International Journal of Mathematics and
Computer Science, **18**(2023), no. 4, 723–730

Probability Sampling-Key to Reliable Cross-sectional Survey Results: An Election Poll Case

Ivy Corazon A. Mangaya-ay

Mathematics and Natural Sciences Department
College of Arts and Sciences
Bohol Island State University
CPG Avenue, 6300 Tagbilaran City
Bohol, Philippines

email: ivycorazon.mangayaay@bisu.edu.ph

(Received April 4, 2023, Accepted May 22, 2023,
Published May 31, 2023)

Abstract

This paper emphasizes the significance of probability sampling in surveys; particularly, in the context of cross-sectional study design. It showcases pre-election survey results that illustrate its importance in producing trustworthy results. To attribute the survey results to the sampling technique, the survey operation design controls the different types of errors, including frame error, nonresponse error, measurement error, and specification error. Ultimately, after minimizing other sources of error, sampling is the sole remaining major source of error. In sample size determination, stratified random sampling is used, and the 'samplingbook' package in *R* Programming Software is used to compute the sample. This study concludes that appropriate sampling is essential in ensuring reliable survey results.

1 Introduction

The basic premise of statistical inference is the use of a sample to generalize a characteristic of a population. However, if one must generalize, a suffi-

Key words and phrases: Probability Sampling, Stratified Random Sampling, Election Poll, Cross-sectional Survey Design.
AMS (MOS) Subject Classifications: 00-xx.
ISSN 1814-0432, 2023, http://ijmcs.future-in-tech.net

cient and appropriate sample size [1] is essential. The sample selection must be scientific and random to have reliable results. Using a random method to choose a sample increases the likelihood that the sample will be highly representative [4].

Election polling is perhaps one of the most prominent applications of statistics, exemplifying the use of random sampling and representing one of the significant successes of the field [6]. It is with election polls that one can assess immediately if the estimates are close to the parameter by comparing it with actual results. An election poll is a concrete example of a cross-sectional survey design, which allows the completion of the survey for a specific time point or within a short time frame. The defining and salient aspect of this design is its reliance on a representative cross-sectional sample of the population, which enables the generalization of the study findings to the larger population [2]. Developing a robust sampling strategy is a crucial element in designing this particular study design as the target population often exhibits heterogeneity. It is essential to have a solid sampling plan in place to address this variability and ensure the selection of a representative sample [3].

Frame Error arises when the sampling frame does not match the target population, as in web-based and internet panel surveys, when those who do not have the internet are excluded. On the other hand, a Nonresponse Error occurs when missing values are systematically related to the response. It is important not only for the selected sample to be representative but also for the respondents. For example, a nonresponse error emerges when supporters of the trailing candidate are less likely to respond to surveys [6]. The last 2016 USA Election failed to accurately predict the outcome; the actual election result differed greatly from the projections of a Clinton victory. According to some experts, the possible culprit is that certain people systematically do not respond to surveys (Pew Research Center, 2016). These are voters who are a pro-Trump segment of the population that choose not to participate in the poll surveys, which have had a larger share of actual voters than what was accounted for in the surveys. Nonresponses, in this case, generate biased results. Furthermore, Measurement Error exists when the survey instrument or questionnaires itself influences the response, such as the phrasing and order of survey items. Finally, specification error happens when a respondent interprets a question differently than the aim intended.

Existing literature has advanced considerably in recent years regarding the advantages of appropriate sampling in achieving reliable results. However, much of this literature heavily relies on proof via simulation or theo-

retical analysis. In this paper, I aim to substantiate the significant role of appropriate sampling in surveys in a more easily digestible manner without the need for extensive mathematical or statistical knowledge. To achieve this, I present a use case of pre-election polling, which allows for a comparison between estimated outcomes and actual results, providing a more convincing illustration for the readers. To highlight and emphasize the role of sampling, the pre-election survey is designed by eliminating other sources of errors such as frame, measurement, specification, and non-response errors.

