Opinion Polls and Statistical Surveys: What They Really Tell Us

Rajeeva L Karandikar and Ayanendranath Basu

In recent times we seem to be having frequent parliamentary elections in addition to lots of assembly elections. Even as this article is being written, we are heading for another parliamentary election. At the same time, exit polls and opinion polls are gaining popularity. The media often reports the results of opinion polls predicting the seat shares of major parties before the actual elections; even exit polls are now commonplace. All major national and state elections in the recent times have been covered by exit polls. The agencies conducting these polls now recruit the services of highly qualified experts for this purpose.

At the same time, however, the general public of our country are largely unaware of the scientific issues involved in a statistically planned opinion survey (or sometimes even what it means). This leads to fantastic claims and counterclaims from politicians (depending on whether their parties are projected to win or lose); some statements go as far as asserting that opinion polls mean nothing and there is no science behind them. While the analysis of exit polls at the Doordarshan Channel was going on after the 1998 parliamentary elections, an otherwise respected political leader claimed that any opinion based on a sample of size 26000 in a country of 26 crore voters cannot give any meaningful information. It is true that no survey, not even the most optimally designed one, can predict the true state of nature with *complete* confidence, and there is always a degree of uncertainty involved. However, when the confusion reaches such a stage where a layman freely questions the findings of a scientific study, it becomes the duty of the scientific community to justify the merits and scope of such techniques. Can a sample of 26000 really lead to any meaningful conclusions in a country so large and diverse as India?

The first step in an exit/opinion poll (or for that matter

Ayanendranath Basu is with the Indian Statistical Institute, Calcutta. He received the MStat. degree from the Indian Statistical Institute in 1986 and the PhD degree from Pennsylvania State University, USA in 1991. After spending four years at the University of Texas, he returned to the Indian Statistical Institute in 1995. His research interests include robust estimation, shape estimation, and HIV modelling.

Rajeeva Karandikar is with the Indian Statistical Institute, New Delhi. He received his PhD from the Indian Statistical Institute in 1982. His research interests include stochastic processes and filtering theory.

When confusion reaches such a stage where a layman freely questions the findings of a scientific study, it becomes the duty of the scientific community to justify the merits and scope of such techniques.

any other statistical survey) is the estimation of some unknown parameter. (See Delampady & Padmawar, *Resonance*, Vol. 1, No. 5 for some simple examples on parameter estimation.) While in the context of elections one can consider the unknown parameter to be the proportion of voters who intend to vote for a particular party (or alliance), indeed this is also the case for other social or economic parameters such as literacy ratios, unemployment rates, proportion of people below the poverty line, proportion of vehicle, house, TV and telephone connection owners and many such other things. If we were to simply estimate a population proportion (the proportion of votes for a given party in the election context), it can be done fairly accurately based on the information of a subset of the population – called the sample, provided the sample was large enough and 'properly' chosen (discussed in detail later). In an election, however, the main interest of the populace (and hence the media) is not in the percentage of votes for a given party (or alliance), but rather in its seat share. This makes the problem complicated, and we need further analysis than the simple statistical estimation of a population proportion.

If the entire population size is small, we can probably enumerate all the individuals and thereby determine the exact proportion of individuals of each different type making up the population. Often, however, the population of interest will be so big that complete enumeration will be totally impractical – perhaps even impossible – due to constraints on essential resources such as money, time, manpower, etc. It is in situations such as this that the idea of a statistical survey becomes relevant. In these cases, statisticians must base their estimate of the unknown proportion on a smaller fraction of the actual population – the so called sample.

In the context of the parliamentary election, if one wants to predict the winner in each constituency of Lok Sabha with a high degree of confidence, one has to sample a reasonably large number, say 500, of individuals from each of our 544 constituencies. This will make the overall sample size prohibitively large, and the entire process will become an unmanageable and time-consuming exercise even if one had

adequate financial resources. It becomes necessary therefore to restrict attention to a set of selected constituencies in some optimum way. The data obtained by sampling from these constituencies is then combined with some other information such as past voting trends and records, and the seat shares are then predicted based on a suitable mathematical model.

Since we get the actual information only from a subset of the total number of constituencies, the selection of the constituencies to be sampled is of great importance. In the United Kingdom, where the procedures for general elections and installation of new governments are very similar to our system, the idea of 'safe seats' is often used in the determination of the constituencies to be sampled. Both the Conservative and the Labour parties have large committed vote banks; voting patterns in UK change slowly, so that the constituencies with overwhelming support for one of the two parties can often be considered safe for the given party at elections in the not too distant future. For prediction purposes, one can therefore concentrate on the remaining marginal seats.

