International Journal of Scientific and Innovative Mathematical Research(IJSIMR)
Volume 11, Issue 1, 2023, PP 13-29
ISSN No. (Online) 2347-3142
DOI: https://doi.org/10.20431/2347-3142.1101002
www.arcjournals.org

Robust & Reliable Statistical Projection During Opinion & Exit Polls in India Based on Sampling Forecast Models

DR. AMBIKA BHAMBANI*

M.A , M.Phil., Ph.D. Associate Professor ,Department of Mathematics , Jesus & Mary College [University of Delhi],Chanakyapuri , New Delhi .

Corresponding Author:** ***DR. AMBIKA BHAMBANI, *M.A , M.Phil., Ph.D. Associate Professor ,Department of Mathematics , Jesus & Mary College [University of Delhi],Chanakyapuri , New Delhi .*

Abstract: *Our experiences, observations and analysis of a situation make it possible to derive inferences from the activities around us. Knowingly or unknowingly we all tend to draw some conclusions by either observing what is happening around us or by the information that we receive from different sources.*

This paper contributes to the literature on each of the key elements for predicting election results more accurately , in a scientific manner , by innovating the frames, training-data and learners needed to generate accurate readings of public opinion.

Keywords: *Constituency , Data , Sampling , t – test , Chi – square test .*

1. INTRODUCTION

Diversity of India : The word “diversity” places more emphasis on differences than on unfairness. It refers to group disparities, or distinctions separating one group of individuals from another. These differences could be biological, religious, linguistic, or anything else. Diversity refers to the variety of races, religions, languages, castes, and cultures.

India as a country with rich cultural diversity, we are referring to the wide variety of social groupings and cultures that call India home. These groups identify primarily through cultural traits like language, religion, sect, race, or caste.

Constituency, basic electoral unit into which eligible electors are organized to elect representatives to a legislative or other public body. The registration of electors is also usually undertaken within the bounds of the constituency.

The present delimitation of constituencies has been done on the basis of 1971 census under the provisions of Delimitation Act, 2001. The Commission is a powerful and independent body whose orders cannot be challenged in any court of law.

For the purpose of constituting the Lok Sabha , the whole country has been divided into 543 Parliamentary Constituencies , each one of which elects one member.

As of 2019, Malkajgiri is the largest Lok Sabha constituency by number of electors with 3,150,303.

Lakshadweep Lok Sabha constituency is a Lok Sabha (lower house of the Indian parliament) constituency, which covers the entire area of the Union Territory of Lakshadweep in India. This seat is reserved for Scheduled Tribes. As of 2014, it is the smallest Lok Sabha constituency by number of voters.

Total Electors 49,922

All constituencies within a state should, ideally, be equal in *population*. To achieve this as nearly as possible, periodic alterations of boundaries are made. Constituencies are most often formed on a geographical basis, but the basis could also be occupational (for example, the university constituencies that once existed in the United Kingdom).

INDIAN POLITY - ELECTIONS SYSTEM :

The **Election Commission of India (ECI)** is in charge of the **election process in India** which includes elections to parliament, state legislatures, and the offices of President and Vice-President. ECI has been an **independent constitutional authority** since **January 25, 1950 .**

- In India, there are three levels of government, i.e.,
  - Centre level,
  - State level, and
  - Local level.
- At centre level, elections are conducted to elect Member of Parliament, which is known as Lok Sabha elections.
- For Lok Sabha election, the whole country is divided into **543 constituencies** and each constituency elects one representative as a Member of Parliament (MP).

Election Campaign

- The main purpose of an election is to give the people a chance to choose their representatives and make a government of their choice who frames policies to address their concerns.
- During election campaigns, voters get the opportunity to have a free and open discussion about who is a better candidate, which party can give a better government, or what are their policies.

- In India, election campaigns take place for two weeks period between the announcement of the final list of candidates and the date of polling.
- During campaigns, the political leaders address election rallies and political parties mobilize their supporters.

- The contesting candidates contact their voters through various methods such as −
  - They advertise in newspapers, radio, television, etc.;
  - They publish pamphlets and distribute them in their respective constituencies;
  - They arrange rallies and give speeches at every public place of their constituencies;
  - They tell their voters about their plan and policies and also ask about their (voters') problems.
  - They try to convince their voters in their favor and appeal them to vote and elect the right candidate.

What are Exit & Opinion polls ?

Opinion poll :

- An **opinion poll is a pre-election survey** to gather voters' views on a range of election-related issues.

An **opinion poll**, often simply referred to as a **survey** or a **poll** (although strictly a poll is an actual election) is a human research survey of public opinion from a particular *sample*. Opinion polls are usually designed to represent the opinions of a population by conducting a series of questions and then extrapolating generalities in ratio or within *confidence intervals*.

Opinion polls for many years were maintained through telecommunications or in person-to-person contact. Methods and techniques vary, though they are widely accepted in most areas. Over the years, technological innovations have also influenced survey methods such as the availability of *electronic clipboards* and Internet based polling. Verbal, ballot, and processed types can be conducted efficiently, contrasted with other types of surveys, systematics, and complicated matrices beyond previous orthodox procedures. Some polling organizations, such as , use *Internet* surveys, where a sample is drawn from a large panel of volunteers, and the results are weighted to reflect the demographics of the population of interest. In contrast, popular web polls draw on whoever wishes to participate, rather than a scientific sample of the population, and are therefore not generally considered professional.

Recently, statistical learning methods have been proposed in order to exploit *social media* content (such as posts on the micro-blogging platform *Twitter*) for modelling and predicting voting intention polls.

Exit Poll :

- An **exit poll, on the other hand, is conducted immediately after people have voted**, and assesses the support for political parties and their candidates.

An **election exit poll** is a *poll* of voters taken immediately after they have exited the *polling stations*. Pollsters – usually private companies working for *newspapers* or *broadcasters* – conduct exit polls to gain an early indication as to how an election has turned out, as in many elections the actual result may take hours to count.

