

Sampling Errors in Political Polls

KEYWORDS

Teaching
Proportions
Sample size

Zbigniew Kmietowic

University of Leeds, Leeds,,England

Summary

This article examines, the sampling error of the lead of one political party over another as observed in a random sample of voters... The sample size needed to achieve a certain precision is also investigated..

◆INTRODUCTION◆

IN HIS article “Political Polls and Errors” published in “Teaching Statistics” in 1992, Harding rightly points out that results of political polls shown on television are often misleading. The same applies, by and large, to reports appearing in daily and weekly newspapers, including the quality papers. His main point is that if the standard error is included among the poll results, it is not indicated that it applies only to a single sample proportion (percentage) of electors supporting a particular political party. Thus an impression is created that standard errors of other sample proportions and sample leads of one party over another (differences between two proportions or percentages) are the same. Harding explains that the standard error given in his example refers to the largest sample proportion (for which the error is largest), and demonstrates that for smaller sample proportions the error is smaller. Thus the standard error given overstates sampling errors of the smaller sample proportions. It would have been better if standard errors had been given for individual sample proportions. As Harding also points out, the situation with respect to the standard error of the sample lead is quite different. Although he does not say so, he implies that the standard error of the sample lead is larger than the standard errors of the two sample proportions defining the lead. This is in fact the case, but Harding does not pursue the point, stating simply that the calculation of the standard error for the sample lead is more complicated.

The purpose of this article is to show that the standard error of the sample lead can be readily calculated, and can be used to obtain the confidence interval for the population lead, test hypotheses about the

population lead and calculate the required sample size for a given accuracy of the sample lead and risk level. These procedures are illustrated with numerical examples using data provided by Harding. The results an illustrations should make the analysis and interpretation of political polls better informed and more interesting. It should be pointed out however, that the results presented here assume that simple random sampling methods are used in collecting the information and that non-response does not arise. Usually more complex sampling methods are used in practice and non-response is invariably present. Some of these issues are discussed later.

◆ EXPECTED VALUE AND VARIANCE OF THE SAMPLE LEAD ◆

It may be shown, see Kmietowicz (1990), that if a random sample of n electors is taken from a very large electoral register, the expected value and variance of the sample lead of one party over another (difference between two sample proportions, which must not be confused with the difference between two proportions obtained from two independent random samples) are given by

$$E(p_x - p_y) = P_x - P_y \quad (1)$$

and

$$\text{var}(p_x - p_y) = (P_x Q_x + P_y Q_y + 2P_x P_y) / n \quad (2)$$

where p_x is the proportion of electors in the sample supporting party x , p_y is the proportion of electors in the sample supporting party y , P_x and P_y being similarly defined but referring to proportions in the population (the electorate), $Q_x = 1 - P_x$ and $Q_y = 1 - P_y$. Note also that p_z is the proportion of electors in the

sample supporting other parties, P_z is the corresponding proportion in the population and that $p_x + p_y + p_z = 1$ and $P_x + P_y + P_z = 1$

Result (1) shows that the sample lead ($p_x - p_y$) is an unbiased estimator of the population lead $P_x - P_y$.

Result (2) shows that $\text{var}(p_x - p_y)$ is larger than both $\text{var}(p_y)$ and $\text{var}(p_x)$ and also larger than their sum, i.e. $\text{var}(p_x) + \text{var}(p_y)$. Thus the sample lead is subject to larger sampling errors than the two sample proportions which define it.

It can be readily shown that if Q_x and Q_y are replaced in result (2) by $(1 - P_x)$ and $(1 - P_y)$ respectively, we obtain an equivalent expression in terms of P_x and P_y alone which is more convenient for some analytical and computational purposes, i.e.

$$\text{var}(p_x - p_y) = [P_x + P_y - (P_x - P_y)^2]/n. \quad (3)$$

It can also be seen from result (3) that for given n , $\text{var}(p_x - p_y)$ will be small if P_x and P_y are both small; the limiting value being zero, which will arise when $P_x = P_y = 0$ and $P_z = 1$. It may be shown that for given n and P_z , $\text{var}(p_x - p_y)$ will be largest when $P_x = P_y = (1 - P_z)/2$, i.e. when P_x and P_y are equal and as large as possible. Substituting these values in result (3) it may be verified that maximum variance for this case is given by $\text{var}(p_x - p_y) = (1 - P_z)/n$. The limiting case arises when $P_z = 0$, and $P_x = P_y = 1/2$ giving maximum $\text{var}(p_x - p_y) = 1/n$.

