Power and Limitations of Opinion Polls

Rajeeva L. Karandikar
Director
Chennai Mathematical Institute
rlk@cmi.ac.in

Question often asked:

- How can obtaining opinion of, say 20,000 voters be sufficient to predict the outcome in a country with over 71 Crore voters?

Probability and Statistics background

Suppose a box contains 100 slips of paper, identicle in all aspetcs and have the number 7 or 8 written on it- 99 of them have one number on it and 1 has the other number on it. The slips of paper are mixed after folding and one slip is drawn and opened. Suppose it has the number 7.

Based on this if someone has to guess the number that dominates, most people will guess it as: 7.

If instead of 99 having one letter, only 95 have one letter and 5 the other, we can draw 3 times and go with the majority: the accuracy level is over 99%

$$\frac{95*95*95+3*5*95*95}{100*100*100} = 0.992750$$

If the gap is lesser, we need to increase number of draws to achieve 99% accuracy.

Probability and Statistics background...

Now consider an assembly constituency with 100000 voters and to make matters simple, suppose there are only two candidates, A and B. Suppose we make all possible lists of n voters, (where n is an odd number). What proportion of lists show A as the winner?

Two candidates A and B. Population size 100000.
Column header is the percentage of support for Candidate A and row header is the size of the list. The Table shows percentage of lists that have Candidate A having majority.

	45	46	47	48	49	50	51	52	53	54	55
101	15.6	21	27.3	34.3	42	50	58	65.7	72.7	79	84.4
151	10.9	16.2	23	31.1	40.3	50	59.7	68.9	77	83.8	89.1
201	7.7	12.8	19.7	28.5	38.8	50	61.2	71.5	80.3	87.2	92.3
251	5.6	10.2	17	26.3	37.6	50	62.4	73.7	83	89.8	94.4
501	1.2	3.6	8.9	18.5	32.7	50	67.3	81.5	91.1	96.4	98.8
751	0.3	1.4	5	13.6	29.2	50	70.8	86.4	95	98.6	99.7
1001	0.1	0.6	2.9	10.2	26.3	50	73.7	89.8	97.1	99.4	99.9
1251	0.1	0.2	1.7	7.8	23.9	50	76.1	92.2	98.3	99.8	99.9
1501	0.1	0.1	1	6	21.9	50	78.1	94	99	99.9	99.9

Probability and Statistics background...

Thus if the winning candidate is getting at least 54% votes (not a close election) and if we take $n \geq 1001$, then 99.4% lists have the winning candidate having majority support.

If the election is closer, with winning candidate getting 53% votes and if we take $n = 1501$ then we have 99% lists have the winning candidate having majority support.

Probability and Statistics background...

What if the total number of voters is 500000 instead of 100000? Suppose winning candidate is getting 53% votes. We needed lists of size $n = 1501$ to ensure 99% lists have the winning candidate having majority support.

Do we need to take $n = 7505$ to have same accuracy now?

Let us go back to $n = 3$

Observe that

$$\frac{95 * 95 * 95 + 3 * 5 * 95 * 95}{100 * 100 * 100} = 0.992750$$

is the same as

$$\frac{95000 * 95000 * 95000 + 3 * 5000 * 95000 * 95000}{100000 * 100000 * 100000} = 0.992750$$

Lesson: Population size does not matter (if repetition is allowed), only list size matters.

Suppose Candidate A has 52% support. The Table below shows percentage of lists that have Candidate A having majority. Column header is the population size and row header is the size of the list.

	10000	25000	50000	100000	250000	500000	1000000	2500000	5000000
401	79.3	79.1	79	78.9	78.9	78.9	78.9	78.9	78.9
601	84.4	84	83.8	83.8	83.7	83.7	83.7	83.7	83.7
1001	90.9	90.2	90	89.8	89.8	89.8	89.7	89.7	89.7
1501	95.4	94.5	94.2	94.1	94	94	94	94	94
1801	97	96.1	95.8	95.7	95.6	95.6	95.5	95.5	95.5
2001	97.7	96.9	96.6	96.5	96.4	96.4	96.3	96.3	96.3
2501	99	98.3	98	97.9	97.8	97.8	97.7	97.7	97.7
3001	99.6	99	98.8	98.7	98.6	98.6	98.6	98.6	98.6
4001	99.9	99.7	99.6	99.5	99.5	99.4	99.4	99.4	99.4

Probability and Statistics background...

