Estimating Vote Counts with Limited Electoral Integrity

Dalson Figueiredo
Federal University of Pernambuco, Brazil

Jose Antonio Gomez Duarte*
Somewhere in Venezuela

Raphael Nishimura
Institute for Social Research, University of Michigan

Walter R. Mebane, Jr.
Department of Statistics, University of Michigan

July 28, 2024

Abstract

This report presents a method for quickly estimating vote counts in contexts with constrained electoral integrity. Using disaggregated data from Venezuelan elections from 2013 to 2021, we stratified polling stations along a seven-point scale ranging from most favorable to the opposition to most favorable to the government coalition. We then describe the sample selection and estimation procedures to obtain participation rate and voting estimates for the 2024 presidential elections based on the data from a sample of polling stations collected on the election night of July 28. This framework ensures reliable outcomes in vote counting and it is easily adaptable for comparisons both within and across countries. To increase the transparency and replicability of our findings, we provide a detailed spreadsheet of all calculations, along with supplementary slides that explain the model's rationale.

Keywords: vote estimation methodology; electoral integrity; Venezuela elections.

1 Introduction

Free and fair elections are the cornerstone of representative democracy and rely on a comprehensive set of procedures, including maintaining a detailed voter registry, accurately counting votes, and resolving disputes peacefully (Hyde, 2011). Conceptually, electoral integrity, founded on the principles of universal suffrage, political equality, and transparency, is essential to ensure that elections genuinely reflect the will of the voters (Birch, 2011; Donno, 2013; Norris, 2013; Przeworski, 2018).

In this technical report, we present a novel method for estimating vote counts in institutional environments where electoral integrity is compromised. Using historical data from the Venezuelan Presidential elections in 2013, Parliamentary elections in 2015 and 2020, and the Regional elections in 2021, we classify polling stations on a seven-point scale, ranging from those most favorable to the opposition (1) to those most favorable to the government coalition (7). This procedure is key for identifying areas more inclined towards change or maintaining the status quo. Our framework simulates vote counts for both incumbent and challenger candidates based on a probabilistic distribution of polling stations expected to report data on election night. We adjust these estimates to account for geographical variations in voter turnout over time, enhancing the precision of the statistical model.

The selection of Venezuela as a case study is justified on the following grounds. First, Venezuela shows one of the lowest scores on the Perceptions of Electoral Integrity (PEI), an index that quantifies the perceived integrity of elections (Frank and Ferran, 2017; Partheymüller et al, 2022). Figure 1 depicts PEI average scores for the 2020 and 2022 election cycles.

*Our contributor from Venezuela requested the use of a pseudonym due to the risks associated with their participation.

