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

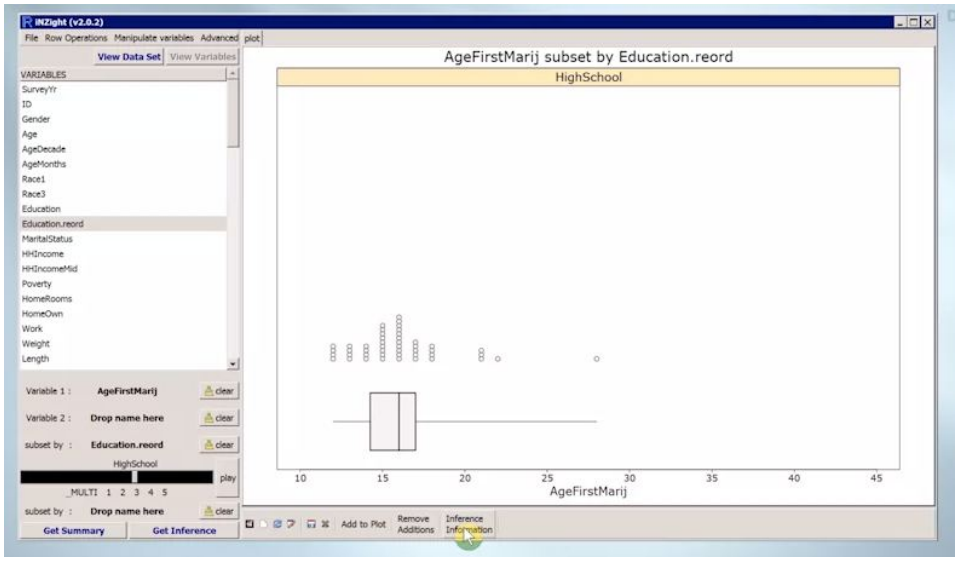
WEEK 6

NUMERIC OUTCOMES by Chris Wild

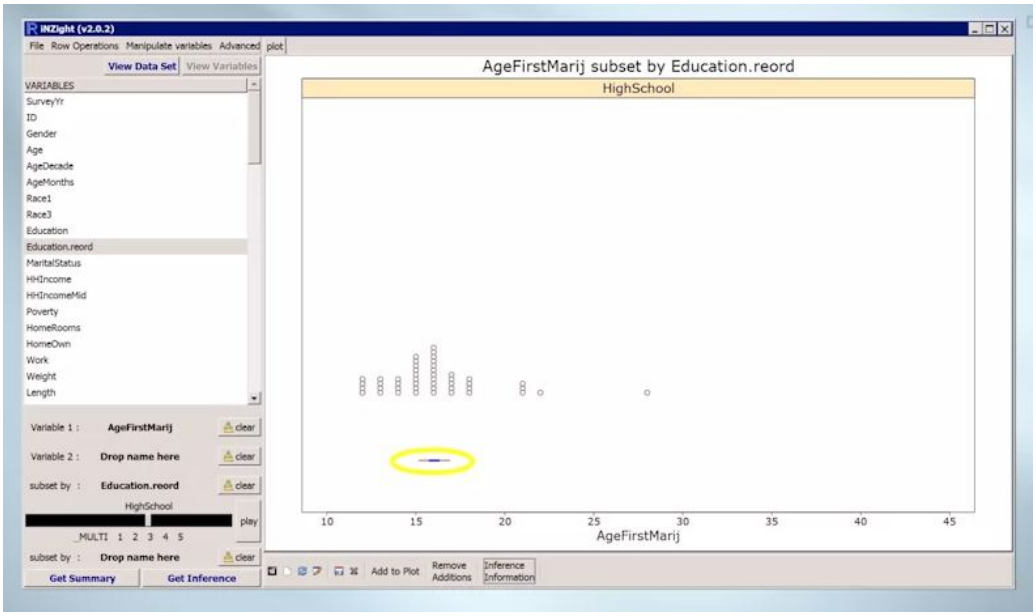
Hi. This week we've been building up the idea of estimating quantities from data with confidence. The idea is to put intervals around the estimates we get from our data to allow for sampling error. Each interval is obtained in a way that makes us pretty sure that the true value of the quantity we're estimating will be in the interval.

