THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

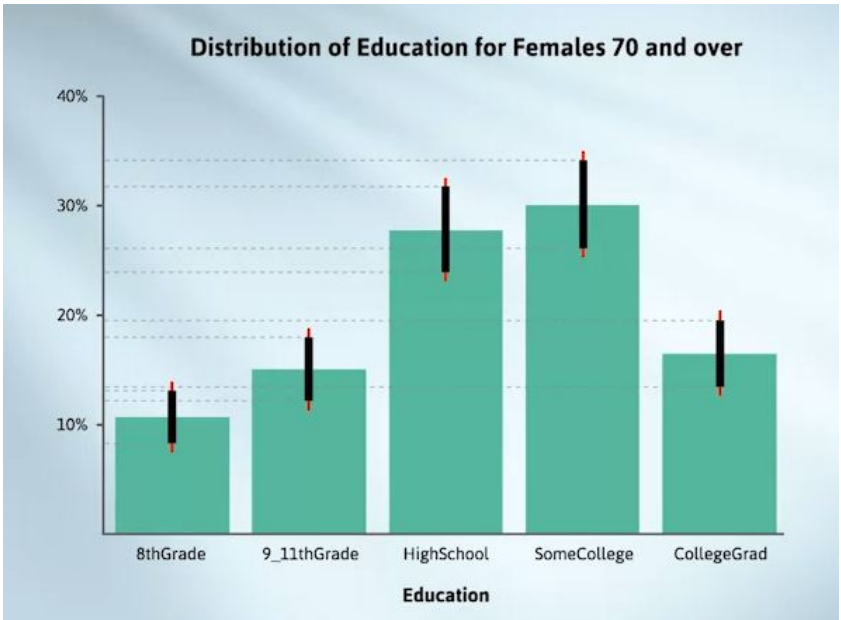
WEEK 6

ESTIMATION WITH CONFIDENCE FOR A CATEGORICAL OUTCOME

by Chris Wild

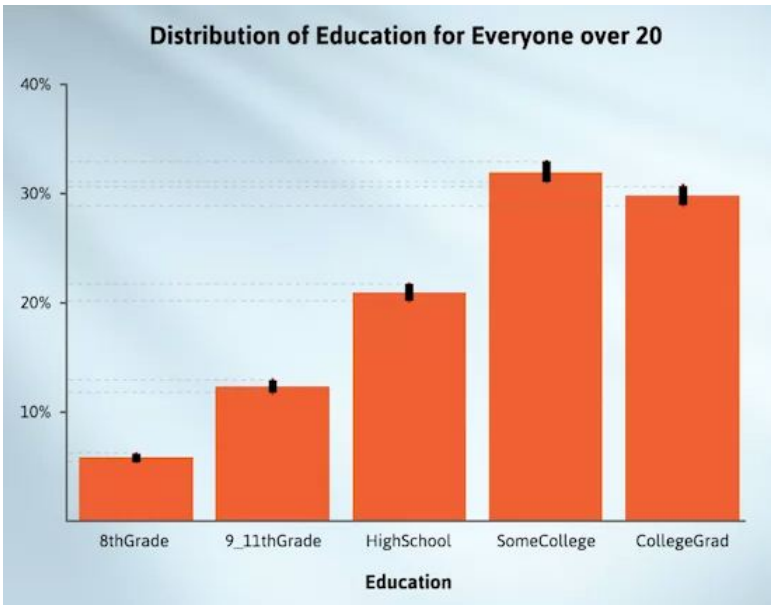
Welcome back. In the last video, we experienced estimation with confidence in the context of numeric outcomes and group means. We'll now move on to focus on a categorical outcome variable.

For this video, we'll use everyone in NHANES-2009-12 who's 20 or older. We'll focus on the Education variable, which we've reordered to form a natural progression as in Week Three.



This is a barchart of Education for just the people who are 70 or older. There are roughly 350 of them. Just as in the last video, we've added confidence intervals and comparison intervals to allow for the uncertainty in estimating the true values due to sampling error. This has produced small lines around the tops of each bar. The thicker black lines are the comparison intervals. The confidence intervals are the longer red lines, which sit underneath the black lines with only their ends showing.

