Estimating parliamentary composition through electoral polls

Frederic Udina[1], Dept. d'Economia i Empresa, UPF
Pedro Delicado[2], Dept. d'Estadística i I.O., UPC

Abstract

Any electoral system has an electoral formula that converts vote proportions into parliamentary seats. Pre-electoral polls usually focus on estimating vote proportions and then applying the electoral formula to give a forecast of the parliament's composition. We here describe the problems arising from this approach: there is always a bias in the forecast. We study the origin of the bias and some methods to evaluate and to reduce it. We propose some rules to compute the sample size required for a given forecast accuracy. We show by Monte Carlo simulation the performance of the proposed methods using data from Spanish elections in last years. We also propose graphical methods to visualize how electoral formulae and parliamentary forecasts work (or fail).

Keywords: d'Hondt rule, electoral formula, forecasting election results, Monte Carlo, sample size, seats apportion.

1 Introduction

Designing and conducting electoral polls have several well known steps. We focus on forecasting the final parliamentary composition. Indeed, front-page news about any pre-electoral poll usually includes predicted parliamentary composition. In this paper we study the problems related to the estimation of parliamentary composition from a statistical point of view.

[1]`udina@upf.es`. Work supported by Spanish grant BFM2000-0807.
[2]`pedro.delicado@upc.es`. Work partially supported by Spanish grant PB98-0919.

Let K parties be contending for a total of M seats in a parliament. Let C be the number of provinces or electoral regions, and let M_j be the number of seats decided by province j, with $\sum_j M_j = M$ seats, the parliament total. After the elections, the proportion of valid votes f_{ij} obtained by party i in province j is known. These proportions are used by the electoral formula in use (see section 3 for an analysis of some of the more usual ones) to apportion the seats among the parties. Let m_{ij} be the number of seats obtained by party i in province j.

The effect of different electoral formulae from the political point of view is a well studied question (see Cox (1997), Taagepera and Soberg (1989) or Benoit (2000)). But to our best knowledge, the statistical problems related to parliament forecasting have not been fully investigated. Brown and Payne (1985) describe the methods used by the BBC for election night forecasting of the 1983 British general election. Methods are very specific because of the special electoral rules in Britain: each of a large number of constituencies (650 in 1983) decides a seat by the majority rule. Bernardo (1984) describes a Bayesian hierarchical model used in the Spanish general elections of October 1982. It focuses on forecasting the proportion of votes and does not study in depth the problems related to forecasting number of seats apportioned to each party we study here.

A typical pre-electoral poll tries to estimate the proportions f_{ij} by fixing a total sample size N and distributing it among the provinces. The distribution rule is usually somewhere between one consisting of the simple division $n_j = N/C$ and the proportional distribution according to the number of potential voters in each province. After conducting the poll, the sample proportions $\widehat{f}_{ij}$ will be the estimators for the unknown proportions. From these estimated proportions, the estimated number of seats $\widehat{m}_{ij}$ can be computed through the electoral formula in use.

The main point of this paper is to show that when estimating parliamentary composition by adding up the $\widehat{m}_{ij}$, a significant bias appears. We show it both graphically and numerically. We design and describe a graphical method based on principal components to visualize the forecast of several electoral polls once the final results are known.

The bias in the prediction of the parliamentary composition depends on the actual proportions of votes (unknown when the poll is being conducted) and the sample size used in each province. In most cases, the bias vanishes whit increasing sample size, but there are some critical values of the proportions that make forecasting the results impossible: the simplest case occurs when two parties have each 50% of voters and they are contending for a seat.

We also study methods for distribution of sample sizes among provinces both using a pilot poll to obtain a first estimate of the unknown proportions and without using it. Using data from electoral polls and elections in Spain from the year 2000, and in Catalonia for the regional parliament in 1999, we study by Monte Carlo simulations the performance of the proposed methods.

We do not address at all in this paper the non-sampling errors that seriously affect polls: miss-responses, abstention detection and no answers or missing data. We will see that even under ideal sampling conditions there remain difficult and interesting problems in the estimation of the parliamentary composition.

In the next section we describe the main problem, the bias in the parliamentary composition forecasts, using a gaphical method we have designed. The device can also be used to show the discrepancy between the forecast of any real electoral poll and the ones resulting from Monte Carlo simulation. In Section 3 we briefly describe and analyze the most commonly used electoral formulae and we study the origin and consequences of the estimation bias.

In Section 4 we study the problem of choosing a sample size for each province to achieve a prefixed level of error estimation. After a summary of conclusions, the mathematical details are included in the Appendix.

2 Visualizing results and polls

To represent graphically the results of some parliament elections, and to compare them with the forecast of pre-electoral poll we use a principal components (PC) based biplot. We illustrate the technique using data from the aforementioned election results. Starting from the known final results, we use the computer to generate B samples (typically B =2,000) drawn from C multinomial distributions, one for each province. We use the same sample sizes as some of the published pre-electoral polls that give enough technical details. Applying the electoral formula to each simulated proportions we obtain a forecasted parliament, a *virtual parliament*, a vector of K integers. Using the two main principal components of this cloud of points we can represent it the better way in a plane (see Figures 1 and 2). There is a relevant pattern in the cloud of points: feasible parliaments have integer coordinates that add up to M so they are arranged in some hyperplanes that are still visible when projected to our PC plane. Actually, for better visualization we only draw some of the points (for instance 500) randomly selected.

