

Policy Watch

No. 7



Winning Voter Confidence: Fixing India's Faulty VVPAT-based Audit of EVMs

K. Ashok Vardhan Shetty



THE HINDU CENTRE

for

Politics and Public Policy

© The Hindu Centre for Politics & Public Policy, 2018

The Hindu Centre for Politics and Public Policy, Chennai, is an independent platform for exploration of ideas and public policies. As a public policy resource, our aim is to help the public increase its awareness of its political, social and moral choices. The Hindu Centre believes that informed citizens can exercise their democratic rights better.

In accordance with this mission, The Hindu Centre's publications are intended to explain and highlight issues and themes that are the subject of public debate, and aid the public in making informed judgments on issues of public importance.

Cover Photo: View of an Election Counting Centre during the 2018 Assembly Elections in Bengaluru on May 15, 2018. Photo: Sampath Kumar G P

All rights reserved.
No part of this publication may be reproduced in any form
without the written permission of the publisher.

Winning Voter Confidence: Fixing India's Faulty VVPAT-based Audit of EVMs

K. Ashok Vardhan Shetty

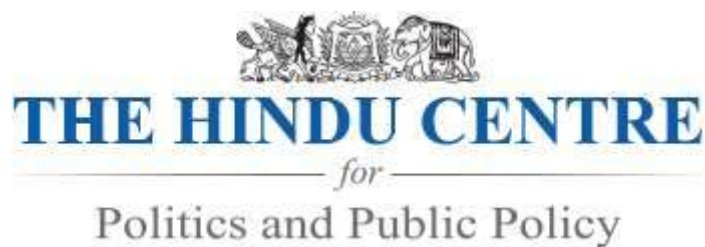


TABLE OF CONTENTS

I	INTRODUCTION	1
II	SOME ODDITIES OF STATISTICAL SAMPLING	5
III	HYPERGEOMETRIC DISTRIBUTION MODEL: AN EXACT FIT FOR EVM SAMPLING	10
IV	THE ‘ONE EVM PER ASSEMBLY CONSTITUENCY’ FALLACY	14
V	ECI MUST SET THE CONTROVERSY AT REST	23
VI	ANNEXURE I	25
VII	ANNEXURE II	26

ABSTRACT

As the world's largest democracy gears up for a season of elections, including the 2019 General Election, there is an urgent need to examine the integrity of the electoral process. Electronic Voting Machines (EVMs) are 'black boxes' in which it is impossible for voters to verify whether their votes have been recorded correctly, and counting mistakes and frauds are undetectable and unchallengeable.

The 'voter verified paper audit trail' (VVPAT) is an additional verifiable record of every vote cast that allows for a partial or total recount independent of the EVM's electronic count. It is a critical safeguard that can help detect counting mistakes and frauds that would otherwise go undetected. The success of the VVPAT audit, however, depends on a proper, statistically acceptable, and administratively viable sample plan.

The Election Commission of India (ECI)'s prescription of a *uniform* sample size of just "*one polling station (i.e. one EVM) per Assembly Constituency*" for all Assembly Constituencies and all States stirs up an avoidable controversy and diminishes voter confidence. The ECI has not made public as to how it arrived at this sample size, and it has also *not clearly specified the population to which this sample size relates*. The latter is important because in the event of a defective EVM turning up in the sample, the hand counting of VVPAT slips will have to be done for all the remaining EVMs of the *specified population*.

In this Policy Watch, **K. Ashok Vardhan Shetty**, a former Indian Administrative Service (IAS) officer, demonstrates that the sample size prescribed by the ECI for VVPAT Audit is a statistical howler that fails to conform to fundamental sampling principles, leading to very high margins of error which are unacceptable in a democracy. By failing to detect outcome-altering miscounts due to EVM malfunction or fraud, it defeats the very purpose of introducing VVPAT. Spending hundreds of crores of rupees on procurement of VVPAT units makes little sense if their utilisation for audit purposes is reduced to an exercise in tokenism.

This report suggests statistically correct—and administratively viable—sample sizes to eliminate the risk of electoral fraud and infuse public confidence in the electoral process. It suggests ways in which the ECI can set the controversy at rest and make a beginning with the elections for 5 States whose counting is scheduled for December 11, 2018.

I. INTRODUCTION

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”¹

H.G. Wells

[1866-1946]

Electronic Voting Machines (EVMs) have many advantages including ease of operation, reduction of invalid votes cast and the speeding up of counting. But they also have some glaring disadvantages. EVMs are ‘black boxes’ in which it is impossible for voters to verify whether their votes have been recorded and counted correctly. There is always some risk of the votes cast being lost due to equipment malfunction. Electronic recounting is meaningless because it will simply yield the same total. Contrary to the claim by the Election Commission of India (ECI), even under election conditions and with all the security features and administrative safeguards in place, it is still possible for a determined attacker, acting in collusion with insiders, to tamper with EVMs and steal votes on a scale large enough to change election outcomes². The problem with EVMs is that counting mistakes and frauds are undetectable and the losers are left with no means to challenge the results.

It follows that EVMs are not fully reliable and there should be an additional verifiable physical record of every vote cast. This is called the ‘voter verified paper audit trail’ (VVPAT). After a voter casts his vote, he gets to view for a few seconds - before it drops into a box - a printed paper slip so that he can verify if his vote has been recorded correctly. It provides a back-up in case of loss of votes due to equipment malfunction, and allows for a partial or total recount of the paper slips independent of the electronic count. In 2013, the Supreme Court passed an order mandating the use of EVMs with VVPAT units and directed the ECI to implement them in a phased manner.

The importance of conceptual clarity

VVPAT is an additional safeguard, a very critical, and final safeguard, which can help detect counting mistakes and frauds that would otherwise go undetected. But VVPAT, by itself, cannot

prevent EVM malfunction or tampering. If it is to have any real security value, it should be backed by a proper *sampling process*. This involves 4 steps:

- (1) Defining the *population*³ clearly in terms of ‘population units’ (polling stations or EVMs) and ‘population boundaries’ (e.g. Assembly Constituency, Parliamentary Constituency, State, country). The population size varies depending upon how the boundaries are set.
- (2) Determining the correct sample size, or what is called the *statistically significant sample size*, of EVMs whose VVPAT slips will be hand counted. The sample size should not only be statistically sound but also administratively viable.
- (3) *Random sampling* of the EVMs, preferably by draw of lots by the candidates or their authorised representatives on the counting day.
- (4) A ‘*decision rule*’, based on the sample results, to determine whether the election results can be declared or the hand counting of VVPAT slips should be done for all the remaining EVMs of the population. The latter entails additional time and effort but is justified by the need to declare the election results correctly without any outcome-altering miscounts due to EVM malfunction or fraud. Two types of decision rules are possible:
 - a) Comparison of the EVM electronic count and the VVPAT hand count for the sample of EVMs to verify if (i) *the two totals tally*, and (ii) *the votes secured by the leading candidate tally*. If both tally, then there is no problem and the election results based on the EVM count can be declared⁴. But if any one or both do not tally, then there is a problem and the hand counting of VVPAT slips should be done for all the remaining EVMs of the population and the election results declared only on the basis of the VVPAT count.
 - b) Adoption of ‘*Lot Acceptance Sampling*’, a statistical quality control technique widely used in industry and trade the world over for assuring the quality of incoming and outgoing goods. The decision, based on *counting the number of defectives in a sample*, can be to accept the lot, reject the lot, or even, for sequential sampling schemes, to take another sample and then repeat the decision process.
An ‘acceptance number’ - ‘c’ - is specified. If the number of defectives found in the sample is less than or equal to ‘c’, the lot is accepted; otherwise, the lot is rejected. Unlike industry and trade where the presence of a few defectives in the sample may be tolerated depending upon the size of the lot and the quality norms, in the election context, the acceptance number ‘c’ will have to be *zero*.

