

Using Bayesian Analysis to Predict Election Results

Samuel Martineau

Abstract

This paper introduces a method to predict election results in a given constituency from partial ballots counts data, such as that available during an election night, the goal being to reproduce the predictions made by large news agencies. The model built throughout this paper is based upon the ideas of Bayesian analysis, and is compared against real-world data cumulated from the 2019 and 2021 Canadian federal elections, as well as two provincial elections from 2022.

Contents

1. Building the Model	1
1.1. Probability of probabilities	2
1.2. Beta Distribution	2
1.3. Building the Likelihood Function	3
1.4. Building Prior Beliefs	4
1.5. Combining Prior Beliefs and Likelihood	5
1.6. Comparing Probability Distributions	7
1.7. Considering the Number of Votes Left	8
2. Analyzing the Model	12
2.1. Collecting Real World Data	12
2.2. Analyzing the Data	12
2.3. Evaluating the Model	13
3. Conclusion	16
4. Bibliography	17
Appendices	19
A. Collected Data	19

List of Tables

1. Predictions from first and second model	12
2. Predictions from first, second and third model	14
3. Values of S tested with the number of random points used and the average error	15

List of Figures

1. A few beta distributions	3
2. Plot of the likelihood function for the leading candidate	4
3. Plot of the probability distribution built from prior knowledge	5
4. The set of distributions D	7
5. The CDFs of the distributions D for all but the leading candidate	7
6. The product of the CDFs of the distributions D for all but the leading candidate	7
7. Probability that the leading candidate at any given share of the votes	7
8. The set of distributions E	9
9. The set of distributions E_{tc}	11
10. Plot of the collected data	12
11. Plot of the model from Section 1.7 with 835 random points	13
12. Plot of the error in the model for $S = 100$	15
13. Plot of the model for $S = 251$	15
14. Plot of the error in the model for $S = 251$	16

1. Building the Model

As with any mathematical problem, a considerable portion of building the model is simply to lay down the assumptions and to split the task into multiple, more specific, problems. To approach this using the tools of conditional probability, understanding why predicting election results even involves random events is a must. The fundamental assumption needed here, from which all of the mathematics will follow, is that each individual casting its vote can be considered an independent random event were the different possibilities are the different candidates in the constituency, with each candidate having a different probability of receiving a vote.

Essentially, the probability that a voter will vote for a given candidate is the final proportion of votes that this candidate will have received in the final results. Furthermore, each vote would be independent of the other ones, because election results are not shown until every polling booth is closed.¹

However, before going any further, the basics of the electoral system of interest shall be outlined. The model will be based on the “first-past-the-post” election system used in provincial and federal elections in Canada. The Canadian electoral systems generally work in the following way:

1. The territory is divided in smaller districts of similar size in terms of population called *constituencies*.
2. During the elections, electors can go cast a vote for their single favourite candidate in their constituency. Each vote will go in a *box*. Each constituency has multiple boxes of an approximately fixed number of votes.
3. Once all the votes have been gathered, the counts start to be released. This phase can take multiple hours, due to the long process of counting every vote.
4. The results are released box by box.

¹For federal elections, due to the large timezone differences, the results of some of the Eastern provinces are compiled before polls close in some of the Western provinces. However, there is, overall, very little overlap.

Then, a few variables need to be defined. Let n be the number of candidates in the constituency.

Let $v = \{v_1, v_2, v_3, \dots, v_n\}$ be the set of the current vote counts for the different candidates, ordered from largest to smallest, where v_1 is the number of votes for candidate 1, v_2 is the number of votes for candidate 2, etc. And let $v_t = \sum_{i=1}^n v_i$ be the total number of votes.

Also, let b_c be the number of ballot boxes counted and b_t be the total number of ballot boxes.

The number of votes left to be counted will also be relevant (if only a few votes are left to be counted, the probability of the lead candidate being elected will be much higher), but it is not a number known in advance. However, it can be approximated by assuming that the number of votes per ballot box is roughly constant. Therefore, let $v_e = \frac{b_t}{b_c} v_t$ be the expected end total number of votes, and let $v_l = v_e - v_t$ be the expected number of votes left to count.

In general, when discussing a certain candidate, it will be referred to as the k th-candidate. For example, the candidate k is considered to currently have v_k votes.

Working with conditional probability, beliefs about the probability each candidate has to win will be most often represented by probability distributions. This idea will be detailed below, notably in Section 1.1.

In this paper, the first goal will be to represent the likelihood of observing the current evidence (the current number of votes) as a function (Section 1.3) and to represent the prior beliefs (what is thought before observing any data about the chances that each candidate has to win) as a probability distribution (Section 1.4). It will then be possible to combine those two pieces of information through the use of Bayes' theorem, which will give a probability distribution representing the probability that a certain candidate will have a certain share of the final votes, assuming the election contains infinitely many votes (Section 1.5). Finally, using this and the number of votes left to be counted, it will be possible to generate a probability distribution representing the expected final number of votes for a given candidate (Section 1.7). This

will give all the information required to compute the probability that each of the candidates has to win over the others.

Therefore, let $D = \{D_1, D_2, D_3, \dots, D_n\}$ be the list of the unknown probability distributions representing the probability that a certain candidate will have a certain share of the votes, where D_1 is the probability distribution for the candidate 1, D_2 for the candidate 2, etc.

Finally, $E = \{E_1, E_2, E_3, \dots, E_n\}$ will represent the list of probability distributions for the final expected number of votes, where E_1 is the distribution for the candidate 1, E_2 for the candidate 2, etc.

Although the sets D and E may look quite cryptic for now, their meaning and utility will become much clearer through the rest of this paper.

Due to the usefulness of specific, visual examples when trying to investigate probability questions, the following variables will be used as a simple and concrete example throughout this paper:

$$\begin{aligned} n &= 5 \\ v &= \{60, 50, 36, 34, 20\} \\ v_t &= 60 + 50 + 36 + 34 + 20 = 200 \\ b_c &= 10 \\ b_t &= 16 \\ v_e &= \frac{16}{10}(200) = 320 \\ v_l &= 320 - 200 = 120 \end{aligned}$$

This is a hypothetical five-candidate election (n), where the leading candidate currently has 60 votes (v_1). Out of the 16 boxes in the constituency (b_t), 10 have been opened (b_c), which allows to predict that there will be around 320 votes in the end (v_e), based on the 200 currently counted votes (v_t).

Although this set of data will be used for numerical and graphical example, this paper will not focus on the computation of specific numerical examples, as the endgoal is to have a generalized computer model. Furthermore, due to their nature, many of the computations discussed here have no analytical solutions, thus computer based approximations will be used.

1.1. Probability of probabilities

A recurrent theme in this paper will be the idea of *probability of probabilities*, an idea which is at the root of many advanced concepts in conditional probability. Below is an example exploring this concept.

Considering a biased coin whose mathematical weight (bias) is unknown, after observing 90 heads and 10 tails out of 100 trials, what should one expect the bias to be?

One might argue that the answer is trivial: to find the weight, divide the number of observed heads (or tails) by the number of throws. This goes with the idea of the *Law of large numbers* [25], that the more trials there are, the more the observed frequency will approach the theoretical (the real) probability.

However, this reasoning is flawed. Yes, $\frac{90}{100} = 0.9$ is the most likely probability, but it is possible that the *true* probability is 0.1, 0.99 or any other value between 0 and 1, exclusively. An event being unlikely does not mean it is impossible.

The better approach is therefore to use probability distributions: instead of trying to define the weight of the coin with a single number, it can be defined as a probability distribution that represents how likely each of the infinitely many possible values of the bias are. That probability distribution would most likely be a beta distribution, which is explored below.

1.2. Beta Distribution

Two reasons make the beta distribution ideal for representing probability of probabilities: its domain is $[0, 1]$ and the area under a beta distribution's Probability Density Function (PDF) over its range, as with any valid probability distribution, is 1. This means that any value on the x -axis represents a possible probability and that the y -value of the distribution at that point represents the probability density that this probability is the true one.

Furthermore, the beta distribution can take a variety of shapes, as its PDF is, most commonly, defined in terms of two shape parameters, α and β , both being positive non-null real numbers. Let X be a distribu-

tion $X \sim \text{Be}(\alpha, \beta)$, where Be is the beta distribution:

$$P(X = x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathcal{B}(\alpha, \beta)}, x \in [0, 1]$$

In the definition of the PDF of the beta distribution, \mathcal{B} is the beta function [20]. Dividing by the beta function has the effect of scaling the numerator in order to make the area under the beta distribution's PDF equal to 1. It is therefore equal to the integral of the numerator:

$$\mathcal{B} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

However, it is more commonly defined as follows, where Γ is the gamma function [21]:

$$\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

This distribution would have a mean of [20]:

$$E(X) = \mu_X = \frac{\alpha}{\alpha + \beta}$$

Finally, the gamma function can be viewed as an expansion of the factorials to the reals (except for integers smaller or equal to 0) while respecting the following identity [23], n being a positive integer²:

$$\Gamma(n) = (n - 1)!$$

The beta distribution will be referred to as $\text{Be}(\alpha, \beta)$ throughout this paper. Here are a few beta distributions plotted, demonstrating some of the various shapes it can take:

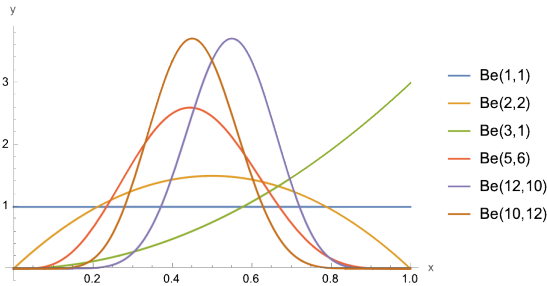


Figure 1: A few beta distributions

In Figure 1, multiple interesting things can be seen,

²A more detailed explanation of the gamma function has been deemed outside of the scope of this investigation.

notably that a $\text{Be}(1, 1)$ distribution is equivalent to a $\text{Uniform}(0, 1)$ distribution [27]³ and that the beta distribution can be both symmetric and highly asymmetric about the average.