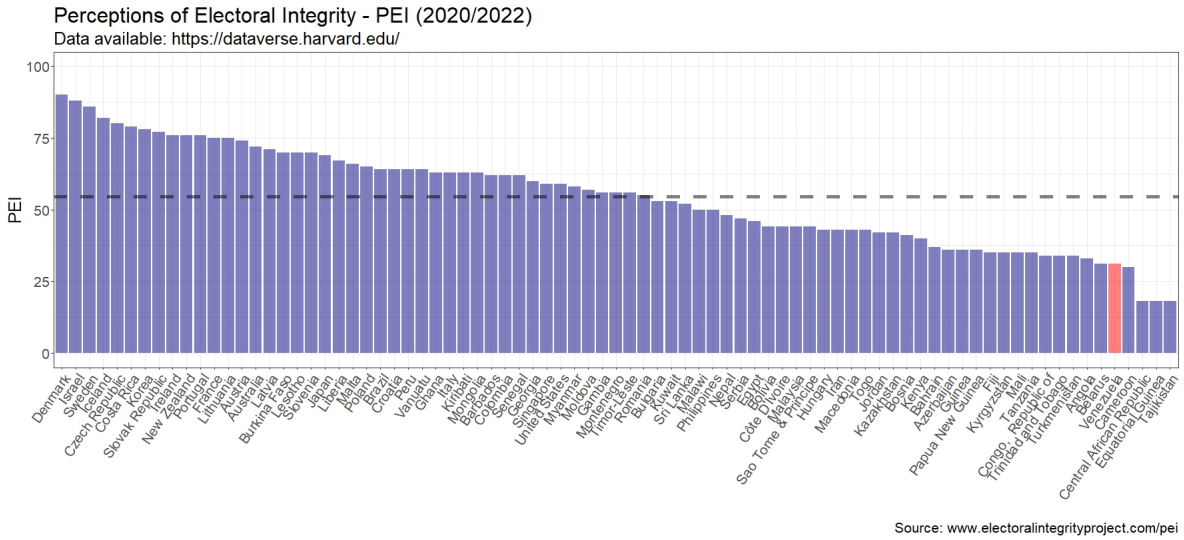

Figure 1: Perceptions of Electoral Integrity

The index ranges from 0 to 100, where a higher score signifies a greater perception of electoral integrity. The dotted line indicates an average score of 54 from a sample of 83 countries. Denmark (90), Israel (88), and Sweden (86) lead the ranking, while Tajikistan, the Central African Republic, and Equatorial Guinea show the lowest, all coming in at 18. Brazil (64) outperforms Italy (63), Colombia (62), and the United States (59). Venezuela (31), lagging behind Angola (33), is tied with Belarus, noted for its significant democratic deficits.

Second, a growing body of literature points to Venezuela as an extreme case of lacking electoral integrity (Harding, 1993; Levin et al, 2009; Levin and Alvarez, 2012; Jiménez and Hidalgo, 2014)[1]. Some countries do not have an independent and impartial Electoral Management Body (EMB). In some nations, practices like vote buying (Schaffer and Schedler, 2007), clientelism (Hicken, 2011), and the involvement of formal institutions such as the military and police pose threats to the integrity of elections (Sawasdee, 2019). Other countries suffer from severe political violence, including murders and kidnappings. In all these scenarios, it is no longer feasible to assume that elections are free and fair.

Finally, Nicolás Maduro is attempting to be reelected for the third time. He took office in March 2013 and has served for almost twelve years. In Adam Przeworski's minimalist concept of democracy, the alternation of power ensures that no group or individual remains indefinitely in control, thus preventing an excessive concentration of power. This alternation of power is precisely what is missing in Venezuela's political regime. More recently, María Corina Machado, a key opposition leader, reported that her security chief, Milciades Ávila, was detained, marking another arrest in a series targeting opposition figures[2].

The opposition is rallying behind Edmundo González, a leading retired diplomat. González decried Ávila's arrest in a video, calling it a regime abduction made under false charges of gender violence. Gonzalo Himiob Santomé, vice-president of a leading Venezuelan human rights NGO, noted that these arrests represent a systemic crackdown on opposition activists and affiliates[3]. Such concerns have only amplified after Maduro's comments on July 19, where he proclaimed that "Venezuela's destiny in the 21st century depends on our victory on July 28. If you do not want Venezuela to fall into a bloodbath, into a fratricidal civil war, a product of the fascists, let us guarantee the greatest success, the greatest victory in the electoral history of our people"[4].

[1] According to Jiménez and Hidalgo (2014), a summary of the alleged electoral irregularities during the Chavismo are available at America's Quarterly.

[2] According to Human Rights Watch, in 2023 more than 270 political figures are in jail. See: HRW Report on Venezuela.

[3] See: Guardian article on Venezuela's opposition arrests

[4] See: Mercopress on Maduro's comments

Given its historical and current conditions, Venezuela presents an excellent case study to evaluate an analytical framework for examining vote counts in institutional settings characterized by compromised electoral integrity. This report makes two important contributions. The first is methodological. Our framework is costless and easily adaptable for cross-country and within-country comparisons. Our second contribution is substantive. After acquiring historical data, we clustered polling stations based on their likely ideological leanings. This classification allowed us to obtain precise stratified estimates, which we used to extrapolate vote counts in settings with limited information. The results are robust and can assist international organizations and third-sector entities committed to ensuring respect for universal suffrage, equal opportunities, and transparency in electoral processes.

The remainder of the document is organized as follows: the next section outlines the data and methodological procedures used in our framework. We conclude with a discussion on the potential implications of our findings for electoral integrity and recommendations for the practical application of our methodology in other electoral contexts.

2 Data and Methods

This section describes the methodology developed to estimate the distribution of votes in the 2024 Venezuelan presidential elections for Nicol´as Maduro (incumbent) and Edmundo Gonz´alez (challenger). Figure 2 depicts the framework, which is divided into four steps.