In other words, the election results can be declared only if no ‘defective EVM’⁵ is found in the randomly drawn sample of EVMs. If even a single defective EVM is detected in the

sample⁶, the hand counting of VVPAT slips should be done for all the remaining EVMs of the population and the election results declared only on the basis of the VVPAT count.

The second option is preferable and easier to implement. For the rest of this paper, it will be assumed that this decision rule will be followed.

Unfortunately, the issue of sampling procedure for VVPAT-based audit of EVMs has received scant attention by policy-makers, the academic community, and most importantly, the voting public in India until recently⁷. This Policy Watch aims to point out the statistical weakness of the procedure that is in place and make the case for statistically significant sample sizes that are also administratively viable. VVPAT-based audits are the final check and remedy against electoral fraud. The ECI, which oversees the largest electoral exercise in the democratic world should ensure that this audit is both infallible and statistically acceptable, and correctly reflect voter-choice.

The error of uniform sample size

The ECI has courted controversy by prescribing a *uniform* sample size of “*one polling station (i.e. one EVM) per Assembly Constituency*” for all Assembly Constituencies and all States. This sample size was adopted in the Assembly Elections for Gujarat and Himachal Pradesh held in November-December 2017; for Meghalaya, Nagaland and Tripura held in February 2018; and for Karnataka held in May 2018.

For reasons best known to it, the ECI has not made public as to how it arrived at this sample size, and it has also *not clearly specified the population to which this sample size relates*. The latter is important because in the event of a defective EVM turning up in the sample, the hand counting of VVPAT slips will have to be done for all the remaining EVMs of the *specified population*.

A mistake with grave consequences

As we shall demonstrate shortly, the sample size prescribed by the ECI is a statistical howler that fails to conform to scrutiny of statistical principles, leading to very high margins of error which are unacceptable in a democracy. It is open to legal challenge on this score. It defeats the very purpose of introducing VVPAT and is fraught with all the risks of conducting elections with paperless EVMs.

In something as important as ensuring the integrity of the election process – a process which in any case takes about 2-3 months from the date of announcement to the date of counting – a delay of a few hours or even a couple of days in hand counting VVPAT slips of a larger sample of EVMs should not matter at all. Spending hundreds of crores of rupees on procurement of VVPAT units makes little sense if their utilisation for audit purposes is reduced to an exercise in tokenism. This could result in the easily avoidable perception that the ECI is afraid that pro-active implementation of VVPAT may show up many EVMs to be defective and raise a question mark about the sanctity of the election process.

II. SOME ODDITIES OF STATISTICAL SAMPLING

“The mind is not designed to grasp the laws of probability, even though the laws rule the universe.”⁸

Steven Pinker

[Johnstone Family Professor of Psychology, Harvard University]

Statistical sampling is fundamental to almost all of our understanding of the world. It provides a means of gaining information about a population without the need to examine the population in its entirety. The latter is usually neither cost-effective nor practicable. No estimate taken from a sample is expected to be exact, and there is likely to be some difference between the sample estimate and the actual population value. ‘Confidence level’ is how certain one wants to be that the population value is within the sample estimate and its associated margin of error. The purpose of statistical sampling is to draw conclusions about a *suitably defined population* on the basis of the *most economic sample* for a *specified level of confidence* in the results.

If I were to tell a layperson that (for a given set of parameters) the sample size required for a population size of one lakh is 458 but the sample size required for a population size of one crore (100 times greater) is only 459, he is likely to think that I am mistaken. It seems counter-intuitive but that is the way statistical sampling theory works! As population size (N) increases, the sample size (n) also increases but at a much slower rate and ‘hits a plateau’ beyond some point so that *further increases in population size have no effect on the sample size*. The following example illustrates how sample size varies with population size.

Let us assume that one per cent of the EVMS used in an election are defective. [It must be remembered that a ‘defective EVM’, according to our definition, is one which has a *mismatch* between the EVM count and the VVPAT count]. Random samples are drawn *without replacement*.⁹ Detecting a defective EVM is treated as a ‘success’. The sample sizes required, for various population sizes, *for 99 per cent probability of detecting at least one defective EVM* are shown in **Table 1**, and are also displayed graphically in **Chart 1**. [All Tables and Charts compiled by author.]

Table 1**How Sample Size varies with Population Size**

<i>Population Size (N)</i>	<i>Sample Size (n)</i>	<i>% of n to N</i>
100	99	99
200	180	90
500	300	60
1,000	368	36.8
2,000	410	20.5
5,000	438	8.76
10,000	448	4.48
20,000	453	2.27
50,000	457	0.91
1,00,000	458	0.46
2,00,000	458	0.23
10,00,000	459	0.05
20,00,000	459	0.02
1,00,00,000	459	0.005

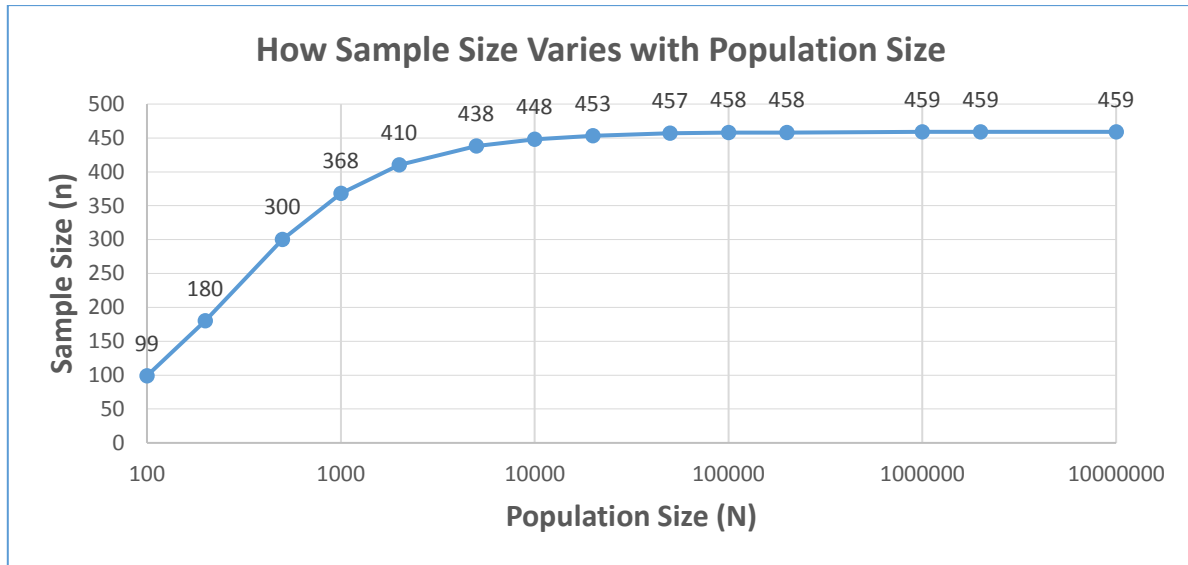
Source: Compiled by author using Hypergeometric Distribution.