Finally, the Cumulative Distribution Function [17] (CDF) of a beta distribution is the regularized beta function [26], notated $\mathcal{I}(z; a, b)$, which is in itself expressed in terms of the incomplete beta function [24], notated $\mathcal{B}(z; a, b)$.⁴

$$P(A \leq z) = \mathcal{I}(z; \alpha, \beta) = \frac{\mathcal{B}(z; \alpha, \beta)}{\mathcal{B}(\alpha, \beta)}$$

1.3. Building the Likelihood Function

The first step is to figure out the probability distribution representing the share of votes each candidate has.

Seeing this from the perspective of each of the candidates, it can be considered that the number of votes received over the total number of votes is a binomial experiment, where a *success* is defined as a vote for that candidate, and a *failure* as a vote given to any other. As a reminder, the Probability Mass Function [10] (PMF), the discrete analogue of the PDF [10], for a binomial distribution Y , $Y \sim \text{B}(m, p)$ ⁵, would be the following, where p is the probability of the event happening and m is the total number of trials:

$$P(Y = x) = \binom{m}{x} p^x (1-p)^{m-x}, x \in \{0, 1, 2, \dots, m\}$$

In the case of this paper, both the number of successful trials, v_k (the current number of votes for the candidate) and the total number of trials, v_t (the current total number of votes) are known. This means that, for the candidate k , with the number of votes v_k , the unknown left is the probability, here p , of receiving a vote distributed from the unknown distribution D_k , D_k being the distribution representing the probability that the candidate will receive the next

³A uniform distribution is a distribution where all values in a given interval (in this case, $[0, 1]$) are equally likely.

⁴A deeper exploration of the regularized and incomplete beta functions not being relevant to the rest of the mathematics, they will not be explained in greater details in this paper.

⁵Here, m is used instead of the typical n in order to avoid confusion with the number of candidates in the constituency.

vote. The above equation can therefore be rewritten as $V_k \sim B(v_t, p)$.

$$P(V_k = v_k \mid D_k = p) = \binom{v_t}{v_k} p^{v_k} (1 - p)^{v_t - v_k}$$

However, as the distribution V_k is not really important, it could also be represented as

$$P(v_k \mid D_k = p) = \binom{v_t}{v_k} p^{v_k} (1 - p)^{v_t - v_k}$$

This answers the question: *What is the probability of observing the evidence v_k given that $D_k = p$?*

As what is really of interest is the unknown distribution D_k , let $L_{D_k}(p)$ represent its likelihood function [19], which will answer the question: *Based solely on the evidence, how likely is it that a certain value of the probability p is the true probability that led to the observed events?*

$$\begin{aligned} L_{D_k}(p) &= P(v_k \mid D_k = p) \\ &= \binom{v_t}{v_k} p^{v_k} (1 - p)^{v_t - v_k} \end{aligned}$$

Here is the plot of this function for the leading candidate ($k = 1$) in the above example, considering it currently has $v_k = v_1 = 60$ votes and that the total number of votes is currently $v_t = 200$:

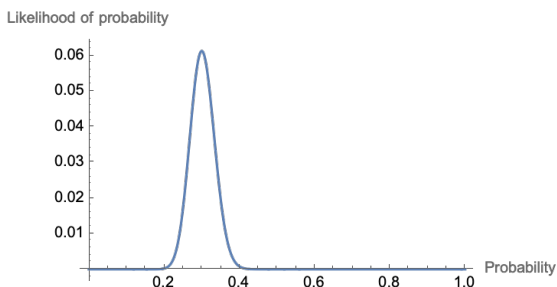


Figure 2: Plot of the likelihood function for the leading candidate

Referring back to Section 1.1, this is an example of a probability distribution representing an unknown probability. One should, however, still expect the mode of the distribution (the maximum of the likelihood function) to be the simple frequency calculation $\frac{v_1}{v_t} = \frac{60}{200} = 0.3$, which can be verified in Figure 2.

However, a key element is still missing before it can be said that this function represents the probability

distribution of the share of the votes a given candidate has, as the prior beliefs [19] are yet to be considered.

1.4. Building Prior Beliefs

Prior beliefs, as the name implies, is what is believed to be the probability distribution before seeing any evidence (the partial election results, in this context). It is expressed in the form of a probability distribution. In the context of elections, there are two ways to approach this: prior ignorance and substantial prior knowledge [7].

Prior ignorance is quite trivial: it is assumed that nothing is known before the election. Therefore, a distribution illustrating that all probabilities are considered to be equally likely is needed. This is the perfect use for the uniform distribution, so it could be said that the prior beliefs about the probability distribution of the share of the votes of a given candidate (D_k) follows a Uniform(0, 1) distribution, also known as a Be(1, 1) distribution.

Substantial prior knowledge unfortunately is not as simple. Commonly, it is considered to be “[when] expert opinion, for example, gives us good reason to believe that some values in a permissible range for $[p]$ are more likely to occur than others”. [6] In this context, expert opinions could be the polls from firms like LÉGER, who usually publish their predictions a few weeks before any major election. An example of such a report could be LÉGER’s *ÉLECTIONS PROVINCIALES : MONTRÉAL ET LAVAL* [8], which contains two key pieces of information:

- The voting intentions (what percentage of people plan to vote for each of the parties).
- The firmness of the intentions (for each party, what percentage of people do not expect to change their minds).

For example, suppose it was known from a report that 35% of the citizens intended to vote for a given party, and that 45% of those people are quite firm about their decision, how could this be transformed into a probability distribution? For the reasons outlined in Section 1.2, it seems reasonable to try building a beta distribution. Therefore, let $U, U \sim \text{Be}(\alpha, \beta)$,

be the prior beliefs distribution about the candidate’s probability of receiving a vote.

First, the expected value (the mean) of the distribution is known to be 35 % (0.35). Then, “quite firm” could be defined as being at $\pm 5\%$ of the mean. The probability of landing in that range must therefore be equal to 45 % (0.45). This is equivalent to stating that the area under the PDF of the distribution in the range $[0.30, 0.40]$ should be equal to 0.45. Here is a system of equation combining both of these facts:

$$\begin{aligned} 0.35 &= E(U) \\ &= \mu_U \\ &= \frac{\alpha}{\alpha + \beta} \end{aligned}$$

And

$$\begin{aligned} 0.45 &= \int_{0.30}^{0.40} P(U = x) dx \\ &= \int_{0.30}^{0.40} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathcal{B}(\alpha, \beta)} dx \end{aligned}$$

To satisfy the requirements of the beta function, both α and β need to be non-null positive reals. As there is no trivial analytical solution to this system of equations, the simplest solution is to resort to numerical approximation to solve for α and β . It is to be noted that this system of equations may not always yield a solution when considering extreme requirements, like having a exceedingly small margin around the mean for the definition of “quite firm”. This, however, is not really an issue as these cases would lead to such intense certain prior beliefs that any evidence would hardly be relevant.

Using WOLFRAM MATHEMATICA [29] or similar software, the solution to this system can be computed to be $\alpha \approx 11.485$ and $\beta \approx 21.330$. This yields the following probability distribution as the prior beliefs:

It is important to keep in mind that this process is quite subjective. In fact, it was here chosen to define “quite firm” as being $\pm 5\%$ of the mean, but it could have chosen to be $\pm 7\%$, $\pm 3\%$, or any other value. This is the main weakness of this process: biases can easily sneak into the statistics if one is not careful.

As the prior beliefs can be represented as a beta

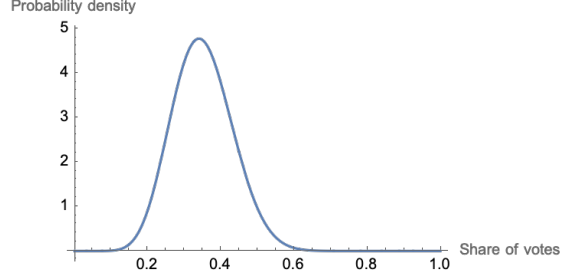


Figure 3: Plot of the probability distribution built from prior knowledge

distribution both in the case of prior ignorance and of prior substantial knowledge, it makes sense to define the prior beliefs for the candidate k as $D_k \sim \text{Be}(a_k, b_k)$ before seeing any of the evidence. For the rest of this paper, all of the prior knowledge about the candidate k will be referred to with the variables a_k and b_k shaping this distribution. The PDF of D_k could therefore be written as

$$P(D_k = p) = \frac{p^{a_k-1}(1-p)^{b_k-1}}{\mathcal{B}(a_k, b_k)}$$

1.5. Combining Prior Beliefs and Likelihood

Now that the prior beliefs and the likelihood function are both formulated, it is time to combine them into the probability distribution for the candidate’s share of the total votes.

