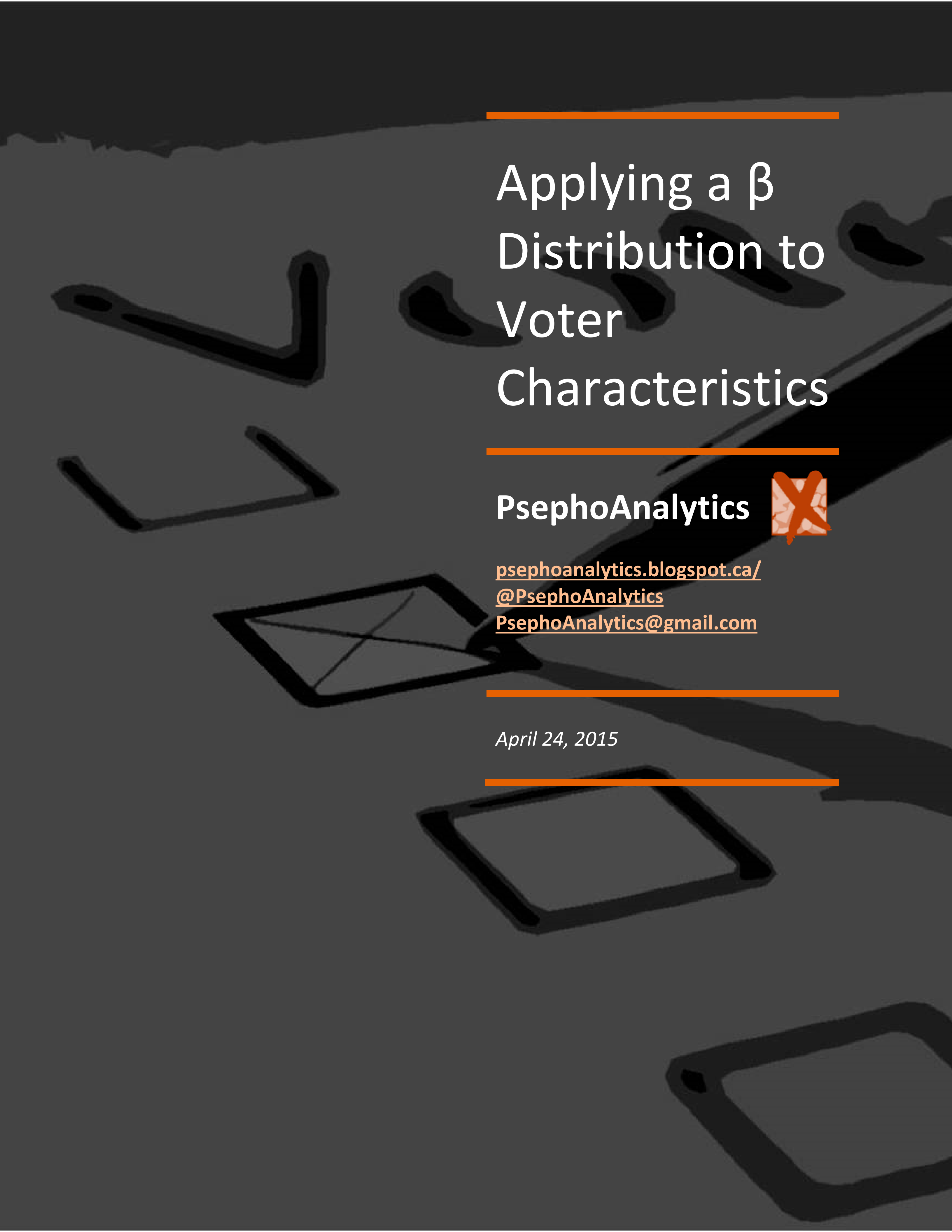


Applying a β Distribution to Voter Characteristics

PsephoAnalytics

psephoanalytics.blogspot.ca/
@PsephoAnalytics
PsephoAnalytics@gmail.com

April 24, 2015

Motivation

The first (and long) step in moving towards agent-based modeling is the creation of the agents themselves. While fictional, they must represent reality – meaning they need to behave like actual people. The main issue in voter modeling, however, is that since voting is private we do not know how *individuals* behave, only collections of voters. So, for example, we could know average beliefs by location based on how that location voted in aggregate. Or, we could go further, by statistically estimating how such differences stem from demographic characteristics to suggest, say, how women of a certain age group and income vote. But that still does not get us to how a *specific* female of a certain age group and income would vote – and we do not want them all to behave the exact same way.