It is seen that when the population size of EVMs is 100, the sample size is 99 i.e. *it is nearly as big as the population size*. When the population size is 1,000, the sample size is 368 and when the population size is 10,000, the sample size is 448. But the ‘sampling fraction’ (n/N) i.e. the sample size relative to the population size is seen to decrease rapidly. The sample size then ‘hits a plateau’ and increases to only 458 for a population size of one lakh; to only 459 for a population size of ten lakhs, and remains at 459 even for a population size of one crore. *In other words, for big populations, the population size is irrelevant to sample size.*

Chart 1 makes the point clearer. [To avoid the crowding of figures at the lower end and for ease of visualisation, the figures are plotted on a logarithmic scale]. In this particular example, it is seen that *increase of population size beyond about 10,000 ($N/n > 20$) has little or no impact on the sample size.*

Chart 1

Graphic Representation of Table 1



The figures in **Table 1** also tell us how *statistical sampling is superior to arbitrary, non-statistical sampling* such as, say, a flat “10 per cent sample” ($n=0.1N$). With statistical sampling, the sample size required is 99 for a population size of one hundred, and just 459 for a population size of one crore. But with a flat “10 per cent sample”, for a population size of one hundred, the sample size is 10 which is too small and statistically incorrect; and for a population size of one crore, it is 10 lakhs which is too big and administratively impractical. Thus, a flat “10 per cent sample” is utterly wrong for small population sizes and is utterly inefficient for very big population sizes.

As Robert Schlaifer, author of a classic text on Statistics, puts it:

“One of the most common ‘vulgar errors’ concerning sampling is the belief that the reliability of a sample depends upon its percentage relationship to the population. Many businessmen operate sampling inspection plans which call for inspection of a certain percentage of each lot – usually 10 per cent. . . however, this policy is completely misguided: unless the sample takes in a really substantial fraction of the population, its reliability depends on its absolute rather than its relative size.”¹⁰

The relevance of the foregoing discussion to VVPAT-based audit of EVMs should be obvious. In the election context, *depending upon how the population is defined, the population size can vary widely as shown in Table 2* below.

Table 2**How population is defined and its effect on population size**

<i>Population Boundary</i>	<i>Population Size (N) (Number of EVMs)</i>
Assembly Constituency	$\approx$ 30 to 300
Parliamentary Constituency	$\approx$ 300 to 1800
A State as a whole	Ranging from 589 (Sikkim) to 1,50,000 (U.P) For 9 States $N < 10,000$ For 20 States $N > 10,000$
India as a whole	$\approx$ 10,00,000

$\approx$ is the symbol for ‘approximately equal’.

The importance of defining the ‘population’

Studying the figures in **Table 1** and **Table 2** together, it is obvious that if the EVMs used in an *Assembly Constituency* are defined as the population, the population size (N) will be very small; the sampling fraction (n/N) will be very big; and the sample size (n) will vary considerably across Assembly Constituencies. The same is true if the EVMs used in a *Parliamentary Constituency* are defined as the population.

If the EVMs in a State as a whole are defined as the population, there is considerable variation in population size from the very small (Sikkim) to the very big (Uttar Pradesh). For the nine smaller States with population size less than 10,000 EVMs, the sampling fraction (n/N) will be quite big and the sample size will vary considerably across the States. For the 20 bigger States with population size greater than 10,000 EVMs, the sample size will ‘hit a plateau’ in the 450s and further increase in population size will have little or no effect on it.

If the EVMs used in India as a whole are defined as the population, due to the ‘plateau effect’, the sample size is just one more than that for U.P.

Chapter 4 will elaborate upon these points and explain why the uniform sample size of “one EVM per Assembly Constituency” for all Assembly Constituencies and all States presently adopted by the ECI is completely off the mark, and with serious implications.

The ECI’s critics have not fared any better. They are also guilty of committing the ‘vulgar error’ (to use Robert Schlaifer’s telling phrase) of demanding arbitrary, non-statistical sample sizes like

**WINNING VOTER CONFIDENCE:
FIXING INDIA'S FAULTY VVPAT-BASED AUDIT OF EVMS**

“10 per cent of the EVMs per Assembly Constituency” for VVPAT-based audit of EVMs. This is precisely what Congress leader Kamal Nath did in a writ petition filed before the Supreme Court¹¹.

Other critics of the ECI have demanded “15 per cent samples” and even “25 per cent samples” under the mistaken impression that a “bigger percentage” guarantees greater accuracy of results. It does not. What guarantees greater accuracy of results is a statistically significant sample size based on a properly defined population and the appropriate probability distribution model.

III. HYPERGEOMETRIC DISTRIBUTION MODEL: AN EXACT FIT FOR EVM SAMPLING

“Probability theory is nothing more than common sense reduced to calculation”.

Pierre-Simon Laplace

[French Mathematician, 1749-1827]

Consider the following two problems:

A: There are 100 fish in a pond. 95 of them are grey and five are green. The fish are caught without replacement. The characteristic of interest here is a green fish, catching which is treated as a ‘success’. If we catch a random sample of, say, three fish, what is the probability that the sample will contain at least one green fish?

B: There are 100 EVMs in an Assembly Constituency. 95 of them are good while five are defective. The characteristic of interest here is a defective EVM, detecting which is treated as a ‘success’. If we pick a random sample of, say, three EVMs, what is the probability that the sample will contain at least one defective EVM?

Problems *A* and *B* are exactly equivalent. They are both classic examples of what is called a *Hypergeometric Probability Distribution*. The probabilities can be calculated using the standard formula for Hypergeometric Distribution¹² or using Excel or an online calculator¹³ or any of the statistical analysis software.

The answer to problems *A* and *B* is that there is only a 14.4 per cent probability of the sample size of three having at least one ‘success’¹⁴.

If we wish to be 99 per cent sure of having at least one ‘success’, then the sample size should be increased to 59¹⁵.

The Hypergeometric Distribution model is an ‘exact fit’ to the EVM problem and should form the basis of the sampling plan for VVPAT-based audit of EVMs¹⁶.

In the fish problem, if the number of green fish in the pond is large, say, 50 out of 100, then it is easy to catch a green fish even if you cast the net narrow. But if the number of green fish in the

pond is very small, say, only five out of 100, then you will have to cast the net much wider in order to catch a green fish.

Therefore, with the Hypergeometric Distribution, as the proportion (P) of the 'characteristic of interest' in the population decreases, the sample size (n) required for detecting at least one 'success' increases. Applied to VVPAT-based audit of EVMs, it means that *the sample size (n) required for detecting defective EVMs is the biggest when the proportion of defective EVMs (P) is assumed to be very small and it gets smaller when P gets bigger*. **Table 3** and **Chart 2** (compiled by the author) make this point clear.

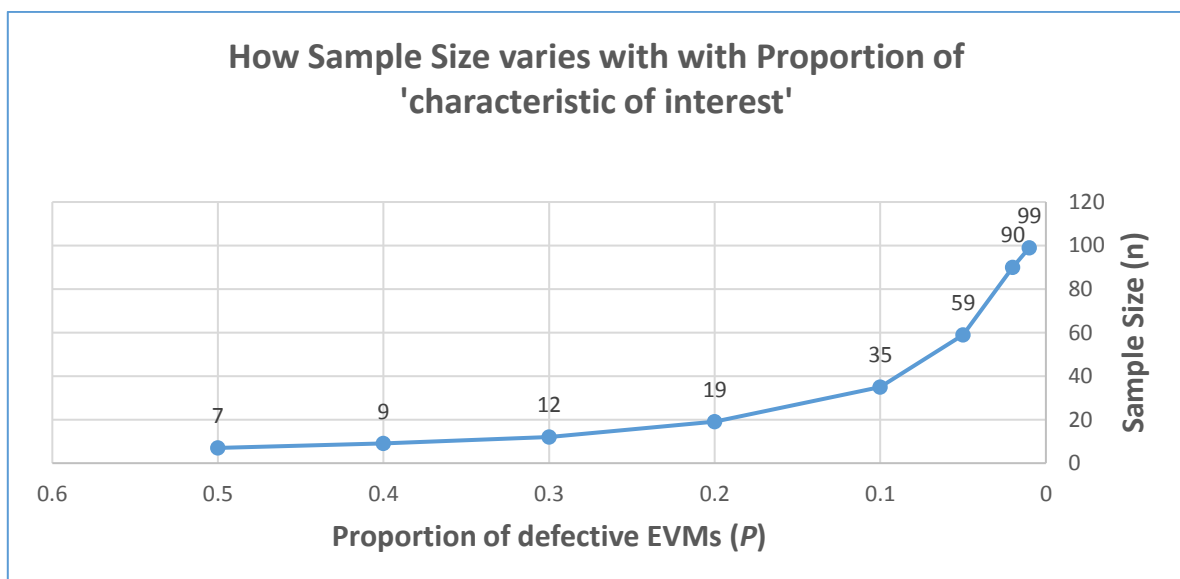
Table 3

How Sample Size varies with the Proportion of the 'characteristic of interest'

Population Size (N) = 100 EVMs.

