Springer Proceedings in Mathematics & Statistics

Isadora Antoniano-Villalobos ·
Ramsés H. Mena ·
Manuel Mendoza · Lizbeth Naranjo ·
Luis E. Nieto-Barajas *Editors*

Selected Contributions on Statistics and Data Science in Latin America

33 FNE and 13 CLATSE, 2018, Guadalajara, Mexico, October 1–5

Springer

Springer Proceedings in Mathematics & Statistics

Volume 301

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Isadora Antoniano-Villalobos ·
Ramsés H. Mena · Manuel Mendoza ·
Lizbeth Naranjo · Luis E. Nieto-Barajas
Editors

Selected Contributions on Statistics and Data Science in Latin America

33 FNE and 13 CLATSE, 2018, Guadalajara, Mexico, October 1–5

Springer

Editors
Isadora Antoniano-Villalobos
Department of Environmental Sciences, Informatics and Statistics
Ca' Foscari University of Venice
Venice, Italy

Manuel Mendoza
Department of Statistics
Instituto Tecnológico Autónomo de México
Mexico City, Mexico

Luis E. Nieto-Barajas
Department of Statistics
Instituto Tecnológico Autónomo de México
Mexico City, Mexico

Ramsés H. Mena
Department of Probability and Statistics
IIMAS, UNAM
Mexico City, Mexico

Lizbeth Naranjo
Department of Mathematics
Facultad de Ciencias, UNAM
Mexico City, Mexico

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-31550-4 ISBN 978-3-030-31551-1 (eBook)
https://doi.org/10.1007/978-3-030-31551-1

Mathematics Subject Classification (2010): 34K12, 34K29, 60J22, 60J25, 60J75, 62D05, 62F15, 62G99, 62H20, 62H25, 62H30, 62H99, 62J12, 62J99, 62M05, 62M09, 62P10, 65C40, 90C90, 92D30

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume includes a selection of peer-reviewed contributions presented at the 33FNE and 13CLATSE meetings held jointly in Guadalajara, Mexico from October 1st to 5th, 2018.

The FNE (Foro Nacional de Estadística) is the official meeting of the Mexican Statistical Association (AME), taking place annually since 1986. The purpose of the FNE is to offer an opportunity for statisticians and practitioners to share the latest developments in research and applications, exchange ideas, and explore opportunities for collaboration. The meetings are complemented by short courses and workshops, to fulfill the mission of AME of promoting the knowledge and good practice of Statistics in the country. The CLATSE (Congreso Latino Americano de Sociedades de Estadística) is the joint statistical meeting of Latin America. Born as a collaboration between the Argentinean (SAE) and Chilean (SOCHE) Statistical Societies, it has grown significantly since its first edition in 1991 in Valparaíso, Chile. It now includes the Statistical Societies of Colombia (SCE), Ecuador (SEE), Perú (SOPEST), and Uruguay (SUE), as well as AME and the Brazilian Statistical Association (ABE). The 33FNE was organized by AME and the University of Guadalajara (UDG), while responsibility for the 13CLATSE was shared by SAE, ABE, SOCHE, SCE, and AME. The joint event was hosted by the UDG University Centre for Exact Sciences and Ingeneering (CUCEI).

Statistical research in Latin America is prolific, and collaborative networks span within and outside the region. A great territorial extension, climatic peculiarities, and political and socioeconomic factors may hinder the international dissemination of the high-quality research output of the region. Additionally, much of the work is typically carried out and published in Spanish, and thus a large portion of the interested public may overlook interesting findings. We hope that this volume will provide access to selected works from Latin American statisticians and their research networks to a wider audience. We are sure that new methodological advances, motivated in part by the challenges of a data-driven world and the Latin American context, will be of interest to academics and practitioners around the world.

The scientific program of the 33FNE and 13CLATSE included a total of 107 oral presentations, organized in 43 contributed and 7 invited sessions, in Spanish and English, plus 4 keynote sessions delivered by Alexandra M. Schmidt (McGill University, Canada), Haavard Rue (KAUST, Saudi Arabia), Abel Rodriguez (University of California Santa Cruz, USA), and Francisco Louzada (University of São Paulo, Brazil). Five short courses on Spacial Statistics (Ronny Vallejos, Federico Santa María Technical University, Chile), Computational methods for Bayesian inference (Hedibert Lopes, Insper, Brazil), Environmental Statistics (Bruno Sansó, University of Santa Cruz, USA), The challenges of Teaching Statistics: New Scenarios at the Undergraduate, Masters and Doctorate levels (María Purifcación Galindo, University of Salamanca, Spain), Statistical Foundations of Machine Learning with STATA (Miguel Ángel Cruz, MultiON Consulting, Mexico), and 45 poster presentations completed the program. The event was preceded by a full day of courses, aimed mainly at interested students, on the topics of Variational Bayes and beyond: Bayesian inference for big data (Tamara Broderick, MIT, USA), Machine Learning (Elmer Garduño, Google Inc., USA), Bayesian computing with INLA (Haavard Rue, KAUST, Saudi Arabia), and Statistical and psychometric intricacies of educational survey assessments (Andreas Oranje, Educational Testing Service, USA).