That is why one of the key elements of our work is the ability to create meaningful differences among our agents – particularly when it comes to the likes of issue positions and political engagement. For example, we are *teaching* our agents what they believe (at least initially). This is somewhat akin to polling, except we are (randomly) assigning these agents what they believe rather than asking, such that it aggregates back to what the polls would have said, on average.

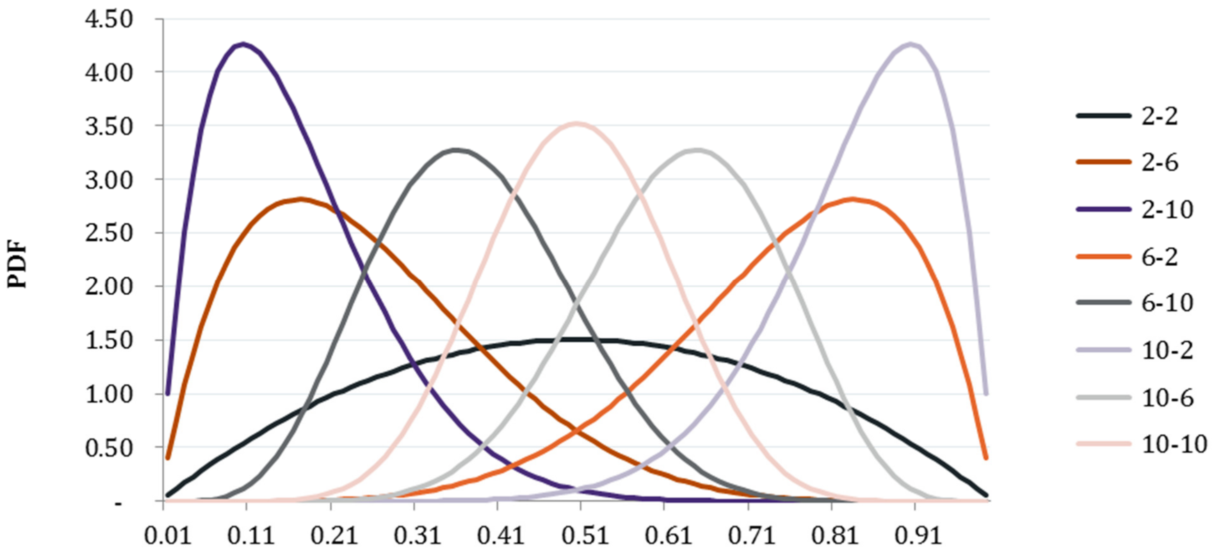
The obvious difficulty is how to do that. In our model, many of our agents' characteristics are limited to values between 0 and 1 (e.g., political positions, weights on given issues). Many standard distributions, such as the normal, would be cut off at these extremes, creating unrealistic "spikes" of extreme behaviour. We also cannot use uniform distributions, as the likelihood of individuals in a group looking *somewhat the same* (i.e., more around an average) seems much more reasonable than them looking uniformly different.

Which brings us to the β distribution.

What is the β distribution?

The β distribution is a type of continuous probability distribution defined between 0 and 1, and is therefore suitable for modeling the random behaviour of percentages and proportions. Such distributions have two (positive) shape parameters – called α and β – that control the shape of the distribution, which can vary widely. For example, Figure 1 shows a number of α-β combinations that "look like" what we want.