Proportion of defective EVMs (P)	Number of defective EVMs in the population	Sample Size (n) required for 99% probability of detecting at least one defective EVM in the sample
0.50	50	7
0.40	40	9
0.30	30	12
0.20	20	19
0.10	10	35
0.05	5	59
0.02	2	90
0.01	1	99

Chart 2



In the case of EVMs employed in an election, the proportion of defective EVMs (P) is *unknown*. It may be zero or 0.01 or 0.02 or 0.10 or whatever. The ECI thinks that P is zero or very close to zero. But just because EVM tampering didn't take place in the past, we can't assume that it won't take place sometime in the future. So even if P was zero or very close to zero in the past, there is no guarantee that it won't be high in the next election. Any debate on the precise value of P is bound to be uninformed and therefore, inconclusive as each one's guess would be as good as the other's.

With the Hypergeometric Distribution model, the debate about the precise value of P is inconsequential because the sample size is the greatest when P is very close to 0 (which is what ECI claims it is), and it becomes lesser as P increases. *So, the sample size calculated for $P = 0.01$ (one per cent) will hold good for all higher proportions of defectives. It therefore obviates the need to make questionable assumptions about the value of P or estimate it based on the data of past trials which may or may not be fully reliable.*

When can rigging be 'successful'

A question may be asked as to why we should not assume a value for P that is less than one per cent, as then the sample size required will be even bigger. *The following thought experiment will show that the actual value of P required for the successful rigging of an election, even in a neck-to-neck contest, needs to be much higher than one per cent.*

In India, the average number of polling stations (N.B. There is one EVM per polling station) per Assembly Constituency is around 240. The actual number of polling stations in an Assembly Constituency varies widely from State to State and sometimes even within a State - from about less than 30 to about 300-plus polling stations. In what follows, the figures are hypothetical but the logic holds good, even if we assume different sets of figures.

On an average, a polling station has about 900 voters attached to it out of whom about 65 per cent may vote. That means about 600 votes may be cast in a typical EVM. Not all of the votes can be 'stolen' (i.e. transferred to the winning candidate) by tampering with the EVM. There are *practical limits* to the maximum percentage of votes of an EVM that may be 'stolen' without attracting the ECI's adverse attention. Let us assume that this is about 20 per cent of the votes cast i.e. 120 votes. Consider an Assembly Constituency where the election is expected to be very close. Let us assume that the contest is only between the candidates of the two main parties and the rest don't matter,

**WINNING VOTER CONFIDENCE:
FIXING INDIA'S FAULTY VVPAT-BASED AUDIT OF EVMS**

and that the votes are 'stolen' only from the rival candidate of the other main party. Clearly, it is not sufficient to tamper with just one EVM to be sure of victory when the number of votes that can be 'stolen' is only 120.

A potential attacker may have to tamper with at least five EVMs in an Assembly Constituency to 'steal' at least $(120 \times 5) = 600$ votes from his rival candidate, which would make him reasonably sure of victory. Even in a large-sized Assembly Constituency with 300 EVMs, five EVMs work out to 1.5 per cent of the total EVMs; for an average-sized Assembly Constituency with 240 EVMs, it is 2.1 per cent of the total; for an Assembly Constituency with 100 EVMs, it is five per cent of the total; for even smaller Assembly Constituencies, the percentage is much higher.

So, our assumption of "one per cent defective EVMs" as the value for P is itself on the lower side, and will yield the most conservative (i.e. biggest) sample size that is adequate for our purpose. Let us recall that for higher values of P, the sample size required is *smaller*.

IV. THE ‘ONE EVM PER ASSEMBLY CONSTITUENCY’ FALLACY

“A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions.”¹⁷

– M.J. Moroney

[Facts from Figures, 1951, p 3]

In Statistics, there are no hard-and-fast rules as to how a population should be defined except that (i) the boundaries of the population should clearly separate items which are of interest to us from items which are not, and (ii) the sampling process is administratively viable.

We now proceed to show that whereas the boundaries for the population of EVMs can be an *Assembly Constituency*, or a *Parliamentary Constituency*, or a *State as a whole*, or *India as a whole*, only one of these populations [a State as a whole] is administratively viable.

It must be remembered that in the event of a defective EVM turning up in the chosen sample of ‘n’ EVMs, *the hand counting of VVPAT slips will have to be done for all the remaining (N – n) EVMs forming part of the population.*

Let:

- W_n represent the administrative workload involved in hand counting VVPAT slips for the chosen sample of ‘n’ EVMs, and
- $W_{(N-n)}$ represent the administrative workload involved in hand counting VVPAT slips of all the remaining (N–n) EVMs in the population.

There has to be a trade-off between W_n and $W_{(N-n)}$. As we shall demonstrate presently, if W_n is small, $W_{(N-n)}$ is big and *vice versa*. Both cannot be small. *The ECI is at liberty to define ‘population’ suitably as long as it is commonsensical and represents the right balance between the administrative workloads W_n and $W_{(N-n)}$.*

**WINNING VOTER CONFIDENCE:
FIXING INDIA'S FAULTY VVPAT-BASED AUDIT OF EVMS**

In all the scenarios that follow, we assume a very low proportion of defective EVMs ($P =$ one per cent or 0.01) and work out the sample sizes required, using the Hypergeometric Distribution model, for 99 per cent probability that the sample will detect at least one defective EVM.

1. **EVMS of an Assembly Constituency as ‘population’:** Let us assume four hypothetical Assembly Constituencies A, B, C and D with 50, 100, 200 and 300 polling stations (EVMS) in them respectively. The results are shown in **Table 4**.

Table 4
Sample Sizes if EVMS of an ASSEMBLY CONSTITUENCY
are the Population

Assembly Constituency	Population Size (N) [Total number of polling stations in the constituency]	Number of defective EVMS in the population @ $P = 0.01$	Sample Size (n) required	% of n to N	<i>Probability that the ECI - prescribed sample size of “one EVM per Assembly Constituency” will fail to detect a defective EVM</i>
A	50	1 [#]	50	100	98%
B	100	1	99	99	99%
C	200	2	180	90	99%
D	300	3	235	78.3	99%

- rounded off to the next highest integer.

EVMS employed in an Assembly Constituency would seem to be the logical choice of ‘population’ for Assembly Elections. But it is seen that *the resulting sample sizes are nearly as big as the respective population sizes leaving little or no scope for statistical sampling!* We may as well have paper ballots and count them 100 per cent instead of having EVMS and hand-counting the VVPAT slips of between 78.3 per cent and 100 per cent of EVMS in each Assembly Constituency!