We'll now take this idea back into our data analysis. We'll look back at some of the NHANES variables that we've worked with previously. But this time we won't just take the estimates our data gives us at face value, we'll use confidence intervals to take into account the uncertainties about where the truth values lie.

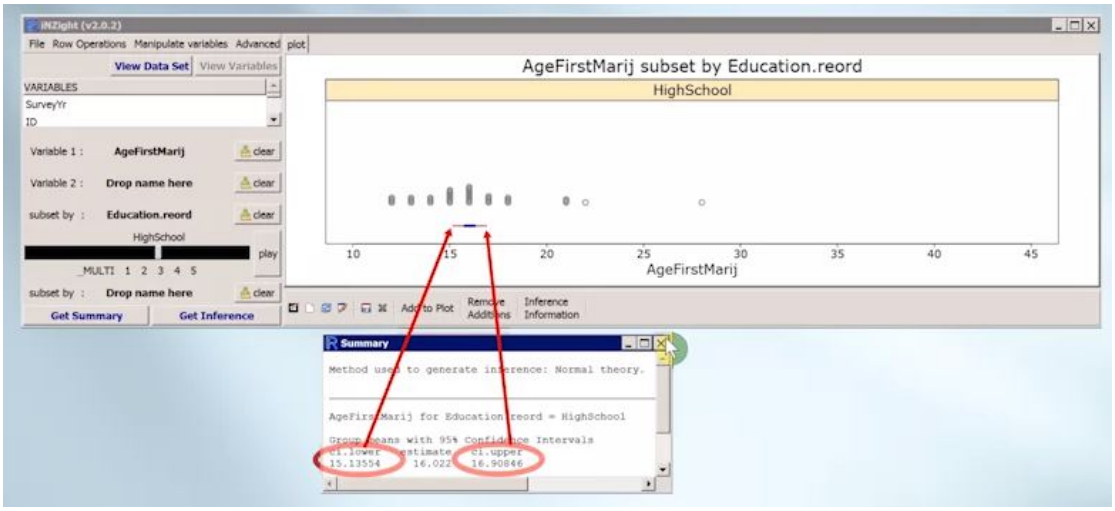
We'll start with the NHANES-1000 data here. My base question is, "At what age did those people who have tried marijuana first try the drug?" I'll be using iNZight.



I'm looking here only at people whose highest educational attainment is completing high school. The median age of first use looks like sixteen. iNZight can be asked to mark up most types of plot with appropriate information about uncertainties via the "Inference Information" button.



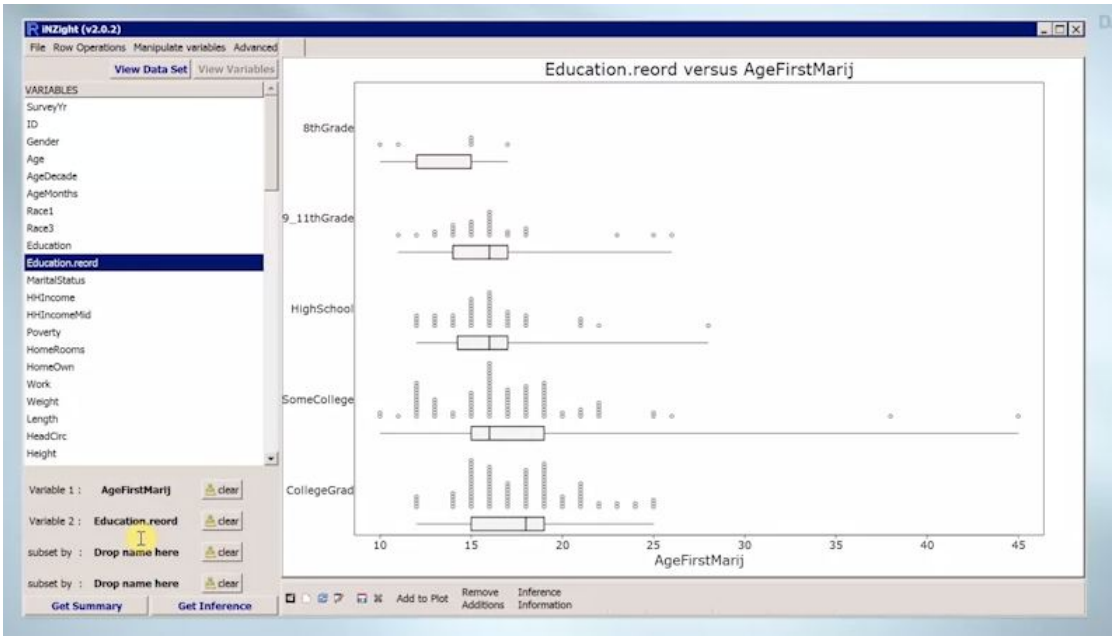
Here I've asked for inference information and just accepted the default settings. Two lines have been added. There's a thick, blue line, which sits on top of thin red line so that only the ends of the red lines are showing. (We'll ignore the thick, blue line for now.) The longer red line is the confidence interval. It gives a confidence interval for the true population mean. It seems to go from about 15 to 17.