This is where Bayes’ theorem comes in. In fact, this theorem gives a systematic method to mix prior beliefs and observed evidence (summarized into the likelihood function) into posterior beliefs. As a reminder, here is the formula for said theorem [18], where A and B are independent random events

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

$P(A | B)$ This represents the *posterior beliefs* about A , considering that B was observed.

$P(B | A)$ This represents the *likelihood* that A happens given the observed evidence for B .

$P(A)$ This represents the *prior beliefs* about A .

$P(B)$ This represents the total probability of B . Essentially, this has the effect of scaling the proba-

bility of $A|B$ such that it lands between 0 and 1. In the case of probability distributions, this ensures that the area under the distribution's curve equals 1 [5].

It is also interesting to note that $P(B | A)$ and $P(A)$ can not only be probabilities, but also probability distributions, making $P(A | B)$ into one too.

As $P(B)$ is simply a scaling constant, the formula can be rewritten as

$$P(A | B) \propto P(B | A)P(A)$$

Which is also commonly known as the fact that [5]:

$$\text{posterior beliefs} \propto \text{likelihood} \times \text{prior beliefs}$$

The beauty of this lies in how clearly it highlights how evidence (likelihood) does not *replace* prior beliefs, but rather *updates* them to form posterior beliefs [15].

Going back to the context of elections, this statement could be rewritten as such:

$$P(D_k = p | v_k) \propto P(v_k | D_k = p)P(D_k = p)$$

$P(D_k = p | v_k)$ This is the target probability distribution D_k (as a function of p).

$P(v_k | D_k = p)$ This is the likelihood function that was derived earlier, $L_{D_k}(p)$.

$P(D_k = p)$ This is the prior beliefs distribution that was derived earlier.

Substituting in allows to compute an expression proportional to $P(D_k = p | v_k)$.

$$\begin{aligned} P(D_k = p | v_k) &\propto P(v_k | D_k = p)P(D_k = p) \\ &\propto \left(\binom{v_t}{v_k} p^{v_k} (1-p)^{v_t-v_k} \right) \\ &\quad \left(\frac{p^{a_k-1} (1-p)^{b_k-1}}{\mathcal{B}(a_k, b_k)} \right) \\ &\propto p^{v_k} (1-p)^{v_t-v_k} \\ &\quad (p^{a_k-1} (1-p)^{b_k-1}) \\ &\propto p^{v_k+a_k-1} (1-p)^{v_t-v_k+b_k-1} \end{aligned}$$

There are three things to notice and recall here: (I) As this distribution represents possible values of a probability p , its domain is $[0, 1]$. (II) As with any

other continuous probability distribution, its area over its range (here, $[0, 1]$) must be equal to 1. (III) The beta distribution matches both the form of the obtained equation and the above two criteria.

Finding the beta distribution corresponding to the above equation is simply a question of identifying the values of its shape parameters. In a beta distribution $\text{Be}(\alpha, \beta)$ whose PDF is expressed as a function of x , x is raised to the power of $\alpha - 1$ and $1 - x$ is raised to the power of $\beta - 1$. Applying this to the above equation, where the distribution's PDF is expressed as a function of p , yields the following coefficients and, therefore, the following distribution:

$$\alpha - 1 = v_k + a_k - 1$$

$$\alpha = v_k + a_k$$

And

$$\beta - 1 = v_t - v_k + b_k - 1$$

$$\beta = v_t - v_k + b_k$$

Therefore

$$D_k | v_k \sim \text{Be}(v_k + a_k, v_t - v_k + b_k)$$

Sadly, as detailed polls for elections dating back multiple years are not trivial to find, prior ignorance will have to be assumed when evaluating the model against real-world data. Remembering that prior ignorance can be represented as a $\text{Be}(1, 1)$ distribution, both a_k and b_k would be equal to 1 in this scenario. The following expression therefore represents the posterior beliefs when one lacks substantial prior knowledge.

$$D_k | v_k \sim \text{Be}(v_k + 1, v_t - v_k + 1)$$

As a reminder, D_k is the distribution representing the probability that the candidate k will receive the next vote, which is equivalent to the share of votes it would get if the election was to run infinitely.

For the sake of visual understanding, here are the computed probability distributions for each of the candidates in the example.

It is interesting to note that both the prior and pos-

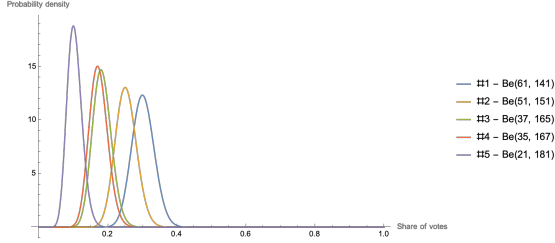


Figure 4: The set of distributions D

terior beliefs are beta distribution when the likelihood function comes from a binomial distribution; thus the beta distribution is a conjugate prior for the binomial distribution [9].

1.6. Comparing Probability Distributions

In Figure 4, it can be seen that, just as one would expect, the more votes a candidate currently has, the more likely it is to have a larger share of the votes. For example, the candidate with the most votes, candidate #1, is associated with the rightmost distribution, while the candidate with the least votes, candidate #5, is associated with the leftmost distribution.

However, this is not the same as the probability that each candidate has to win. For now, it will be assumed that elections are infinite and that winning means having the greatest share of votes in the long run.⁶

This would mean that a candidate's probability of winning is the probability that its probability distribution for the share of votes (D_k) is "bigger" than all the other candidates' distributions. But how exactly could "bigger" be quantified? For the following steps, visual examples will be crucial, thus the leading candidate will be used as an example.

First, consider the probability that a candidate k will have less than a certain share r of the votes, $P(D_k \leq r)$ ⁷. Plotting this for all candidates excluding the leading one yields Figure 5.

Since all of the distributions originate from independent events, the probability that all these four distributions will be smaller than r can be found by

⁶This assumption will be revisited in Section 1.7.

⁷When working with continuous distributions, $P(D_k \leq r)$ is equivalent to $P(D_k < r)$.

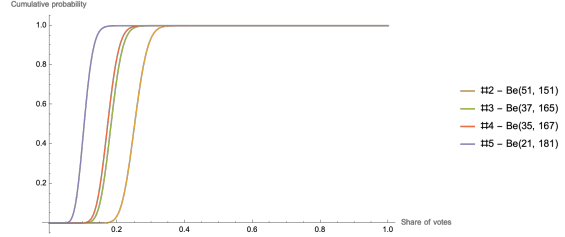


Figure 5: The CDFs of the distributions D for all but the leading candidate

merely multiplying them together. Plotting this leads to Figure 6.

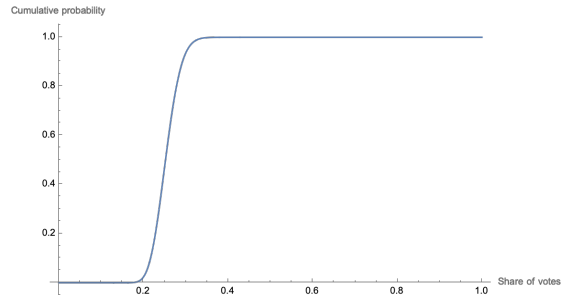


Figure 6: The product of the CDFs of the distributions D for all but the leading candidate

From the distribution of the leading candidate, D_1 , the probability that it will have some share r of the votes is known. Therefore, considering the independence of the events, the probability that all other candidates will have a share of the votes smaller than r (as shown in Figure 6), *and* that the leading candidate will have that share of the votes (D_1 's PDF evaluated at r) can be found by simple multiplication.

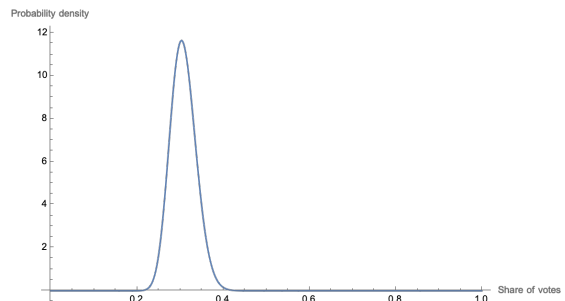


Figure 7: Probability that the leading candidate at any given share of the votes

Finally, the total probability that the leading candidate will have a larger share of votes than all the other candidates can be obtained by calculating the

area under this curve over the course of its domain. This would indicate that the leading candidate has a probability of approximately 0.86658 of winning. Performing the calculations for all the candidates yields approximately the following results: (1) 0.86658 (2) 0.13183 (3) 0.00012 (4) 0.00004 (5) 0.00000.