We thank all participants who brought scientific quality to the events and made the experience rewarding. A special recognition is due to the local organizers Humberto Gutiérrez Pulido and Abelardo Montesinos López (UDG, Mexico), and to Leticia Ramírez Ramírez (CIMAT, Mexico) of the organizing committee. The international quality of the event would not have been achieved without the hard work of the members of the Scientific Committees: Eduardo Gutiérrez Peña (UNAM, Mexico), Abelardo Montesinos López (UDG, Mexico), Lizbeth Naranjo Albarrán (UNAM, Mexico), and Luis Enrique Nieto Barajas (ITAM, Mexico), for the 33FNE; Jorge Luis Bazán (University of São Paulo, Brazil), Ramón Giraldo (National University of Colombia), Manuel Mendoza (ITAM, Mexico), Orietta Nicolis (University of Valparaiso, Chile), and Lila Ricci (National University of Mar del Plata, Argentina), for the 13CLATSE.

Venice, Italy — Isadora Antoniano-Villalobos
Mexico City, Mexico — Ramsés H. Mena
Mexico City, Mexico — Manuel Mendoza
Mexico City, Mexico — Lizbeth Naranjo
Mexico City, Mexico — Luis E. Nieto-Barajas

Contents

A Heavy-Tailed Multilevel Mixture Model for the Quick Count in the Mexican Elections of 2018

Michelle Anzarut, Luis Felipe González and María Teresa Ortiz

Abstract Quick counts based on probabilistic samples are powerful methods for monitoring election processes. However, the complete designed samples are rarely collected to publish the results in a timely manner. Hence, the results are announced using partial samples, which have biases associated to the arrival pattern of the information. In this paper, we present a Bayesian hierarchical model to produce estimates for the Mexican gubernatorial elections. The model considers the poll stations poststratified by demographic, geographic, and other covariates. As a result, it provides a principled means of controlling for biases associated to such covariates. We compare methods through simulation exercises and apply our proposal in the July 2018 elections for governor in certain states. Our studies find the proposal to be more robust than the classical ratio estimator and other estimators that have been used for this purpose.

Keywords Bayesian calibration · Hierarchical model · Model-based inference · Multilevel regression · Poststratification · Zero-inflated model

1 Introduction

In this paper, we present one of the statistical models used in the quick count of the 2018 Mexican elections. Mexico is a Federal State that comprises 32 states. The government system is presidential; the president and the governor of each state are elected for a 6-year term by the population. The candidate who wins a plurality of

M. Anzarut (✉) · L. F. González · M. T. Ortiz
Instituto Tecnológico Autónomo de México, Río Hondo 1, Altavista 01080, CDMX, Mexico
e-mail: michelle@sigma.iimas.unam.mx

L. F. González
e-mail: luis.gonzalez@itam.mx

M. T. Ortiz
e-mail: maria.ortiz@itam.mx

I. Antoniano-Villalobos et al. (eds.), *Selected Contributions on Statistics and Data Science in Latin America*, Springer Proceedings in Mathematics & Statistics 301,
https://doi.org/10.1007/978-3-030-31551-1_1

votes is elected and no president nor governor may be reelected. Each state has its own electoral calendar, and in some cases, the federal and state elections coincide.

The National Electoral Institute (INE) is a public, autonomous agency with the authority for organizing elections. The INE organizes a quick count the same night of the election. The quick count consists of selecting a random sample of the polling stations and estimating the percentage of votes in favor of each candidate. With highly competed electoral processes, the rapidity and precision of the quick count results have become very important. Even more, the election official results are presented to the population a week after the election day. Therefore, the quick count prevents unjustified victory claims during that period.

The election of 2018 was qualified as the largest election that has taken place in Mexico, with 3,400 positions in dispute. For the first time, quick counts were made for nine local elections for the governor position, simultaneous to a quick count for the presidential federal election. The INE creates a committee of specialists in charge of the quick count, whose responsibilities encompass, mostly, the sample design, and the operation of statistical methods to produce the inferences. The inferences are presented as probability intervals with an associated probability of at least 0.95.

