 --> 00:29:58.480 But I can't estimate them for the non-selected sample

750 00:29:58.480 --> 00:30:00.520 because I might observe X-bar

751 00:30:00.520 --> 00:30:03.100 but I don't observe anything having to do with you.

752 00:30:03.100 --> 00:30:05.660 'Cause I have no Y information.

753 00:30:05.660 --> 00:30:07.500 So in order to identify this model

754 00:30:07.500 --> 00:30:08.870 and be able to come up with estimates

755 00:30:08.870 --> 00:30:10.210 for all of these parameters,

756 00:30:10.210 --> 00:30:13.460 we have to make an assumption about the selection mechanism.

757 00:30:13.460 --> 00:30:16.070 So we assume that the probability of selection

758 00:30:16.070 --> 00:30:19.070 into my sample is a function of U.

759 00:30:19.070 --> 00:30:20.690 So we're allowing it to be not ignorable.

760 00:30:20.690 --> 00:30:23.170 Remember that's underlying Y and X,

761 00:30:23.170 --> 00:30:25.450 that proxy which is a function of Z.

762 00:30:25.450 --> 00:30:29.520 So that's observed and V, those other variables.

763 00:30:29.520 --> 00:30:30.940 And in particular, we're assuming

764 00:30:30.940 --> 00:30:33.910 that it's this funny looking form of combination

765 00:30:33.910 --> 00:30:35.150 of X and U.

766 00:30:35.150 --> 00:30:38.490 That depends on this sensitivity parameter phi.

767 00:30:38.490 --> 00:30:41.010 So phi it's one minus phi times X

768 00:30:41.010 --> 00:30:42.790 and phi times U.

769 00:30:42.790 --> 00:30:44.640 So that's essentially weighting

770 00:30:44.640 --> 00:30:46.780 the contributions of those two pieces.

771 00:30:46.780 --> 00:30:48.750 How much of selection is dependent

772 00:30:48.750 --> 00:30:50.330 on the thing that I observe

773 00:30:50.330 --> 00:30:52.860 or the proxy builds off the auxiliary variables

774 00:30:52.860 --> 00:30:56.120 and how much of it is depending on the underlying latent U

775 00:30:56.120 --> 00:30:57.020 related to Y,

776 00:30:57.020 --> 00:30:58.360 that is definitely not observed

777 00:30:58.360 --> 00:30:59.680 for the non-selected.

778 00:30:59.680 --> 00:31:00.513 Okay.

779 00:31:00.513 --> 00:31:01.650 And there's a little X star here,

780 00:31:01.650 --> 00:31:03.170 that's sort of a technical detail.

781 00:31:03.170 --> 00:31:04.800 We're rescaling the proxy.

782 00:31:04.800 --> 00:31:07.070 So it has the same variance as U,

783 00:31:07.070 --> 00:31:08.920 very unimportant mathematical detail.

784 00:31:10.090 --> 00:31:13.110 So we have this joint distribution

785 00:31:13.110 --> 00:31:15.570 that is conditional on selection status.

786 00:31:15.570 --> 00:31:18.860 And in addition to, we need that one assumption

787 00:31:18.860 --> 00:31:19.693 to identify things.

788 00:31:19.693 --> 00:31:21.840 We also have the latent variable problem.

789 00:31:21.840 --> 00:31:24.430 So latent variables do not have separately identifiable

790 00:31:24.430 --> 00:31:26.160 mean and variance, right?

791 00:31:26.160 --> 00:31:27.040 So that's just...

792 00:31:27.040 --> 00:31:28.649 Outside of the scope of this talk

793 00:31:28.649 --> 00:31:29.690 that's just a fact, right?

794 00:31:29.690 --> 00:31:31.020 So without loss of generality

795 00:31:31.020 --> 00:31:33.620 we're gonna set the variance of the latent variable

796 00:31:33.620 --> 00:31:35.350 for the select a sample equal to one.

797 00:31:35.350 --> 00:31:38.230 So it's just the scale of the latent variable.

798 00:31:38.230 --> 00:31:42.210 So what we actually care about is a function of you, right?

799 00:31:42.210 --> 00:31:44.590 It's the probability Y equals one marginally

800 00:31:44.590 --> 00:31:46.400 in my entire population.

801 00:31:46.400 --> 00:31:47.910 And so the probability Y equals one,

802 00:31:47.910 --> 00:31:49.930 is a probability U is greater than zero.

803 00:31:49.930 --> 00:31:51.340 That's that relationship.

804 00:31:51.340 --> 00:31:54.910 And so it's a weighted average of the proportion

805 00:31:54.910 --> 00:31:56.180 in the selected sample

806 00:31:56.180 --> 00:31:59.870 and the proportion in the non-selected sample, right?

807 00:31:59.870 --> 00:32:00.703 These are just...

808 00:32:00.703 --> 00:32:02.480 If U has this normal distribution

809 00:32:02.480 --> 00:32:03.900 this is how we get down to the probability

810 00:32:03.900 --> 00:32:04.900 U equals zero.

811 00:32:04.900 --> 00:32:06.523 Like those are those two pieces.

812 00:32:07.570 --> 00:32:09.780 So the key parameter that governs

813 00:32:09.780 --> 00:32:13.750 how this MUBP works is a correlation, right?

814 00:32:13.750 --> 00:32:16.810 It's the strength of the relationship between Y

815 00:32:16.810 --> 00:32:18.280 and your covariates.

816 00:32:18.280 --> 00:32:22.170 How good of a model do you have for Y, right?

817 00:32:22.170 --> 00:32:24.080 So remember we think back to that example

818 00:32:24.080 --> 00:32:26.440 of what if I had no biases Z.

819 00:32:26.440 --> 00:32:28.440 Or if Y wasn't related to Z,

820 00:32:28.440 --> 00:32:31.720 well, then who cares that there is no bias in Z.

821 00:32:31.720 --> 00:32:34.260 But we want there to be a strong relationship

822 00:32:34.260 --> 00:32:38.973 between Z and Y so that we can kind of infer from Z to Y.

823 00:32:39.820 --> 00:32:42.560 So that correlation in this latent variable framework

824 00:32:42.560 --> 00:32:45.750 is called the biserial correlation of the binary X

825 00:32:45.750 --> 00:32:46.920 and the continuous.

826 00:32:46.920 --> 00:32:49.839 I mean, sorry, the binary Y and the continuous X, right?