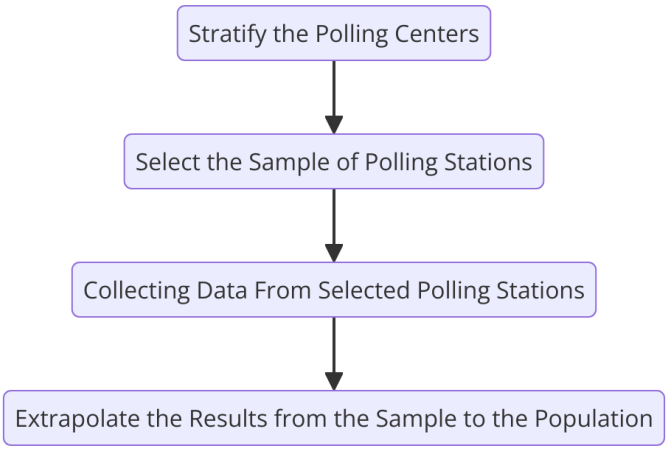

Figure 2: Vote count estimation step by step

2.1 Stratify The Polling Centers

The first step is to stratify the polling centers using historical data from four election cycles: the Presidential elections in 2013, the Parliamentary elections in 2015 and 2020, and the Regional elections in 2021. For each of these elections, we ordered voting centers according to the percentage of votes obtained by the government coalition. As the opposition boycotted the 2020 election, we instead ordered these centers by participation rate, as a 0 indicates that all eligible voters in a center followed the opposition's boycott. Then, we averaged the relative positions of each voting center across the four elections. This ranking placed each voting center on a scale from 0 to 1, with those closest to 0 being the most favorable to the opposition. For new voting centers, the following steps were taken: first, we determined the origin of the citizens registered at each center. If more than two-thirds of these citizens came from a previously existing voting center in the same municipality, we assigned the new center the rank calculated for the original center. Otherwise, we assigned the new voting center a rank of 1, as most new centers are located in areas challenging for the opposition. With the current data in the electoral registry, we ranked the voting centers based on their average relative positions across the four elections and stratified them in that order into seven strata such that each stratum has approximately the same number of registered voters.

2.2 Select the Sample of Polling Stations

The second step is to use information from Step 1 to select the sample. We selected a systematic sample of 1,500 polling stations. We started by ordering the polling stations by strata and then by geographical location (State, Municipality, Parrish, voting center, and Polling station number). Each polling station had a consecutive number from 1 to 30,026. Then, we determined a random seed between 1 and 20. That will be the first polling station in the sample. We then add to the seed value a sampling interval of 20.01733 (= 30,026/1500) and round it. That was the second polling station in the sample. This process is repeated until the selection number exceeds 30,026; thus obtaining 1,500 polling stations.

2.3 Collecting Data from Selected Polling Stations

On the night of July 28, 2024, after the presidential elections, our team will collect official data from the 1,500 polling stations selected in the sample. This information will be incorporated into a spreadsheet and dashboard. The data will be gathered using the automated tally sheets produced by the voting machines, which are audited according to Venezuelan laws and electoral regulations. For each polling station, we will have the total number of voters who participated in the election, the total number of votes for each candidate, and the total number of null/blank votes.

2.4 Extrapolate the Results from the Sample to the Population

In our framework, we will use information from data collected in the sample of polling stations to estimate the participation rate and distribution of votes for each candidate of the target population of 30,026 polling stations with 95% confidence intervals. As described above, we selected 1,500 polling with systematic sampling across the seven strata to provide data for this statistical inference. However, for various reasons, including lack of cellular communication and interference or intimidation, the team might not be able to collect data in some of the selected polling stations. In order to provide statistical inference, we will tackle this missing data issue using two approaches.

First, during the night of July 28, 2024, while the team collects data in real-time and obtain partial results dynamically, we will assume that any unobserved polling station follows a missing at random mechanism conditioned to strata. That is, within the stratum, we would not expect systematic differences with respect to participation and voting distribution between observed and unobserved polling stations. Any differential missingness would be explained by the strata and will be adjusted through stratum-level weights. While this could be considered a plausible assumption, given how polling stations were stratified, creating very homogenous strata, there might still be some unexplained systematic differences between observed and unobserved polling stations within the stratum. To account for those, we also propose a second approach next.

