Special Proceeding of 20th Annual Conference of SSCA
held at Pondicherry University, Puducherry during January 29-31, 2018; pp 09-16

Power and Limitations of Opinion Polls in the Context of Indian Parliamentary Democracy

Rajeeva Karandikar
Director, Chennai Mathematical Institute

Final Version Received on 06 September, 2018

Abstract
How can obtaining opinion of, say 10000 voters be sufficient to predict the outcome of an election in a country with over 80 million voters? Do the opinion polls conducted say a month before the election accurately predict what is to happen on the voting day? Do respondents tell the truth? In this article, I will answer these questions and share my own experiences with opinion polls and exit polls in India over last 2 decades.

Key words: Opinion Poll, Exit Poll, Sample Survey, Indian Parliament.

1 Introduction

Let us begin with some facts about Indian political reality. India has parliamentary democracy, with the country being divided into 543 constituencies. Each constituency elects a representative to the *Lok Sabha* (also called lower house) on the basis of first past the post the post basis - candidate getting highest number of votes in the constituency gets elected. A party or an alliance of parties having support of majority of members of the 543 member Lok Sabha names the Prime Minister and forms the government.

The objective for a nationwide opinion poll at the time of Lok Sabha elections is to predict the number of seats for (major) political parties or pre-election allies in the Parliament.

The important questions that need to be answered are:

- Which party or a pre-election alliance of parties will get the maximum number of seats in the house.
- Will this number be more than 272 (the half way mark)?

This talk was delivered as MN Das Memorial Lecture on 29 January 2017 during the 20th Annual Conference of SSCA held at Pondicherry University, Pudducherry.

Corresponding author: Rajeeva Karandikar
Email: rlk@cmi.ac.in

Same applies to a opinion poll at a state level at the time of state legislature election.

To get an idea of the numbers involved, let us note that the total no of eligible voters in India (during last Parliamentary poll in 2014) was 834,101,479). The actual votes polled (2009) was 553,801,801.

Let p be the proportion of voters in India that prefer a party**A** to other parties (or independent candidates). If we have a random sample of n voters and $\hat{p}_n$ denotes the proportion of voters that sample who prefer **A** to other parties / candidates in the sample, then

$$\sqrt{\tfrac{n}{p(1-p)}}(\hat{p}_n - p) \tag{1.1}$$

has standard normal distribution for large n. People who have not studied statistics / probability find it difficult to digest that the (distribution of) error $(\hat{p}_n - p)$ does not depend upon sample size. In common perception, the sampling fraction $\frac{n}{N}$, where N is the total electorate, should determine accuracy of the estimate. Of course, that is not so. I am surprised that even folks who studied statistics at undergraduate level or even graduate level forget this and question me as to how a sample of say 10,000 across India with an electorate of over 800 million, (80 lakhs) will tell me anything about the mood of the nation.

It follows from (1.1) that for large n

$$P\Big(|\frac{\sqrt{n}(\hat{p_n} - p)}{\sqrt{p(1-p)}}| > 1.96\Big) \sim 0.05$$

and thus with $n = 2500$, we can see that for all p

$$P\Big(|(\hat{p_n} - p)| > 0.02\Big) \sim 0.05. \tag{1.2}$$

Likewise, if $n = 10000$, we can conclude that for all p

$$P\Big(|(\hat{p_n} - p)| > 0.01\Big) \sim 0.05. \tag{1.3}$$

Thus if we get a nationwide sample of size 10000, we can estimate p within 1% with 95% probability.

Assuming that we have got a random sample of respondents and their voting intentions, we can get good estimates of vote percentages for major political parties. However, the public interest as well as the interest of the media is in forecast of the seats and not in forecast of vote percentages. And it is easy to see that the seats depend not only on the vote percentages but on the distribution of the votes across constituencies.

If we have a random sample of size 2500 from a constituency with just two candidates, assuming that the winner has 52% or more support, equation (1.2) implies that with 95% probability, the winning candidate will have majority support in the sample as well. Hence if we simply declare the candidate getting majority support in the sample as the winner, we will be right with 95% probability. If there are more than two candidates, the same is true if the gap between the winner and the runner up is at least 4%.

So we could do a sample survey with 2500 respondents in each of 543 constituencies and get a fiarly accurate prediction. However this would require a sample of over 1,300,000 voters and this would be very expensive. Moreover since it is to be done in at most a week's time, this would require a huge manpower. Getting trained and reliable manpower to accomplish such a task is very difficult if not impossible. So we have to find an alternate method.