Moreover, in the event of a ‘defective EVM’ turning up in the chosen sample, the number of the remaining EVMS in the population whose VVPAT slips need to be counted i.e. $(N - n)$ is very less in this case. But this advantage is more than negated by the fact that the sample sizes are nearly

as big as the population sizes. In other words, workload W_n is enormous even if workload $W_{(N-n)}$ is very less.

So, EVMs used in an Assembly Constituency are not an appropriate choice for ‘population’.

The last column of **Table 4** shows why the ECI-prescribed sample size of “one EVM per Assembly Constituency” is utterly wrong. *The probability that the sample will not detect a defective EVM is 99 per cent!*¹⁸ (It is 98% for Assembly Constituency A only because of the rounding off).

2. EVMs of a Parliamentary Constituency as ‘population’: A Parliamentary Constituency typically comprises about six Assembly Constituencies and may have between 300 and 1800 polling stations. Consider four hypothetical Parliamentary Constituencies P, Q, R and S with 300, 600, 1200 and 1800 polling stations in them. The results are shown in **Table 5**.

Table 5
Sample Sizes if EVMs of a PARLIAMENTARY CONSTITUENCY
are the Population

Parliamentary Constituency	Population Size (N) [Total number of polling stations in the constituency]	Number of defective EVMs in the population @ P = 0.01	Sample Size (n) required	% of n to N	<i>Probability that the ECI - prescribed sample size of “one EVM per Assembly Constituency”[#] will fail to detect a defective EVM.</i>
P	300	3	235	78.3	94.1%
Q	600	6	321	53.5	94.1%
R	1200	12	381	31.75	94.1%
S	1800	18	405	22.5	94.1%

- This works out to a sample size of six EVMs per Parliamentary Constituency as per ECI norms.

EVMs employed in a Parliamentary Constituency would seem to be the *logical* choice for ‘population’ for Parliamentary Elections. But it is seen that the resulting sample sizes are very big relative to the respective population sizes and do not serve the purpose of statistical sampling i.e. workload W_n involved in the hand counting of VVPAT slips for the chosen sample size (n) is enormous. In the event of a defective EVM turning up in the chosen sample, the number of the

remaining EVMs in the population whose VVPAT slips need to be counted, $(N - n)$, is also quite large i.e. workload $W_{(N-n)}$ is also considerable.

So, EVMs of Parliamentary Constituency are not an appropriate choice for ‘population’. It is not administratively viable on both counts [W_n as well as $W_{(N-n)}$]. The last column of **Table 5** shows why the ECI-prescribed sample size of “one EVM per Assembly Constituency” is seriously wrong even in this case. *The probability that it will fail to detect a defective EVM is 94.1 per cent.*

3. EVMs used in a State as a whole as ‘population’: Let us consider the five States that will have Assembly Elections in November-December 2018 – Mizoram, Chhattisgarh, Telangana, Rajasthan, and Madhya Pradesh. The results are shown in **Table 6**.

Table 6
Sample Sizes if EVMs of a STATE AS A WHOLE are the Population

State	Number of Assembly Constituencies	Population Size (N) [Total number of polling stations in the State]	Sample Size (n) required for the State as a whole	% of n to N	Average Number of EVMs per Assembly Constituency whose VVPAT slips should be hand counted	Probability that the ECI-prescribed sample size of “one EVM per Assembly Constituency” [#] will fail to detect a defective EVM
Mizoram	40	1164	370	31.79	10	65.6%
Chhattisgarh	90	23672	455	1.92	5	40.3%
Telangana	119	32574	455	1.40	4	30.1%
Rajasthan	200	51796	457	0.88	2	13.3%
Madhya Pradesh	230	65341	457	0.70	2	9.9%

- This works out to a sample size of 40 EVMs for Mizoram as a whole, 90 EVMs for Chhattisgarh as a whole, 119 EVMs for Telangana as a whole, and so on as per ECI norms.

As the population size of EVMs is very small for Mizoram, the sampling fraction (n/N) is big but this is inevitable. For the remaining 4 States, the sampling fraction is very reasonable and is administratively viable. The average number of EVMs to be hand counted per Assembly Constituency is also indicated (fractions rounded off to the *next higher integer*). It is seen that the administrative workload W_n involved in the hand counting of VVPAT slips for the chosen sample size is minimal.

Since the sample size is for *a State as a whole*, in the event of a defective EVM turning up in the chosen sample, the VVPAT slips of all the remaining EVMs in the population (*throughout the State*) will need to be hand counted and not just EVMs of the particular Assembly Constituency in which the defective EVM was detected. The workload $W_{(N-n)}$ involved in the hand counting of VVPAT slips for the remaining $(N - n)$ EVMs is considerable. As already indicated, there has to be a trade-off between W_n and $W_{(N-n)}$; both can't be small. Whereas W_n is unavoidable, $W_{(N-n)}$ is contingent upon a defective EVM being discovered which may be rare. It is preferable to have a small or reasonable W_n and a large $W_{(N-n)}$ than *vice versa*.

Moreover, the purpose of VVPAT is not just to detect fraud but also to deter it. The knowledge that if a defective EVM turns up, full hand count of VVPAT slips of all EVMs will be done is a sufficient deterrent for any likely fraudster. It will also put pressure on the two EVM manufacturers (Bharat Electronics Limited and Electronics Corporation of India Limited) to improve the quality of their EVMs and VVPAT-units so that instances of malfunctioning of EVM or VVPAT unit are negligible.

The average number of EVMs to be hand counted per Assembly Constituency, which is just 'two for Rajasthan and Madhya Pradesh, may seem 'very small' and create a doubt in the mind of a layperson about its correctness. But when it is remembered that the sample size is for the "State as a whole" [457 for both States] and that the discovery of even a single defective EVM *anywhere in the State* among the sample of 457 will entail the hand counting of VVPAT slips of all the remaining EVMs in all the Assembly Constituencies of the State, our layperson will realise that the sample size is correct.

The last column of **Table 6** shows why the ECI-prescribed sample size of "one EVM per Assembly Constituency" is seriously wrong even in this case. *The probability that it will fail to detect a defective EVM varies from 9.9 per cent for Madhya Pradesh to 65.6 per cent for Mizoram.*

4. **EVMS of India as 'population':** The results are shown in **Table 7**:

Table 7

Sample Size if INDIA AS A WHOLE is the Population

Unit	Number of Assembly Constituencies in India	Population Size (N) [Total number of polling stations in India]	Sample Size (n) required for India as a whole	% of n to N	Average Number of EVMS per Assembly Constituency whose VVPAT slips should be hand counted	<i>Probability that the ECI-prescribed sample size of "one EVM per Assembly Constituency" # will fail to detect a defective EVM</i>
INDIA	4120	10,00,000	459	0.045	0.11 [rounded off to 1].	Almost ZERO

- This works out to a sample size of 4,120 EVMS (after the rounding off) for India as a whole.

It would appear that the ECI has arrived at its sample size of "one EVM per Assembly Constituency" by treating EVMS in India as a whole as 'population'. The ECI-prescribed sample size will work correctly only in this case. But the ECI as well as its statistical advisors seem to have overlooked two crucial aspects: First, since the sample size is for 'India as a whole', in the event of a defective EVM turning up in the chosen sample, the VVPAT slips of all the remaining EVMS in the population (i.e. throughout India) will need to be hand counted, and not just EVMS of the particular Assembly Constituency in which the defective EVM was detected. Can the ECI keep the declaration of results throughout India on hold and order the hand counting of all the remaining 99.96 per cent of EVMS in the country? Surely not. When EVMS used in the country as a whole are treated as the 'population', W_n becomes very small but this small sample size comes at a big 'price', viz. $W_{(N-n)}$ is too large and just not administratively viable in the event of a defective EVM turning up in a sample anywhere in the country.

