Mathematics Internal Assessment

Using Bayesian Analysis to Predict Election Results

Samuel Martineau

December 19, 2022

Contents

1 Introduction **1**

2 Collecting Real-World Data **1**

3 Analyzing the Data **2**

4 Building the Model **3**
4.1 Probability of probabilities . . . 3
4.2 Understanding the Beta Distribution . . . 4
4.3 Building the Likelihood Function . . . 4
4.4 Building Prior Beliefs . . . 5
4.5 Combining Prior Beliefs and Likelihood . . . 6
4.6 Comparing Probability Distributions . . . 7
4.7 Considering the Number of Votes Left . . . 9

5 Analyzing the Model **11**

6 Conclusion **11**

7 Biblography **12**

1 Introduction

For as long as I can remember, I have been fascinated by politics, from the power dynamics that have shaped recent history to the magnificent system in which we live, a democracy. Although democracies are not without their flaws, particularly when we consider the current voting system used in Canada, they are inarguably the best political system ever created by mankind.

An highly interesting event that results from a democratic election is the night right after where the nation awaits for the final results, slowly receiving updates for the current ballots count for different constituencies. While this is happening, news agencies are trying to use their current data to predict the final results. This process of highly confidently predicting the final results of constantly updating data while trying to make that prediction as soon as possible has long been a source of interrogation for me. Impressively, news agencies are ridiculously fast at forming their predictions, like when Radio-Canada successfully predicted that the Coalition Avenir Québec would form a majority government less than 11 minutes after results started to come in for the Québec 2022 election [1]. Furthermore, although they occasionally make wrong predictions [2], this is exceedingly rare.

In short, I started to wonder about how news agencies could be so fast and so accurate. This paper will be my attempt at building a model to make electoral predictions, so that I can better understand the seemingly magical tools that they used. It is to be noted that my goal here is not to reverse engineer how existing systems work, as I do not have access to the same data that news agencies have. I will instead try to build a simple tool that would allow anyone to simply insert the current ballot counts in their constituency and see the probability that each of the candidates has to win.

The model will based on the "first-past-the-post" election system used in provincial and federal elections in Canada. Furthermore, to verify the accuracy of my model, I will need to compare it to past election data. The data I chose to collect was sampled from Quebecois, Ontarian and Canadian elections (at the provincial or federal level) from the past few years, since those are the elections I have most interacted with, as a Quebecer currently living in Ontario.

Out of all the possible ways to approach such a problem, the one I found the most interesting was to model the situation as a conditional probability problem, as it is a very theorical approach and I was curious to know if it could accurately represent the real world. Other approaches, such as regression or hypothesis testing, would be quite interesting extensions to this paper.

2 Collecting Real-World Data

Before trying to model the situation, we should first gather past data, so that we can test the model with real-world examples while developping it. As we are interested in the partial results of past elections (while the ballots are still being counted) instead of the final results, there is not much publicly available data. Fortunately, Radio-Canada has public archives of all the election nights they streamed on YouTube over the last few years.

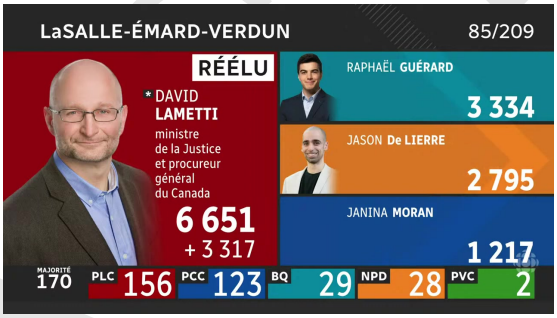

Figure 1: A sample frame from Radio-Canada's presentation of the 2021 federal election [3]

This means that we can look at every time a constituency was shown on screen, record the current ballot counts and, using public records, which of the candidates really won in the end. We should also note how many boxes have already been counted versus the total number of boxes in the constituency. Here are the elections I chose to gather data from:

- Canada (Federal), 2019; *Sources: [4], [5]*
- Canada (Federal), 2021; *Sources: [3], [5]*
- Ontario (Provincial), 2022; *Sources: [6], [7]*
- Quebec (Provincial), 2022; *Sources: [1], [8]*

At first, I attempted to collect the data by hand, with custom software to assist me in the menial task. However, I realized that this endeavour would know no ends and that I had to find a better solution. This lead me to fully automate the task using a mix of optical character recognition (OCR) and of color recognition. Although the OCR was not always perfect, my code had several failcheck to make sure the collected data was as reliable as possible. Here are a few caveats about the data collection:

- Only the candidates shown by Radio-Canada are counted. To match this, when looking up the end total vote count, only the top five candidates were considered.[1]

[1] From my observations, Radio-Canada never displays more

- The OCR could only capture the frames where the data was shown in full screen, which means not all data points were captured.

