Le, Kien T., Abdoulaye Diop, and Darwish Al-Emadi. 2013. "Practical Guidance on Calculating Sampling Error in Election Polls." *Survey Practice* 6 (1). https://doi.org/10.29115/SP-2013-0006.

ARTICLES

Practical Guidance on Calculating Sampling Error in Election Polls

Kien T. Le[1], Abdoulaye Diop[1], Darwish Al-Emadi[1]

[1] Social and Economic Survey Research Institute, Qatar University

Keywords: sample designs, sampling error, election polls

https://doi.org/10.29115/SP-2013-0006

Survey Practice

Vol. 6, Issue 1, 2013

In election poll results, the proportions favoring candidates and the survey sampling error are usually reported. However, it is hard to assess if the gap between any two candidates is statistically significant or not based on this information. This article provides an alternative measurement of sampling error for this assessment purpose. We detail the calculation steps in STATA and SPSS programs to handle polls based on simple random sampling and also polls based on more complicated designs.

Introduction

Sampling error (also known as the margin of error) is the error caused by observing a sample instead of the whole population (Groves 2009, 56). In election polls based on simple random sampling (SRS), the sampling error at the 5 percent significance level is calculated with this expression: $1.96\sqrt{Var/\mathrm{n}}$ and $Var=p(1-p)$ where Var is the variance, n is the number of respondents, and p is the proportion favoring a candidate. When p is evaluated at 50 percent, this sampling error achieves the highest value at $1.96 \times 0.5/\sqrt{\mathrm{n}}$. In practice, most polls are based on more complicated design than the SRS design (e.g., use of clusters, strata, multiple stages, and weights). Consequently, the calculation of the error has to take into account the sample design and is usually larger than the one with SRS.

This sampling error, either with SRS or with a more complicated design, is usually reported in election polls. However, it is unclear how this reported error together with the proportions favoring candidates can be used to assess the gap between any two candidates. The following is an excerpt from a report on a Gallup survey of Republican registered voters that appeared on the CNN website in December 2011[1]:

> Gingrich now has the backing of 29% of Republican voters ... Meanwhile, his closest competitor, Mitt Romney, has made modest gains in the last week, up to 24% ... with a sampling error of plus or minus four percentage points.

In this excerpt, the gap between the two Republican candidates is 5 percentage points (=29 percent−24 percent) while the reported sampling error is 4 percentage points. Some people may come to a conclusion that this gap is

1 http://politicalticker.blogs.cnn.com/2011/12/15/gallup-poll-unsteady-december-for-gop-field/. As you may know, Mitt Romney eventually won this race for the Republican n

statistically significant since it is larger than the error. By contrast, others may come to an opposite conclusion that this gap is not statistically significant since it is less than two times the error (2×4 %percent = 8 percent).[2] The latter conclusion is based on the overlap between the confidence intervals for the two candidates. The confidence interval, which is calculated by adding and subtracting the reported sampling error from the proportion favoring a candidate, ranges from 25 percent to 33 percent for Gingrich and from 20 percent to 28 percent for Romney, so there is an overlap from 25 percent to 28 percent between them.

The logic behind both of these conclusions is incorrect because the reported sampling error can only be used to compare one proportion favoring a candidate to a specific number. It cannot be used to compare one proportion to another proportion since both proportions have sampling variations. Note that the proportions favoring the two candidates are from the same sample and are highly correlated with each other. If the proportion favoring one candidate increases, the proportion favoring the other is likely to decrease.

In order to correctly assess the gap between two candidates, it is necessary to calculate the sampling error of the gap itself, and there are statistical textbooks that provide this kind of sampling error. For example, Daniels, McKean, and Kapenga (2001) offer this formula for the calculation: $1.96\sqrt{Var/\mathrm{n}}$ and $Var = p1 + p2 - (p1 - p2)^2$ where $p1$ and $p2$ are, respectively, the proportions of respondents favoring candidate 1 and candidate 2.[3] Unfortunately, this formula only applies to polls with SRS while most polls in practice are based on sampling that differs from SR and it is not clear from these books how to calculate the sampling error in this case.

This article presents a simple way to calculate this sampling error. In polls based on SRS, our calculation yields the same formula as the one presented in the textbooks. In polls with sample designs different from SRS, there is no simple formula and the calculation has to be performed with the help of a statistical software. In this case, we detail the calculation steps in STATA and SPSS programs.

Calcuation

Suppose we have a poll of three candidates. The respondents' choices of the candidates are shown in the column 2 of Table 1. Assume that candidates 1 and 2 are the top two candidates, and we want to assess if there is any statistically significant gap between them. In order to calculate the sampling error for the gap, we add column 3 for candidate 1. The values in this column are 1 if

2 See http://blogs.wsj.com/numbersguy/whats-a-statistical-tie-anyway-234/.

3 This formula can be found at http://www.stat.wmich.edu/s160/book/node64.html.

Table 1 Hypothetical poll data.

Respondent ID (1)	Respondent choice (2)	Added column for candidate 1 (3)	Added column for candidate 2 (4)	Difference between candidates 1 and 2 (5)=(3)−(4)
1	Candidate 1	1	0	1
2	Candidate 3	0	0	0
3	Candidate 2	0	1	−1
..	.....	...	...	...
N	Candidate 1	1	0	1

respondents select him and 0 otherwise. In a similar way, we add column 4 for candidate 2. Finally, column 5 represents the difference between columns 3 and 4.