In India, the situation is quite different. Our election process involves a very large number of national and regional parties. Politics in India is largely personality-based rather than issue-based. Voter moods appear to be more easily swayed here. The political parties keep splitting and regrouping in different formations; old alliances are broken up and new alliances are created. All this causes voter loyalties to shift often and by wide margins, and there are very few safe seats. In addition, major events which have large impacts on voter's perception often have a very regional nature in a diverse country like India. All these make the use of the safe seat idea a rather shaky one in the Indian context.

We now discuss one by one the two stages of estimating the seat share of a particular party; these are (i) estimating the proportion of votes for a given party in a given region and (ii) estimating the corresponding number of seats based on a suitable model.

As the statistician will not have complete information about the population, it is improbable that the estimate offered on the basis of the sample observations will actually be exactly equal to the desired target.

This problem may also be described in terms of urns and balls – as there being an urn with a very large number of balls of k different colours, and we have to draw a sample of appropriate size to estimate the actual proportion of balls of each colour in the urn. As the statistician will not have complete information about the population, it is improbable that the estimate offered on the basis of the sample observations (whatever it may be, usually it is the observed proportion based on the sample) will actually be exactly equal to the desired target – unless that happens to be a lucky coincidence. Allowing for a margin of error, one must therefore try to minimize the error in some average sense. One of the keys to this generally lies in the technique that is employed in choosing the sample on which the statistician's inference is to be based – it must be 'representative' of the pattern of the true population. In the simplest case this is done through 'simple random sampling' (equal probability sampling) where one chooses a sample of a given size from the population in such a way that each individual in the population has the same chance of being included in the sample. This is necessary to diminish the possibility of the sample drawn being biased in the sense of being concentrated on a particular segment of the entire population. This randomness is a property of the sample collection method and not of the actual sample. It is not possible to say just by looking at the sample whether the particular sample has been chosen randomly or through some systematic procedure. In fact even the random sampling procedure may occasionally result in a biased sample, the conclusions based on which may be far off from the truth; however the randomness of the procedure ensures that this is unlikely.

We can illustrate this with the following. Consider a population of N individuals. Let B be the set

$$B = \{1, 2, \ldots N\}$$

and suppose for each $i \in B$, we have $a_i \in \{0, 1\}$. Let $M = \sum_{i=1}^{N} a_i$ and $p = M/N$. Then p is the proportion of i for which $a_i = 1, i = 1, 2, \ldots, N$. We want to estimate this unknown proportion p based on a sample of size n from B.

Let $S = \{(i_1, i_2, \ldots, i_n) : \quad i_1, i_2, \ldots, i_n \in B\}$. For each $(i_1, i_2, \ldots, i_n) \in S$, let

$$X((i_1, i_2, \ldots, i_n)) = \sum_{j=1}^{n} a_{i_j} \quad \text{and}$$

$$\hat{p}((i_1, i_2, \ldots, i_n)) = \frac{X((i_1, i_2, \ldots, i_n))}{n}.$$

Each element in S is a sample and X is the number of i in the sample for which $a_i = 1$ and $\hat{p}$ is the proportion of i in the sample for which $a_i = 1, i = 1, 2, \ldots, n$. In sampling with replacement, where any element of the population can appear more than once in the sample, the total number of samples (number of elements in S) is N^n (since for each of the n draws there is a choice of N elements to choose from).

It is not possible to say just by looking at the sample whether the particular sample has been chosen randomly or through some systematic procedure.

Here we present the following identities with brief explanations:

$$\sum_{S} [X((i_1, i_2, \ldots, i_n))] = nMN^{n-1}. \qquad (1)$$

This follows from the fact that when we consider all possible samples, each of the N members of the population will appear the same number of times among them. Since there are N^n samples in all, and there are n individuals in each sample, this number will be nN^{n-1}. Since there are M elements in B for which $a_i = 1$, the above identity follows.

$$\sum_{S} \left[X((i_1, i_2, \ldots, i_n))^2\right] = nMN^{n-1} + n(n-1)M^2N^{n-2}. \qquad (2)$$

For this identity, notice that

$$X((i_1, i_2, \ldots, i_n))^2 = \sum_{j=1}^{n} a_{i_j}^2 + \sum_{j=1}^{n} \sum_{k=1}^{n} a_{i_j} a_{i_k},$$

where $j \neq k$ in the second (double) summation. Since $a_{i_j} \in \{0, 1\}$, $\sum_S [X((i_1, i_2, \ldots, i_n))^2] = nMN^{n-1} + n(n-1)\sum_S a_{i_1} a_{i_2}$. Since there are N^n samples, and M/N of them have 1 in the first position, and M/N of them have 1 in the second position, identity (2) follows.

