Proc Indian Natn Sci Acad **86** No. 4 December 2020 pp. 1461-1479
© *Printed in India*. DOI: 10.16943/ptinsa/2020/157320

Lecture

Mathematics and Elections[#]

RAJEEVA L KARANDIKAR
Chennai Mathematical Institute, Chennai, India

(Delivered on 17 December 2019)

Statistical ideas have been impacting various aspects of our lives for at least a century. Starting with Mendel's experiments with pea plant in the 1850s, these ideas grew to what is now known as genetics. These have had tremendous influence on agriculture. The interplay between agriculture and statistics is so deep that several statistical concepts - applicable in various contexts- have names coming from agricultural experiments (treatment, split-plot design, ...). Another aspect where statistical ideas have had a major role is drug and vaccine discovery via randomized control trials (RCT). These are in news today during the time of COVID-19 pandemic and everyone is waiting for results of RCTs underway on various candidate treatments/vaccines.

Mathematical and Statistical ideas have played a role in understanding various things that we observe. Apart from natural sciences (physics, biology, medicine, agriculture), it has had a huge impact in behavioural sciences. So much so that 4 papers written by mathematician John Nash (on what is now called Game theory) led to his being awarded the Nobel Memorial Prize in Economics - though he never wrote any other paper on economics.

In this article, I am going to talk about statistical ideas and elections. One issue that I will be discussing is: What determines the accuracy of an estimate based on a random sample: *sample size* or *sampling fraction* (the latter is defined as ratio of sample size and population size). Most people intuitively feel that sampling fraction is what should determine accuracy. However, I will explain as to why that is wrong and it is the sample size that determines the accuracy.

One obvious connection of statistics and elections is opinion polls, in which I have been involved for over 2 decades. Talking about opinion polls to people often leads to two kinds of reaction : one of astonishment- as to how talking to such a small fraction (far less than 0.1%) of population can give us any insight into the ground reality. The other reaction is that of dismissing it completely- comparing it with astrology, saying if 20 astrologers make generalised predictions, one of them is going to get it right ! So no big deal. Well. My reply has always been that I cannot comment on poll conducted by others because the methodology of sampling and analysis is not published. What I can say with confidence is that our sampling methodology is good and our analytical framework is good and so mostly we should be correct though occasionally we may be off the mark. I will explain the methodology developed by us in the second section and will also give a complete account of the predictions made and the actual outcomes for a period of 10 years.

In the last section I will write about the issue of sampling of Electronic Voting Machines (EVM), to cross verify paper ballot count and EVM count, in order to show to the skeptics that EVMs have not been tampered with. I was a member of the three member team constituted by the Election Commission of India (ECI) and our report had been the basis of ECI's reply to the Supreme Court. I will explain our thinking on this issue.

**Author for Correspondence: E-mail: rlk@cmi.ac.in, rkarandikar@gmail.com*

[#]Lecture delivered in the Mathematics and Society Symposia during the INSA Anniversary General Meeting, 2019 at CSIR-NIO, Goa

How can a Sample of 40000 in a Country with 800 Million Voters Suffice

In this section, I am going to focus on the issue of accuracy of a sampling scheme. Let me start with an experiment that I have conducted at several places when I give a talk on my election work.

Suppose I have a box containing 100 slips of paper, all identical, folded. I tell the audience that each slip has a letter on it, **A** or **B**. I also tell the audience that 99 of them have one letter and one is an outlier, having the other letter - so either scenario 1: 99 are **A** one is **B**, or scenario 2: 1 is **A** and 99 are **B**. I mix the slips and walk to the audience, asking someone to draw one slip, open it and read the letter written on it to the audience. Suppose it turns out that it is **B**. Now I ask the audience to factor this information and make a guess as to which of the two scenarios they think it is : And invariably the answer is scenario 2: 1 is **A** and 99 are **B**.

Let us pause and think what is the thought behind most people picking scenario 2. Some will say it is obvious or it is common sense. Those who have studied probability/statistics would say, if scenario 1 was true, the probability of what was observed is 0.01, while if scenario 2 was true, the probability of what was observed is 0.99. So they pick scenario 2. If one thinks about it, similar thought process is at the back whenever we are making decisions under uncertainty.

Now let us change the experiment. This time there are 5 slips with one letter and 95 with the other. If we draw one slip (after mixing) as earlier, very likely the answer will be same (the observed letter being the dominant one), except that the person answering may not be that confident as the two probabilities now are 0.05 and 0.95, so the ratio of probabilities (of the observed event under the two scenarios respectively, called the likelihood ratio) has dropped to 19 from 99. What if instead of one draw we repeat the experiment 3 times, after each draw, put the slip back in the box and draw again. Now if we choose the majority rule, whichever letter occurs more in our trials, we say 95 of that letter - so if we have 2 or 3 **A**, our guess would be: 95 **A** and 5 **B**. Note that I have chosen an odd number of trials so we will not have a tie. Now a simple probability calculation (number of favorable cases as a proportion of total number of cases) will give the following: Let us call 95 **A** and 5 **B** as scenario 1 and 95 **B** and 5 **A** as scenario 2.

The probability of observing 2 or 3 **A**s under scenario 1 is :

$$\frac{95*95*95+3*5*95*95}{100*100*100}=0.99275 \quad (1)$$

while the probability of observing 2 or 3 **A**s under scenario 2 is :

$$\frac{5*5*5+3*95*5*95}{100*100*100}=0.00725 \quad (2)$$

so again the likelihood ratio is over 99 and we can confidently say that if we observed 2 or 3 **A**s, then we must have scenario 1.

Let us use the phrase that *accuracy is over 99%* if probability of observing the given event under one scenario is above 99% while that under the other scenario is less than 1%.

Now if instead we have 80 with one letter and 20 with the other, increasing the number of repeated trials would give us likelihood ratio of 99 or more and again we can go with the majority rule confidently. It can be seen with some computation, that repeating it 15 times would achieve accuracy of over 99%.

What if instead of 100 we had 10000 slips with 95% having one letter and 5% having the other. For 100 slips we needed 3 draws to achieve likelihood ratio of 99. With 10000 slips, would we need 300 (linear scaling) or 20 (square root scaling)? Well. Neither - 3 draws will still be enough. Here, the probability of observing 2 or 3 **A**s in 3 draws under scenario 1: 95% **A** and 5% **B** is :

$$\frac{9500*9500*9500+3*500*9500*9500}{1000*1000*1000}=0.99275 \quad (3)$$

while the probability of observing 2 or 3 **A**s under scenario 2: 95% **B** and 5% **A** is :

$$\frac{500*500*500+3*9500*500*500}{1000*1000*1000}=0.00725 \quad (4)$$

so again the accuracy is over 99%.