The information system starts at 6 p.m. and, every 5 min, collects all the sample information sent. Thus, the system produces a sequence of accumulative files used to determine the available percentage of the sample and its distribution over the country. The partial samples are analyzed with the estimation methods to track the trend of the results. Notice that the partial samples have a potential bias associated to the arrival pattern of the information. Generally, the quick count results that are made public use one of these partial samples, since the complete sample takes too long to arrive. The committee reports a result when certain conditions are met, such as the arrival of a large part of the sample and the stability in the estimates.

In addition to the partial samples being biased, it has been observed in recent elections that the complete planned sample hardly ever arrives. Studying the missing data of the 2012 elections, we note that the two main reasons for this missingness are communication problems in nonurban areas and the weather conditions in certain regions, especially heavy rain. Therefore, we must assume that the data is not missing completely at random. As a consequence, the probability that a polling station is not reported may depend on the response we intend to measure.

The context of the analysis is then as follows. We have a stratified sample designed by the committee, so we know the inclusion probabilities and the strata weights, which are proportional to the inverse of the inclusion probabilities. The key challenge is that we have to estimate with incomplete and biased samples, which may imply limited (or null) sample size in some strata and where the missingness is not completely at random. This is where model-based inference brings out its usefulness.

For a population with N units, let $Y = (y_1, ..., y_N)$ be the survey variables and $I = (I_1, ..., I_N)$ be the inclusion indicator variables, where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included. Design-based inference for a finite population quantity $f(Y)$ involves the choice of an estimator $\hat{f}(Y)$. The estimator is a function of the sampled values Y_{inc}, and usually is unbiased for f with respect to I.

Hence, the distribution of I remains the basis for inference. As an implication, design-based methods do not provide a consistent treatment when there is nonresponse or response errors. Model-based inference means modeling both Y and I. The model is used to predict the non-sampled values of the population, and hence finite population quantities. A more detailed explanation can be found in [6].

Let Z denote known design variables; we apply Bayesian model-based inference, meaning that we specify a prior distribution $P(Y|Z)$ for the population values. With this, we have the posterior predictive distribution $p(Y_{\text{exc}}|Y_{\text{inc}}, Z, I)$ where Y_{exc} are the non-sampled values. We make the inference for the total number of votes in favor of each candidate based on this posterior predictive distribution. Occasionally, it is possible to ignore the data collection mechanism. In such a case, the mechanism is called ignorable (see [4, p. 202]). This means that inferences are based on the posterior predictive distribution $p(Y_{\text{exc}}|Y_{\text{inc}}, Z)$, which simplifies the modeling task. However, it also means we are assuming that, conditional on Z, the missing data pattern supplies no information. With a complete sample, the mechanism would be ignorable by including the strata variable in the model. In this setting, there is always missing data. Hence, we include in the analysis all the explanatory variables we consider relevant in the data collection mechanism or in the vote decision. Note that as more explanatory variables are included, the ignorability assumption becomes more plausible.

The proposed model is a *heavy-tailed multilevel mixture model* (denoted as heavy-MM model). This model, defined later on, is a Bayesian multilevel regression, where the dependent variable has a heavy-tailed distribution with a mass at zero. We tested the heavy-MM model using data from 2006 and 2012 gubernatorial elections in the states of Chiapas, Morelos, and Guanajuato. The model was used, among others, to estimate the results in the quick count of the 2018 elections in those three states. In this paper, to show the process of model building, we use the data of Guanajuato.

The outline of the paper is as follows. In Sect. 2, we describe the sample design. In Sect. 3, we define the proposed model. In Sect. 4, we describe the estimation method and calibration. In Sect. 5, we develop the application of the model to the elections of 2018. Finally, in Sect. 6, we give some concluding remarks and future research directions.

2 Sample Design

The sample design was stratified where, within each stratum, we selected polling stations by simple random sampling without replacement. To define the strata, we considered possible combinations of the following variables:

- Federal district: Units in which the territory of Mexico is divided for the purpose of elections.
- Local district: Units in which the territory of each state is divided for the purpose of elections.

- Section type: Urban, rural, or mixed.

In addition, as a comparative point, we also considered simple random sampling without stratification.