827 00:32:49.839 --> 00:32:52.650 There's lots of different flavors of correlation,

828 00:32:52.650 --> 00:32:54.890 biserial is the name for this one

829 00:32:54.890 --> 00:32:57.330 that's a binary Y and a continuous X

830 00:32:57.330 --> 00:33:00.130 when we're thinking about the latent variable framework.

831 00:33:00.130 --> 00:33:01.470 Importantly, you can estimate

832 00:33:01.470 --> 00:33:03.560 this in the selected sample, right?

833 00:33:03.560 --> 00:33:06.200 So I can estimate the correlation between you and X

834 00:33:06.200 --> 00:33:07.450 among the selected sample.

835 00:33:07.450 --> 00:33:08.800 I can't for the non-selected sample,

836 00:33:08.800 --> 00:33:11.700 of course, but I can for the selected sample.

837 00:33:11.700 --> 00:33:14.070 So the non-identifiable parameters

838 00:33:14.070 --> 00:33:15.483 of that pattern-mixture model, here they are.

839 00:33:15.483 --> 00:33:17.170 Like the mean for the latent variable,

840 00:33:17.170 --> 00:33:18.570 the variance for the latent variable

841 00:33:18.570 --> 00:33:21.740 and that correlation for the non-selected sample

842 00:33:21.740 --> 00:33:24.130 are in fact identified when we make this assumption

843 00:33:24.130 --> 00:33:26.330 on the selection mechanism.

844 00:33:26.330 --> 00:33:30.070 So let's think about some concrete scenarios.

845 00:33:30.070 --> 00:33:32.050 What if phi was zero?

846 00:33:32.050 --> 00:33:33.110 If phi is zero,

847 00:33:33.110 --> 00:33:35.340 we look up here at this part of the formula,

848 00:33:35.340 --> 00:33:37.610 well, then phi drops out it.

849 00:33:37.610 --> 00:33:40.300 So therefore selection only depends on X

850 00:33:40.300 --> 00:33:43.200 and those extra variables V that don't really matter

851 00:33:43.200 --> 00:33:45.690 because V isn't related to X or Y.

852 00:33:45.690 --> 00:33:49.700 This is an ignorable selection mechanism, okay.

853 00:33:49.700 --> 00:33:51.510 If on the other hand phi is one,

854 00:33:51.510 --> 00:33:53.500 well, then it entirely depends on U.

855 00:33:53.500 --> 00:33:55.070 X doesn't matter at all.

856 00:33:55.070 --> 00:33:57.590 This is your worst, worst, worst case scenario, right?

857 00:33:57.590 --> 00:34:00.090 Where whether or not you're in my sample only depends

858 00:34:00.090 --> 00:34:03.817 on U and therefore only depends on the value of Y.

859 00:34:03.817 --> 00:34:06.797 And so this is extremely not ignorable selection.

860 00:34:06.797 --> 00:34:09.510 And of course the truth is likely to lie

861 00:34:09.510 --> 00:34:11.210 somewhere in between, right?

862 00:34:11.210 --> 00:34:13.040 Some sort of non-ignorable mechanism,

863 00:34:13.040 --> 00:34:15.960 a phi between zero and one, so that U matters

864 00:34:15.960 --> 00:34:17.790 but it's not the only thing that matters.

865 00:34:17.790 --> 00:34:19.890 Right, that X matters as well.

866 00:34:19.890 --> 00:34:20.723 Okay.

867 00:34:20.723 --> 00:34:22.250 So this is a kind of moderate,

868 00:34:22.250 --> 00:34:23.410 non-ignorable selection.

869 00:34:23.410 --> 00:34:26.070 That's most likely the closest to reality

870 00:34:26.070 --> 00:34:28.263 with these non-probability samples.

871 00:34:30.120 --> 00:34:32.520 So for a specified value of phi.

872 00:34:32.520 --> 00:34:34.610 So we pick a value for our sensitivity parameter.

873 00:34:34.610 --> 00:34:36.230 There's no information in the data about it.

874 00:34:36.230 --> 00:34:40.340 We just pick it and we can actually estimate the mean of Y

875 00:34:40.340 --> 00:34:43.250 and compare that to the selected sample proportion.

876 00:34:43.250 --> 00:34:45.100 So we take this select a sample proportion,

877 00:34:45.100 --> 00:34:47.480 subtract what we get as the truth

878 00:34:47.480 --> 00:34:49.540 for that particular value of phi,

879 00:34:49.540 --> 00:34:51.610 and that's our measure of bias, right?

880 00:34:51.610 --> 00:34:54.110 So this second piece that's being subtracted

881 00:34:54.110 --> 00:34:54.943 here depends on phi.

882 00:34:54.943 --> 00:34:56.850 Right, it depends on what your value

883 00:34:56.850 --> 00:34:58.040 of your selected parameter is,

884 00:34:58.040 --> 00:35:00.860 or selection for your sensitivity parameter is.

885 00:35:00.860 --> 00:35:03.270 So in a nutshell, pick a selection mechanism

886 00:35:03.270 --> 00:35:05.500 by specifying specifying phi,

887 00:35:05.500 --> 00:35:07.270 estimate the overall proportion,

888 00:35:07.270 --> 00:35:10.057 and then subtract to get your measure of bias.

889 00:35:10.057 --> 00:35:12.060 And again, we don't know whether we're getting

890 00:35:12.060 --> 00:35:13.730 the right answer because it's depending

891 00:35:13.730 --> 00:35:15.170 on the sensitivity parameter

892 00:35:15.170 --> 00:35:18.670 but it's at least going to allow us to bound the problem.

893 00:35:18.670 --> 00:35:20.750 So the formula is quite messy,

894 00:35:20.750 --> 00:35:24.020 but it gives some insight into how this index works.

895 00:35:24.020 --> 00:35:26.660 So this measure of bias is the selected sample

896 00:35:26.660 --> 00:35:29.450 mean minus that estimator, right?

897 00:35:29.450 --> 00:35:31.760 This is the overall mean of Y

898 00:35:31.760 --> 00:35:33.910 based on those latent variables.

899 00:35:33.910 --> 00:35:35.560 And what gets plugged in here

900 00:35:35.560 --> 00:35:36.750 importantly for the mean

901 00:35:36.750 --> 00:35:39.030 and the variance for the non-selected cases

902 00:35:39.030 --> 00:35:42.030 depends on a component that I've got colored blue here,

903 00:35:42.030 --> 00:35:44.490 and a component that I've got color red.

904 00:35:44.490 --> 00:35:46.090 So if we look at the red piece

905 00:35:46.090 --> 00:35:48.930 this is the comparison of the proxy mean for the unselected

906 00:35:48.930 --> 00:35:50.450 and the selected cases.