In the same graphic we represent the vectors corresponding to the parties by projecting the variable vectors onto the representation plane. We draw them with their origins at the average point of the cloud of virtual parliaments. By also drawing the point that corresponds to the real parliament, we

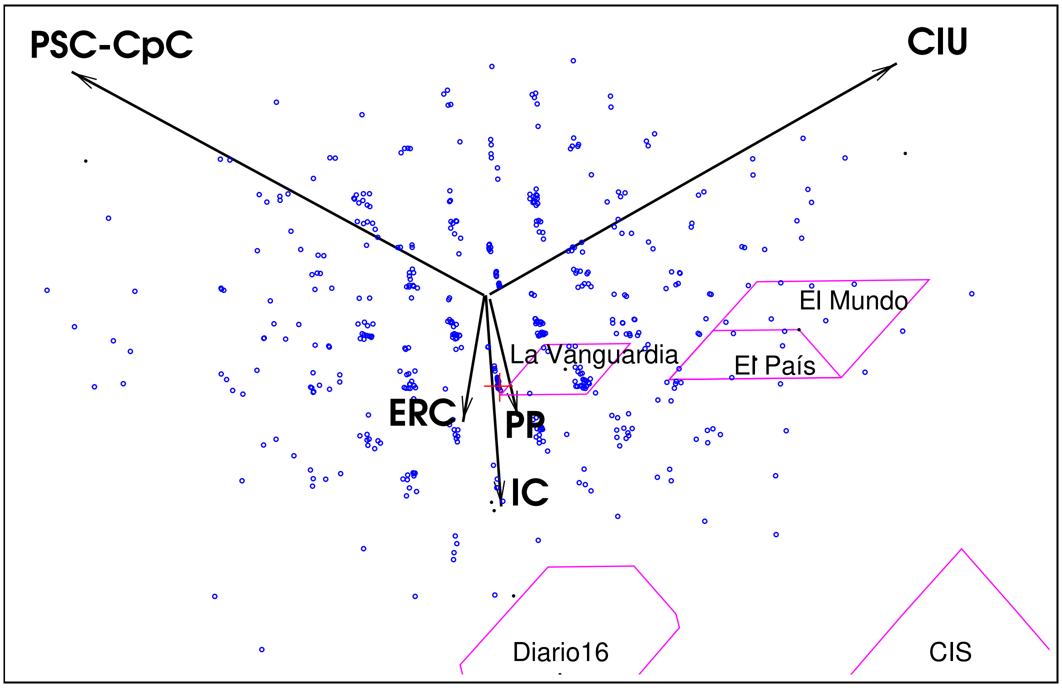

Figure 1: Biplot based on principal components that shows polls and results for Catalan parliamentary elections in 1999, in terms of seats apportioned to the parties. Points in the cloud represent (a sample of) 2,000 parliaments obtained from simulated polls based on the final results. The arrows, originating at the point average forecasted parliament, represent the direction favouring the respective parties. A cross, located above the "PP" label, marks the position of the real parliament showing the bias of the estimation. The polygons represent the forecast given by polls published in several newspapers a week before the elections.

show visually that there is a significant bias in the estimation of parliamentary composition. In the cases depicted in the figures, one can evaluate the bias roughly as being between one half and one third of the sample variability.

To incorporate the parliaments forecasted by pre-electoral polls we use the confidence intervals they give for the number of seats forecast for each party. We compute all the feasible parliaments that fit the given confidence intervals and draw the convex hull of the projected points. If any of the polls used a different global sample size from the one we used to generate the virtual parliaments we correct the distance to the origin of the corresponding polygon accordingly.

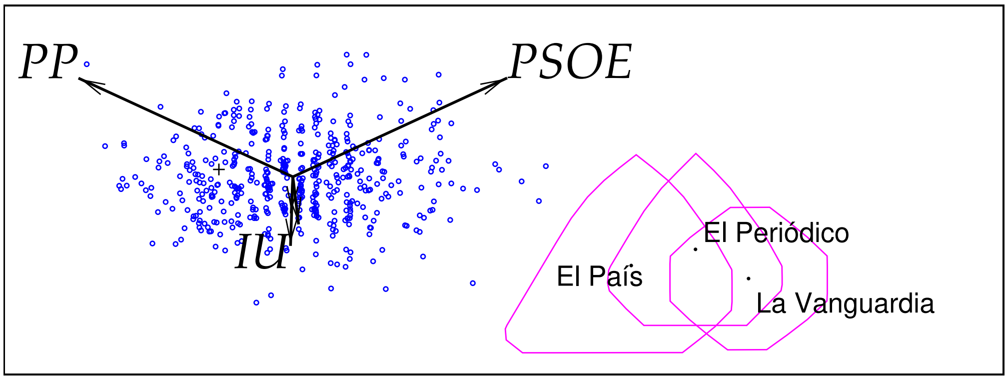

Figure 2: This graphic is similar to the one in Figure 1. It shows polls and results for general elections in Spain for the year 2000 parliament. Arrows originate at the centre of the cloud formed by parliaments obtained from simulated polls. To the left of this point, the real parliament has been marked with a small cross. The bias is apparent. Some published pre-electoral polls are displayed as in Figure 1, while others are too far away to be included.

3 Electoral formulae: an estimation bias problem

We concentrate here on a single province, where K parties obtained proportions of votes $(f_1, f_2, \ldots, f_K)$ and there are M seats to be apportioned among them. An electoral formula is a rule for translating these proportions to a seat allocation $(m_1, m_2, \ldots, m_K)$ such that $\sum m_i = M$. Most electoral formulae (see Taagepera et al. (1989)) are *proportional rules* that attempt to make the averages f_i/m_i similar among the parties. The most frequently used proportional rules work as follows.