While it is true that the precision of the estimate increases with the size of the sample, given that the sample fraction is small, the accuracy of estimation will not depend on the population size or the value of the sample fraction.

Hence

$$\frac{1}{N^n}\sum_S [\hat{p}((i_1, i_2, \ldots, i_n))] = p$$

$$\frac{1}{N^n}\sum_S \left[\hat{p}((i_1, i_2, \ldots, i_n))^2\right] = \frac{p}{n} + \frac{n-1}{n}p^2.$$

All these lead to the result

$$\frac{1}{N^n}\sum_S \left[(\hat{p}((i_1, i_2, \ldots, i_n)) - p)^2\right]$$

$$= \frac{1}{N^n}\sum_S \left[\hat{p}((i_1, i_2, \ldots, i_n))^2\right] - p^2 = \frac{p(1-p)}{n} \qquad (3)$$

As a result of (3), we have (by applying the well-known Chebyshev's inequality)

$$\frac{1}{N^n}\#\{(i_1, i_2, \ldots, i_n) \in S : |\hat{p}((i_1, i_2, \ldots, i_n)) - p| > \epsilon\}$$

$$\leq \frac{1}{\epsilon^2}\frac{p(1-p)}{n}. \qquad (4)$$

By choosing n large enough, we can make the right hand side in (4) small, say less than 0.05 and thus ensure that in more than 95% of the samples the difference between observed proportion $\hat{p}$ and true proportion p is less than (a pre-chosen) $\epsilon > 0$.

So, if we can find a mechanism to choose one of the N^n possible samples in such a manner that each sample has equal probability of being chosen, we will get a representative sample with more than 95% probability. This is *random* sampling with replacement.

Using central limit theorem it can be shown that for large n (greater than 100 say)

$$P\{|\hat{p}((i_1, i_2, \ldots, i_n)) - p| > \frac{1}{\sqrt{n}}\} \leq 0.05, \qquad (5)$$

where the *probability* refers to the probability allocation corresponding to random sampling with replacement. (See Karandikar, *Resonance*, Vol. 1, No. 2, and Ramasubramanian,

Resonance, Vol. 2, No. 6 and No. 7 for some discussion on basics of probability theory and limit theorems.)

In sampling without replacement a particular object is allowed to appear at most once in the sample. However, when n, N are such that n/N is small (say less than 0.01), then the computations given above for the sampling with replacement case continue to be valid for practical purposes if we consider sampling without replacement. Here the set of all samples of size n is the set of all subsets of B with exactly n elements and we choose one of these as our sample with equal probability. This is *random* sampling without replacement. We will refer to the ratio n/N as the sample fraction.

Normally in problems such as estimating the percentage of voters favouring a particular party, the sample size is generally a very small (practically negligible) fraction of the true population size so that the results of sampling with replacement are applicable even when the actual sampling is without replacement. It is commonly believed that the sample size must reach a certain proportion of the true population size for a given degree of accuracy in estimation to be attained. This is incorrect. While it is true that the precision of the estimate increases with the size of the sample (see equation 4), given that the sample fraction is small, the accuracy of estimation will not depend on the population size or the value of the sample fraction. Besides the sample size n, the accuracy of the estimation process depends on the actual value of the true proportion, with true proportions closer to 1/2 being more difficult to estimate accurately than more extreme values closer to 0 or 1. In the election context, for two different constituencies where the proportions of voters favouring the candidate of a particular party are roughly the same, one will require the same sample size to estimate these proportions with a given degree of accuracy – even if the total population sizes are quite different (one may be several times the size of the other). Thus estimation of vote proportions can be done more or less equally efficiently with the same sample size, be it in a small city or in a large state like Bihar or for the whole of India! And so constituency-wise prediction and estimation of number of seats in parliament

Estimation of vote proportions can be done more or less equally efficiently with the sample size, be it in a small city or in a large state like Bihar or for the whole of India!

are mammoth tasks. The implementation procedure of the data collection for the purpose of estimating the proportion of voters favouring a particular party can then be described in the following steps: first depending upon the resources and desired accuracy, one determines the nationwide target sample. Next, the proportion of constituencies that should be sampled for the opinion poll is decided upon (say 15–20%). That many constituencies are then selected from all the constituencies via random sampling. A selected number (say 8–12) of polling booths are then randomly selected from each constituency. The nationwide sample size is divided into target sample size for each constituency. Then, from the voters' list of these polling booths, the required number of individuals are chosen at random. This then forms the random sample on which our proportion estimation methods are to be based.