The study controlled the frame error sample on the list of voters from Commission on Election (COMELEC) records. The survey questionnaire is designed with neutrality and is meant to be impartial with no traces of a party or candidate affiliation to reduce measurement error. Moreover, it resembles an election ballot wherein the respondents were asked about the preferred candidate should the actual election be held during the conduct of the study to get rid of specification error issues. Similar to the actual ballot, candidates are listed in alphabetical order and the respondents are to shade the circle beside the preference. Furthermore, a random replacement of non-response is the strategy to weaken if not diminish the impact of non-response.

2 Sample Size and Sampling Technique

The pre-election survey utilizes stratified sampling with the number of voting populace in each barangay, town, and district as the stratifying variables for town, district, and province-level elections, respectively.

$$n_o = \frac{1}{(\frac{d}{z_{a/2}})^2} \sum_{h=1}^{L} W_h P_h Q_h.$$

If $\frac{n_o}{N} < 0.05$, then $n = n_o$. Otherwise,

$$n = \frac{n_o}{1+\frac{1}{N(\frac{d}{z_{a/2}})^2} \sum_{h=1}^{L} W_h P_h Q_h}$$

N = Total number of voters
L = Total number of barangay (town-wide) or town (district-wide)
N_h = Number of voters that belong in the h^{th} stratum (barangay, town, district)
$W_h = h^{th}$ stratum weight = $\frac{N_h}{N}$
d = Margin of error
α = Level of significance

$Z = z$ value at $\alpha/2$
P_h = the proportion of the garnered votes in the past (latest) election
$Q_h = 1 - P_h$.
R Programming Software was used in the computation of the sample using the 'samplingbook' package. The package implements the above formula using the 'stratasize' function. The function asks to specify the margin of error, the level of confidence, and the data frame containing the list of barangay (town-level) or town (district-level), number of voters and the standard deviation of the percentage of votes of the winning candidate in the latest election.

2.1 Sampling Technique

The sample frame used is the list of voters from COMELEC. The selection of respondents used a randomization algorithm via R. The same method chose reserved respondents as replacements for unavailable respondents. A respondent can be categorized as unavailable when s/he: refuses to answer, has transferred residence, is dead, could not be found in the specified address, or is currently in prison. To control the number of non-response, only a few reserved respondents (for replacements) are allocated.

2.2 Survey Questionnaire

The administered pre-election survey used a concise questionnaire, consisting of a single page. It was patterned after the actual election ballots, with the names of the candidates arranged in alphabetical order. It typically asked how the respondents would vote "if the election were held today". The questionnaire is designed to be straightforward to prevent misinterpretation or indication of allegiance and bias to a candidate or a party. It went through pilot testing, to test its understandability and legibility.

Each town (barangay) in a district (town) has different color codes for printed survey questionnaires. The color indicates the town a respondent belongs to without necessarily writing the town's name in the questionnaire-again to avoid biased responses. The respondents have the feeling of discomfort in responding when their identity is exposed-especially in a district or town where the constituents know each other.

2.3 Data Collection

1. There was an enumerator's briefing prior to the study. They are to pilot the survey questionnaire to practice ease in conducting the study.

2. The respondents themselves will respond to the questionnaire using the paper-pencil method. The only time the enumerator speaks to the respondent is when the former tries to convince the latter to participate in the political survey. This is to eliminate the possible effect of the interviewer's attitude on the responses.

3. The respondents do not need to write their names on the survey questionnaires to avoid their discomfort, thus eliciting more honest answers.

4. The respondents' utmost confidentiality was tightened by letting them drop the survey questionnaire in a secured box. Even the enumerators do not have access to their responses.

5. The questionnaires must be administered to the respondents as ordered in the list (randomly generated respondents)

6. If the respondent is unavailable, the enumerator shall proceed to the subsequent respondents as sequenced on the list.

7. A field evaluator randomly checked if, indeed, the enumerator went to the field and administered the pre-election survey questionnaire properly.

2.4 Data Processing

The data preparation and data management are in Excel. To check the completeness of the data, frequencies were run in each item. There was a random check of whether the encoded data and the filled-out survey questionnaire matched.

