Opinion polls, Exit polls and Early seat projections

Rajeeva L. Karandikar

Cranes Software International Ltd.
rajeeva.karandikar@cranessoftware.com

About this talk

In this talk we will see that some simple mathematics and statistics, lots of common sense and a good understanding of the ground reality or domain knowledge together can yield very good forcast or predictions of the outcome of elections based on opinion polls and exit polls.

These statistics, common sense and domain knowledge are the ingredients that go into psephology - the science (?) of opinion polls and poll projections.

Contents

We will discuss various issues involved in psephology with focus on the Indian context. I will also share my experiences over the last 10 years (and views on applications of statistics to real world questions). We will begin with a brief general discussion on sampling theory (with apologies to experts). We will also address various questions that are raised by skeptics and politicians all the time.

The questions we will address

- Is there a science behind opinion polls?
- I do not believe in gambling. Why should I believe in opinion polls?
- Do opinion polls conducted well ahead (say a month) of the polling date have predictive power as far as final results are concerned?
- Is a sample size of, say $20,000$ sufficient to predict the outcome in a country with over 60 crore voters?

How can opinion of a small fraction reveal the true result?

Most surveys in India including the ones conducted by me end up interviewing less than 0.05% voters. For an all India survey, it may be much less.

This is one question which is most difficult to answer to a layman. Or a related question- say we have conducted a nationwide survey with 10000 respondents. Then the media agency funding say Gujarat poll would want us to use only about 500 to 700 respondents. After all, Gujrat's population is less than one fifteenth of the country!

For the skeptics, with apologies to experts

If an urn contains M balls identical in all aspects except for colour, with K of them being orange and rest being green. If the balls are mixed and without looking, one of them is drawn, then the chance of it being orange is $\frac{K}{M}$.

This is based on the premise that each ball has an equal probability of being drawn and since there are K orange balls, the required probability is $\frac{K}{M}$.

For the skeptics, with apologies to experts...

Suppose $M = 10000$ and K is either 9900 or 100, so either 9900 are orange or 9900 are green. One ball is drawn (after mixing, without looking) and its colour is found to be green. We are to make a decision about K - choose out of the two possibilities: $K = 100$ or $K = 9900$.

If $K = 100$, then probability of drawing a green ball is 0.99 whereas if $K = 9900$, then probability of drawing a green ball is 0.01.

For the skeptics, with apologies to experts...

Based on this, we can say: colour of most of the balls (9900) is likely to be green. This is what common sense tells us and can be justified in various ways. The story would not change if K is either $99,000$ or $1,000$ with $M = 100,000$.

This is the only idea from probability theory or statistics that is needed to answer most of the questions of skeptics as we will see.

For the skeptics, with apologies to experts...

Now consider a constituency, say Bangalore South and to make matters simple, suppose there are two candidates, A and B. Suppose there are M voters, $V_1, V_2, \ldots, V_M$.

Suppose $a_i = 1$ if V_i prefers A (to B) and $a_i = 0$ if V_i prefers B (to A). Suppose p is the proportion of voters who prefer A to B *i.e.*

$$K = a_1 + a_2 + \ldots + a_M, \;\; p = \frac{K}{M}.$$

For the skeptics, with apologies to experts...

The interest is in deciding if $K > (M/2)$ or $K < (M/2)$ (for simplicity, let us assume M is an odd integer) Who will win the election A or B? . Of course if we observe each a_i, we will know the answer. Can we have a decision rule even if we observe only a few a_i's?

For the skeptics, with apologies to experts...

Let $\mathscr{S}$ denote the collection all n-tuples $(i_1, i_2, \ldots, i_n)$ of elements of $\{1, 2, \ldots, M\}$. For $(i_1, i_2, \ldots, i_n) \in \mathscr{S}$ let

$$f(i_1, i_2, \ldots, i_n) = a_{i_1} + a_{i_2} + \ldots + a_{i_n}$$
$$g(i_1, i_2, \ldots, i_n) = \frac{1}{n} f(i_1, i_2, \ldots, i_n).$$

Note that $g(i_1, i_2, \ldots, i_n)$ is the proportion of 1's in the sample $\{i_1, i_2, \ldots, i_n\}$

For the skeptics, with apologies to experts...

We are going to estimate

$$\frac{\#\{(i_1, i_2, \ldots, i_n) \in \mathscr{S} :\mid g(i_1, i_2, \ldots, i_n) - p \mid > \varepsilon\}}{M^n}$$

It can be easily seen (high school algebra) that

$$\sum_{\mathscr{S}} f(i_1, i_2, \ldots, i_n) = nKM^{n-1}$$

$$\sum_{\mathscr{S}} (f(i_1, i_2, \ldots, i_n))^2$$

$$= nKM^{n-1} + n(n-1)K^2M^{n-2}.$$

For the skeptics, with apologies to experts...