Looking At a Glance :

Table1. *General Elections 2014 – Failed in Predicting a Majority for BJP.*

Seat Forecast	BJP allies	Congress allies	Others
ABP - AC Nielson.	**281**	**97**	**165**
CNN-IBN-CSDS	**276 - 282**	**92 - 102**	**150 - 159**
Headlines Today-CICERO	**261 - 283**	**101 - 120**	**152 - 162**
India TV- C-Voter	**289**	**101**	**153**
News 24 -	**340**	**70**	**133**

Today's Chanakya			
Times Now-ORG.	**249**	**148**	**146**
Actual result	**326**	**60**	**157**

Note: Seat predictions based on exit polls conducted during the elections .

Table2. *General Elections 2019 – Correct Estimation of NDA 3.0 Seats.*

Seat Forecast	**BJP allies**	**Congress allies**	**Others**
India Today-AXIS My India	**339 - 365**	**77 - 108**	**69 - 95**
Today's Chanakya	**350**	**95**	**97**
News18-Ipsos	**336**	**82**	**124**
Times Now-VMR	**306**	**132**	**104**
India-News	**298**	**118**	**127**
Republic-C Voter	**287**	**128**	**127**
ABP Nielsen.	**277**	**130**	**135**
Actual result	**353**	**91**	**98**

Potential for Inaccuracy :

In India, it is logistically and statistically difficult to obtain representative samples of voters: surveys have to be designed in multiple languages; rural voters may only be reached by face-to-face interviews due to lack of communications technology; alliances involve local parties and hence in-depth knowledge of local politics is necessary. There are daunting challenges that need to be addressed: within each state, voters speak some subset of the 23 officially recognised languages, and uncountable number of unofficial dialects. According to the World Bank, around 65% of the population lives in rural areas in India.

CSDS election studies continue to be the only reliable resource material and reference that can aid in unravelling the flaws and limitations of survey research and its elements. The fallibility of opinion polls in estimating the correct vote share exposed its fault lines, as survey errors led to inaccurate election forecasts in national and state elections with impunity. The accuracy levels of elections surveys, unlike other opinion polls is comprehendible ? and easy to test on the yardsticks of the closeness of vote estimates with the actual vote share and the representativeness of sample with the demographic profile of the electorate. The review of media opinion polls based on election forecasting in earlier section show statistical miscalculation on both these comparative parameters.

The error in vote share and imbalanced sample profile coupled with erroneous seat predictions point towards flaws in opinion poll design, fieldwork implementation at ground zero and statistical forecasting models (Rai, 2021) . Thus, it becomes pertinent to dig deeper into election survey discourse to find the inherent fallacies and statistical limitations of polls and its adverse impact on elections forecasting.

A sample of 4,000 voters in 543 parliamentary constituencies can predict the seats accurately, but a sample size of over 21 lakhs would be impractical as it will entail a huge cost and require an army of trained and reliable field enumerators. Thus, predicting seats on a mathematical model of vote shares ascertained at state level from a cluster of 10-12 assembly constituencies increases the possibility of modelling error. The other limitation of a survey done well ahead of actual polling day is that though it measures the opinion of the whole population, what really counts is the group that actually goes out and votes. The CSDS election data reveals that the propensity to vote is much lower among the urban, upper middle class and upper class, college educated and high-income groups. The electorate is quite volatile and voting intentions undergo massive swings as voting day approaches in India. These two factors mean that the predictive power of any election opinion poll done weeks ahead of the poll is limited and fallible, as all it can measure is the mood of the nation at the time of the poll (Karandikar, 2014). Voters may change their minds between a poll and the Election Day, and this is the main reason why polls taken 6 months before an election have a much poorer predictive record than those taken close to election date (Northcott, 2015). The traditional polls are snapshots of public opinion at a certain point in time and do not provide predictions. The routine interpretation of polling results as election day forecasts can result in poor predictions particularly if the election is still far way, because public opinion can be difficult to measure and remains fragile over the course of an election campaign. (Campbell, 1996).

The voting intentions of a sample serve as a proxy for those of a population and the main reason for an unrepresentative sample is the sampling error, as small samples can lead to misleading flukes. A major issue for pollsters is to ensure that the samples are in appropriate balance with respect to various demographic variables, and if required use balancing procedures to put relevant weights. In addition to sampling errors and systematic bias, the phenomenon of herding can lead to forecasting error. Most polling agencies at the end of a campaign, it is widely suspected 'herd' and report headline figures closer to the industry mean, presumably to avoid the risk of standing out as having missed the final result by an unusually large margin. Some sensitivity to this turns out to be optimal for accurate election prediction (Northcott, 2015). The vote share estimates of an election poll and seat predictions can be fully correct, but as part of media manipulation different figures could be publicly released and the subsequent error be blamed on a faulty projection model. The main problems concern the unwarranted and misleading inferences drawn from polls by their readers and users, often an audience that may not be well aware of the limitations of statistical methodology. There are several statistical polls run by the media under the pressure of deadlines and to puff up a poll by published findings that may excite readers. There is no surety that opinions polls is fallacy-free and it is on critical thinking public to become aware of the biases and fallacies and to take a 'Buyer Beware' attitude. Social statistics are needed to conduct intelligent public deliberations and set social policies in a democracy, but activists, the media and private agencies can and often do use 'Mutant Statistics' as tactics to manipulate public opinion (Douglas, 2007) .