Second, EVMS employed in 'India as a whole' can be treated as the 'population' only for an all-India Parliamentary Election; not for individual State Assembly Elections. When we have an Assembly Election for Mizoram or Telangana or Madhya Pradesh, the ECI should treat only the EVMS used in the 'State as a whole' as the 'population'. In that case, the sample size should be 370 for Mizoram; 455 for Telangana; and 457 for Madhya Pradesh which works out to an average of 10 EVMS per Assembly Constituency for Mizoram; four for Telangana; and two for Madhya Pradesh. So, the ECI-prescribed sample size of "one EVM per Assembly Constituency" which may be appropriate for 'India as a whole' is illogical and inappropriate if used for Assembly Elections. So EVMS used in the country as a whole are also not an appropriate choice for 'population'.

What should the ECI do?

As already stated, the ECI is at liberty to define the ‘population’ suitably as long as it is logical, statistically sound, administratively viable, and represents a proper trade-off between W_n and $W_{(N-n)}$. It is evident from the foregoing discussion that EVMs used in ‘Assembly Constituency’, ‘Parliamentary Constituency’ or ‘the country as a whole’ are NOT suitable choices for ‘population’. *The only suitable choice, both for Assembly and Parliamentary Elections, are EVMs used in ‘a State as a whole’.*

Is the ECI worried that the administrative workload $W_{(N-n)}$ involved in the hand counting of VVPAT slips *all over a State* on discovery of a stray defective EVM anywhere in the State is too much? It shouldn’t be worried for 2 reasons:

- (i) The ECI’s present sample size holds good only when EVMs used in ‘India as a whole’ are treated as the ‘population’. In the event of a defective EVM turning up anywhere in India, the hand counting of VVPAT slips must be done for VVPATs of all EVMs *in all constituencies throughout India*. In other words, the *status quo* is much worse.
- (ii) The ECI has claimed ‘perfect tallying’ between EVM electronic counts and VVPAT hand counts in 843 constituencies in the past Assembly elections where VVPAT-units were deployed and its sample size of “one EVM per Assembly Constituency” was adopted. If this was indeed the case, the ECI has nothing to worry about as the biggest sample size for a State is only 458. But the correctness of the ECI’s claim is open to question. *First*, there is a *bias in sample selection* when the defective VVPAT units that couldn’t be replaced are left out from the population from which the sample of one EVM per Assembly Constituency is chosen. Since the percentage of defective VVPAT units on polling day was reportedly as large as 20 per cent, and *the polling went ahead in many of these polling stations without the VVPAT units*, the legitimacy of the population is open to question. *Second*, the ECI’s minuscule sample size of “one EVM per Assembly Constituency” had very high margins of error and would have missed out on many defective EVMs which a larger, statistically sound sample may have detected.

If the ECI wants greater accuracy, it should go in for a sample size that will have **99.9 per cent** probability of detecting at least one defective EVM. The sample sizes for the five States are indicated in **Table 8**.

Table 8

Sample Sizes using A STATE AS A WHOLE as the Population

Percentage of defective EVMs (P) is assumed as 1%.

*Probability of detecting at least one defective EVM is chosen as **99.9%**.*

State	Number of Assembly Constituencies	Population Size (N) [Total number of polling stations in the State]	Sample Size (n) required for the State as a whole	% of n to N	Average Number of EVMs per Assembly Constituency whose VVPAT slips should be hand counted
Mizoram	40	1164	508	43.64	13
Chattisgarh	90	23672	677	2.86	8
Telangana	119	32574	680	2.09	6
Rajasthan	200	51796	683	1.32	4
Madhya Pradesh	230	65341	685	1.05	3

The sample sizes and the *average* number of EVMs per Assembly Constituency whose VVPAT slips are to be hand counted are relatively greater in this case but are still reasonable and administratively viable.

Sample size determination is not a purely statistical exercise. Since elections are the bedrock of democracy and the *perceptions* of political parties and voters are important, the ECI would do well to opt for 99.9 per cent probability that the sample will detect at least one defective EVM.

The average number of EVMs to be hand counted per Assembly Constituency have been indicated in **Table 6** and **Table 8** so as to give an ‘order-of-magnitude’ figure vis-a-vis the present figure of one EVM per constituency. Since the sample is for a State as a whole and since the number of polling stations per Assembly Constituency may vary widely even within a State, the ECI may apportion the total sample among the various Assembly Constituencies *in proportion to the number of polling stations in each constituency and round off fractions to the next higher integer*. The rounding-off is likely to increase the sample size for each constituency slightly which is a good thing.

The State-wise sample sizes required have been worked out and are shown in **Annexure I** (for 99% probability of detecting at least one defective EVM) and **Annexure II** (for 99.9% probability).

It is best that the ECI do the necessary calculations and communicate to the Chief Electoral Officer (CEO) of each State the sample size for hand counting of EVMs' VVPAT slips (1) for the State as a whole, and (2) for each Assembly Constituency. Unless there is a significant change in the number of polling stations, the ECI should permanently 'fix' the sample size for the State as a whole and for each Assembly Constituency for all future elections.

There may be a problem for by-elections where an Assembly Constituency or a Parliamentary Constituency will have to be taken as the population and the sampling fraction for VVPAT-based audit will be very large as seen in **Table 4** and **Table 5**. But the ECI usually groups together several Assembly Constituencies and Parliamentary Constituencies for which by-elections have to be conducted. *The total EVMs used in all these by-elections put together* may be taken as the population which will yield an administratively viable sample size for VVPAT-based audit.

V. ECI MUST SET THE CONTROVERSY AT REST

*"There are two possible ways to approach phenomena. The first is to rule out the extraordinary and focus on the "normal." The examiner leaves aside "outliers" and studies ordinary cases. The second approach is to consider that in order to understand a phenomenon, one needs to first consider the extremes - particularly if, like the Black Swan, they carry an extraordinary cumulative effect."*¹⁹

- Nassim Nicholas Taleb

[Distinguished Professor of Risk Engineering, NYU Tandon School of Engineering]

Most people expect all swans to be white because that's what their experience tells them; a black swan is by definition a surprise. According to Nassim Nicholas Taleb, a "Black Swan Event" is characterized by the following three attributes. *First*, it is an outlier, as it lies outside the realm of regular expectations, because nothing in the past can convincingly point to its possibility. *Second*, it carries an extreme impact. *Third*, it will seem obvious in hindsight with people asking why the warning signs were not noticed sooner. In sum: rarity, extreme impact, and retrospective (though not prospective) predictability.

The Great Depression of 1929, the precipitous demise of the Soviet bloc during 1989-91, the global financial crisis of 2008, and the Punjab National Bank-Nirav Modi scam of 2018 were some typical Black Swan Events. History is replete with them. Our inability to predict the course of history is due to our inability to predict Black Swan Events. According to Taleb, no matter how hard we try, it is very likely that the next Black Swan Event will also take us by surprise. So, while we should prepare for the specific threats that we envision we should not forget to also prepare for the unexpected.

Rigging of an election through EVM fraud fits Taleb's depiction of a Black Swan Event. The "unexpected" that the ECI should prepare for is EVM fraud. It may have a very low (but non-zero) probability and it may be unpredictable in terms of time and place. However, if EVM fraud were to occur, the damage to the sanctity of the electoral process will be immense. There is no point in regretting or rationalising after the event.

