ELECTORAL RESEARCH: A DISCUSSION OF SAMPLE SCENARIOS BUILT FROM THE DICHOTOMIC AND POLITOMIC DISTRIBUTIONS

Julio Cesar Guimarães de Paula[1]

Abstract: This paper aims to discuss the sample design used in electoral research in Brazil. From the point of view of statistical theory, every sample design is constructed from a probability distribution. Thus, the estimation of categorical variables is related to two types of probability distribution: dichotomous and polytomous. The first is used in two categories (Binomial Distribution), while the second, in three or more. The work shows that the use of Binomial Distribution by the institutes reduces the sample magnitude with impacts on errors and confidence intervals. Thus, it is proposed to use the Multinomial Distribution indexed to Bonferroni's corrections for raising the quality of electoral estimates in Brazil.

Keywords: Electoral Research; Sample research design; Probability distribution; Bonferroni Correction.

Introduction

Electoral surveys can be defined as quantitative surveys in which structured questionnaires are usually applied to a sample supposedly representative of the electorate. These surveys are currently configured as the best information voters and candidates have regarding the direction of the electoral process. Thus, electoral polls in contemporary democracies serve as mechanisms for vocalizing preferences and ensuring different wills in the public space, enabling a congruent relationship between the state and citizens.

In this sense, we currently perceive the emphasis Brazilian media gives to research institutes. These institutes lead public discussions on the assertiveness of their surveys. A question arises in this scenario of doubts regarding the quality of the results published by the institutes: should research anticipate electoral results? Or should they be understood as a portrait of a specific moment in a larger process? These problematizations consider important aspects of the consolidation of electoral preferences in mass democracies, such as Brazil. Little consolidated preferences and structuring of the party system and high electoral volatility project electoral results as extremely difficult. This does not exempt research institutes from a more detailed analysis of the methodological foundations used in their surveys. Thus, the present work fits in this horizon.

In this light, this study proposes an analysis of sample errors based on dichotomous and polytomous distributions. To achieve this goal, the work beyond this introduction is divided as

[1] Doctorate student in Political Sciences at the Universidade Federal de Minas Gerais (UFMG) and researcher of the Legislative Studies Center, of the same institution (CEL-DCP-UFMG). ORCID: https://orcid.org/0000-0003-1243-225X. Email: juliogcpaula@gmail.com

follows. The first section presents a historical survey on the dissonances between polls and electoral results and discusses the consolidation of sampling techniques from the North American experience. The second section presents a few fundamental facts of electoral research in Brazil.

The following section focuses on the fundamentals of evaluating electoral polls in Brazil. The third section discusses the dichotomous and polytomous distributions of probability and the sample designs constructed from these distributions. In the following section, Bonferroni corrections are calculated for scenarios with more than two categories. Finally, we present the final considerations.

1. Notes on sampling: the dissonance between research and results

The Literary Digest magazine was founded in 1890. It has conducted electoral polls that aimed to predict the results of the US presidential elections. The magazine correctly indicated Woodrow Wilson's victory in the 1916 elections and the results in the four subsequent elections. However, in the 1936 elections, the magazine's methodology did not reach the expected results. The survey wrongly indicated the victory of Republican candidate Alfred Landom over the re-election candidate, Democrat Franklin D. Roosevelt (FERRAZ, 1996). The design proposed by the researchers was based on sending response cards to voters. These were sampled based on the phone records and car owners. According to data from that time, around ten million cards were sent, reaching a response rate of approximately 25% of the total sample. The difference between the data and the electoral result reached 19 percentage points. The error in the estimates was attributed to the inconsistency of the sampling method used. The universe of car owners was not representative of the universe of voters. Many voters hit by the 1929 crisis did not have cars and could not be sampled (ABEP, 2008).

In this sense, the mistaken estimation made for the 1936 election motivated a wide debate in American society about electoral research's methodological foundations. The scholars at the time showed that the bias of a small representative sample and the low proportional participation of the sampled population was why the result indicated by the magazine differed from the result obtained by the ballot boxes (FERRAZ, 1996; ABEP, 2008). In the same election, George Gallup and his research institute, the American Institute of Public Opinion, matched the results with a demographically representative sample of 3,000 interviews. The survey conducted by Gallup was the starting point for using scientific methods in the preparation and conduction of electoral research (ABEP, 2008).

In the 1948 elections, the scientificity of the sampling procedures was also questioned. The most prominent American research institutes did not indicate the victory of Democrat Harry Truman for the Presidency. The explanations for the error were based on two dimensions. The first regarding the volatility of voter preferences. The second, the distance between the polls and the elections, since the surveys were conducted two weeks before the election, which did not attain the changes in preferences in the final period of the electoral process (FERRAZ, 1996; ABEP, 2008). The other dimension of errors is linked to a discussion of the basics of sampling. The use of probabilistic samples and quota sampling was at the heart of the debate. The Social Science Research Council report strongly defended probabilistic sampling (FERRAZ, 1996). After that, the form of samplings became the core of discussions regarding inferential robustness in electoral polls. It is important to note that inference is how information obtained from a sample of the population of interest is generalized. Thus, the inferences' quality can be affected by two factors in this generalization process, non-sampling errors and sampling errors (ABEP, 2008).

