**Transcript: Calculating Interquartile Range in R Commander**

We are going to go over how to calculate the interquartile range in R Commander as well as producing boxplots. These are both great options when you have data that are potentially skewed or otherwise not normally distributed.

The data set we're going to be using today is the SLID data set which you've encountered before in an assignment. It looks at labour force participation in Ontario. If we look at our data set using “View data set”, we'll be able to see we have “wages”—this is in dollars per hour—education in years, age, sex, and first or primary language.

When we look at income or wages, we might expect this to be skewed because we know that most people make in the low to middle income range, and then there are a few people who make a lot of money that's well beyond the majority. So this is a great opportunity to look at some of the resistant indicators and why they might be better than things like just looking at the mean.

The first thing we want to do is look at our interquartile range. To do that, we can go to “Statistics”, “Summaries”, and “Numerical summaries...”. I keep losing it. There we go. OK, so the variable that we're going to use is “wages”. And we'll go into “Statistics” and we can select the statistics that we want to look at. We can keep all of this as is. If you just wanted the interquartile range, you could just select that, but we're going to compare it to the mean and standard deviation just so you can see some of the advantages of using the interquartile range.

Once we click “OK”, you can see here that our mean wage is 15.55. Our standard deviation is 7.88 and our minimum is 2.3 and our maximum is 49.92. So, as we thought, this distribution of income seems to be pretty far skewed to the right. You can see that 75% of people earn $19.80 per hour or less, but the highest is $49.9.

And just to demonstrate, you don't need to follow along with this, but I will show you a histogram as we went over in the data visualization lab just so you can kind of see this visually. And as you can see there, there's our positive skew, with most people in this lower range, and then a few up here at the top.

So, the statistic that we want to look at is our interquartile range, which is “IQR” here of 10.565, and that's going to be much more informative than the range as a whole. Remember that R Commander doesn't actually give us the range, but we can calculate it pretty easily just by doing the maximum minus the minimum. In this case 49.92 - 2.3; you could get out a calculator to do that, but it's 47.62. We might look at that and say, wow, there's this huge variability in how much income people are earning in Ontario in this survey. But if we look at our interquartile range, it's only 10.565. And how we got that is by doing the 75th percentile minus the 25th percentile. If you plug that into your calculator, you would get the same thing as the IQR here. And this is giving us a much better estimate of the true spread of values within this data set, because it's excluding these high outliers on this end. And if there were low outliers, it would also be excluding those, although there don't seem to be too many.

So that's an example of how the interquartile range can be really handy when we calculate descriptive statistics with skewed distributions. Another one that we went over last week with descriptive statistics, but that we can look at again is the median. This 14.09 is going to better reflect the centre than the mean of 15.55.

That's how you look at resistant indicators in R Commander.