These results are very simple and may be useful in practice, e.g. when P_x and P_y are unknown, and are both assumed to be close to $1/2$, variance of the sample lead is given by $1/n$. Note also that when $P_z = 0$, i.e. when $P_x + P_y = 1$, variance of the sample lead is based on the binomial distribution and is given by $\text{var}(p - q) = 4PQ/n$ (using the more common binomial notation) which the reader should be able to verify for him/herself. The reader should also be able to show that $\text{var}(p - q)$ is maximised when $P = Q = 1/2$ which gives maximum $\text{var}(p - q) = 4PQ/n = 4(1/2)(1/2)/n = 1/n$.

This agrees with the result given above. It is interesting to note that in the binomial case $\text{var}(p - q)$ and $\text{var}(p)$ are both maximised when $P = Q = 1/2$. Another interesting relationship between sampling errors of the sample lead and sample proportions can be stated in terms of their standard errors, (SE), i.e.

$$SE(p_x - p_y) \leq SE(p_x) + SE(p_y) \quad (4)$$

Result (4) states that the standard error of the sample lead cannot exceed the sum of the standard errors of the two sample proportions which define it. Result (4) is not difficult to prove and the reader

is invited to do this for him/herself. Note that as $SE(p_x)$ and $SE(p_y)$ are maximised when $P_x = 1/2$ and $P_y = 1/2$,

$$SE(p_x) + SE(p_y) = \sqrt{\frac{1}{2} \frac{1}{2} \frac{1}{n}} + \sqrt{\frac{1}{2} \frac{1}{2} \frac{1}{n}} = \frac{1}{\sqrt{n}}$$

for this case, and thus maximum $SE(p_x - p_y) = 1/\sqrt{n}$. This agrees with the result deduced above, i.e. that maximum $\text{var}(p_x - p_y) = 1/n$. The result can be used to calculate required sample size for given accuracy of the sample lead and given risk level, when values of the population proportions are not known exactly, but are both believed to be close to $1/2$. Such a deliberate overestimation of the required sample size may be desirable when non-response is expected and is assumed to be small. Then non-response may approximately balance the original overestimation of the sample size, ensuring that the sample lead is of required accuracy. Result (4) can also be used to obtain an approximate value of the standard error of the sample lead when the exact formula is forgotten. This can be done by evaluating $SE(p_x)$ and $SE(p_y)$ and adding them together, using best available estimates of P_x and P_y . If such an approximation is used to calculate the required sample size for given accuracy and risk level, the small overestimation of the required sample size may again offset some non-response. Result (4) can also be used to check the calculation of $SE(p_x - p_y)$ using result (3) when $SE(p_x)$ and $SE(p_y)$ are available (perhaps from earlier calculations).

The uses of the results mentioned above and other applications described in the following sections are based on the assumption that the sample lead is normally distributed. This is the case provided the random sample is large, (i.e. ≥ 25) and population proportions P_x and P_y are not extreme (i.e. close to 0 or 1). Moreover, an estimate of the variance of the sample lead can be obtained by using results (2) or (3) and replacing the population proportions by their sample estimates.

◆ CONFIDENCE INTERVAL ◆ FOR THE POPULATION LEAD

Data provided in Harding (1992) will now be used to calculate the standard error of the sample lead of the Labour Party (x) over the Conservative Party (y) and the 95% confidence interval for the population lead. Assuming that Harding's data came from a simple random sample chosen from a very large

population of electors, we have using result (3):
 $SE(P_x - P_y) = \sqrt{[.44 + .38 - (.44 - .38)^2]/1087} = .0274$.