Non-sampling errors are linked to research logistics, the inadequate definition of the population of interest, poorly designed questionnaires (questions that induce certain answers, lack of objectivity, inadequate order, inaccessible vocabulary, etc.), and poorly trained interviewers. On the other hand, sampling errors are linked to the sampling plane's construction, size, homogeneity, and stratification. In this sense, representative samples are those in which the proportion of both types of errors is minimized or quantified. In the case of sample errors, surveys conducted using probabilistic methods should always include an error in their estimates, the so-called margin of error in electoral surveys (FERRAZ, 1996; ABEP, 2008). The next section will examine these assumptions in Brazilian electoral surveys based on the understanding of the importance of probabilistic samples and estimating errors.

2. The basis of electoral research in Brazil

In the Brazilian case, the debate on the fundamentals of electoral research and its ability to predict the polls' results is recent. The creation of the Brazilian Institute of Public Opinion, IBOPE, in 1942, is the founding landmark of research and the beginning of the debate on the quality of inferences in electoral contexts (*Idem*). The years of democratic interruption cooled the debate, which was only resumed after the stabilization of the electoral calendar, after the 1982 elections for state governments. Thus, discussions regarding how the polls reflect electoral results have also begun to guide Brazilian elections.

Discussions regarding possible errors and sample inconsistencies also guided the Brazilian debate on the validity of the inferences verified by the surveys. The multi-stage procedure[2] is generally adopted in the voting intention surveys carried out in Brazil, with stratification[3] and clusters[4] in the first stages (regions, municipalities, census sectors),

[2] A sampling procedure can be performed in many stages, in which case sampling is conducted in multiple stages. The objective is to combine the different types of sampling using the advantages of each type. In a three-stage sampling, for example, the first two stages allow us to employ randomization techniques, using sampling by quotas in the last.

[3] Stratified sampling is the appropriate design when the researcher intends to study a population based on specific characteristics. The objective is to define representative groups, combining random sampling and stratification. The stratified sample was developed to increase the accuracy of the sampling process, reducing the degree of heterogeneity present in the simple random sample. Sample stratification is characterized by the smaller variation of data within each stratum than between strata.

[4] The lack of lists available for large populations is a difficulty in using simple random sampling. Cluster sampling can

incorporating quotas by sex[5], age, education level, etc. in the second stage, defined according to IBGE and the Superior Electoral Court criteria. Quotas are used to interview individuals who have a low probability of response, thereby avoiding possible bias in the sample. For example, EAP (Economically Active Population) and Non-EAP quotas are defined to impose that individuals who work and those who do not belong to the sample (ABEP, 2008).

According to King *et al.* (2001), the rates of nonresponses in electoral surveys range between 50% and 90%. According to the authors, these high rates are related to two factors: (1) the topics covered in the questionnaire, such as racism, inequality, adherence to democracy, issues that, depending on how they are presented, can increase the rate of nonresponses; and (2) where the surveys are conducted. In light of this, Smith (1983) compares nonresponse rates with official data, such as those obtained by demographic censuses. In the wake of studies that consider socioeconomic variables with variables independent of nonresponse rates, Henkel (2012) found important results when evaluating students' nonresponse rates from Pará state schools to research that sought to assess the positioning of students on aspects of the Brazilian political system. According to the author, the sociodemographic structure, the respondents' life cycle, and their experiences regarding public policies help explain nonresponse rates.

Quota sampling at its limit (MOSER *et al.*, 1953) mitigates the likelihood of nonresponses by bringing into the scenario individuals who would have difficulty being researched. The method used - of establishing quotas within the sectors probabilistically selected - can be considered approximately high-weighted, and its function is to avoid the possible distortions possibly introduced by the interviewers if there were no quotas. Although it is widely known that quota sampling does not enable calculating the sampling error (or margin of error), given that it does not meet the principles of statistical randomness, research institutes consciously adopt the multi-stage sampling model, involving quotas in the last stage, as an approximate model to ideal model (from a probabilistic perspective).

Problems such as the time used, the cost of conducting the survey, and the impositions of Electoral Law[6], which requires the declaration of the margin of error when registering the survey with the Electoral Justice, are the main arguments used by the institutes for choosing this method.

correct this problem. The researcher builds multiple selection stages where the initial stages are called clusters and the same principle of randomness applies to each one. A cluster is a unit that agglomerates individuals. When randomly drawing a cluster, the logic is the same as the random drawing of individuals.

[5] Annex (1) shows the sampling designs used by the main research institutes in the last Presidential elections.

[6] The restrictions were suspended in 1988, based on resources presented by the media and research production, and in 1990, they were removed from the legislation (Resolution 16.402/1990). The electoral legislation advanced to the field of information regulation, providing transparency both egarding the agents involved in the political process and the methodological parameters of data production. In the most recent changes occurred with the partial Political Reform conducted in 2005 and 2006, which defined new rules for the conduct of electoral campaigns and the dissemination of polls, valid from the 2008 municipal elections. This law (law 11.300/06) defined the restriction of disclosure for the period of 15 days prior to the election. However, on November 8th, 2007, Resolution 22.623 of the Superior Electoral Court established an open regulation, defining that surveys conducted before the day of elections may be released at any time, including on the day of the elections.