Let us start with examining the opinion polls conducted elsewhere in the world. In the US, at the time of election, the entire focus is on the Presidential election with winner-take-all in each state. There is election for house of representatives which is like the election to Lok Sabha in India, but that takes a back seat.

Indeed, the Indian system is similar to (indeed derived from) the UK model. So when I got involved the first time, I had examined in depth the methodology used in UK (I interacted extensively with Prof Clive Payne, who had at that time been analyzing polls for BBC for over 2 decades). We concluded that the methods used in UK by him will not work in India. See [1], [2]. The two main reasons are (i) non-availability of socio-economic profiles of constituencies - the census data is organized at district level and not at parliamentary constituency level. (ii) differences in voter behaviour in UK and in India. If we denote by ρ the volatility of public opinion- the proportion of voters who changed their vote from one election to the next (typical gap of 5 years between elections, sometimes could be less in case of mid-term election), then experts believe that ρ is small in UK while in India ρ is believed to be rather high. While in recent times, even in UK the volatility seems to have increased, it is still very low as compared to that in India.

2 Model for Voting Behavior

So what we need to do is to build a model for voting behavior - not of individual voters but at the level of constituencies. Indeed, if C denotes the list of constituencies and P denotes the list of parties, and y_{ij} denotes the proportion of votes for i^{th} party in the j^{th} constituency, then we need a model that would give us an estimate for

$$\{y_{ij}\ :\ \ i \in P,\ j \in C\}.$$

Then we would be able to predict winners in each constituency and thereby get a forecast for the party position in the Lok Sabha.

To build a model, let us note some features of the Indian reality. We have already stated that socio-economic profile of constituencies is not available and thus cannot be incorporated directly in the model. Also, while socio-economic factors do influence voting behavior, groups with similar socio-economic background vote differently in different states- even in neighboring states: one can identify several pairs of constituencies with similar profiles in say Karnataka and Tamilnadu but in Tamilnadu the regional parties - DMK, AIDMK dominate and the national parties - Congress and BJP have marginal presence whereas in Karnataka it is the national parties that dominate. Similar situation prevails in other regions as well- say UP and Bihar. Thus there is hardly any national effect while the state is an important factor.

Let x_{ij} denote the proportion of votes for i^{th} party in the j^{th} constituency in the previous election and let $z_{ij} = y_{ij} - x_{ij}$ denote the *swing* in the i^{th} constituency for the j^{th} party.

As a first step we assume that for each j, the swing z_{ij} is constant as i varies in a state. Or one can refine it and assume that the swing is constant over a region in a state (the region could be a geographic region within a state, such as Vidarbha in Maharashtra or a political region such as rural Bengal in West Bengal).

Let S denote the states and R denote regions. For a state s and a region r let us denote by $\bar{x}_{is}$, $\bar{x}_{ir}$ the proportion of votes in the previous election in the state s and region r respectively for the party i. The quantities $\bar{y}_{is}$, $\bar{y}_{ir}$ are defined analogously for the upcoming election and the swing across a state/ region is defined by

$$\bar{z}_{is} = \bar{y}_{is} - \bar{x}_{is}$$

$$\bar{z}_{ir} = \bar{y}_{ir} - \bar{x}_{ir}.$$

Finally, if s_j denotes the state and r_j denotes the region to which a constituency j belongs, our model can be written as

$$y_{ij} = x_{ij} + \alpha_s \bar{z}_{is_j} + (1 - \alpha_s)\bar{z}_{ir_j} + \text{ error.} \tag{2.1}$$

Here α_s is a number between 0 and 1 and is chosen based on political perception of how strong is the regional effect in the state.

Note that (2.1) can be thought of as a regression model. The absence of any nationwide effect translates to the model not having any intercept.

