Robust Bayesian Statistics for Bernoulli Trials

Raleigh Bracewell Isadore Priour
Tivy High School

April 2025

1 Abstract

Bayesian Analysis allows for a complete description of one's beliefs about the outcomes of events. However, different people have different beliefs (priors). Robust Bayesian Analysis allows one to start from a baseline prior, for example an uninformative prior, and accepts all beliefs obtainable from that prior given one saw at most the degree of prior certitude of beliefs, d, more points of evidence as reasonable. Inside the collection of reasonable distribution, nothing can be said about which prior/posterior is the *correct* one. This paper goes through the derivation of the beta distribution as a the conjugate prior for the binominal distribution. Then it derives formula and inequalities for Robust Bayesian Analysis of Bernoulli Trials. In order to explain a graphing algorithm, which the author has implemented in python, for the maximum and minimum reasonable probabilities about the chance of success, p_s, of the binominal distribution. The author uses the last 30 years of the S&P500 as an example to illustrate the methods covered in the paper.

Contents

2 Bayesian Inference for Bernoulli Trials

This paper uses the nonstandard notation of $x!$ instead of the cumbersome $\Gamma(x+1)$ notation.

Imagine someone just started out as a vacuum salesperson going door to door trying to figure out the chance[1] of someone opening the door and buying their product, p_s. They have never sold a vacuum before and assume the probability of the chances of success are equal likely; therefore, their pdf for the chance of success is $f(x) = 1$ for $x \in [0, 1]$ where x is the chance(probability) of success.

Suppose the first two houses they went to did not buy a vacuum and let $x = p_s$ therefore $1\text{-}x = 1\text{-}p_s = p_f$ where p_f is the probability of failure. Given x the probability of getting 2 failures and 0 successes is $(1\text{-}x)^2$. The (probability that $p_s = x) \propto (1\text{-}x)^2$. Integrating ones gets $\int_0^1 (1\text{-}x)^2 dx = \text{-}\frac{1}{3}(1\text{ - }x)^3 + C\Big|_0^1 = \frac{1}{3}$ therefore $\Pr(p_s = x|2 \text{ failures})$[2] $= 3(1\text{-}x)^2$. In general, $\Pr(X|\text{Data}) = \dfrac{\Pr(\text{Data}|X)\Pr(X)}{\Pr(Data)}$ (Baye's Theorem). This is the called the posterior distribution with $\Pr(p_s = x) = 1$ being the prior distribution. Now

[1] probability is the correct term; however, this creates esoteric sentences like the probability that the probability is 20%

[2] $\Pr(X|Y)$ means the probability of X given Y

assume one has a successes, b failures, and $n = a+b$, the total number of observations. Then $\Pr(p_s = x) \propto \Pr(\text{Success})^a \Pr(\text{Failure})^b = x^a(1\text{-}x)^b$. With $\Pr(p_s = x) = \dfrac{x^a(1\text{-}x)^b}{\int_0^1 x^a(1\text{-}x)^b dx}$ let $g(a,b) = \displaystyle\int_0^1 x^a(1\text{-}x)^b dx$

$$g(a,b) = \int_0^1 x^a(1\text{-}x)^b dx = \frac{x^{a+1}}{a+1}(1\text{-}x)^b\Big|_0^1 - \int_0^1 \frac{\text{-}b}{a+1}x^{a+1}(1\text{-}x)^{b\text{-}1}dx$$

$$= \frac{b}{a+1}g(a+1, b\text{-}1) \quad \text{for } a,b \geq 0$$

$$= \frac{b}{a+1}\frac{b\text{-}1}{a+2}\cdot\ldots\cdot\frac{2}{b+a\text{-}1}\frac{1}{b+a}g(a+b,0)$$

$$= b!\frac{a!}{(b+a)!}\int_0^1 x^{a+b}dx = \frac{a!b!}{(a+b)!}\frac{1}{a+b+1} = \frac{a!b!}{(a+b+1)!}$$

$$g(a,b) = \int_0^1 x^a(1\text{-}x)^b dx = \frac{a!b!}{(a+b+1)!}$$

Therefore $\dfrac{(a+b+1)!}{a!b!}\displaystyle\int_0^1 x^a(1\text{-}x)^b dx = \frac{(a+b+1)!}{a!b!}\frac{a!b!}{(a+b+1)!} = 1$

So $\Pr(p_s = x) = \dfrac{(a+b+1)!}{a!b!}x^a(1\text{-}x)^b = \dfrac{(n+1)!}{a!b!}x^a(1\text{-}x)^b$

From this we form the Beta distribution $B(x;a,b) = \dfrac{(a+b+1)!}{a!b!}x^a(1\text{-}x)^b$

Here are Some Statistics of the Beta Distribution:

$$\mathbb{E}[B(a,b)] = \int_0^1 \frac{(a+b+1)!}{a!b!}x^a(1\text{-}x)^b x dx = \int_0^1 \frac{(a+b+1)!}{a!b!}x^{a+1}(1\text{-}x)^b dx$$

$$= \int_0^1 \frac{(a+b+1)!}{a!b!}\frac{a+1+b+1}{a+1}\frac{a+1}{a+1+b+1}x^{a+1}(1\text{-}x)^b dx$$

$$= \int_0^1 \frac{(a+1+b+1)!}{(a+1)!b!}\frac{a+1}{a+1+b+1}x^{a+1}(1\text{-}x)^b dx$$

$$= \frac{a+1}{a+1+b+1}\int_0^1 \frac{(a+1+b+1)!}{(a+1)!b!}x^{a+1}(1\text{-}x)^b dx$$

$$\mathbb{E}[B(a,b)] = \frac{a+1}{a+b+2}$$

$$\mathbb{E}[B(a,b)^2] = \int_0^1 \frac{(a+b+1)!}{a!b!} x^a (1\text{-}x)^b x^2 dx = \int_0^1 \frac{(a+b+1)!}{a!b!} x^{a+2} (1\text{-}x)^b dx$$

$$= \int_0^1 \frac{(a+b+1)!(a+b+2)(a+b+3)(a+1)(a+2)}{a!b!(a+1)(a+2)(a+b+2)(a+b+3)} x^{a+2} (1\text{-}x)^b dx$$

$$= \int_0^1 \frac{(a+b+3)!}{(a+2)!b!} \frac{(a+1)(a+2)}{(a+b+2)(a+b+3)} x^{a+2} (1\text{-}x)^b dx$$

$$= \frac{(a+1)(a+2)}{(a+b+2)(a+b+3)} \int_0^1 \frac{(a+b+3)!}{(a+2)!b!} x^{a+2} (1\text{-}x)^b dx$$

$$\mathbb{E}[B(a,b)^2] = \frac{(a+1)(a+2)}{(a+b+2)(a+b+3)}$$

$$\text{Var}[B(a,b)] = \mathbb{E}[B(a,b)^2] \text{-} \mathbb{E}[B(a,b)]^2$$

$$= \frac{(a+1)(a+2)}{(a+b+2)(a+b+3)} - \left(\frac{a+1}{a+b+2}\right)^2$$

$$= \frac{a+1}{a+b+2} \left(\frac{a+2}{a+b+3} - \frac{a+1}{a+b+2}\right)$$

$$= \frac{a+1}{a+b+2} \left(\frac{(a+2)(a+b+2) \text{-} (a+1)(a+b+3)}{(a+b+3)(a+b+2)}\right)$$

$$= \frac{a+1}{a+b+2} \left(\frac{b+1}{(a+b+3)(a+b+2)}\right)$$

$$\text{Var}[B(a,b)] = \frac{(a+1)(b+1)}{(a+b+2)^2(a+b+3)}$$

$$\text{Mode for } a,b>0 \Leftrightarrow \underset{x}{\text{argmax}} \ln(B(x;a,b)) \Leftrightarrow \frac{\partial \ln(B(x;a,b))}{\partial x} = 0$$