Thus

$$\frac{1}{M^n}\sum_{\mathscr{S}} g(i_1, i_2, \ldots, i_n) = \frac{K}{M} = p$$

$$\frac{1}{M^n}\sum_{\mathscr{S}} (g(i_1, i_2, \ldots, i_n))^2 = \frac{p}{n} + \frac{(n-1)p^2}{n}.$$

For the skeptics, with apologies to experts...

It follows that

$$\frac{1}{M^n}\sum_{\mathscr{S}}\Big(g(i_1,i_2,\ldots,i_n)-p\Big)^2=\frac{p(p-1)}{n}.$$

Thus

$$\frac{\#\big\{(i_1,i_2,\ldots,i_n)\in\mathscr{S}:\mid g(i_1,i_2,\ldots,i_n)-p\mid>\varepsilon\big\}}{M^n}$$

$$\leq\frac{1}{\varepsilon^2}\frac{p(1-p)}{n}.$$

If this is not true then the previous equality would be violated.

For the skeptics, with apologies to experts...

Thus by choosing a large sample, one can ensure that in most samples, the sample proportion and true proportion differ by a small quantity. Thus if a large sample is selected at random, the probability that the sample proportion and true proportion differ by a small quantity is close to 1.

For the skeptics, with apologies to experts...

Writing $\hat{p}_n = g(i_1, i_2, \ldots, i_n)$ - the sample proportion, we get

$$\text{Prob}(|\,\hat{p}_n - p\,| > \varepsilon) \leq \frac{1}{\varepsilon^2}\frac{p(1-p)}{n}$$

This estimate for $p = 0.5, \varepsilon = 0.05$ yields

$$\text{Prob}(|\,\hat{p}_n - p\,| > 0.05) \leq \frac{100}{n}$$

Note that the estimate does not depend on M.

For the skeptics, with apologies to experts...

The probability estimate given above can be improved (using central limit theorem). It can be shown that, to ensure that

$$\text{Prob}(|\,\hat{p}_n - p\,| > 0.05) \leq 0.01,$$

we need to take $n = 640$ while to ensure that

$$\text{Prob}(|\,\hat{p}_n - p\,| > 0.02) \leq 0.01,$$

we need to take $n = 4000$, to ensure that

$$\text{Prob}(|\,\hat{p}_n - p\,| > 0.01) \leq 0.01,$$

we need to take $n = 16000$.

For the skeptics, with apologies to experts...

These estimates do not depend on M, but only on n. Thus the accuracy of a sampling scheme does not depend on the sampling proportion $\frac{n}{M}$, but on sample size n. Indeed, if we wish to estimate the percentage votes for the political parties across the nation, a sample of size 16000 will give an estimate correct up to 1% with probability over 99%.

Importance of Random Sampling

The argument given above can be summarized as: "Most samples with size 16000 are representative of the population and hence if we select one randomly, we are likely to end up with a representative sample".

In colloquial English, the word random is also used in the sense of arbitrary (as in Random Access Memory- RAM). So some people think of a random sample as any arbitrary subset.

Importance of Random Sampling ...

Failure to select a random sample can lead to wrong conclusions. In 1948, all opinion polls in US predicted that Thomas Dewey would defeat Harry Truman in the presidential election. The problem was traced to choice of sample being made on the basis of randomly generated telephone numbers and calling the numbers. In 1948, the poorer sections of the society went unrepresented in the survey.

Importance of Random Sampling ...

Today, the penetration of telephones in US is almost universal and so the method now generally works in US. It would not work in India even after the unprecedented growth in telecom sector, as poorer section are highly under represented among people with telephone and thus a telephone survey will not yield a representative sample.

Importance of Random Sampling ...

Another method used by market research agencies is called quota sampling, where they select a group of respondents with a given profile - a profile that matches the population on several counts, such as Male/Female, Rural/Urban, Education, Caste, Religion etc. Other than matching the sample profile, no other restriciton on choice of respondents is imposed and is left to the enumerator.

Importance of Random Sampling ...

However, in my view, the statistical guarantee that the sample proportion and population proportion do not differ significantly doesn't kick in unless the sample is chosen via randomization. The sample should be chosen by suitable randomization, perhaps after suitable stratification.

This costs a lot more than the quota sampling! But is a must.

Predicting seats for parties

Well. Following statistical methodology, one can get a fairly good estimate of percentage of votes of the major parties, at least at the time the survey is conducted.

However, the public interest is in prediction of number of seats and not percentage votes for parties.

Predicting seats for parties

It is possible (though extremely unlikely) even in a two party system for a party 'A' with say 26% to win 272 (out of 543) seats (majority) while the other party 'B' with 74% votes to win only 271 seats ('A' gets just over 50% votes in 272 seats winning them, while 'B' gets 100% votes in the remaining 271 seats).