A straightforward verification that can be performed to ensure the mathematical reasoning was not blatantly incorrect is to add the above numbers and verify they sum to 1, as it is known that a candidate will be elected (the probability of any candidate being elected is the sum of the probability of each candidate to be elected), which they do.⁸ In other words, the probability that a candidate will win is mutually exclusive and complementary to the probability that any of the other candidates will.

These steps can be summarized in a more general form, assuming the search for the probability that a candidate k will win. First, the probability that all other candidates would have a share smaller than r of the votes was multiplied.

$$\prod_{\substack{i=1 \\ i \neq k}}^n P(D_i \leq r)$$

Next, that expression was multiplied by the probability density that the candidate k would have that share r of the votes.

$$P(D_k = r) \prod_{\substack{i=1 \\ i \neq k}}^n P(D_i \leq r)$$

Finally, the area under the curve was computed.

$$\int_{-\infty}^{\infty} P(D_k = r) \prod_{\substack{i=1 \\ i \neq k}}^n P(D_i \leq r) dr$$

However, given that D_k is a beta distribution, $P(D_k = r)$ is null for all values outside of the interval $[0, 1]$, thus the bounds of the integral can be

⁸Adding the numbers displayed here results in finding 1.00001 as the sum instead. This deviation is merely due to the numbers being calculated with more significant figures than displayed here.

limited.

$$\int_0^1 P(D_k = r) \prod_{\substack{i=1 \\ i \neq k}}^n P(D_i \leq r) dr$$

More generally, the following is the formula for calculating the probability that a certain probability distribution X_k will have a greater value than all other distributions in the set X , containing n elements, considering the PDF of the distribution X_k has non-zero values only in the interval $[a, b]$. This expression was largely inspired from *What is $P(X_1 > X_2, X_1 > X_3, \dots, X_1 > X_n)$?* [28]⁹.

$$P\left(\bigcap_{i=1}^n X_k \geq X_i\right) = \int_a^b P(X_k = x) \prod_{\substack{j=1 \\ j \neq k}}^n P(X_j \leq x) dx$$

It is to be noted that there is no analytical solution to the above equations for sets of distributions that contain more than two elements [28]. Therefore, numerical integration will be needed in order to find the probability that a certain candidate will win.

1.7. Considering the Number of Votes Left

Up to this point, it has been assumed that there is some sort of infinite election where a candidate wins if the distribution of their share of the votes in the long run is larger than that of all other candidates. However, in a real-world election, there is a fixed number of votes, a fact which needs to be taken into account to accurately model the situation.

The first thing that needs to be known is the probability that a certain candidate will gain a certain number of votes over the number of votes left, v_l . As one may notice, this looks quite a bit like a binomial experiment: (I) there is a fix number of trials (the number of votes left) (II) there are only two possible states for each trial (*success* being the candidate gaining a vote and *failure* being another candidate gaining it) (III) each trial has the same probability of having

⁹Although it originally came from a mathematics discussion forum, I believe I have provided a sufficient justification for this formula.

a specific outcome.

The only problem is that the probability of gaining a vote cannot be formulated directly, but rather elicited as a probability distribution, $D_k | v_k$ (for the candidate k), as shown above. Although this may seem like an issue, it actually is not. What needs to be done is to combine the binomial distribution described above to the probability distribution $D_k | v_k$ into a combined *predictive distribution*. In this case, because a beta distribution and a binomial distribution are being combined, the output distribution will be a beta-binomial distribution [16], notated here $\text{BetaBin}(\alpha, \beta, m)$, where α and β are the shape parameters of the underlying beta distribution and m is the number of trials¹⁰.

The following demonstration of the combination of both distributions is a more detailed version of the one included in *Bayesian Statistics, Simulation and Software — The Beta-Binomial Distribution* [1]. The first step is to find the *simultaneous distribution* of the beta and binomial distributions. This means weighing the binomial distribution, $X \sim B(m, p)$, as a function of the probability p , by the probability that the beta distribution, $Y \sim \text{Be}(\alpha, \beta)$, will equal p . This process is extremely similar to what was done when trying to form posterior beliefs from a binomial likelihood and a beta prior.

$$\begin{aligned} P(X = x | Y = p) &= P(X = x)P(Y = p) \\ &= \left(\binom{n}{x} p^x (1-p)^{n-x} \right) \\ &\quad \left(\frac{p^{\alpha-1} (1-p)^{\beta-1}}{\mathcal{B}(\alpha, \beta)} \right) \\ &= \frac{\binom{n}{x}}{\mathcal{B}(\alpha, \beta)} p^{x+\alpha-1} (1-p)^{n-x+\beta-1} \end{aligned}$$

Then, the predictive distribution, the distribution of interest, can be found by integrating the above over the range of p , $[0, 1]$.

$$P(X = x) = \int_0^1 \frac{\binom{n}{x}}{\mathcal{B}(\alpha, \beta)} p^{x+\alpha-1} (1-p)^{n-x+\beta-1} dp$$

¹⁰Here, m is used instead of the typical n in order to avoid confusion with the number of candidates in the constituency.

$$= \frac{\binom{n}{x}}{\mathcal{B}(\alpha, \beta)} \int_0^1 p^{x+\alpha-1} (1-p)^{n-x+\beta-1} dp$$

One may recognize from Section 1.2 that the leftover integral is the denominator of the PDF of a beta distribution $\text{Be}(x + \alpha, n - x + \beta)$, which can be expressed in terms of the beta function, as follows:

$$\begin{aligned} P(X = x) &= \frac{\binom{n}{x}}{\mathcal{B}(\alpha, \beta)} \int_0^1 p^{x+\alpha-1} (1-p)^{n-x+\beta-1} dp \\ &= \frac{\binom{n}{x}}{\mathcal{B}(\alpha, \beta)} \mathcal{B}(x + \alpha, n - x + \beta) \\ &= \binom{n}{x} \frac{\mathcal{B}(x + \alpha, n - x + \beta)}{\mathcal{B}(\alpha, \beta)} \end{aligned}$$

Considering this, an expression can now be found for the probability distribution $E_k | v_k$ that the candidate k will receive a certain number of votes over the rest of the counting process, using v_l as the number of trials and the parameters from $D_k | v_k$, the probability distribution for the k th-candidate to receive the next vote, for the underlying beta distribution.

$$E_k | v_k \sim \text{BetaBin}(v_k + a_k, v_l - v_k + b_k, v_l)$$

Plotting this distribution for each of the candidates gives us Figure 8.

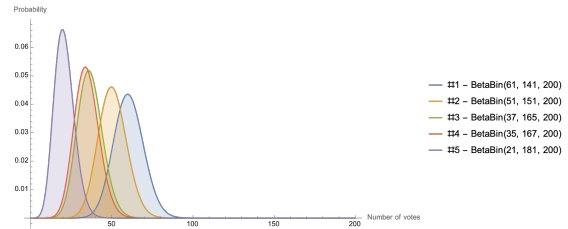


Figure 8: The set of distributions E

Carrying forward, the PDF of the distribution $E_k | v_k$ will be notated using functional notation to facilitate the representation of the operations that will be done on it.

$$\begin{aligned} E_k(x) &= \binom{v_l}{x} \frac{\mathcal{B}(x + v_k + a_k, v_l - x + v_t - v_k + b_k)}{\mathcal{B}(v_k + a_k, v_t - v_k + b_k)} \\ &= \binom{v_l}{x} \frac{\mathcal{B}(x + v_k + a_k, v_e - x - v_k + b_k)}{\mathcal{B}(v_k + a_k, v_t - v_k + b_k)} \end{aligned}$$

Comparing these probability distributions, however, would not be the full story. In fact, what needs to be taken into account is not only the number of votes

each candidate is expected to get, but also the current number of votes of each candidate. This can be done by translating the above function to the right by the candidate's current number of votes, v_k . The set of the translated distributions will be referred to as E_t and the distribution of the candidate k as E_{tk} .

$$\begin{aligned} E_{tk}(x) &= E_k(x - v_k) \\ &= \binom{v_l}{(x - v_k)} \frac{\mathcal{B}((x - v_k) + v_k + a_k, v_e - (x - v_k) - v_k + b_k)}{\mathcal{B}(v_k + a_k, v_t - v_k + b_k)} \\ &= \binom{v_l}{(x - v_k)} \frac{\mathcal{B}(x + a_k, v_e - x + b_k)}{\mathcal{B}(v_k + a_k, v_t - v_k + b_k)} \end{aligned}$$

An important fact to keep in mind is that E_k , and therefore E_{tk} , are discrete probability distributions. The problem with this is that discrete probability distributions are much harder to compute than continuous ones. This is because modern computational mathematics engines, like WOLFRAM MATHEMATICA [29] have many more tricks to optimize integrals (used in continuous distributions) than sums (used in discrete distributions). Furthermore, the formula derived in Section 1.6 to compare probability distributions is only built for continuous distributions.

The good news is that the beta-binomial distribution, $\text{BetaBin}(\alpha, \beta, n)$, can be computed for non-integer values, as all the functions and operations it depends on also are.

