**Choosing a sample size when estimating a proportion**

Before a survey or an experiment is conducted, we need to choose n, the number of trials or the sample size. How do we pick n?

If hypothesis testing is being performed, one could choose n with respect to power. Alternatively, if estimating π is the main interest, a sample size can be chosen so that

 ≤ E

with a certain level of confidence and where E represents the allowable error. Notice that  shows how far the estimate, , is from the population parameter. The E stated here is the margin of error which is often stated in polls or surveys presented in the news media!

Sample size to estimate π: We can be approximately (1-α)100% confident that the error will be less than the specified amount E when the sample size is approximately



Where does this result come from? Earlier, we saw that



If we let  = E and set



solving for n produces the desired result.

Problem: What is π? Use a conservative solution by using the π maximizes π(1–π). Why is this desirable?

Examine the table below.

|  |  |
| --- | --- |
| π | π(1–π) |
| 0.1 | 0.09 |
| 0.4 | 0.24 |
| 0.45 | 0.2475 |
| 0.5 | 0.25 |
| 0.55 | 0.2475 |

Therefore, π = 0.5 maximizes π(1 – π). One can also show this by taking the derivative of π(1 – π), setting it equal to 0, and then solving for π.

Another way to think about this is we want the value of n needed to guarantee  ≤ E no matter what π is.

Sample size to estimate π (restated): We can be **AT LEAST** (1-α)100% confident that the error will not exceed a specified amount E when the sample size is



Example: Political Polls

Suppose you want to determine the percentage of registered voters who will vote for the Republican or Democratic nominee in a presidential election. How large of a sample should be taken so that the margin of error is less than ±3.5% with 95% confidence?



If this had resulted in a non-integer value (like 784.2), the next largest integer would need to be chosen (785).

Suppose in the resulting poll Democratic nominee has 51% and the Republican nominee has 49%. Who is leading?

Note about the margin of error:

The margin of error needs to be added or deducted from both candidates in a two-person race. For example, suppose candidate A had 55% and candidate B had 45% with a margin of error of ±5%. This is still an “even race” since candidate A could be as low as   
55%-5%=50% and candidate B could be as high as 45%+5%=50%.

Example: Clinton and Dole 1996 presidential race

There were two polls taken by a TV network right before and right after Bob Dole chose Jack Kemp as his running mate. Below are the results.

|  |  |  |
| --- | --- | --- |
|  | **Before** | **After** |
| Clinton | 53% | 52% |
| Dole | 37% | 39% |

After presenting the results of the poll, a TV news anchorperson said, “Kemp has helped Dole in the presidential race.” Suppose E = 4%. What do you think about this statement?

Note about the confidence level:

There are many polls taken and usually the margin of error is given. Why do organizations taking the polls and stating the results rarely (if ever) mention the confidence level?

Example: Hagel and Nelson 1996 Nebraska senate race

This article appeared in the Omaha World-Herald shortly before the November 1996 election. The Omaha World-Herald had the Gallup Organization take a poll to estimate the proportion of people who would vote for the candidates. Focus on the part entitled “Here’s How W-H Poll Was Conducted”.

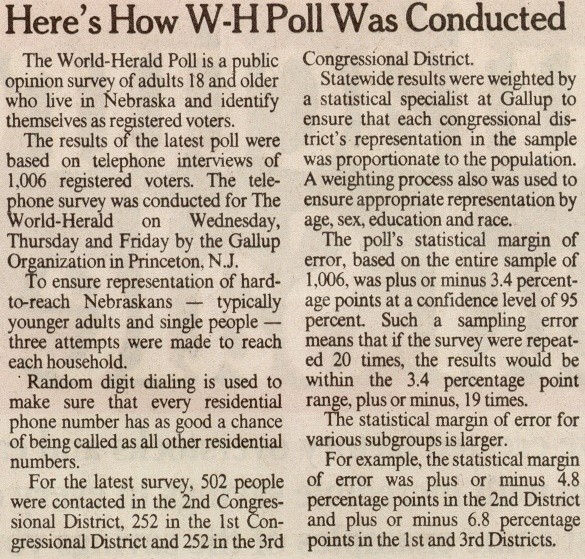
Notes:

* Hagel won the 1996 election. Notice he was “behind” 49% to 44% in late October.
* Nelson won an election in 2000 for the other Nebraska seat in the senate.









From using our formulas and “part” of the information given here (α = 0.05, E = 0.034),



Why isn’t this 1,006? Without more information, I do not have the exact answer. Most likely, it is due to the extra variability which would be caused by sampling a different number of people from the 3 congressional districts.

Final comments:

* Internet polls
* Scientific vs. non-scientific polls