After all the data has been collected, we will work with a more complete, yet not fully observed sample of the polling stations to develop further adjustment weights to account for any additional systematic differences between observed and unobserved polling stations within strata. For such, we will use contextual and voting history variables of each polling station. By conditioning on these additional variables, we expect that the missingness at random assumption to become more plausible and provide more robust results against bias due to non-observation of certain polling stations selected in the sample.

Estimation procedures for the participation rate and voting distribution along with their 95% confidence intervals (margin of error) are provided in the Appendix of this report and accounts for the sample design features.

3 Conclusion

This report presented a methodology for quick estimating vote counts in contexts where electoral integrity is compromised, using Venezuela as a case study. By stratifying the polling stations on a seven-point scale, selecting pooling stations with a probability sample, obtaining official election data from the selected polling stations, and providing estimation procedures that accounts for the sample design, our approach enhances the accuracy of estimating electoral outcomes and provides insights into the challenges of conducting elections in politically volatile environments.

The findings underscore the importance of maintaining electoral integrity to ensure that election outcomes reflect the true will of the people. This study's methodology can be instrumental for electoral observers, policy-makers, and researchers aiming to understand and mitigate the impacts of compromised electoral processes. The detailed breakdown of voting patterns across different strata offers a clear view of how political biases and logistical challenges can affect the accuracy of vote counts.

According to Levin and Alvarez (2017), "data availability and data quality are also issues for those who wish to study election integrity and fraud" (Levin and Alvarez, 2017: 5). For future research, we recommend applying this methodology to other regions with similar challenges to validate its effectiveness and adaptability. Additionally, further refining the model to incorporate real-time data and developing predictive analytics could offer more timely insights for election monitoring.

Overall, this research advances the literature on electoral integrity and provides practical tools for enhancing transparency and fairness in elections, thereby strengthening democratic processes in Venezuela and potentially other similar contexts.

4 Appendix

4.1 Estimation of the margin of error

4.1.1 Participation rate

The participation rate in the election can be estimated as

$$r = \sum_{h=1}^{7} W_h r_h$$

where $W_h = \frac{X_h}{X}$ is the stratum-level weight, X_h is the number of registered voters in the population in stratum h, X is the total number of registered voters in the population, and

$$r_h = \frac{y_h}{x_h} = \frac{\sum_{i=1}^{n_h} y_{hi}}{\sum_{i=1}^{n_h} x_{hi}}$$

is the stratum-level participation rate estimator. Here, n_h is the number of polling stations observed in stratum h, y_{hi} and x_{hi} are, respectively, the number of voters that participated in the election and the number of registered voters in polling station i observed in stratum h.

Since the polling stations were selected using a systematic sample with a single start, we cannot use a design-based sampling variance estimator. Instead, we propose a model-based approach, in which we use a stratified random model to account for the (implicit) stratification of the polling stations by the seven strata. In this case, the sampling variance estimator for r is given by:

$$\operatorname{var}(r) = \sum_{h=1}^{7} W_h^2 \operatorname{var}(r_h)$$

Also, because polling stations are unequal in size (number of registered voters), the estimator r and its corresponding stratum estimators r_h are ratio means and, therefore, do not have a closed-form sampling variance estimator. Hence, we propose using Taylor series approximation to estimate the sampling variance of r_h (Wolter, 2007):

$$\operatorname{var}(r_h) = \frac{1}{x_h^2} \left\{ \operatorname{var}(y_h) + r_h^2 \operatorname{var}(x_h) - 2 r_h \operatorname{cov}(y_h, x_h) \right\}$$

While, within each stratum, there is possibly some additional implicit stratification due to the ordering of the polling stations by geographical location, in order to anticipate missing data in some polling stations, we estimated $\operatorname{var}(y_h)$, $\operatorname{var}(x_h)$, and $\operatorname{cov}(y_h, x_h)$ using a Simple Random Sample model (Wolter 2007; Kish 1965):

$$\operatorname{var}(y_h) = \frac{n_h}{n_h - 1} \left\{ \sum_{i=1}^{n_h} y_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} y_{hi}\right)^2}{n_h} \right\}$$

$$\text{var}(x_h) = \frac{n_h}{n_h - 1} \left\{ \sum_{i=1}^{n_h} x_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} x_{hi}\right)^2}{n_h} \right\}$$

$$\text{cov}(y_h, x_h) = \frac{n_h}{n_h - 1} \left\{ \sum_{i=1}^{n_h} y_{hi} x_{hi} - \frac{\left(\sum_{i=1}^{n_h} y_{hi}\right)\left(\sum_{i=1}^{n_h} x_{hi}\right)}{n_h} \right\}$$

where: - n_h is the number of polling stations in stratum h, - y_{hi} is the number of voters who participated in the election at polling station i in stratum h, - x_{hi} is the number of registered voters at polling station i in stratum h.