907 00:35:50.450 --> 00:35:52.310 This is that bias in Z.

908 00:35:52.310 --> 00:35:54.120 The selection bias in Z,

909 00:35:54.120 --> 00:35:55.340 and it's just been standardized

910 00:35:55.340 --> 00:35:56.940 by its estimated variance, right?

911 00:35:56.940 --> 00:35:58.790 So that's how much selection bias

912 00:35:58.790 --> 00:36:01.510 was present in Z via X, right.

913 00:36:01.510 --> 00:36:04.800 Via using it to predict in the appropriate regression.

914 00:36:04.800 --> 00:36:07.850 Similarly, down here, how different is the variance

915 00:36:07.850 --> 00:36:10.400 of the selected and unselected cases for X.

916 00:36:10.400 --> 00:36:12.960 How much bias, selection bias is there in estimating

917 00:36:12.960 --> 00:36:14.160 the variance?

918 00:36:14.160 --> 00:36:16.270 So we're going to use that difference

919 00:36:16.270 --> 00:36:18.563 and scale the observed mean, right?

920 00:36:18.563 --> 00:36:21.530 There's that observed the estimated mean of U

921 00:36:21.530 --> 00:36:24.360 in the selected sample and how much it's gonna shift

922 00:36:24.360 --> 00:36:26.430 by is it depends on the selection,

923 00:36:26.430 --> 00:36:28.770 I mean, the sensitivity parameter phi,

924 00:36:28.770 --> 00:36:30.810 and also that by serial correlation.

925 00:36:30.810 --> 00:36:33.920 So this is why the by serial correlation is so important.

926 00:36:33.920 --> 00:36:36.810 It is gonna dominate how much of the bias

927 00:36:36.810 --> 00:36:39.543 in X we're going to transfer over into Y.

928 00:36:41.700 --> 00:36:44.090 So if phi were zero,

929 00:36:44.090 --> 00:36:45.470 so if we wanna assume

930 00:36:45.470 --> 00:36:47.690 that it is an ignorable selection mechanism,

931 00:36:47.690 --> 00:36:49.520 then this thing in blue here,

932 00:36:49.520 --> 00:36:52.300 think about plugging zero here, zero here, zero everywhere,

933 00:36:52.300 --> 00:36:54.500 is just gonna reduce down to that correlation.

934 00:36:54.500 --> 00:36:56.460 So we're gonna shift the mean of U

935 00:36:56.460 --> 00:36:58.900 for the non-selective cases

936 00:36:58.900 --> 00:37:03.020 based on the correlation times that difference in X.

937 00:37:03.020 --> 00:37:05.880 Whereas if we have phi equals one,

938 00:37:05.880 --> 00:37:09.403 this thing in blue turns into one over the correlation.

939 00:37:10.350 --> 00:37:12.070 So here is where thinking about the magnitude

940 00:37:12.070 --> 00:37:13.330 of the correlation helps.

941 00:37:13.330 --> 00:37:15.227 If the correlation is really big, right?

942 00:37:15.227 --> 00:37:17.270 If the correlation is 0.8, 0.9,

943 00:37:17.270 --> 00:37:19.850 something really large than phi and...

944 00:37:19.850 --> 00:37:22.060 I mean, sorry, then rho and one over rho

945 00:37:22.060 --> 00:37:23.423 are very close, right?

946 00:37:23.423 --> 00:37:25.940 0.8 and 1/0.8 are pretty close.

947 00:37:25.940 --> 00:37:28.710 So if we're thinking about bounding this between phi

948 00:37:28.710 --> 00:37:30.160 equals zero and equals one,

949 00:37:30.160 --> 00:37:32.580 our interval is gonna be relatively small.

950 00:37:32.580 --> 00:37:34.620 But if the correlation is small,

951 00:37:34.620 --> 00:37:37.200 the correlation were 0.2, oh, oh, right?

952 00:37:37.200 --> 00:37:38.700 We're gonna get a really big interval

953 00:37:38.700 --> 00:37:40.100 because that correlation,

954 00:37:40.100 --> 00:37:42.770 we're gonna shift with the factor of multiplied by 0.2

955 00:37:42.770 --> 00:37:44.260 but then one over 0.2.

956 00:37:44.260 --> 00:37:46.200 That's gonna be a really big shift

957 00:37:46.200 --> 00:37:48.200 in that mean of the latent variable U

958 00:37:48.200 --> 00:37:49.843 and therefore the mean of Y.

959 00:37:51.290 --> 00:37:52.760 So how do we get these estimates?

960 00:37:52.760 --> 00:37:54.900 We have two possibilities. We can use what we call

961 00:37:54.900 --> 00:37:57.540 modified maximum likelihood estimation.

962 00:37:57.540 --> 00:37:58.373 It's not true.

963 00:37:58.373 --> 00:38:00.080 Maximum likelihood because we estimate

964 00:38:00.080 --> 00:38:01.960 the biserial correlation with something

965 00:38:01.960 --> 00:38:03.840 called a two step method, right?

966 00:38:03.840 --> 00:38:07.180 So instead of doing a full, maximum likelihood,

967 00:38:07.180 --> 00:38:11.590 we kind of take this cheat in which we set that mean of X

968 00:38:11.590 --> 00:38:14.520 for the selected cases equal to what we observe,

969 00:38:14.520 --> 00:38:16.070 And then conditional not to estimate

970 00:38:16.070 --> 00:38:17.800 the by serial correlation.

971 00:38:17.800 --> 00:38:18.670 Yeah.

972 00:38:18.670 --> 00:38:21.920 And as a sensitivity analysis we would plug in zero one

973 00:38:21.920 --> 00:38:23.410 and maybe 0.5 in the middle

974 00:38:23.410 --> 00:38:25.313 as the values sensitivity parameter.

975 00:38:26.160 --> 00:38:28.840 Alternatively, and we feel is a much more attractive

976 00:38:28.840 --> 00:38:30.810 approach is to be Bayesian about this.

977 00:38:30.810 --> 00:38:34.120 So in this MML estimation,

978 00:38:34.120 --> 00:38:37.560 we are implicitly assuming that we know the betas

979 00:38:37.560 --> 00:38:38.680 from that probate regression.

980 00:38:38.680 --> 00:38:42.480 That we're essentially treating X like we know it.

981 00:38:42.480 --> 00:38:43.770 But we don't know X, right?

982 00:38:43.770 --> 00:38:44.820 That probate regression,

983 00:38:44.820 --> 00:38:47.240 those parameters have error associated with them.

984 00:38:47.240 --> 00:38:48.086 Right?