First, the choose function has a continuous expansion, which can be expressed as follows [22].

$$\binom{x}{y} = \begin{cases} 0 & y < 0 \\ \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)} & 0 \leq y \leq x \\ 0 & x < y \end{cases}$$

Although it is common not to set restrictions on this expression, they keep the function closer to its original meaning, which is useful in this concrete context, as the idea that it is impossible to have fewer than 0 votes or more than the maximum number of votes is still needed.

Second, the beta function is perfectly well defined for both integer and non-integer values, except for non-positive integers. However, when examining each of

the parameters of the beta functions in the expression, one can realize that they will never be nonpositive as long as the number of votes considered is between the current number of votes, v_k , and the maximum number of votes the candidate could get, $v_k + v_l$, keeping in mind that a_k and b_k will always be greater than 0, due to restrictions on the parameters of the beta function.

$\mathbf{x + a_k \leq 0}$ This implies that $x \leq -a_k$, but it makes no sense to consider the probability that a certain candidate will *lose* votes.

$\mathbf{v_e - x + b_k \leq 0}$ This implies that $x \geq v_e + b_k$. However, it does not make sense to consider the probability that a candidate will have more votes than are expected in the end for all candidates.

$\mathbf{v_k + a_k \leq 0}$ This implies that $v_k \leq -a_k$, but a candidate will always have a non-negative vote count.

$\mathbf{v_t - v_k + b_k \leq 0}$ This implies that $v_k \geq v_t + b_k$, but it is not possible for a candidate to have more votes than the total amount.

For impossible number of votes, the most logical thing is to define the function as having a value of 0, to indicate the impossibility of such an event happening.

The continuous version of E_{tk} and the continuous version of the set E_t will be respectively denoted E_{tck} and E_{tc} . This yields the following expression.

$$E_{tck}(x) = \begin{cases} 0 & x < v_k \\ \binom{v_l}{(x - v_k)} \frac{\mathcal{B}(x + a_k, v_e - x + b_k)}{\mathcal{B}(v_k + a_k, v_t - v_k + b_k)} & v_k \leq x \leq v_k + v_l \\ 0 & v_k + v_l < x \end{cases}$$

To consider $E_{tk}(x)$ for non-integer values of x , there is one last problem which needs to be fixed. Whereas continuous probability distributions use areas to determine probability, discrete ones use sums. This means that $E_{tck}(x)$ needs to be rescaled to ensure that the area under its PDF in the interval $[v_k, v_k + v_l]$ (the interval on which it is non-zero) is equal to 1. This can be achieved by dividing the function by its integral over that interval. For the sake of clarity, the following demonstration will assume $x \in [v_k, v_k + v_l]$ as it is the only part of the function which will be

affected by the rescaling.

$$\begin{aligned}
E_{tck}(x) &= \frac{E_{tk}(x)}{\int_{v_k}^{v_k+v_l} E_{tk}(t) dt} \\
&= \frac{\binom{v_l}{x-v_k} \mathcal{B}(x+a_k, v_e-x+b_k)}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \mathcal{B}(v_k+a_k, v_t-v_k+b_k)} dt \\
&= \frac{\binom{v_l}{x-v_k} \mathcal{B}(x+a_k, v_e-x+b_k)}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \mathcal{B}(t+a_k, v_e-t+b_k) dt} \\
&= \frac{\left(\frac{1}{\mathcal{B}(v_k+a_k, v_t-v_k+b_k)} \right)}{\left(\frac{1}{\mathcal{B}(v_k+a_k, v_t-v_k+b_k)} \right)} \\
&= \frac{\binom{v_l}{x-v_k} \mathcal{B}(x+a_k, v_e-x+b_k)}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \mathcal{B}(t+a_k, v_e-t+b_k) dt} \\
&= \frac{\binom{v_l}{x-v_k} \Gamma(x+a_k) \Gamma(v_e-x+b_k)}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \frac{\Gamma(t+a_k) \Gamma(v_e-t+b_k)}{\Gamma((t+a_k)+(v_e-t+b_k))} dt} \\
&= \frac{\binom{v_l}{x-v_k} \Gamma(x+a_k) \Gamma(v_e-x+b_k)}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \Gamma(t+a_k) \Gamma(v_e-t+b_k) dt} \\
&= \frac{\left(\frac{1}{\Gamma(v_e+a_k+b_k)} \right)}{\left(\frac{1}{\Gamma(v_e+a_k+b_k)} \right)} \\
&= \frac{\binom{v_l}{x-v_k} \Gamma(x+a_k) \Gamma(v_e-x+b_k)}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \Gamma(t+a_k) \Gamma(v_e-t+b_k) dt}
\end{aligned}$$

As a reminder, v_k is the number of votes of the candidate k , with a_k and b_k being the parameters of the beta distribution representing the prior beliefs about its share of the votes.

Keeping in mind the domain restrictions on the above expression, the following is the actual function:

$$E_{tck}(x) = \begin{cases} 0 & x < v_k \\ \frac{\binom{v_l}{x-v_k} \Gamma(x+a_k) \Gamma(v_e-x+b_k)}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \Gamma(t+a_k) \Gamma(v_e-t+b_k) dt} & v_k \leq x \leq v_k + v_l \\ 0 & v_k + v_l < x \end{cases}$$

Plotting this continuous and translated set of distributions gives us Figure 9.

The distributions have now been translated by the candidates' current vote counts. Furthermore, just as one would expect, there is very little difference in the shape of each distribution, because the discrete plots already had so many points that they looked

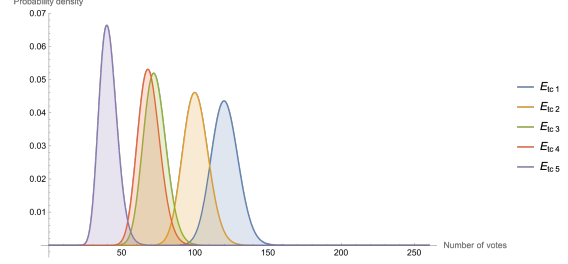


Figure 9: The set of distributions E_{tc}

continuous. The only noticeable change is the scale, due to the rescaling done above.

This is finally a set of continuous probability distributions taking into account the current vote counts and the number of votes left to be counted. However, before using the formula derived in Section 1.6, the CDF of E_{tck} also needs to be known.

Once again, the only relevant interval is $[v_k, v_k + v_l]$, as the cumulative probability of having less than the current number of votes is 0 and the cumulative probability of having more than the possible number of votes is 1.

$$\begin{aligned}
P(E_{tck} \leq x) &= \int_{v_k}^x E_{tck}(r) dr \\
&= \int_{v_k}^x \frac{\binom{v_l}{r-v_k} \Gamma(r+a_k) \Gamma(v_e-r+b_k)}{\left(\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \Gamma(t+a_k) \Gamma(v_e-t+b_k) dt \right)} dr \\
&= \frac{\int_{v_k}^x \binom{v_l}{r-v_k} \Gamma(r+a_k) \Gamma(v_e-r+b_k) dr}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \Gamma(t+a_k) \Gamma(v_e-t+b_k) dt}
\end{aligned}$$

Including the restrictions, the full definition of the CDF of E_{tck} would therefore be the following:

$$P(E_{tck} \leq x) = \begin{cases} 0 & x < v_k \\ \frac{\int_{v_k}^x \binom{v_l}{r-v_k} \Gamma(r+a_k) \Gamma(v_e-r+b_k) dr}{\int_{v_k}^{v_k+v_l} \binom{v_l}{t-v_k} \Gamma(t+a_k) \Gamma(v_e-t+b_k) dt} & v_k \leq x \leq v_k + v_l \\ 1 & v_k + v_l < x \end{cases}$$

Remembering the equation from Section 1.6, it is now possible to replace the terms with the expressions found in this section.

$$P\left(\bigcap_{i=1}^n E_{tck} \geq E_{tci}\right) = \int_{v_k}^{v_k+v_l} P(E_{tck} = x) \prod_{\substack{j=1 \\ j \neq k}}^n P(E_{tcj} \leq x) dx$$

Using this formula in the context of the example yields the following predictions.

Table 1: Predictions from first and second model

Candidate #	Section 1.6	Section 1.7
1	0.86658	0.96604
2	0.13183	0.03395
3	0.00012	0.00000
4	0.00004	0.00000
5	0.00000	0.00000

It can be observed that the predictions shift significantly once the number of remaining votes is taken into account. The probability of the first candidate winning is augmented by approximately 10 percentage points, consequently reducing the probability of other candidates' victory. This is plausible, considering the leading candidate not only has a greater probability of gaining a vote than its rivals, but also because it does not need to catch up to anyone.

2. Analyzing the Model

2.1. Collecting Real World Data

To analyze the accuracy of the model, comparing it to past real world data is a must. However, as the required data points are partial results (while the ballots are still being counted), there is very little publicly available data. Fortunately, Radio-Canada has public archives of all broadcast election nights from the last few years.

By using optical character recognition software, data from the following elections has been gathered.