We can click Get Inference to find the actual values. The estimate of the true mean from our data is 16.02, and the lower and upper 95% confidence limits are 15.135 and 16.908.

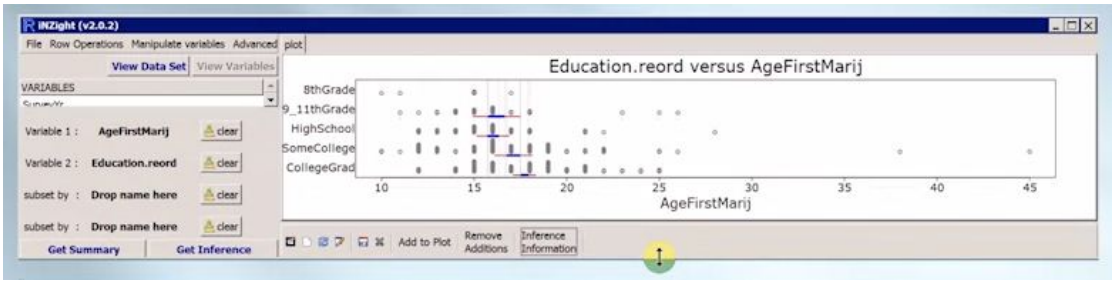
I'd translate this to say with 95% confidence the true mean age of first use is somewhere between about 15.1 and 16.9.

Let's now compare "Age of First Marijuana Use" for the different levels of education.



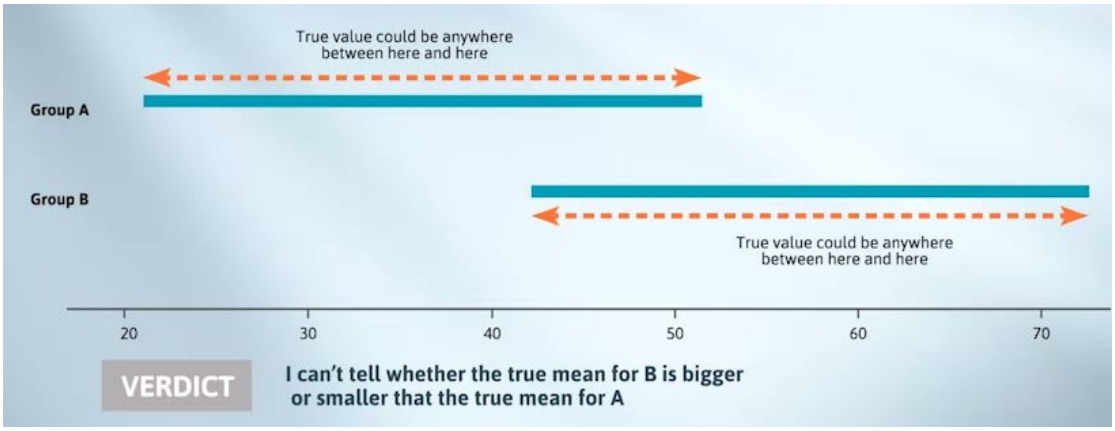
I've reordered the education categories to make them logical. It looks like the mean-age-of-first-use is increasing with educational attainment (going down the plot).