985 00:38:48.086 --> 00:38:49.430 And you can imagine that the bigger your selected sample,

986 00:38:49.430 --> 00:38:51.490 the more precisely estimating those betas,

987 00:38:51.490 --> 00:38:52.900 that's not being reflected

988 00:38:52.900 --> 00:38:55.880 at all in the modified maximum likelihood.

989 00:38:55.880 --> 00:38:57.420 So instead we can be Bayesian.

990 00:38:57.420 --> 00:39:00.520 Put non-informative priors on all the identified parameters.

991 00:39:00.520 --> 00:39:01.920 That's gonna let those,

992 00:39:01.920 --> 00:39:04.640 the error in those betas be propagated.

993 00:39:04.640 --> 00:39:07.430 And so we'll incorporate that uncertainty.

994 00:39:07.430 --> 00:39:11.160 And we can actually, additionally put a prior on phi, right?

995 00:39:11.160 --> 00:39:11.993 So we could just say

996 00:39:11.993 --> 00:39:14.300 let's have it be uniform across zero one.

997 00:39:14.300 --> 00:39:15.133 Right?

998 00:39:15.133 --> 00:39:17.540 So we can see what does it look like if we in totality,

999 00:39:17.540 --> 00:39:20.360 if we assume that phi is somewhere evenly distributed

1000 00:39:20.360 --> 00:39:21.610 across that interval.

1001 00:39:21.610 --> 00:39:22.870 We've done other things as well.

1002 00:39:22.870 --> 00:39:25.860 We've taken like discreet priors.

1003 00:39:25.860 --> 00:39:28.960 Oh, let's put a point mass on 0.5 and one

1004 00:39:28.960 --> 00:39:29.940 or other different, right?

1005 00:39:29.940 --> 00:39:31.883 You can do whatever you want for that prior.

1006 00:39:32.880 --> 00:39:34.560 So let's go back to the example

1007 00:39:34.560 --> 00:39:36.090 see what it looks like.

1008 00:39:36.090 --> 00:39:38.300 If we have the proportion ever married for females

1009 00:39:38.300 --> 00:39:40.340 on the left and males on the right,

1010 00:39:40.340 --> 00:39:42.950 the true bias is the black dot.

1011 00:39:42.950 --> 00:39:45.070 And so the black is the true bias.

1012 00:39:45.070 --> 00:39:49.540 The little tiny diamond is the MUBP for 0.5.

1013 00:39:49.540 --> 00:39:52.030 An so that's plugging in that average value.

1014 00:39:52.030 --> 00:39:55.780 Some selection mechanism that depends on why some,

1015 00:39:55.780 --> 00:39:56.850 somewhere in the middle.

1016 00:39:56.850 --> 00:39:57.993 So we're actually coming pretty close.

1017 00:39:57.993 --> 00:40:00.210 That happens to be, that's pretty close.

1018 00:40:00.210 --> 00:40:01.750 And the intervals in green

1019 00:40:01.750 --> 00:40:04.040 are the modified maximum likelihood intervals

1020 00:40:04.040 --> 00:40:06.120 from phi equals zero to phi equals one,

1021 00:40:06.120 --> 00:40:08.240 and the Bayesian intervals are longer, right?

1022 00:40:08.240 --> 00:40:09.073 Naturally.

1023 00:40:09.073 --> 00:40:10.840 We're incorporating the uncertainty.

1024 00:40:10.840 --> 00:40:12.920 Essentially these MUBP,

1025 00:40:12.920 --> 00:40:14.767 modified maximum likely intervals are too short.

1026 00:40:14.767 --> 00:40:17.103 And we admit that these are too short.

1027 00:40:18.350 --> 00:40:21.300 If we plug in all zeros and all ones

1028 00:40:21.300 --> 00:40:25.380 for that small proportion of my NSFG population

1029 00:40:25.380 --> 00:40:27.310 that we aren't selected into the sample,

1030 00:40:27.310 --> 00:40:31.160 we get huge bounds relative to our indicator.

1031 00:40:31.160 --> 00:40:31.993 Right?

1032 00:40:31.993 --> 00:40:33.560 So remember when I showed you that slide, that bounded,

1033 00:40:33.560 --> 00:40:36.810 we know the bias has to be between these two values.

1034 00:40:36.810 --> 00:40:37.790 That's what's going on here.

1035 00:40:37.790 --> 00:40:39.320 That's what these two values are.

1036 00:40:39.320 --> 00:40:41.480 But using the information in Z

1037 00:40:41.480 --> 00:40:43.260 we're able to much, much narrow

1038 00:40:43.260 --> 00:40:45.780 or make an estimate on where our selection bias is.

1039 00:40:45.780 --> 00:40:47.670 So we got much tighter bounds.

1040 00:40:47.670 --> 00:40:48.503 An important fact here

1041 00:40:48.503 --> 00:40:50.420 is that we have pretty good predictors.

1042 00:40:50.420 --> 00:40:52.620 Our correlation, the biserial correlation

1043 00:40:52.620 --> 00:40:54.360 is about 0.7 or 0.8.

1044 00:40:54.360 --> 00:40:55.850 So these things are pretty correlated

1045 00:40:55.850 --> 00:40:58.650 with whether you've been married, age, education, right?

1046 00:40:58.650 --> 00:41:00.400 Those things are pretty correlated.

1047 00:41:01.310 --> 00:41:04.370 Another variable in the NSFG is income.

1048 00:41:04.370 --> 00:41:07.890 So we can think about an indicator for having low income.

1049 00:41:07.890 --> 00:41:10.130 Well, as it turns out those variables

1050 00:41:10.130 --> 00:41:13.810 we have on everybody; age, education, gender,

1051 00:41:13.810 --> 00:41:16.150 those things are not actually that good of predictors,

1052 00:41:16.150 --> 00:41:18.720 of low income, very low correlation.

1053 00:41:18.720 --> 00:41:21.040 So our index reflects that.

1054 00:41:21.040 --> 00:41:23.380 Or you get much, Y, your intervals.

1055 00:41:23.380 --> 00:41:25.940 Sort of closer to the Manski bounds.

1056 00:41:25.940 --> 00:41:28.770 And in fact, it's exactly returning one of those bounds.

1057 00:41:28.770 --> 00:41:32.930 The filling in all zeros bound is returned by this index.

1058 00:41:32.930 --> 00:41:34.750 So that's actually an attractive feature.

1059 00:41:34.750 --> 00:41:35.583 Right?