What is worse, without a credible VVPAT-based audit of EVMs, the fraud may be undetectable and may be carried on with impunity. The ECI should, therefore, move out from its comfort zone

and focus on “outlier” events like EVM fraud. The risk of EVM fraud, howsoever remote, is something the political parties and voters of India will never accept – not because they overestimate the risk but because the cost of the catastrophe is too dreadful to contemplate.

More than 100 years after H.G. Wells wrote that statistical understanding will one day be as necessary for efficient citizenship as reading and writing, a shocking lack of statistical understanding continues to persist among citizens in India today. The ECI prescribing a patently wrong sample size of “one EVM per Assembly Constituency” for all Assembly Constituencies in all States and managing to get away with such a statistical howler for so long is a case in point.

It is important that the ECI must set the controversy at rest and implement the Supreme Court’s order of 2013 properly both in letter and spirit. It should adopt the statistically correct sample sizes of EVMs for hand counting VVPAT slips, suggested in this paper, starting from the Assembly Elections for Mizoram, Chhattisgarh, Telangana, Rajasthan, and Madhya Pradesh due in November–December 2018. If the ECI persists with its statistically incorrect sample, an adverse inference is liable to be drawn against it and it may lose the perception battle in the minds of the political parties and voters.

Annexure I

State-wise Sample Sizes for 99% probability that the sample will detect at least one defective EVM

EVMS in the State as a whole are assumed as 'population'

Percentage of defective EVMS (P) is assumed as 1%.

Sl.No.	State	Number of Assembly Constituencies in the State	Population Size (N) = Total Number of Polling Stations (EVMS) in the State	Sample Size (n) for the State	Average Number [@] of EVMS whose VVPAT slips are to be hand counted <i>per Assembly Constituency</i>
1	Sikkim	32	589	315	10
2	Mizoram	40	1164	370	10
3	Goa	40	1642	409	11
4	Nagaland	60	2194	413	7
5	Arunachal Pradesh	60	2562	414	7
6	Manipur	60	2794	422	8
7	Meghalaya	60	3082	424	8
8	Tripura	60	3174	424	8
9	Himachal Pradesh	68	7521	446	7
10	Jammu & Kashmir	87	10035	450	6
11	Uttarakhand	70	10854	450	7
12	Haryana	90	16357	451	6
13	Kerala	140	21498	454	4
14	Punjab	117	22615	454	4
15	Chhattisgarh	90	23672	454	6
16	Jharkhand	81	24803	455	6
17	Assam	126	24890	455	4
18	Telangana	119	32574	455	4
19	Odisha	147	35959	455	4
20	Andhra Pradesh	175	39970	456	3
21	Gujarat	182	50128	457	3
22	Rajasthan	200	51796	457	3
23	Karnataka	224	56696	457	3
24	Bihar	243	65337	457	2
25	Madhya Pradesh	230	65341	457	2
26	Tamil Nadu	234	65616	457	2
27	West Bengal	294	77247	458	2
28	Maharashtra	288	91329	458	2
29	Uttar Pradesh	403	150000	458	2
INDIA		4120	About 10,00,000	459	1

[@] - Rounded off to the next higher integer.

Annexure II

State-wise Sample Sizes for 99.9% Probability that the sample will detect at least one defective EVM

EVMs in the State as a whole are assumed as 'population'

Percentage of defective EVMs (P) is assumed as 1%.

Sl. No.	State	Number of Assembly Constituencies in the State	Population Size (N) = Total Number of Polling Stations (EVMs) in the State	Sample Size (n) for the State	Average Number [@] of EVMs whose VVPAT slips are to be hand counted per Assembly Constituency
1	Sikkim	32	589	461	15
2	Mizoram	40	1164	508	13
3	Goa	40	1642	574	15
4	Nagaland	60	2194	589	10
5	Arunachal Pradesh	60	2562	595	10
6	Manipur	60	2794	608	11
7	Meghalaya	60	3082	613	11
8	Tripura	60	3174	614	11
9	Himachal Pradesh	68	7521	659	10
10	Jammu & Kashmir	87	10035	667	8
11	Uttarakhand	70	10854	669	10
12	Haryana	90	16357	672	8
13	Kerala	140	21498	677	5
14	Punjab	117	22615	678	6
15	Chhattisgarh	90	23672	679	8
16	Jharkhand	81	24803	678	9
17	Assam	126	24890	678	6
18	Telangana	119	32574	680	6
19	Odisha	147	35959	680	5
20	Andhra Pradesh	175	39970	681	4
21	Gujarat	182	50128	683	4
22	Rajasthan	200	51796	683	4
23	Karnataka	224	56696	684	4
24	Bihar	243	65337	685	3
25	Madhya Pradesh	230	65341	685	3
26	Tamil Nadu	234	65616	684	3
27	West Bengal	294	77247	685	3
28	Maharashtra	288	91329	685	3
29	Uttar Pradesh	403	150000	686	2
INDIA		4120	About 10,00,000	688	1

[@] - Rounded off to the next higher integer.

Endnotes

-
- ¹ In his presidential address to the American Statistical Association in 1950, Samuel S. Wilks said, "Perhaps H.G. Wells was right when he said 'Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.'" The quote was then published in the Association's journal in 1951. This is the form in which it is popularly quoted. But H.G.Wells' original quote which appeared in his book *"Mankind in the Making"* (1903) was as follows: "The great body of physical science, a great deal of the essential fact of financial science, and endless social and political problems are only accessible and only thinkable to those who have had a sound training in mathematical analysis, and the time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex world-wide States that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write."
- ² Shetty, K.A.V. 2018. *"Making Electronic Voting Machines Tamper-proof: Some Administrative and Technical Suggestions"*, The Hindu Centre for Politics and Public Policy, Policy Watch No. 6, published on August 30, 2018 and updated on October 3, 2018. Please see Chapter VI *"The Vulnerability of Indian EVMs"*, Chapter VII *"Three Security Loopholes"* and Chapter VIII *"ECI's Administrative Safeguards are not Foolproof"*.
- ³ In Statistics, the *population*, or *universe*, refers to the complete set of elements (persons or objects) that possess some common characteristic which is of interest to the researcher. e.g. all persons with HIV-AIDS in a city; all EVMs used in an election, etc. A *sample* is a subset of the population consisting of one or more elements drawn from the population. Based on the sample results, the researcher can make inferences or extrapolations from the sample to the population.
- ⁴ Let us assume that 300 EVMs were used in an election. A sample of three EVMs is drawn randomly. As per the EVM electronic count, let the total votes polled in these three EVMs put together be 1,800 and the votes secured by the leading candidate be 600. If the hand count of VVPAT slips for these three EVMs also yields the same total of 1,800 votes and the same number of 600 votes for the leading candidate, then there is no possibility of any EVM malfunction or fraud. The results of the election (for 300 EVMs put together) can be declared based on their EVM electronic count.
- ⁵ A 'defective EVM' is defined as one which has a mismatch between the 'EVM count' and the 'VVPAT count'. The mismatch may be due to EVM malfunction or EVM tampering or VVPAT-unit malfunction or mistakes in the hand counting of VVPAT slips. In the event of a mismatch, *at least one recounting* of the VVPAT slips of the particular EVM may have to be done to rule out mistakes in hand counting. The VVPAT total as per the recount should tally either with the EVM count or the previous VVPAT count. If it doesn't tally with either, further recounts should be done until the last VVPAT count matches either with the EVM count or one of the previous VVPAT counts.