Since column 5 is the difference between the two candidates, the variance or standard error of this column can be used to derive the sampling error between the two candidates with this expression: $1.96x\sqrt{Var/\mathrm{n}}$ where *Var* is the variance of the last column. Most election polls employ a certain sample design that may include the use of clusters, stratification, multistage sampling, or combination of some or all of these designs. Also, these designs usually call for the use of sample weights to correct for probability of selection and non-response rate. In addition, they include post-stratification weights to adjust for some demographic characteristics.

In any of these designs, the variance of the last column can be calculated with the help of statistical software packages such as STATA and SPSS (with the complex sample module). In these packages, one first needs to set up the sample design so that STATA or SPSS knows what kind of design is used in the poll. Then, one simply requests the variance of the last column. The following is an example of the calculation in STATA for an election poll based on the following sample design: the sample is stratified by region and then in each *region* cluster sampling is applied where cluster is the geographical *area* (e.g., village, district or county). The sample *weight* is calculated to account for selection probability and non-response. Using STATA, we can calculate the variance for the last column as follows:[4]

> **svyset** *area* [**pweight** =*weight*], **strata** (*region*)
> **mean** *last_column*

where *area* is the name of the cluster variable, *weight* is the name of the sample weight variable, *region* is the name of the stratum variable, and *last_column* is the variable name of the last column in Table 1.

4 STATA is able to understand most sampling designs used in practice including multistage sampling. The variance is calculated in STATA by one of the following methods: boot

In this example, the first command " **svyset** " is used to let STATA know the sample design used in the poll while the second command " **mean** " requests the standard error (which is $\sqrt{Var/\mathrm{n}}$) of the last column. By multiplying this standard error with 1.96, we get the sampling error between the two candidates at the 5 percent significance level.[5] By comparing the gap between the two candidates to this sampling error, one can know if the gap is statistically significant or not. For example, in the excerpt from the CNN website, the gap between the two candidates is 5 percentage points. If the sampling error from our calculation is <5 percentage points, one can conclude that the gap between the two candidates is significant. By contrast, if the sampling error is more than 5 percentage points, one can arrive at the opposite conclusion.

In the case of SRS, it is not difficult to show that the variance of the last column in Table 1 is simply $p1 + p2 - (p1 - p2)^2$ which is also the formula mentioned above (in Daniels, McKean, and Kapenga 2001) for election polls with SRS design.[6] Since the last term $(p1 - p2)^2$ is usually very small compared to the first two terms, this variance can be reduced to $p1 + p2$ and so the sampling error can be derived with this expression: $1.96x\sqrt{(p1 + p2)/\mathrm{n}}$. This is an interesting case since we can directly compare this sampling error to the commonly reported one. If we choose to compare two top candidates in the poll then $p1 + p2$ is likely >25 percent. Also, $p1 + p2$ is always <1 since even in polls with only two candidates, some respondents would select "undecided." Therefore, our sampling error in SRS is likely to range from $1.96x0.5/\sqrt{\mathrm{n}}$ to $1.96\sqrt{\mathrm{n}}$. The lower bound is the reported (maximum) sampling error while the upper bound is twice the reported error. That means conclusions based on comparing the gap between two candidates to the reported error or twice the reported error could be quite misleading.

Conclusion

This article presents a simple way to assess the gap between the two candidates in election polls. The assessment is based on calculating the sampling error for the gap between the two candidates, and the calculation can be used in polls based on any sampling design (either SRS or more complicated sampling). By comparing the gap between the two candidates to this sampling error, one can instantly know if one candidate is really ahead of the other.

---

5 For SPSS users, you first need to add the Complex Sample Module in order to correctly calculate the variance. Next, the sample design can be set up by selecting "Analyze\Complex Samples\Prepare for Analysis ..." from the menus. Follow SPSS instructions to declare region as the stratum variable, area as the cluster variable, and weight as the sample weight variable. Finally, the standard error for the last column can be calculated by selecting "Analyze\Complex Samples\Descriptive ..." from the menus.

6 To derive the variance of the last column in the SRS, we start with this standard formula: $Var = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$ where xi is the value in the last column of Table 1 and $\bar{\mathrm{x}} = \sum_{\mathrm{i}} \mathrm{x_i}/\mathrm{n}$ is the mean value of the last column. By looking at the last column, we can see that the first term in the variance $\left(\frac{1}{\mathrm{n}}\sum_i \mathrm{x_i^2}\right)$ is simply $p1 + p2$ while the second term $(\bar{\mathrm{x}}^2)$ is $(p1 - p2)^2$. So the variance is $p1 + p2 - (p1 - p2)^2$.

REFERENCES

Daniels, A.J., J.W. McKean, and J.A. Kapenga. 2001. *Statistics and Data Analysis*. Kalamazoo, MI: Western Michigan University. http://www.stat.wmich.edu/s160/book/book.html.

Groves, Robert M. 2009. *Survey Nonresponse*. Edited by Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little. Hoboken, NJ: Wiley Press. https://www.amazon.com/Survey-Nonresponse-Robert-M-Groves/dp/0471396273.