1. Exclude from the seat distribution parties that have vote proportions less than a fixed threshold $\delta \geq 0$.
2. Choose a non-decreasing sequence of denominators $d = (d_1, d_2, \ldots d_M)$.
3. Form all the $K \times M$ quotients f_i/d_j, $i = 1 \ldots K$, $j = 1 \ldots M$.
4. Select the M largest quotients and give the corresponding parties a seat for each of its largest quotients.

The choice of the denominator sequence d controls the proportionality of the rule (see Benoit (2000) for a study of proportionality). As an extreme case, if $d = (1, 1, 1, \ldots)$ the rule gives all the seats to the most voted party, it has no proportionality at all. The so called *d'Hondt rule* (PR-HR) takes $d = (1, 2, 3, \ldots)$. It is the rule used in Spain (with $\delta = 0.03$) and other European countries. It is also used in the U.S. to distribute congress seats among states according to population size. It gives more chances of obtaining seats to medium-sized parties and less to small parties. The *Sainte-Lagüe rule* (PR-SL) takes $d = (1, 3, 5, 7, \ldots)$ and makes it easier for small parties to obtain a first seat. Among the commonly used formulae, the *modified Sainte-Lagüe rule* (PR-MSL) has maximum proportionality, see Benoït (2000). It takes the sequence $d = (1.4, 3, 5, 7 \ldots)$. To ensure that every party having more votes than δ has at least a seat, one could use the sequence $d_j = 1 + M(j - 1)$. In the rest of the paper we use the d'Hondt rule (except in Figure 5 for comparison) but a similar analysis could be performed using any other rule.

In Appendix A.1 we include a mathematical formulation of these proportional rules and in Figures 3 to 4 we give a graphical idea about how they work. In Figure 3 we show the case of two parties contending for five seats. The horizontal axis is proportion of votes for party 1, party 2 having the rest. It can be seen that there are some values of f_1 that make the number of seats jump, and that there are intervals where the seat allocation is constant. The rule used there is PR-HR. Figure 4 shows the seat allocation when there are three parties in the game. The triangle depicted is the R^3 simplex $\{(f_1, f_2, f_3) \in R^3 | f_1 + f_2 + f_3 = 1\}$. So each point corresponds to a feasible combination of vote proportions. This is a so-called *ternary diagram* or *barycentric coordinate space*, see Aitchison (1986). Triangular coordinates are in use: a given point has proportions measured by the distances to the sides of the triangle. It can be seen that the regions of constant seat allocation (we name it *constant seat allocation cells*, CSA-cells) are convex polygons with 4 to 6 sides. See Appendix A.1 for mathematical details.

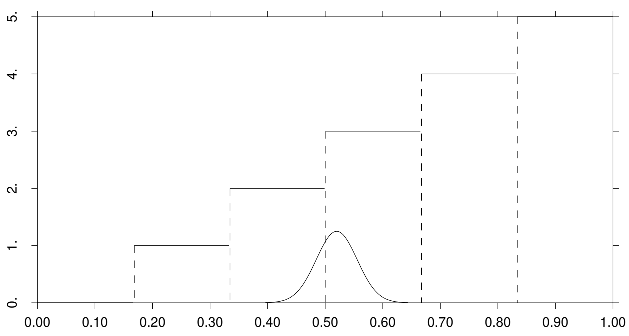

Figure 3: In a province like Cáceres in the Spanish elections, in 2000, only two parties were effectively contending for five seats. The horizontal axis is the proportion of votes for one of them. On the vertical axis, the number of seats apportioned to it. For a population proportion of 52%, the real result there, the small bell-shaped curve shows the sample proportion distribution for a sample size of $n_i = 199$, as used in one of the main electoral polls.

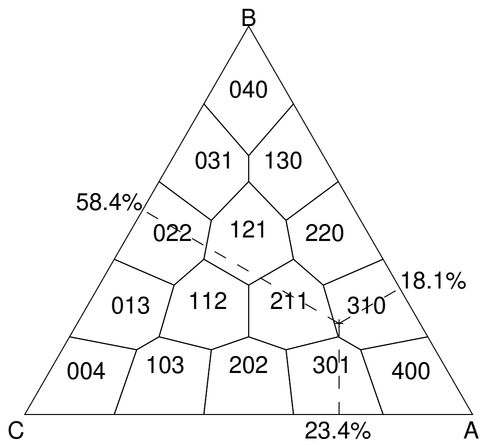

Figure 4: Three parties are contending for four seats. Points in the triangle correspond to three proportion of votes adding up to one, in triangular coordinates. Polygons shown in the triangle are sets of proportions with constant seat allocation (CSA cells) labeled with the number of seats for parties A, B, and C using the d'Hondt rule. The marked point is for the results in Ourense, Spanish elections 2000. It gives three seats to party A, one to party B and none to C.

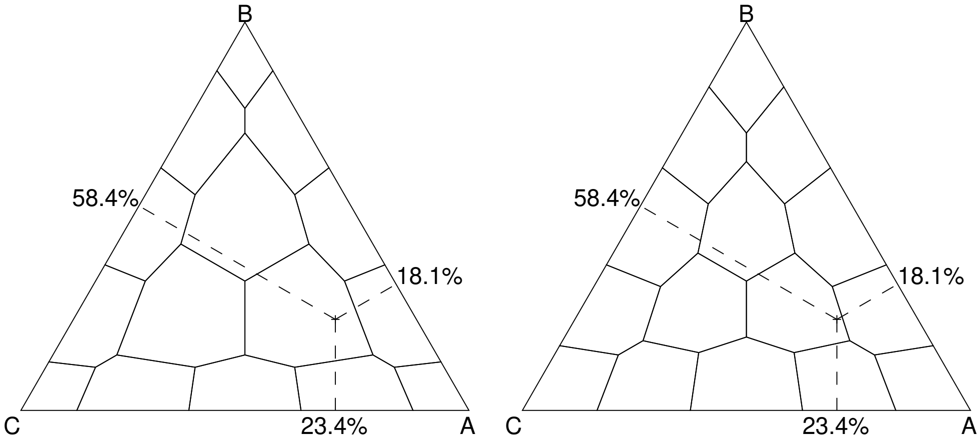

Figure 5: The graphs are similar to Figure 4 but different seat apportioning rules are in use. On the left we use the Sainte-Lagüe rule and on the right the modified Sainte-Lagüe rule. Note that the main difference is the way small and medium size parties are favoured.