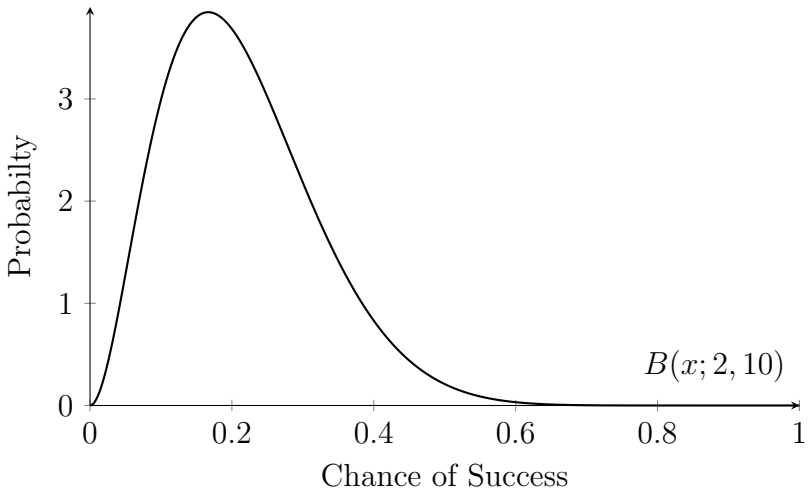
$$\ln(B(x;a,b)) = \ln((a+b+1)!) \text{-} \ln(a!) \text{-} \ln(b!) + a\ln(x) + b\ln(x)$$

$$\frac{\partial \ln(B(x;a,b))}{\partial x} = \frac{a}{x} \text{-} \frac{b}{1\text{-}x} = 0$$

$$a(1\text{-}x) \text{-} bx = 0 \therefore x = \frac{a}{a+b}$$

$$\text{Mode of } B(a,b) = \frac{a}{a+b} \quad \text{for } a,b>0$$

Suppose the salesperson went to 10 more houses and sold 2 vacuums, then their pdf for the chance (probability) of success is $B(x;2,10)$.



3 Framework for Robust Bayesian Statistics

However, there are many problems in which general consensus on the exact prior is not universal. For example, lets say one was modeling the height of Redwood trees as normally distributed with a prior mean and standard deviation. Now suppose one person modeled $\sigma \propto \frac{1}{\sigma}$ while a second person modeled standard deviation as $\sigma = \sigma e^{\sigma}$. Both of these are reasonable priors with the first one being scale invariant while the second excludes implausible scales such as the tree's standard deviation being a mile. If data is collected and the posteriors for standard deviation disagree on parameters, such as the mean or credibility intervals, then said parameters cannot be derived from the data to much confidence since different reasonable starting points result in discord. There is a popular phrase in statistics, *let the data speak for itself*[3], by incorporating multiple reasonable priors and showing what these posteriors agree allows *the data to speak for itself*, because of how one interprets the data does not significantly affect their conclusion. This is my definition of Robust Bayesian Statistics(RBA) having statistical procedures that given a little more or less data does not significantly change the posterior.

The problem this paper will be focused on is robust bayesian statistics for the beta distribution. We will determine the likelihood that S&P500 will be higher at the end of the next trading day than it was at the last trading

[3]This phrase is popular in frequentist statistics not bayesian

day based off the last 30 years the S&P500. The answer is surprisingly low between 52%–55%.

For robust bayesian statistics, define the reasonable set $\mathcal{R}(\theta) = \{\, \pi(\theta) \mid \pi \in \Pi\}$ for a parameter, θ, with Π being the set of all reasonable probability distribution functions. For many statistics $\mathcal{R}(T) = [\inf \mathcal{R}(T), \sup \mathcal{R}(T)]$[4]. However, the determination of any statistic, T, cannot go beyond the precision of $\mathcal{R}(T)$. This means robust bayesian statistics is incapable of giving pointwise estimates. As annoying as this is, it is a good thing. Imagine asking five respected moderate expert economic forecasters about their economic projections for the next 6 months. With three of the economist predicting moderately strong economic growth, one of them predicting slightly negative growth, and one of them predicting moderately strong negative economic growth, it is tempting to omit the moderately negative prediction and report that economic forecasters anticipate negligible to moderate economic growth. If years of school and being in the field producing respected economic predictions cannot lead one to the conclusion that the economy will have a more positive outlook, then the data itself is not strong enough to suggest that conclusion. Bayesians generally do not like the phrase *let the data speak for itself* because one's knowledge about the situation (one's prior) changes what they take away from the data (their posterior). Now suppose one rotated an object with 3 sections: *T,S,H* and got *S,T,H*. Now suppose it was spinner with equal are sides labeled *T,S,H*. This result would be unsurprising and would not give one knowledge that change their preconceived beliefs about the spinner, that each side is probably approximately equally likely as the others. However, if one took out a random nickel with H for heads, T for tails, and S for the side/edge. One would be extremely surprised about getting *S,T,H*, but would still believe that the coin will still most likely land head or tails and that observing the coin landing on its side was an extremely rare[5] observation.It would be ridiculous to conclude from those 3 coins tosses that the probability of a nickel landing on its edge is about as likely as it landing on its head. It is only rational to incorporate one's knowledge about events in their model which can led the exact same data data leading to radically different conclusion as shown here. Fortunately, robust bayesian statistics does allow *the data to speak for itself* because if all reasonable viewpoints (priors)

[4] interval may include or exclude its endpoints

[5] however it is not as rare as one would think with a study predicting the probability to be about 1 in 6,000

led to the a small $\mathcal{R}(\theta)$ set then one has usable clear conclusions which were not the results of someone's prior but the nature of the data. While it is still annoying to have a small range for a parameter instead of an exact number such as the beliefs about the mean of $B(x; a, b)$ ranging from .68 to .72, the real annoying truth is we, humans, want more precision than the data we collected can give us. Going back to the vacuum salesperson example if the actual chance (probability) of success is $.310\overline{0}$ it would take 32868 observations to be 95% certain that the probability of successes is within $.31 \pm .005$. Nonetheless, if a person had a 100 experts in a room there would be very little they could unanimously agree on[6]. If someone just had a bayesian and a frequentist they could not even agree on the correct definition of probability[7]. Therefore, it does matter on what priors one admits as reasonable to the problem with one aiming to incorporate the minimal range that includes all of the different arguable priors for some parameter. A good way to construct the collection of reasonable priors is to start with a baseline prior and then defining the reasonable priors as the class of posteriors achievable based on seeing at most d, prior certitude of beliefs, additonal observations.

