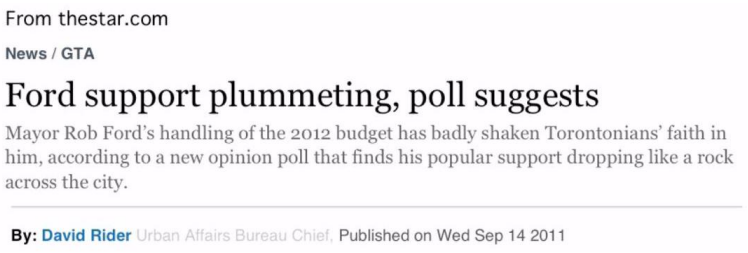
Confidence Intervals

Sample Size for Estimating a Proportion

The goal of this lecture is to learn how to determine the sample size needed to estimate a proportion. When planning any study, determining the number of observational or experimental units required is an extremely important consideration. If the sample size is too large we will waste time and resources and unnecessarily expose our observational or experimental units to any risks or inconveniences that may be associated with being studied. On the other hand, if our sample size is too small we may not have have precise enough estimates to make meaningful conclusions.

Example 1
Here is an example of a typical poll published on the Toronto Star newspaper's website in 2011.


From thestar.com

News / GTA

Ford support plummeting, poll suggests

Mayor Rob Ford's handling of the 2012 budget has badly shaken Torontonians' faith in him, according to a new opinion poll that finds his popular support dropping like a rock across the city.

By: David Rider Urban Affairs Bureau Chief, Published on Wed Sep 14 2011

Figure 1: Toronto Star headline for Rob Ford poll

> *"The Forum Research survey of 1,046 Torontonians conducted Monday after the release of city manager Joe Pennachetti's recommended budget cuts, pegs Ford's support at 42 per cent – a big drop from 57 per cent on June 1, and 60 per cent in late February... The automated telephone poll's margin of error is plus or minus 3 per cent, 19 times out of 20."*

Rob Ford became mayor of Toronto on December 1st, 2010. He was elected on a platform that he called 'Respect for Taxpayers' which included the elimination of any spending in the city budget that he felt was unnecessary. The reality of implementing budget cuts is always difficult and controversial. As the headline suggests, the people of Toronto's support of cost cutting dropped as they saw programs important to them being eliminated.

Let p represent the true but unknown proportion of people in Toronto who approved of the mayor's budget. From this excerpt we know that $n = 1046$ and $\hat{p} = 0.42$. Since 19 times out of 20 is 0.95, the stated margin of error of 3% is for a 95% confidence interval (CI).

$$95\% \text{ CI for } p = 0.42 \pm 0.03 = [0.39, 0.45] = [39\%, 45\%]$$

Let's step back now to the planning stages of the poll. A 3% margin of error for a proportion has become a standard value used in most surveys. This results in a confidence interval for p with a length of 6%. Previously, we saw that the margin of error is given by the formula

$$\text{Margin of error for } p = z_{\alpha/2}\sqrt{p(1-p)/n}, \tag{1}$$

where $z_{\alpha/2}$ is a quantile from a standard Normal distribution.

While planning a survey we can decide on the margin of error and the confidence level that will be appropriate. This helps us determine how many people should be included in the survey. Suppose we choose a 3% margin of error for a 95% confidence interval for p.

For a 95% CI, $\alpha = 0.05$ implies $\alpha/2 = 0.05/2 = 0.025$. Therefore, the required normal quantile is $z_{0.025} = 1.96$. Since we do not know p and we have not yet collected any data we cannot estimate p by $\hat{p}$ in the margin of error formula. The typical course of action is to take the worst case scenario which will give us the largest possible margin of error. In this way, whatever our estimate of p turns out to be, we will have a margin of error that is no larger than our goal of 3% (or whichever we have planned).

Which value of p gives the largest possible margin of error? The margin of error depends on the quadratic term $p(1-p)$ and is maximized at the maximum of $p(1-p)$. Figure **??** is a plot of the function $p(1-p)$ showing all the possible values as p ranges from 0 to 1. The maximum value occurs when $p = 0.5$.

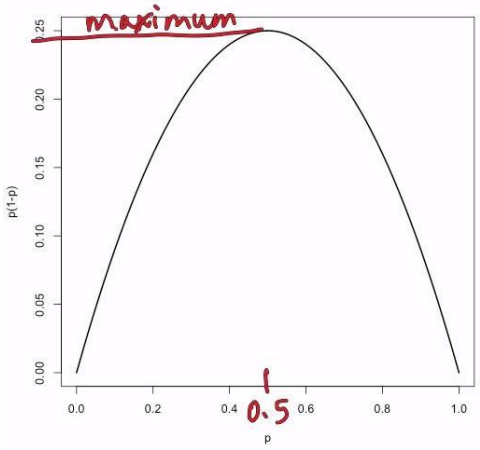


Figure 2: The maximum of the function $p(1-p)$ occurs at $p = 0.5$

Now we can substitute all the known quantities and solve for n in equation (**??**).