Note that the population proportions in result (3) were replaced by their sample estimates. The approximate 95% confidence limits for the population lead ($P_x - P_y$) are:

$$(p_x - p_y) \pm 1.96SE(p_x - p_y) = (.44 - .38) \pm 1.96(.0274) = .06 \pm .0537$$

where 1.96 is the 97.5 percentile of the standard normal distribution. As the confidence interval does not include zero, the null-hypothesis claiming that $P_x - P_y = 0$, can be rejected at the 5% significance level, i.e. the test indicates that the Labour Party has a lead over the Conservative Party in the population at the 5% significance level.

Note also that the sampling error of the lead is $1.96SE(p_x - p_y) = 1.96(.0274) = .0537$, while the sampling error of the proportion of electors supporting the Labour Party, is:

$$1.96SE(p_x) = 1.96\sqrt{.44 \times .56/1087} = .0295$$

(which is larger than that for the Conservative Party), i.e. the first sampling error is 82% larger than the second ($.0537/.0295 = 1.82$). The same applies to the comparison of standard errors. Thus the sampling error of the lead is 82% larger than the sampling error of the proportion supporting the Labour Party, and to imply (as some TV and newspaper reports do) that they are the same is very misleading.

Note also that the above results satisfy the inequality given in result (4), as $.0274 < .0151 + .0147 = .0298$. Moreover, the maximum possible value of the standard error of the sample lead is given by

$$1.96SE(p_x - p_y) = 1/\sqrt{n} = 1/\sqrt{1087} = .0303$$

which arises when $P_x = P_y = 1/2$. and is slightly larger than .0298 in this case.

◆TEST OF HYPOTHESIS ◆

As an illustration of the use of the results obtained in Section 2 to test a hypothesis, let us assume that at the last local elections, the Labour Party lead over the Conservative Party was 10% (Labour Party support standing at 46%) and that subsequent polls have indicated a decline. Do the results of the poll given in Harding (1992) support the hypothesis that the Labour lead has declined?

Here the null-hypothesis is $H_0: (P_x - P_y) = 0.10$ and the alternative hypothesis is $H_1: (P_x - P_y) < 0.10$ where x and y are defined as in Section 3. Note that this is a one tail test because the direction of change in the lead is indicated. Using result (3) and the population parameters underlying the null-

hypothesis, we have:

$$\text{var}(P_x - P_y) = [.46 + .36 - (.46 - .36)^2]/1087 = .0007452$$

and, therefore, $SE(p_x - p_y) = .0273$.

The value of the standardised normal variable (the test statistic) is

$$Z = [(p_x - p_y) - (P_x - P_y)]/SE(p_x - p_y) = [(.44 - .38) - (.46 - .36)] / .0273 = -1.465$$

The critical value of Z for a one-tail test and the 5% significance level (say) is 1.6449. As $1.465 < 1.6449$ H_0 should not be rejected in favour of H_1 at the 5% significance level, i.e. the decline in the lead of the Labour Party over the Conservative Party since the local elections is not significant at the 5% level.

◆REQUIRED SAMPLE SIZE◆

Results (2) or (3) can also be used to find the required sample size which will ensure that the standard error of the sample lead is equal to a specified amount, e.g. 1%. Suppose that when the poll reported in Harding (1992) was planned, it was believed that $P_x = P_y = .40$. How large a sample should have been taken to ensure that $SE(p_x - p_y) = .01$? Using result (3), we have $\sqrt{[.40 + .40 - (.40 - .40)^2]/n} = .01$ which gives $n = 8000$. Thus the sample should have been more than 7 ($8000/1087 = 7.36$) times larger to achieve the required accuracy.

An alternative approach to the determination of the required sample size is to make the sample sufficiently large to ensure that a given difference between the sample lead and the population lead will be significant in the corresponding test of hypothesis at a given level of significance. Continuing with the previous example, let us find the minimum sample size which would have ensured that a sample lead of 3% would have indicated a lead of one party over another in the population, except for the 5% risk that this might not have been so. The required sample size could have been obtained by solving the following inequality which is based on the standard normal deviate used in the test of hypothesis above, i.e.

$$\frac{(p_x - p_y) - (P_x - P_y)}{\sqrt{\{P_x + P_y - (P_x - P_y)^2\}/n}} \geq 1.96$$

As $(p_x - p_y) = .03$, $P_x = .40$ and $P_y = .40$, the inequality becomes

$$(.03 - 0)/\sqrt{[.40 + .40 - (.40 - .40)^2]/n} \geq 1.96$$

which gives $n \geq 3415$. Thus if a sample of 3415 electors had been taken, a sample lead of 3% would have indicated a significant lead of the Labour Party over the Conservative Party, except for the 5% risk that

this might not have been so.

