Project 2: Urn models using the Beta-binomial Distribution

Craig Brooks

March 2023

In this project, we have an urn that contains a distribution of marbles where the average ratio is known to be 32 % White and 68 % Black. To confirm this, we draw from this urn and obtain 340 White and 660 Black marbles.

To our left is a population of urns that follow a beta-binomial distribution with $(\alpha, \beta)_{H0} = \{1, 2\}$, and a population of urns to our right following the same distribution, but instead have parameters $(\alpha, \beta)_{H1} = \{4, 9\}$. Our goal: to determine the likeley population our particular urn most likely came from.

Probability Distribution

While the marbles are multinomially distributed within urns, the conjugate prior follows a Beta-binomial distribution. The key difference between a 'normal' multinomial distribution and the beta-binomial version is that the *probabilities* for each marble color follows a distribution based on a parameters α and β.

The marginal probability for the beta-binomial distribution is

$$Pr(x|n, \alpha, \beta) = \int_0^1 Bin(x|n, \beta) Beta(p|\alpha, \beta) dp$$

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

where x is the category count (in this case White marbles), p is the probability of picking White/Black, n is the number of draws (without replacement), and $(\alpha, \beta)_{Hi}$ are the 'priors' that we know, which essentially the dispersion of the probabilities inside the respective urn populations.

We can write this explicitly as

$$Pr(x|n, \alpha, \beta) = \binom{n}{x} \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(x+a)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Where n is the number of observations per trial and x is the count for the category, in particular, the categories are the colors White and Black. In

our case, the distribution of white marbles will be a good candidate by which to distinguish a population, given that the distribution from $(\alpha, \beta)_{H0}$ and $(\alpha, \beta)_{H1}$ are so different from each other.

In practice, we will use write the beta-binomial likelihood functions using the Beta function.

$$Pr(x|n, \alpha, \beta) = \binom{n}{x} \frac{B(x + \alpha, n - x + \beta)}{B(\alpha, \beta)}$$

The expectation value of the mean μ in the beta-binomial distribution is

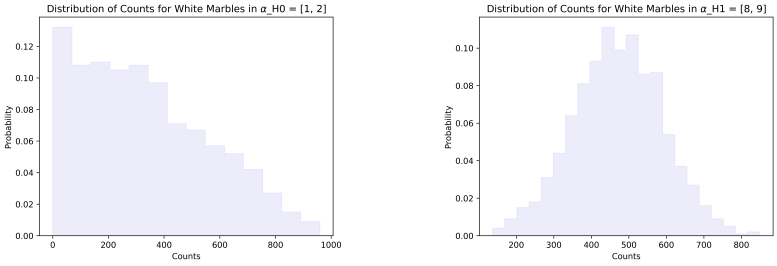
$$E[\mu] = \frac{\alpha}{\alpha + \beta} = np$$

The variance for the beta-binomial distribution can be calculated by

$$E\left[\sigma^2\right] = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + n)} = np(1 - p)\frac{\alpha + \beta + n}{\alpha + \beta + 1}$$

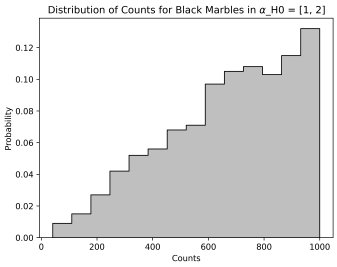
For our experiment, we will set $n = 1000$ and x is the number of White marbles observed, and all statistical values will be for x unless otherwise stated. For $(\alpha, \beta)_{H0} = [1, 2]$, we can calculate $E[\mu] = 300$, so $E[p] \approx .30$. In the case that $(\alpha, \beta)_{H1} = [8, 9]$, then $E[\mu] = 470$ therefore $E[p] \approx .47$. The square root of the variance $\sqrt{\sigma^2}$ for $(\alpha, \beta)_{H0} = 236$ while in $(\alpha, \beta)_{H0} = 118$

Below are histograms representing the distribution of marbles for each color within the $(\alpha, \beta)_{H0}$) and $(\alpha, \beta)_{H1}$ populations, respectively. We notice that in each population, the distributions of the counts for each color are significantly different, although there may be considerable overlap. For example, if we look at the White marbles from $(\alpha, \beta)_{H0}$, they appear to follow a Poisson distribution with $p \approx .3$ as the most probable outcome. while $(\alpha, \beta)_{H1}$ is approximately a normal distribution with the most probable fraction $\approx .5$

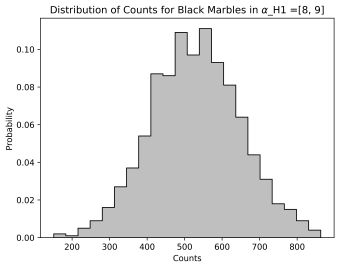


((a)) $\mathbf{H}_0$) ((b)) $\mathbf{H}_1$

Figure 1: Distribution of White marbles for each population



((a)) $\mathbf{H}_0$



((b)) $\mathbf{H}_1$

Figure 2: Distribution of Black marbles for each population

Experiment

First, we define two populations of urns each with `N_urn_pop` $= 1000$ based on their alpha parameters, which we call `alpha_H0` $= \{1, 2\}$ and `alpha_H1` $= \{8, 9\}$ using a Dirichlet random generator (the Dirichlet distribution is the beta-binomial distribution for $K = 2$). These alpha parameters determine the dispersion of each color with the respective population, but each individual urn would follow the usual multinomial distribution. From there. we can generate the individual urns by passing the arrays of the distributions into a function `Category` which by default generates urns with `N_marbles_urn = 10000` This function initially assigns `1, 2,` where `1 = 'White'` and `2 = 'Black'`