The dataset being unreasonably long for embedding in the present document, I have placed it, along with other relevent files for this project, on the platform GitHub at the following address:

`https://github.com/Samuel-Martineau/BayesianAnalysisElectionsPredictions`

3 Analyzing the Data

In the end, the full dataset is 603 rows long and contains data from 228 different constituencies. Here are a few interesting metrics from it vizualized:

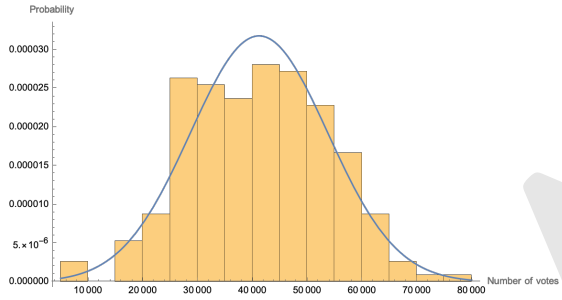

Figure 2: Distribution of the final total vote counts

In Figure 2, the collected data for the total amount of votes at the end of the election for each constituency is shown as an histogram in orange. Instead of showing the count, the histogram shows the probability of a data point landing in a given bin. In blue is a normal distribution with the mean and standard deviation of the data, showing that the end total vote count seems to be somewhat normally distributed.

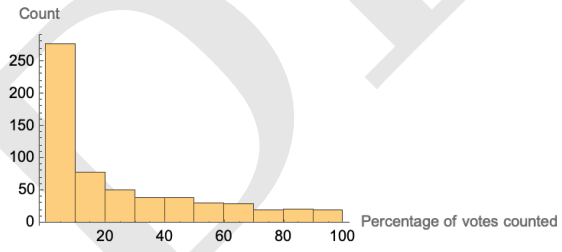

Figure 3: Distribution of the percentages of votes counted

In Figure 3, the collected data shows how the percentage of votes counted is distributed in an histogram. We can notice how the vast majority of the data shown by Radio-Canada is shown when not many votes have been counted.

To compare our statistical model to real-world data, a plot showing the probability of being elected based on the collected data will be quite useful. However, it is impossible to show all the useful dimensions of our data (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 (6 for the independent variables and 1 for the dependent variable). Therefore, we need a way to group some of these axes together. The solution I found to this problem is to use the percentage of votes counted and the percentage lead of the leading candidate as axes, as these are arguably the two main intuitive factors when trying to predict if the leading candidate will be elected.

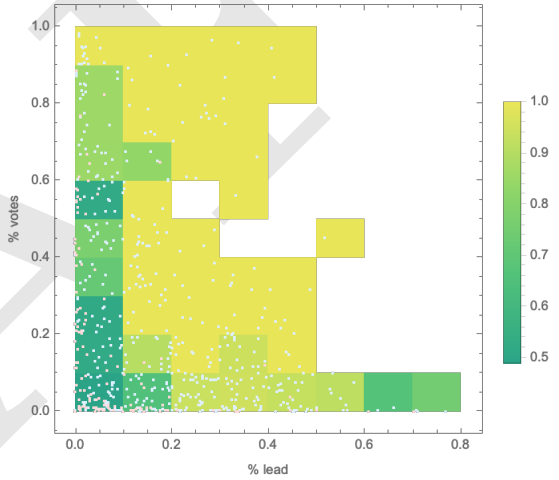

Figure 4: Plot of the collected data

In Figure 4, the plot described above, all 603 data points are placed on a plot with the axes described above. Their color depends on wether or not the candidate was elected in the end (blue if elected, red if not). Then, each axis is divided into 7 segments, separating the points into 49 bins. The bins are finally colored based on the ratio of blue points over the total number of points in them. So that they can be visually compared, all graphs of this type throughout this exploration will use the same colour scale. This essentially calculates the empirical probability of being elected for points inside that bin. However, we need to keep in mind that the axes used here are not a direct representation of our original data. Our representation taking only the relative difference of the first and second candidate into account, the plot assumes that all the other factors average out. Therefore, it is only reliable when many data points are in bin, which explains why there is some random variation in the

than the top five candidates. Furthermore, the candidates not shown by Radio-Canada probably have so little votes that they would have little to no impact on the final results.

colors of the graph.

As we would expect, we see that the higher the percentage lead and the percentage of votes counted are, the higher the probability of being elected is. However, this plots gives us valuable information about how does that trend behave at different points. For example, we can see that the probability of being elected is still quite high even when very little votes are counted and there is very little percentage lead.

4 Building the Model

As with any mathematical problem, a considerable portion of building the model is simply to lay down our assumptions and to split the task into multiple, more specific, problems. To approach this using the tools of conditional probability, we first need to understand why predicting elction results even involves random events. The fundamental assumption we need to do here, from which all of the mathematics will follow, is that we can consider each indivual casting its vote as 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.

Let's unpack this. Essentially, we can imagine that the probability that a voter will vote for a given candidate is the final proportion of votes that that candidate will have received in the final results. Furthermore, each vote would be independent of the other ones, because election results aren't shown until every polling booth is closed.[2]

Let's start by defining a few variables. Let n be the number of candidates in the constituency. Let $v = \{v_1, v_2, v_3, \ldots, v_n\}$ be the set of the current vote counts for the different candidates, ordered from largest to smallest. Also, let $v_t = \sum_{i=1}^{n} v_i$ be the total number of votes, let b_c be the number of ballot boxes counted, b_t be the total number of ballot boxes and let $l = \frac{v_1 - v_2}{v_t}$ be the relative lead of the first candidate. 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, we can approximate it by assuming 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. Also, let $D = \{D_1, D_2, D_3, \ldots, D_n\}$ be the list of the unknown probability distributions, we are searching for,represent the probability for each of the candidates to receive a certain vote and let $E = \{E_1, E_2, E_3, \ldots, E_n\}$ be the probability distributions representing the probability that a given candidate will receive a certain number of votes over the rest of the counting process. The elements of these lists, D and E, are unknowns we are searching for. The idea behind D is detailed in Section 4.1. Finally, let $w = \{w_1, w_2, w_3, \ldots, w_n\}$ be the set of probabilies that each candidate has to win.