Comparing (1) and (3), we see that each factor in each term of (1) got multiplied by 100 to get (3) and hence the ratio is same. Now, little thought should convince the reader that even if we had a billion slips with 95% having one letter and 5% the other, drawing 3 slips successively after mixing and then going with the majority rule will still give us a likelihood ratio greater than 99 and posterior probability of being correct under non-informative prior remains 0.99275, in other words accuracy of over 99%.

Thus from this little experiment, we see that *sample size* (here 3) determines the accuracy and not the sampling fraction $\left(\frac{3}{100}, \frac{3}{10000}\right)$.

Moving to election contest, consider a constituency with K registered elec- tors. Suppose there are only two candidates **A** and **B**. Suppose their respec- tive supports are Ka and Kb. If we write the names of electors on slips of paper, one name per slip, mix the slips and draw one slip, seek her/his preference among **A** and **B** and record the same, put back the slip and repeat the experiment $n = 2m + 1$ times. The total number of possible outcomes of this experiment is K^n. Simple calculation would yield that number of cases out of these that have *exactly j supporters of A and n – j supporters of B* is

$$\binom{n}{j}(K_a)^j(K_b)^{n-j}.$$

Thus writing $p = \frac{K_a}{K}$ and noting that $\frac{K_b}{K} = 1 - p$, the proportion of total K cases that have at least $m + 1$ supporters of **B** and at most m supporters of **A** is

$$p_b = \sum_{k=0}^{m}\binom{2m+1}{k}p^k(1-p)^{(2m+1-k)} \tag{5}$$

and likewise the proportion of total cases that have at least $m + 1$ supporters of **A** and at most m supporters of **B** is

$$p_a = \sum_{k=m+1}^{2m+1}\binom{2m+1}{k}p^k(1-p)^{(2m+1-k)} = 1 - p_b. \tag{6}$$

Suppose the winner has at least 52% support *i.e.* $\max\{p, (1 - p)\} \geq \alpha = 0.52$. Using a simple python program one can obtain smallest m such that

$$\beta = \sum_{k=0}^{m}\binom{2m+1}{k}(1-\alpha)^k\alpha^{(2m+1-k)} \geq 0.99 \tag{7}$$

and then

$$\max\{p_a, p_b\} = \beta.$$

The smallest m such that (7) holds can also be obtained using *R*, using the built-in function *pbinom*. The smallest m turns out to be 1690 so that $n = 2m + 1 = 3381$, and then $\beta = 0.99016$.

Thus then for sample of size $n \geq 3381$, if $p \leq 0.48$ (so that **B** is going to win the election), $pb \geq 0.99$ and $pa \leq 0.01$ and if $p \geq 0.52$ (so that **A** is going to win the election), then $pa \geq 0.99$ and $pb \leq 0.01$.

Thus all we need to do is pick electors randomly from among all registered voters, obtain her/his preference and record the same and repeat the experiment $n = 3381$ times. Whoever gets more votes among these n electors, predict that the same candidate will win. As explained above, the accuracy of this prediction will be at least 99%.

Now in these calculations, the population size K got knocked off early on it has no role.

Of course if the election is a close one, say winner is getting just over 51% votes so that $\max\{p, (1 - p)\} \geq \alpha = 0.51$, we would need $n = 4105$ and if the election was even narrower, namely $\max\{p, (1 - p)\} \geq \alpha = 0.505$ so that winning margin is just about 1%, then we need $n = 16423$ to get accuracy of 99%. Thus we see that sample size matters and not the sampling fraction. The other factor that matters is the gap in the support for the two candidates.

In the discussion above we had considered what is called *Simple Random Sampling With Replacement* (SRSWR), where after drawing one item we put it back and draw again from the whole population. However, in practice, there is no point in choosing the same item again. At each stage we do not put back the item drawn and randomly pick from the rest. This is known as *Simple Random Sampling With Out Replacement* (SRSWOR). The resulting probabilities can be computed using hypergeometric distribution rather than binomial. The desired accuracy may be achieved in a little smaller n (as compared to SRSWR). The computations may be done using

python or *R*. It can be seen that if population size is 1,00,000 with one having 48% support while the other having 52% support, sample of size 3269 will suffice, while if the population size is 10,00,000, sample of size 3369 will achieve accuracy of 99%. Thus for SRSWOR, population size does have a minor impact on probabilities, and it can be proven that *n* obtained via SRSWR will always suffice whatever be the population (in this case, the *n* obtained for SRSWR was 3381).

We have discussed the problem of predicting the winner in a two candidate election. Now let us change the question. Suppose a candidate wishes to know her/his support level. If we say she/he is ahead in the race, and if so, then by what margin. So, if the proportion of electors who prefer this candidate is *p*, we would like to *estimate p*. Intuitively, if we have a random sample of size *n* where *n* is large, the observed proportion *pn* of supporters of this candidate among the sample is a good guess. It can be shown that it is the best estimate. And using the *central limit theorem* one can show that say when *n* is greater than say 200,

$$P\left(\mid p_n - p \mid > t\sqrt{\tfrac{p(1-p)}{n}}\right) \approx 2(1-\Phi(t)) \quad \forall t > 0 \tag{8}$$

where Φ is the distribution function of the standard normal distribution. Note that $\Phi(1.64) = 0.95$, $\Phi(1.96) = 0.9975$, $\Phi(2.58) = 0.995$. For $0 \leq p \leq 1$, $p(1 \leq p) \leq \frac{1}{2}$, and thus we can conclude that

$$P\left(\mid p_n - p \mid > \frac{1.29}{\sqrt{n}}\right) \leq 0.01 \tag{9}$$

Thus for example with $n = 4161$, we have

$$P(\mid p_n - p \mid > 0.02) \leq 0.01. \tag{10}$$

The above relations are true irrespective of the population size and are con- servative estimates if population is not large and sampling is without replacement. The relation (10) is often expressed as *the margin of error is 2% at 99% confidence.*

Remember, all the calculations given above that lead to the *statistical guarantee* that if $\hat{p}_n$ is the observed proportion of supporters for a party in a random sample of size *n* and *p* is the true proportion of supporters for the party, then the conclusion (9) relies on the sample being chosen randomly. This can be summarised as *Most samples with size n* = 4161 *(say) are representative of the population and hence if we select one randomly, we are likely to end up with a representative sample*.