Thus good estimate of vote percentages does not automatically translate to a good estimate of number of seats for major parties.

Predicting seats for parties ...

Thus in order to predict the number of seats for parties, we need to estimate not only the percentage votes for each party, but also the distribution of votes of each of the parties across constituencies. And here, independents and smaller parties that have influence across few seats make the vote-to-seat translation that much more difficult. Let us first examine prediction of a specified seat.

Predicting seats for parties...

Consider the task of predicting the winner in a Lok Sabha election in Madras South Constituency (which has over 20 lakh voters). Suppose that the difference between true support for the two leading candidates is over 4 percent votes. By generating a random sample of size 4000 and getting their opinion, we can be reasonably sure (with 99% probability) that we can pick the winner. Indeed the same is true even if the constituency had larger number of voters.

Predicting seats for parties...

So if we can get a random sample of size 4000 in each of the 543 constituencies, then we can predict winner in each of them and we will be mostly correct (in constituencies where the contest is not a very close one).

But conducting a survey with more than 21 lakh respondents is very difficult: money, time, reliable trained manpower,.... each resource is limited.

Predicting seats for parties...

One way out is to construct a model of voter behavior. While such a model can be built, estimating various parameters of such a model would itself require a very large sample size. Another approach is to use past data in conjunction with the opinion poll data. In order to do this, we need to build a suitable model of voting behavior- not of individual voters but for percentage votes for a party in a constituency.

The Indian reality

To make a model, let us observe some features of the Indian democracy.

Voting intentions are volatile- in a matter of months they can undergo big change. (Example: Delhi in March 98, November 98, October 99) This is very different from the situation in UK where voting intentions are very stable, and thus methods used in UK can not be used in India, though superficially, the Indian political system resembles the one in UK.

The Indian reality

This is where domain knowledge plays an important role. A model which works in the west may not work in Indian context if it involves human behavior. And having all the data relating to elections in India (since 1952) will not help. The point is that large amount of data cannot substitute understanding of ground realities.

The Indian reality...

While the behavior of voters in a constituency may be correlated with that in adjacent constituencies in the same state, the voting behavior in one state has no (or negligible) correlation with that in another state. (The behavior is influenced by many local factors.)

The socio-economic factors do influence the voting pattern significantly. However, incorporating it directly in a model will require too many parameters.

The Indian reality...

It is reasonable to assume that the socio-economic profile of a constituency does not change significantly from one election to the next. So while the differences in socio-economic profiles between two constituencies are reflected in the differences in voting pattern in a given election, the change from one election to the next in a given constituency does not depend on the socio-economic profile.

The Model

So we make an assumption that the change in the percentage of votes for a given party from the previous election to the present is constant across a given state.

The resulting model is not very accurate if we look at historical data, but is a reasonably good approximation- good enough for the purpose- namely to predict the seats for major parties at national level.

The Model...

The change in the percentage of votes is called swing. Under this model, all we need to do via sampling is to estimate the swing for each party in each state and then using the past data we will have an estimate of percentage votes for each party in each state.

Here we can refine this a little- we can divide the big states in regions and postulate that the swing in a seat is a convex combination of swing across the state and swing across the region.

Predicting the Winner

Here enters one more element. We need to predict the winner in each constituency and then give number of seats for major parties.

Suppose in one constituency with nly two candidates, we predict 'A' gets 50.5%, 'B' gets 49.5%, in another constituency we predict that 'C' gets 54% votes, 'D' gets 46% votes, in both cases, the sample size is say 625. It is clear that while winner between 'A' and 'B' is difficult to call, we can be lot more sure that 'C' will win the second seat.

Predicting the Winner...

What is the best case scenario for 'B'- that indeed 'A' and 'B' have nearly equal support with 'B' having a very thin lead, and yet a random sample of size 625 gives a 1% lead to 'A'. This translates to : in 625 tosses of a fair coin, we observe 316 or more heads. The probability of such an event is 0.405 (using normal approximation). So we assign 'B' a winning probability of 0.402 and 'A' a winning probability of $1 - .405 = 0.595$.

Predicting the Winner...

This can be extended to cover the case when there are three candidates 'A','B' and 'C' getting significant votes, say 36%, 33%, 31% respectively. Now we will asiign probabilities to the three candidates, adding upto one. First the best case scenario for 'C', then the best case scenario for 'B'.

Predicting the Number of seats

Summing over the probabilities over all the 543 seats we get the expected number of seats for each party. This method gives reasonable predictions at state level and good predictions at the national level. This model was validated (in run up to 1998 elections, we treated 1991 data as given and assumed that of the 1996 election, only statewide and region wide vote percentages was given and the model used to predict seats. This was close to the real outcome).