3.1 The bias in estimating the number of seats

The origin of the bias shown in figures 1 and 2 can be clearly seen in simple cases. When there are only two parties, as in Figure 3, if the real proportion in favour of the first one is close to one of the jumps, a significant part of the samples drawn from the population would predict the wrong number of seats. In the situation depicted in Figure 3, as many as 28.6% of the samples would predict two seats, so the expected number of seats predicted by sampling would be $0.714 \times 3 + 0.286 \times 2 = 2.714$ with a bias of -0.286. We use here the normal approximation to the binomial distribution and similar approximation to the multinomial distribution in the following discussion.

Figure 6 shows a case where there are three parties with proportions of votes $49.7\%, 39.3\%$, and 11.0%. The point is very close to two edges of its CSA-cell (the one that is below, corresponding to 5 seats for A, 3 for B and one for C). Sampling from these proportions is very unstable from the point of view of seats allocation : more than 50% of the samples would give a wrong forecast. The ellipses centered on the point with these coordinates are level curves of the joint distribution of the sample proportions when the sample size is $n_i = 337$ (used by one of the main electoral polls published before the elections). Levels have been chosen so that the probabilities inside the ellipses are, respectively, $0.90, 0.95, 0.99$. These level curves can be computed using the χ^2 distribution with two degrees of freedom, see Section 4.2. Most of the samples give proportions that fall outside the correct constant seat

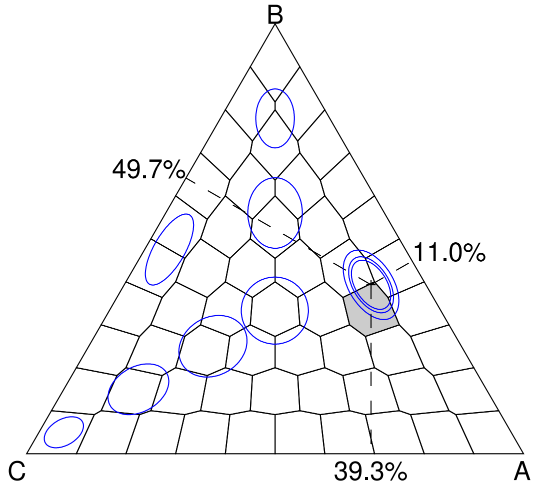

Figure 6: The graphic shows the case of Asturias in Spanish elections and polls in the year 2000. Triangle and polygons are as in Figure 4. The ellipses are level curves of the sample proportion distribution for sample size $n = 337$ and several population proportion values. Around the marked point (final results in Asturias) we draw the level curves containing $0.90, 0.95,$ and 0.99 of the probability. The other ellipses contain 0.99 of the probability for other values of the population proportions.

allocation cell. The average predicted seat allocation will give more seats to parties B and C, and less to party A than the ones apportioned according to the real proportion of votes.

The magnitude of the bias vector, defined as the difference between average seat allocation by the samples and real seat allocation, depends on the sample size but mainly on the real proportion of votes: If the point with real proportions falls close to the centre of its CSA-cell, a small sample size can be enough to get a good forecast, but if it falls close to any cell edge, a bigger sample size is needed. In the singular cases when these proportions fall on a cell edge, the bias will not disappear even if the sample size increases to infinity.

This is a serious problem for parliamentary forecasting. There are some opinion poll firms that conclude that parliamentary forecasts ought never be published. Others publish them giving some confidence intervals for the seats apportioned to each party; in the examples we have studied, however, the confidence level is not clearly stated. Good confidence intervals could be computed by simulation as described in Section 3.3.

3.2 Estimating the bias by the parametric bootstrap

One possible way to correct the bias or to compute confidence intervals when the real proportions are unknown is by the parametric bootstrap. We apply this technique described in Efron and Tibshirani (1993) in the following way.

1. Let S_0 be an electoral poll conducted in every province with sample sizes n_i for a total sample size of N. Let $\widehat{f}_{ij}, i = 1 \dots K, j = 1 \dots C$ be the resulting sample proportions for each party in each province. Let $\widehat{m}_{ij}$ be the number of seats apportioned to party i in province j according to these results.

2. Repeat B_1 times:

   (a) For each province, draw a sample S_1, using the same sample sizes as in S_0, from a multinomial distribution with proportions $\widehat{f_{ij}}$.

   (b) Use the resulting sample proportions to apportion seats $\widehat{m}^*_{ij}$.

3. Using the bootstrapped seats $\widehat{m}^*_{ij}$ obtained in step 2 compute its mean $\bar{m}^*_{ij}$ over all the B_1 bootstrapped samples. Then, $\bar{m}^*_{ij} - \widehat{m}_{ij}$ is a good estimator of the unknown bias $\widehat{m}_{ij} - m_{ij}$. The new bootstrap estimate of m_{ij} is then
$$\widehat{m}^B_{ij} = \widehat{m}_{ij} - (\bar{m}^*_{ij} - \widehat{m}_{ij}).$$

Table 1: Bias (first row) and estimated bias (second row) for total of seats for each one of 12 parties.