This statistical guarantee does not kick in if the sample is not a random sample. In colloquial English, the word *random* is also used in the sense of *arbitrary* (as in Random Access Memory-RAM). So often people think of a random sample as any arbitrary subset. Indeed, once in an interview I was asked that how come I rely on data from CSDS in my work on opinion polls (Center for Studies in Developing Societies, more about CSDS later) since they declare on their web site that they do a Randomized survey!

Let me also clarify that using auxiliary information one can improve upon the accuracy that simple random sample provides for a given sample size. To illustrate, suppose we wish to estimate support among students in a large school towards a provision proposed by the authorities. If based on some discussions the perception is that opinions among boys and girls is different, it would be better to distribute the sample proportionately among boys and girls and have a random sample among boys and a random sample among girls of the required size. So if boys are 55% and girls are 45% among all students, and a sample of size 300 is planned, we should choose a random sample of size 165 among boys and random sample of size 135 among girls. This is called stratified sampling - the girls here constitute a strata and the boys constitute another strata. Here the gain in accuracy is significant if the difference in opinion in the two strata is significant. Thus stratification can help but once the strata are identified, one must pick the respondents randomly within each strata.

Opinion Polls in the Indian Context

Let me now come to to predicting outcome of elections in India. I will discuss National elections (elections to Lok Sabha, the lower house of the Indian Parliament) based on Opinion polls but same ideas apply to predicting outcome of elections to a state legislature.

As in all applications of statistics to real world problems, one must start with the question : *What is the objective of such an exercise?* The second step

is to examine what data is available, and what additional data can be obtained towards fulfilling the objective. The third step is to use domain knowledge to build a model that links the data that we would have with the outcome of interest.

In the national elections, the interest of the public is in number of seats won by major parties and not percentage of votes polled by major parties. This is so because it is the number of seats in the parliament that determines as to who forms the government.

So clearly *the objective of the opinion poll is to predict number of seats for major political parties (or alliances)*. Let us note that the country is di- vided among 543 constituencies, each one electing its representative to the Lok Sabha. The candidates could be affiliated to a political party or be independent.

We now come to relevant data that is available and can be collected. As seen in the previous section, if we can get a random sample of size 4200 in each of the 543 constituencies, we can predict each outcome and also the total number of seats. But this would involve a sample of over 22 lakh (2.2 million) voters- which is very difficult given the resources (money and trained manpower). So clearly we need to do something more.

Each Lok Sabha constituency is made up of several segments, each segment being a constituency for the state legislative assembly (or Vidhan Sabha). In the 2019 Lok Sabha polls, the 543 Lok Sabha Constituencies were divided into 4125 assembly segments. Each segment is further divided into polling booths, the number in each segment varies from 8 to 1023, only 468 out of 4125 segments had less than 100 booths and only 23 segments had over 500 booths. The total number of booths across the country was 10,37,496.

We have the historical data on elections - for each constituency, the votes obtained by each candidate and the party affiliation of the candidate, if any. This is also available at assembly segment level. Since the introduction of Electronic Voting Machines, the data is also available at the pooling booth level, though not easy to obtain.

Discussions with political scientists or political activists would reveal the following ground realities of Indian politics. Socio-economic attributes such as gender, religion, caste, education, economic status do influence propensity to vote. To put it differently, let us classify the population on any of the attributes or a combination there of (say into $= 1, 2, \ldots m$ classes) and there are n major parties, and create a two way table with $a(i, j)$ being the proportion of the population in the *ith* class who intend to vote for the *jth* party. Thus rows represent socio-economic class and columns represent political parties. Then the rows of the matrix $a(i, j)$ would be quite different from each other. Likewise, each state is a separate political entity. The political history of a state influences the voting intentions of the residents a lot. Thus, if we take two constituencies on the boundary of, say, Tamilnadu and Karnataka, while the profile of the two constituencies may be very similar, the political preferences are likely to be very different. In the constituency in Karnataka, the main two contenders could be national parties while in the constituency in Tamilnadu, it is likely that the regional parties hold sway. One can see similar differences in neighbouring constituencies in Bihar and UP.

Let us observe that estimating vote shares across the country will not suffice. There is no magic formula that would yield number of seats for parties if we know percentage votes of parties.

The seats won would depend upon the distribution of votes across constituencies. For example, suppose there are only four parties A, B, C and D and their nationwide support levels are 38%, 33%, 17% and 12%. If the votes are distributed across the country in such a way that the parties have exactly same level of support in each constituency, then party A will win all the seats as it has highest support in every constituency! Of course this is unlikely to happen. It is also possible that a party with maximum votes nationwide among all parties may not get the largest number of seats. This has not happened in Indian parliamentary elections, but has happened in some state legislature elections. In Karnataka 2018 election for Vidhan Sabha (Leg- islative Assembly) elections, while the Indian National Congress got 38.14% votes and 80 seats, The Bhartiya Janata Party got 36.35% votes and 104 seats! This is because in each constituency, the candidate getting maximum votes wins. It does not matter if the winning candidate

won by a thin margin or a huge margin.

Thus to estimate number of seats in an upcoming Lok Sabha election, one needs to estimate not only the vote share for the parties across the country but also need a measure of how these votes are distributed.

Let us briefly examine as to what is done in other countries. The US political system is very different- it is presidential with winner take all at state level (for most states). Our system, with winner take all at a constituency level, is similar to the one in UK, perhaps because it was based on the model in UK. I examined carefully the methodology adopted by psephologists in UK. The profile of constituencies on socio-economic attributes is readily available in UK and is an important input in the prediction model. In India, the Census data provides population profile on (most) socio-economic attributes but this is organised at district level. And there is no mapping of districts to constituencies- most constituencies consist of parts of two or more districts. Thus the socio-economic profile of population in a constituency is not available.

There is one more crucial difference. Borrowing a term from finance, let us define *volatility of public opinion* (over a period of time, say 5 years) as the proportion of the electorate who change their voting intention over last 5 years. Based on expert opinion and some supporting data, it is clear that volatility in UK is rather low. In India, based on expert opinion and backed by data that we had collected years ago, volatility is very high. We had estimated in 1998 that volatility over 2 weeks just preceding the Lok Sabha election was about 30%. These two factors mean that methods used in UK are not appropriate for Indian scenario, even though the electoral systems are the same.

One must remember that electrons will spin the same way in India as they do in UK but the same cannot be postulated about human minds. Thus for a physical phenomenon, if a model works in the west, we can accept it as it is. But if a phenomenon involves human choices and human interactions, a model that is working well in the west can at best be a good starting point. One has to test it and attempt to improve it with local data.