$$\begin{aligned}
\text{Margin of error for } p &= z_{\alpha/2}\sqrt{p(1-p)/n} \\
0.03 &= 1.96\sqrt{0.5(1-0.5)/n} \\
1.96(0.5) &= 0.03\sqrt{n} \\
n &= \left(\frac{1.96(0.5)}{0.03}\right)^2 \\
n &\doteq 1067
\end{aligned}$$

This sample size of 1067 is very close to the sample size of 1048 used in the survey. What if we wanted a smaller margin of error, say 1%? Then

$$\begin{aligned}
0.01 &= 1.96\sqrt{0.5(1-0.5)/n} \\
n &\doteq 9604 \\
n &\doteq 9 \times 1067
\end{aligned}$$

Note that reducing the margin of error to one-third of its previous value requires increasing the sample size nine-fold.

Similarly, if we wanted to reduce the margin of error by a half, we need four times the original sample size.

$$\begin{aligned}
0.015 &= 1.96\sqrt{0.5(1-0.5)/n} \\
n &\doteq 4268 \\
n &\doteq 4 \times 1067
\end{aligned}$$

How would the required sample size change if we wanted a 99% confidence interval instead of a 95% confidence interval? We need a larger sample size to have more confidence. Let's use a margin of error of 3%. For 99% confident interval, $\alpha = 0.01$ so $\alpha/2 = 0.01/2 = 0.005$. Therefore the normal quantile is $z_{0.005} = 2.576$. The sample size required is

$$\begin{aligned}
0.03 &= 2.576\sqrt{0.5(1-0.5)/n} \\
n &\doteq 1843
\end{aligned}$$

If we wanted 90% confidence interval with a margin of error of 3% we would need fewer subjects. We use the normal quantile $z_{0.05} = 1.645$ and calculate the sample size

$$\begin{aligned}
0.03 &= 1.645\sqrt{0.5(1-0.5)/n} \\
n &\doteq 752
\end{aligned}$$

Example 2
Here is another city of Toronto poll from the Toronto Star.


From thestar.com

News / City Hall

Massive poll shows Toronto is united against Ford's proposed cuts

One of the biggest polls ever conducted in Toronto shows residents from every corner of the city are overwhelmingly against Mayor Rob Ford's cuts. A survey of nearly 13,000 show over three-quarters of Torontonians want their local councillor to protect services rather than comply with the mayor's wishes.

By: Robyn Doolittle Urban Affairs Reporter, Published on Fri Sep 16 2011

Figure 3: Toronto Star headline for massive Toronto poll

> *"A Forum Reseach telephone survey of nearly 13,000 people reveals that more than three-quarters of Torontonians want their local councillor to protect services rather than comply with the mayor's wishes. And only 27 per cent of residents say they would vote for Rob Ford if an election was held tomorrow.*
>
> *More significantly, because of the poll's size, Forum was able to provide the first authoritative assessment of support on a ward-by-ward level.*
>
> *Forum's poll, which was paid for by CUPE Local 79, one of two major unions at city hall, questioned 12,848 Toronto residents on Tuesday using a random dial, push-button response, phoning system. The margin of error is plus or minus 0.9 per cent, 19 out of 20 times."*

Why did this survey include so many people in their sample? The researchers wanted to be able to estimate the proportion of Ford supporters in each of the 44 wards in Toronto. The article claims that the margin of error of this poll, with 95% confidence, is 0.9%. This margin of error may be true for a proportion calculated using data from all 12,848 respondents but it is not true on a ward-by-ward basis.

If we assume that there are equal numbers of respondents in each ward, then there are $12848/44 = 292$ respondents per ward. The margin of error is then 5.7%.

$$\text{Margin of error for 95\% CI for } p = 1.96\sqrt{0.5(1-0.5)/292} = 0.057$$

This situation also illustrates something we must always keep in mind with confidence intervals. For a 95% confidence interval, we expect it to cover the population proportion p 95% of the time. In other words, 95% of such intervals based on random samples from the population will include p. The other 5% of the time the confidence interval does not contain p. Unfortunately we never know whether we are in the 95% or in the 5% for the single sample we collect.

With 44 confidence intervals, one for each ward, we have a pretty good chance of missing the true p in at least one of those confidence intervals. Since we expect to miss about 5% of

the time, in 44 confidence intervals, we expect to miss the true proportion about twice. And we do not know in which ward we have failed to capture p.

Whenever there are multiple polls and we have a 95% confidence interval for each one we need to be cautious in interpreting the results. We expect that 5% of the confidence intervals will miss the parameter we are trying to estimate. This is not necessarily a problem since it is exactly what we expect from the randomness in the sampling process.

Finally, remember that no matter how large the sample size for a poll is, the resulting confidence interval is not likely to capture the population proportion if the survey suffers from non-response bias or includes improperly worded questions which may lead to response bias.