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
-2.51	2.04	0.45	-0.85	-0.26	0.39	0.86	-0.08	0.19	-0.02	-0.11	-0.09
-0.45	0.48	0.01	0.03	-0.12	0.00	0.08	0.01	0.12	-0.04	-0.03	-0.09

We conducted Monte Carlo simulations to evaluate the performance of this bias estimation method using data from the 2000 Spanish general elections (see Delicado and Udina (2001) for details of the electoral results and poll data). There were $K = 12$ parties and $C = 52$ provinces (most of the parties have no real presence in many provinces). We used a total sample size of N =15,000 distributed among provinces by giving a fixed quota of 100 to each one and apportioning the rest in proportion to the size of the electoral census (These were the sizes used by some of the main electoral polls). Using the results of the elections, we produced B_0 =1,000 polls like S_0 described in step 1 above. We applied the above described procedure to each one of these 1,000 polls using B_1 =1,000. Table 1 lists the average bias given by these polls in the first row. This row numerically reflects the same bias that can be seen graphically in Figure 2. In the second row we list the average (over B_0 samples) of the bootstrap estimates of the bias. Note that the magnitude of the bias is severely underestimated, but the direction of the estimation is mostly correct: figures in table 1 have a correlation of 0.91.

3.3 Bootstrapped confidence intervals

Using a similar simulation setup, we computed for every simulated poll S_0 (out of $B_0 = 1000$ polls as described in step 1 above) a *confidence interval* for the composition of the parliament as follows, for a fixed *nominal* confidence level $1 - \alpha$.

1. Let $\mathcal{P}_0$ be the parliament obtained from S_0. Take B_1 =1,000 parliaments $\mathcal{P}_l$ ($l = 1 \ldots B_1$) obtained from samples like S_1 in step 2 (section 3.2), compute distances d_l from $\mathcal{P}_0$ to $\mathcal{P}_l$ and compute
$$d_\alpha = \min \Big\{ d_l \mid \#\{d_k | d_k \leq d_l\} \geq (1-\alpha) B_1 \Big\}.$$
2. Then we take all the parliaments that have $d_l \leq d_\alpha$ and for each party we compute the interval that covers all the seats apportioned to that

party in these parliaments. This finally gives a set of K intervals, which we call the confidence interval for the parliament $\mathcal{P}_0$ (CIP).

It is expected that at least a proportion of $(1 - \alpha)$ of these CIP that are computed from all the simulated polls S_0 effectively contain the real parliament. Table 2 shows that this is really so for data from Spanish 2000 elections. We used the Euclidean distance (similar results were obtained for other distances). We report Monte Carlo results for nominal confidence levels of 90% (a reasonable one) and 60%, which gives intervals of comparable width to those of the main published pre-electoral polls (typical interval widths were $6, 7, 2, 1, 1, 2$ for the six bigger parties). Note that none of these published polls gave any statement about the coverage of the given intervals. Our Monte Carlo studies, using data from the published polls, show a very wide range of coverages (see Delicado and Udina (2000) for further details).

Table 2: Effective coverage of parliament confidence intervals computed by bootstrap using simulated polls. The mean width of the intervals for bigger parties is also reported. The width used in some of the main published pre-electoral polls is comparable with the 60% nominal intervals.

Nominal coverage $(1-\alpha)$	Real coverage	Mean interval width for 6 biggest parties					
90%	92.7%	10.7	10.7	5.0	6.7	3.7	2.7
60%	62.3%	6.6	6.6	4.6	5.6	3.5	2.6

4 Choosing the sample size

Electoral polls generally choose the sample size for each province based on the number of potential voters. In general, it is proportional to the number of seats in the province. But two provinces with similar numbers of seats can have different numbers of contending parties and, even if the number of parties is the same, they can differ in the proportions of voters for each party. From our point of view, when a pre-electoral poll is designed, regional sample sizes should be assigned accordingly to the difficulty of estimation in each electoral region.

Section 3 shows that the difficulty with estimating the number of seats in a given electoral region depends jointly on the number of seats, the number of parties and the exact values of the proportion of votes for the parties involved. In this section we propose two ways for choosing the sample size for each region. The first one is derived from a probabilistic analysis. Geometric arguments are the basis for the second one.

4.1 Probabilistic rule for choosing the sample size

A first approach to decide the sample size in a region is to choose n big enough to ensure that, with a probability greater than $(1-\alpha)$, the right seats apportion is predicted.

If the real vote proportions are known or if some estimations of them are available (from the results of a previous election, or from a pilot poll), standard multivariate techniques can be used to compute the sample size required to ensure a desired precision in seat estimation. The main idea is as follows. The $1-\alpha$ fraction of estimated proportions that are closest (in Mahalanobis distance) to the real proportions, form an ellipsoid which volume decreases with n. So this ellipsoid would be completely included into the right CSA-cell for large values of n.

The following theorem gives the rule expression (see a more precise statement and a sketch of the proof in the Appendix A.2).