Back to Indian elections. As a first step, taking into account the ground realities and limitations of data, we came up with a model for predicting votes for major alliances in every constituency.

We will consider pre-poll alliances rather than parties since during the last three decades, we have seen pre-poll alliances capture the largest block of seats in the parliament. So wherever a major party has not entered into a pre-poll alliance, the word alliance refers to this stand alone party.

Suppose the states are numbered from 1 to m, major alliances from 1 to k and constituencies from 1 to n. Let $w(i, j)$ denote the votes polled by candidate of *jth* alliance in *ith* constituency in the previous Lok Sabha election and let $z(i)$ be the total votes in *ith* constituency in that election. Let $s(i)$ be the index of the state to which *ith* constituency belongs. Let $x(s)$ be the total number of valid votes cast in the *sth* state in the previous Lok Sabha elections and let $u(t, j)$ be the total votes for the candidates of the *jth* alliance out of $x(t)$ (in the previous elections in the *sth* state). Then

$$x(t) = \sum_{i:s(i)=t} z(i), \qquad u(t, j) = \sum_{i:s(i)=t} w(i, j)$$

and the vote share (percentage of votes) of *jth* party in the state s is

$$\theta(s, j) = \frac{u(s, j)}{x(s)}.$$

Suppose we have conducted an opinion poll and assuming that it is a fairly representative poll in each state, let $y(t)$ be the total number of votes in the *sth* state in the opinion poll and let $v(s, j)$ be the total votes for the candidates of the *jth* alliance out of $y(s)$ (in the same opinion poll). The vote share (percentage of votes) of *jth* party in the state s in the opinion poll is

$$\pi(s, j) = \frac{u(s, j)}{y(s)}.$$

Let us note that $v(s, j)$ has binomial distribution and when $y(s)$ is large, say greater than 50, $\pi(s, j)$ has normal distribution.

Let $\xi(s, j)$ denote the (estimated) change in vote share for *jth* alliance in *sth* state from last election,

also called *swing*, then:

$$\xi(s, j) = \pi(s, j) - \theta(s, j).$$

Now our model is : *swing or the change in vote share for each alliance in a constituency* is constant across a state. This can be called the uniform swing model. The rationale for this model is: from one election to next, the socio-economic profile of constituencies does not change drastically and so the differences among constituencies due to these factors are already captures in the previous election.

With this model, we immediately have an estimate of vote share (or percent- age votes) $p(i, j)$ of j^{th} alliance in i^{th} constituency:

$$p(i, j) = \left(\frac{w(i, j)}{z(i)}\right) + \xi(s(i), j)$$

recall $s(i)$ is the index of the state to which *ith* constituency belongs.

Yet another way of describing this model is- *estimated level of support for an alliance is as in the opinion poll, and the way this support is distributed across a state is as in the previous election.*

This model can be modified by dividing large states into smaller politically homogenous regions, (often geographical), based on expert opinion. Suppose *sth* state is divided into $g(s)$ regions ($g(s)$ can be 1). We can model that the swing (change of votes) in a constituency from previous election is a convex combination of swing across the region and the swing across the state to which the constituency belongs, and the coefficient is dependent on the state: let $r(i)$ be the region in the state $s(i)$ to which *ith* constituency belongs. Our modified model is:

$$p(i, j) = \left(\frac{w(i, j)}{z(i)}\right) + a(s(i)) * \rho(r(i), s(i), j) + (1 - \alpha(s(i))) * \xi(s(i), j)$$

where $\rho(r, s, j)$ is the swing in *rth* region in *sth* state for party j given by

$$\rho(r, s, j) = \frac{\sum_{s(i)=s, r(i)=r} v(i, j)}{\sum_{s(i)=s, r(i)=r} y(i)} - \frac{\sum_{s(i)=s, r(i)=r} w(i, j)}{\sum_{s(i)=s, r(i)=r} z(i)}$$

The coefficients $\alpha(s)$ are chosen based on expert opinion, $0 \leq \alpha(s) \leq 1$.

Thus to come up with an estimate of vote shares of major parties/alliances in each constituency under our swing model, we need the following:

(i) votes for major parties/alliances in each constituency in the previous election along with the total votes polled

(ii) estimated vote shares across states and regions for major parties/alliances from the opinion poll.

We will need to conduct an opinion poll where the sample in each region is representative of that region.

One of the methods used for surveys in the context of opinion polls has been telephonic surveys, where the investigators choose telephone numbers randomly from a list and call to get opinion of respondents. However, if used in India, this method would end up having a sample where rural, economically weaker sections of the society are under represented and thus the resulting sample is not representative of the population. So we have avoided this.

Another method of sampling used by market research agencies is called quota sampling. In this, one starts with population profile on attributes that seem to have an influence on voting intention, typically socio-economic attributes such as age, gender, cast, religion, education level, and try to get a sample whose profile comes close to that of the population on these attributes. The hope is that since profiles match on these attribute, the voting intentions would also match. But in my view, this is only a hope and it may be way off the mark.

Over last 5-7 years, several people who know me have said to me, why don't you move on. Door-to-door survey is methodology from the past. In current times, use social media data or run an opinion poll on twitter - Facebook and within hours you can generate a very large sample. Well, yes but we can be sure that the sample so collected will not be representative of the Indian electorate. Rural, underprivileged class will be absent or have minimal presence in the chosen sample and thus outcome will be unreliable. This shows that the current trend

of just pushing available data on some com- putational engine without giving thought as to what is it that the collected data represents could lead to incorrect conclusions.

I still recommend randomly chosen sample chosen from the voters list followed by a door-to-door survey.

I was roped into opinion poll analysis as a statistical consultant in 1997 by Yogendra Yadav, who then was professor at CSDS - Center for Studies in Developing Societies. We had developed or rather refined the sampling methodology being used by CSDS. Later from 2005 when the news channel CNN-IBN was launched, till the 2014 Lok Sabha elections, I and CSDS worked as a team and were commissioned by CNN-IBN and our findings were aired on CNN-IBN. Later I will give the full record of our work over these 10 years.

The core interest of CSDS team in this exercise has been to understand political behaviour and analyse its connection with socio-economic background.

Thus it was important to get a representative sample across the country as well as states (and the regions identified within large states). It was clear that we need to take states as strata. I will describe the sampling scheme we had arrived at. The same has been used by CSDS over the years.

Let us now come to how the list of voters is organised. In each constituency, there is a list of polling booths and for each polling booth, there is list of voters for that booth. The voters list contains name, address, gender and age.