Thus, the dissemination of the polls is legally conditioned to the registration of the following information in the Electoral Court with a minimum period of five days before the results are known:

1) The research contractor;

2) The value and origin of the resources;

3) The methodology and the period for conducting the research;

4) The sample plan and weighting regarding the interviewee's sex, age, education and economic level, work location, interval of confidence, and error margin;

5) The internal control and verification system, data collection, and fieldwork inspection;

6) The complete questionnaire applied or to be applied;

7) The name of who paid for the work;

8) Documents proving the company's registration;

9) The name of the statistician responsible for the research and its registration with the Regional Statistics Council;

10) Registration number of the company responsible for research at the Regional Statistics Council.

The following section proposes a discussion on the criteria that underlie the evaluation of the electoral polls' results based on the understanding of how Brazilian research institutes format sample designs and how the Electoral Law regulates their realization and dissemination.

3. How to evaluate electoral research in Brazil

It is well known that electoral polls have become popular in Brazil in recent decades (MENDES, 1991; SILVA *et al.*, 2019). At the same time, there is an increase in disbelief regarding their results (ALMEIDA; 2008; BRAUN; 2009). Concepts such as sampling and inference have entered the population's collective unconscious and direct criticism of the dissonances between polls and electoral results. The book written by Frederick Mosteller (1949) entitled *The Pre-election Polls of 1948: Report to the Committee on Analysis of Pre-Election Polls and Forecasts,* is a canonical work in this area and the result of many studies on the errors of election polls in the 1948 American elections. Mosteller *et al.* (1949) presented eight forms to measure the accuracy of electoral surveys. The methods proposed by the authors can be divided into two groups: *(i) those focused on the difference between the absolute percentages of votes obtained by the candidates and those estimated by the institutes and (ii) those that address the relative distances between the candidates.*

In this sense, Gramacho (2013; 2015) works use the methodology proposed by the North American statistician to understand the Brazilian case. The author uses "Method 3", described by Mosteller *et al.* (1949), based on the Brazilian political system's idiosyncrasies, especially the multiparty system. This method, denominated MM3, consists of measuring each electoral survey's error from the difference between the average of the absolute values of the estimated intention to vote for each candidate and the percentage of valid votes obtained by the candidate (GRAMACHO, 2013). The calculation process is as follows:

*1) Percentage without decimals of the estimate of votes made by the X institute for the candidat*es - ***(Voting intention, V.I.)***;

2) Percentage without decimals of the result obtained by the candidates in the elections ***(Total Vote, T.V.)****;*

3) Extract the absolute value of the difference between ***(V.I. – T.V. = Error)****;*

4) The arithmetic mean of these differences is calculated: ***Mosteller Method 3 (MM3)****.*

For illustrative purposes, Tables 1, 2, 3, and 4 show the MM3 calculations for both main research institutes in Brazil, IBOPE and DATAFOLHA, for the 2018 elections[7]. IBOPE presents an inferior MM3 value compared to DATAFOLHA. The calculations are presented below.

Variables (*V.I. 1, 2, 3*) present the voting intentions for presidential candidates in 2018. Variable (*Average V.I.*) represents the average of the voting intentions. *T.V.* represents the votes obtained by the respective candidates. The Error is the difference between the variables (*Average V.I.*) and (*T.V.*). The MM3 calculated for the IBOPE surveys was 2.34, 0.34 above the 2 points reported by the institute. The institute's methodology is presented in Annex (1).

Table 1 – Calculation of the MM3 for the IBOPE 2018 ELECTIONS (1st Round)

Candidates	Party	V.I. 1	V.I. 2	V.I. 3	Mean V.I.	T.V.	Error
Jair Bolsonaro	PSL	38	41	45	41.3	46	-4.7
Fernando Haddad	PT	28	25	28	27.0	29	-2.0
Ciro Gomes	PDT	12	13	14	13.0	12	1.0
Geraldo Alckmin	PSDB	8	8	4	6.7	4	2.7
João Amoêdo	NOVO	4	3	3	3.3	2	1.3
Marina Silva	REDE	3	3	2	2.7	1	1.7
Henrique Meirelles	MDB	2	2	1	1.7	1	0.7
Guilherme Boulos	PSOL	2	2	1	1.7	0	1.7
Cabo Daciolo	PATRIOTA	2	2	1	1.7	1	0.7
Alvaro Dias	PODEMOS	1	1	1	1.0	0	1.0
João Goulart Filho	PPL	0	0	0	0.0	0	0.0
Vera Lúcia	PSTU	0	0	0	0.0	0	0.0
Eymael	DC	0	0	0	0.0	0	0.0
Mosteller Method 3 (MM3): 2.34							

Source: The author, with data obtained from the "Poder 360°" website.

[7] The 2018 elections had special characteristics due to several factors. The candidacy of the candidate for the Labor Party (PT), former president Luís Inácio Lula da Silva, was not granted by the electoral justice due to the Clean Record Law. The non-approval of the ex-president's candidacy put his candidate for vice-president in the dispute, former Minister of Education and former Mayor of São Paulo, Fernando Haddad. This situation meant that the name of the candidate of the Labor Party (PT), Fernando Haddad, only appeared in the electoral polls as of the 09/24/2018.