We work with these in the same way as before. There are lots of comparison where we can't see which bar is really higher, that is which of the true percentages is bigger. For example, we can't tell for this subgroup which percentage is really bigger between "8th Grade" and "9-11th Grade", or between "High School" and "Some College", or between "9th-11th Grade" and "College Graduate". We can, however, tell that the true percentage with "Some College" is considerably higher than the percentage who are "College Graduates"or "8th Grade" or "9-11th Grade".



But when we look at everyone over 20 (nearly 7,000 people instead of only 350) the intervals are much narrower. The true heights of the bars are fairly precisely estimated. And there's no overlap here between any of the comparison intervals. The true population percentages are clearly in the same order as we see with the bars.

Let's now look at some of the relationships we investigated in Week Three and see if the features we highlighted then can still be taken seriously once we take sampling error into account.

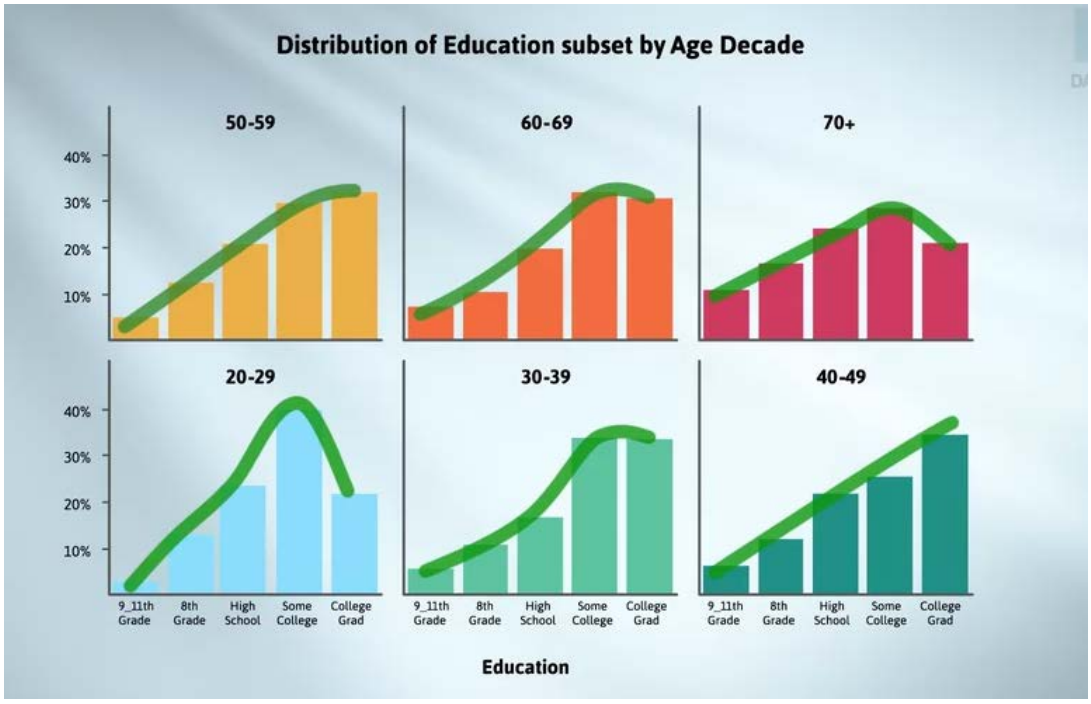
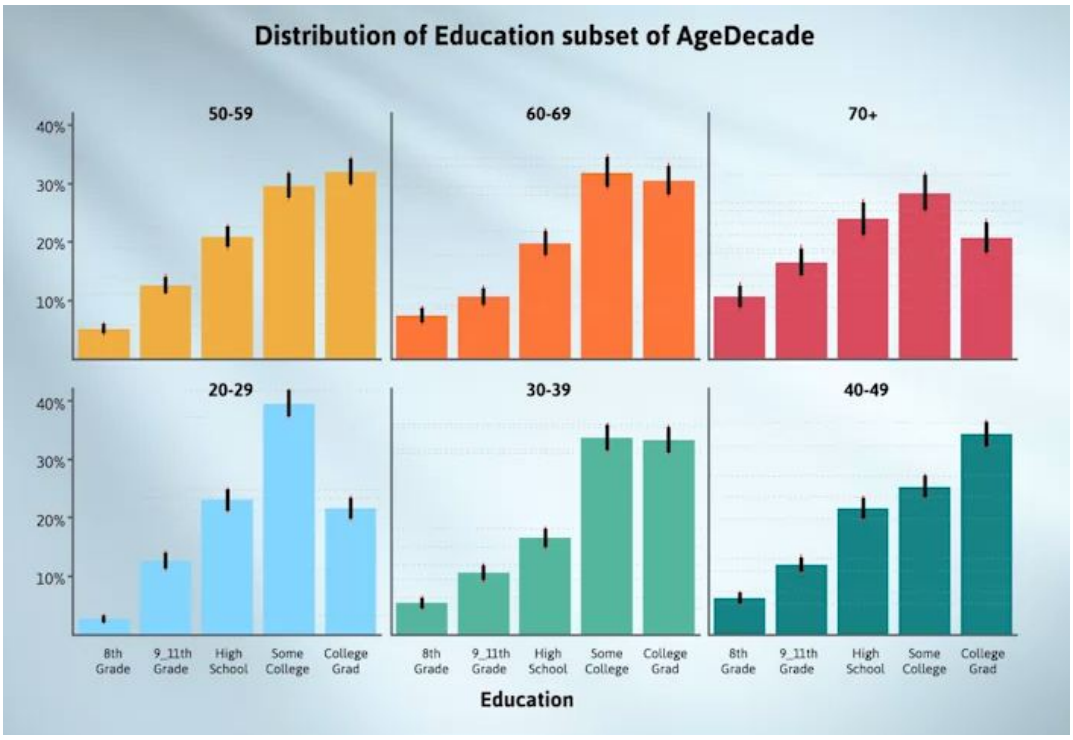
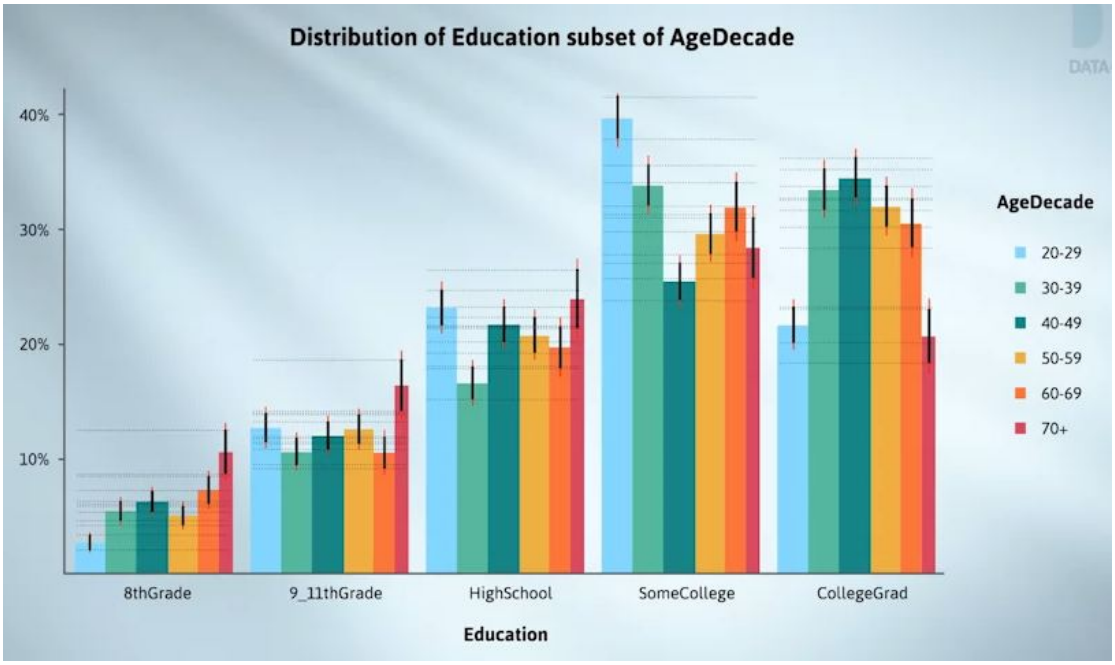


Image from Week 3 (Relationships between categorical variables)



We looked at separate graphs for the Education outcomes for each decade of Age. What were some of the main features we highlighted?