So accuracy is determined by list size and does not depend upon population size (once list size is less than 0.1% of population size)

A list is what is called a sample and once sample is chosen we can talk to the voters on the list and see who is ahead in the sample. Based on this we can make a prediction about winner in an election.

Probability and Statistics background...

Thus by choosing a large sample, one can ensure that in most samples (99%), the winner in the sample is also the winner in the constituency. Thus if a large sample is selected at random, we can pick the winner with 99% probability

Importance of Random Sampling

The argument given above can be summarized as: "Most samples with size say 4000 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 restriction 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.

Predicting seats for parties...

If we get a random sample of size 4000 in each of the 543 constituencies, then as explained earlier, we can predict winner in each of them. 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.

Let us look at what is done elsewhere.

The Indian reality

US

UK

The Indian reality

At the face of it, the Indian system is very similar to the British system and so it would appear that the methodology used in Britain can be used in India too.

When I got involved in this exercise the first time in 1997, Professor Clive Payne - statistician and psephologist with over 20 years of experience of analyzing polling data for BBC was specially flown into Delhi and we discussed at great length his methods and the ground realities in India and concluded that it would not be appropriate to use the same.

The Indian reality...

The main reason is that voting intentions in UK are very stable across time whereas in India they are very volatile. If we define ρ as the proportion of people who changed their vote from previous election to the present, then ρ is very small in UK whereas in India it could be very high.

This is what experts believe. Indeed, in 1998 we had funds to conduct an opinion poll just before the election exercise started and then another round a day after the actual voting. There was a gap of 8 days for a third of the country, 16 days for another third and about 25 days for the remaining third.

The Indian reality...

And we found that as high as 30% voters in our sample had changed their mind! Thus ρ is at least 25%.

So the methods used by in UK cannot be used by us in India.

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.

Predicting seats for parties...

We need to create a model for voter behavior. We don't have to model individual voter preference, a model for vote percentage for major parties in each of the constituencies would suffice.

While one can create a model that incorporates various socio-economic parameters such as caste, religion, economic status, educational level etc., the number of parameters is too large as the behavior varies from one state to the other.

Predicting seats for parties...

Moreover, the constituency profile on these parameters is not available. The census data in India is available at district level while a constituency could include parts of several districts.

We work with a very crude model that assumes that the Change in votes - called Swing- for a given party from the previous election to the present is uniform across a state. This is based on the premise that constituency profile on socio economic factors does not change drastically over the 5 years (perhaps true for most of the constituencies).

Predicting seats for parties...

We can refine this a little and assume that the swing in a constituency for a given party is a convex combination of the swing across the state and swing across sub-region. We could also add division of the state according to some other criterion as a factor in the model.

When we try to validate this model using past data, we see that this is a very bad model if we look at the microscopic level- namely compare predicted votes and actual votes in each constituency. *However, if we use the model for our objective- to predict to seats, it yields satisfactory results.*

Design of sample survey

So the task is to estimate the swing at state level and perhaps across regions within a state. The crux of the matter is to get a sample that is reasonably distributed across the country.

How do we choose the sample?

Well the data on voters is organized as follows. We have list of constituencies (where constituencies in a state come together) and then in each constituency we have a list of polling stations (here adjacent booths come in a cluster) and then for each booth we have voters list, with neighbourhoods forming clusters.

Design of sample survey...

Thus we have chosen to undertake multi-stage circular random sampling- first we choose (say 20%) of the constituencies, then pick 4-6 booths and then in each booth pick 30-50 voters - at each stage the choice is via circular random sampling, also known as systematic sampling.