Due to the usefulness of specific, visual examples when trying to investigate statistics, let's use the following variables as a simple and concrete example:

$$
\begin{aligned}
n &= 5 \\
v &= \{60, 50, 36, 34, 20\} \\
v_t &= 60 + 50 + 36 + 34 + 20 = 200 \\
b_c &= 10 \\
b_t &= 16 \\
l &= \frac{60 - 50}{200} = 0.05 \\
v_e &= \frac{16}{10}(200) = 320 \\
v_l &= 320 - 200 = 120
\end{aligned}
$$

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. The details of the numerical calculations will therefore be left as an exercise to the reader. Furthermore, due to their nature, many of the computations discussed here have no analytical solutions, which is why computer based approximations will be favoured.

4.1 Probability of probabilities

A reccurant theme in this paper will be the idea of *probability of probabilities*. Although this may seem like an utterly nonsensical statement at first, it is actually at the root of many advanced concepts in conditional probability. In order to explore this idea, let's use an example situation.

Considerin, a biased coin whose weight is unknown, after observing 90 heads and 10 tails out of 100 trials, what is the expected weight of the coin? One might argue that the answer is trivial: to find the weight, we divide the number of observed heads over the number of trials. This goes with the idea of the *Law of large numbers* [9] that the more trials are observed, the more the observed frequency will approach the theorical, the real, probability.

[2] For 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.

However, I would argue that this reasoning is flawed. Yes, $\frac{90}{100}$ is the most likely probability, considering only this piece of evidence, but it is not impossible that the *true* probability is $\frac{1}{100}$, $\frac{99}{100}$ or any other value between 0 and 1. 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, we can define 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 we will explore below.

4.2 Understanding the Beta Distribution

As we will heavily rely on it, it is important that we understand the beta distribution. The reason why it is so useful is that its domain is $[0, 1]$. As with any valid continous probability distribution, the area under a beta distribution's Probability Density Function (PDF) over its range is 1. Those two facts make it ideal for representing probability of probabilities. Furthermore, the beta distribution can take a variety of shapes, as its PDF is most commonly defined in terms of two parameters, α and β. It's definition is based on the beta function, here called $\mathcal{B}$ [10]. Let's define a distribution X such that $X \sim \text{Be}(\alpha, \beta)$, 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. 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. However, it is more commonly defined as follows, where Γ is the gamma function [11]:

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

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 [12], n being a positive integer, (a more detailed explanation of the gamma function has been deemed outside of the scope of this investigation):

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

The beta distribution will be refered 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:

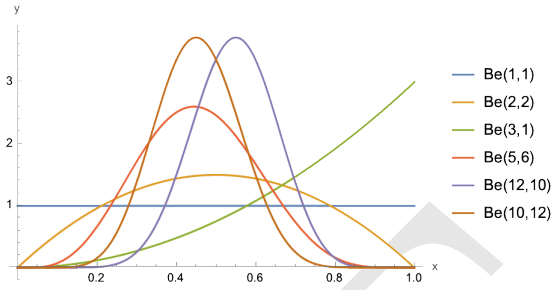

Figure 5: A few beta distributions

In Figure 5, we can see multiple interesting things, notably that a $\text{Be}(1, 1)$ is equivalent to a $\text{Uniform}(0, 1)$ distribution [13][3] and that the beta distribution can be both symmetric and highly asymmetric.

The mean of a distribution $A \sim \text{Be}(\alpha, \beta)$ is [10]:

$$E(A) = \mu_A = \frac{\alpha}{\alpha + \beta}$$

And the variance of the same distribution is [10]:

$$Var(A) = \sigma_A^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Finally, the Cumulative Distribution Function [14] (CDF) of a beta distribution is the regularized beta function [15], notated $\mathcal{I}(z; a, b)$, which is in itself expressed in terms of the incomplete beta function [16], notated $\mathcal{B}(z; a, b)$.[4]

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

Now that we understand the beta distribution, we can go back to building the model.

4.3 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 persepective of each of the candidates, we can consider the number of votes received over the total number of votes as 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 [17] (PMF), the discrete analogue of the PDF [17], for a binomial distribution Y, $Y \sim \text{B}(n, p)$, would be

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

[4] A deeper exploration of the regularized and incomplete beta functions not being relevant to the rest of the mathematics, I will not explore them in greater details.

the following, where p is the probability of the event happening and n is the total number of trials:

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

In our case, we know 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). This means that, for candidate k, with number of votes v_k, the unknown left is the probability, here p, distributed from the unknown distribution D_k. As with any other probability, p lies between 0 and 1. We can therefore rewrite the above equation by building a binomial distribution $V_k \sim \mathrm{B}(v_t, p)$.

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

As what really interests us is the unknown distribution D_k, we can rewrite this as its likelihood function [18], $L_{D_k}(p)$, 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 lead to the observed events? As with any probability, p must lie between 0 and 1.

