

I've Been Framed! Sampling and The Margins of Errors

BY: JANE TANG, VP ANALYTICAL INTELLIGENCE – VISION CRITICAL

Part of the Series “Take Me Back To The Stats: Where Data and Insight Collide”



You've probably seen polls that are based on a “random sample of the Canadian population”. Or, to be more accurate, you've probably seen polls that SAY they are based on a “random sample of the Canadian population”. And while that expression is commonly used, it's not exactly true. It simply isn't possible to get a true random sample of the whole population. Nobody knows the names of all Canadians, or exactly how many of us there are – not even Statistics Canada, despite their best efforts to count as many heads as possible every five years.

What the pollster is really saying is that he has a frame of the Canadian population from which he can draw a random sample. What is a “frame”? It's a list of potential candidates, used as a substitute for the (unobtainable) list of everyone in the population. When a substitution like this happens, a frame becomes imperfect. This immediately introduces a new concern – how well does the frame match the population?

How a pollster constructs his frame is largely dependent on how they intend to ask their questions, or the mode of survey. A pollster that uses telephone calls to gather data is going to start with the telephone list, and is not likely to include in their frame people who don't have a phone. Or people who only have a cell phone, or have an unlisted number. Pollsters that use internet panel samples is going to start with the list of people using the internet, and are unlikely to include in their frame people that don't have some way of getting online.

Once a pollster has selected a sample from the frame he constructed, they face another significant obstacle: getting people to respond. People may not answer the phone, or decline to participate once they find out who is calling. They may start to ignore

all of those survey requests piling up in their email inbox. At the end of the day, the best a pollster can hope for is to obtain responses from a subset of their chosen subset of the population.

How does this relate to margin of error of the sample? The first consideration is the measure of sampling error between the sample respondents and the frame. This can be calculated when the properties of the frame are known. When potential respondents from the frame is selected into the sample with known probabilities, the size of the sampling error for any statistics derived from the survey data is largely a function of sample size. Given a sufficiently large sample size, discrepancies can be corrected for. "Our sample doesn't have enough responses from left-handed shepherders living in Ontario, so we'll increase the weighting on that group. This decreases the overall efficiency of the sample, and we account for that by the increased margin of error rates around the sample estimates."

The next consideration takes us back to where we started – how well does the frame match the population? In many cases the target population isn't going to be the entire population of Canada anyway. Children and Canadians living aboard are routinely excluded. But once you've defined your target population, you still want to get a good fit between your target and your frame.

For many years the best substitute for a list of the whole population has been a frame consisting of everyone's home phone numbers. Historically, most Canadians had a home phone, and not too many of them were unlisted. This frame can be accessed easily, using "Random Digit Dialing", or RDD. Despite its name, RDD does not involve computers generating truly random phone numbers. Instead, the first eight digits of each phone number are drawn from a bank of known phone numbers, and only the last few digits are dialed randomly. This system worked well, but is starting to fray at the edges as technology changes. - More people are dropping their landlines every year, and ever more people use call display to filter out unknown callers.

Online polling (using panels of 'registered respondents') draws from a different frame – people who have internet access. In the past, the fit between this frame and the population has not been considered as good as the fit that could be obtained by calling people on the phone. After all, far more people had phones than had computers. But this is not as true as it once was. More people are getting online every year, while more people are dispensing with their landline telephone. And most of the people who've discarded their landline have internet access. Complicating things further, they may use an internet phone, whose location can be difficult to pin down – that Edmonton phone number might ring a VOIP phone located in Oshawa or New York. Furthermore, response rates for online panels are higher than response rates for phone surveys. After all, if a market researcher phones you just as you're sitting down to dinner you're probably not going to want to take "a few minutes" to talk to them. But an email invitation to participate in a survey can easily be delayed to a more convenient time.

This means that it's no longer a given that phone polling provides a better frame than online polling. It can credibly be argued that online polls provide an adequate fit for the entire population. Online polling results are used repeatedly to predict results in several major elections in Canada. This provides supporting evidence for this assertion. Given all of these complications, can you believe the published results (and margins of error), that you come across in the media? We know that these numbers aren't really based on random samples of the entire population. But as long as the inference is limited to the frame itself (e.g. the online panel; people with home phones who are willing to talk to pollsters) it is the statistical truth. Things get a bit grayer if that inference is expanded to cover the general population. How gray things get depends on how good of a fit there is between the frame and the population. As we've seen, both of approaches (random dialing and online panels) have limitations, but with care can be made to work.