Theorem 1 *Let $p=(p_1,\dots,p_K)^t$ the vector of population proportions and $\hat{p}$ the sample proportions obtained by simple random sampling with sample size n. For x in the simplex $\sum_{i=1}^K x_i = 1$, let $H(x)$ be the vector of seats apportioned by the d'Hondt rule to x, and set $H(p)=(h_1,\dots,h_K)$. Let α be the maximum proportion of polls with wrong seat estimation admitted. Then taking*

$$n \geq n_\alpha = \frac{1}{D(p)}\chi^2_{K-1,\alpha},$$

we have

$$Prob(H(\hat{p}) = H(p)) \geq 1-\alpha.$$

$D(p)$ is the distance (measured by Σ^-, a generalized inverse of the covariance matrix Σ of p) from p to the frontier of its CSA-cell, and it is computed by

$$D(p) = \min_{i,j} \frac{(h_j+1)^2p_i^2 + h_i^2p_j^2 - 2h_i(h_j+1)p_ip_j}{(h_j+1)^2(p_i-p_i^2) + h_i^2(p_j-p_j^2) + 2h_i(h_j+1)p_ip_j}.$$

where minimization is done over all indices satisfying

$$1 \leq i \leq k, 1 \leq j \leq k, i \neq j, h_i > 0.$$

In practice, the sample sizes for several regions have to be chosen jointly and they must add up to a given N. Then α is calibrated iteratively in order to have

$$\sum_j n_{j,\alpha} = N,$$

where $n_{j,\alpha}$ is the sample size obtained from the theorem for region c_j. Observe that N growths with $(1-\alpha)$.

As an example, we compute the sample size needed in Asturias in Spanish election 2000. The proportions p_i are known and we can see that it is a very difficult region to be predicted, as it is shown in Figure 6. To get the right apportionment with probability 0.95, the sample size proposed by the theorem is in the order of 300,000, clearly unreachable in practice.

As this example and the Monte Carlo studies (see section 4.2.1) point out, the probabilistic rule presented above is very *conservative* in the sense that, even in the worst proportions configuration, it provides the right sample size for having no errors (with probability greater than $1-\alpha$). The price for this exactness is that the proposed sample sizes can be enormous for difficult cases: regions with real proportions of voters near the CSA-cell frontier.

Next subsection proposes a more practical rule: it does not need pilot estimation of proportions and proposes more realistic sample sizes.

4.2 Geometric rule for choosing the sample size

Two provinces with similar census sizes and even similar number of seats can have a different number of contending parties. This is the case in Barcelona and Madrid in Spain, with 31 and 34 seats in the parliament, respectively: while in Barcelona five parties get parliamentary representation, only three do it in Madrid.

We take into account the number of seats and the number of parties to measure the *a priori* difficulty of seat allocation estimation of a region. A rule for choosing the sample size is based on this measure.

For K parties and M seats, the number of ways seat allocation can be done is the number of ways to form K groups from M identical objects. This number is

$$\mathrm{NC}(K,M) = \binom{M+K-1}{M}.$$

So, for example, there are 52,360 possibilities in Barcelona, and only 630 in Madrid. The total *volume* of the K-dimensional simplex is that of the solid in $\mathcal{R}^d$, $d=K-1$, given by

$$\mathrm{Volume}(\{x \in \mathcal{R}^d : x_i \geq 0 \text{ and } \sum x_i \leq 1\}) = \frac{1}{d!}.$$

The average volume of a CSA-cell is $A(K,M) = (d!\,\mathrm{NC}(K,M))^{-1}$. Let us compare this average volume with the volume covered by the sample variability of the sample proportions.

Fix a confidence level $1-\alpha$ and a sample size n. The ellipsoid containing probability $1-\alpha$ is

$$\{x \in R^d : x^t \Sigma_n^{-1} x \leq k^2\}, \qquad k^2 = \chi^2_{d,\alpha}$$

where Σ_n, the variance-covariance matrix of the proportions, is given by

$$\Sigma_n = n^{-1}(\sigma_{ij}), \quad \sigma_{ii} = p_i(1-p_i), \quad \sigma_{ij} = -p_i p_j \qquad (i,j = 1\dots d,\, i \neq j)$$

The volume of the ellipsoid is (see Johnson-Wichern (1998), p. 132)

$$V(\alpha, n) = \frac{2\pi^{d/2}}{d\,\Gamma(d/2)} |\Sigma_n|^{1/2} k^d$$

which can be written as $n^{-d/2}V(\alpha,1)$. As long as the proportions for the parties are unknown, we take a *worst case* approach: set $p_i = 1/K$ for all i.

In this framework, one could decide to take sample sizes in each province so that the $1-\alpha$ ellipsoid covers the volume of a single cell. This will ensure roughly that the error in the seat estimation would be no greater than one seat for each party in that province. But this rule can require a too large total sample size. To avoid this problem, we propose the following sample size allocation rule.

Geometric rule:

1. Fix α.
2. Set G, the *number of contiguous cells to be covered*, initially equal to one.
3. Determine, for each province, a sample size n_i so that

$$n_i = \left(\frac{G^g A(K,M)}{V(\alpha,1)}\right)^{-2/d}$$

   using as K the number of parties that have some chance of obtaining at least one of the M_i seats in the province (the exponent g is discussed below).
4. Add up all the sample sizes $N = \sum n_i$ and if it is too big, increase G or α by adjusting them to get the desired total sample size.

The exponent g needs to account for the dimensionality in each province. We assign it so that doubling G the number of covered cells includes all the contiguous cells to the initial one. This gives $g = \log(K-1)!/\log 2$.

Table 3: Performance of the geometric rule in several provinces with different number of parties and seats, and difficulty level ($\alpha = 0.05$). At right, percentage of polls (out of 2,000) that give respectively $0, 1, 2$, or more than 2 misaportioned seats are listed in each case.

				% of polls with wrong seats			
Province	# parties	# seats	n_i	0	1	2	3+
Asturias	3	9	399	46.25	53.35	0.40	0.00
Badajoz	3	6	203	94.05	5.95	0.00	0.00
Barcelona	6	31	2892	73.65	25.90	0.45	0.00
Ceuta	2	1	16	92.00	8.00	0.00	0.00
Madrid	3	34	4565	90.40	9.60	0.00	0.00
Ourense	4	4	113	55.70	44.30	0.00	0.00

A surprising result from this geometric rule, confirmed by Monte Carlo results below, is that it is *more difficult* to get the right seats apportion when there are fewer parties in the game. Precisely, to get the right seats (accepting one misapportioned seat, and $\alpha = 0.01$) in a province with 30 seats and 3 contending parties, a sample size of 5,525 is needed. For the same number of seats and 6 parties, the required sample size is just 3,712.