4 RBA for Bernoulli Trials

Given the observation of a success, b failures and having d be the prior certitude of beliefs. $\Pi = \{B(a + x, b + y) \mid x, y \in \mathbb{R}_{\geq 0}, x + y \leq d\}$ If ones does not want to have the reasonable priors branch from the uninformative prior, $B(0, 0)$, but from $B(a', b')$ is equivalent to them observing $a + a'$ successes and $b + b'$ failures

$$\mathcal{R}(\text{mean}) = \left[\frac{a+1}{a+b+1+d}, \frac{a+1+d}{a+b+1+d}\right]$$

$$\mathcal{R}(\text{mode}) = \left[\frac{a}{a+b+d}, \frac{a+d}{a+b+d}\right]$$

Variance and standard deviation are tricker, because variance does not decrease monotonically with respect to a or b

$$\text{Var} = \frac{(a+1)(b+1)}{(a+b+2)^2(a+b+3)} \quad \mu = \mathbb{E}[B(a,b)] = \frac{a+1}{a+b+2} \quad n = a + b$$

[6] Unless it was about theorems with mathematicians

[7] It is the bayesian one

$$\text{Var} = \frac{\mu \cdot (1\text{-}\mu)}{n+3} \quad d\text{Var} = \frac{1\text{-}2\mu}{n+3}d\mu\text{-}\frac{\mu(1\text{-}\mu)}{(n+3)^2}dn$$

$$d\mu = \frac{1\text{-}\mu}{n+2}da\text{-}\frac{\mu}{n+2}db \quad dn = da + db$$

$$\frac{\partial\,\text{Var}}{\partial a} = \frac{(1\text{-}\mu)(1\text{-}2\mu)}{(n+3)(n+2)}\text{-}\frac{\mu(1\text{-}\mu)}{(n+3)^2} = \frac{1\text{-}\mu}{n+3}\left(\frac{n+3\text{-}(3n+8)\mu}{(n+2)(n+3)}\right)$$

We want to find the values where an increase in a does not decrease variance given b

$$\frac{\partial\,\text{Var}}{\partial a} = \frac{1\text{-}\mu}{n+3}\left(\frac{n+3\text{-}(3n+8)\mu}{(n+2)(n+3)}\right) \geq 0$$

$$n+3\text{-}(3n+8)\mu \geq 0$$

$$\frac{n+3}{3n+8} \geq \mu$$

$$\frac{1}{3} + \frac{1}{9n+24} \geq \mu$$

$$\frac{1}{3} + \frac{1}{9n+24} \geq 1\text{-}\frac{b+1}{n+2}$$

$$\frac{1}{9n+24} + \frac{b+1}{n+2} \geq \frac{2}{3}$$

$$(9n+24)b + 10n + 26 \geq (6n+16)(n+2)$$

$$6n^2 + (18\text{-}9b)n + 6\text{-}24b \leq 0$$

$$n \leq \sqrt{(\frac{3}{4}b + \frac{7}{6})^2\text{-}\frac{1}{9}} + \frac{3}{4}b\text{-}\frac{3}{2}$$

$$a + b \leq \sqrt{(\frac{3}{4}b + \frac{7}{6})^2\text{-}\frac{1}{9}} + \frac{3}{4}b\text{-}\frac{3}{2}$$

$$a \leq \sqrt{(\frac{3}{4}b + \frac{7}{6})^2\text{-}\frac{1}{9}}\text{-}\frac{1}{4}b\text{-}\frac{3}{2}$$

For the case a<b choose d such that a+d is the closest possible to

$$\sqrt{(\frac{3}{4}b + \frac{7}{6})^2\text{-}\frac{1}{9}}\text{-}\frac{1}{4}b\text{-}\frac{3}{2}$$

For the case b<a choose d such that b+d is the closest possible to

$$\sqrt{(\frac{3}{4}a + \frac{7}{6})^2\text{-}\frac{1}{9}}\text{-}\frac{1}{4}a\text{-}\frac{3}{2}$$

for the case a=b variance is already at it maximum with respect to μ

Credible intervals are still needed in robust bayesian statistics. To determine $\mathcal{R}(\text{Crl})$ and ones just needs to find the beta distribution with the lowest lower bound, b_l, and the beta distribution with the highest upper bound, b_u, given the credible intervals, with $\mathcal{R}(\text{Crl}) = [b_l, b_u]$. More formally, $\mathcal{R}(\text{Crl}_\alpha(\theta)) = \bigcup_{\pi \in \Pi} \text{Crl}_\alpha(\theta; \pi)$. Analysis of the Equal-Tailed Interval (ETI) will be discussed in the next section. ***Add Section link***

5 Analytical Computational Methods for Graphing

Unlike normal bayesian statistics in which one plots their posterior, we plot the regions of values $\pi(\theta)$ can take. Define the upper posterior as $\pi_u(x) = \sup \mathcal{R}(\Pr(p_s = x))$ and the lower posterior $\pi_l(x) = \inf \mathcal{R}(\Pr(p_s = x))$. Using a_s = number of successes observed, b_f = number of failures observed, $n = a_s + b_f$, and $m = n + d$ Also define $B_f(x) = \frac{(m+1)!}{a_s!(m\text{-}a_s)!} x^{a_s}(1\text{-}x)^{m\text{-}a_s}$ and $B_s(x) = \frac{(m+1)!}{(a_s+d)!(m\text{-}a_s)!} x^{a_s+d}(1\text{-}x)^{m\text{-}a_s}$. π_u starts as B_f and stays B_f until $\frac{\partial B_f}{\partial a}\big|_{x,n} = 0$ and then continuously transition throughout the a values until the point where to $\frac{\partial B_s}{\partial a}\big|_{x,n} = 0$. Another way to put it is given x, n we are finding the a value that maximizes $B(x; a, m\text{-}a)$ given $a \in [a_s, a_s + d]$ with B_f and B_s being the optimal values until the a enters and leaves the a value at the points $\frac{\partial B_f}{\partial a}\big|_{x,n} = 0$ and $\frac{\partial B_s}{\partial a}\big|_{x,n} = 0$. However to differentiate B_f with respect to a we need to find the derivative of x! To do this we first derive its extension into $\mathbb{R}_{\geq 0}$

This is not a truly formal proof since it assumes

$\lim_{n\to\infty} \ln((n+x)!) \to \ln(n!) + x \cdot \ln(n)$

$$\ln(n!) = \ln\left(\prod_{k=1}^{n} k\right) = \sum_{k=1}^{n} \ln(k)$$

$$\ln((n+x)!) = \ln((M+x)!) - \sum_{k=n+1}^{M} \ln(k+x)$$

$$\ln((n+x)!) = \lim_{M\to\infty} \ln(M!) + x\ln(M) - \sum_{k=n+1}^{M} \ln(k+x)$$

$$\ln((n+x)!) = \lim_{M\to\infty} \ln(n!) + \sum_{k=n+1}^{M} \ln(k) - \sum_{k=n+1}^{M} \ln(k+x) + x\ln(M)$$

$$\ln((n+x)!) = \ln(n!) + \lim_{M\to\infty} \sum_{k=n+1}^{M} \ln\left(\frac{k}{k+x}\right) + x\ln\left(\frac{k}{k\text{-}1}\right) + x\ln(n)$$

$$\frac{d\ln((n+x))!}{dx} = \lim_{M\to\infty} \sum_{k=n+1}^{M} -\frac{1}{k+x} + \ln\left(\frac{k}{k\text{-}1}\right) + \ln(n)$$

$$\frac{d\ln((n+x))!}{dx} = \lim_{M\to\infty} (\text{-}H_{M+x} + H_{n+x} + \ln(M)\text{-}\ln(n) + \ln(n)) \therefore$$

$$\frac{d\ln((n+x))!}{dx} = H_{n+x}\text{-}\gamma^{[8]} \quad \text{and} \quad \frac{dx!}{dx} = (H_x\text{-}\gamma)x!$$

