<u>Electoral Polls and Statistical Confidence</u>

Jorge Luis Romeu, Ph.D.
Emeritus, State University of New Yor
201 Rugby Rd. Syracuse NY 13203
Email: romeu@cortland.edu

November 5, 2019

The Electoral Season always brings up the *vote intention* polls or surveys that candidates, pundits and journalists use, and sometimes misuse to assess who is ahead in each political contest. We often hear or read that "candidate X is ahead, with 47% of vote intention, to only 45% for his opponent". Such statement is statistically incomplete and incorrect. In addition, there are other technical complications.

For example, many voters use cell instead of land phones, often unlisted and difficult to reach by surveys. Others use answering machines and caller Id devices as screening mechanisms, to identify callers before deciding to take their call. This limits access to many potential survey subjects, something that can seriously invalidate or bias, the survey findings.

Such problems have resulted in some polls grossly overstating or understating the status of an election. Recent cases include the US 2016 presidential election, and the British referendum on BREXIT. Let's review the basic premises upon which polls are based.

Confidence, the internal conviction that some situation or statement will hold, is one of them. For example, we may have 90% confidence that it will rain today, because in 90 out of the last 100 days, with the same humidity conditions as today, it has rained.

Another is *randomness.* In polls and surveys we want to *estimate the population* vote intention proportion or percent, for a given candidate, *based upon a sample* of voters. But samples vary (we take one sample, somebody else may take a different one). And hence, so do their results. Therefore, instead of giving a single value for the percentage vote intention, we derive a *confidence interval*.

Confidence intervals define a range of values where the population percent vote intention will be. It is based upon a center value, plus/minus a *margin of error*. For example, say a *confidence interval* for vote intention, for one candidate, may be [44% to 50%].

Such range comes from a center value 47%, plus or minus 3%, yielding 47-3=44% to 47+3=50%. The 3% is the *margin of error*, and depends on (1) the *sample size* (how many random voters have been interviewed for the poll) and (2) of the *coverage probability* or the percentage of times we assume that the poll will be right.

For example, we say that "candidate X has a vote intention of 47%, with a margin of error of 3%, for a coverage probability or confidence of 95%". This means that, 95% of the times that we take a poll or survey for voter intention, the population percentage for this candidate will be between 44% and 50%. This is the correct and complete way to state the results of a voter intention poll or survey.

Poll results are valid if the sample of voters was taken randomly from *all possible voters*. If sample was not taken randomly, or some voters were not considered into this sample, poll results are invalid.

Such occurred during the 1936 presidential election. A legendary poll result stated that FDR would be unquestionably defeated (which he was not). The technical problem was that the pollsters took a telephone-based survey, at a time when only wealthy people could afford to own a phone. And wealthy people hated FDR with a passion. The poll results inflated the percent of wealthy voters in the population. Results were wrong because the sample was not *representative* of the voter population. Poor subject access to the survey can seriously invalidate (or bias) its results.

With the advent of cellular phones, caller ID, and answering machines much has changed in the world of phone banks and phone polling. Before, most phones were listed in the telephone book, and almost everyone owned one. Therefore, taking a random sample from the phone book provided a good representation of the population. Today, many voters use unlisted cell phones and caller Id devices as screening mechanisms. Hence, they can't be easily reached by the survey

Now, consider the problem of comparing voter intentions between two candidates. For the previous example, one candidate got 47% voter intention from the poll. Assume that the opposite candidate got 45%. The difference between the two (47-45=2) is within the *margin of error of* 3%. The results constitute *a statistical tie*. It is incorrect, in such case, to state that one candidate is ahead of the other.

In some polls the margin of error or the coverage probability are not always disclosed. The above example underlines the importance of always requesting the margin of error and the coverage probability, with a poll result:

When the election is very close and polls are unable to distinguish which candidate is really ahead we need to take a larger sample -or to wait until Election Day, to find out!

<u>Note</u>: Dr. Jorge Luis Romeu is a professional statistician with 40+ years of experience in teaching, research and consulting. He retired Emeritus from SUNY Cortland, worked as Senior Engineer, for IIT Research Institute, in the Reliability Analysis Center, Rome Lab, and was, for 16 years, a Research Professor, at Syracuse University. He currently teaches part time graduate statistics courses at SU.

Technical Appendix

Such margins of error are always possible because the samples drawn are "large" (greater than 20). For example say that a poll has been taken over a sample of size n=1064. Then, the sample proportion "p" is approximately Normal, and the *margin of error* is:

$$z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 1.96\times\sqrt{\frac{0.47\times 0.53}{1064}} = 0.03; or .3\%$$

Therefore, a *95% Confidence Interval* (CI) is the *survey proportion +/- margin of error*:

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.47 \pm 1.96\times\sqrt{\frac{0.47\times 0.53}{1064}} = 0.47 \pm 0.03$$

Using the Normal Standard percentile (1.96) we obtain a confidence of 95%.

Thefore, to be complete and informative, and possibly just valid, any poll result must include (a) the margin of error (say ± 3%) and (b) the confidence (say 95%).

The confidence depends of the random sample size (n) taken in the poll, as well as the quality of the data. The "poll precision" or margin of error (H) depends on such size. For example, for a confidence of 95%, a 3% margin of error (H = 0.03) and a true but unknown voter intention p = 0.47, we would require a sample of size n = 1064::

$$n = \frac{(z_{\alpha/2})^2\,\bar{p}(1-\bar{p})}{H^2} = \frac{1.96^2 * 0.47 * 0.53}{0.03^2} = 1063.27 \approx 1064$$