4.2.1 Monte Carlo results

We tested the geometric rule using data from Spanish 2000 election. Setting $G = 1$ and $\alpha = 0.05$ we simulate in each case $B =$2,000 samples using the real results of the election. We list in Table 3 the probabilities of getting the right seat apportionment (or misapportioning some seat) in a poll with sample size computed by the geometric rule just described. Note that Asturias and Ourense have results that made it really difficult to get the right seat distribution, as it can be seen in figures 4 and 6. Badajoz and Madrid are relatively easy, in the sense that the vote proportions point is quite centered in its CSA-cell.

To compare the performance of the geometric rule in different settings, other Monte Carlo studies are reported in tables 4 and 5. In Table 4 number of parties, number of seats and proportions of votes are chosen so that all parties have equal proportions and equal number of votes. This corresponds geometrically to having the point in the center of the central CSA-shell in the simplex (refer to figure 6). We take $\alpha = 0.05$ and choose the sample size by the geometric rule (with $G = 1$). Then we simulate $B =$5,000 samples from

Table 4: Performance of sample size geometric rule in several artificial settings with different number of parties and seats ($\alpha = 0.05$). Proportions of votes are equal for the contending parties, so the point is in the center of a CSA-cell. At right, percentage of polls (out of 5,000) that give respectively $0, 1, 2$, or more than 2 misaportioned seats are listed in each case. Geometric rule (G-rule) is in use, sample sizes given by the probabilistic or conservative rule are listed in column C-rule.

			% of polls with wrong seats				
K	M	G-rule	0	1	2	3+	C-rule
10	30	2081	89.74	10.14	0.12	0.00	4213
6	30	2724	93.20	6.76	0.04	0.00	4041
5	30	2974	93.66	6.28	0.06	0.00	4022
5	15	860	94.02	5.98	0.00	0.00	1176
5	5	156	93.92	6.06	0.02	0.00	227
3	30	3594	94.90	5.10	0.00	0.00	3965
3	15	986	95.08	4.92	0.00	0.00	1090
3	6	203	94.92	5.08	0.00	0.00	228

Table 5: Number of parties and seats, and sample size are the same as in Table 4. Proportions of votes are set here to locate the point near the corner of its CSA-cell. At right, percentage of polls (out of 5,000) that give respectively $0, 1, 2$, or more than 2 misaportioned seats are listed in each case.

							% of polls with wrong seats			
K	M	Prop. of votes (seats)					0	1	2	3+
5	30	.35(11)	.263(8)	.240(8)	.088(2)	.059(1)	27.12	68.70	4.18	0.00
5	30	.208(7)	.206(6)	.205(6)	.204(6)	.177(5)	27.24	72.46	0.30	0.00
5	15	.212(4)	.211(3)	.210(3)	.209(3)	.158(2)	20.48	79.00	0.52	0.00
5	5	.224(2)	.223(1)	.221(1)	.220(1)	.110(0)	18.54	80.84	0.62	0.00
3	30	.345(11)		.343(10)		.312(9)	39.46	60.44	0.10	0.00
3	15	.354(6)		.352(5)		.294(4)	34.98	64.88	0.14	0.00
3	5	.336(2)		.333(2)		.331(1)	32.30	67.52	0.18	0.00

the appropriate multinomial distribution. Since the point is in the center, we expect $100(1-\alpha)\%$ of the samples giving the correct seat apportionment, and this is confirmed by the results of the simulation. For each setting, we also report in Table 4 the sample size given by the probabilistic or conservative rule (see 4.1). Since the point is in the center of the CSA-cell, result are similar. Note that when number of parties (dimensionality of the simplex) increases, C-rule gives sample sizes too big. This is mostly due to the well known *sphere effect* associated with the *curse of dimensionality*: the proportion of the volume of a cube filled by the inscribed sphere decreases when dimensionality increases.

In Table 5 we choose the proportions of votes for the parties so that the points are located very near a corner of the CSA-cell. We list in the table for each case the proportions of votes and the number of seats for each party. The first case corresponds to a vertex of a cell close to the border of the simplex, while the rest correspond to vertices of the central CSA-cell. Sample sizes are computed by the geometric rule as before. We do not report here the samples sizes given by the conservative rule: they would be really unrealistic. As before, we simulate B =5,000 samples and we report the percentage of samples that give the correct seat apportionment, or that give 1, 2, or more that 2 incorrectly apportioned seats. Simulation results are satisfactory.

5 Concluding remarks

We have presented graphical tools to evaluate the results of pre-electoral polls. The problem of bias in allocation seats estimation has been pointed out as an essential one. The graphical study of the bias problem indicates that the difficulty in estimating the seats allocation in a province depends on several parameters. When all them are taken into account, two different rules for choosing the sample size are obtained, in the hope they reduce the estimation bias. The second rule is the most advisable, as a simulation study indicates.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.

Benoit, Kenneth (2000). Which electoral formula is the most proportional? a new look with new evidence. *Political Analysis*, **8:4**, 381–388.

Bernardo, José M. (1984). Monitoring the 1982 spanish socialist victory: A bayesian analysis. *JASA*, **79**, 510–515.

Brown Ph., Payne C. (1984). Forecasting the 1983 british general election. *Statistician*, **33, No. 2**, 217–228.

Cox, Gary W. (1997). *Making votes count. Strategic coordination in the World's Electoral Systems.* Cambridge U. Press.

