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 number, only 95 have one number 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.

Importance of Random Sampling ...

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 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...

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

Predicting seats for parties...

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...

Can be refined further

Gives reasonable results on backtesting.

Design of sample survey

Multi stage systematic sampling

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 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.

Predictive power of pre-election poll?

Volatility of opinion.

Not all eligible voters vote!

Predictive power of pre-election poll?

Any prediction based on pre-election poll does not predictive power as far as final results are concerned.

How high is refusal rate?

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

Refusal rate about 8-10%.

Do we correct for lying?

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

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.