If we are targeting say about 30000 voters nationwide (this is chosen based on resources available), we can choose 108 constituencies (one in five), in each chosen constituency, 8 polling booths and in each chosen booth 35 voters.

The list of constituencies is organised such that each state is a cluster and thus if we do a circular random sampling or systematic random sampling to choose 108 constituencies. In this, we draw a random number between 1 and 543, say we get 292. We pick constituency number 292 and then every 5^{th} going upto 542. Then we skip 543, 1, 2,3, and pick 4, and then again every 5^{th}. So we imagine that 1 to 543 are arranged in a circle with 543 following by 1 and the simply skip 4 pick 5^{th}. It can be seen that all clusters will be approximately proportionately represented.

In each constituency we have a list of polling stations which are numbered by the election commission in such a way that a large neighbourhood is covered by booths with consecutive numbers, and if we pick booths via circular random sampling, we will have a nice spread across the constituency. Finally, within a booth, a household is a cluster, the neighbours are next cluster, one building in a society is one cluster. Thus we again pick respondents using circular random sampling in each booth.

This design can be called, multi-stage circular random sampling. We have seen that over the years, we have generally got a representative sample. We see from the above discussion that the sampling design to be used depends upon how the sampling frame (information about the population) is made available to us. Also, it be noted that we do not make efforts to balance or match profile on any socio-economic attribute. We trust that randomness will ensure that profiles would match reasonably on all attributes. We verify this on socio-economic variables, and if the sample profile on these variables is not approximately same as that of the population, it indicates that something is not right-either sampling has not been done as specified or there has been a data entry problem and we correct the same.

Now once we have a sample, using past data and the mathematical model described above, we can estimate the vote share that each party is likely to get in the next election in each constituency. Now one way to predict seats for parties nationwide is to simply predict that the party getting largest vote share in a constituency will win the seat and then count the winners to get nationwide seats. However, it is clear that in constituencies where the leading two parties have predicted vote share of 43% and 34% respectively, we will be confident in calling the winner while this will not be the case in another constituency where the leading two have predicted vote shares of 37% and 36% respectively. Somehow this confidence has to be accounted for while estimating predicted seats. One way of doing this is the following.

Let us fix a constituency and let party *a* have

the maximum predicted vote share in that constituency and party b be the runner-up, with Y_a and Y_b representing the respective predicted vote shares. Let $X = Y_a - Y_b$. Since the swing in each state and region for each party has normal distribution (indeed, the joint distribution of all swings has a multivariate normal distribution), it follows that X has normal distribution. The mean μ would be positive if the leading candidate is winning and would be negative if the winner is the candidate who is trailing. If we use a normal prior with mean zero and variance γ^2 for the unknown parameter of interest μ then the posterior distribution of μ would be normal with mean X and variance $\xi^2 = \text{var}(X) + \gamma^2$. Thus posterior probability that the leading candidate in this constituency will win is

$$\Phi\left(\frac{X}{\xi}\right).$$

where Φ is the distribution function of standard normal distribution. The posterior probability that candidate who is second in race will win is

$$\left(1 - \Phi\left(\frac{X}{\xi}\right)\right).$$

The choice of ξ would be done based on the sample size for the vote share estimates. For $\xi = 6$, the constituency where leading two parties have predicted vote share of 43% and 34% respectively, the posterior probabilities are 93% and 7% while if we are less confident about our vote share estimates and choose $\xi = 9$, then the posterior probabilities are 84% and 16%. The constituency where the predicted vote shares are 37% and 36% respectively, the probabilities with $\xi = 6$ are 57% and 43% and with $\xi = 9$ are 54% and 46%.

Adding posterior probabilities of win for an alliance across all constituencies (in India for Lok Sabha and in a state for Vidhan Sabha) yields a good estimate of seats for the alliance. Here the parameter $\hat{\imath}$ has to be chosen by the analyst using experience and the sample size for the poll.

The methodology described above is essentially what we (I and the CSDS team, lead by Yogendra Yadav and Sanjay Kumar) have followed both for national and state elections. We have generally chosen ξ in the range 6 to 10.

There are some variations which I will now describe. In the swing model, like regions, we could bring in division of state by other attributes such as rural/urban or reserved/general. We could even bring in phase of election, as in large states, elections are held in 4 or more phases these days. Also, in assigning posterior probability of win, we could bring in top 3 parties instead of top 2.

We also have to factor in making and breaking of alliances, merger or split of parties and arrival of new parties. Using expert opinion, we try to simulate a distribution of votes in the scenario where the alliance/party picture is as it is today while the voters preferences were as they were at the time of the previous election. This distribution is then used instead of actual votes ($w(i, j)$) in the swing model.

In my view, opinion poll conducted well before polling begins can at best yield a snapshot of the mood of the nation at the time of survey and cannot predict as to how the mood will sway in the remaining time. Some pollsters do tracking poll, namely conduct polls say every week leading to the voting day and then extrapolate the trend from the last poll to the voting day. This assumes that the trend is linear which I think is not valid. There is far bigger churn in the last few days. Thus we have refrained from doing so. Further, any such survey gives mood of the entire electorate while what matters is the 60-70% (approximately) voters who go and cast their vote. Thus the predictive power as far as final outcome is concerned for such pre-election opinion polls is rather low (even with the best of methodology).

Exit polls address both these issues - respondents are interviewed as they exit the booth after voting. However, randomisation of respondents is very difficult if not impossible and as a result other biases may creep in to our sample. We have generally avoided this.

We prefer what we call post-poll - survey after the voting is over. Given the multi phase elections which have become the norm in India, we interview people between one and three days after they have voted (but well before counting day) in all but the last phase. For the last phase we could do an exit poll or make the final projection one day after polling is over.

I will now give a complete account of our predictions on CNN-IBN over a ten year period,

2005-2014 - as long as this arrangement lasted, with CSDS conducting the poll (a post-poll) and me doing the analysis as far as seats are concerned. These are based on post-poll as explained above.

Bihar Legislative Assembly, 2005 (October)

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
JDU-BJP	36	37	127-137	147
RJD+	31	31	72-80	65
Others	33	32	29-39	31

• We Got it right!
• We were the only poll projecting Majority for JDU-BJP combine.
• Predicted vote share within the 3% margin of error.
• Underestimation of the winning alliance's seats

Assam Legislative Assembly, 2006

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
Congress	31	31	52-60	53
BJP	11	12	10-15	10
AGP	22	20	25-31	24
Others	36	37	26-35	39

• We got both votes and seats right.
• Most others got it wrong.