Delicado, P. and F. Udina (2001). ¿Cómo y cuánto fallan los sondeos electorales? *Document de Recerca. Departament d'Estadística i Investigació Operativa, UPC.*, **2001/02**, 1–26.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman and Hall.

Johnson and Wichern (1998). *Applied Multivariate Statistical Analysis, 2nd edition.* Prentice Hall.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate analysis.* London: Academic Press [Harcourt Brace Jovanovich Publishers]. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Taagepera, R. and M. Soberg Shugart (1989). *Seats and Votes. The Effects and Determinants of Electoral Systems.* Yale U. Press.

A Mathematical details

A.1 Proportional rules

Let K parties be competing for M seats (in a single province). Let δ be the minimum proportion to obtain any seat. Let $(f_1, f_2, \ldots, f_K)$ be the proportion of votes obtained respectively by the parties. We have

$$0 \leq f_i \leq 1, (i = 1, \ldots, K), \qquad \sum_{i=1}^{K} f_i = 1.$$

Let d_j, $j = 1 \ldots M$, be a non-decreasing sequence. Define the quotients $q_{i,j}$, for $i = 1, \ldots, K$ and $j = 1, \ldots, M$ by

$$\begin{array}{ll} \text{If } f_i < \delta, & q_{i,j} = 0 \\ \text{If } f_i \geq \delta, & q_{i,j} = f_i/d_j \end{array}$$

A quotient $q_{i,j}$ deserves a seat if and only if it is one of the M greatest among all the KM quotients, or equivalently, if there are more than $M(K -$

M) quotients smaller than itself[3]. So the rule is an application S from the simplex

$$\{f \in R^K | \sum_{i=1}^{K} f_i \leq 1\}$$

onto the discrete set of possible parliamentary configurations

$$\{m \in Z^K | \sum_{j=1}^{K} m_j = M\}$$

defined by

$$\begin{aligned} &S_d(f_1, \ldots, f_K) = (m_1, \ldots, m_K) \Longleftrightarrow \\ &\quad \forall i = 1, \ldots K, \quad m_i = \max\{j = 1, \ldots, M | Q(i,j) > KN - N\} \\ &\quad\ \text{where } Q(i,j) = \#\{q_{k,l} < q_{i,j} : k = 1, \ldots, K, l = 1, \ldots, M\} \end{aligned} \tag{1}$$

Note that $Q(i,j)$ is the number of quotients less than $q_{i,j}$. The definition can be written in the following non-closed form that may be more useful:

$$\begin{aligned} &S_d(f_1, \ldots, f_K) = (m_1, \ldots, m_K) \Longleftrightarrow \\ &\quad \forall i, j \in \{1, \ldots, K\}, i \neq j, \quad m_i = 0 \text{ or } \frac{f_i}{d_{m_i}} > \frac{f_j}{d_{m_j+1}} \end{aligned} \tag{2}$$

which states that the last quotient of party i that got a seat must be bigger than the first quotient of party j that does not got one.

With this definition it is easy to understand that the regions with constant seat allocation (which we call CSA-cells) are limited by hyper-planes and thus are convex polyhedra. Each one is limited by the inequalities appearing in (2), up to $K(K-1)$ of them can be effective, and in the case $m_i = 0$ the effective inequality is simply the boundary of the simplex. We have seen in figures 3, 4, and 6 examples of such regions in the plane, when $K = 3$.

A.2 Sample size needed to get the right seats

Following notation in theorem 1, $\hat{p}$ is n^{-1} times a multinomial random variable, $\hat{p} \sim n^{-1} M_k(n; p_1, \ldots, p_k)$ that can be approximated, for n big enough, by a multivariate normal with mean $p = (p_1, \ldots, p_k)^t$ and $k \times k$ covariance matrix $n^{-1}\Sigma$, where $\Sigma = Diag(p) - pp^t$.

[3]Ties among quotients are a set of very small probability. In such improbable cases, electoral laws usually give the seat to the party with greater absolute number of votes (and by lottery if these are equal).

Σ has, in general, rank $K-1$. If $\lambda_1, \dots, \lambda_{K-1}$ are the non-zero eigenvalues of Σ, it is possible to find a $K \times (K-1)$ matrix C with $C^t C = I_{K-1}$ such that

$$\Sigma^- = C\, Diag(\lambda_1, \dots, \lambda_{K-1})^{-1}\, C^t$$

is a generalized inverse of Σ, i. e., it satisfies $\Sigma\Sigma^-\Sigma = \Sigma$ (see Mardia, Kent, and Bibby (1979)).

Σ^- defines a metric in the simplex in R^K. If we consider a hyperplane $x^t b = \beta$, it can be shown that the probability that p and $\hat{p}$ are on the same side of the hyperplane is greater than

$$P((\hat{p}-p)^t n \Sigma^- (\hat{p}-p) \leq D^*) = P(\chi^2_{K-1} \leq D^*)$$

where D^* is the distance from p to the hyperplane, given by

$$D^* = n \frac{(\beta - p^t b)^2}{b^t \Sigma b}.$$

In our problem, the hyperplanes are given by (see (2) above)

$$\frac{p_i}{h_i} = \frac{p_j}{h_j+1} \iff (h_j+1)p_i - h_i p_j = 0$$

and we have

$$D^* = n \frac{(h_j+1)^2 p_i^2 + h_i^2 p_j^2 - 2h_i(h_j+1)p_i p_j}{(h_j+1)^2(p_i - p_i^2) + h_i^2(p_j - p_j^2) + 2h_i(h_j+1)p_i p_j}.$$

If we want all the inequalities to hold, we need to take the minimum of these distances, as stated in theorem 1.