To evaluate this at noninteger points we next derive the continuation of the Harmonic Numbers[9]

$$\lim_{M\to\infty} H_{M+x} \to H_M + \frac{x}{M}$$

$$H_x = H_{M+x}\text{-}(H_{M+x}\text{-}H_x)$$

$$H_x = \lim_{M\to\infty} \sum_{k=1}^{M} \frac{1}{k} + \frac{x}{M}\text{-}\sum_{k=1}^{M} \frac{1}{k+x} = \lim_{M\to\infty} \sum_{k=1}^{M} \frac{1}{k}\text{-}\frac{1}{k+x} = \sum_{k=1}^{\infty} \frac{1}{k}\text{-}\frac{1}{k+x}$$

We are now ready to calculate $d\ln(B)$

$$\frac{\partial \ln(B)}{\partial a} = \frac{\partial}{\partial a}(\ln((a+b+1)!)\text{-}\ln(a!)\text{-}\ln(b!) + a\ln(x) + b\ln(1\text{-}x))$$

$$\frac{\partial \ln(B)}{\partial a} = H_{a+b+1}\text{-}\gamma - (H_a\text{-}\gamma) + \ln(x) \quad \text{Likewise}$$

$$\frac{\partial \ln(B)}{\partial b} = H_{a+b+1}\text{-}H_b + \ln(1-x) \therefore$$

$$d\ln(B) = (H_{a+b+1}\text{-}H_a + \ln(x))\,da + (H_{a+b+1}\text{-}H_b + \ln(1-x))\,db$$

Local extrema happen where $dB = 0$

---

[8] This is the Euler-Mascheroni Constant $\gamma = \lim_{x\to\infty}(H_x - \ln(x)) \approx 0.5772156649\ldots$

[9] H_n denotes the Harmonic numbers with $H_n = \sum_{k=1}^{n} \frac{1}{k}$ and $H_n = \frac{1}{n} + H_{n\text{-}1}$

$$\implies H_{a+b+1}\text{-}H_a + \ln(x) = H_{a+b+1}\text{-}H_b + \ln(1-x) = 0$$

$$H_b - H_a = \ln\left(\frac{1}{x} - 1\right) = 0 \quad a = b \text{ and } x = \frac{1}{2}$$

Now given the constraint $a + b = m$ using lagrange multipliers one gets

$$H_{m+1}\text{-}H_a + \ln(x) = H_{m+1}\text{-}H_{m\text{-}a} + \ln(1-x) \Leftrightarrow H_{m\text{-}a}\text{-}H_a = \ln\left(\frac{1}{x} - 1\right)$$

$$x(a) = \frac{1}{e^{H_{m\text{-}a}\text{-}H_a} + 1} = \frac{e^{H_a\text{-}H_{m\text{-}a}}}{1 + e^{H_a\text{-}H_{m\text{-}a}}} = 1\text{-}\frac{1}{1 + e^{H_a\text{-}H_{m\text{-}a}}}$$

Given these constraints B_f is the maximum for $0 \geq x \geq x_f = 1\text{-}\dfrac{1}{1 + e^{H_{a_s+d}\text{-}H_{b_f}}}$

and B_s being the maximum for $1 \leq x \leq x_s = 1\text{-}\dfrac{1}{1 + e^{H_{a_s}\text{-}H_{b_f+d}}}$

for $x \in [x_f, x_s] \quad \pi_u(x(a)) = B(x(a); a, m\text{-}a) \quad$ for $a \in [a_s, a_s + d]$

with $x(a) = 1\text{-}\dfrac{1}{1 + e^{H_a\text{-}H_{m\text{-}a}}}$ and $\pi_u(x(a)) = \dfrac{(m+1)!}{a!(m\text{-}a)!}\dfrac{e^{a(H_a\text{-}H_{m\text{-}a})}}{(1 + e^{H_a\text{-}H_{m\text{-}a}})^m}$

Given the constraint that $a + b = m$, the only extrema is a maximum therefore $\pi_l(x) = \min(B_s(x), B_f(x))$ with π_l starting at $B_s(x)$ and transitioning to $B_f(x)$ at x_c where $B_s(x_c) = B_f(x_c)$

$$\Leftrightarrow \frac{(m+1)!}{a_s!(b_f+d)!}x_c^{a_s}(1\text{-}x_c)^{b_f+d} = \frac{(m+1)!}{(a_s+d)!(b_f)!}x_c^{a_s+d}(1\text{-}x_c)^{b_f}$$

$$\Leftrightarrow \frac{(a_s+d)!}{a_s!}(1\text{-}x_c)^d = \frac{(b_f+d)!}{b_f!}x_c^d \Leftrightarrow \sqrt[d]{\frac{(b_f+d)!a!}{b_f!(a_s+d)!}}x_c = 1\text{-}x_c$$

$$\Leftrightarrow x_c = \frac{1}{1 + \prod_{k=1}^{d}\sqrt[d]{\frac{b_f+k}{a_s+k}}} \quad \text{and} \quad x_c \in (x_f, x_s)$$

If x_c was not then there would be a region were $\pi_u(x) = \pi_l(x)$

From this one sees there are 4 sections[10] :

$0 \leq x \leq x_f \;\; \pi_u = B_f$ and $\pi_l = B_s$

$x_f \leq x \leq x_c \;\; \pi_u(x(a)) = B(x(a); a, m\text{-}a)$ and $\pi_l = B_s$

$x_c \leq x \leq x_s \;\; \pi_u(x(a)) = B(x(a); a, m\text{-}a)$ and $\pi_l = B_f$

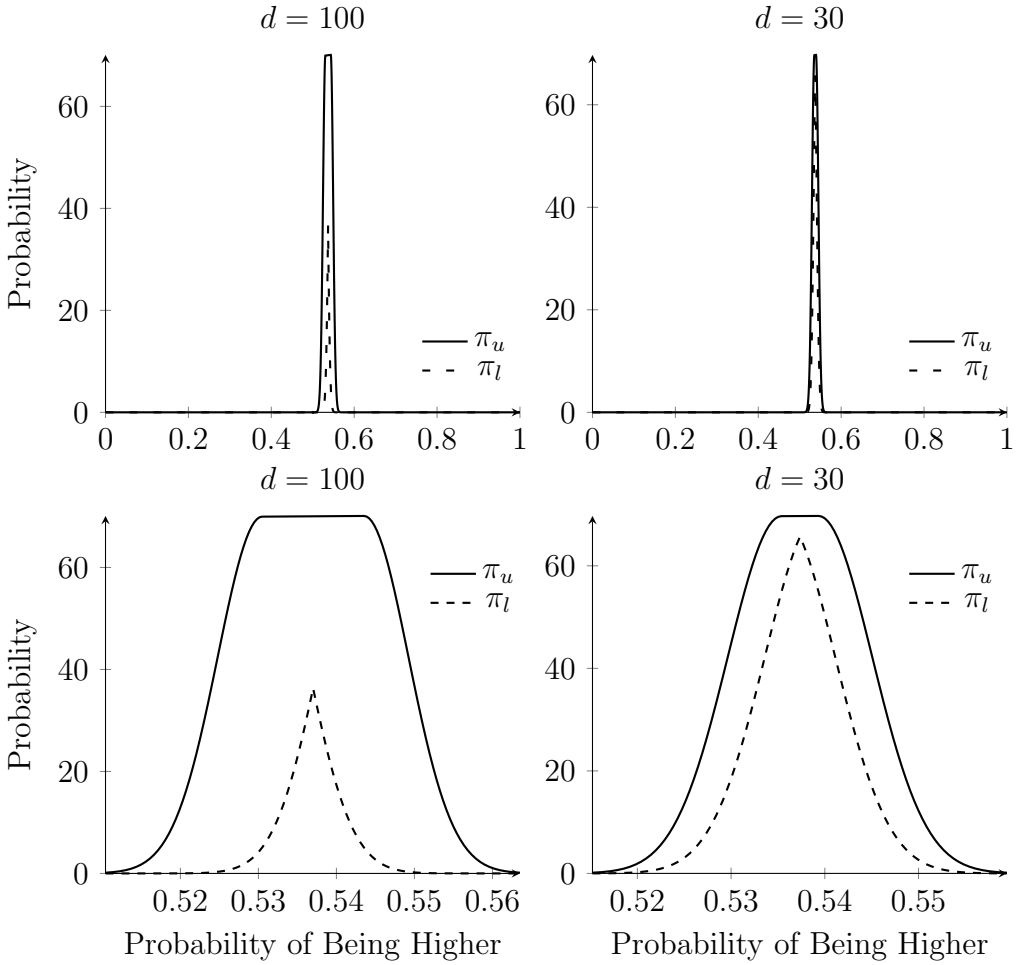
$x_s \leq x \leq 1 \;\; \pi_u = B_s$ and $\pi_l = B_f$