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

Here is the plot of this function for the leading candidate ($k = 1$) in our example (using $v_k = v_1 = 60$ and $v_t = 200$):

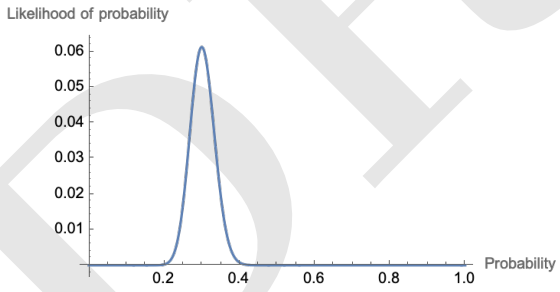

Figure 6: Plot of the likelihood function for the leading candidate

Referring back to Section 4.1, this is an example of a probability distribution representing an unknown probability. We should however still expect the mode of our distribution, its maximum, to be the simple frequency calculation $\frac{v_1}{v_t} = \frac{60}{200} = 0.3$, which we can verify in Figure 6.

However, we are still missing a key element before being able to say that this function represents the probability distribution of the share of the votes a given candidate has, as we still need to consider our prior beliefs [18].

4.4 Building Prior Beliefs

Our prior beliefs, as the name implies, is what we believe the probability distribution to be before seeing the evidence (the partial election results, in our context). We express it in the form of a probability distribution. In our context, there are two ways we can approach this: prior ignorance and substantial prior knowledge [19]. This process of quantifying our prior beliefs is often referred to as prior elicitation [20].

Prior ignorance is really quite easy: we assume we know nothing before the election. Therefore, we need a distribution illustrating that we consider all probabilities to be equally likely. This is the perfect use for the uniform distribution, so we would say that our prior beliefs about the probability distribution of the share of the votes a given candidate follows a Uniform$(0, 1)$ distribution (also known as a Be$(1, 1)$ distribution).

Substantial prior knowledge is quite a bit less trivial. First, let's define exactly what it means. Commonly, we will say we have substantial prior knowledge "[when] expert opinion, for example, gives us good reason to believe that some values in a permissable range for [p] are more likely to occur than others." [20] In our case, expert opinions could be the polls from firms like Léger, who usually publish there results a few weeks before any major election. An example of such a report could be Léger's *Élections provinciales : Montréal et Laval* [21], 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 don't expect to change their minds).

For example, suppose we knew 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 we transform this into a probability distribution? For the reasons outlined in Section 4.2, it seems reasonable to try building a beta distribution. Let's therefore define our prior beliefs distribution as $U \sim \mathrm{Be}(\alpha, \beta)$. First, we know that our expected value (the mean of the distribution) should be 0.35. Then we could define "quite firm" as saying that the probability of being at $\pm 5\%$ of the mean. The probability of landing in that range must therefore be equal to 0.45. This is equivalent to stating that the area under the PDF of our distribution in the

range $[0.30, 0.40]$ should be equal to 0.45. Let's write a system of equation using both of these facts:

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

And

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

As there is no trivial solution to this system of equation, the most efficient solution is to resort to numerical approximation to solve for α and β. Using Wolfram Mathematica [22] or similar software, we can find that this system is solved by $\alpha \approx 11.485$ and $\beta \approx 21.330$. This gives us the folowing probability distribution as our prior beliefs:

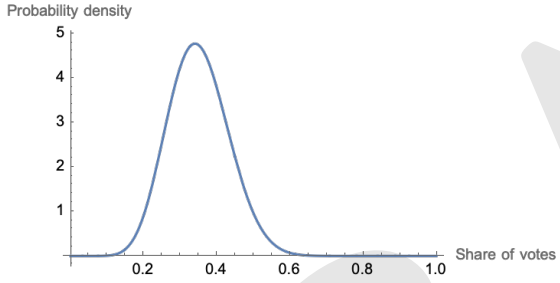

Figure 7: Plot of the probability distribution built from prior knowledge

It is important to keep in mind that this process is quite subjective. In fact, we chose to define "quite firm" as being ±5% of the mean, but we could have chosen ±7%, ±3% or any other value. This is the main weakness of this process: our biases can easily sneak into our statistics if we are not careful.

As our prior beliefs can be represented as a beta distribution no matter if we have prior ignorance or prior substantial knowledge, it makes sense to define our prior beliefs for the candidate k as $D_k \sim \mathrm{Be}(a, b)$ before seeing any of the evidence. Therefore, using proper notation, we can write our prior beliefs as follows:

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

4.5 Combining Prior Beliefs and Likelihood

Now that we know how to form our prior beliefs and our likelihood function, it is time to combine them into the probability distribution for the share of votes of a candidate.

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.[5] As a reminder, here is the formula for said theorem [23], where A and B are independent random events:

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

However, I dislike this depiction of Bayes' theorem as it abstracts and hides its true beauty. Exploring each of the terms leads us to the following:

$\boldsymbol{P(A \mid B)}$ This represents our *posterior beliefs* about A, considering that B happened.

$\boldsymbol{P(B \mid A)}$ This represents the *likelihood* that A happens given the observed evidence for B.

$\boldsymbol{P(A)}$ This represents our *prior beliefs* about A.

$\boldsymbol{P(B)}$ This represents the total probability of B. Essentially, this has the effect of scaling the probability of A 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 [24].

Given that $P(B)$ is simply a scaling constant and that there are much more eloquent ways to describe the other terms of the formula, I believe the following is a much more elegant way to describe Bayes' theorem [24]:

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

The beauty of this lies in how clearly it highlights how evidence (likelihood) doesn't replace our prior beliefs, but rather updates them to form our posterior beliefs. This is further illustrated by popular mathematics communicator Derek Muller, more commonly known for his YouTube channel *Veratisium*, and his description of the theorem as a way to "update [our] beliefs systematically" [25].