Note that if sampling is from a finite population without replacement, the finite population correction, $(N-n)/(N-1)$, where N = size of the population, should be used, i.e. results (2) and (3) should be multiplied by the correction. Such an adjustment reduces the variance and the standard error of the sample lead, as well as the required sample size. As $0 \leq (N-n)/(N-1) \leq 1$ the smaller is the finite population correction, the larger is the reduction.

◆CONCLUDING REMARKS ◆

The results given above can be extended to stratified sampling. They provide an unbiased estimate of the population lead and the standard error. They can also be adapted to deal with overlapping classifications. In a political poll, an elector can vote for only one political party, but in an industrial survey, a household can buy products made by more than one company. The variance of the sample lead of the proportion of households buying products of one company over another for this case can be obtained by adapting the results presented above. Details of both extensions can be found in Kmietowicz (1990).

The field of possible applications can also be extended. It was mentioned above that the sample lead of one company over another can arise in an industrial survey. Other examples include the sample lead of one product over another in market research surveys, one programme over another in TV viewing and radio listening surveys, one class over another in social investigations, etc.

As was mentioned earlier, Harding was right to point out that the sampling error of 3% given in the TV discussion of results of a political poll applies only to the largest sample proportion, but not to the other sample proportions or the sample lead, as seemed to be implied. In fact, the sampling error of the lead for this particular poll was nearly 5.4%, assuming the data came from a simple random sample. Thus the sampling error of the lead was 82% higher than the sampling error of the proportion of electors supporting the Labour Party (2.95%). Moreover, the sampling error of 3% (more exactly 2.95%) shown on TV applies only to the sample proportion of Labour Party supporters (44%). As Harding's calculations show (and I hope I am interpreting them correctly), the sampling error declines as sample proportion falls, e.g. it is 2.1% for the Liberal Democratic Party and only 0.8% for the Green Party.

Finally there is the question of the unusual sample size (1087). It is probably due to non-response which

was not included on the TV data panel quoted in Harding (1992). If the planned sample size was 1100, say, non-response could have easily brought it down to 1087. If an elector refuses to declare his/her voting intention, he/she should not be classified as voting for other parties or as an undecided voter. If this is done, proportions of voters supporting other parties or being undecided will be overestimated (biased upwards), while proportions of the individually named parties will be underestimated (biased downwards). The usual practice in such cases is to exclude non-respondents and to base the estimates on respondents alone. Such a procedure results in the reduction of the sample size, but produces unbiased estimates of the proportions for the respondents. This does not mean that the estimated proportions are necessarily unbiased estimators of the corresponding population proportions. This would be the case if non-response was random. Unfortunately this is not likely. Some non-respondents may have good reasons for refusing to disclose their voting intentions, e.g. a trade unionist supporting the Conservative Party may refuse to declare his voting intention if he does not want his colleagues to know about it, especially if the interviewing is conducted in front of other workers. Such a behaviour will produce a downward bias in the estimate of the proportion of voters supporting the Conservative Party in the electorate. If the number of non-responding electors was given, it might be possible to estimate the bias and to remove it from the original estimate, e.g. if it was known from previous studies that 60% of non-respondents normally vote Conservative, a corrected estimate of the proportion of electors supporting the Conservative Party could be obtained. Some information about voting intentions of non-respondents may be obtained by comparing actual results of elections with predictions obtained from polls taken just before an election, provided, of course, a very late change in voting intentions does not occur.

Acknowledgement

I should like to thank the anonymous referee for helpful comments.

References

- Harding D. (1992), Political Polls and Errors
Teaching Statistics, **14**(2), 6.
Kmietowicz Z. W. (1990), The significance of the
Lead in Opinion Polls, *Journal of Applied
Statistics*, **17**(1), 9-30.