For each combination, we computed the estimation precision with a 95% of probability and with different sample sizes using the databases of gubernatorial elections of 2012. The details may be consulted in [1]. The more variables used in the stratification, the smaller the estimation error. The same applies to the sample size, the greater the sample the smaller the error. Nevertheless, there are some other important criteria that need to be evaluated, for example, the total number of strata, the average number of polling stations within each stratum, and the number of strata with few voting stations. Moreover, we also took into consideration the average number of polling stations to be reported by field employees, and the percentage of field employees in charge of more than one polling station. The aim of evaluating all of these criteria is to find a balance that minimizes the errors without jeopardizing the collection of the sample.

After considering all the alternatives, we decided to use the local district with a sample of 500 units, giving rise to 22 strata with an average of 300 polling stations each. Finally, we set the sample size for each stratum proportionally to its size.

3 A Multilevel Model

Multilevel models are appropriate for research designs where data are nested. The model we propose is based on the well-known multilevel regression and poststratification model, which has a long history (see, for example, [10]), but its modern-day implementation can be traced to [9]. The central idea of the multilevel regression and poststratification model is to use multilevel regression to model individual survey responses as a function of different attributes, and then weight the estimates to estimate at the population level. The multilevel regression requires a set of predictors and the choice of a probability distribution. In this section, we discuss those two topics.

3.1 Predictors

For the elections, the INE does a geographic subdivision of the country in electoral sections. The electoral sections can be labeled as urban, rural, or mixed. Within each section, a basic polling station is installed. Additionally, other types of polling station may be installed which are:

- Adjoint polling station: They are installed when the number of voters in the section is greater than 750.

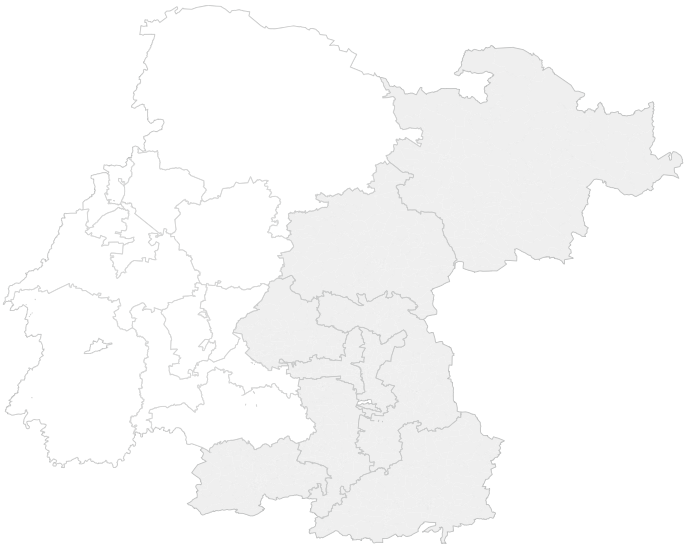

Fig. 1 State of Guanajuato divided by local district; the gray and white indicate the two regions considered as predictors in the heavy-MM model

- Extraordinary polling station: They are for the residents of a section that, because of conditions of communication routes or sociocultural matters, have difficult access to the polling stations.
- Special polling station: They are installed so that the voters outside the section corresponding to their home can vote.

As a consequence, at most 750 citizens are registered as potential voters on every station. The file with the names and photographs of these citizens is called the nominal list.

After testing all the available predictors, Table 1 summarizes the ones that we choose to use. The specification of the region variable can be found in Fig. 1. In addition, it is natural to consider the interaction of section type with section size. We model all the variables in Table 1, except strata, as regression coefficients without multilevel structure.

While exit polls and past election results could be strong predictors, we cannot include them in the model since it is considered to be politically unacceptable for a quick count organized by the electoral authority.

3.2 Multilevel Model with Normal Probability Distribution

Once we have established the predictors, we move to the task of defining the probability distribution assumed for the total number of votes of each candidate. A common

Table 1 Predictors of the multilevel regression model

Predictor	Levels	Notation
Section type	Rural Urban or mixed	Rural ·
Polling station type	Basic or contiguous Special or extraordinary	· typeSP
Section size	Small (less than 1000 voters) Medium (from 1000 to 5000 voters) Large (more than 5000 voters)	· sizeM sizeL
Region	East West	regionE ·
Strata	22 local districts	·

assumption based on asymptotic theory is to use a normal distribution. One contribution that follows this direction is [7]. They propose a Bayesian parametric model where the number of people in favor of a candidate divided by poll and stratum has a normal distribution. Based on the same idea, the first model we raise is the multilevel model with normal probability distribution.