From this we can see that the Equal-Tailed Credible Interval has the nice form $\mathcal{R}(\text{ETI}_{1\text{-}\alpha}(p_s)) = [Q_f(\alpha/2), Q_s(1-\alpha/2)]$, where Q_f, Q_s are the quantile functions for B_f, B_s respectively. This is true because B_f has the biggest cumulative probability function of all distributions in Π and therefore it has the lowest quantile function of all distributions in Π. Likewise, B_s has the lowest cumulative probability function and biggest quantile function of all distributions in Π.

5.1 Example: S&P500

Looking at the last 30 years of the S&P500 in order to determine the likelihood of the S&P500 being higher the next trade day than the trade day before. I analyzed 2 cases using $d = 100$ and $d = 30$. For case $d = 100$ what one can say about this probability is still true given one saw up to 100 additional days of S&P500 daily increases or decreases. For $d = 30$ what one can say about this probability is still true given one saw up to 30 additional days of S&P500 daily increases or decreases. Lastly, the baseline prior is the initially belief that are probabilities are equally likely $B(x; 0, 0)$

[10]The sections all us non-strict inequalities because at those points both definitions give the same answer



From this even with ±100 days of increases or decrease the probability of the S&P500 being higher the next trading day than the current trading day is between 51% and 56%. From this over a general 30 year period of the S&P500 behavior implies that generally the S&P500 is only slightly biased towards increase from day to day.

The graph also show that a naive plotting algorithm that divides the 4 sections of the graph into equal spaced intervals will result in extremely bad graphs, because for the intervals outside $[x_f, x_s]$ the algorithm will sample 1 or 2 points with non-negligible probability, and the rest of the points it samples will have almost zero probability density.

I implemented robust bayesian analysis of Bernoulli trials in python us-

ing matplotlib; however, the algorithm can be effectively implemented in any program or software capable of scientific computing and creating graphs. If someone was plotting the curve $y = 2x + 2$ for $x \in [0, 2]$ they would only need to provide the points:$(0, 2), (2, 4)$ to their graphing software as the line drawn between these two points will be the linear function that they wanted to plot. Therefore, what matters when plotting a function is not the function but the difference between the function and the linear interpolation between the points sampled. The following algorithm calculates the value of the next x point at which the area between the linear approximation and quadratic approximation reaches some threshold and the samples from there. Mathematically, $f(x_0 + x) \approx f(x_0)x + f^{'}(x_0)x + \frac{f^{''}(x_0)}{2}x^2$ and we want to find $\Delta x > 0$ such that $\int_0^{\Delta x} |f(x_0 + t) \text{-} (f(x_0) + f^{'}(x_0)t)|dt = \varepsilon$ using the second order approximation we get $\int_0^{\Delta x} |f(x_0)x + f^{'}(x_0)t + \frac{f^{''}(x_0)}{2}t^2 \text{-} (f(x_0) + f^{'}(x_0)t)|dt = \varepsilon$ so $\int_0^{\Delta x} \frac{|f^{''}(x_0)|}{2}t^2 dt = \varepsilon$ or $\frac{|f^{''}(x_0)|}{6}\Delta x^3 = \varepsilon$ so $\Delta x = \sqrt[3]{\frac{6\varepsilon}{|f^{''}(x_0)|}}$ However this works to well. Since given some data $f = \pi_u$ will be the same function. Therefore given f, $\Delta x_{\text{avg}} \propto \varepsilon^{1/3}$ Now the Beta distribution approaches a stable distribution[11] so asymptotically, using regression from different values of d,n,p_s, and an, where an is approximately the asymptotical number of points sampled for π_u, $\varepsilon = 6.61889/an^3$ The last thing we have to do is calculate f'' which comes in two variants x as the independent variable for $x \notin [x_f, x_s]$ and where x is a function of a for $x \in [x_f, x_s]$. We use this algorithm to sample x points for π_u and π_l for $x \in [x_f, x_s]$ However, from the S&P500 we see that as n becomes large, π_u for $x \in [x_f, x_s]$ becomes flat and so we sample x individually for π_u and π_l, because in this region where π_u is flat π_l is the most interesting.

[11] The normal distribution given $\lim_{n\to\infty} a, b \to \infty$

$$\frac{\partial B(x;a,b)}{\partial x} = \frac{\partial}{\partial x}\left(\frac{(a+b+1)!}{a!b!}x^a(1\text{-}x)^b\right)$$
$$= \frac{(a+b+1)!}{(a\text{-}1)!b!}x^{a\text{-}1}(1\text{-}x)^b\text{-}\frac{(a+b+1)!}{a!(b\text{-}1)!}x^a(1\text{-}x)^{b\text{-}1}$$
$$\frac{\partial B(x;a,b)}{\partial x} = \left(\frac{a}{x}\text{-}\frac{b}{1\text{-}x}\right)B(x;a,b)$$
$$\frac{\partial^2 B(x;a,b)}{\partial x^2} = \frac{\partial}{\partial x}\left(\left(\frac{a}{x}\text{-}\frac{b}{1\text{-}x}\right)B(x;a,b)\right)$$
$$= \left(\text{-}\frac{a}{x^2}\text{-}\frac{b}{(1\text{-}x)^2}\right)B(x;a,b)\left(\frac{a}{x}\text{-}\frac{b}{1\text{-}x}\right)^2 B(x;a,b)$$
$$\frac{\partial^2 B(x;a,b)}{\partial x^2} = \left[\left(\frac{a}{x}\text{-}\frac{b}{1\text{-}x}\right)^2\text{-}\left(\frac{a}{x^2}+\frac{b}{(1\text{-}x)^2}\right)\right]B(x;a,b)$$
recall for $x\in[x_f,x_s]\quad x(a) = 1\text{-}\dfrac{1}{1+e^{H_a\text{-}H_{m\text{-}a}}}$ but to compute $\dfrac{dx}{da}$
we first need to be able to compute$\dfrac{dH_t}{dt}$ recall $H_t = \displaystyle\sum_{k=1}^{\infty}\frac{1}{k}\text{-}\frac{1}{k+t}$
$$\frac{dH_t}{dt} = \psi_1(t)^{[12]} = \sum_{k=1}^{\infty}\frac{1}{(k+t)^2}\ \text{In general}\ \frac{d^nH_t}{dt^n} = \psi_n(t)^{[13]} = \sum_{k=1}^{\infty}\frac{(\text{-}1)^{n+1}n!}{(k+t)^{n+1}}\quad\therefore$$
$$\frac{dx}{da} = \frac{e^{H_a\text{-}H_{m\text{-}a}}}{(1+e^{H_a\text{-}H_{m\text{-}a}})^2}(\psi_1(a)+\psi_1(m\text{-}a)) = \frac{\psi_1(a)+\psi_1(m\text{-}a)}{1+e^{H_a\text{-}H_{m\text{-}a}}}x$$
$$\frac{d^2x}{da^2} = \frac{\psi_2(a)\text{-}\psi_2(m\text{-}a)}{1+e^{H_a\text{-}H_{m\text{-}a}}}x\text{-}\frac{e^{H_a\text{-}H_{m\text{-}a}}(\psi_1(a)+\psi_1(m\text{-}a))^2}{(1+e^{H_a\text{-}H_{m\text{-}a}})^2}x+\frac{(\psi_1(a)+\psi_1(m\text{-}a))^2}{(1+e^{H_a\text{-}H_{m\text{-}a}})^2}x$$
$$1\text{-}2x = 1\text{-}\frac{2e^{H_a\text{-}H_{m\text{-}a}}}{1+e^{H_a\text{-}H_{m\text{-}a}}} = \frac{1\text{-}e^{H_a\text{-}H_{m\text{-}a}}}{1+e^{H_a\text{-}H_{m\text{-}a}}}\quad\therefore$$
$$\frac{d^2x}{da^2} = \frac{x}{1+e^{H_a\text{-}H_{m\text{-}a}}}\left((1\text{-}2x)(\psi_1(a)+\psi_1(m\text{-}a))^2+\psi_2(a)\text{-}\psi_2(m\text{-}a)\right)$$
$$y(a) = B(x(a);a;m\text{-}a) = \frac{(m+1)!}{a!(m\text{-}a)!}\left(\frac{e^{a(H_a\text{-}H_{m\text{-}a})}}{(1+e^{H_a\text{-}H_{m\text{-}a}})^m}\right)$$
$$\ln y(a) = \ln((m+1)!)\text{-}\ln(a!)\text{-}\ln((m\text{-}a)!)+a(H_a\text{-}H_{m\text{-}a})\text{-}m\ln\left(1+e^{H_a\text{-}H_{m\text{-}a}}\right)$$