We have already ruled out large sample in each constituency. However, in each state or a large enough region, we should have enough sample size so as to estimate vote percentages for major parties in each state and region. This and the model helps us estimate y_{ij} as follows:

Based on the sample survey we estimate $\bar{y}_{is}, \bar{y}_{ir}$- namely percentage of votes for a party in every state and every region. Denoting the estimated values by $\hat{y}_{is}$ and $\hat{y}_{ir}$, we have the estimated swing in region r and state s for j^{th} party being

$$\hat{z}_{is} = \hat{y}_{is} - \bar{x}_{is}$$

$$\hat{z}_{ir} = \hat{y}_{ir} - \bar{x}_{ir}.$$

and then the estimated vote percentage for i^{th} party in the j^{th} constituency $\hat{y}_{ij}$ is given by

$$\hat{y}_{ij} = x_{ij} + \alpha_s \hat{z}_{is_j} + (1 - \alpha_s)\hat{z}_{ir_j} + \text{ error.} \tag{2.2}$$

3 Design of the Survey

What we can hope for is that in the survey, we have large enough sample size in each of the states- say with 20 or more constituencies. If we have a sample size of about 30,000 nationwide, then in a cluster of 20 constituencies we can expect to have 1000 samples and this is adequate to give a reasonable estimate of percentage of votes for major parties in a state, or in a geographic sub-region of a state provided it has 20 or more constituencies. As explained above, our model would then give estimates of percentage of votes for major parties in all constituencies.

So the aim is to get a sampling scheme that gives proper representation to all states / sub-regions.

Typically the opinion polls are conducted in US and UK by randomly generating telephone numbers and calling them. Even with the recent growth in number of telephone connections in India, the spread is still lopsided and a opinion poll by this method will lead to a sample that would be biased towards, rich, educated, urban population. Some agencies in India are doing such opinion polls. We have preferred to do door-to-door polls with sample selected following standard statistical techniques.

In India, the data on voters is organized as follows. We have a list of constituencies (where contiguous constituencies come together) and then in each constituency we have a list of polling booths (once again adjacent booths coming together) and then for each booth we have voters list, with neighborhoods forming clusters. Given this fact, we have chosen to undertake multi-stage circular random sampling- first choose (say 20%) of the constituencies, then pick 8 booths and then in each booth pick 50 voters - at each stage the choice is via circular random sampling, also known as systematic sampling.

Since contiguous constituencies appear together in the list, this scheme ensures that all states or sub-regions are suitably represented. We repeat the process for picking booths and then get voters list for the selected booths to pick the respondents, once again using circular random sampling at these two stages as well. The investigators are asked to go door to door and get the response from the selected voters.

Experience has shown that this methodology gets us a sample that is fairly representative of the population on various socio - economic parameters such as caste, religion, education, income etc. Since we are not force fitting this but rather, achieved through our sampling techniques, it gives us confidence that our vote share estimates will also be close to the true vote share.

It should be noted that the choice of the sampling scheme is dependent upon the sampling frame. If for example, the lists of constituencies, polling booths in a constituency, and voters list in a booth were by alphabetical order, circular sampling would be inappropriate.

A question remains on the estimation of standard deviation of the error in (2.2) (from which we can obtain the estimate of standard deviation of y_{ij} using the sample size at state level and regional level). Extensive backtesting has shown that the model (2.2) has a large error variance. The empirical data suggests a standard error of about 4 to 8 percent for the estimate of y_{ij}, depending upon the sample size involved.

The error variance being large confirms that the model is rather crude and if we make constituency wise predictions several will be incorrect.

4 Vote-to-Seat Conversion

Once we have the estimates $\hat{y}_{ij}$, one approach to predicting number of seats for a party would be to count number of constituency where that party is predicted to have highest vote share and take that as its predicted seats. But we can do better.

Consider two constituencies with only two candidates and suppose in the first constituency, the predicted vote shares of first and second party are $\hat{y}_{11} = 0.51$, $\hat{y}_{21} = 0.49$ and in the second constituency, the predicted vote shares are $\hat{y}_{12} = 0.60$, $\hat{y}_{22} = 0.40$ (and all estimates have a standard error of 0.08). Now based on our data, we can say that party 1 will very likely win in constituency 2, but we can't say so for constituency 1- all we can say is that party 1 is more likely to win than party 2 in constituency 1. We need to take this into account while making seat estimates. Our approach: the best case scenario for party 2 in constituency 1 is that while it is marginally ahead, a sample of the specified size throws up a deficit of 0.01. The probability of such an event is approximately equal to (here Z denotes a standard normal variate)

$$P(Z > \frac{0.01}{0.08}) = 0.45.$$

Thus we assign party 2 in 1^{st} constituency a win probability of 0.45 and party 1 in 1^{st} constituency a win probability of 0.55. On the other hand, noting that

$$P(Z > \frac{0.1}{0.08}) = 0.11$$

we assign party 2 in 2^{nd} constituency a win probability of 0.11 and party 1 in 2^{nd} constituency a win probability of 0.89.