In the absence of forecasting models of India polling firms in the public domain, it is quite difficult to assess the status of research and development in the field of accuracy and advancement. In contrast, the USA from the 1970s onwards witnessed an addition of a wide range of successful election forecasting techniques in the literature on electoral forecasting. The structural models were primarily applied in two-party and presidential democracies, but it runs into difficulty in more complex and multiparty democracies. In contract, the popular dynamic linear model (Jackman, 2005) is tried and tested and has shown that reasonable forecasts can be made despite the complexity of the party systems and the emergence of new and smaller parties. A novelty is set in motion, with the introduction of cyclical changes in party support in the model through a seasonal component before they are seen in the polls.

Opinion formed against Opinion polls & Exit polls :

There is a wide opinion formed against Opinion polls & Exit polls.

Some People feel that , It is high time that such a charade is put to an end forever. Election after election, the absurd pretence has been thrust on the people in the form of the so-called exit polls and countless hours of mundane analysis on something which is a far cry from reality. Soon after the polling is over, broadcasting channels go gung ho-over their predictable election outcome without any iota of shame when they are so off the mark when the final results come in.

What has often been seen that a media channel propounding the cause of a particular political party tends to defend as well as predicting its victory? And when the shocking election results are announced, the channel would not eat the crow. It is an open secret that most of the Indian media channels, barring a very few, are biased and they tend to overzealously favour a particular political party (no prizes for guessing), often jeopardising their authenticity.

When the ludicrous exit polls are beamed, it has a cascading effect. Besides the usual spectacle in the news channels, there is eruption on social media. They too join the party, take over from the channels and start expressing their political views. Amidst all this frenzy, who gets martyred? *Investors!* The *stock exchanges* too react to the exit polls and sometimes a fortune is wiped out. The markets open disastrously if a defeat is predicted for the incumbent ruling party or the possibility of a rag-tag coalition government.

Why is the Election Commission (EC) against these polls?

- Both kinds of polls **can be controversial if the agency conducting them is perceived to be biased.**
- The projections of **these surveys can be influenced by the choice, wording and timing of the questions,** and by the nature of the sample drawn.
- Political parties **often allege that many opinion and exit polls are motivated and sponsored by their rivals**, and could have a distorting effect on the choices voters make in a protracted election, rather than simply reflecting public sentiment or views.

Need for Scientific Approach i.e., Statistical Methods for Predictions :

Contrary to their foreign counterparts, media opinion polls on elections in India have focused more on predicting the number of seats that major political parties are going to win or lose in the elections rather than on understanding the key issues facing the electorate.

A recent sting operation on polling agencies also revealed that seat prediction figures are on occasion manipulated in favour of their clients.

Thus election surveys have been reduced to a media gimmick used only to predict the outcome of election results that quite often end up wrong or off the mark.

Similarly, exit polls, in which the voters are interviewed outside the polling booth after they have voted, do not always reveal voters' actual decisions.

This happens because there can be a fear that a correct revelation of the party they voted for could be used by other political parties in subsequently identifying and targeting them.

Most Indian polls go wrong because their sampling methodology is poor which makes the sample profile unrepresentative. Though a scientific and representative sample determines the accuracy of the survey, there is no guarantee that a forecast based on the survey will be right. A survey has its limitations as it cannot capture the diverse and nuanced complexities and undercurrents of electoral behaviour and choices in India.

Over time, a number of theories and mechanisms have been offered to explain erroneous polling results. Some of these reflect errors on the part of the pollsters; many of them are statistical in nature. Others blame the respondents for not giving candid answers .

Inferential Statistics is built on the foundation of Probability theory , and has been remarkably successful in guiding opinion about the conclusions to be drawn from data.

The case study why predictions are not always very accurate , could be well understood with Mathematical lens , based on Sampling.

Population or Universe : It is defined as the collection of individuals or the totality of observations

Sample : It is the part of the population selected for drawing conclusions regarding population.

Sampling : Sampling is the process of selecting the sample from the population .

Advantages of Sampling : To increase speed and to reduce cost and labour

Types of Sampling :

1.Unbiased Sampling or representative Sampling or Random Sampling :

ln which the selection of individuals from the population is random

2 . Purposive Sampling : ln which the selection is according to some purpose

3. Mixed Sampling : Above 1.& 2.combined.

Simple Random Sampling :

lt is defined as the process which consist in selecting the

individuals from the population in such a way that each individual in the population has the same chance of selection . The most important method of getting a random sample is with the help of Trippet's nos.

Sample Size : The number of members in the sample is called the sample size . 1f the number is less than or equal to 30 (i.e., 5 30), the sample is regarded as small sample , otherwise the large sample.

Parameter : The given population has some characteristics' such as mean, variance etc. , which are called the parameters of the population .

Statistic : Statistical measures obtained from the sample observations alone is called a statistic Statistic may thus be regarded as an estimate of parameter , obtained from the

sample and is a function of the sample variable alone . It varies from sample to sample

Sampling Distribution :

For each of the samples , a statistic , say t eg. . x , s^2 etc. . can be

computed which obviously varies from sample to sample . The aggregate of the various values of the statistic under consideration so obtained (one from each sample) may be grouped into a frequency distribution , which is called the sampling distribution of the statistic .

Note : The sampling distribution tends to the normal distribution , if the number of the samples is large.

Standard Error : The standard deviation of the sampling distribution of a statistic is called the standard error & is written as S.E

Standard error of a statistic may be reduced by increasing the sample size but this results in corresponding increase in cost ,labour& time etc.

Probable Error : *is defined as (0.67449) S. E*

Tests of Significance :

Tests of Significance are the tests of significance of statistical hypothesis .We test whether the difference between the sample proportion (or values) and the population proportion (or values) is so small as could be due to fluctuations of sampling or large enough as to signify evidence against the hypothesis .