⁶ Should the discrepancy of even a single vote or single digit votes between the EVM count and VVPAT count (even after following the recount procedure stated in Endnote 5 above) lead to the designation of the EVM as ‘defective’? Ideally, *yes*. Or, should the ECI ignore minor discrepancies of *not more than, say, five votes* in order to avoid the huge administrative workload of hand counting VVPAT slips of all the remaining EVMs of the population? Whether to ignore such minor discrepancies or not in cases where there will be no change in election outcomes is a *policy decision* to be made by the ECI in consultation with various political parties and other stakeholders.

⁷ Chapter 5 titled “*Perfunctory Implementation of VVPAT*” of Policy Watch no. 6 “*Making Electronic Voting Machines Tamper-proof: Some Administrative and Technical Suggestions*” written by the author was one of the first papers in India to deal with the issue of sampling plan of EVMs for VVPAT-based audit. In that paper, sample sizes were calculated using ready reckoners based on the Normal Distribution model. The Normal Distribution model is a reasonably ‘good fit’ to the EVM problem but the Hypergeometric Distribution model (which is used in the present paper) is even better for the following three reasons:

- (i) It is an ‘exact fit’ to the EVM problem;
- (ii) It yields a more economic (i.e. smaller) sample size; and
- (iii) In the Normal Distribution model – for a given confidence level and a given margin of error – the sample size is maximum when the ‘Proportion of defectives’ (P) in the population is assumed to be 0.5 and decreases significantly as the value of P decreases and approaches zero. But in the Hypergeometric Distribution, the exact reverse is the case i.e., the sample size is maximum when P is close to zero and decreases significantly as P increases. So, irrespective of what the true value of P is, if we calculate the sample size for P very close to zero such as $P = 0.01$ (which is what the ECI thinks it is), then this holds good for all the other scenarios where P is higher. We do not need to make any questionable assumptions about the value of P as in the Normal Distribution model nor do we need to extrapolate trends based on questionable past empirical data.

⁸ Pinker, S. 1997. “*How the Mind Works*”, W.W.Norton & Co.

⁹ When a sample is drawn *without replacement from a finite population*, the probability of occurrence of the various outcomes is given by the *Hypergeometric Probability Distribution* model.

Note: A ‘probability distribution’ is a mathematical function that gives the probability of occurrence of different possible outcomes in an experiment. The simplest case is the ‘uniform distribution’ in which all outcomes have an equal probability of occurrence. Apart from Hypergeometric Distribution, Binomial Distribution, Poisson Distribution, and Normal Distribution are some of the most commonly used probability distribution models.

¹⁰ Schlaifer, R. (1959) “*Probability and Statistics for Business Decisions – An Introduction to Managerial Economics under Uncertainty*”, McGraw-Hill Book Company, Inc.

¹¹ Supreme Court of India, 2018. Writ Petition (civil) no. 935 of 2018 in *Kamal Nath vs Election Commission of India*. Oct. 12.

¹² In Hypergeometric Distribution, the probability of finding 'x' successes in a sample of size 'n' drawn from a population of size 'N' with 'M' successes is given by the formula:

$$\text{Prob (x, n, M, N)} = \frac{{}^M C_x \cdot {}^{(N-M)} C_{(n-x)}}{{}^N C_n}$$

¹³ The online Casio calculator available at <https://keisan.casio.com/exec/system/1180573201> is very useful for calculating probabilities under Hypergeometric Distribution. Enter the known values of population size (N) and 'successes' in the population (M), where $M = N \cdot P$ where P is the 'proportion of the characteristic of interest'. Try out different values of sample size (n) in the calculator such that the probability that $x = 0$ (of not finding any 'success' in the sample) is *less than the specified level*, say, less than 0.01 or 0.001; or, which is the same thing, the probability of finding at least one 'success' in the sample is greater than 0.99 or 0.999.

¹⁴ In the online Casio calculator referred to above, enter $N = 100$, $M = 5$, $n = 3$, $x = 0$ (not finding even a single 'success'). The probability of ' $x = 0$ ' is 0.856. Or, the probability of getting at least one 'success' is $[1 - 0.856] = 0.144$ i.e. 14.4%.

¹⁵ In the same calculator, enter $N = 100$, $M = 5$, $x = 0$ (not finding even a single 'success'). Enter increasing values of 'n' till the probability of ' $x = 0$ ' becomes less than 0.01. It is seen that the probability of ' $x = 0$ ' is 0.011 for $n = 58$, and is 0.0099 for $n = 59$. So, with a sample size of 59, the probability of not getting a single 'success' is less than 1%. Or, the probability of getting at least one 'success' is 99%.

¹⁶ The superiority of the Hypergeometric Distribution model to the Normal Distribution model has already been discussed in Endnote 7. The Binomial Distribution is applicable to *infinite populations* or where the samples are taken *with replacement*. In Binomial Distribution, the sample size (n) is *independent* of the population size (N) and depends on the proportion of the characteristic of interest (P) and the confidence level (C). The formula for sample size is:

$$n = \ln(1 - C) / \ln(1 - P) \text{ where 'ln' stands for natural logarithm.}$$

For $C = 0.99$ and $P = 0.01$, $n = \ln(1-0.99) / \ln(1-0.01) = \ln(0.01) / \ln(0.99) = 458.21$, rounded off to 459 (the next highest integer).

Only the Hypergeometric Distribution gives the correct, economic sample sizes *for finite populations*. In the example discussed in pages 2-4 (please see Table 1), with Hypergeometric Distribution, $n = 448$ when $N = 10,000$; $n = 457$ when $N = 50,000$; $n = 458$ when $N = 1,00,000$ and $n = 459$ when $N = 5,00,000$. So, as the population size (N) increases, the sample size (n) as per the Hypergeometric Distribution model approaches the value given by the Binomial Distribution model (459). The Binomial Distribution model is a reasonably 'good fit' when the population size is very large but is not suitable for smaller, finite populations.

¹⁷ **Moroney, M.J. 1951.** *"Facts from Figures"*, Penguin, London.

¹⁸ In the online Casio calculator in end note 11, enter $N = 300$, $M = 3$, $n = 1$ and $x = 0$. The probability of $x = 0$ (i.e. of not finding a single 'success') is 0.99. That is, the ECI-prescribed sample size will miss a defective EVM 99% of the time. Repeat the calculations for $N = 200$, $N = 100$ and $N = 50$ to get the figures for the last column of Table 4.

¹⁹ **Taleb, N, N. 2007.** "The Black Swan: The Impact of the Highly Improbable", *Random House*.



About the Author

K. Ashok Vardhan Shetty is a former Vice-Chancellor of the Indian Maritime University, Chennai, a Central University under the Ministry of Shipping. Before assuming charge as the Vice-Chancellor, Shetty was a member of the Indian Administrative Service (IAS), Tamil Nadu Cadre, of the 1983 batch. He held a number of key assignments including Registrar, University of Madras, Director of Collegiate Education; District Collector, Viluppuram; Director of Rural Development; Managing Director, Tamil Nadu State Marketing Corporation, (TASMAC); Secretary, Chief Minister's Secretariat; Principal Secretary, Rural Development and Panchayat Raj Department; Principal Secretary, Municipal Administration and Water Supply, among others. Successful project implementation was his forte. He was commended by the Government of Tamil Nadu several times.

Shetty has published several articles on public administration, management, E-Government, popular science, and popular mathematics in leading English and Tamil newspapers such as The Hindu, The Hindu - Tamil, The Hindustan Times, Indian Express, The Hindu BusinessLine, and (the now defunct magazine) Science Today. His earlier contribution to The Hindu Centre for Politics and Public Policy can be accessed at "Making Electronic Voting Machines Tamper-proof: Some Administrative and Technical Suggestions".

He can be contacted at shetty25@hotmail.com