- Canada (Federal), 2019; *Sources:* [2], [11]
- Canada (Federal), 2021; *Sources:* [2], [12]
- Ontario (Provincial), 2022; *Sources:* [3], [13]
- Quebec (Provincial), 2022; *Sources:* [4], [14]

For each constituency of each election, the on-screen data about current ballot counts, as well as the number of boxes counted versus the total number of boxes in

the constituency was recorded. This was then cross-referenced with public records to identify the final winner in each case. This data was then combined into the dataset, spanning 603 rows about 228 distinct constituencies, available in Appendix A.

2.2. Analyzing the Data

To compare the statistical model to real-world data, a plot showing the probability of being elected based on the empirical data will be quite useful. However, it is impossible to show all the useful dimensions of the dataset (the vote count for each of the candidates and the percentage of votes counted) in a single plot, as this would require a 7-dimensional graph in a five-candidate election (6 for the independent variables and 1 for the dependent variable). This problem can be solved by using the percentage of votes counted and the percentage lead of the leading candidate as axes, as these embedded many of the other axes within them.

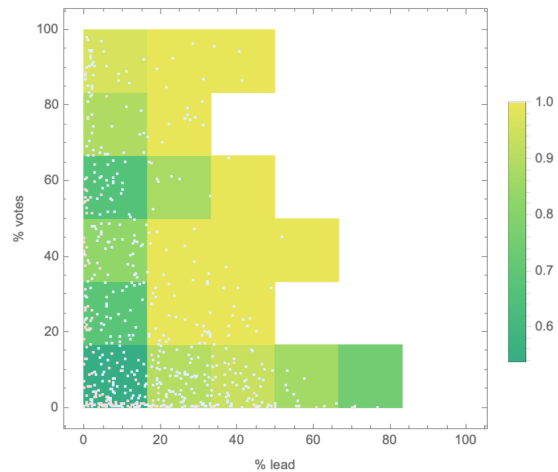


Figure 10: Plot of the collected data

Figure 10 shows exactly this. It was built by first plotting all 603 data points on a plot with the axes described above. These points were then coloured based on whether or not the leading candidate was elected in the end (blue if elected, red if not). The axes

were then separated into 6 segments each, creating 36 bins. Finally, the bins were coloured based on the ratio of blue points (situations where the lead won) over the total number of points (total number of situations). For example, if in the upper left bin, there are 28 points, with only one red. This means that out of 28 observed situations with 0% to 10% of votes counted and 90% to 100% lead, only once did the leading candidate not win. The probability of the leading candidate winning if the situation is in that bin is therefore $\frac{27}{28} \approx 0.9643$, which means the bin will be yellow. The cells that do not contain any points were left white. This makes this plot a two-dimensional histogram of the probability of a lead candidate winning if it lands in a specific bin.

However, it is important to keep in mind to keep in mind that the axes used here are not a direct representation of the original data. The representation taking only values derived from the raw data into account, the plot assumes that all the other factors average out. Therefore, it is only reliable when many data points are in each bin, which explains why there is some random variation in the colours of the graph. This random variation introduces a source of error when working with the data: the size of the bins (derived from the number of bins) can change the observed trends. The number 36 was chosen here as a tradeoff between having enough bins to observe trends, while having each bin contain quite a few points.

As one can see, for very low percentages of votes counted, there is quite a bit of random variation in the probability of being elected. However, as the percentage of votes and the percentage of lead increases, the probability of the lead being elected increases. This is represented by the graph being more and more yellow towards the top-right corner.

2.3. Evaluating the Model

Due to the time-cost of the expression found in Section 1.7, plotting it in a continuous manner is not really feasible, especially when taking into account that it would need to be averaged over many other factors.

Therefore, another method is required to visualize it. The chosen solution was to generate random points in the 6-dimensional space (a number of votes for each of the five candidates and the final total number of votes), feeding them through the function found above and finally graphing them in the same fashion as the real world data in Figure 10. The only difference in the graph is that the bins were coloured based on the average of the points they were containing, as each point already represented the probability for the lead candidate to be elected.

The total number of votes was generated based on the normal approximation of the total number of votes per constituency from dataset (mean of 43 476 and standard deviation of 13 106), truncated to a reasonable range, 8000 to 80 000.

The principal downside to using random points is that it allows for some random variation in the graphs, which is why the graphs below were made with as many points as possible.

Plotting the graph described above yields Figure 11.

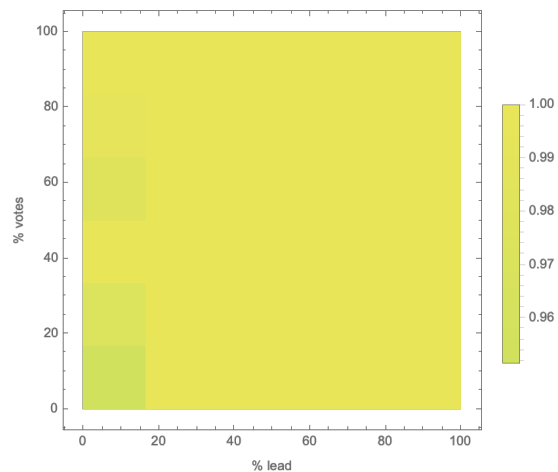


Figure 11: Plot of the model from Section 1.7 with 835 random points

Sadly, Figure 11 shines the light on an important issue: the model is over confident compared to what should be expected based on real-world data. For example, the model predicts a 0.95 probability of winning even when only 0% to 16.67% of votes were counted and the leading candidate only had 0% to 16.67% of lead. This means that each vote fed the model carries too much certainty.

The simplest fix for this would therefore be to scale down the number of votes given to the model by a certain scaling constant, S , in order to diminish their importance. The rationale for this probably lies in the fact that each vote was assumed to be completely independent from all others, even though this is probably not the case in real-life, where many factors influence the relation between different votes.

Examples of this may include: (I) opinions varying between different geographic or demographics parts of the constituency (II) herd mentality taking place (III) individuals trying to account for the failures of the first-past-the-post voting system (not voting for their favourite candidate in order to prevent a candidate they dislike from getting into office), although this is probably more a humanities question than a mathematical one.

Applying this fix to the model is fortunately quite trivial. In fact, the only adjustment required is to divide each value of the set of votes per candidate v by the scaling constant S before calculating the total number of votes v_t , the number of expected votes v_e and the expected number of votes left to be counted v_l .

For example, using the example data from Section 1 and a scaling constant of $S = 10$ would yield the following values.

$$\begin{aligned}
 n &= 5 \\
 v &= \left\{ \frac{60}{10}, \frac{50}{10}, \frac{36}{10}, \frac{34}{10}, \frac{20}{10} \right\} \\
 &= \{6, 5, 3.6, 3.4, 2, 1\} \\
 v_t &= 6 + 5 + 3.6 + 3.4 + 2 = 20 \\
 b_c &= 10 \\
 b_t &= 16 \\
 v_e &= \frac{16}{10}(200) = 32 \\
 v_l &= 32 - 20 = 12
 \end{aligned}$$

As justified earlier in Section 1.7, it is perfectly valid to use non-integer values in the function, as it does not rely on any integer-only functions or operations.

With this scaling back of $S = 10$, the model would

now yield the following probabilities as its predictions.

Table 2: Predictions from first, second and third model

Candidate #	Section 1.6	Section 1.7	Section 2
1	0.86658	0.96604	0.66200
2	0.13183	0.03395	0.25740
3	0.00012	0.00000	0.04492
4	0.00004	0.00000	0.03321
5	0.00000	0.00000	0.00246

The scaled-down model produces probabilities much closer to each other. The model calculated a much smaller probability for the first candidate to win and a much larger one for all the others.

The next step would therefore be to find the value of S that maximizes the accuracy of the model¹¹. First, a metric defining how *good* a certain value of S is will be required. A sufficient way to evaluate this could be to generate a plot of the model for a certain value of S , and look at the difference, in each bin, between the calculated probability for the leading candidate winning and the real world probability from the equivalent bin. Then, the average of the absolute value of these differences could be used as the metric for the value of S . The goal would then simply be to find the value of S that minimizes this average error. This can be visualized in Figure 12.

In Figure 12, one can see how the error varies bin per bin, from approximately 0.00 at 33.33% to 50% of lead and 83.33% to 100% of votes counted up to approximately 0.25 at 0% to 16.67% of lead and 50% to 66.67% of votes counted. Calculating the average of the different bins in this plot would yield an average error of approximately 0.0596.

However, it is important to realize that this graph is susceptible to quite a few sources of error¹²:

1. The real world data probably contains quite a few anomalies due to the relatively small dataset gathered (approximately 600 points divided in

¹¹For the sake of brevity, the following steps will be a simple attempt at optimizing this parameter. However, a more rigorous and complete working of the optimal value would make a most interesting extension to this paper.

¹²Although a quantitative way to handle these error sources would be most helpful, such a thing has been deemed outside of the scope of this paper.