Then, the margin of error for a 95% confidence interval for the overall participation rate in the election, r, can be estimated as

$$1.96 \times \sqrt{\text{var}(r)}$$

Also, the margin of error for a 95% confidence interval for the participation rate in the election in stratum h, r_h, can be estimated as

$$1.96 \times \sqrt{\text{var}(r_h)}$$

4.1.2 Proportion of voters who voted for each candidate

The calculations for the proportion of voters who voted for each candidate (or voted null/blank) follow the same approach, except that y_{hi} and x_{hi} are, respectively, the number of voters who voted for a given candidate (or voted null/blank) and the number of voters who participated in the election at polling station i observed in stratum h.

Additionally, $W_h = \hat{W}_h = \frac{X_h p_h}{\sum_{h=1}^{7} X_h p_h}$, where p_h is the participation rate estimate in the election in stratum h, as computed above.

4.1.3 Steps to compute the margin of error

1. Compute the variance of the stratum-specific participation rate estimator $\text{var}(r_h)$ for each stratum according to the estimator r_h.

2. Compute the weight for each stratum W_h according to the estimator.

3. Compute the variance of the overall participation rate r as follows:

$$\text{var}(r) = \sum_{h=1}^{7} W_h^2 \, \text{var}(r_h)$$

4. Compute the margin of error for a 95% confidence interval for r using:

$$1.96 \times \sqrt{\text{var}(r)}$$

4.2 The Electoral Integrity Project

The Electoral Integrity Project (EIP) is an independent academic initiative that studies global electoral processes to identify their flaws. It conducts innovative, policy-relevant research, comparing elections worldwide to explore issues like security and declining trust. Founded in 2012, the EIP is led by Dr. Holly Ann Garnett and Professor Toby S. James, and affiliated with the Royal Military College of Canada and the University of East Anglia. Figure 3 shows the relationship between per capita income and the Perception of Electoral Integrity index.