For this . we follow the ***six step procedure*** , as given below :

Step 1 : Null hypothesis : We set up an hypothesis about the population . This hypothesis is called the null hypothesis . since it asserts that there is no difference

between the sample and the population regarding the particular matter under consideration

Step 2 : Alternative hypothesis : Contrary to null Hypothesis is Alternative Hypothesis

Step 3 : α : Level of significance or amount of risk :

The probability level below which we reject the hypothesis . is called the level of significance . Usually we take 5% level of

significance .

i.e.. P { | z | > 1.96 } = 0.05 and sometimes P { | z | > 3 = 0.0027

Step 4: W: Critical Region or Region of Rejection:

$W = \{ z : | z | \geq 1.96 \}$

Note if | z | > 1.96 , then W is significant

If | z | > 3 , then W is highly significant .

where 1.96 is the value of z , corresponding to 5 % level of significance .

& 3 is the value of z corresponding to 0.27 % level of significance

Step 5: Test Statistic or Critical Ratio or Test Ratio :

z_0 (say): value of standardized variate z under H_0

Step 6: Inference or Judgement:

If $z_0 \notin W$, Accept H_0 at α .

If $z_0 \in W$, Reject H_0 at α .

1. If $z_0 > 1.96$ (W is significant) .

2. If $z_0 > 3$. (W is highly significant) .

Confidence Limits or Fiducial Limits :

The values $\bar{x} - 1.96\, \sigma / \sqrt{n}$ and $\bar{x} + 1.96\, \sigma / \sqrt{n}$

are called fiducial limits Or confidence limits for the mean

of the population corresponding the given sample .

Confidence Interval :

The interval

$\bar{x} - 1.96\, \sigma / \sqrt{n}$ to $\bar{x} + 1.96\, \sigma / \sqrt{n}$

is called confidence interval .

Probable Limits : $E (X) \pm 3\, S.E (X)$

Or

$E (X) \pm 3 \sqrt{Var (X)}$

Where Var (X) stands for Variance of X.

Note : To compute these probable limits . we use the sample estimate p for P which generally proves to be satisfactory for

$N \geq 30$.

Probable limits are always found when Null hypothesis is rejected.

Now we can define sampling as a process through which we choose a smaller group to collect data that can be the best representative of the population. We can extend our results obtained from the sample group to the entire population. There are a number of ways in which the sampling process can be carried out.

Random Sampling

Random sampling as the name suggests is a process to arbitrarily choose our sample from population group

Stratified Random Sampling :

Stratified random sampling is a process to gather data by separating the actual population into the distinct subset or strata, and then choosing simple random samples from each stratum.

After the population has been stratified or grouped, you may use simple random sampling to gather the complete sample. The next step after you have identified the sample is to collect the relevant data.

The whole process of data collection can be summarized in the following flow chart:

*Sample : Choose your sample across age, profession and family background.
* Data collection

*Data analysis

* Drawing inferences.

From Where to Get Data ?

Data is found around us in many forms. You can get data from direct observations, interview surveys experiments and festing , and so on. There are data already collected by someone which can give us very useful insight or we can collect our own data specifically based on our needs. Data can be collected through direct sources or through indirect sources.

Based on the sources of collection, data can be broadly classified in two categories

1.PrimaryData

2.SecondaryData

When we collect data directly from the field or from the individuals, the data is called Primary Data .

Primary data is also called first hand data. It is the original data and collected with a specific purpose in mind. Primary data can be collected through Survey, observations case studies, etc.(source) Sometimes if it is not possible to collect the original data, then we can look for already collected data. Such data is called Secondary Data. Secondary data is the data that has already been collected and readily available for other uses. It can be obtained through available records such as official reports newspapers. research articles, etc.

There may be situations where we may use either primary data or secondary data or both.

We can now summarize the differences between the primary & Secondary data :

Primary Data	**Secondary Data**
Original and New	Reused & old
Primary sources	Secondary sources
Customized as per the need of the research/project Questions	May not be directly linked sith targeted research/project Questions
High on Reliability	Low on Reliability
Less economical in terms of time , manpower and money.	Highly economical in terms of time , manpower & money

Looking above in the table & doing the comparisons , for the elections predictions purpose , the primary data is highly recommendable to get results near to accuracy , as being more reliable , though collecting primary data is less economical , and time consuming.

How to Collect Data?

Data collection is a process of planning for and obtaining useful information about the project question you would like to explore. Data collection will allow you to develop factual basis making decisions. It can involve multiple choices by the data collectors. When you plan your data collection scheme, you must focus on the relevant parameters to reduce ambiguity.

The following questions can be asked while developing the Data Collection Scheme: Why do I want to collect data?

*What is the purpose of collecting data?

*What will be the source of data collection?

*What type of data should I collect?

*Who will collect the data?

• What tools should I use to collect data?

There are many methods used to collect data.

The following are two most popular methods:

1. Surveys

2. Direct Observation .

Survey is the most commonly used method in social science projects/research. You can collect primary data through surveys. Survey can be done in face-to-face mode and indirect mode.

Some of the common mediums to do survey are:

• Face-to-face/door-to-door visits

•Telephone
•Internet
• Interview

Focus group (group of 6-7 people participating through discussion, debate, suggestion to the questions asked in the survey)

For elections purpose , the primary data , collected through Face-to-face/door-to-door visits , is the most reliable data collection.

How to Develop a Good Questionnaire?

The data you collect can provide meaningful information only when you prepare the right kind questions. A wrongly worded question can mislead the entire data. A good survey requires meticulous planning about the sample, time-line and nature of questions to be asked in the survey It is always recommended to plan questions in the form of Questionnaire before going for the survey (questionnaire is a list of questions to be asked during the survey).

The questionnaire can have close ended and open ended questions.

Close ended questions are designed by giving prior options to the respondents. It makes the possible responses limited and focussed . It gives respondent little or no freedom to move beyond the given options.

Open ended questions allow respondents to speak their minds and answers the questions as per their claity.

It gives creative and diverse responses.

For predictions in Opinion poll , a combination of Open and closed ended questions is recommended in the Questionnaire , while for the Exit poll , the closed ended questionnaire is sufficient.