Let's add uncertainties to the plot, just as we did before. There's clearly a lot of overlap between the confidence intervals.



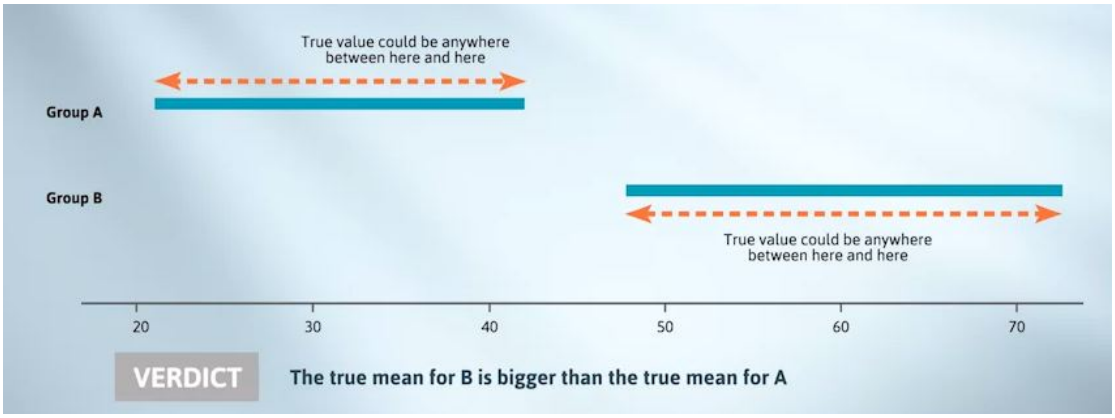
Squashing the graph up like this, makes overlap between intervals easier to see.

If the true value could be anywhere in the interval, then -- when intervals overlap -- we can't tell which true value is larger.



Non-overlapping intervals

Whereas if there's no overlap, then it's obvious which is bigger.



Non-overlapping intervals

There's a slight complication in this. As you know, 95% confidence intervals catch the true value for 95% of samples taken, so the error rate is 5%, or one sample in 20.

If we want to compare means visually in a way that has the same error rate, it turns out that we should use shorter intervals. These shorter intervals are the thick, blue lines (inside the red lines) which iNZight calls comparison intervals. How they're calculated is beyond this course.

So when we're making comparisons and looking for overlap we should use the intervals given by the thick, blue lines. Then, if there's no overlap, it's obvious which group has the bigger true value. But if there is overlap, we can't tell which of the true values is larger. We'll work with just this for a while before doing anything more sophisticated.

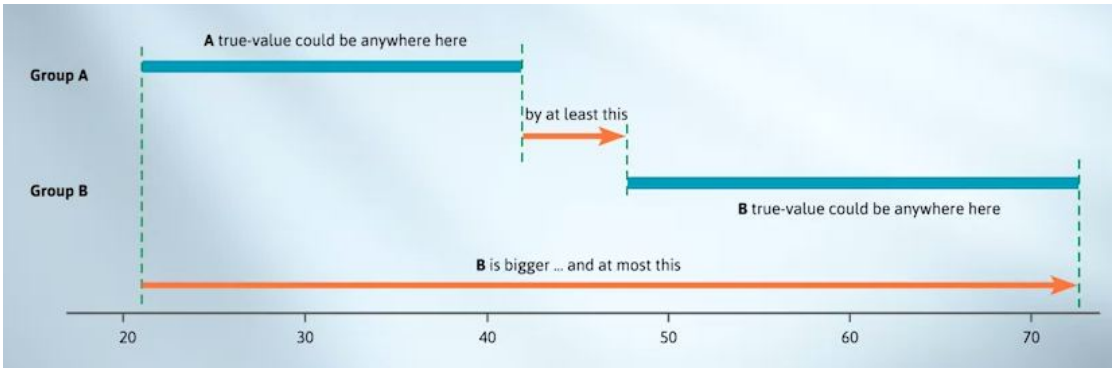
No intervals have been drawn for the "8th Grade" because there's not enough data there. There is substantial overlap between the "9-11th Grade" people and those

who finished high school, so we can't tell which group really has the highest mean age of first use.