Now that the populations of urns are created, we are now ready to sample from one of the populations using `color_samples` (by default, we select 100 urns and sample 1000 marbles from each urn) and take the numbers generated using the `Category` function and convert them to the appropriate `'White', 'Black'` strings. For each urn in our sample, we draw `N_picks` marbles for `N_Trials` from each and calculate the counts/fractions of each color. This data is then written into a file called `alpha_H0_D_{draws}_T_{trials}`

Next, we have a second script called `dirichlet_analysis.py` that will allow us to calculate the log-likelihood, where the likelihood function is

$$\mathcal{L} = Pr(x|n, \alpha, \beta)$$

We select the rejection zone where the log-likelihood $= 1$. We can accomplish this by indexing on the log-likelihood ratio, then pass the index to the data to determine the x-value corresponding to the ratio to define the rejection region.

Hypothesis

We will define the null hypothesis $\mathbf{H}_0$ and the alternative hypothesis $\mathbf{H}_1$ as.

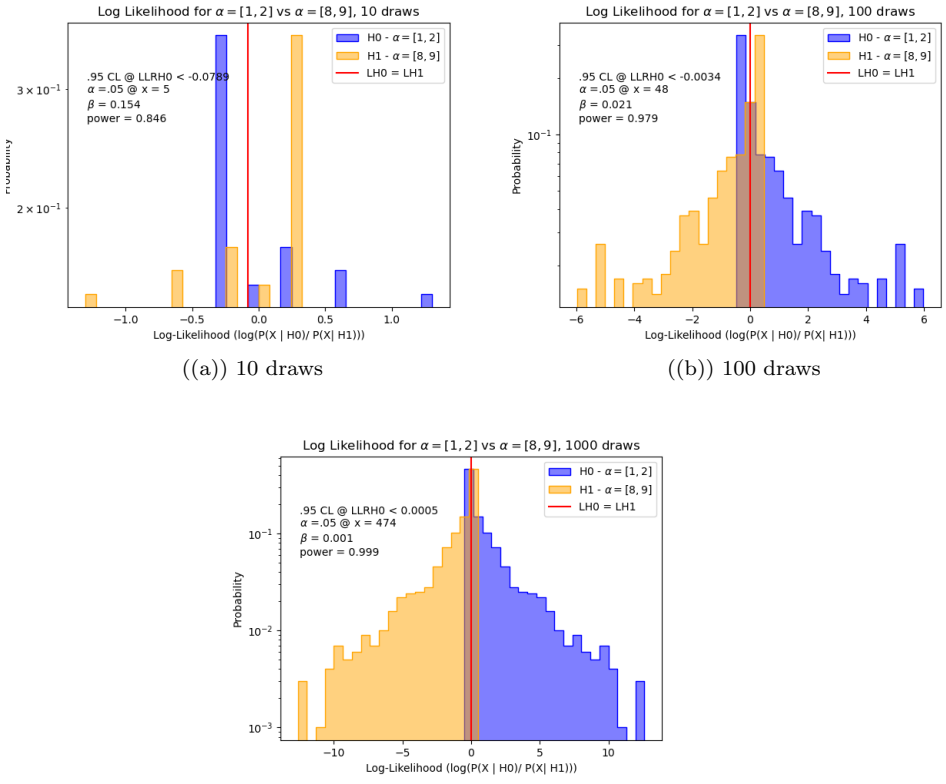
$$\mathbf{H}_0 : X \sim \alpha_0 = [1, 2] \rightarrow p < .47 \quad , \quad \mathbf{H}_1 : X \sim \alpha_1 = [8, 9] \rightarrow p >= .47$$

Analysis

Consider our observation that $x = 340$ *White* and $1 - x = 660$ *Black*. We can calculate the log of the ratio of the likelihoods, assuming $\mathbf{H}_0$ is true, as

$$\sum \log\left(\frac{P(\mathbf{x}|n, \alpha_0, \beta_0)}{P(\mathbf{x}|n, \alpha_1, \beta_1)}\right)$$

Figure 3 shows histograms plotting the log-likelihood for x given $\alpha_{H0} = \{1, 2\}$ and $\alpha_{H1} = \{8, 9\}$. Despite sparse data when 10 per urn are drawn, it becomes clear that as the number of trials increases, we are able to distinguish $\mathbf{H}_0$ from $\mathbf{H}_1$ using our generated data-set.



((a)) 10 draws

((b)) 100 draws

((c)) 1000 draws

Figure 3: Log-Likelihood plots for 10, 100 and 1000 draws per urn

For $\alpha = .05$, the rejection region was calculated to be when $x = 5$ for 10 draws, $x = 48$ for 100 draws, and $x = 474$ for 1000 draws, well above our observation of 360 White marbles for 1000 draws. As a result of this experiment, it is determined there is insufficient evidence to reject the null hypothesis $\mathbf{H}_0$

Significance and power of the test

For 10 marbles, we calculate a $\beta = .154$ with size $1 - \beta = .846$ In the trials where we drew 100 marbles per urn, we calculate a $\beta = .021$ with a size $1 - \beta = .979$. When we drew 1000 marbles per urn, we determined $\beta = .001$ and size $1 - \beta = .999$.