But how could we apply this to our variables? Let's rewrite this in terms of our variables and explore each of the terms, for $p \in [0, 1]$:

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

$\boldsymbol{P(D_k = p \mid v_k)}$ This is the probability distribution D_k as a function of p we are searching for.

$\boldsymbol{P(v_k \mid D_k = p)}$ This is the likelihood function we derived earlier, $L_{D_k}(p)$.

$\boldsymbol{P(D_k = p)}$ This is the prior beliefs distribution we derived earlier.

[5] A justification for Bayes' theorem has been deemed outside of the scope of this investigation.

As we can see, all of our work is really coming in together. Let's substitute the terms with our findings from the previous subsections.

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

There are three things to notice and recall here: (I) As this distribution represents possible values of a probability p, it's domain must be $[0, 1]$. (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 equation we obtained and the above two criterias.

Therefore, we know that this will end up as a beta distribution without needing to calculate $P(v_k)$.

Finding the beta distribution corresponding to our above equation is simply a question of identifying the values of the unknown parameters. In a beta distribution $\text{Be}(\alpha, \beta)$ 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 our example, where the distribution is expressed in function of p, we get the following coefficients and, therefore, the following distribution:

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

And

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

Therefore

$$D_k \sim \text{Be}(v_k + a, v_t - v_k + b)$$

Sadly, as detailed polls for elections dating back multiple years are not trivial to find, we will have to assume prior ignorance for the rest of this investigation. Remembering that prior ignorance can be represented as a $\text{Be}(1, 1)$ distribution, we know that both a and b would be equal to 1 in this scenario. The following expression therefore represents our posterior beliefs when we lack substantial prior knowledge.

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

The process of deriving prior beliefs, observing evidence to build a likelihood function and combining those two distributions together is commonly referred to as *bayesian analysis* [18], hence the name of this paper.

For the sake of visual understanding, let's visualize our findings for each of the candidates in our example.

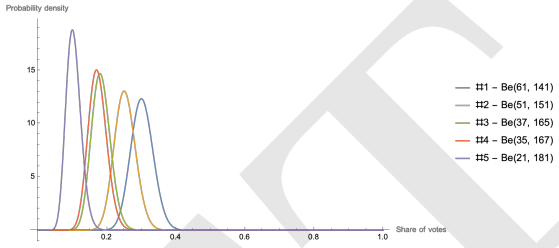

Figure 8: The set of distributions D

It is interesting to note that both our prior and posterior beliefs are beta distribution when the likelihood function comes from a binomial distribution. In bayesian analysis terminology, we would therefore say that the beta distribution is a conjugate prior for the binomial distribution [26].

4.6 Comparing Probability Distributions

In Figure 8, we can see that, just as we would expect, the more votes a candidate currently has, the more likely it is to have a larger share of the votes. However, we still don't have the concrete probability that each candidate has to win. For now, let's assume 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 to win is the probability that its probability distribution is "bigger" than all the other candidates' distributions. But what exactly does "bigger" mean here? And how could we quantify it? For the following steps, visual examples will be crucial. Let's use the leading candidate as our example.

First, let's consider the probability that some candidate k will have less than a certain share r of the votes, $P(D_k \leq r)$. Plotting this for all candidates except the leading one gives us Figure 9.

As all of our distributions come from independent events, we can find the probability that all distributions will be smaller than r by simply multiplying them together. Plotting this leads to Figure 10:

From the distribution of the leading candidate, D_1, we know the probability that it will have some share r of the votes. Therefore, keeping in mind we are working with independent events, we can find the probability that all other candidates will have a share

[6] This assumption will be revisited in Section 4.7.

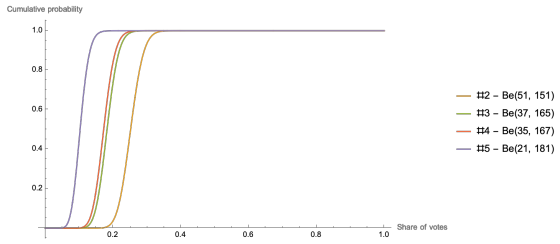

Figure 9: The CDFs of the distributions D for all but the leading candidate

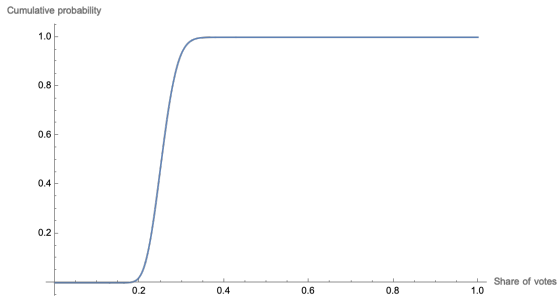

Figure 10: The product of the CDFs of the distributions D for all but the leading candidate

smaller than r (what we see in Figure 10) and that the leading candidate will have that share of the votes (D_1's PDF evaluated at r) by simply multiply them together.