- There was a big dropoff in percentages in the 70 plus group going from "Some College" to "College Graduate". That change is clearly still evident, even if we consider the uncertainties. The intervals are widely separated.
- And then there was an even bigger dropoff in the "20 -29" group. That change is clearly real too. The intervals are even more widely separated.



Final still from explanatory animation sequence

We also looked at the relationship between Education and AgeDecade using side-by-side bar charts. We highlighted these features:

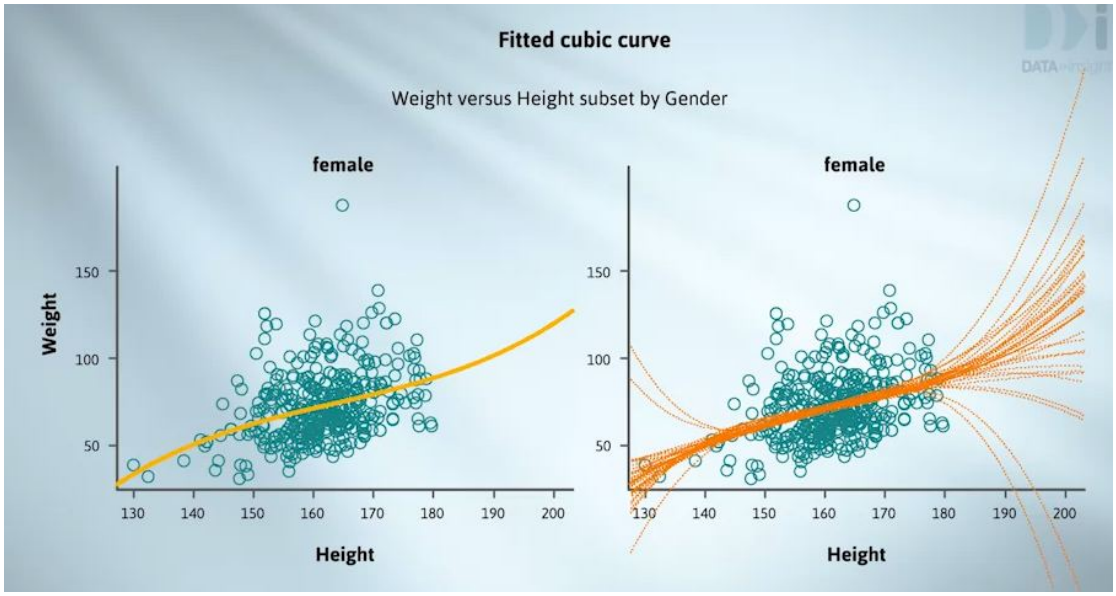
- For the 70 and over group, the low percentage of college graduates. That's still clearly supported.
- The high percentage with "8th Grade" or "9-11th Grade" education compared with all the other age groups. That's also still clearly supported.
- The high percentage with "High School" compared with close age groups. There's overlap between the intervals there, so we can't really draw that conclusion.

We highlighted these features for the "20-29" age group (reading from the right):

- The low percentage of "College Graduates" compared with everyone except the 70 and over group. Clear separations, so clearly supported.
- The high percentage with "Some College" compared with all other groups--clearly supported.

- The high percentage with "High School" compared with everyone except the 70 and over group -- only supported for about half of the other groups.
- The high percentage with "9-11th Grade" compared with everyone except the 70 and over group -- not supported for any of the comparisons because of overlap with all other comparison intervals.

That brings us to the end of our treatment of categorical variables, but we're not finished quite yet. I want to pick up a loose end from Week Four. Let's go back to these plots.



Now that you know what a bootstrap resample is, I can explain what they are and what they're for. In the left-hand plot, the trend is summarised using a smoother. The right-hand plot is what we get from iNZight when we ask for "Inference Information" for a scatterplot that has a trend curve drawn on it. So what are these other curves on the right hand plot?

Well, 40 bootstrap resamples have been taken from the data, and for each resample, the same type of curve has been computed and added to the graph. So we get 40 additional curves from the 40 resamples. Statisticians often describe something that looks like this as a nest of curves.

Now for each resample there will be some data points that aren't used. We can feel confident that we know fairly precisely where the true trend should be positioned at places where all the curves in the nest are very close together. We shouldn't have any confidence in the fitted trend in regions where the curves are far apart, because there, small changes in the data are making big changes in the result.

That brings us to the end of this video and the end of this discussion of estimation with confidence.