The data presented in Table 2 follow the same logic as the previous Table. Variables (*V.I. 1, 2, 3*) present the voting intentions for presidential candidates in 2018. Variable (*Average V.I.*) represents the average of the voting intentions. T.V. represents the votes obtained by the respective candidates. The Error is the difference between the variables (*Average V.I.*) and (*T.V.*). The MM3 calculated for DATAFOLHA surveys was 4.2, 2.2 above the 2 points reported by the institute. The institute's methodology is presented in Annex (1).

Table 2 – Calculation of the MM3 for the DATAFOLHA 2018 ELECTIONS (1st Round)

Candidates	Party	V.I. 1	V.I. 2	V.I. 3	Mean V.I.	T.V.	Error
Jair Bolsonaro	PSL	33	39	40	37	46	-9
Fernando Haddad	PT	21	26	25	24	29	-5
Ciro Gomes	PDT	11	13	15	13	12	1
Geraldo Alckmin	PSDB	9	9	8	9	4	5
João Amoêdo	Novo	8	4	3	5	2	3
Marina Silva	Rede	5	3	3	4	1	3
Henrique Meirelles	MDB	4	2	2	3	1	2
Guilherme Boulos	Podemos	3	2	2	2	0	2
Cabo Daciolo	Psol	2	1	1	1	1	0
Alvaro Dias	Patriota	2	1	1	1	0	1
João Goulart Filho	PSTU	2	0	0	1	0	1
Vera Lúcia	DC	0	0	0	0	0	0
Eymael	PPL	0	0	0	0	0	0
Mosteller Method 3 (MM3): 4.2							

Source: The author, with data obtained from the "Poder 360º" website.

The 2nd Round of the 2018 elections took place between 10/07 and 10/28. IBOPE surveys were conducted on October 15th, 23rd, and 27th. The error estimated by the institute's methodology was two percentage points, and the MM3 was below 1.4. This indicates that the 2nd round estimates are more assertive than those of the 1st Round.

Table 3 – Calculation of the MM3 for the IBOPE 2018 ELECTIONS (2nd Round)

Candidates	Party	V.I. 1	V.I. 2	V.I. 3	V.I. 4	Mean V.I.	T.V.	Error
Jair Bolsonaro	PSL	59	57	56	54	56.5	55.13	1.4
Fernando Haddad	PT	41	43	44	46	43.5	44.87	-1.4
Mosteller Method 3 (MM3): 1.4								

Source: The author, with data obtained from the "Poder 360º" website.

The DATAFOLHA surveys were conducted on October 10th, 18th, 25th, and 27th. The error estimated by the institute's methodology was two percentage points, and the MM3 was at the 1.9 thresholds. DATAFOLHA data were also more assertive in the 2nd Round. Statistical explanations for such assertiveness will be presented in the next section.

Table 4 – Calculation of the MM3 for the DATAFOLHA 2018 ELECTIONS (2nd Round)

Candidates	Party	V.I. 1	V.I. 2	V.I. 3	V.I. 4	Mean V.I.	T.V.	Error
Jair Bolsonaro	PSL	58	59	56	55	57	55.13	1.9
Fernando Haddad	PT	42	41	44	45	43	44.87	-1.9
Mosteller Method 3 (MM3): 1.9								

Source: The author, with data obtained from the "Poder 360º" website.

According to Gramacho (2013), considering multiparty contexts, as is the case in Brazil, MM3 has an important limitation, indicating a set of results of each electoral survey and not specifically for the individual candidates. In response to this limitation, the author developed the MM3C, which is the *Error Estimation Method for each Candidate*, using the same calculation process as MM3 up to the third step of the four mentioned above.

Using both methods, MM3 and MMEC, Gramacho (2013; 2015) developed two works to discuss the inconsistencies between the electoral polls and the results of the ballot boxes in the majority of 2010 and 2014. The results reveal errors above the margins reported to the Electoral Justice. The largest discrepancies were found in *(i) surveys conducted with greater advance, (ii) conducted in the 1st Round, (iii) conducted in uncompetitive disputes, (iv) conducted when the number of candidates is reduced, and (v) in governor elections*. However, an assessment of the quality of electoral polls based exclusively on the correct prediction of the electoral result is at least hasty.

Therefore, the construction of predictive models for electoral results should not only consider voting intentions. Thus, some dimensions can be used to build more assertive models, such as *(i) Previous electoral research, (ii) History of white and null votes; (iii) Voting history consolidated in specific regions of the country; (iv) Party and candidate rejection patterns; (v) Party indices and consolidation of the party system*. Thus, evaluating the degree of assertiveness of election polls would neglect a range of factors that interfere with voting intentions.

In light of this, Gramacho (2013) estimates a linear regression model in which the dependent variables are the MM3 values calculated for the 153 electoral polls analyzed by the author. Of the independent variables mobilized, the highest scores estimated in the four models presented are related to the variable 2nd Round, which refers to the Round in which the election is held. Many factors can explain this result, including the adequacy of the probability distribution to the sample design.