2.5 Computation of the Percentage of Votes

The Overall Mean Percentage of votes was computed with the number of registered voters as weights and the following formula was used:
$\bar{Y}_{st} = \frac{\sum_{h=1}^{15} N_h Y_h}{N}$,
where

Table 1. Pre-election vs Actual Election Results

MATCH	POPULATION SIZE	SAMPLE SIZE	MARGIN OF ERROR %	CONFIDENCE LEVEL %	CANDIDATE	PRE-ELECTION	ACTUAL %
A	233 942	173	8	95	A.1*	85.45	70.95
					A.2	4.85	16.18
					A.3	9.70	12.87
B	284 216	127	15	95	B.1*	49.59	56.89
					B.2	39.02	38.16
					B.3	8.13	4.38
					B.4	0	0.34
					B.5	3.25	0.23
C	303 915	131	9	95	C.1*	75.91	65.03
					C.2	23.36	34.06
					C.3	0.73	0.52
D	9920	175	6	90	D.1*	68.88	59.84
					D.2	31.22	40.16
E	9920	175	6	90	E.1*	50.03	54.78
					E.2	49.97	45.22
F	887771	374	5	95	F.1*	59.77	64.37
					F.2	16.82	18.09

$\bar{Y}_{st}$ is the mean percentage of votes using stratified random sampling as a sampling technique,
N_h is the number of registered voters in the h^{th} town,
Y_h is the percentage of votes in h^{th} district, and
N is the total number of registered voters.

3 Comparison of the Polling and Actual Election Results

Table 1 displays the survey results of each candidate in different races with the associated margin of error, and the actual percentage of votes they garnered in the actual election. Here, the candidate with an asterisk (*) next to their name won the election.

Results show that not only the pre-election polling was able to successfully predict the winning candidate, but the percentage of votes garnered was also able to capture the percentage of votes within the margin of error. T-test for the dependent sample signifies that there is no significant difference ($t_{16} = -0.176, p = 0.83$) between pre-election and actual results. Indeed, the sampling method is suitable and efficient. Although probability samples

are not flawless and may encounter non-sampling errors and non-response, they do eliminate potential sources of bias such as the likelihood of advocacy groups skewing poll results by influencing others to abstain from participating [5].

The accuracy of survey data and findings can be compromised by non-sampling errors but when these are controlled for, the results can align with the actual numbers as evidenced by the pre-election survey results. To prevent errors in election surveys, it is crucial to address them during the design phase [7].

4 Conclusion

Probability sampling has played a significant role in achieving pre-election results that accurately reflect the true numbers. The election poll case substantiates claims that appropriate sampling technique is crucial to survey research, especially in cross-sectional studies. Furthermore, it is essential not only to select an adequate sample size but also to ensure that responders are randomly selected in order to make valid population generalizations.

While this paper underscores the significance of random sampling, it is worth noting that other methodological factors can impact the quality of data and results. Future research endeavors may also consider magnifying the lenses on other sources of error.

The findings and results of this study have many practical implications for cross-sectional studies that aim to achieve generalization. It is common to theses, dissertations, and even institutional researches to overlook the importance of sampling methodology, often simply reporting a formula. However, the quality of results depends on the quality of input and sampling is of great influence in this regard.

References

[1] M. A. Memon, H. Ting, J. H. Cheah, R. Thurasamy, F. Chuah, T. H. Cham, The sample size for survey research: Review and recommendations. Journal of Applied Structural Equation Modeling, **4,** no. 2, (2020), 1–20.

[2] A. Omair, Selecting the appropriate study design for your research: Descriptive study designs. Journal of Health Specialties, **3,** no. 3, (2015), 153.

[3] X. Wang, Z. Cheng, Cross-sectional studies: strengths, weaknesses, and recommendations. Chest, **158,** no. 1, (2020), S65–S71.

[4] K. Levin, Study design III: Cross-sectional studies. Evid. Based Dent., **7,** 24-25, (2006). https://doi.org/10.1038/sj.ebd.6400375

[5] S. L. Lohr, Sampling: design and analysis. CRC Press, 2021.

[6] H. Shirani-Mehr, D. Rothschild, S. Goel, A. Gelman, Disentangling bias and variance in election polls. Journal of the American Statistical Association, **113,** (522), (2018), 607–614.

[7] R. S. Kenett, D. Pfeffermann, D. M. Steinberg, Election Polls-A Survey, A Critique, and Proposals, Annual Review of Statistics and Its Application, **5**, (2017), 1–24.