1060 00:41:35.583 --> 00:41:37.810 We're sort of bounded at the worst possible case

1061 00:41:37.810 --> 00:41:39.410 on one end of the bias

1062 00:41:40.496 --> 00:41:42.260 but we are still capturing the truth.

1063 00:41:42.260 --> 00:41:44.150 The Manski bounds are basically useless,

1064 00:41:44.150 --> 00:41:45.650 right in this particular case.

1065 00:41:47.210 --> 00:41:50.278 So that's a toy example.

1066 00:41:50.278 --> 00:41:53.060 Just gonna quickly show you a real example,

1067 00:41:53.060 --> 00:41:54.010 and I'm actually gonna to skip

1068 00:41:54.010 --> 00:41:55.190 over the incentive experiment,

1069 00:41:55.190 --> 00:41:57.070 which well, very, very interesting

1070 00:41:57.070 --> 00:41:59.160 is there's a lot to talk about,

1071 00:41:59.160 --> 00:42:01.943 and I'd rather jump straight to the presidential polls.

1072 00:42:03.210 --> 00:42:07.633 So there's very much in the news now,

1073 00:42:07.633 --> 00:42:08.466 and over the past several years,

1074 00:42:08.466 --> 00:42:10.900 this idea of failure of political polling

1075 00:42:10.900 --> 00:42:12.417 and this recent high profile failure

1076 00:42:12.417 --> 00:42:14.930 of pre-election polls in the US.

1077 00:42:14.930 --> 00:42:17.500 So polls are probability samples

1078 00:42:17.500 --> 00:42:20.035 but they have very, very, very low response rates.

1079 00:42:20.035 --> 00:42:21.100 I don't know how much you know about how they're done,

1080 00:42:21.100 --> 00:42:23.100 but they're very, very low response rate.

1081 00:42:23.100 --> 00:42:25.230 But think about what we're getting at in a poll,

1082 00:42:25.230 --> 00:42:28.450 a binary variable, are you going to vote for Donald Trump?

1083 00:42:28.450 --> 00:42:29.283 Yes or no?

1084 00:42:29.283 --> 00:42:30.520 Are you gonna vote for Joe Biden?

1085 00:42:30.520 --> 00:42:31.353 Yes or no?

1086 00:42:31.353 --> 00:42:32.186 These binary variables.

1087 00:42:32.186 --> 00:42:33.750 We want to estimate proportions.

1088 00:42:33.750 --> 00:42:35.550 That's what political polls aimed to do.

1089 00:42:35.550 --> 00:42:37.350 Pre-election polls.

1090 00:42:37.350 --> 00:42:40.620 So we have these political polls with these failures.

1091 00:42:40.620 --> 00:42:43.580 So we're thinking, maybe it's a selection bias problem.

1092 00:42:43.580 --> 00:42:45.390 And that there is some of this people

1093 00:42:45.390 --> 00:42:49.210 are entering into this poll differentially,

1094 00:42:49.210 --> 00:42:51.730 depending on who they're going to vote for.

1095 00:42:51.730 --> 00:42:52.760 So think of it this way,

1096 00:42:52.760 --> 00:42:54.130 and I'm gonna use Trump as the example

1097 00:42:54.130 --> 00:42:55.320 'cause we're going to estimate,

1098 00:42:55.320 --> 00:42:56.153 I'm gonna try to estimate

1099 00:42:56.153 --> 00:42:57.498 the proportion of people who will vote

1100 00:42:57.498 --> 00:43:01.900 for Former President Trump in the 2020 election.

1101 00:43:01.900 --> 00:43:04.320 So, might Trump supporters

1102 00:43:04.320 --> 00:43:07.120 just inherently be less likely to answer the call, right?

1103 00:43:07.120 --> 00:43:10.760 To answer that poll or to refuse to answer the question

1104 00:43:10.760 --> 00:43:13.440 even conditional demographic characteristics, right?

1105 00:43:13.440 --> 00:43:15.900 So two people who otherwise look the same

1106 00:43:15.900 --> 00:43:19.730 with respect to those Z variables, age, race, education,

1107 00:43:19.730 --> 00:43:22.160 the one who's the Trump supporter, someone might argue, 1108 00:43:22.160 --> 00:43:24.260 you might be more suspicious of the government

1109 00:43:24.260 --> 00:43:25.820 and the polls, and not want to answer

1110 00:43:25.820 --> 00:43:28.460 and not come into this poll, not be selected.

1111 00:43:28.460 --> 00:43:30.910 As it would be depending on why.

1112 00:43:30.910 --> 00:43:35.240 So the MUBP could be used to try to adjust poll estimates.

1113 00:43:35.240 --> 00:43:37.810 Say, well, there's your estimate from the poll

1114 00:43:37.810 --> 00:43:40.200 but what if selection were not ignorable?

1115 00:43:40.200 --> 00:43:41.690 How different would our estimate

1116 00:43:41.690 --> 00:43:43.440 of the proportion voting for Trump?

1117 00:43:44.700 --> 00:43:47.790 So in this example, our proportion of interest

1118 00:43:47.790 --> 00:43:51.300 is the percent of people who are gonna vote for Trump.

1119 00:43:51.300 --> 00:43:52.950 The sample that we used

1120 00:43:52.950 --> 00:43:54.420 are publicly available data

1121 00:43:54.420 --> 00:43:56.390 from seven different pre-election polls

1122 00:43:56.390 --> 00:44:00.530 conducted in seven different states by ABC in 2020.

1123 00:44:00.530 --> 00:44:02.760 And the way these polls work

1124 00:44:02.760 --> 00:44:04.830 is it's a random digit dialing survey.

1125 00:44:04.830 --> 00:44:07.770 So that's literally randomly dialing phone numbers.

1126 00:44:07.770 --> 00:44:08.650 Many of whom get

1127 00:44:08.650 --> 00:44:10.340 throughout 'cause their business, et cetera,

1128 00:44:10.340 --> 00:44:12.960 very, very low response rates, 10% or lower.

1129 00:44:12.960 --> 00:44:16.810 Very, very, very low response rates to these kinds of polls.

1130 00:44:16.810 --> 00:44:19.290 They do, however, try to do some weighting.

1131 00:44:19.290 --> 00:44:20.810 So it's not as if they just take that sample and say,

1132 00:44:20.810 --> 00:44:23.490 there we go let's estimate the proportion for Trump.

1133 00:44:23.490 --> 00:44:24.730 We do waiting adjustments

1134 00:44:24.730 --> 00:44:28.300 and they use what's called inter proportional fitting

1135 00:44:28.300 --> 00:44:32.820 or raking to get the distribution of key variables

1136 00:44:32.820 --> 00:44:35.660 in the sample to look like the population.