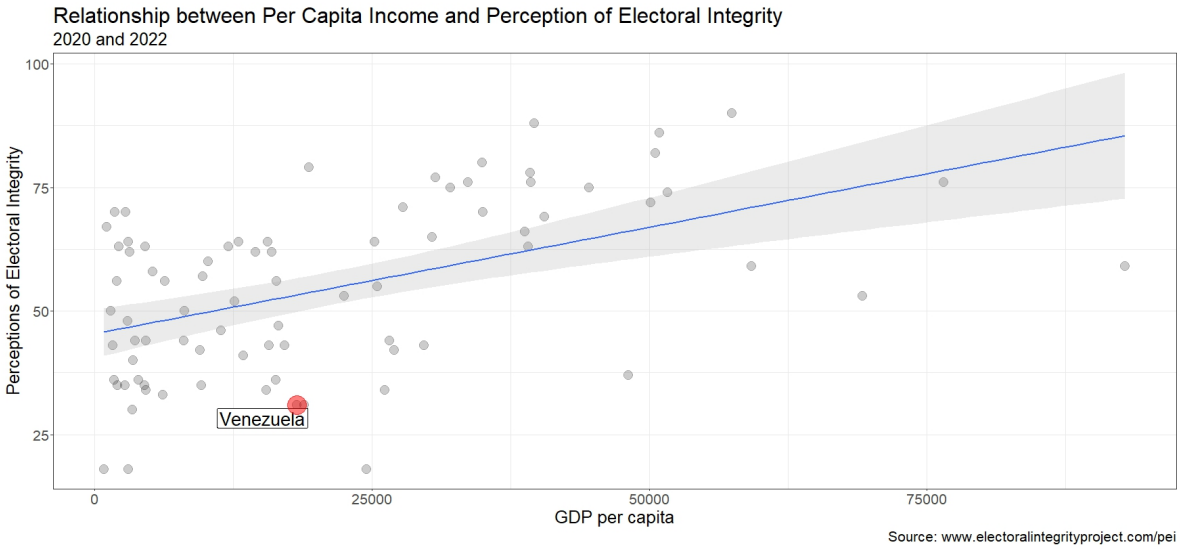

Figure 3: Correlation between income and the Perception of Electoral Integrity

The graph illustrates the relationship between GDP per capita and perceptions of electoral integrity for the years 2020 and 2022. It shows a general trend where countries with higher GDP per capita tend to have higher perceptions of electoral integrity, indicated by a positive slope in the regression line. Venezuela, however, is a noticeable outlier, marked distinctly below the regression line and confidence band. Despite having a moderate GDP per capita, Venezuela's perception of electoral integrity is significantly lower than expected, at around 31, which underscores its unique challenges in electoral fairness and transparency compared to other nations with similar economic standings.

The data presents a comparative view of perceived electoral integrity across several countries, with Venezuela highlighted for its notably low score. Considering American countries, this places Venezuela at the bottom of the list among the countries surveyed, such as Costa Rica, which leads with a score of 79, and Brazil and Peru, both scoring 64. Venezuela's position underscores significant concerns about its electoral processes compared to other nations in the Americas, reflecting ongoing challenges in ensuring fair and transparent elections.

4.3 Free and Fair Elections in Venezuela

Bishop and Hoeffler (2016) created a dataset including ten variables of election quality for all leadership elections for the period 1975–2011. These include the legal framework, the role of electoral management bodies, the protection of electoral rights, the accuracy of the voter register, and ballot access. Additionally, the fairness of the campaign process, media access, the integrity of the voting process, the conduct of officials, and the transparency in the counting of votes are also pivotal. These dimensions collectively define the robustness and credibility of an election. Figure 4 summarizes two key variables for elections in Venezuela from 1978 to 2006.

Year	Was election free?	Was election fair?
1978	No	Yes
1983	No	Yes
1988	No	No
2000	Yes	Yes
2006	No	No

Figure 4: Free and fair elections in Venezuela (1978 - 2006)

The data set summarizes the perceived freedom and fairness of elections in various years from 1978 to 2006. In 1978 and 1983, the elections were not free but were considered fair. By 1988, both freedom and fairness were absent. A positive change occurred in 2000, where the election was deemed both free and fair. However, by 2006, the situation regressed, with the election being neither free nor fair, indicating significant fluctuations in electoral integrity over these years.

4.4 Turnout in Venezuela

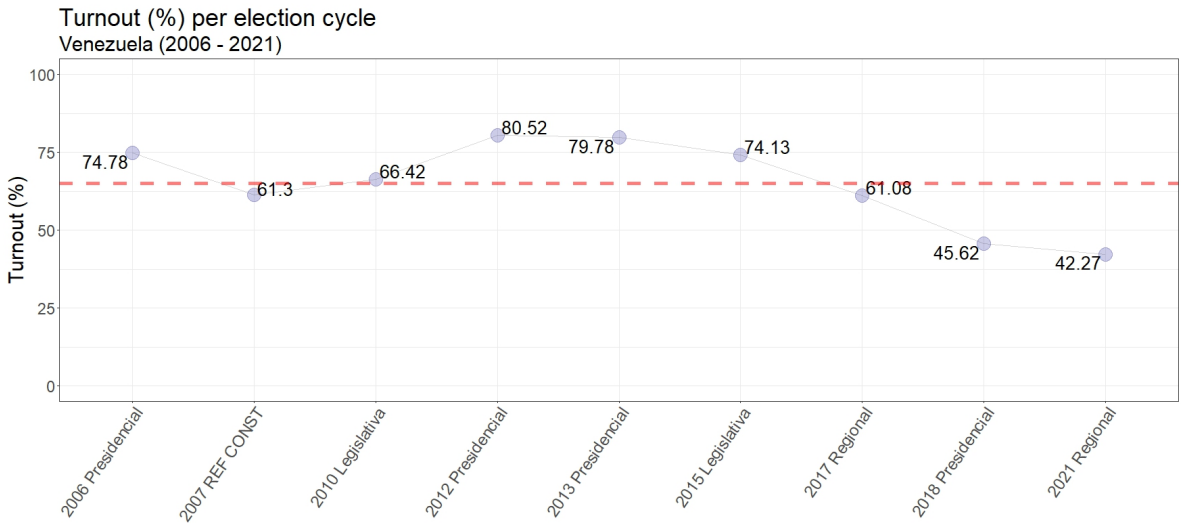

Figure 5: Turnout in Venezuela (2006-2021)