Whenever you go to collect the questionnaire-based data, people may not be very willing to as the questions you put to them. They may ask you the need and purpose of the questionnaire. It is, therefore important to provide all relevant information to the respondents regarding your study.

However , this is not precisely be the situation during polls in India . In India people indulge in political discussions , with all interest & enthusiasm , everywhere , whether that is home , offices , markets , or in small gatherings , may be at café or small tea stalls or during their walks & workouts. Among the factors that impact the results of Opinion Polls, are the wording and order of the questions being posed by the surveyor. Among the factors that impact the results of Opinion Polls, are the wording and order of the questions being posed by the surveyor.

Nonresponse bias :

Since some people do not answer calls from strangers, or refuse to answer the poll, poll samples may not be representative samples from a population due to a non-response bias. Response rates have been declining, and are down to about 10% in recent years. Because of this selection bias, the characteristics of those who agree to be interviewed may be markedly different from those who decline. That is, the actual sample is a biased version of the universe the pollster wants to analyze. In these cases, bias introduces new errors, one way or the other, that are in addition to errors caused by sample size. Error due to bias does not become smaller with larger sample sizes, because taking a larger sample size simply repeats the same mistake on a larger scale. If the people who refuse to answer, or are never reached, have the same characteristics as the people who do answer, then the final results should be unbiased. If the people who do not answer have different opinions then there is bias in the results. In terms of election polls, studies suggest that bias effects are small, but each polling firm has its own techniques for adjusting weights to minimize selection bias.

Response bias :

Survey results may be affected by response bias, where the answers given by respondents do not reflect their true beliefs. This may be deliberately engineered by unscrupulous pollsters in order to generate a certain result or please their clients, but more often is a result of the detailed wording or ordering of questions (see below). Respondents may deliberately try to manipulate the outcome of a poll by e.g. advocating a more extreme position than they actually hold in order to boost their side of the argument or give rapid and ill-considered answers in order to hasten the end of their questioning. Respondents may also feel under social pressure not to give an unpopular answer.

Some people responding may not understand the words being used, but may wish to avoid the embarrassment of admitting this, or the poll mechanism may not allow clarification, so they may make an arbitrary choice. Some percentage of people also answer whimsically or out of annoyance at being polled.

A common technique to control for this bias is to rotate the order in which questions are asked. Many pollsters also split-sample. This involves having two different versions of a question, with each version presented to half the respondents.

Unlocking the Power of Data :

Large Data :

If the data is very large it becomes difficult to handle it manually. We then take help of the available technology to manage huge data.

There are a number of software's available that can help you in organizing the data. Two of these are Microsoft Excel that comes with Microsoft Office and the other openofficeorg-cal which is an open software.

Visualization of Data:

Organizing the data in such a form is called a **frequency table or frequency distribution.** Since the data is discrete we may also call such a distribution as discrete frequency distribution. frequency means the number of times an item or a number or any quantity is repeated.

Continuous frequency distribution :

To manage huge data , a frequency table and a continuous frequency distribution is recommended.

But it is not always an easy task to accurately communicate your ideas, either verbally or visually. The pictorial representation of data was revolutionary , through tables & graphs .

Graphs convey comparative information in ways that no tables of numbers or written accounts could ever do. The eye can perceive trends, differences and association instantly that a brain would need few minutes to understand from a table. This makes it attractive to businessmen, scientists and other people. The chart allows the number to speak to all and transcends all boundaries. The data may be represented graphically through by ***Pie chart , Bar graph , Bar chart , Dot plot , Line Graph , Histogram.*** When quantitative information in a graphical form, is presented , one should consider which type of graph has been used, what trend or data in the pattern to emphasize and how to construct the actual graph. The graph may give a misleading impression to the reader if some aspect of a graph is distorted. Sometimes , during presentation a graph is distorted by shortening the vertical axis , by scaling and axis manipulation. In other words, the scale on the vertical axis may not start from zero.

Data Analysis :

We Analyse the Predictions on a larger set of data [Large Sample Tests , In Sampling] , we talk about the predictions of the results , with 5 % level of risk , the result value falling in Critical Region or the region of Rejection , or falling in 95% confidence Interval , not with 100% certainty.

Inferences drawn or Conclusions :

Our experiences, observations and analysis of a situation make it possible to derive inferences from the activities around us. Knowingly or unknowingly we all tend to draw some conclusions by either observing what is happening around us or by the information that we receive from different sources.

But , we should also be careful not to generalize by perceptions that have been created.

Margin of error due to sampling :

The possible difference between the sample and whole population is often expressed as a margin of error – usually defined as the radius of a 95% confidence interval for a particular statistic. One example is the percent of people who prefer product A versus product B. When a single, global margin of error is reported for a survey, it refers to the maximum margin of error for all reported percentages using the full sample from the survey. If the statistic is a percentage, this maximum margin of error can be calculated as the radius of the confidence interval for a reported percentage of 50%. Others suggest that a poll with a random sample of 1,000 people has margin of sampling error of ±3% for the estimated percentage of the whole population.

A 3% margin of error means that if the same procedure is used a large number of times, 95% of the time the true population average will be within the sample estimate plus or minus 3%. The margin of error can be reduced by using a larger sample, however if a pollster wishes to reduce the margin of error to 1% they would need a sample of around 10,000 people.

Sampling Error :

Polls based on samples of populations are subject to ***sampling error*** which reflects the effects of chance and uncertainty in the sampling process. Sampling polls rely on the ***law of large numbers*** to measure the opinions of the whole population based only on a subset, and for this purpose the absolute size of the sample is important, but the percentage of the whole population is not important (unless it happens to be close to the sample size).

In practice, pollsters need to balance the cost of a large sample against the reduction in sampling error and a sample size of around 500–1,000 is a typical compromise for political polls. (Note that to get complete responses it may be necessary to include thousands of additional participators) .