Figure 1 – A selection α-β combinations that generate “useful” distributions

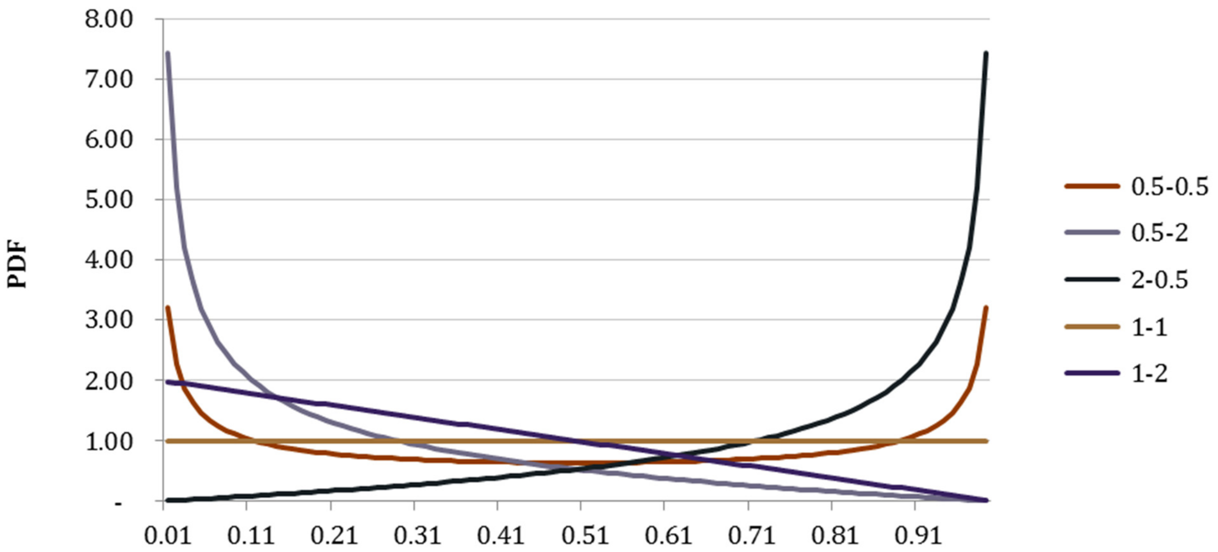


A few things to notice here:

1. When α>β the shape skews to the right and when α<β the shape skews to the left;
2. The higher the values for α and β, the tighter the distribution about the mean; and
3. When α=β, the distribution is symmetric about 0.5.

However, certain combinations can also generate shapes that do not “look like” what we want (see Figure 2). We therefore need to avoid these combinations whenever possible.

Figure 2 – Examples of α-β combinations that generate "not useful" distributions



Our objective is to generate random distributions of behaviour around specified averages. Given that there are many β distributions that could generate the specified average, we need to fix either α or β to try to generate "appropriate" shapes. (As we will explain in the following sections, this will not always be possible.)

Applying the β distribution to voter characteristics

For the following discussion, we will use the first behaviour in our model that requires the β distribution – namely voter engagement (represented by *e*). We have turnout characteristics from our recent study of voter turnout. This allows us to derive average turnout rates by sex, age, education, family type, immigrant status, labour force status, and home ownership combination (9,072 in all), into which each voter agent could be placed. Given that we have a few million voter agents (for Toronto alone), 9,072 types would not provide enough differentiation among agents (i.e., every male/30-34 year old/university grad… would behave the exact same way). So we need to randomly assign each voter agent an initial *e* about their demographic cohort's average – which we do by randomly choosing a value from an inverted β distribution.

What does this look like?

It turns out that it depends quite a lot on which values we choose for fixing α and β. Given we know the average *e* (and the average of the distribution, given α and β), we need to fix one of α or β. For example, if we fix α, then β is defined as:

$$e = \frac{\alpha}{\alpha + \beta} \rightarrow \beta = \alpha \cdot \left(\frac{1 - e}{e}\right)$$

Alternatively, if we fix β, then α is defined as:

$$e = \frac{\alpha}{\alpha + \beta} \rightarrow \alpha = \beta \cdot \left(\frac{e}{1 - e}\right)$$