1137 00:44:35.660 --> 00:44:37.620 So they use census margins for, again,

1138 00:44:37.620 --> 00:44:40.460 it's gender as binary, unfortunately,

1139 00:44:40.460 --> 00:44:43.913 age, education, race, ethnicity, and party identification.

1140 00:44:44.800 --> 00:44:46.870 So, because we're doing this after the election

1141 00:44:46.870 --> 00:44:47.730 we know the truth.

1142 00:44:47.730 --> 00:44:50.250 We have access to the true official election outcomes

1143 00:44:50.250 --> 00:44:51.210 in each state.

1144 00:44:51.210 --> 00:44:53.780 So I know the actual proportion of why.

1145 00:44:53.780 --> 00:44:56.590 And my population is likely voters,

1146 00:44:56.590 --> 00:44:58.460 because that's who we're trying to target

1147 00:44:58.460 --> 00:44:59.427 with these pre-election polls.

1148 00:44:59.427 --> 00:45:02.290 You wanna know what's the estimated proportion

1149 00:45:02.290 --> 00:45:04.950 would vote for Trump among the likely voters.

1150 00:45:04.950 --> 00:45:07.000 So the tricky thing is that population

1151 00:45:07.000 --> 00:45:09.930 is hard to come by summary statistics.

1152 00:45:09.930 --> 00:45:11.170 Likely voters, right?

1153 00:45:11.170 --> 00:45:13.440 It's easy to get summary statistics from all people

1154 00:45:13.440 --> 00:45:16.030 in the US or all people of voting age in the US

1155 00:45:16.030 --> 00:45:17.467 but not likely voters.

1156 00:45:18.380 --> 00:45:21.340 So here Y is an indicator for voting for Trump.

1157 00:45:21.340 --> 00:45:24.310 Z is auxiliary variable in the ABC poll.

1158 00:45:24.310 --> 00:45:25.410 So all those variables I mentioned

1159 00:45:25.410 --> 00:45:27.480 before gender age, et cetera.

1160 00:45:27.480 --> 00:45:29.270 We actually have very strong predictors

1161 00:45:29.270 --> 00:45:32.260 of why basically because of these political ideation,

1162 00:45:32.260 --> 00:45:33.980 party identification variables, right?

1163 00:45:33.980 --> 00:45:36.820 Not surprisingly the people who identify as Democrats,

1164 00:45:36.820 --> 00:45:39.263 very unlikely to be voting for Trump.

1165 00:45:40.670 --> 00:45:44.080 The data set that we found that can give us population level

1166 00:45:44.080 --> 00:45:47.630 estimates of the mean of Z for the non-selected sample

1167 00:45:47.630 --> 00:45:49.890 is a dataset from AP/NORC.

1168 00:45:49.890 --> 00:45:51.700 It's called their VoteCast Data.

1169 00:45:51.700 --> 00:45:54.690 And they conduct these large surveys

1170 00:45:54.690 --> 00:45:57.770 and provide an indicator of likely voter.

1171 00:45:57.770 --> 00:46:00.370 So we can basically use this dataset

1172 00:46:00.370 --> 00:46:02.280 to describe the demographic characteristics

1173 00:46:02.280 --> 00:46:03.520 of likely voters,

1174 00:46:03.520 --> 00:46:07.503 instead of just all people who are 18 and older in the US.

1175 00:46:08.520 --> 00:46:10.260 The subtle issue is of course,

1176 00:46:10.260 --> 00:46:12.530 these AP VoteCast data are not without error,

1177 00:46:12.530 --> 00:46:15.070 but we're going to pretend that they are without error.

1178 00:46:15.070 --> 00:46:16.530 And that's like a whole other papers.

1179 00:46:16.530 --> 00:46:17.363 How do we handle the fact

1180 00:46:17.363 --> 00:46:19.350 that my population data have error?

1181 00:46:19.350 --> 00:46:22.610 So we're gonna use the unweighted ABC poll data

1182 00:46:22.610 --> 00:46:25.530 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 small.

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 small,

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, Minnesota,

1200 00:46:59.270 --> 00:47:01.550 North Carolina, Pennsylvania, and Wisconsin.

1201 00:47:01.550 --> 00:47:05.960 So this horizontal line here at zero that's no bias, right?

1202 00:47:05.960 --> 00:47:08.140 So it's estimated, the ABC poll estimate

1203 00:47:08.140 --> 00:47:09.490 would have no bias.

1204 00:47:09.490 --> 00:47:12.920 And we can see then in Arizona where sort of overestimated

1205 00:47:12.920 --> 00:47:14.060 and in the rest of the states

1206 00:47:14.060 --> 00:47:16.277 we've got underestimated the support for Trump.

1207 00:47:16.277 --> 00:47:19.140 And so that was really the failure was the underestimation

1208 00:47:19.140 --> 00:47:20.290 of the support for Trump.

1209 00:47:20.290 --> 00:47:23.880 Notice that our Bayesian bounds

1210 00:47:23.880 --> 00:47:26.230 cover the true bias everywhere except

1211 00:47:26.230 --> 00:47:27.920 in Pennsylvania and Wisconsin.

1212 00:47:27.920 --> 00:47:30.430 And so Wisconsin had an enormous bias,

1213 00:47:30.430 --> 00:47:32.570 or they way under called the support for Trump

1214 00:47:32.570 --> 00:47:34.470 in Wisconsin by 10 percentage points.

1215 00:47:34.470 --> 00:47:35.410 Huge problem.

1216 00:47:35.410 --> 00:47:36.850 So we're not getting there

1217 00:47:36.850 --> 00:47:39.880 but notice that zero is not in our interval.

1218 00:47:39.880 --> 00:47:42.760 So our bounds are suggesting

1219 00:47:42.760 --> 00:47:45.530 that there was a negative bias from the poll.

1220 00:47:45.530 --> 00:47:47.660 So even though we didn't capture the truth,

1221 00:47:47.660 --> 00:47:49.260 we've at least crossed the threshold

1222 00:47:49.260 --> 00:47:52.360 saying very likely that you are under calling

1223 00:47:52.360 --> 00:47:54.023 the support for Trump.

1224 00:47:55.280 --> 00:47:59.200 So how do estimates using the MUBP compared to the ABC poll?

1225 00:47:59.200 --> 00:48:02.830 Well, we can use the MUBP bounds to basically shift

1226 00:48:02.830 --> 00:48:04.570 the ABC poll estimates.

1227 00:48:04.570 --> 00:48:07.740 So we're calling those MUBP adjusted, right?