When we have three or more parties, we look at the top three parties in a constituency and distribute the probability of win among them. As an example, if the top 3 parties estimated votes are 0.32, 0.26 and 0.24 with each having standard error of 0.08), then the estimates of probability of win are respectively 0.55, 0.26, 0.19.

Once we have *predicted probability of win* in all constituencies for all parties, we can add the probabilities across constituencies to get an estimate of number of seats for a party.

One problem arises with this approach - if alliances change from one election to the other, we need to factor that in. If two parties who fought separately form an alliance, we cannot simply add the votes to get votes for the alliance. Some proportion of voters may nto like the alliance and may change their vote. We have to use political judgement to create a simulated vote share distribution taking into account current alliances using the vote shares in the previous election.

5 Some Comments

Any opinion poll can at best measure the mood of the country as a whole while for actual results, only those who vote count. This creates a question mark on prediction based on opinion poll. Moreover, in India voting intentions are volatile and so even if the opinion poll accurately measure the mood of nation at the time of the poll, the mood could change when the voting day approaches.

One can correct for differential voter turnout across various social classes but to model and measure the churn in voting intention seems almost impossible.

Some agencies resort to tracking poll- several polls done say with a gap of one week each and then estimate the trend and extrapolate. This assumes that the change is stationary - again an assumption that is questionable specially in the Indian context.

An exit poll addresses the two issues - the sampling is done on the day of the poll choosing voters as they come out of the booth. However, randomization of respondents is difficult to achieve. One can choose polling booths via randomization but we cannot have our investigator stand there with a list asking people to check if their name is in the list. One has to leave the choice of respondents to the investigator by giving some thumb rules. Our experience shows that this method gives a biased sample on socio-economic attributes that we can measure.

We have mostly resorted to day-after poll- given the gap between the voting day and counting day, we collect data following the sampling scheme mentioned above (multi stage circular sampling) by interviewing the targeted voters a day after they have voted. We can still come out with a projection for the Lok Sabha or Vidhan Sabha before actual counting begins. Or when the polling is in several phases, we can do the day after poll in areas that had polling in all but the last phase, and do an exit poll in the last phase. Then we can have more reliable result on the day polling ends.

In our day after poll, we ask the respondents if they have voted and also see the ink mark on the finger. For the purpose of seat predictions, we ignore the opinion of those who do not have the ink mark. We have been fairly successful with this strategy, often getting much closer to the actual result then the other surveys.

A technique that is often used in such surveys is to detect possible bias in the sample by asking a question about the voting in the previous election, determining the bias (since the results from the previous election are available) and than taking it as a measure of bias in our sample. For example, if in 2019 just before the next Lok Sabha poll we ask respondents about their current voting preference and whom did they vote for in 2014, and our estimate is that currently 34% vote share for NDA but about 43% recall having voted for NDA in 2014. The actual vote share of NDA in 2014 was 38, so we could adjust the current vote estimate to $\frac{34*38}{42} = 30$ percent.

However, what we have observed is that in India, there is always over reporting for whoever won the last election. This is so even when the party which won the previous election is currently losing. So in our view, this can be attributed to tendency to align with the winner. Thus we have refrained from making this correction. We have seen that if we were using such correction, we would have been worse off in terms of our prediction in most cases.

A question often asked to me is:*Do respondents answer question about their voting preferences?*

Of course many respondents will not answer a question on voting preference if asked face to face. We do ask other questions face to face. And we carry a old style ballot paper and a sealed cardboard box with a slit and ask respondents to go to a corner, mark their preference on the paper, fold the same, and put it in the box. With this method, refusal rate about 8-10%.

Do respondents hide the truth and do we correct for the same?

When it comes to detecting hiding the truth (or lying), one approach would be to fit a model that uses answers to other questions. I do believe that voting is based on complex set of parameters and in Indian context, fitting such models will not work. For example, various other questions may point towards party the voter likes but the voting may be influenced by the candidates more than the party. So we have refrained from doing so.

To conclude, let me say that following statistical principles, we have had a fairly good success rate in forecasting the election outcomes in India.

References

Karandikar, R. L. (1999). Opinion Polls and Statistics. *Calcutta Statistical Association Bulletin*, **49**, 193-194.

Karandikar, R. L., Payne, C. and Yadav, Y. (2002). Predicting the 1998 Indian parliamentary election. *Electoral studies*, **21**, 69-89.