Tamil Nadu Legislative Assembly, 2006

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
AIADMK+	35	40	64-74	69
DMK+	45	45	157-167	163
DMDK	10	8	2-6	1
Others	10	7	0	1

• Got seat projections on the dot.
• We had over estimated AIADMK+ seats.
• Some others had predicted AIADMK victory.

Kerala Legislative Assembly, 2006

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
LDF	51	49	107-117	98
UDF	41	43	25-31	42
Others	8	8	0-1	0

• We got vote share right, within error bounds.

• Seats were off the mark.

West Bengal Legislative Assembly, 2006

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
LF	53	50	230-240	235
INC	16	15	17-23	24
TMC+	27	29	32-40	31

• We hit bulls eye as far as seats is concerned.
• Vote estimates within the 3% error margin.

Punjab Legislative Assembly, 2007

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
SAD-BJP	41	45	50-60	68
Congress	41	41	50-60	44
Others	18	14	3-9	5

• We got it wrong. Failed to predict SAD-BJP victory.
• Underestimated SAD-BJP vote share.

Uttarakhand Legislative Assembly, 2007

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
Congress	31	30	21-27	21
BJP	34	32	33-39	35
Others	35	38	8-12	14

• We Got it right and so did many others!

Uttar Pradesh Legislative Assembly, 2007

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
SP	25	25	99-111	97
BSP	29	30	152-168	206
BJP+	22	18	80-90	52
Congress	11	9	25-33	22
Others	13	18	21-27	26

• We failed to predict BSP getting absolute majority.
• Vote share estimates were on the dot for SP, BSP.
• Did predict that BSP is far ahead of SP unlike many others.

Gujarat Legislative Assembly, 2007

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	47	49	92-100	11
Congress+	42	39	77-85	76
Others	11	12	3-7	23

• Got vote shares right, within 3% margin of error.
• Predicted BJP victory, under estimated the seats.

Karnataka Legislative Assembly, 2008

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	30	34	79	110
Congress	35	35	86	80
JDS	21	19	45	28
Others	14	12	14	6

• Under estimated BJP votes.
• Failed to predict BJP getting more seats than congress.
• Very difficult situation to predict - Congress got more votes while BJP got lot more seats.

Lok Sabha 2009

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
UPA	36	36	210-225	262
NDA	28	24	180-195	159
Others	36	40	130-146	122

• Over estimated NDA vote share.
• Under estimated UPA seats.
• We could predict that UPA can form government without Left support.

Bihar Legislative Assembly, 2010

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
JDU-BJP	46	39	185-201	206
Congress	9	8	6-12	4
RJD-LJP	27	26	22-32	25
Others	18	27	9-19	8

• Got it right. Most other polls were suggesting a closer contest.
• We were the only poll showing JDU-BJP could touch 200.
• We were the only poll which showed RJD-LJP could be below 30.
• Overestimated the vote share for JD-BJP but the modelling error and sampling error were in opposite direction and neutralised each other!.

Assam Legislative Assembly, 2011

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
Congress	36	39	64-72	78
BJP	9	11	7-11	5
AGP	18	16	16-22	10
AIUDF	13	13	11-17	18
Others	24	21	12-20	15

• We got it right! The only poll to predict Congress victory. All experts and other polls were predicting defeat for the Congress, maximum 45 seats.
• Vote shares within the 2-3% margin of error.

Kerala Legislative Assembly, 2011

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
LDF	46	45	69-77	68
UDF	45	46	63-71	72
Others	9	9	0	0

• We predicted close finish, though picked wrong winner.
• Others were predicting much bigger victory for UDF.

Tamil Nadu Legislative Assembly, 2011

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
DMK+	44	39	102-114	31
AIADMK+	46	52	120-132	203
BJP Front	3	2	0	0
Others	7	7	0	0

• We had said comfortable victory for AIADMK+, though hugely under- estimated votes and seats.
• Some other surveys were predicting DMK victory!

West Bengal Legislative Assembly, 2011

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
Left	40	41	60-72	62
TMC+	50	48	222-234	227
Others	10	11	0	5

• We got it right!
• Most others predicted Majority for TMC+.
• We were on the dot in predicting seats.

Uttarakhand Legislative Assembly, 2012

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
Congress	39	34	31-41	32
BJP	32	33	22-32	31

• Both seat predictions were good, actual outcome within the range.
• We had overestimated the Congress votes by nearly 5%.

Punjab Legislative Assembly, 2012

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
SAD+BJP	41	42	51-63	68
Congress	40	40	48-60	46

• We predicted the winner.
• Others predicted a comfortable victory for SAD-BJP, but race was tight as votes show.
• With a 2% vote advantage, SAD-BJP got 22 more seats!

Manipur Legislative Assembly, 2012

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
Congress	30	42	24-32	42
TMC	14	17	7-13	7

• First ever poll in North East, got the winner right.
• We underestimated votes and seats for Congress.
• We predicted that TMC will be number 2, and regional parties would get wiped out.

Uttar Pradesh Legislative Assembly, 2012

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
SP	34	29	232-250	224
BSP	24	26	65-79	80
BJP+	14	15	36-44	47
Congress	12	12	28-38	28

• One of our biggest success stories.
• We were the only ones predicting SP getting absolute majority.
• Over estimated SP votes by 5%.

Gujarat Legislative Assembly, 2012

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	48	48	129-141	116
Congress+	36	39	37-45	60
Others	16	13	4-10	6

• Predicted BJP victory and so did others.
• Under estimated Congress votes and seats.

Himanchal Pradesh Legislative Assembly, 2012

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
Congress	41	43	29-35	36
BJP	40	38	29-35	26

• Vote estimates were within 3% margin.
• Error in vote estimated translated in under estimation of Congress seats.

Karnataka Legislative Assembly, 2013

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	23	20	39-49	40
Congress	37	37	117-129	122
JDS	20	20	34-44	40
Others	20	23	14-22	21

• Actual seats within predicted range.
• Vote estimates within 3% error margin.

Madhya Pradesh Legislative Assembly, 2013

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	41	45	136-146	165
Congress	35	36	67-77	58
Others	24	19	13-21	7

• Predicted massive victory for BJP.
• Under estimated votes and seats for BJP.

Rajasthan Legislative Assembly, 2013

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	43	45	126-136	162
Congress	33	33	49-57	21
Others	24	22	12-20	16

• Predicted massive victory for BJP.
• Under estimated votes and seats for BJP.