[12] This is the trigamma function

[13] Using nonstandard notation that k start at 1 and not at 0 for the polygamma functions,ψ_n

$$\frac{d\ln y}{da} = \gamma\text{-}H_a\text{-}\gamma + H_{m\text{-}a} + H_a\text{-}H_{m\text{-}a} + a(\psi_1(a) + \psi_1(m\text{-}a))\text{-}mx(\psi_1(a) + \psi_1(m\text{-}a))$$

$$\frac{d\ln y}{da} = (a\text{-}mx)(\psi_1(a) + \psi_1(m\text{-}a))$$

$$\frac{d^2\ln y}{da^2} = \frac{d}{da}\left((a\text{-}mx)(\psi_1(a) + \psi_1(m\text{-}a))\right)$$

$$= (1\text{-}m\frac{\psi_1(a) + \psi_1(m\text{-}a)}{1 + e^{H_a\text{-}H_{m\text{-}a}}}x)(\psi_1(a) + \psi_1(m\text{-}a)) + (a\text{-}mx)(\psi_2(a)\text{-}\psi_2(m\text{-}a))$$

Surprisingly $\frac{d^2f}{dx^2} \iff 1/\frac{d^2x}{df^2}$ Proof: $f(x) = x^2 \quad \frac{d^2f}{dx^2} = 2 \quad 1/\frac{d^2x}{df^2} = -4x^3$

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{da}\left(\frac{dy}{dx}\right)\cdot\frac{da}{dx} = \frac{d}{da}\left(\frac{dy}{da}\frac{da}{dx}\right)\cdot\frac{da}{dx}$$

$$= \left(\frac{d^2y}{da^2}\frac{da}{dx} + \frac{dy}{da}\cdot\frac{d}{da}\left(\frac{da}{dx}\right)\right)\cdot\frac{da}{dx}$$

$$= \frac{d^2y}{da^2}(\frac{da}{dx})^2 + \frac{dy}{da}\frac{da}{dx}\cdot\frac{d}{da}\left(\frac{1}{\frac{dx}{da}}\right) = \frac{d^2y}{da^2}\left(\frac{da}{dx}\right)^2 + \frac{dy}{da}\frac{da}{dx}\cdot\left(-\frac{1}{\left(\frac{dx}{da}\right)^2}\frac{d^2x}{da^2}\right)$$

$$\frac{d^2y}{dx^2} = \frac{d^2y}{da^2}\left(\frac{da}{dx}\right)^2\text{-}\frac{dy}{da}\left(\frac{da}{dx}\right)^3\frac{d^2x}{da^2}$$

$$\frac{d^2\ln y}{da^2} = \frac{d}{da}\left(\frac{y'}{y}\right) = \frac{y''}{y}\text{-}\frac{(y')^2}{y^2} = \frac{y''}{y}\text{-}\left(\frac{d\ln y}{da}\right)^2 \quad \therefore$$

$$\frac{d^2y}{dx^2} = y\left(\frac{d^2\ln y}{da^2} + \left(\frac{d\ln y}{da}\right)^2\right) \quad \therefore$$

$$\frac{d^2y}{dx^2} = y\left(\frac{da}{dx}\right)^2\left(\frac{d^2\ln y}{da^2} + \frac{d\ln y}{da}\left(\frac{d\ln y}{da}\text{-}\frac{da}{dx}\frac{d^2x}{da^2}\right)\right)$$

Now that we know all of the derivatives we are ready to describe the implementation:

$h \leftarrow \sqrt[3]{6\varepsilon}$

$x \leftarrow x_f$

$\pi_u, \pi_l = B_f, B_s$

While $x \geq 0$:

Sample π_u, π_l

$$x \leftarrow x - \frac{h}{\sqrt[3]{\left|\frac{d^2\pi_u}{dx^2}(x)\right|}}$$

$a \leftarrow a_s$

$\pi_u(a) = \pi(x(a), a, m-a)$

While $a \leq a_s + d$:

$x \leftarrow x(a, m-a)$

Sample π_u

$$a \leftarrow a + \frac{da}{dx}\frac{h}{\sqrt[3]{\left|\frac{d^2\pi_u}{dx^2}(x)\right|}}$$

$x \leftarrow x_c$

While $x \geq x_f$:

Sample π_l

$$x \leftarrow x - \frac{h}{\sqrt[3]{\left|\frac{d^2\pi_l}{dx^2}(x)\right|}}$$

$x \leftarrow x_c$

$\pi_u, \pi_l = B_s, B_f$

While $x \leq x_s$:

Sample π_l

$$x \leftarrow x + \frac{h}{\sqrt[3]{\left|\frac{d^2\pi_l}{dx^2}(x)\right|}}$$

While $x \leq 1$:

Sample π_u, π_l

$$x \leftarrow x + \frac{h}{\sqrt[3]{\left|\frac{d^2\pi_u}{dx^2}(x)\right|}}$$

The author's implementation of the code can be found at *github.com/coolnicecool/Robust-Bayesian-Analysis*

The last thing we will show is that highest density value of π_u is equal to value of the π with mode furthest from .5 which is $B_f(\frac{a_s}{m})$ when $2a_s \leq n$ and $B_s(\frac{a_s+d}{m})$ when $2a \geq n$. As well as, Mode $B_f \leq x_f$ and $x_s \leq$ Mode B_s.