The sample design guided the discussions on the research quality in Brazil and the United States. However, in addition to a discussion of sample characteristics, we propose a debate on sample error, popularly known as a *margin of error*. Consider the Brazilian majority elections that take place in two rounds. The first Round presents more than two candidates invariably. The variable of interest is by polytomous definition, that is, it has three or more possibilities. The sample calculation used by research institutes considers a dichotomous distribution. This choice has important impacts on the size of the samples and error estimation, which compromise the inferences, not necessarily predicting the results. Thus, the next section will be devoted to both distributions' formal foundations and their impact on the sample size.

4. Dichotomous and polytomous distributions

The estimation of proportions takes place in two types of questions: dichotomous and polytomous. Dichotomous questions are those with two items, and polytomous questions have more than two items. The statistical theory advanced concerning dichotomous questions. However, the questions are polytomous in electoral scenarios, that is, they present more than two possibilities to the respondents. The possibilities of abstaining increase the number of categories, even in a second-round scenario where there are only two candidates, (SILVA, 2012; ASSUMÇÃO, 2017).

4.1. Dichotonous questions

The sampling of proportions in dichotomous questions is based on the binomial distribution, which refers to a random experiment that consists of repeated attempts that present only two possible results (Bernoulli's attempts) and has the following characteristics (AGRESTI *et al.*, 2012):

1) Attempts are independent, that is, the result of one does not change the result of the other;
2) Each experiment replicate admits only two results: success or failure;
3) The probability of success (p) in each attempt is constant;
4) The odds for both categories are the same for each observation;
5) The probabilities are represented by π for category 1 and (1 - π) for category 2.

The random variable **X** has $\boldsymbol{n}$ parameters and $\boldsymbol{\theta} \boldsymbol{\epsilon}$ **[0,1] se X (ω) ϵ {0,1,....,n}** in the Binomial Distribution , with

$$\mathbb{P}\,(\mathbf{X} = \boldsymbol{K}) = \frac{\boldsymbol{n}!}{\boldsymbol{K}!(\boldsymbol{n}-\boldsymbol{k})!}\,\boldsymbol{\theta}^{\boldsymbol{k}}(\mathbf{1}-\ \boldsymbol{\theta})^{\boldsymbol{n}-\boldsymbol{k}}. \tag{4.1}$$

The expectation and variance are given by[8]:

[8] The average of a discrete random variable is the weighted average of the possible values of X, where the weights are probabilities. Likewise, the variance uses f(x), with a weight to multiply each square deviation $(x - \mu^2)$. (ASSUMÇÃO, 2017).

$$\boldsymbol{\mu = E(X) = np} \tag{4.2}$$

$$\boldsymbol{\sigma^2 = V(x) = np(1 - p)} \tag{4.3}$$

The calculation of the size of sample ***n,*** for dichotomous cases, is modeled by the following equation:

$$\boldsymbol{n} = \frac{\boldsymbol{N\ .\ z^2 .p.q}}{\boldsymbol{(N-1)\ .e^2 + z^2 .p.q}} \tag{4.4}$$

4.2. Polytomous questions

In polytomous questions, the distribution that underlies the sampling process is multinomial, which is a generalization of the binomial distribution to more than two proportions. Situations that can be modeled by the probability above represent multiple-choice questions, whether single or multiple answers, Likert scale, numerical scale, etc. (SILVA, 2012; ASSUMÇÃO, 2017). Therefore, with the probabilities **$\boldsymbol{\theta}$1,..., $\boldsymbol{\theta}$2, satisfying 0 ≤ $\boldsymbol{\theta}$i ≤ 1, for i=1,...,n,** e $\sum_{\mathbf{i=1}}^{\mathbf{k}} \boldsymbol{\theta}_{\mathbf{1=1}}$, then, the joint probability of obtaining the quantities (**$\mathbf{n}_1$, ..., $\mathbf{n}_k$**), from a sample of size m, is given by:

$$\mathbb{P}\left(\boldsymbol{N} = \left(\boldsymbol{n_{1,} n_2, \ldots . n_k}\right)\right) = \frac{\boldsymbol{m!}}{\boldsymbol{x_1!x_2 \ldots . x_k!}} \boldsymbol{\theta_1^{n1}\ \theta_2^{n2} \ldots .\ \theta_3^{n3}} \tag{4.5}$$

Thus, the equations of the sampling parameters (expectation and variance) used to sample dichotomous questions are valid for sampling polytomous questions. Due to the equivalence shown above, the expectation of ***nᵢ*** is **m·$\boldsymbol{\theta_1}$** and its variance is **m·$\boldsymbol{\theta_i}$·(1-$\boldsymbol{\theta_i}$)**, which are equivalent to the binomial case.

Since the re-democratization, the voter has been subjected to electoral scenarios with more than two candidates. This situation must be modeled by a multinomial distribution and would have a sample size equal to:

$$\boldsymbol{n = \underset{p.q}{max.}} \frac{\boldsymbol{N\ .\ z^2 .p.q}}{\boldsymbol{(N-1)\ .e^2 + z^2 .p.q}} \tag{4.6}$$

Where **N** is the population size, **q** is equal to **(1-p)**, **e** is the margin of error, and **z** is the standardized normal distribution factor corresponding to the level of significance α. Product **p.q** is normally obtained from the history of previous works or, when entirely unknown, replaced by 0.25, the maximum value that will provide a conservative calculation of the sample size.