Another way to reduce the margin of error is to rely on ***poll averages***. This makes the assumption that the procedure is similar enough between many different polls and uses the sample size of each poll to create a polling average.

Another source of error stems from faulty demographic models by pollsters who weigh their samples by particular variables such as party identification in an election. One may underestimate a victory or a defeat of a particular party candidate that saw a surge or decline in its party registration relative to the previous election cycle.

A caution is that an estimate of a trend is subject to a larger error than an estimate of a level. This is because if one estimates the change, the difference between two numbers *X* and *Y*, then one has to contend with errors in both *X* and *Y*. A rough guide is that if the change in measurement falls outside the margin of error it is worth attention.

Method of Sampling and Size of Sample:

The most significant feature in the preparation of an opinion poll survey is the sample size and method of sampling. The sample size for any national- and state-level election study depends upon the level of analysis one intends to do. Thus if one wants to analyse the voting behaviour and attitudes of voters only at the state level, a survey of 1,500 respondents would be good enough. But if one also wants to do a region-wise analysis in the state, then the sample size should be bigger as there should be sufficient number of cases for disaggregate analysis.

For Large Sample tests [i.e., for $N \geq 30$] :

Test 1 : To test the significance of single proportion , $p = x/n$:

$z = [x - n P] / \sqrt{ n P Q }$

where $\sqrt{ n P Q } = S.E$

x = Number of Successes .

n = Number of independent Bernoullian trials .

P = Probability of success in each trial (to be tested under the hypothesis) .

Or

$z = [x / n - P] / \sqrt{ \{ P Q / n \} }$

where $S.E = \sqrt{ \{ P Q / n \} }$.

As an example , if tested on a sample of 1000 people verdict & 5195 people give their preference for a particular candidate or a particular party , one can apply this test , to conclude or draw the inference , whether the people's verdict is unbiassed one or not .

The predictions could be made , based on 5% , level of risk.

The probable error and Standard error could also be determined.

Test 2 : To test the difference between two sample proportions , from two populations :

p_1 and p_2 = proportions of two large samples of an attribute 'A' .

n_1 and n_2 = sample sizes of two samples .

$z = [(p_1 \sim p_2) \sim (P_1 \sim P_2)] /$

$\sqrt{ \{ P_1 Q_1 / n_1 + P_2 Q_2 / n_2 \} }$

Where $S.E = \sqrt{ \{ P_1 Q_1 / n_1 + P_2 Q_2 / n_2 \} }$

Under the Null hypothesis ,

$P_1 = P_2 = P$ (say)

Where $P = \{ n_1 p_1 + n_2 p_2 \} / \{ n_1 + n_2 \}$.

$Q = 1 - P$

$z_0 = (p_1 \sim p_2) / \sqrt{ \{ P Q / (1 / n_1 + 1 / n_2) \} }$.

As an example ,

if tested on a sample of 500 persons from town A , 200 are in favour of a particular party & 400 persons from town B , 250 are in favour of the same particular party , as that of town A , where , Town A & town B , both chosen from same constituency.

The test could be applied to discuss whether the data reveal a significant difference between the samples A and B , so far as the proportion of same party candidates are concerned.

Peoples verdict or preference for a particular party , could be tested , to conclude or draw the inference , whether the people's verdict is unbiassed one or not and the predictions could be done , based on 5% , level of risk.

Test 3 : To test the significance of the difference between $\bar{x}$ and μ , where $\boldsymbol{\sigma}$ **is known :**

$z = (\bar{x} \sim \mu) / (\boldsymbol{\sigma} / \sqrt{n})$

$\bar{x}$ = sample mean

μ = population mean

$S.E = \boldsymbol{\sigma} / \sqrt{n}$

As an example , if a sample of 400 individuals have given average marks to a party say , 67.47 .Can it be reasonably be regarded as sample from large population with mean 67.39 and S.D 1.30 ?

If yes , then the results could be predicted accurately , though with 5% level of risk .

Test 4 : To test the significance of the difference between two sample means , where

σ is known :

$Z = [(\bar{x}_1 \sim \bar{x}_2) \sim (\mu_1 \sim \mu_2)] / \sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}$

$\bar{x}_1$ & $\bar{x}_{2.}$ = means of two samples respectively.

μ_1 & μ_2 = population means

But , since under the Null hypothesis , the two samples are assumed to be taken from the same population , therefore , $\mu_1 = \mu_2$)]

$z_0 = (\bar{x}_1 \sim \bar{x}_2) / \sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}$

$S.E = \sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}$

As an example , if the means of two single large samples of 1000 & 2000 members are 67.5 and 68.0 respectively , in favour of a particular party , may be scaled on from 100 marks , 10 questions asked , each question having marked from a score of 10 marks , in a questionnaire , can the samples be regarded as drawn from the same population [same constituency] , of S.D 2.5 ?

Again , it could be tested at 5 % level of significance.

The above tests could be applied on large samples .As the constituency size is vast , the above large sample tests could be proved very useful , if applied by a well equipped Statistician.

Small Sample Tests :

Even the importance & usefulness of small sample tests could not be ignored , and could be proved very useful , in making the predictions.

Even small- samples can be used to obtain estimates of national vote intentions that are well within acceptable deviations from actual vote outcomes.

Thus sample sizes for any election survey depend upon the level of disaggregate data one requires for analysis. As long as the sample size is representative, seat predictions based on even a small sample size can come up with accurate results.Though The method of sampling used in an election survey and its accuracy also plays an important role in making a reasonably accurate election forecast.

The combination of methods of Large samples and small samples both , depending upon which method is suitable for the tested criteria , should be focussed as a deciding factor.

Yogendra Yadav has argued that there is no guarantee that a bigger sample size will get you the right result. Bigger surveys only multiply errors 10 times *(Nagaraj 2008)* **.** Both of them have emphasised on small samples , on the grounds that On the contrary, a survey based on a large but unrepresentative size will yield a wrong seat forecast.