In this to pick say 108 out of 543 constituencies, we randomly pick a number between 1 and 543, say 378 then begining with 378, we include every 5th constituency: we generate the list 378, 383, ..., 543, 5, 10, ..., 370.

Design of sample survey...

The circular random sampling or systematic sampling ensures that various parts of the country are well represented.

If the lists were say alphabetical by constituency name, polling booth name and voter name, then circular random sampling would not be the best scheme, perhaps simple random sampling at each stage would be better.

Thus what sampling scheme to use depends upon how the master data is organised.

Design of sample survey...

We have generally found that sample obtained by this method is fairly balanced- the sample profile on various socio-economic parameters matches the population profile obtained from the census data at state level.

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 only 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 0.405 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.

The prediction...

I have earlier commented upon the fact that as voting day approaches, there seems to be a huge churn in voting intention in India, perhaps driven by some events or speeches given by leaders and media coverage etc.

Moreover, an opinion poll at best can give the pulse of the general voter but what counts is the voters who go and vote. It has been observed in India that fewer educated, rich, urban voters vote as compared to poorer, less educated rural voters (comparison in %).

The prediction...

These two factors put a big question mark on the predictive power of any poll done ahead of the polling day.

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.

How high is refusal rate?

Another question is: do respondents answer question about their voting preferences?

Of course respondents will not answer a question on voting preference if asked face to face. 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.

Refusal rate about 8-10%.

Do we correct for lying?

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. For example, if someone in Mumbai likes the central NDA government, likes the state government, says that Modi government is much better than the previous UPA government, and yet says he will vote for congress- may be classified by various such models as a liar. However, his local NDA candidate may be a candidate whom the respondent dislikes intensely and hence may be voting for Congress.

Our answer

I firmly believe that voting intention is a very complex process and trying to fit any model and using the same to *correct* the respondents answer is unlikely to improve our estimate- indeed, it may lead us away from the truth.

Exit Poll

Exit polls were devised to correct both these effects: the gap between the opinion poll and date of voting and also that only between 50% and 70% voters actually vote.

Here, voters are asked questions as they exit the polling booth. However, here randomly selecting voters is almost impossible.

Day after poll...

What we prefer to do is the following:
The polling in India is of late divided in several phases, lasting may be over a month. This is so that the security forces can be moved from one area to another to ensure smooth conduct of polls. After the last phase is over, the counting is done after 2 or three days gap.

So we conduct proper randomized poll day after the voting (door-to-door) with the multi stage circular sampling.

Our Track record

Let me mention that the media hypes these projections as *the truth, the whole truth and nothing but the truth.*

Actually, the polls should be seen as giving an indication, as to who is likely to win, will anyone get majority and so on.

And it also gives a deeper insight into why people voted the way they did.

Our Track record

Let me come to our (CNN-IBN - CSDS - RLK) track record.

Period November 2005 - May 2014.

Our track record...

By my own assessment, **we were not good on 4 occasions** (off the mark and others did better than us) - (i) Punjab 2007 (ii) Gujarat 2007, (iii) Karnataka 2008, (iv) Gujarat 2012.

On the following **8 occasions we were good** (generally on track and as good as others) (i) Kerala 2006 (ii) Uttarakhand 2007 (iii) Uttar Pradesh 2007 (iv) Lok Sabha 2009 (v) Tamilnadu 2011 (vi) Himachal Pradesh 2012 (vii) Uttarakhand 2012 (viii) Lok Sabha 2014.

Our track record...

And on the following **16 occasions we were very good** (estimates on the dot or close and better or as good as others) (i) Bihar 2005 (ii) Assam 2006 (iiii) Tamil Nadu 2006 (iv) West Bengal 2006 (v) Bihar 2010 (vi) Assam 2011 (vii) Kerala 2011 (viii) West Bengal 2011 (ix) Uttar Pradesh 2012 (x) Punjab 2012 (xi) Manipur 2012 (xii) Karnataka 2013 (xiii) Madhya Pradesh 2013 (xiv) Rajasthan 2013 (xv) Chhatisgarh 2013 (xvi) Delhi 2013.