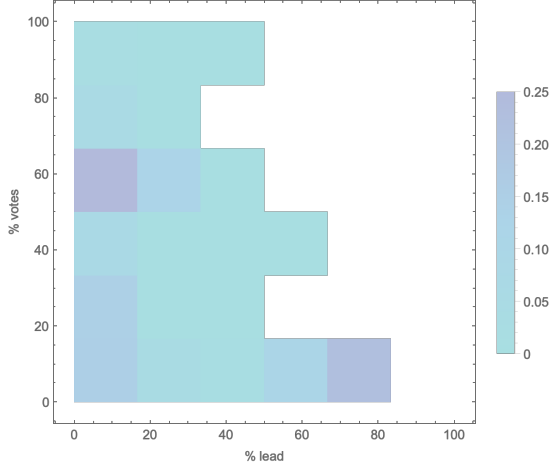


Figure 12: Plot of the error in the model for $S = 100$

36 bins only leaves about 16 points per bin, with some having much less). For example, the bin at 0% to 16.67% of lead and 50% to 66.67% in the real world data plot (Figure 10) does seem to have an abnormally low probability of the leading candidate being elected compared to its neighbours.

2. This also means that changing the number of bins would probably change the average error in the plot due to point moving over boundaries.
3. The model plot being generated from random points, it is also somewhat susceptible to random error.

Calculating the average error for some values of S gives the following results¹³:

Table 3: Values of S tested with the number of random points used and the average error

Value of S	Number of random points	Average error
1	835	0.0999
3	2865	0.0972
100	5000	0.0596
251	5000	0.0449
376	5000	0.0498
500	5000	0.0495
544	5000	0.0542
587	5000	0.0550
1000	5000	0.0818

¹³Smaller values of S have less random points, as they are much more expensive to compute.

Those values were selected quite randomly within a reasonable range of values for S (1 to 1000), while using an approximate binary-search inspired algorithm (starting at the extremes of the reasonable range of values and recursively testing values in their middle). The point $S = 251$ was also selected, as it is approximately the average number of votes per ballot box in the collected dataset.

Out of these points, $S = 251$ seems to be the optimal value, having the smallest average error. This seems to indicate that even though individual votes are not truly independent, individual ballot boxes can be taken to be. This gives an outcome really quite close to the real world data, with an average error of only (approximately) 0.0449. This version of the model and its error can be visualized in Figures 13 and 14.

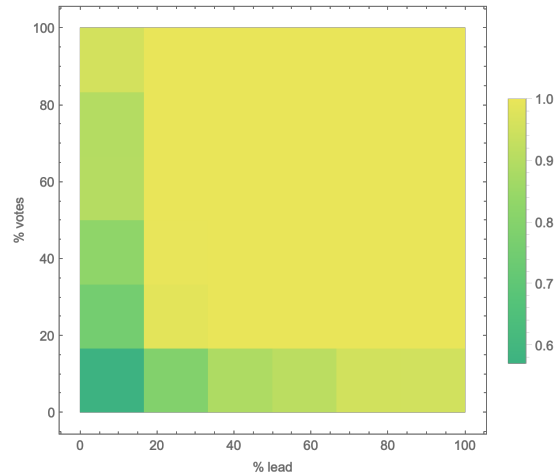


Figure 13: Plot of the model for $S = 251$

As one can see, the mathematical model with $S = 251$ produces an output (Figure 13) really quite similar to the real-world data (Figure 10), to the exceptions of some anomalies.

This is in terms shown in Figure 14 by a mostly blue graph, indicating a very small per bin difference and, therefore, a very small average difference.

Figure 13 also shows the expected general trend: the more votes are counted and the more lead the leading candidate has, the higher are its chances of winning. One can also observe other, more specific, trends, such as the fact that the probability of being

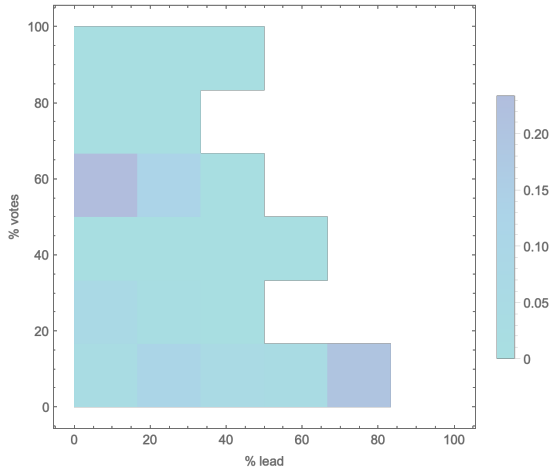


Figure 14: Plot of the error in the model for $S = 251$

elects gets really quite close to 1 as soon as more than approximately 16.67% of votes are counted and that there is more than 16.67% lead.

3. Conclusion

In conclusion, thanks to the tools of conditional probability, it was possible to build a mathematical model to calculate the probability that a certain candidate in a given constituency has to win. To do so, a likelihood function summarizing the probability of observing the current evidence (the number of votes for each candidate) was first built. This was followed by the construction of prior beliefs using prior elicitation to summarize what was thought about each candidate before observing any election results, based on, for example, survey data. It was then possible to combine those two pieces of information using Bayes' theorem to obtain a probability distribution representing the probability that a certain candidate had a certain probability of gaining the next vote. Using a translated beta-binomial distribution, it was then possible to find the final expected number of votes for each candidate, which could then be compared to find the probability that a certain candidate had to win. Finally, it was observed that scaling back the number of votes, by a factor found to be near 251, was required in order to make the model's output much closer to the collected real world data. This led to an average error of less than 0.045.

4. Bibliography

- [1] “Bayesian Statistics, Simulation and Software — The beta-binomial distribution.” (May 28, 2005), [Online]. Available: <https://people.math.aau.dk/~slb/kurser/r-11/betabin.pdf> (visited on 12/22/2022).
- [2] “Elections Canada.” (n.d.), [Online]. Available: <https://www.elections.ca/> (visited on 12/17/2022).
- [3] “Elections Ontario.” (n.d.), [Online]. Available: <https://www.elections.on.ca/en.html> (visited on 12/17/2022).
- [4] “Élections Québec.” (n.d.), [Online]. Available: <https://www.electionsequbec.qc.ca/> (visited on 12/17/2022).
- [5] L. Fawcett. “MAS2317/3317: Introduction to Bayesian Statistics — Chapter 2: Bayes’ Theorem for Distributions — Lecture Notes.” (2015), [Online]. Available: <https://www.mas.ncl.ac.uk/~nlf8/teaching/mas2317/notes/chapter2.pdf> (visited on 12/18/2022).
- [6] L. Fawcett. “MAS2317/3317: Introduction to Bayesian Statistics — Chapter 3: Prior Elicitation — Lecture Notes.” (2015), [Online]. Available: <https://www.mas.ncl.ac.uk/~nlf8/teaching/mas2317/notes/chapter3.pdf> (visited on 12/18/2022).
- [7] L. Fawcett. “MAS2317/3317: Introduction to Bayesian Statistics — Chapter 3: Prior Elicitation — Slides.” (2015), [Online]. Available: <https://www.mas.ncl.ac.uk/~nlf8/teaching/mas2317/notes/slides3.pdf> (visited on 12/18/2022).
- [8] Léger. “Élections provinciales : Montréal et laval - 10 septembre 2022.” (Sep. 10, 2022), [Online]. Available: <https://leger360.com/fr/intentions-de-votes/elections-provinciales-montreal-et-laval-10-septembre-2022/> (visited on 12/19/2022).
- [9] J. Orloff and J. Bloom. “Beta Distributions.” (n.d.), [Online]. Available: <https://math.mit.edu/~dav/05.dir/class14-prep-a.pdf> (visited on 12/22/2022).
- [10] PennState: Statistics Online Courses. “7.2 - Probability Mass Functions — STAT 414.” (n.d.), [Online]. Available: <https://online.stat.psu.edu/stat414/lesson/7/7.2> (visited on 12/18/2022).
- [11] Radio-Canada Info. “Revoyez la soirée électorale fédérale 2019.” (Oct. 21, 2019), [Online]. Available: <https://www.youtube.com/watch?v=D3tqBHh5IjY> (visited on 12/17/2022).
- [12] Radio-Canada Info. “Soirée électorale 2021 au Canada.” (Sep. 20, 2021), [Online]. Available: <https://www.youtube.com/watch?v=E3PKiPiwW8o> (visited on 12/17/2022).
- [13] Radio-Canada Info. “Élections en Ontario 2022 : La soirée électorale.” (Jun. 2, 2022), [Online]. Available: <https://www.youtube.com/watch?v=JNHZ4hzuN90> (visited on 12/17/2022).
- [14] Radio-Canada Info. “Élections Québec 2022 : La soirée électorale.” (Oct. 3, 2022), [Online]. Available: <https://www.youtube.com/watch?v=EkV28Wew-1Q> (visited on 12/17/2022).
- [15] Veritasium. “How To Update Your Beliefs Systematically - Bayes’ Theorem.” (Apr. 5, 2017), [Online]. Available: <https://www.youtube.com/watch?v=R13BD8qKeTg> (visited on 12/23/2022).
- [16] E. W. Weisstein. “Beta Binomial Distribution.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/BetaBinomialDistribution.html> (visited on 12/23/2022).
- [17] E. W. Weisstein. “Distribution function.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/DistributionFunction.html> (visited on 12/21/2022).

