THE UNIVERSITY OF TEXAS AT AUSTIN
SCHOOL OF INFORMATION

MATHEMATICAL NOTES FOR LIS 397.1
INTRODUCTION TO RESEARCH IN
LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys

Last revised: 2004 Feb 2

PUBLIC OPINION POLLING
AND
SAMPLE-SIZE DETERMINATION

Public opinion polling, such as the Gallup Poll, is an especially prevalent example of the process of making inferences, from observations on a sample, about parameters in a population. Often the parameter of interest is the proportion of the population that will vote for a certain candidate.

(Note: It is important to bear in mind that here we are using "population" in its technical sense: the set from which the sample is drawn and about which inferences can be made. Bear in mind also that a fundamental problem in sampling is to ensure that the sample drawn is from the desired population. In political opinion polling, for example, pollsters face a major problem in trying to identify which of those persons interviewed are in fact not only registered for voting but also likely to actually vote, so that the sample can be made truly representative of the population of *voters*--not merely representative of adult citizens in general, or even merely of adult citizens who have registered but will not necessarily vote.)

Political polls frequently aim at questioning about 1,500 persons who intend to vote. Nationwide polls try to draw their sample of around 1,500 from all parts of the country and from all socioeconomic classes. We shall ignore the obvious, and severe, problems of identifying and interviewing the sample, and simply look at the inference process.

In a poll of preferences for candidates in a two-candidate contest, the goal is to estimate the proportion, π, of voters who will vote one way, and hence also the proportion, $1-\pi$, of voters who will vote the other way. Pollsters do this as you would expect, by using the relative frequency with which interviewees favor one of the candidates as the estimate, P, of that candidate's likely proportion, π, of the vote. But, unfortunately, the results of such polls are often reported in the press in terms of point estimates, not as confidence intervals (CIs). You can readily see, for example, that if the point estimate of Candidate A's vote is 51% = P but the 95% CI for π is, say $.48 < \pi < .54$, then A could easily have as little as 48% of the potential vote and hence would by no means be really assured of success.

Although polls are now reported with a statement of their probable imprecision more frequently than in the past, you still often see polls reported without such a statement. To fix the imprecision involves both the *standard error (SE) of the sample proportion* and also the *confidence factor* corresponding to the level of confidence that is desired for the statement that is to be made about the probable range of values for the population proportion π. For conven-ience, we shall use the factor corresponding to 95% confidence, so that the approximate CI will be:

$$\text{sample proportion} \pm 2(\text{SE})$$

Note that the width of this CI is 4(SE) or, equivalently, that its half-width is 2(SE). The half-width is often called the *margin of probable error*, and we shall use PE to represent it. Thus,

PE = 2(SE)

For the kind of problem exemplified by a two-candidate race or a simple for-or-against issue--i.e., for a proportion in a *binomial*, or two-category, population--it can be proved (though we omit the proof here) that if we have taken a sample of size *n* and have found a relative frequency of $k / n = P$ of votes for Candidate A, then the best estimate $s_{\overline{P}}$ of the standard error, SE, of the sample proportion is

$$s_{\overline{P}} = \sqrt{\frac{P(1-P)}{n}}$$

This means that the half-width of the CI, or margin of probable error, PE, is

$$PE = 2(SE) = 2s_{\overline{P}} = 2\sqrt{\frac{P(1-P)}{n}}$$

from which, by squaring both sides, we find that in general

$$(PE)^2 = 4\frac{P(1-P)}{n}$$

From this, first multiplying both sides by n and then dividing both sides by $(PE)^2$, we find that

$$n = 4\frac{P(1-P)}{(PE)^2}$$

Note that the sample size needed is inversely proportional to the *square* of the desired margin of error.

In political polling *P* and 1-*P* are typically fairly close to each other, and hence to 1/2. Hence, it will be satisfactory to use the approximation

$$\sqrt{P(1-P)} = \sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)} = \sqrt{\frac{1}{2}\times\frac{1}{2}} = \frac{1}{2}$$

With this approximation, the margin of probable error is

$$PE = 2\sqrt{\frac{P(1-P)}{n}} = \frac{2\sqrt{P(1-P)}}{\sqrt{n}} = \frac{2\sqrt{\frac{1}{2}\times\frac{1}{2}}}{\sqrt{n}} = \frac{2\left(\frac{1}{2}\right)}{\sqrt{n}} = \frac{1}{\sqrt{n}}$$

Hence, in a close race, the margin of probable error in the reported point estimate P derived from interviews with 1,500 intended voters is

$$PE = \frac{1}{\sqrt{1500}} = \frac{1}{38.73} = 0.0258 = 2.58\%$$

Suppose pollsters wanted to bring their margin of error (measured in terms of a 95% confidence interval) down to 1%. The size *n* of the sample necessary to do this would be given by

$$1\% = .01 = \frac{1}{\sqrt{n}}$$

from which it follows that

$$\sqrt{n} = \frac{1}{.01} = 100$$

and hence that n = 10,000. To sample 10,000 intended voters would be much more expensive and, perhaps worse, much more time-consuming, than sampling only 1,500. You can see why the pollsters, usually pressured by both money and time, are often willing to settle for around 1,500 interviews, with which their probable error is, after all, less than 3%. $\pm$

Note. It is important to recognize that the process being employed here to estimate sample size is usable only in the special circumstance of a *binomial* variable whose probability *P* of success can be assumed to be *close* to 0.5. In other words, this shortcut method of determining sample sizes *cannot* be applied when *P* is *not* close to 0.5.

In presidential elections, the polling (like the voting) is complicated by the Electoral College. If the approximately 1,500 interviews often used in opinion polling were distributed among the various states so as to reflect each state's proportion of the 538 votes in the Electoral College, then, since Texas has 32 electoral votes, approximately (32/538)(1,500) = 89 interviews would be taken in Texas. If only these 89 interviews were used to estimate the proportion of the Texas vote going to Candidate A, the associated probable error would be

$$PE = \frac{1}{\sqrt{89}} = \frac{1}{9.43} = 10.6\%$$

Hence, from such a poll's results Candidate A should not feel reasonably assured of carrying Texas unless his or her proportion of the Texas vote in the 89 interviews was at least P = 60.6%. For if it was something less, say 60%, then the 95% confidence interval, .600$\pm$ 0.106 = (.494, .706), would include values less than .50. To put it another way, if the *true* proportion of Texas voters favoring Candidate A was only, say, 49.5%, it could still happen rather easily that in a sample of only 89 interviewees the proportion favoring A would be as high as 60%.

Because of these complications, opinion polling in presidential elections usually involves more than 1,500 interviews; and the states with large electoral votes are given special attention, to reduce the margin of error in these states.