Figure 5 illustrates the voter turnout percentages for various elections in Venezuela from 2006 to 2021. It shows a fluctuating trend in voter engagement over time. Notably, the turnout peaked in 2012 for the presidential election with approximately 80.52% of registered voters participating. However, this was followed by a sharp decline in recent years, with the 2021 regional election recording only 42.27% turnout, marking the lowest in the depicted period. The dashed red line represents the average turnout across these elections, which is around 65%. The general decline in turnout in the later years could indicate growing voter apathy or potential issues with electoral integrity and public trust in the electoral process. This downward trend in participation, especially in the context of decreasing figures in the last two elections, suggests significant challenges in mobilizing the electorate in Venezuela.

4.5 Relevant literature on election fraud and sample design

- Alvarez, R. Michael, Thad E. Hall, and Susan D. Hyde. 2008. "Introduction: Studying Election Fraud." In R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde, editors, Election Fraud: Detecting and Deterring Electoral Manipulation, Washington, D.C.: The Brookings Institution Press, pages 1-17.,

- Alvarez, R. Michael and Jonathan N. Katz. 2008. "The Case of the 2002 General Election." In R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde, editors, Election Fraud: Detecting and Deterring Electoral Manipulation, Washington, D.C.: The Brookings Institution Press, pages 149-61.

- Collier, Paul and Pedro C. Vicente. 2012. "Violence, Bribery, and Fraud: The Political Economy of Elections in Sub-Saharan Africa." Public Choice, Volume 153, pages 117-47.

- Lehoucq, Fabrice. 2003. "Electoral Fraud: Causes, Types and Consequences." Annual Review of Political Science, Volume 6, pages 233-56.

- Figueiredo Filho, D., Silva, L., Carvalho, E. (2022). The forensics of fraud: Evidence from the 2018 Brazilian presidential election. Forensic Science International: Synergy, 5, 100286.

- Kish, L. (1965). Survey Sampling. New York: Wiley.

- Levin, Ines, Gabe A. Cohn, Peter C. Ordeshook, and R. Michael Alvarez. 2009. "Detecting Voter Fraud in an Electronic Voting Context: An Analysis of the Unlimited Reelection Vote in Venezuela." Proceedings of the 2009 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE '09).

- Magaloni, Beatriz. 2010. "The Game of Electoral Fraud and the Ousting of Authoritarian Rule." American Journal of Political Science, Volume 54, pages 751-65.

- Mebane, Walter R., Jr. 2008. "Election Forensics: The Second-Digit Benfords Law Test and Recent American Presidential Elections." In R. Michael Alvarez, Thad E. Hall, and Susan D. Hyde, editors, Election Fraud: Detecting and Deterring Electoral Manipulation, Washington, D.C.: The Brookings Institution Press, pages 162-81.

- Mebane, W. R. (2011). Comment on "Benford's Law and the detection of election fraud". Political Analysis, 19(3), 269-272.

- Mebane, Walter R., Jr. 2010. "Fraud in the 2009 presidential election in Iran?" Chance, Volume 23, pages 6-15.

- Mebane, W. R., Kalinin, K. (2009). Comparative election fraud detection. In APSA 2009 Toronto Meeting Paper.

- Myagkov, Mikhail, Peter C. Ordeshook, and Dimitri Shakin. 2009. The Forensics of Election Fraud: Russia and Ukraine. New York: Cambridge University Press.

- Rozenas, A. (2017). Detecting election fraud from irregularities in vote-share distributions. Political Analysis, 25(1), 41-56.

- Wand, J. N., Shotts, K. W., Sekhon, J. S., Mebane, W. R., Herron, M. C., Brady, H. E. (2001). The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. American political science review, 95(4), 793-810.

- Tucker, Joshua A. 2007. "Enough! Electoral Fraud, Collective Action Problems, and Post-Communist Colored Revolutions." Perspectives on Politics, Volume 5, pages 535-51.

- Vickery, Chad and Erica Shein. 2012. "Assessing Electoral Fraud in New Democracies: Refining the Vocabulary." Washington, D.C., International Foundation for Electoral Systems.

- Wolter, K. M. (2007). Introduction to Variance Estimation. Berlin: Springer-Verlag.