Last year, the liberal party obtained 39.47% of all seats. To make our calculations easier, we may consider this as 40% of all seats. Thus my first question would be,

Will the liberal party gain 40% of all votes in Canada during the 2019 election?

The population for this would be the actual proportion of voters who chose the liberal party as their first vote of choice in 2019.

We can visualize this data by forming a histogram, which showcases the different percentages of voters for their respective parties. By displaying the percentage of votes each party received in 2019, we can visually compare which party has the largest pool of votes. We have the 'votechoice' variable available to us, which shows the first choice for every voter. With this data, it will be easier to gauge the possibilities of a sweep by the liberal party.

With the help of the hypothesis test, we can test this data to see if our question rings true. For this test, our null hypothesis would be the assumption that the liberal party gains 40% of all votes in Canada during the 2019 election. Of course we must include the alternate hypothesis here as well, i.e., the assumption that the liberal party will not gain 40% of all votes in Canada in 2019. Next, we would calculate the test statistic using our sample data and begin our simulations in the upcoming step. When it is time to simulate, we simulate under the assumption that the null holds true. When plotting the histogram for this simulated data, we will find our null at the middle of our histogram. Using this, we can see what sorts of values are usual for the statistic we are calculating. When plotting our histogram, we should add vertical lines that represent our test statistic, and another that is the mirror of our test statistic. With the help of these vertical lines, we can calculate the p-value, which is the proportion of simulated values which are either like our test statistic or more extreme. The p-value allows us to determine whether we can reject or accept the null, depending on the size of the p-value. A smaller p-value indicates strong evidence against the null, which means we would have to reject our null. On the other hand, a larger p-value indicates weak evidence against the null, which means we would be able to accept the null.

Are the means for liberal party voters the same in Quebec and Toronto?

The population for this would be the true proportion of liberal party voters in Toronto and Quebec in 2019.

We can visualize this data by forming two boxplots, one for the liberal party voters in Toronto, and the other for the liberal party voters in Quebec. Forming two boxplots would allow us to compare the data for either city easily. We would first have to use the variable 'votechoice', to filter out all voters for any party other than the liberal party. Then using this data, we would have to further filter it with the 'province' variable, which can be used to separate the liberal party voters in Toronto and Montreal. By using these two variables, we can effectively plot the separate boxplots for the liberal party voters in either city. Comparing this data can help to direct efforts of campaigning to whichever city seems to require it, in order to increase votes.

To investigate the answer to this question, we could use randomization tests. Also known as two tailed hypothesis tests. For such data, we can form a null hypothesis that there is no difference between the means of the liberal party voters in Montreal and Toronto. The next step would be to determine the alternate hypothesis and calculate the test statistic (the difference in means of the voters in both cities) from the sample. Moving on, we would have to use RStudio to create multiple simulations, without replacement, of this same data being calculated for shuffled values. It is after this part that we plot our obtained simulated values and calculate the p-value from this. Finally with the help of the p-value, we can determine whether it is possible to accept of reject the null, i.e., we can determine whether there is no difference between the averages of liberal party voters in the city of Toronto and Montreal.

What is the range of plausible values for the average age of liberal party voters?

The population for this would be the true proportion of all voters of the liberal party in Canada in 2019.

We could visualize this data from a histogram. We would use the 'age' variable to carry out observations with this data. Using a histogram would allow us to easily visualize the spread of voters for the liberal party according to age. This could help with the campaigning program, i.e., it could possibly help to direct the campaigning towards a certain age group.

To investigate the answer to this question, we could use bootstrapping to help us. In bootstrapping, we require an initial sample of size n from the population. The next step would be to re-sample this original sample multiple times into samples of size n. This should be done with replacement, so that we do not obtain the value of the original sample. Then for each of these bootstrap samples (the re-samples) we should calculate a test statistic. Taking all these bootstrap statistics, we can create a distribution called a bootstrap sampling distribution. It is this sampling distribution that will give us the range of plausible values for the population parameter. However, we have to find this using percentiles, i.e., by taking a confidence interval. For a 95% confidence interval, the calculated interval will include the parameter for 95% of possible samples. That is, the middle 95% of values will be taken. Hence we will be able to obtain a range of plausible values for the average age of voters for the liberal party.