Thus a smaller representative sample can help make an accurate predication as compared to a bigger unrepresentative sample. The method of sampling used in surveys done by market research organisations is usually not the most scientific method of survey research.

This becomes clear when measuring voting preferences and intentions of castes and communities. For example, in UP, a majority of the Dalit's, especially the Jatavas, have been voting for the BSP while the majority among the upper-caste voters have been supporters of the BJP and Congress. Similarly the voters belonging to the Yadav's have been staunch supporters of the Samajwadi Party. Thus, the sample in UP should be representative of these caste and communities, approximating the percentage of the population of these communities in the state.

So if a sample survey fails to gather the opinion of any important caste and community, the election predictions will be highly vulnerable to failure**.** *Yogendra Yadav (2008)* has argued that the method through which the sample is selected is crucial for the survey.

Some relevant , useful , efficient & accurate small sample tests for polling :

Test 1 - t - Test : To test the significance of the difference between two correlation coefficients of the two samples :

Assuming the

Null Hypothesis (H_0) : r_1 does not differ significantly from r_2 .

$t = (z_1 - \mathbf{z_2}) \ / \ \sqrt{[\, 1 / (n_1 - 3) + 1 / (n_2 - 3) \,]}$

where

$z_1 = ½ [\log_e (1 + r_1) / (1 - r_1)]$

$= 1.1513 [\log_{10} (1 + r_1) / (1 - r_1)]$

$z_2 = ½ [\log_e (1 + r_2) / (1 - r_2)]$

$= 1.1513 [\log_{10} (1 + r_2) / (1 - r_2)]$

If $| t | > 1.96$, the difference is significant at 5 % level , i.e., $t_0 \in$ W or W is significant.

r_1 differs significantly from r_2 .

Null hypothesis will be rejected in this case.

This test could be used to check whether the samples collected from the population of the same constituency , are correct or not .

As an example , the first of two samples consists of 23 pairs and gives a correlation coefficient of 0.5 , while the second of 28 pairs has a correlation coefficient of 0.8 . Are these values significantly differ ?

If the calculated value of t , in this case will come out to be 1.83 < 1.96 , therefore , the difference of the two samples correlation coefficient is not significant.

Test 2 : Chi square test of goodness of fit :

If O_i (i = 1 , 2 , ….., n) is a set of observed frequencies and E_i

(i = 1 , 2 , ….., n) is the corresponding set of expected frequencies , then Karl – Pearson's Chi – square is given by

$\chi_i{}^2 = \boldsymbol{\sum (O_i - E_i)^2 / E_i}$

If calculated value of χ^2 < tabulated value of χ^2

, for (n – 1) d.f , then accept H_0 .

Contingency Table : Let the data be classified into p classes , A_1 , A_2 , …., A_p ,

according to the attribute A , and q classes , B_1 , B_2 , …., B_q ,

according to the attribute B , so that there will be p q classes , according to both the attributes A & B. Let O_{ij} denotes the cell frequency of the cell , belonging to both the classes A_i and B_j **;** i = 1 , 2 , ….., p ; j = 1 , 2 , ….., q .

Let (A_i) and **(** B_j **)** be the totals of the frequencies belonging to both the classes A_i and B_j respectively**.**

Then the data can be set out in the form of the table , called the p × q contingency table , with p rows & q columns.

B--> **A ↓**	B_1	B_2	B_j	B_q	**Totals**
A_1	$\mathbf{O_{11}}$	$\mathbf{O_{12}}$	$\mathbf{O_{1j}}$	$\mathbf{O_{1q}}$	**(A_1)**
A_2 ….	$\mathbf{O_{21}}$	$\mathbf{O_{22}}$	$\mathbf{O_{2j}}$	$\mathbf{O_{2q}}$	**(A_2)**
A_i …..	$\mathbf{O_{i1}}$	$\mathbf{O_{i2}}$	$\mathbf{O_{ij}}$	$\mathbf{O_{iq}}$	**(A_i)**

A_p	O_{p1}	O_{p2}	O_{pj}	O_{pq}	(A_p)
Totals	(B_1)	(B_2)	(B_j)	(B_q)	**N**

$$E [O_{Ij}] = e_{Ij} = [(A_I) \times (B_j)] / N$$

Degrees of freedom for a contingency table (d.f) = v { greek symbol , nu } = (p - 1) (q – 1)

As an example , the two main contesting parties , in a particular contingency may be taken row – wise & the for the different criterions on which they have to be judged viz., say poverty , price rise , women relayed problems , education , health , economy , etc . could be taken column – wise , and a contingency table be formed .

The chi – square test could give the very appropriate & useful results.

2. Conclusion

The goal of sampling strategies in survey research is to obtain a sufficient sample that is representative of the population of interest. It is often not feasible to collect data from an entire population of interest (e.g., all individuals with interest to vote ; therefore, a subset of the population or sample is used to estimate the population responses . A large random sample increases the likelihood that the responses from the sample will accurately reflect the entire population. In order to accurately draw conclusions about the population, the sample must include individuals with characteristics similar to the population.Also , amongst the small sample test , the t – test & Chi – square test , are good measures to identify the accuracy of forecasts efficiently.

Survey research is a useful and legitimate approach to research that has clear benefits in helping to describe and explore variables and constructs of interest. Survey research, like all research, has the potential for a variety of sources of error, but several strategies exist to reduce the potential for error. Advanced practitioners aware of the potential sources of error and strategies to improve survey research can better determine how and whether the conclusions from a survey research study apply to practice.

References

[1] Butler, David, Lahiri, Ashok and Roy, Prannoy (1995): India Decides. Elections 1952-1995, Delhi: Books & Things, (1995).