1228 00:48:07.740 --> 00:48:09.850 So we've got the truth is...

1229 00:48:09.850 --> 00:48:11.590 The true proportion who voted for Trump

1230 00:48:11.590 --> 00:48:14.360 are now these red triangles

1231 00:48:14.360 --> 00:48:17.290 and then the black circles are the point estimates

1232 00:48:17.290 --> 00:48:19.810 from three different methods of estimation,

1233 00:48:19.810 --> 00:48:21.450 of obtaining an estimate.

1234 00:48:21.450 --> 00:48:24.720 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.

1237 00:48:30.340 --> 00:48:32.330 Is MUBP adjusted.

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

1260 00:49:24.660 --> 00:49:26.900 we would we captured that.

1261 00:49:26.900 --> 00:49:29.060 So the important idea here is, you know,

1262 00:49:29.060 --> 00:49:33.630 before the election, we wouldn't have these red triangles.

1263 00:49:33.630 --> 00:49:35.620 But it's important to be able to see

1264 00:49:35.620 --> 00:49:38.600 that this is saying you're under calling

1265 00:49:38.600 --> 00:49:39.870 the support for Trump

1266 00:49:39.870 --> 00:49:42.330 if there were a non-negligible selection, right?

1267 00:49:42.330 --> 00:49:44.070 So it's that idea of a sensitivity analysis?

1268 00:49:44.070 --> 00:49:46.130 How bad would we be doing?

1269 00:49:46.130 --> 00:49:48.780 And what we would say is in Minnesota and Wisconsin

1270 00:49:48.780 --> 00:49:49.960 we'd be very worried

1271 00:49:49.960 --> 00:49:53.083 about under calling the support for Trump.

1272 00:49:56.280 --> 00:49:59.370 So what have I just shown you?

1273 00:49:59.370 --> 00:50:00.530 I'll summarize.

1274 00:50:00.530 --> 00:50:03.900 The MUBP is a sensitivity analysis tool

1275 00:50:03.900 --> 00:50:07.780 to assess the potential for non-ignorable selection bias.

1276 00:50:07.780 --> 00:50:11.640 If we have a phi equals zero, an ignorable selection,

1277 00:50:11.640 --> 00:50:14.110 we can adjust that away via weighting

1278 00:50:14.110 --> 00:50:15.850 or some other method, right?

1279 00:50:15.850 --> 00:50:18.140 So if it's not ignorable, I mean, if it is ignorable

1280 00:50:18.140 --> 00:50:20.530 we can ignore the selection mechanism.

1281 00:50:20.530 --> 00:50:22.750 On the other extreme if phi is one,

1282 00:50:22.750 --> 00:50:23.970 totally not ignorable,

1283 00:50:23.970 --> 00:50:26.030 selection is only depending on that outcome

1284 00:50:26.030 --> 00:50:27.560 we're trying to measure.

1285 00:50:27.560 --> 00:50:29.550 Somewhere in between we've got the 0.5.

1286 00:50:29.550 --> 00:50:31.970 That if you really needed a point estimate

1287 00:50:31.970 --> 00:50:33.610 of the bias, that would be 0.5.

1288 00:50:33.610 --> 00:50:36.630 And in fact, that's what this black dot is.

1289 00:50:36.630 --> 00:50:40.003 That's the adjustment at 0.5 for our adjusted estimator.

1290 00:50:41.420 --> 00:50:45.210 This MUBP is tailored to binary outcomes,

1291 00:50:45.210 --> 00:50:47.923 and it is an improvement over the normal base SMUB.

1292 00:50:47.923 --> 00:50:48.980 I didn't show you the,

1293 00:50:48.980 --> 00:50:51.930 so the results from simulations that basically show

1294 00:50:51.930 --> 00:50:54.550 if you use the normal method on a binary outcome

1295 00:50:54.550 --> 00:50:56.020 you get these huge bounds.

1296 00:50:56.020 --> 00:50:58.180 You go outside of the Manski bounds, right?

1297 00:50:58.180 --> 00:51:01.010 'Cause it's not properly bounded between zero and one,

1298 00:51:01.010 --> 00:51:03.300 or your proportion isn't properly bounded.

1299 00:51:03.300 --> 00:51:05.910 And importantly, our measure only requires

1300 00:51:05.910 --> 00:51:08.140 summary statistics for Z,

1301 00:51:08.140 --> 00:51:11.160 for the population or for the non-selected sample.

1302 00:51:11.160 --> 00:51:13.750 So I don't have to have a whole separate data set

1303 00:51:13.750 --> 00:51:15.660 where I have everybody who didn't get selected

1304 00:51:15.660 --> 00:51:16.493 into my sample,

1305 00:51:16.493 --> 00:51:19.703 I just need to know the average of these co-variants, right.

1306 00:51:19.703 --> 00:51:23.380 I just needs to know Z-bar in order to get my average

1307 00:51:23.380 --> 00:51:25.580 proxy for the non-selected.

1308 00:51:25.580 --> 00:51:27.100 With weak information,

1309 00:51:27.100 --> 00:51:30.410 so if my model is poor then my Manski bounds

1310 00:51:30.410 --> 00:51:31.560 are gonna be what's returned.

1311 00:51:31.560 --> 00:51:34.200 So that's a good feature of this index.

1312 00:51:34.200 --> 00:51:35.670 Is that it is naturally bound

1313 00:51:35.670 --> 00:51:38.000 unlike the normal model version.

1314 00:51:38.000 --> 00:51:41.020 And we have done additional work to move

1315 00:51:41.020 --> 00:51:43.140 beyond just estimating means and proportions

1316 00:51:43.140 --> 00:51:45.950 into linear regression and probate progression.

1317 00:51:45.950 --> 00:51:48.360 So we've have indices of selection bias

1318 00:51:48.360 --> 00:51:49.630 for regression coefficients.

1319 00:51:49.630 --> 00:51:52.780 So instead of wanting to know the mean of Y

1320 00:51:52.780 --> 00:51:54.900 or the proportion with Y equals one,

1321 00:51:54.900 --> 00:51:57.210 what if you wanted to do a regression of Y

1322 00:51:57.210 --> 00:51:58.700 on some covariates?

1323 00:51:58.700 --> 00:52:01.590 So we have a paper out in the animals of applied statistics

1324 00:52:01.590 --> 00:52:04.750 that extends those two regression coefficients.

1325 00:52:04.750 --> 00:52:06.740 So I believe I'm pretty much right on the time

1326 00:52:06.740 --> 00:52:09.240 I was supposed to end, so I'll say Thank you everyone.