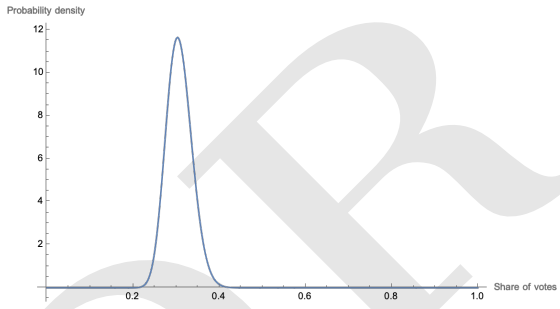

Figure 11: Probability that the leading candidate at any given share of the votes

Finally, we can get the total probability that the leading candidate will have a bigger share of votes than all the candidates by calculating the area under the above curve over the course of its range. This would give us that the leading candidate has approximately 86.658% chances of winning. Doing the calculations for all the candidates gives us approximately the following results, the set of probabilities w: (1) 86.658% (2) 13.183% (3) 0.0116% (4) 0.0044% (5) 0.0000%.

A simple verification we can do to ensure our mathematical reasoning was not blatantly wrong is simply to add the above numbers and verify they add up to 1, as we know that a candidate will be elected, which they do.[7] 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.

An important question left unanswered is why was the area under the curve not 1. Of course, we know intuitively that this couldn't be the case, but all other continous probability distributions encountered in this paper had an area of 1, leading to the question: What is different here? What they all had in common is that they considered *how* an even that we know will happen will happen. However, here, the candidate is not certain to win, which is why the total probability, the area under the curve, is less than one.

Let's summarize the steps we did in a more general form, assuming we are searching for the probability that a candidate k will win. First, we multiplied the probability that all other candidates would have a share smaller than r of the votes.

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

Then, we multiplied that expression by the probability 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, we took the area under the curve.

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

Furthermore, since D_k is a beta distribution, $P(D_k = r)$ is 0 for all values outside of the interval $[0, 1]$ and we can therefore limit the bounds of the integral.

$$\int_{0}^{1} P(D_k = r) \prod_{\substack{i=1\\i\neq k}}^{n} P(D_i \leq r)\,\mathrm{d}r$$

Therefore, 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

[7] Adding the numbers displayed here leads to finding 1.00001 as the sum instead. This deviation is simply due to the fact that the numbers were calculated with more significant figures than displayed here.

has non-zero values only in the interval $[a, b]$. This expression is largely inspired from [27][8].

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

It is to be noted that there is no analytical solution to the above equations for set of distributions that contain more than two elements [27]. Therefore, numerical integration will be needed in order to find the probability that a certain candidate will win.

4.7 Considering the Number of Votes Left

Up to here, we assumed some sort of infinite election where a candidate won if the distribution of his share of the total votes was bigger than the one of all the other candidates. However, in a real world election, there is a fix number of votes. But how could we take this into account?

What we first need to know is the probability that a certain candidate will gain a certain number of votes over the number of votes left, v_l. As we may notice, this looks quite a bit like a binomial experiment: (I) we have a fix number of trials (the number of votes left) (II) we have 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 a specific outcome.

The only problem is that we do not have a probability of gaining a vote, but rather a probability distribution. Although this may seem like an issue, it actually isn't. What we need to do is to combine them into a combined *predictive distribution*. In our case, because we have a beta distribution and a binomial distribution, the distribution we will obtain will be a beta-binomial distribution [28], notated here $\mathrm{BetaBin}(\alpha, \beta, n)$, where α and β are the parameters of the underlying beta distribution and n is the number of trial.

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* [29]. The first step is to find the *simultaneous distribution* of the beta and binomial distributions. This means weighing the binomial distribution, $X \sim \mathrm{B}(n, p)$, as a function of the probability p, by the probability that the beta distribution, $Y \sim \mathrm{Be}(\alpha, \beta)$, will equal p. This process is extremely similar to what we did when trying to form our posterior beliefs from a binomial likelihood and a beta prior.

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

Then, we can find the predictive distribution, what we are actually searching for, by integrating the above over the range of p, $[0, 1]$.

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

We may recognize from Section 4.2 that the integral we are left with is the denominator of the PDF of a beta distribution $\mathrm{Be}(x + \alpha, n - x + \beta)$, which can be expressed in terms of the beta function, as follows:

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

This expression is therefore the PDF of the $\mathrm{BetaBin}(\alpha, \beta, n)$ distribution. Considering this, we can now find an expression for the probability 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 for the underlying beta distribution.

$$E_k \mid v_k \sim \mathrm{BetaBin}(v_k + 1, v_t - v_k + 1, v_l)$$

Plotting this distribution for each of the candidates gives us the following:

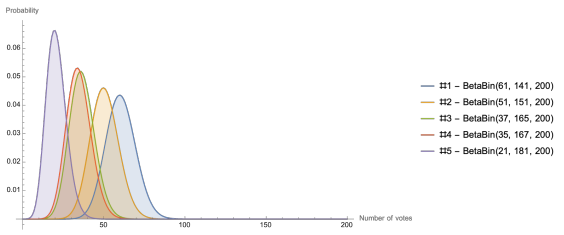

Figure 12: The set of distributions E

[8] Although it originally came from a mathematics discussion forum, I believe I have provided a sufficient justification for this formula

An important fact to keep in mind is that E_k is a discrete probability distribution. The problem with this is that discrete probability distribution are much harder to compute than continous ones. This is because modern computational mathematics engine, like Wolfram Mathematica [22] have many more tricks to optimize integrals (used in continuous distributions) than sums (used in discrete distributions). Furthermore, the formula derived in Section 4.6 to compare probability distributions is only built for continous distributions, which would mean we couldn't use it to compare our distributions for the expected number of votes to be gained.