A somewhat controversial issue often associated with the question of sampling is whether or not to use a quota system for different subgroups of the population. The idea here is to get fixed representations from genders, age categories, socio-economic classes, castes etc. In particular such quota surveys are popular in market researches and as a result have some application in opinion polls as well. In the election context, however, our intention is to determine the vote share based on the voting expectations of the entire population, and quota samples can possibly increase the bias of the method. We maintain that it is more appropriate to do overall random sampling than quota sampling in this context. A properly conducted opinion poll based on random selection of constituencies in each region followed by random selection of respondents should give appropriate representation to each group – this was the case in the *CSDS-India Today* opinion poll published in February 1998 (one of the authors, R L Karandikar, was associated with it).

Next we have to deal with the second issue, that of estimating the number of seats for each party. To do this in a meaningful way it becomes necessary to use some appropriate mathematical model. One such model assumes that the change in voting intention from the previous elec-

tion to the present for a particular party (or alliance) is uniform over small homogeneous regions – such as states, parts of states, or other suitably defined geographical regions. Then, by using random sampling we can estimate this change and thus estimate the votes for each party/alliance in each state/region by applying this uniform change over all the individual seats of this region.

Once we have estimated the proportion of votes for each party, we can assign a probability of winning a given seat for each party. If the estimated proportions of votes for the top two parties in a given constituency differ by a big margin, we can assign probability of 1 for the leading party. If the top two are equal or only marginally separated, we can assign a winning probability of 0.5 for each of the top two contenders. For the intermediate cases, we can assign probabilities in between the above two extremes keeping in mind the sample size on which the estimate of proportion of votes is based. Eventually, the sum of probabilities of winning the seats represents the expected number of seats the party is likely to win.

The statistical part of the problem is thus to estimate the percentage of votes for each party in each state. Clearly to achieve reliable figures for each state (or region) it will be necessary to have a reasonably sized sample from the state. We suggest that to achieve this, a sample size of approximately 2000 be selected from each state so that a certain level of accuracy is assured for the state. Nationwide, this will translate to a sample size of about 50000.

It is also necessary to make a distinction between opinion polls and exit polls. While the opinion poll can give an estimate of the possible seat share based on the voter mood on that particular day, there is no way to predict, even among those who identify themselves as likely voters, as to who will go out and actually vote on the voting day. The random sampling technique provides a representative sample for the entire population, not necessarily for those who will actually go out to vote. In addition, people do change their minds between the poll date and the voting date. The opin-

The random sampling technique provides a representative sample for the entire population, not necessarily for those who will actually go out to vote.

Can a sample of 26000 individuals generate any meaningful information for a country as large as India?

ion poll itself provides no objective way to assess the change in the voter's mood between the day of the poll and the day of voting. This problem is avoided in the case of the exit poll. Here the problem of who will come out to vote is taken care of since we sample only from those who have actually voted. Besides the interviewer gets all the subjects at one place rather than having to do a door to door survey and hence the costs are expected to be lower. However since the interviewer has to go by some thumb rule in this case (such as every 20th voter or so) and cannot be done on the basis of a random sample drawn from the voters list, the sampling error may be higher.

Finally, we come back to the question proposed earlier. Can a sample of 26000 individuals generate any meaningful information for a country as large as India? On hindsight it may be noted that the Doordarshan exit poll for the parliamentary elections in 1998 predicted the *nationwide* seat share of the different parties quite accurately (as well as most of the *state-wise* figures), but the *individual predictions* for the states of Tamil Nadu, Maharashtra and Rajasthan were not quite consistent with the final tallies – although when added up the inconsistencies were evened out. This is because while a sample size of 26000 is quite reasonable if one is interested in only the nationwide prediction of vote share, it will not necessarily give good predictions for each state or subregion where the sample size is much smaller. In our opinion, a nationwide sample of 50000 or more, with sample sizes of 2000 or more for the individual states will result in a good nationwide prediction and reasonable predictions in each of the states.

This of course requires that the sampling is done in a manner consistent with statistical theory and a reasonable method is used to translate the estimated vote percentages to estimated seats. It is interesting to observe that in spite of the large number of parties and regional diversity, it is still possible to come up with a meaningful prediction at the national level with a sample of around 50,000.

Address for Correspondence
Rajeeva L Karandikar
Statistics & Mathematics Unit
Indian Statistical Institute
7, SJS Sansanwal Marg
New Delhi 110 016, India.

Ayanendranath Basu
Applied Statistics Unit
Indian Statistical Institute
203, BT Road
Calcutta 700 035, India.