$$\begin{aligned}\max \pi_u &= \max\{\max\{\pi(x) \mid \pi \in \Pi\} \mid x \in [0,1]\} \\ &= \max\{\max\{\pi(x) \mid x \in [0,1]\} \mid \pi \in \Pi\}\end{aligned}$$

$$\max \pi_u = \max\left\{ B\left(\frac{a}{m}; a, m\text{ - }a\right) \,\middle|\, a \in [a_s, a_s + d] \right\}$$

$$y(a) = B\left(\text{Mode } B(a, m\text{ - }a); a, m\text{ - }a\right) = \frac{(m+1)!}{a!(m\text{ - }a)!}\frac{a^a(m\text{ - }a)^{(m-a)}}{m^m}$$

$$\ln y(a) = \ln((m+1)!)\text{ - }\ln(a!)\text{ - }\ln((m\text{ - }a)!) + a\ln(a) + (m\text{ - }a)\ln(m\text{ - }a)\text{ - }m\ln(m)$$

Since ln is a strictly increasing function and $y(a) > 0$ this implies

$$\max \pi_u = \max\{\ln y(a) \mid a \in [a_s, a_s + d]\} \text{ which occurs when } \frac{d\ln y}{da} = 0$$

$\frac{d\ln y}{da} = \gamma$ - H_a - $\gamma + H_{m\text{-}a} + \ln(a) + 1$ - $\ln(m$ - $a)$ - 1

$\frac{d\ln y}{da} = H_{m\text{-}a}$ - $\ln(m$ - $a)$ - $(H_a$ - $\ln(a))$

Let $f(x) = H_x$ - $ln(x)$ so $\frac{d\ln y}{da} = f(m$ - $a)$ - $f(a)$

$\ln y$ achieves a local extrema when $f(a) = f(m$ - $a)$ This only happens when $a = m\text{-}a \Leftrightarrow a = \frac{m}{2}$ because f is a strictly decreasing function and therefore injective[14]

To see that f is strictly decreasing we will show that $f'(x) < 0$ for all $x > 0$

$$f'(x) = \frac{dH_x}{dx} - \frac{1}{x} = \sum_{k=1}^{\infty} \frac{1}{(k+x)^2} - \frac{1}{x} < \int_0^{\infty} \frac{1}{(k+x)^2} dk - \frac{1}{x}$$

$$< -\frac{1}{k+x} + C\Big|_{k=0}^{k=\infty} - \frac{1}{x} = \frac{1}{x} - \frac{1}{x} = 0 \quad \therefore$$

$f'(x) < 0$ so f is a strictly decreasing and injective function

for $a < \frac{m}{2}$ $\quad m$ - $a > a$ so $0 < f(m$ - $a) < f(a)$

Since $\frac{d\ln y}{da} = f(m$ - $a)$ - $f(a)$ $\quad$ This shows $\frac{d\ln y}{da}$ is decreasing from $a = 0$ to $a = \frac{m}{2}$ and similarly that $\frac{d\ln y}{da}$ is increasing from $a = \frac{m}{2}$ to $a = m$

Therefore, $\ln y(a)$ achieves a global minimum at $a = \frac{m}{2}$ for $a \in [0, m]$ and symmetrically increase from $\frac{m}{2}$, $\ln y(\frac{m}{2} - a) = \ln y(\frac{m}{2} - a)$. Therefore, π_u's maximum given $a \in [a_s, a_s + d]$ is $B_f(\frac{a_s}{m})$ when $2a_s \leq n$ and $B_s(\frac{a_s+d}{m})$ when $2a \geq n$

Finally, we will show Mode $B_f \leq x_f$ and $x_s \leq$ Mode B_s. By showing that $x(a)$ such that $\frac{\partial B(x; a, n-a)}{\partial a}\Big|_{x=x(a),n}$ follows the inequalities below:

$$\text{Mode } B(a, n-a) < x(a) < \frac{1}{2} \quad \text{when } a < \frac{n}{2}$$

$$\frac{1}{2} = x(a) = \text{Mode } B(a, n-a) \quad \text{when } a = \frac{n}{2}$$

[14]Suppose $f(a) = f(m$ - $a)$ since f is injective this implies $a = m$ - a

$$\frac{1}{2} < x(a) < \text{Mode } B(a, n-a) \quad \text{when } a > \frac{n}{2}$$

Recall $x(a) = 1\text{-}\frac{1}{1+e^{H_a\text{-}H_{n\text{-}a}}} \therefore \ln\left(\frac{x(a)}{1\text{-}x(a)}\right) = H_a\text{-}H_{n\text{-}a} = \sum_{k=1}^{\infty} \frac{1}{k+n\text{-}a} - \frac{1}{k+a}$

assume $a > \frac{n}{2}$ so the sum is positive and therefore

$$\ln\left(\frac{x(a)}{1\text{ - }x(a)}\right) = \sum_{k=1}^{\infty} \frac{1}{k+n\text{ - }a} - \frac{1}{k+a} < \int_0^{\infty} \frac{1}{k+n\text{ - }a} - \frac{1}{k+a} dk$$

$$\ln\left(\frac{x(a)}{1\text{ - }x(a)}\right) = \sum_{k=1}^{\infty} \frac{1}{k+n\text{ - }a} - \frac{1}{k+a} < \ln\left(\frac{x+n\text{ - }a}{x+a}\right) + C\Big|_{x=0}^{x=\infty}$$

$$\ln\left(\frac{x(a)}{1\text{ - }x(a)}\right) = \sum_{k=1}^{\infty} \frac{1}{k+n\text{ - }a} - \frac{1}{k+a} < \ln\left(\frac{a}{n\text{ - }a}\right)$$

$$\ln\left(\frac{x(a)}{1\text{ - }x(a)}\right) < \ln\left(\frac{a}{n\text{ - }a}\right) \Leftrightarrow \frac{x(a)}{1\text{ - }x(a)} < \frac{a}{n\text{ - }a} \Leftrightarrow x(a) < \frac{a}{n\text{-}a} - \frac{a}{n\text{-}a}x(a)$$

$$\Leftrightarrow \frac{n}{n\text{-}a}x(a) < \frac{a}{n\text{-}a} \Leftrightarrow x(a) < \frac{a}{n} = \text{Mode } B(a, n-a)$$

If $a = \frac{n}{2}$ then $x(a) = 1\text{-}\frac{1}{1+e^{H_{n/2}\text{-}H_{n\text{-}n/2}}} = 1 - \frac{1}{2} = \frac{1}{2}$

and Mode $B(\frac{n}{2}, n) = \frac{1}{2}$

Finally, if $a < \frac{n}{2}$ then $\ln\left(\frac{x(a)}{1\text{-}x(a)}\right) = H_a\text{-}H_{n\text{-}a} = -(H_{(n\text{-}a)}\text{-}H_{n\text{-}(n\text{-}a)})$

$$\ln\left(\frac{x(a)}{1\text{-}x(a)}\right) = -\ln\left(\frac{x(n\text{-}a)}{1\text{-}x(n\text{-}a)}\right) > -\ln\left(\frac{(n\text{-}a)}{n-(n\text{-}a)}\right) = -\ln\left(\frac{n\text{-}a}{a}\right)$$

$$\frac{x(a)}{1\text{-}x(a)} > \frac{a}{n\text{-}a} \implies x(a) > \frac{a}{n} = \text{Mode } B(a, n\text{-}a) < \frac{1}{2} \square$$