However, because there's complete separation, we can see that both college groups have a higher mean age of first use than either of the pre-college groups. It's a close call where there's an overlap between the comparison intervals for the two college groups.

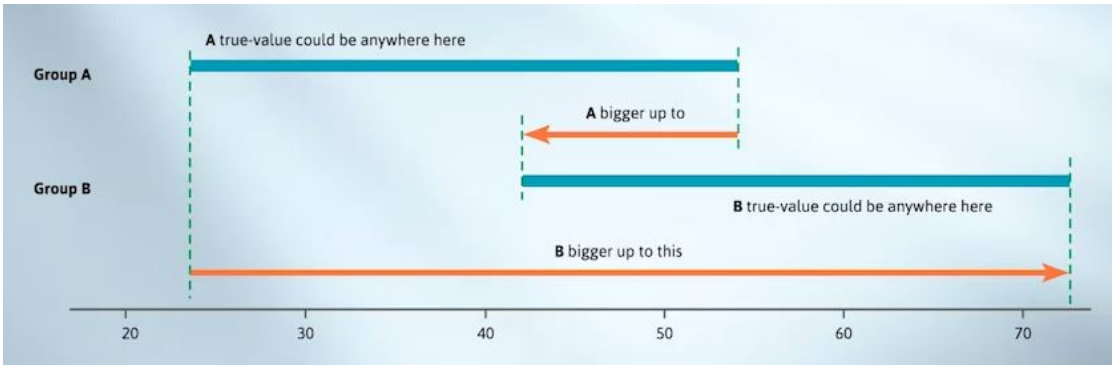
However, the mark-ups of our plots provide only approximate impressions. Their job is to let us see a whole lot of things quickly and in the context of everything else. We can then go to the numeric information provided by "Get Inference" for details and to confirm our impressions.

We can go further and see an interval for the difference.



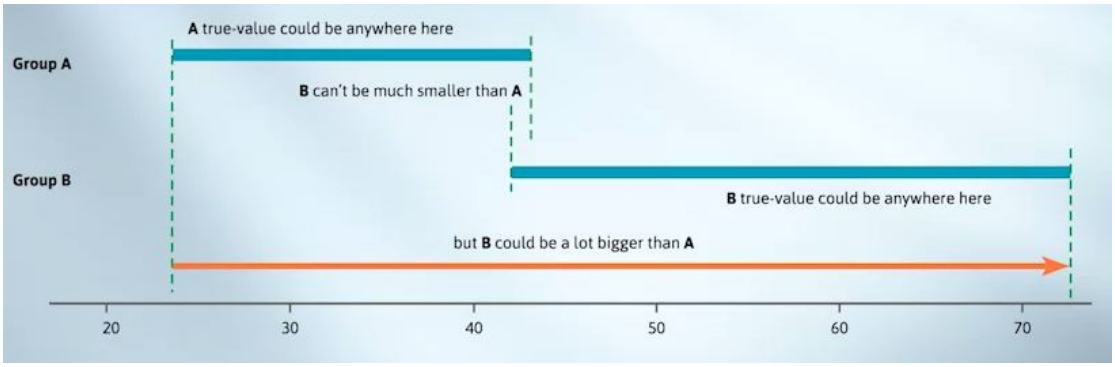
True differences with non-overlapping intervals

Here "B" is clearly bigger by at least this and at most this.



True differences with overlapping intervals

Here it could go either way. "A" could be bigger by up to this much, or "B" could be bigger by up to this much.



Differences with overlapping intervals

When there's very little overlap, "B"can't be much smaller than "A" but it could be quite a lot bigger. We'll talk about reading and interpreting this output in the article that follows. That completes this video. Next time we'll apply our estimation with confidence ideas to categorical variables, and finally, trends on scatter plots.