1327 00:52:09.240 --> 00:52:11.170 And I'm happy to take questions.

1328 00:52:11.170 --> 00:52:12.250 I'll put on my references

1329 00:52:12.250 --> 00:52:15.423 of my meeny, miny fonts, yes.

1330 00:52:19.810 --> 00:52:21.960 Robert Does anybody have any questions?

1331 00:52:25.610 --> 00:52:26.443 From the room?

1332 00:52:33.498 --> 00:52:34.331 So.

1333 00:52:36.340 --> 00:52:37.820 Dr. Rebecca Let me stop my share.

1334 00:52:37.820 --> 00:52:38.653 Student Hey.

1335 00:52:39.630 --> 00:52:41.360 I have a very basic one,

1336 00:52:41.360 --> 00:52:43.740 mostly more of curiosity (indistinct)

1337 00:52:43.740 --> 00:52:45.360 Sure, sure.

1338 00:52:45.360 --> 00:52:47.260 What is it that caused the...

1339 00:52:49.970 --> 00:52:53.710 We know after the fact that in your example

1340 00:52:53.710 --> 00:52:56.907 that there was the direction of the bias,

1341 00:52:56.907 --> 00:53:01.907 but why is it that it only shifted in the Trump direction?

1342 00:53:02.570 --> 00:53:03.403 Why?

1343 00:53:03.403 --> 00:53:05.520 You don't know in advance if something is more likely

1344 00:53:05.520 --> 00:53:06.353 or less likely?

1345 00:53:07.831 --> 00:53:08.664 Okay.

1346 00:53:08.664 --> 00:53:09.497 So excellent question.

1347 00:53:09.497 --> 00:53:11.330 So that is effectively,

1348 00:53:11.330 --> 00:53:14.750 the direction of the shift is going to match...

1349 00:53:14.750 --> 00:53:16.673 The direction of the shift in the mean of Y,

1350 00:53:16.673 --> 00:53:18.410 when the proportion is going to match

1351 00:53:18.410 --> 00:53:20.250 the shift in X, right?

1352 00:53:20.250 --> 00:53:25.080 So if what you get as your mean for your proxy,

1353 00:53:25.080 --> 00:53:28.440 for the non-selected sample is bigger

1354 00:53:28.440 --> 00:53:29.760 than for your selected sample

1355 00:53:29.760 --> 00:53:31.100 then your proportion is gonna get shifted

1356 00:53:31.100 --> 00:53:32.130 in that direction?

1357 00:53:32.130 --> 00:53:32.963 Right.

1358 00:53:32.963 --> 00:53:36.660 It's only ever going to shift it to match the bias in X.

1359 00:53:36.660 --> 00:53:37.493 Right?

1360 00:53:37.493 --> 00:53:38.910 And so then, which way that shifts Y

1361 00:53:38.910 --> 00:53:40.530 depends on what the relationship

1362 00:53:40.530 --> 00:53:45.530 is between the covariates Z and X in the probate regression.

1363 00:53:45.610 --> 00:53:49.380 But it will always shift it in a particular direction.

1364 00:53:49.380 --> 00:53:51.980 I will notice that I fully admit,

1365 00:53:51.980 --> 00:53:54.990 our index actually shifted the wrong direction

1366 00:53:54.990 --> 00:53:56.520 in one particular case.

1367 00:53:56.520 --> 00:53:57.353 Right?

1368 00:53:57.353 --> 00:53:58.823 So actually in Florida,

1369 00:54:00.165 --> 00:54:02.170 we actually shifted down when we shouldn't.

1370 00:54:02.170 --> 00:54:03.003 Right.

1371 00:54:03.003 --> 00:54:05.270 So here's the way to estimate and we're shifting down,

1372 00:54:05.270 --> 00:54:06.790 but actually the truth is higher.

1373 00:54:06.790 --> 00:54:08.810 So we're not always getting it right

1374 00:54:08.810 --> 00:54:12.500 we're getting it right when that X is shifting

1375 00:54:12.500 --> 00:54:13.710 in the correct direction.

1376 00:54:13.710 --> 00:54:14.543 Right?

1377 00:54:14.543 --> 00:54:16.750 So it isn't true that we always...

1378 00:54:16.750 --> 00:54:19.080 It's true that it always shifts the direction of X,

1379 00:54:19.080 --> 00:54:21.540 but it's not a hundred percent true that X

1380 00:54:21.540 --> 00:54:23.740 always shifts in the exact same way as Y.

1381 00:54:23.740 --> 00:54:25.030 Just most of the time.

1382 00:54:25.030 --> 00:54:28.950 There was evidence of underestimating the Trump support,

1383 00:54:28.950 --> 00:54:31.600 and that was in fact reflected in that probate regression,

1384 00:54:31.600 --> 00:54:33.150 right in that relationship.

1385 00:54:33.150 --> 00:54:36.320 The people who replied to the poll were older,

1386 00:54:36.320 --> 00:54:38.860 they were higher educated, right?

1387 00:54:38.860 --> 00:54:39.780 And so those older,

1388 00:54:39.780 --> 00:54:42.660 higher educated people in aggregate

1389 00:54:42.660 --> 00:54:45.080 were less likely to vote for Trump.

1390 00:54:45.080 --> 00:54:47.740 So that's why we ended up under calling the support

1391 00:54:47.740 --> 00:54:49.290 for Trump when we don’t account

1392 00:54:49.290 --> 00:54:52.480 for that potential non-ignorable selection bias.

1393 00:54:52.480 --> 00:54:53.637 Good question though.

1394 00:54:54.520 --> 00:54:56.400 Robert Go it, Thank you.

1395 00:54:56.400 --> 00:54:59.460 Any other questions (indistinct)

1396 00:55:09.360 --> 00:55:10.193 Anybody?

1397 00:55:15.900 --> 00:55:18.750 I know I talk fast and that was a lot of stuff

1398 00:55:18.750 --> 00:55:21.093 so you know, like get it.

1399 00:55:21.093 --> 00:55:23.070 (indistinct)

1400 00:55:23.070 --> 00:55:23.903 Alright.

1401 00:55:23.903 --> 00:55:25.800 Well, Andridge, Thank you again.

1402 00:55:25.800 --> 00:55:26.882 And.

1403 00:55:26.882 --> 00:55:29.882 (students clapping)

1404 00:55:32.950 --> 00:55:33.783 Thank you.

1405 00:55:33.783 --> 00:55:34.960 Thank you for having me.

1406 00:55:34.960 --> 00:55:35.793 Robert Yeah.