[2] Chatterjee, Somdeep and Kamal, Jai (2020): ‘Voting for the Underdog or Jumping on the Bandwagon?: Evidence from India’s Exit Poll Ban’, Public Choice, July 2020.

[3] Cook, T. & David L.DeMets , (2008) . Introduction to Statistical Methods for Clinical Trials (Chapman & Hall/CRC Texts in Statistical Science) 1st Edition. Chapman and Hall .

[4] Deshpande, Rajeshwari (2019): ‘Why Do We Need Election Studies in Political Science Classrooms?’, Studies in Indian Politics, 7(2) , 262–266, 2019.

[5] Dixon, John (2006): ‘Eric P. W. Da Costa: Polling Pioneer of India’, International Journal of Public Opinion Research, Vol. 18 No. 1, Oxford University Press, 2006.

[6] Franklin, C. F (2008): ‘How Pollsters Affect Poll Results’, Political Arithmetic, 2008. Available at http://politicalarithmetik.blogspot.com/search?house+effect

[7] Graefe, Andreas (2014): ‘Accuracy of Vote Expectation Surveys in Forecasting Elections’, The Public Opinion Quarterly, Vol. 78, Oxford University Press, 2014.

[8] Jackman, Simon (2005): ‘Pooling the polls over an election campaign’, Australian Journal of Political Science, Volume 40, 2005 - Issue 4, pp.499-517, 2005.

[9] Jaffrelot, Christophe and Verniers, Gilles (2009): ‘India’s 2009 Elections: The Resilience of Regionalism and Ethnicity’, South Asia Multidisciplinary Academic Journal (Online), 3, 2009. https://doi.org/10.4000/samaj.2787

[10] Karandikar, Rajeeva L. (2014): ‘Elections 2014, Power And Limitations Of Opinion Polls: My Experiences’, The Hindi Centre For Politics And Public Research, 2014.

[11] Karandikar, Rajeeva L., Payne, Clive and Yadav, Yogendra (1998): ‘Predicting the 1998 Indian parliamentary election’, Electoral Studies, Volume 21, Issue 1, pp. 69-89, March 2002.

[12] Kondo, Norio (2007) : Election Studies in India, Institute of Developing Economies: Discussion Paper No.098, 2007. URL: https://www.ide.go.jp/English/Publish/Download/Dp/098.html

[13] Kothari, Rajni (2002): Memoirs. Uneasy is the Life of the Mind, Delhi: Rupa & Co., 2002.

[14] Kumar, Sanjay, Rai, Praveen and Gupta, Pranav (2016) : ‘Do surveys influence results?’, Seminar, 684, August 2016.

[15] Kumar, Sanjay and Rai, Praveen (2013) : Measuring Voting Behaviour in India, Delhi: Sage Publications, 2013.

[16] Lokniti Team (2004): ‘National Election Study 2004: An Introduction’, Economic and Political Weekly, pp. 5373-81, 18 December 2004.

[17] Nagaraj, Anuradha (2008): "Psephology Is Not a Science Like Microbiology.. It's Poll Studies. But Everyone Thinks Only of Seat Forecasts",The Indian Express, 27 January, available at http://archive.indianexpress.com/news/-pse-phology-is-not-a-science-like.-.., accessed on 11 April 2014

[18] Nath, Tripti and R Suryamurthy (1999) : "Poll Pundits or Punters?", The Tribune, 28 August, available at http://www.tribuneindia.com/1999/99aug28/saturday/headı.htm, accessed on 11 April 2014

[19] Narain, Iqbal; Pande, K.C.; Sharma, M.L.; Rajpal, Hansa (1978): Election Studies in India: An Evaluation, New Delhi: Allied Publishers, 1978.

[20] Northcott, Robert (2015): ‘Opinion Polling and Election Predictions’, Philosophy of Science, Vol. 82, No. 5, The University of Chicago Press, pp. 1260-1271, December 2015.

[21] Ponto , J , (2015) , Understanding and Evaluating Survey Research , Journal of the Advanced Practitioner in Oncology , Harborside Press .

[22] Rai, Praveen (2021): ‘Demystifying the Bandwagon Effect of Election Opinion Polls in India’, Academia Letters, Article 3042, August 2021.

[23] Rai, Praveen (2021): ‘Psephological Fallacies of Public Opinion Polling’, Economic & Political Weekly, Volume LVI, No. 28, 10 July 2021.

[24] Rai, Praveen (2014): ‘Fallibility of Opinion Polls in India’, Economic & Political Weekly, Volume XIIX no. 18, 3 May 2014.

[25] Roy, Prannoy and Sopariwala, Dorab R (2019): The Verdict: Decoding India’s Elections, Penguin Random House India, 2019.

[26] Sanjay Kumar & Praveen Rai , 2013 , Measuring Voting Behaviour in India

[27] Traugott, Michael W (2014): ‘Public Opinion Polls and Election Forecasting’, Political Science and Politics, Vol. 47, No. 2, 2014, pp.342-344, April 2014.

[28] Traugott, M. W (2012): ‘Data Quality from Low Cost Data Collection’, unpublished paper presented at the annual conference of the Association for Public Opinion Research, 2012.

[29] Walther, Daniel (2015): ‘Picking the winner(s): Forecasting elections in multiparty systems’, Electoral Studies, Volume 40, pp. 1-13, 274 Praveen Rai

[30] Yadav, Yogendra (2008): ‘Whither Survey Research? Reflections on the State of Survey Research on Politics in Most of the World’, Malcolm Adise-shiah Memorial Lecture, Chennai

Citation: *DR. AMBIKA BHAMBANI, Robust & Reliable Statistical Projection During Opinion & Exit Polls in India Based on Sampling Forecast Models International Journal of Scientific and Innovative Mathematical Research (IJSIMR), vol. 11, no. 1, pp. 13-29, 2023. Available : DOI: https://doi.org/ 10.20431/2347-3142.1101002*