Chhatisgarh Legislative Assembly, 2013

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	42	41	45-55	49
Congress+	38	40	32-40	39
Others	20	19	7-13	2

• Vote vote and seat predictions right.

Delhi Legislative Assembly, 2013

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
BJP	33	34	32-42	31
Congress	23	24	09-17	8
AAP	27	30	13-21	28
Others	17	12	1-5	3

• Got the vote shares right and overall story right.
• The seats for AAP were underestimated.

Lok Sabha 2014

	Estimated vote share (%)	Actual vote share (%)	Projected seats	Actual seats
NDA	40	40	274-286	336
UPA	26	23	92-102	60
Others	34	37	157-168	147

• Over estimated UPA vote share.
• Under estimated NDA seats.
• Like most others, predicted comfortable NDA victory. One poll did far better than us and predicted accurately seats for NDA.

So by my own assessment, we were off the mark on 4 occasions (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 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.

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.

Several other media entities and poll agencies have conducted polls across these years. Since the methodology is not publicly available, (we do have partial information about sample size, sometimes about sample profile and sometimes partial information about sampling methodology), I cannot com- ment on them. All I can say is that across time, we have a very good track record, as the above exhaustive list of all prediction made by us during November 2005 and May 2014.

These outcomes show the power as well as limitations of opinion polls. With correct methodology, one can mostly get the basic story right- who will get largest number of seats, would this party (or alliance) cross the half way mark comfortably or would it be around that number or will it fall short of majority mark significantly.

Moreover, the opinion polls give an insight into why people have voted the way they did- what were the issues that decided their vote etc. CSDS brings out detailed studies on such questions based on the opinion polls. This is the only way one can get insights into the mind of the voter. I will conclude this section with some more observations on polls in Indian context.

One question about any survey is what do we do with missing observations, namely respondents who answered other questions but refused to answer the question on who they would vote for. Related question is - can we tell if the respondent is telling truth about who they would vote for and if it is a lie - can we correct the same.

Answer to both the questions above consists in trying to deduce preference of the voter from the socio-economic attributes of the respondents and answers to some of the other political questions, such as important issues that decided their vote, performance of government, should the government get another chance, etc. We have tried to fit such models to voting behaviour and our conclusion is that it does not predict voting intention well enough to discard a response if it does not match with prediction from our model and replace it by what the model yields. We have discarded missing values and taken the answer given by respondent about voting intention as it is.

I am also asked as to why we ignore the candidates in our analysis. While our surveys reveal that at least 20-30% voters do consider candidate as the single most important factor when choosing whom to vote, it is difficult to incorporate the same in our swing model. And our belief is that if an alliance/party has fielded large number of poor candidates, whom even those who normally support the alliance/party find unacceptable, the effect of such choice on voting intention would be reflected in our survey.

I had earlier mentioned that while opinion polls represent the entire population, exit polls are able to pick respondents who have voted. How do we handle this in post-poll? We ask respondents, apart from the other questions including which candidate they prefer in their constituency, if they voted and if they say no, we ignore their preference as far as predicting seats is concerned. CSDS takes into account their answer in other analysis. However, we found that while our sample profile matched the population on socio-economic attributes, always our sample would have far higher percentage of voters who said they voted than the voter turnout on election day. So after we ask if they have voted, we ask the respondents to show if the indelible ink is still there on their finger. Of course, we tell them that this question is to test

efficacy of the ink. And we ignore all those whose finger does not show the ink mark.

One of the standard techniques used in opinion polls is to what is called correct for bias using voter recall. All respondents are asked whom they voted for in the last election. An imbalance in this as compared to actual outcome of the previous election is seen as bias in the sample and correct the same. So for example, in the US, if the sample in a state has 53% respondents recalling that they voted for democratic party candidate while actually only 46% had voted for the candidate, then the vote share in the sample for the 53 democratic candidates in multiplied by $\frac{46}{53}$ to get the corrected estimate.

However, we have found that in India, whoever won the election last time seems to have a much higher recall than actual vote share. This has been so across states and across the years. Even when whoever won last time is no longer popular and is about to lose big. Thus, while the Left front had 49% vote share in 2006 and in 2011 they were down to 41% (our estimate was 40%), over 58% respondents seem to recall having voted fo the left front candidate! Like wise in Tamil Nadu in 2011, DMK with its allies were on way out paving way for a massive victory for AIADMK, but much larger percent than the DMK+ vote share (45%) in 2006 seem to recall having voted for them. So there seems to be a psychological undercurrent that make people believe the they voted for the winner. This is the only explanation for the observed phenomenon. In any case, explanations apart, we have not used the correction for recall in Indian context. We focus all our effort in getting a representative sample via randomisation.

The EVM-VVPAT Controversy

In the year leading to the 2019 Lok Sabha elections, there were many questions raised on Electronic Voting Machines (EVM) used in India for elections. To give the background, let me recall the history of EVMs in India.

Election commission had taken initiative to introduce EVMs and the first ever usage was in Paravoor assembly constituency in Ernakulam district of Kerala in the assembly elections in 1982 (in 50 out of 123 booths). The Congress had been a big supporter of EVMs, but the Congress candidate lost the election and challenged it in Kerala High court. The High court did not accept his plea and this was challenged in the supreme court. The supreme court set aside the election and ordered re-poll in the booths where EVMs were used on the grounds that Representation of The People Act, which governs elections, did not have provision for EVM.

The said act was amended in 1989 and use of EVM in elections was given legal status. The Election Commission of India (ECI) involved faculty from IIT Bombay to design the EVMs and Public Sector Undertakings (PSU) BEL and ECIL to manufacture the EVMs.

The EVMs were used in June1999 elections to the Goa assembly without a hitch. In the mid-term poll for Lok Sabha that followed in September-October 1999, the election commission decided to hold poll for 46 constituencies spread across 17 states entirely using EVMs. Subsequently, in 2003 the election commission announced that from then on, all elections to state legislatures and the parliament will be held using EVMs. Thus the 2004 Lok Sabha poll was held entirely using EVMs. In 2009 Dr Subramanian Swamy had filed a petition in the Delhi High court seeking a directive from the court to the ECI to include a paper trail of the vote cast by a voter. The High court refused to intervene and the judgement was challenged in the supreme court. One of the arguments had been that such paper trail would mean that in case a recount was ordered by the courts, the paper ballots can be counted and the count matched against the EVM count.