We model each candidate independently, let X_k be the number of votes in favor of a candidate in the k-th polling station, and then

$$X_k \sim \mathsf{N}(\mu_k, \sigma_k^2)\, \mathsf{I}_{[0,750]}, \tag{1}$$

with mean $\mu_k = n_k\theta_k$ and variance $\sigma_k^2 = n_k\psi^2_{\text{strata}(k)}$. The indicator function $\mathsf{I}_{[0,750]}$ is one if the value is in the interval [0, 750] or zero otherwise. The term n_k is the size of the nominal list in the polling station, θ_k represents the proportion of people in the nominal list of the k-th polling station who voted for the candidate, and the variance $\psi^2_{\text{strata}(k)}$ is assumed to be constant in the corresponding stratum.

We fit a multilevel regression model for the parameter θ_k,

$$\begin{aligned}\theta_k = \text{logit}^{-1}(&\beta^0 + \beta^{\text{rural}} \cdot \text{rural}_k + \beta^{\text{rural\_sizeM}} \cdot \text{rural}_k \cdot \text{sizeM}_k \\ &+\beta^{\text{sizeM}} \cdot \text{sizeM}_k + \beta^{\text{sizeL}} \cdot \text{sizeL}_k + \beta^{\text{regionE}} \cdot \text{regionE}_k \\ &+\beta^{\text{strata}}_{\text{strata}(k)} + \beta^{\text{typeSP}} \cdot \text{typeSP}_k).\end{aligned}$$

Finally, we adjust a model to the stratum level,

$$\beta_j^{\text{strata}} \sim \mathsf{N}\left(\mu^{\text{strata}}, \sigma^2_{\text{strata}}\right),$$

Where μ^{strata} is given a $\mathsf{N}(0, 10)$ initial distribution, and σ^2_{strata} is given a $\mathsf{U}(0, 5)$ initial distribution. For the rest of the coefficients, we also assign a $\mathsf{N}(0, 10)$ initial distribution.

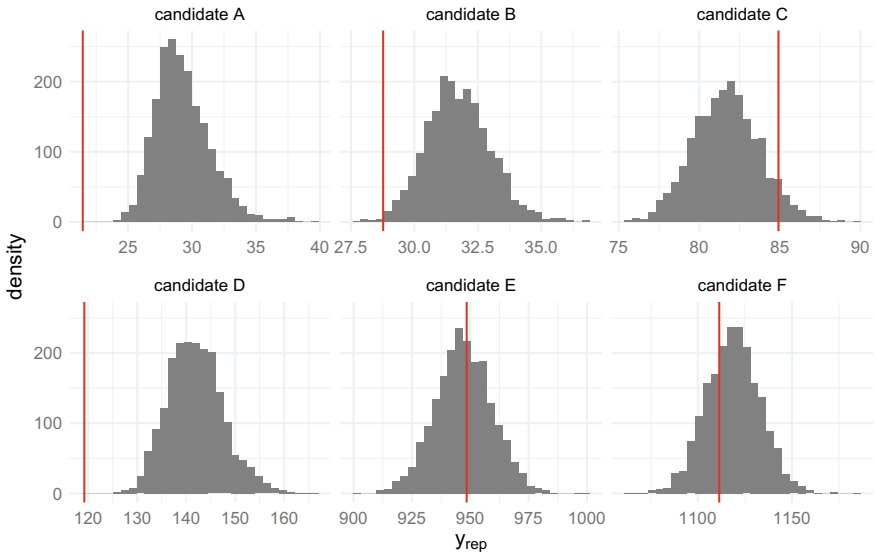

Fig. 2 Posterior predictive distributions of the total number of votes (in thousands) in the 2012 gubernatorial elections of Guanajuato using the multilevel model with normal probability distribution. The red line indicates the total number of votes observed

By adding predictors at the stratum level, we reduce the unexplained variation within each stratum and, as a result, we also reduce the total variation, producing more precise estimates.

As a first step to evaluate the model, we perform a posterior predictive check using the 2012 data. This check helps us test the richness of the model to capture the relevant structure of the true data generating process (see [2]). Figure 2 shows the posterior predictive distributions of the total number of votes. Clearly, the truncated normal distribution produces a bad fit. This is in part due to the longer tails of observed data compared to the normal distribution, in particular with smaller candidates, which tend to have districts where they are extremely popular compared to the rest of the state. Therefore, we need to use another type of probability distribution.

3.3 The Heavy-MM Model

We need a distribution that is also bell-shaped, but with heavier tails than the normal distribution. A natural choice is the t-distribution. However, we also need to catch the high number of zero votes in some polling stations. This leads us to the *heavy-MM model*, which is a multilevel model with a zero-inflated probability distribution. In this model, we replace distribution in Eq. (1) for