5. Probability distributions and Bonferroni corrections

In a discussion analogous to what is intended in this work, Silva (2012: 125, our translation) "*indicated that the estimation of intervals of confidence for the* $\boldsymbol{k}$ *classes must consider that the precision estimates are given simultaneously for the* $\boldsymbol{k}$ *classes. It would mean distributing the global significance level* $\boldsymbol{\alpha}$ *over the* $\boldsymbol{k}$ *estimation intervals*". The author uses the Bonferroni method to correct the level of significance. Generally speaking, the method equally distributes the level of global significance among the **k** interest categories. For example, in the last presidential elections, we had 13 candidates. If we consider white votes, undecided nulls, and abstentions, we will have at least 15 classes to be tested with a global significance level of 0.95 **(α = 0.05)**. The level of significance run by Bonferroni for each class will be $\boldsymbol{\alpha_{15}}$ **= α/k** = 0.05/15 = 0.0033. The ratio between the **α** value (for a global significance level of 0.95) and the number of **k** decreases its $\boldsymbol{\alpha_k}$ value as categories are introduced in the tests. Thus, Bonferroni's corrections cause three immediate impacts: *(i) decrease in the rejection area* ***(z)****; (ii) increase in the interval of confidence, and (iii) the consequent increase in the sample* ***(n)***.

The examples below present a hypothetical situation for the sample calculation, using Bonferroni's corrections for categorical data with **k=15** as a reference. The universe is the number of voters in the municipality of Belo Horizonte, the confidence level is the most used by research institutes, 95%, and the margin of error is 2%. The calculations consider formulas (3.4) and (3.6) as references.

Example 1:

The value **z** to be inserted in example 1 corresponds to the ratio between the level of global significance (0.05) and the 15 **k** categories in question **(0.05/15 = 0.003334)**. This value will represent a **z = 2.712986**. The value **p.q = 0.25** corresponds to the most pessimistic scenario of the sample construction.

$$\boldsymbol{n} = \frac{\mathbf{1{,}956{,}410\ .\ 2.712986^2\ .\ 0.25}}{\mathbf{(1{,}956{,}410 - 1)\ .\ 0.02^2 + 2.712986^2\ .\ 0.25}} = \mathbf{4589.39} \cong \mathbf{4589}$$

Example 2:

The equation below presents the same sample calculation for two categories. The **z** value to be inserted in equation 2 corresponds to the ratio between the global significance level (0.05) and the 2 **k** categories in question **(0.05/2 = 0.025)**. This value will represent a **z = 1.959964**.

$$\boldsymbol{n} = \frac{\mathbf{1,956,410\ .\ 1.959964^2\ .\ 0.25}}{\mathbf{(1,956,410-1)\ .\ 0.02^2 + 1.959964^2\ .\ 0.25}} = \mathbf{2397.97} \cong \mathbf{2398}$$

Example 1 presented the correction of global significance (0.05) by the 15 categorical introductions in the equation. The corrections raised the sample magnitude to a value almost twice as high as the calculation without corrections exemplified in example 2. It is worth noting that the value (n) in the second example is very close to the sample calculations conducted by the Brazilian research institutes that revolve around 2000 respondents.

Table 5 illustrates the calculations above and shows the values of Bonferroni's corrections for **α** for the **k** categories. Thus, the correction values for the **k** categories are more conservative than the uncorrected estimates, which mitigates the possibilities of type I error.

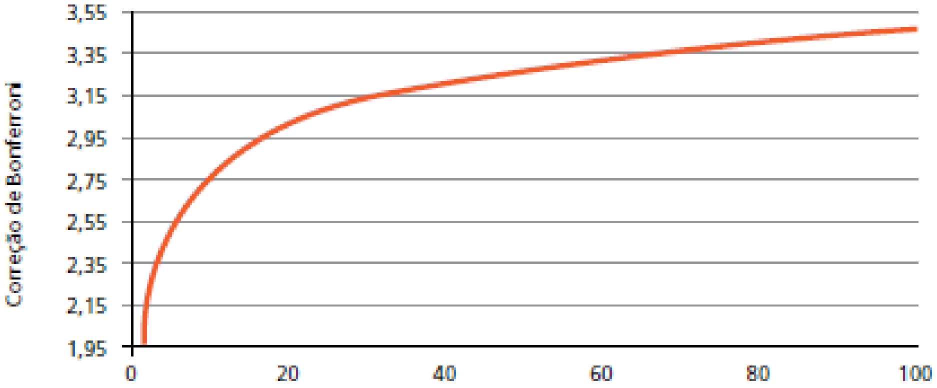
Table 5 – α Bonferroni correction

k	α = 5%	α = 10%
2	0.05000	0.10000
3	0.02500	0.05000
4	0.01667	0.03333
5	0.01250	0.02500
6	0.01000	0.02000
7	0.00833	0.01667
8	0.00714	0.01429
9	0.00625	0.01250
10	0.00556	0.01111

Source: Silva (2012).