Design of sample survey

The crux of the matter is to get random sample that is reasonably distributed across the country. The method we followed in 1998: we decided to sample in 20% (108) constituencies: so we picked a random number between 1 and 5 (say we get 3) and then go on picking constituencies numbered $3, 8, 13, 18, ...$ in the election commission list.

Design of sample survey

In the list contiguous constituencies occur together and hence the method described above, known as systematic sampling or circular sampling gives an even spread across the country.

Then we got a list of polling booths in each constituency and picked 4 polling booths again by circular random sampling. Finally, we got the voters list in these booths and picked 35 voters in each chosen polling booth by circular sampling.

Design of sample survey

The enumerators were asked to go door to door (3 times if necessary) and get the opinion.

I would like to add that about $\frac{1}{3}$ of the country voted 9 days after our poll, another $\frac{1}{3}$ 16 days after the poll and for the rest $\frac{1}{3}$, the gap was 24 days.

The prediction

Our realized sample was about 9600 and while the sample profile on attributes like caste, religion, income level, education level, rural/urban etc matched the national profile (coming from census figures), our prediction was that BJP and allies would be the leading group with 214 seats (this was published in India Today).

The actual result was 251.

The prediction...

We had clearly said that our assessment was that if the elections were held the day our survey was held, the BJP and allies would get 214. However, no one sees the details and conclude that we had predicted (or forecast) that they would get 214 in the actual election.

The prediction...

Some others do claim to correct for this effect.

They conduct what is called a tracking poll, where polls are conducted every week for say 6 to 8 weeks prior to the polls and then the trend is extrapolated to get the prediction on what is to happen on election day.

However, the churn nearer to the voting day is much more than in previous weeks and this method has no basis.

The prediction...

In 1998, we also asked the same respondents (as in our opinion poll) whom they voted for a day after the poll and based on this our prediction, obtained before counting started was 250

We found that as many as 30% voters had changed their mind.

The prediction...

This raises a question about predictive power of any opinion poll in India: voting intentions are far too volatile.

There is the added problem of opinion poll generating a sample from whole population whereas only about 55%-60% actually vote.

Exit polls

Both these problems are addressed by an Exit poll, where we interview respondents as they exit the polling booth.

However, in an exit poll, we cannot ask respondents from a previously generated list. We can at best choose polling booths via multi stage circular sampling and then give a thumb rule, such as pick every 10^{th} voter, to the enumerator. This may introduce bias in the sample.

Exit polls

Beginning with November 2005 Bihar assembly polls, we (myself and Yogendra Yadav) have undertaken exit polls for CNN-IBN, where given the multi-phase polls that have become the norm, we conduct a proper randomized poll in all but the last phase and in the last phase we do an exit poll.

Exit polls

We have had a fair track record, in Bihar we predicted a majority for Mr Nitish Kumar (which no other poll did). Likewise, in Assam 2006. In Bengal and Tamil Nadu (2006) our prediction was on the dot- much better than any other poll. In Kerala 2006 too we had picked the winner, thought we had overestimated the margin. In UP and Gujarat 2007, we had the winner but underestimated their seats.

Early Seat Projection - ESP

Once the counting starts and early trends are reported, the interest of general viewer is in the big picture- the seats for various parties in the house. So if 100 out of 543 constituencies are reporting and if party A has got 40 and no other party has got more than 20, what does it mean for the full house- does it mean no one will get majority?

Early Seat Projection - ESP...

Remember, the 100 constituencies reporting cannot be construed as a random sample out of 543. We had built a model to take this into account and make a forcast of the final seats in the parliament during 1998 counting, which went on for 3 days. About 8 hours after counting started, we had started making projections. This treated the opinion poll projections as prior and given the observations, we computed the posterior.

Early Seat Projection - ESP...

The experiment was very successful in 1998 and 1999 (when counting lasted 2 days). By 2004, the counting was over in 4 hours (most of it) and so we had a small window of opportunity. In 2004, all the opinion polls were quite off the mark and yet our ESP was quite accurate- by 9.30am we had gone on air (on Aaj Tak) with numbers close to final picture.

Final thoughts

So to sum up, proper use of statistical techniques and some domain expertise can give very remarkable results.

Opinion polls/ exit polls do serve a much larger purpose than forcasting the final outcome- it gives an insight into why people have voted the way they did.

Final thoughts

All through the 1990's, as the education level increased the support for Congress came down and support for BJP went up. Same is the story if one looked at economic class. In 2004, this trend was not so prominent. Congress had made up lot of lost ground among the educated and affluent.

If the political parties use the opinion polls as a feedback mechanism to gauge public opinion and act accordingly (too much to expect!).