6 Discussion

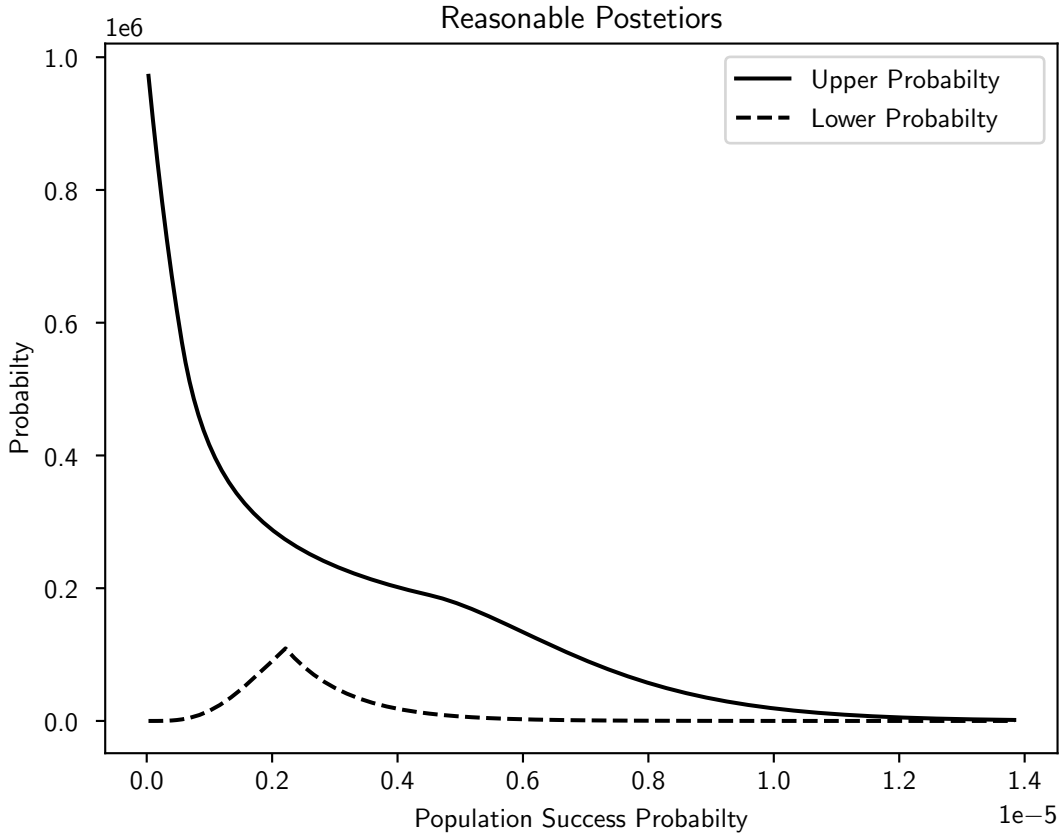
6.1 implementation

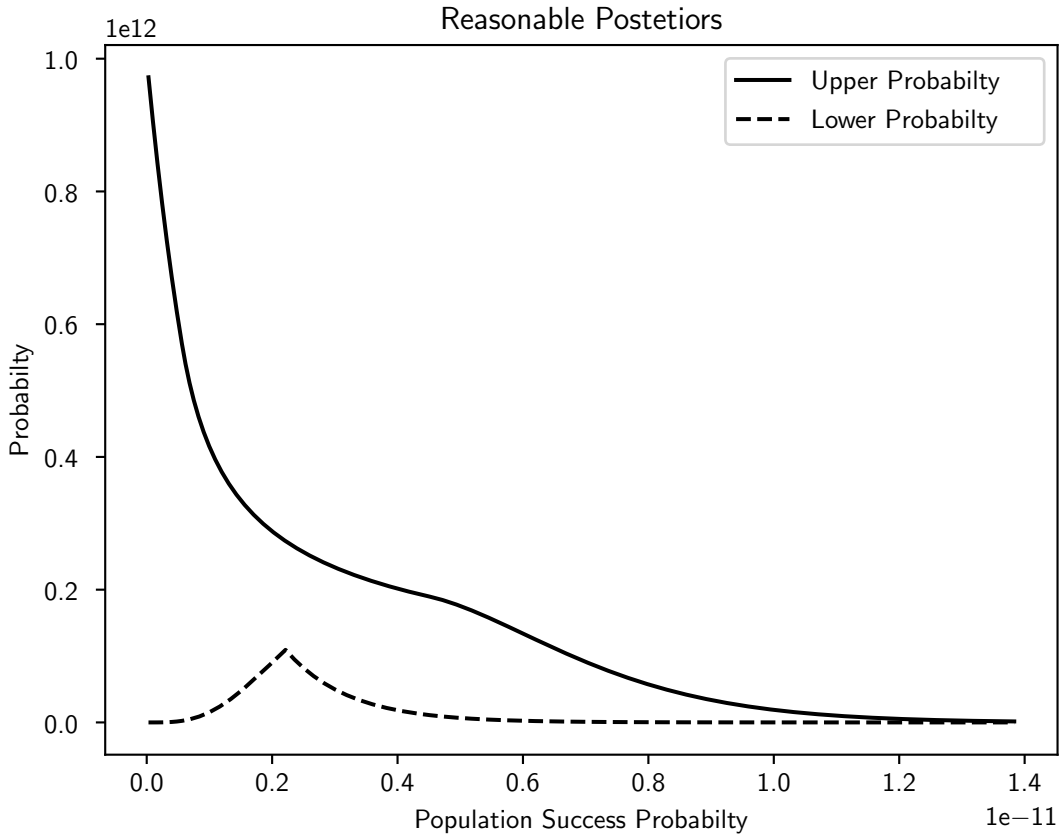
As said previously my implementation can be found on github. The implementation I have written can handle up to $n \approx 2^{63} \approx 10^{22}$ after that due to numpy and secret uses of 64 signed integers the graphing code becomes wonky; however, the actually statistics work for $n \approx 10^{150}$ after that

python is unable to handle the numerical values. I also tried where instead of worrying about cumulative error ones worries about the point in which the difference between the linear and quadratic approximation reach some threshold and where the angle between 3 points reached some threshold from 180°. Both methods worked significantly worse than the original cumulative error method.

6.2 Case where $p_s = 0, 1$

This is actually really intreating because π_u and π_l approach the same 2 relative curves that are most certainly not the normal distribution. Heres an example with $d = 4$





Anyways the case where $p_s = 0$ or $p_s = 1$ create interesting limiting curves

6.3 Choosing Degree of Prior Certitude

This is a very interesting question. Considering the S&P500 for our discussion, 30 years is along time with my data set containing 7559 trade days. An initial thought is to first start with the baseline prior and then calculate one's posterior and have $d = \lambda\sqrt{n * \mathbb{E}[\text{Var}]}$. However, this feels wrong since one cannot define their statistical procedure and ones is changing the hyperparameters based on the data. What one can do instead is set $d = \lambda\sqrt{n}$. For the S&P500 analysis $d = 100$ is $\lambda \approx 1.15$ and for $d = 30$ is $\lambda \approx 0.345$. However, this again is not the point of RBA since the amount of reasonable priors do not increase proportionally to $\sqrt{n}$ portion

6.4 $[x_f, x_s]$ region

From the Analysis of the S&P500 we see that it gets a flat top. π_u for $x \in [x_f, x_s]$ becomes flat after around 50 to 100 datapoints given one gets both success and failures. This means that for large n only 2 or 3 points are actually sampled and since the region is bounded between Mode B_f. Moreover, the length of $[x_f, x_s]$ is less than $\frac{d}{n+d}$ while $\sigma \propto \frac{1}{\sqrt{n}}$ so the interesting transition region of π_u from B_f to B_s becomes negligible as n becomes large.