The **z** value of the normal distribution, despite increasing significantly, does not explode with the larger quantity of items **k**, therefore, not compromising the sample size. In Graph 1, for **α** = 5%, the values of **z** as a function of **k** were plotted using Bonferroni's formulation, which is the most conservative option resulting in the highest **z**. Considering k = 2 as a comparison parameter, where, conventionally, z = 1.9599, we can verify how reduced the advance of z is with the increase of k, with z still just 3.4780, when k reaches 100. In this case, the increase in the sample size, considering the population to be large enough, would be 216% (= (3.4780/1.9566) 2-1), even considering an increase from k = 2 to k = 100 (SILVA, 2012).

Graph 1 – Variation of the Bonferroni correction for the number of k items, for α = 0.05



Source: Silva (2012)

Conclusions

Electoral polls are a part of the democratic routine of modern societies. With the dissemination of information on the direction of electoral processes, they reduce informational asymmetries between individuals. Thus, the figures presented are discussed and questioned not only by specialists but also by society. The side effect of this popularization is the increase in questions regarding the assertiveness of pre-election surveys. The discussion is linked to the role of research in predicting the outcome of the ballot boxes. An important assumption of the present work is to consider the electoral polls as staked surveys, given that their results reflect specific circumstances and should not be measured only by the assertiveness or not of the electoral results.

This work does not propose a discussion of electoral research based on the results of the ballot boxes, but a discussion of the sampling logic employed by the institutes. The calculations performed in Examples 1 and 2 show a clear sample underrepresentation in the design made from a binomial distribution. This under-representation may not impact scenarios where the consolidation of votes and institutionalization of the party system have more robust rates. However, in Brazil, where preferences are increasingly volatile, and the party system is undergoing high deterioration, underestimated samples can provide us with a blurred description of the electoral scenarios.

We are aware of the financial and logistical costs of increasing the sample size for the institutes. However, Brazilian democracy poses new questions and new challenges. Recent experiences have shown that the democratic process is increasingly complex, and electoral research must accompany this path, which seems increasingly inexorable.

References

ASSOCIAÇÃO BRASILEIRA DE EMPRESAS DE PESQUISA. **Publicação de Pesquisas Eleitorais**. São Paulo: ABEP, 2008.

AGRESTI, Alan; FINLAY, Barbara. **Métodos estatísticos para as ciências sociais**. 4. ed. Porto Alegre: Penso, 2012.

ALMEIDA, Alberto Carlos de. **A cabeça do eleitor**: estratégia de campanha, pesquisa e vitória eleitoral. São Paulo: Record, 2008.

ASSUMÇÃO, R. **Fundamentos Estatísticos da Ciência de Dados**: voltado para aplicações. BOOKWEBSITE.COM, 2017.

BRAUN, Cecilia; Maíra STRAW, C. (Org.). **Opinion Pública**: una mirada desde América Latina. Buenos Aires: Planeta, 2009.

FERRAZ, Cristiano. **Crítica Metodológica às Pesquisa Eleitorais no Brasil,** Dissertação de Mestrado, UNICAMP, 1996.

GRAMACHO, Wladimir. G. À margem das margens? A precisão das pesquisas pré-eleitorais brasileiras em 2010. **Opinião Pública**, v. 19, n. 1, p. 65-80, 2013.

GRAMACHO, Wladimir G. A pesquisa governamental de opinião pública: razões, limites e a experiência recente no Brasil. **Revista do Serviço Público**, v. 65, n.1, p. 49-64, 2014.

HENKEL, Karl. Análise da não resposta em *surveys* políticos. **Opinião Pública**, Campinas, v. 18, n. 1, p. 216-238, jun., 2012.

KING, Gary, et. al. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. **American Political Science Review**, v. 95, n. 1, p.49-69, 2001.

MENDES, Antonio Manuel Teixeira. O papel das pesquisas eleitorais. **Novos Estudos Cebrap**, São Paulo, v. 1, n. 29, p. 28-33, mar. 1991

MOSTELLER, Frederick. Measuring the error. In: MOSTELLER, Frederick, et. al. **The pre-election polls of 1948**. Report to the committee on analysis and pre-election polls and forecast. New York: Social Science Research Council, 1049, p. 54-80.

Moser, Claus, et. al. An experimental study of quota sampling. **Journal of the Royal Statistical Society**: Series A 116, p.349–405, 1953.

SILVA, Ângelo Henrique Lopes da. Estimativa de proporções em questões politômicas. **Revista do TCU**, n.125, p.18-27, 2012.

SILVA, Bruno Fernando da; GONCALVES, Ricardo Dantas. Pesquisas eleitorais afetam receitas de campanha: a correlação entre expectativa de vitória e financiamento de campanha em disputas ao Senado. **Revista de Sociologia e Política**, Curitiba, v. 27, n. 71, p.2-17, 2019.

SMITH, T. W. The hidden 25 percent: an analysis of nonresponse on the 1980 General Social Survey. **Public Opinion Quarterly**, v. 47, n. 3, p. 386-404, 1983.