- [18] E. W. Weisstein. “Bayes’ theorem.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/BayesTheorem.html> (visited on 12/19/2022).
- [19] E. W. Weisstein. “Bayesian analysis.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/BayesianAnalysis.html> (visited on 12/18/2022).
- [20] E. W. Weisstein. “Beta distribution.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/BetaDistribution.html> (visited on 12/18/2022).
- [21] E. W. Weisstein. “Beta function.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/BetaFunction.html> (visited on 12/18/2022).
- [22] E. W. Weisstein. “Binomial Coefficient.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/BinomialCoefficient.html> (visited on 01/05/2023).
- [23] E. W. Weisstein. “Gamma function.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/GammaFunction.html> (visited on 12/18/2022).
- [24] E. W. Weisstein. “Incomplete beta function.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/IncompleteBetaFunction.html> (visited on 12/21/2022).
- [25] E. W. Weisstein. “Law of Large Numbers.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/LawofLargeNumbers.html> (visited on 12/23/2022).
- [26] E. W. Weisstein. “Regularized beta function.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/RegularizedBetaFunction.html> (visited on 12/21/2022).
- [27] E. W. Weisstein. “Uniform distribution.” (n.d.), [Online]. Available: <https://mathworld.wolfram.com/UniformDistribution.html> (visited on 12/18/2022).
- [28] whuber. “What is $P(X_1 > X_2, X_1 > X_3, \dots, X_1 > X_n)$?” (Nov. 21, 2012), [Online]. Available: <https://stats.stackexchange.com/a/44142> (visited on 01/05/2023).
- [29] “Wolfram Mathematica: Modern Technical Computing.” (n.d.), [Online]. Available: <https://www.wolfram.com/mathematica/> (visited on 12/19/2022).

Constituency	Boxes Counted	Total Boxes	R-C Elected	First	First Count	Second	Second Count	Third	Third Count	Fourth	Fourth Count	Fifth	Fifth Count	Total Votes	End Winner	End Total Votes
RICHMOND ARTHABASKA	5	270	False	CPC	266	BQ	194	LPC	181	GPC	48	Other	0	689	CPC	58638
RICHMOND ARTHABASKA	25	270	False	CPC	1433	BQ	913	LPC	727	GPC	193	Other	0	3265	CPC	58638
RICHMOND ARTHABASKA	45	270	False	CPC	2979	BQ	1850	LPC	1152	GPC	369	Other	0	6350	CPC	58638
RIMOUSKI NEIGETTE	1	220	False	LPC	33	BQ	13	CPC	9	NDP	4	Other	0	59	BQ	45767
TEMISCOUATA LES BASQUES	145	220	False	BQ	10266	NDP	7719	LPC	5853	CPC	2123	Other	0	25961	BQ	45767
TEMISCOUATA LES BASQUES	150	227	False	BQ	11570	LPC	10360	NDP	2558	CPC	2278	Other	0	26766	BQ	58184
RIVIERE DES MILLE ILES	30	272	False	BQ	2021	LPC	1063	CPC	453	NDP	332	Other	0	3809	BQ	60101
RIVIERE DU NORD	90	272	False	BQ	7177	LPC	3087	CPC	1634	NDP	1109	Other	0	13007	BQ	60101
ROSEMONT LA PETITE PATRIE	1	223	False	LPC	30	BQ	5	NDP	4	Other	2	Other	0	41	NDP	60206
ROSEMONT LA PETITE PATRIE	11	223	False	NDP	1192	LPC	613	BQ	589	GPC	168	Other	0	2562	NDP	60206
ROSEMONT LA PETITE PATRIE	75	223	False	NDP	7697	LPC	4425	BQ	4291	GPC	1088	Other	0	17501	NDP	60206
SAANICH GULF ISLANDS	2	238	False	GPC	88	LPC	58	CPC	54	NDP	18	Other	0	218	GPC	68150
SAANICH GULF ISLANDS	35	238	False	GPC	2844	CPC	1245	LPC	1080	NDP	701	Other	0	5870	GPC	68150
SAINT HYACINTHE BAGOT	1	247	False	BQ	118	LPC	50	CPC	45	NDP	43	Other	0	256	BQ	55914
SAINT HYACINTHE BAGOT	55	247	False	BQ	4607	LPC	2363	NDP	1995	CPC	1708	Other	0	10673	BQ	55914
SAINT JEAN	80	256	False	BQ	12864	LPC	9097	CPC	3136	NDP	1952	Other	0	27049	BQ	61875
SAINT JOHN ROTHESAY	3	170	False	LPC	116	CPC	79	GPC	19	NDP	18	Other	0	232	LPC	41253
SAINT MAURICE CHAMPLAIN	5	281	False	LPC	238	BQ	144	CPC	55	PPC	14	Other	0	451	LPC	58414
SAINT MAURICE CHAMPLAIN	210	281	False	LPC	12930	BQ	11812	CPC	5629	NDP	1910	Other	0	32281	LPC	58414
SALABERRY SUROIT	90	286	False	BQ	13573	LPC	8830	CPC	2864	NDP	2126	Other	0	27393	BQ	62903
SHEFFORD	220	271	False	BQ	17216	LPC	16266	CPC	5348	NDP	2744	Other	0	41574	BQ	60913
SHEFFORD	268	271	False	BQ	22752	LPC	21647	CPC	7245	NDP	3559	Other	0	55203	BQ	60913
SHERBROOKE	95	261	False	LPC	4151	NDP	3607	BQ	3227	CPC	1307	Other	0	12292	LPC	59726
SHERBROOKE	145	261	False	LPC	6422	NDP	6210	BQ	5349	CPC	2056	Other	0	20037	LPC	59726
SHERBROOKE	204	261	False	NDP	10072	LPC	10014	BQ	8923	CPC	3462	Other	0	32471	LPC	59726
SHERBROOKE	220	261	False	LPC	11291	NDP	11105	BQ	9823	CPC	3830	Other	0	36049	LPC	59726
SHERBROOKE	255	261	False	LPC	15845	NDP	15338	BQ	14007	CPC	5689	Other	0	50879	LPC	59726
SOUTH SHORE ST. MARGARETS	5	260	False	CPC	170	LPC	153	NDP	52	GPC	25	Other	0	400	LPC	52518
SOUTH SHORE ST. MARGARETS	100	260	False	LPC	7974	CPC	5458	NDP	3081	GPC	2156	Other	0	18669	LPC	52518
ST. JOHN'S EST	1	182	False	NDP	50	CPC	41	LPC	29	GPC	1	Other	0	121	NDP	45072
ST. JOHN'S EST	15	182	False	NDP	1145	LPC	867	CPC	536	GPC	42	Other	0	2590	NDP	45072
ST. JOHN'S EST	50	182	False	NDP	4454	LPC	3148	CPC	1757	GPC	157	Other	0	9516	NDP	45072
ST. JOHN'S SUD MOUNT PEARL	29	185	False	LPC	2567	NDP	1650	CPC	786	GPC	90	Other	0	5093	LPC	40666
ST. JOHN'S SUD MOUNT PEARL	30	185	False	LPC	2816	NDP	1743	CPC	895	GPC	96	Other	0	5550	LPC	40666
SYDNEY VICTORIA	130	196	False	CPC	7193	LPC	7048	NDP	5053	Other	3962	Other	0	23256	LPC	40565
TOBIQUE MACTAGUAC	1	184	False	CPC	17	LPC	10	GPC	4	NDP	1	Other	0	32	CPC	38201
TOBIQUE MACTAGUAC	30	184	False	CPC	2273	LPC	819	GPC	460	NDP	261	Other	0	3813	CPC	38201
TORONTO CENTRE	95	257	False	LPC	7748	NDP	3261	CPC	1665	GPC	959	Other	0	13633	LPC	54512
TROIS RIVIERES	23	260	False	BQ	777	LPC	710	CPC	687	NDP	389	Other	0	2563	BQ	60538
TROIS RIVIERES	125	260	False	BQ	5983	LPC	4986	CPC	4624	NDP	2991	Other	0	17678	BQ	60538
TROIS RIVIERES	220	260	False	BQ	12871	LPC	12011	CPC	11554	NDP	7536	Other	0	43972	BQ	60538
UNIVERSITY ROSDALE	27	207	False	LPC	2534	CPC	925	NDP	856	GPC	357	Other	0	4672	LPC	57391
VANCOUVER GRANVILLE	8	205	False	LPC	189	Other	116	NDP	98	CPC	85	Other	0	488	Ind	53032
VANCOUVER GRANVILLE	50	205	False	Other	1959	CPC	1942	LPC	1836	NDP	1086	Other	0	6823	Ind	53032
VANCOUVER GRANVILLE	77	205	False	Other	3300	CPC	3149	LPC	3025	NDP	1675	Other	0	11149	Ind	53032
VANCOUVER GRANVILLE	175	205	False	Other	9749	LPC	8361	CPC	7100	NDP	4646	Other	0	29856	Ind	53032
WINNIPEG CENTRE	85	175	False	GPC	8606	NDP	4976	LPC	4474	CPC	2435	Other	0	20491	NDP	31724