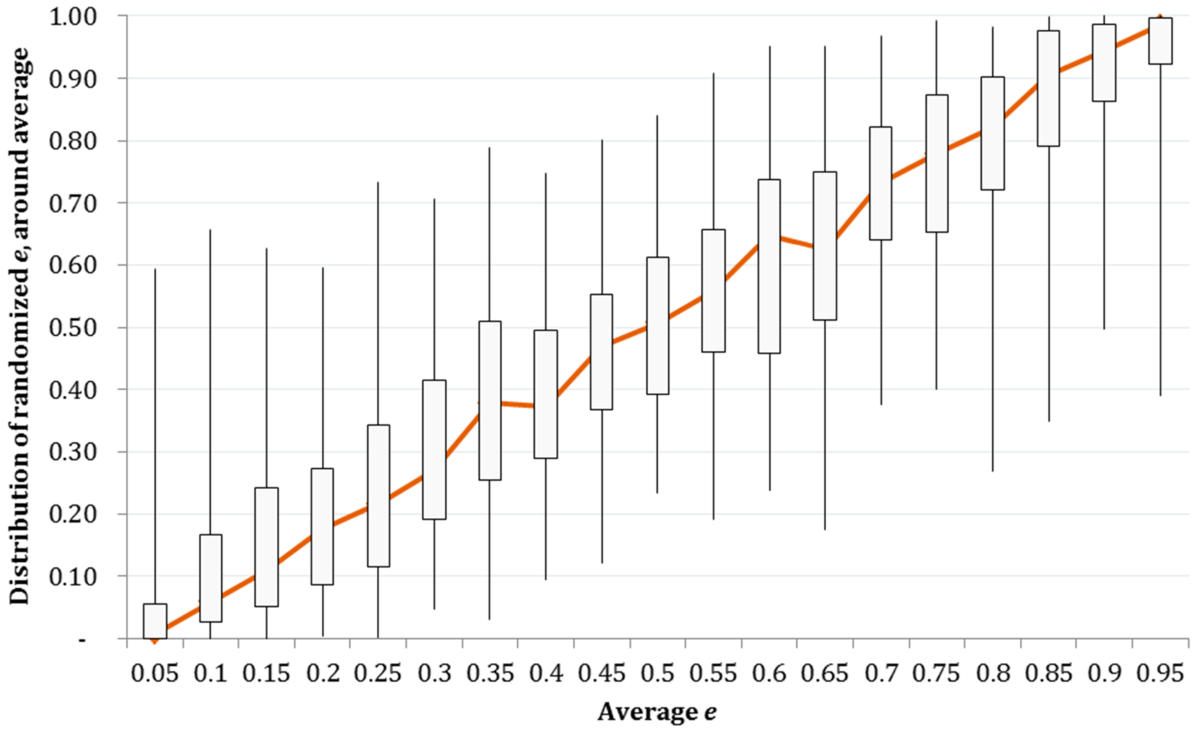
Given the skewing we saw in Figure 1, we may want to flip between fixing α or β depending on where each demographic cohort's *e* sits relative to 0.5. As we move towards the extremes, we want long tails heading the opposite direction. That is, while the cohorts may be increasingly extreme, we want to allow for some members of the cohort to be moderate.

For example, when *e*<0.5 it may be more appropriate to have α defined by β (i.e., fix β, let α move freely). This is because, in such cases, *e*/(1-*e*)<1, meaning α<β, skewing the distributions increasingly to the left. Alternatively, when *e*>0.5 it may be more appropriate to have α fixed and let β move freely. This is because, in such cases, (1-*e*)/*e*<1, meaning α>β, skewing the distributions increasingly to the right. (When α=β, it does not matter which is deemed fixed and which is deemed free.)

Switching which parameter is fixed, depending on the cohort's average *e*, allows the bulk of the individual voter agents in increasingly extreme cohorts to remain around the cohort average, but does not force them all to do so. Further, it ensures that the ranges and distribution shapes are symmetric around 0.5.

To illustrate this, Figure 3 shows a box plot of 100 random draws around a series of specified *e* values (e.g., each representing the average of a demographic cohort) from an inverted β distribution, where either α or β are fixed, as discussed above. The orange line represents the median values, which would be a perfectly straight line with zero sampling error. The boxes show the quartiles on either side of the median – meaning that 50% of the values lie within the box. The vertical lines represent the remaining quartiles, meaning the entire distribution lies between the outer tips of the lines.

Figure 3 – Box plot of β distribution, α=5 when e< 0.5; β=5 when e>0.5; α=β=5 when e=0.5