The ECI informed the Supreme Court that they have been conducting trials of Voter Verifiable Paper Audit Trail (VVPAT) system designed by experts and that the same had been used in bye election to Noksen (ST) Assembly Constituency in the State of Nagaland in 2013. The ECI reported that the trial had been successful and that it envisages it introduction across the ountry, once the budget for manufacture of the large number of VVPAT machines is approved and the same can be manufactured. Based on this assurance, the Supreme court did not pass any directives.

Thus it is to be noted that while the petition had only asked for introduction of VVPAT and the ECI

had promised the same, there had been no demand for suo-moto counting of paper trail ballots, unless court orders as a result of election petition.

The ECI, in order to install confidence in general public, decided that in each assembly segment, one booth will be selected at random and the paper ballots would be counted and matched with EVM before declaration of results.

However, dis-satisfied candidates went on asking that there should be more VVPATs should be cross verified. Petitions were filed in various courts seeking direction to ECI to count VVPAT paper ballots in 10% or 20% of booths. This demand went on increasing and culminated with the demand for cross- checking of 50% of booths and was backed by all opposition parties just before the 2019 Lok Sabha elections. In 2018, the ECI had sought expert opinion from statistics experts on the number of booths that need to be sampled to be confident that EVMs are performing fine. I was a member along with my former colleague Abhay Bhatt from Indian Statistical Institute. The committee constituted by ECI also had Onkar Prosad Ghosh, of the Central Statistical Office, Government of India.

The Technical expert committee was confident about the design and so was the ECI based on its past experience. I would like to mention that the EVMs do not have any networking hardware. There is no ethernet port, no wifi or bluetooth capability, and thus it is not possible to alter or tamper the memory remotely.

Other important feature is that the names of candidates on the EVM appear in an order that is determined by the same convention that had been followed since the sixties for the order on the ballot paper. First, the candidates of the national parties appear, in an alphabetical order of their names, then candidates of state parties (again in an alphabetical order of their names) and lastly the rest, again in an alphabetical order of their names. Thus, the order gets determined only after the last date of withdrawal of nomination. This makes it impossible to centrally tamper with EVMs, even if it were feasible.

The EVMs are distributed across the constituencies via randomisation. The EVMs used in India consist of two units: BU, the balloting unit and CU, the control unit. The BUs and CUs are distributed independently and on the day of polling, the two are connected. If one of them has been tampered or replaced, the handshake between the two units will fail and the pair will have to be replaced. Nonetheless, the count of paper ballots in some chosen EVMs was to statistically install confidence in public in general and in judiciary (subsequent to the spate of court cases) that EVMs are tamper resistant.

So the question we considered was the following. What should be the same size *n*, so that if a random sample of size *n* does not discover any mismatches - the ECI can confidently claim that EVM malfunctioning, if any, is negligible.

We can draw an analogy from legal domain. When police charge a person with a serious crime, say murder, the premise that the judge has to start with is that the accused is not guilty. Police present the evidence collected, which may include matching blood sample, or finger prints, testimony of an eye witnesses and such other evidence that they may have. Of course today, the evidence can include DNA fingerprint matching, but of course, the defence may offer some other explanation. Rarely would we have evidence that (mathematically) proves that the accused indeed committed the murder. The law states that the police have to prove beyond reasonable doubt that the accused is the murderer. What this means is the evidence should be such that the chances that the accused is not the murderer and yet such evidence emerges is very low.

So here, since ECI wishes to claim that EVMs are tamper proof, the ECI should produce evidence, such that the chances of its occurrence is very low if indeed EVMs were tampered.

So the crux is to decide (or choose) two small numbers $\alpha > 0$ and $p^* > 0$ and then obtain n such that *if a random sample of size n is chosen among all the polling booths in use in the country during Lok Sabha elections, and if the count of paper ballot in these n booths matches the EVM count, then the ECI can claim with probability at least* $(1 - \alpha)$ *that proportion of defectives, if any, is less than* $p^* - > 0$.

Hers is roughly what we thought on choice of α and p^*: If someone is able to come up with a way of

tampering the EVMs, the tampering will be on such a scale so as to tilt significantly the national outcome. So, our contention was that if we can confidently claim that less than 2% of the EVMs have some fault (due to manufacturing defect, procedural lapse by officials or due to tampering) we can conclude that there is no systematic tampering of EVMs across the country. This threshold of 2% can be debated but this is what we took and clearly stated in our report. The level of confidence $(1 - \alpha)$ we took was very high- 99.993665752%. (α = 0.00006334248). How did we come up with this number- α = 0.00006334248 is the probability of observing a deviation of over 4ξ when sampling from a normal population. This is of course extremely rare.

So the choice of n is determined by the equation:

$$(1-p)^n \leq \alpha \forall p \geq p^*$$

or in other words

$$n = \log(\alpha) / \log(1 - p^*).$$

This yields n = 479.

Here we have ignored the fact that sampling will be SRSWOR and thus ignored the finite sample correction. We can compute the exact probability using hyper-geometric distribution using the exact number of polling booths in the county (which was 10,38,000) and see the n = 479 is the correct choice.

I would like to re-emphasise the point made earlier, that sample size is what is relevant and not the sampling proportion; a fact overlooked by some critics.

The election commission, while agreeing with out approach and analysis had a different twist. They wanted to stick to its policy of one booth chosen at random in each assembly segment and wanted us to evaluate the confidence if this policy yielded the conclusion that no booth is tampered with high confidence. Indeed our calculations showed that *if there are at least 1% defective EVMs spread across the constituencies in whatever possible way, the chances that ECs sampling scheme will fail to detect a single defective EVM is less than 0.00000001*.

Thus, if EC adopts one randomly drawn EVM per constituency and finds that there are no defectives in the chosen sample, then with 99.99999% confidence we can conclude that proportion of defectives is less than 1% in the population.

The final hearing in the Supreme Court on the petition by all the opposition parties seeking direction to ECI to count paper ballots in 50% booths was held on April 8, 2019. Our report was presented to the court, as part of the final affidavit by the ECI. The Supreme court accepted this report, but suggested to ECI to increase counting paper ballots in five randomly chosen booths per segment instead of just one.

We informed ECI that if follows the Supreme Court's suggestion of verifying 5 randomly chosen EVMs in each assembly segment and if no defective is found, then ECI can say with 99.9999999999% confidence that the proportion of defectives is less than 0.5%, that the proportion of defectives is less than 0.25% with 99.9999980958 % confidence and that that the proportion of defectives is less than 0.2% with 99.9998479152 % confidence. Indeed, in our calculations, we have been very conservative and thus the confidence level is at least what is reported here, and is probably much higher.

This episode shows that very simple mathematics along with understanding of the domain can have very important policy ramifications.