Annex (1)

IBOPE INTELIGENCIA PESQUISA E CONSULTORIA LTDA

Survey methodology

Quantitative research, which consists of conducting personal interviews, with the application of a structured questionnaire with a representative sample of the studied electorate. Survey conducted in the state of Rio de Janeiro.

Sampling plan

Sample plan and weighting regarding sex, age, education and economic level of the respondent; interval of confidence and margin of error:

Representative of voters in the studied area prepared in three stages. In the first stage, a probabilistic design of the surveyed municipalities was established using the PPT method (Probability Proportional to Size), based on the population of voters (TSE 2018, excluding abstention of the 1st rounds of 2010 and 2014) of each municipality. In a second stage, within the selected municipalities, the polling locations were selected using the PPT method, based on the number of voters from each location. In the third stage, within the chosen voting locations, the respondents were selected through sample quotas, proportional to the significant variables, namely: Sex and Age, according to the profile of the voters. The survey is self-weighted due to the adopted sampling methodology. In other words, the proportions of the universe surveyed are provided for in the sample, with no need for any weighting regarding gender, age, and education and economic level. The estimated interval of confidence is 99% and the maximum estimated margin of error, considering a simple random sampling model, is 03 (three) percentage points higher or lower than the results found in the total sample.

Internal control and verification system

The research was conducted by a properly trained team of interviewers and supervisors hired by IBOPE INTELIGÊNCIA PESQUISA E CONSULTORIA LTDA. After the fieldwork, the questionnaires are subjected to an inspection of approximately 20% (twenty percent) of the questionnaires applied by the interviewers to verify the responses and the adequacy of the respondents to the sample parameters.

DATA FOLHA: INSTITUTO DE PESQUISAS LTDA

Survey methodology

Quantitative research, by sampling, with the application of a structured questionnaire and a personal approach at population flow points. The survey universe was represented by the Brazilian electorate as a whole aged 16 or over.

Sampling plan

Sampling plan and weighting regarding sex, age, education and economic level of the respondent; interval of confidence and margin of error:

Universe: Brazilian electorate, aged 16 or over. Sample size: The expected sample is 18,060 interviews. Sampling technique: The sample is stratified by geographic region and nature of the municipalities (capital, metropolitan region, or countryside). At a first stage, the municipalities that will be part of the survey are drawn in each stratum. At a second stage, the neighborhoods and approach points where the interviews will be applied are drawn. Finally, the respondents are selected at random to answer the questionnaire, according to sex quotas and age group. The strata sizes were disproportionate in this sample to allow details of the following units of the federation and its capitals: SP, RJ, MG, in addition to the Federal District (DF). The correct proportions will be restored through weighting in the final results. The data used to define and select the sample are based on data provided by the TSE – Superior Electoral Court (electorate of August 2018) and IBGE (2018 estimate). The data regarding sex and age group are: Male: 47%, female: 53%, 16 to 24 years of age 15%, 25 to 34 years of age 21%, 35 to 44 years of age 21%, 45 to 59 years of age 24%, and 60 years of age or more 19%. Weighting of results: During the data processing, weighting was conducted regarding the proportion of each municipality in the sample for the correct representation of the geographic regions. The possible weighting for correction in the sizes of the segments is foreseen considering the variables sex and age group. For the variables education and economic level (monthly family income), the factor foreseen for weighting is 1 (results obtained in the field). Physical area: Interviews were conducted in 341 municipalities located in the following states: Acre, Alagoas, Amazonas, Amapá, Bahia, Ceará, Federal District, Espírito Santo, Goiás, Maranhão, Minas Gerais, Mato Grosso, Mato Grosso do Sul, Pará, Paraíba, Paraná, Pernambuco, Piauí, Rio de Janeiro, Rio Grande do Norte, Rondônia, Roraima, Rio Grande do Sul, Santa Catarina, São Paulo, Sergipe, and Tocantins. The complete list of municipalities and neighborhoods surveyed will be forwarded to this court until the seventh day following the date of registration of the survey, according to Resolution 23.549/2017 of the TSE, in art.2° paragraph 6. Margin of Error: The maximum expected margin of error is 2 percentage points higher or lower, considering a 95% level of confidence. The intervals of confidence are calculated considering the results obtained for a 95% level of confidence.

Internal control and verification system

The researchers involved this research are trained by the Institute and receive specific instructions for each project carried out. The collection is done using a tablet and electronic questionnaire. At least 30% of the questionnaires of each researcher are verified, either on the spot by field supervisors or, later, by telephone. Internally, all material is verified and coded. A data consistency process is conducted before final processing and issuing the results.

Annex (2)

- *Sampling based on a Binnomial Distribution*

$$\boldsymbol{n = \frac{Z^2_{1-\alpha/k}}{4\ x\ \varepsilon^2}}$$

- *Sampling based on a Multinomial Distribution*

$$\boldsymbol{n = \frac{Z^2_{1-\alpha/2k}}{4\ x\ \varepsilon^2}}$$

Where:

- $\boldsymbol{Z^2_{1-\alpha/2}}$ = Tabulated value of the standard normal curve;

- n = Sample size;

- ε = Maximum error admitted;

- 1 – α = Level of confidence;

- K = Number of categories.