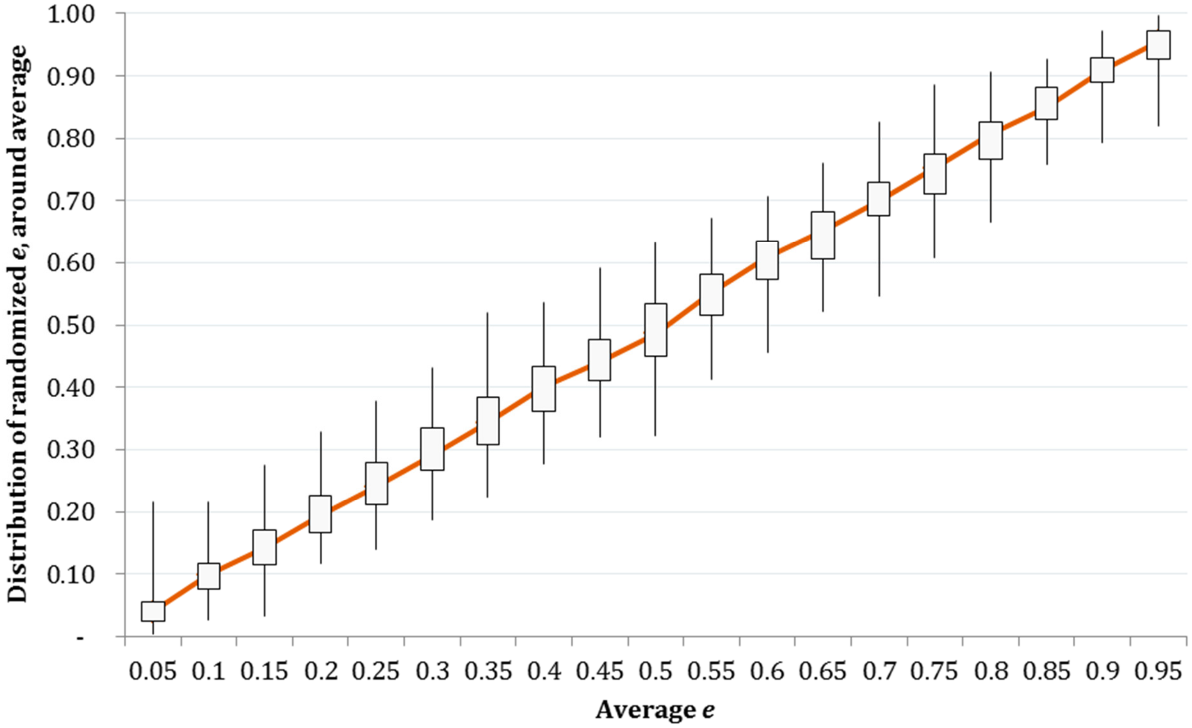


A few things to notice:

1. While the full range for each group is quite wide (a good thing for our purposes), the distribution is relatively tight around the median; and
2. As the average *e* values tend to the extremes, they start to skew heavily as expected. In addition, though, the shape of the distribution flips to one of the "ugly" ones shown in Figure 2 at the very extremes, exacerbating this. (More on this in the next section.)

Similarly, Figure 4 shows a much tighter set of distributions, by significantly increasing α and β.

Figure 4 – Box plot of β distribution, α=50 when e< 0.5; β=50 when e>0.5; α=β=50 when e=0.5



In comparison to Figure 3, these distributions are much tighter with narrow ranges. Further, the skew is much less extreme, and the shape of the distributions are now highly unlikely to "flip".

Which brings us to the trade-off we'll need to make.

The trade-off between high and low fixed values for α and β

As we alluded to in Figures 3 and 4, it is possible for the shape of our distributions to "flip" at extreme average *e* values. This is because by fixing the cohort's average *e* (as defined empirically) and one of α or β (as chosen), it becomes possible for the free moving α or β to drop to or below 1. This occurs when

$$e \geq \frac{\alpha}{1+\alpha}$$

if β is free (i.e., determined by α on the right side of 0.5) or

$$e \leq \frac{1}{1+\beta}$$

if α is free (i.e., determined by β on the left side of 0.5).

In the first case, as $\alpha \rightarrow \infty$, the right-hand side of the inequality tends to 1, meaning it becomes less and less likely that a cohort's average *e* will be high enough to flip the distribution shape. Similarly, in the second case, as $\beta \rightarrow \infty$, the right-hand side of the inequality tends to 0, meaning it becomes less and less likely that a cohort's average *e* will be low enough to flip the distribution shape.

The trade-off then becomes between picking fixed values for α or β that are:

1. Low enough such that our range around each cohort's average *e* value is sufficiently wide to ensure our agents behave differently to one another; and
2. High enough that the likelihood of a cohort's average *e* value flipping the distribution is low, making voters in such groups more extreme.

This will require that, in some cases, we simply accept that groups that are extreme on average are made up of very extreme voters. The balance then becomes mostly one of judgment of "reasonableness", which we hope to at least somewhat empirically limit.

Conclusion

We set out to pick a family of distributions that would help us randomly assign values to agent characteristics (limited to be between 0 and 1), from a given cohort average. Our main concern with most families of distributions (e.g., the normal) was that they would force us to truncate at 0 and 1, creating extreme spikes in some of our agents' behaviour. We also wanted to avoid the uniform distribution, as we felt that agents within a cohort would more often behave like their cohort's average than not.

The β distribution gives us such a family, and is flexible enough to align empirically with our data. That said, there is great diversity in the potential shapes of these distributions, and in (likely) very extreme cases, the shape will not "look like" what we would expect. Therefore, one of our goals will be to somewhat constrain our selection of fixed values for α and β, based on as much empirical data as possible, to ensure we get this balance right.