The good news is that the beta-binomial distribution $\text{BetaBin}(\alpha, \beta, n)$ can be computed for non-integer values. In fact, both the choose function, $\binom{n}{x}$ and the beta function, $\mathcal{B}(\alpha, \beta)$ are perfectly well defined for non-integer values.

This however, introduces the strange idea of our candidates having non-integer vote counts. The important thing to realize is that this doesn't affect the shape of the distribution, as we are not changing the underlying function, which means tha we will still be able to meaningfully compare them.

Carrying forward, I will represent the PDF of the distribution E_k using functional notation, to facilitate the representation of the operations we need to do on it. Therefore, we currently have the following, considering that $x \in \mathbb{N}$.

$$E_k(x) = \binom{v_l}{x} \frac{\mathcal{B}(x + v_k + 1, v_l - x + v_t - v_k + 1)}{\mathcal{B}(v_k + 1, v_t - v_k + 1)}$$

If we are to consider $E_k(x)$ for non-integer values of x, there is one last problem we need to fix. Whereas continous probability distributions use area to determine probability, discrete ones use sums. This means that we need to rescale $E_k(x)$ to ensure the area under its PDF in the interval $[0, v_l]$ is equal to 1, instead of its sum at integer values. This can be achevied by dividing the function by its integral on that range. The continous version of E_k and the continous version of the set E will be respectively denoted E_{ck} and E_c for the sake of clarity.

$$\begin{aligned}
E_{ck}(x) &= \frac{E_k(x)}{\int_0^{v_l} E_k(t)\,\mathrm{d}t} \\
&= \frac{\binom{v_l}{x} \frac{\mathcal{B}(x + v_k + 1, v_l - x + v_t - v_k + 1)}{\mathcal{B}(v_k + 1, v_t - v_k + 1)}}{\int_0^{v_l} \binom{v_l}{t} \frac{\mathcal{B}(t + v_k + 1, v_l - t + v_t - v_k + 1)}{\mathcal{B}(v_k + 1, v_t - v_k + 1)}\,\mathrm{d}t} \\
&= \frac{\left(\frac{1}{\mathcal{B}(v_k + 1, v_t - v_k + 1)}\right) \binom{v_l}{x} \mathcal{B}(x + v_k + 1, v_l - x + v_t - v_k + 1)}{\left(\frac{1}{\mathcal{B}(v_k + 1, v_t - v_k + 1)}\right) \int_0^{v_l} \binom{v_l}{t} \mathcal{B}(t + v_k + 1, v_l - t + v_t - v_k + 1)\,\mathrm{d}t} \\
&= \frac{\binom{v_l}{x} \mathcal{B}(x + v_k + 1, v_l - x + v_t - v_k + 1)}{\int_0^{v_l} \binom{v_l}{t} \mathcal{B}(t + v_k + 1, v_l - t + v_t - v_k + 1)\,\mathrm{d}t}
\end{aligned}$$

We must however keep in mind that for values of x outside of the range $[0, v_l]$, this function must have a value of 0, as it is impossible for a candidate to lose votes our to gain more votes than are left, assuming our value for v_l is correct.

Figure 13 is the plot of the above function for all of our candidates. Just as we would expect, the shape of the graph is exactly the same as in Figure 12, but the scale is slightly different, due to the rescaling. In fact, due to the hundreads of points our previous graph had, it isn't even obvious that we switched from a discrete plot to a continous one.

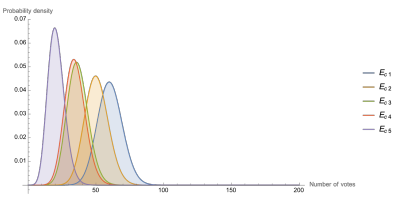

Figure 13: The set of distributions E_c

Comparing these probability distributions, however, would not be the full story. In fact, we not only want to take into account the number of votes each candidate is expected to get, but also the current lead of each candidate. This can be done by translating each of the above functions to the right by their current number of votes. The set of the translated distributions will be referred to as E_{ct} and the distribution of the candidate k as E_{ctk}. Therefore, we have the following.

$$\begin{aligned}
E_{ctk}(x) &= E_{ck}(x - v_k) \\
&= \frac{\binom{v_l}{(x-v_k)}\mathcal{B}((x-v_k)+v_k+1, v_l-(x-v_k)+v_t-v_k+1)}{\int_0^{v_l}\binom{v_l}{t}\mathcal{B}(t+v_k+1, v_l-t+v_t-v_k+1)\,\mathrm{d}t} \\
&= \frac{\binom{v_l}{x-v_k}\mathcal{B}(x+1, v_l-x+v_t+1)}{\int_0^{v_l}\binom{v_l}{t}\mathcal{B}(t+v_k+1, v_l-t+v_t-v_k+1)\,\mathrm{d}t} \\
&= \frac{\binom{v_l}{x-v_k}\mathcal{B}(x+1, v_e-x+1)}{\int_0^{v_l}\binom{v_l}{t}\mathcal{B}(t+v_k+1, v_l-t+v_t-v_k+1)\,\mathrm{d}t}
\end{aligned}$$

Now that we have applied the translation, this expression is only valid for values of x in the interval $[v_k, v_k + v_l]$ and the function has a value of 0 for all other values of x. Plotting this function for all of our candidates leads to the following graph:

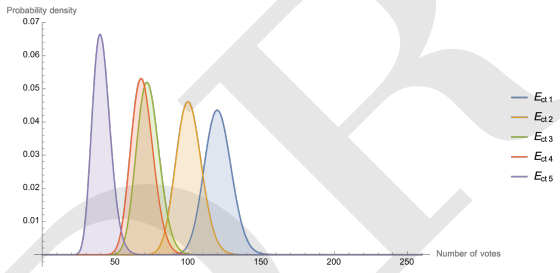

Figure 14: The set of distributions E_{ct}

We now finally have a set of continous 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 4.6, we also need to find the CDF of E_{ctk}.

$$\begin{aligned}
P(E_{ctk} \leq x) &= \int_0^x E_{ctk}(r)\,\mathrm{d}r \\
&= \int_0^x \frac{\binom{v_l}{r-v_k}\mathcal{B}(r+1, v_e-r+1)}{\left(\int_0^{v_l}\binom{v_l}{t-v_k}\mathcal{B}(t+1, v_e-t+1)\,\mathrm{d}t\right)}\,\mathrm{d}r \\
&= \frac{\int_0^x \binom{v_l}{r-v_k}\mathcal{B}(r+1, v_e-r+1)\,\mathrm{d}r}{\int_0^{v_l}\binom{v_l}{t-v_k}\mathcal{B}(t+1, v_e-t+1)\,\mathrm{d}t}
\end{aligned}$$

Keeping in mind the interval on which the expression for E_{ctk} is valid, this expression for the CDF is also only valid over $[v_k, v_k + v_l]$. For values of $x < v_k$, the CDF equals 0 and for values of $x > v_k + v_l$, it equals 1.

5 Analyzing the Model

6 Conclusion

7 Biblography

[1] Radio-Canada Info, director, *Élections Québec 2022 : La soirée électorale*, Oct. 3, 2022. [Online]. Available: https://www.youtube.com/watch?v=EkV28Wew-1Q (visited on 12/17/2022).

[2] Z. P. ICI.Radio-Canada.ca, "Marie-Josée Savard déclarée gagnante par erreur : mais que s'est-il passé? — Élections municipales au Québec 2021," *Radio-Canada.ca*, [Online]. Available: https://ici.radio-canada.ca/nouvelle/1838339/erreur-prediction-victoire-course-mairie-quebec-medias-radio-canada-tva-marie-josee-savard-bruno-marchand (visited on 11/27/2022).

[3] Radio-Canada Info, director, *Soirée électorale 2021 au Canada*, Sep. 20, 2021. [Online]. Available: https://www.youtube.com/watch?v=E3PKiPiwW8o (visited on 12/17/2022).

[4] Radio-Canada Info, director, *Revoyez la soirée électorale fédérale 2019*, Oct. 21, 2019. [Online]. Available: https://www.youtube.com/watch?v=D3tqBHh5IjY (visited on 12/17/2022).

[5] "Elections Canada." (n.d.), [Online]. Available: https://www.elections.ca/ (visited on 12/17/2022).

[6] Radio-Canada Info, director, *Élections en Ontario 2022 : La soirée électorale*, 2022. [Online]. Available: https://www.youtube.com/watch?v=JNHz4hzuN90 (visited on 12/17/2022).

[7] "Elections Ontario." (n.d.), [Online]. Available: https://www.elections.on.ca/en.html (visited on 12/17/2022).

[8] "Élections Québec." (n.d.), [Online]. Available: https://www.electionsquebec.qc.ca/ (visited on 12/17/2022).

[9] E. W. Weisstein. "Law of Large Numbers." (), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/23/2022).

[10] E. W. Weisstein. "Beta Distribution." (n.d.), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/18/2022).

[11] E. W. Weisstein. "Beta Function." (n.d.), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/18/2022).

[12] E. W. Weisstein. "Gamma Function." (n.d.), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/18/2022).

[13] E. W. Weisstein. "Uniform Distribution." (n.d.), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/18/2022).

[14] E. W. Weisstein. "Distribution Function." (), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/21/2022).

[15] E. W. Weisstein. "Regularized Beta Function." (), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/21/2022).

[16] E. W. Weisstein. "Incomplete Beta Function." (), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/21/2022).

[17] "7.2 - Probability Mass Functions — STAT 414," PennState: Statistics Online Courses. (), [Online]. Available: https://online.stat.psu.edu/stat414/lesson/7/7.2 (visited on 12/18/2022).

[18] E. W. Weisstein. "Bayesian Analysis." (), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/18/2022).

[19] 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).

[20] 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).

[21] "Élections provinciales : Montréal et Laval - 10 septembre 2022," Léger. (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).

[22] "Wolfram Mathematica: Modern Technical Computing." (), [Online]. Available: https://www.wolfram.com/mathematica/ (visited on 12/19/2022).

[23] E. W. Weisstein. "Bayes' Theorem." (), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/19/2022).

[24] 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).

[25] Veritasium, director, *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).

[26] J. Orloff and J. Bloom, *Beta Distributions*. [Online]. Available: https://math.mit.edu/~dav/05.dir/class14-prep-a.pdf (visited on 12/22/2022).

[27] whuber. "What is $P(X_1 \geq X_2, X_1 \geq X_3, ..., X_1 \geq X_n)$?" (), [Online]. Available: https://stats.stackexchange.com/q/44142.

[28] E. W. Weisstein. "Beta Binomial Distribution." (), [Online]. Available: https://mathworld.wolfram.com/ (visited on 12/23/2022).

[29] *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).