Robust Bayesian Statistics for Bernoulli Trials

Raleigh Bracewell Isadore Priour
Raleigh@1337maps.com
Tivy High School

April 2025

1 Abstract

Bayesian Analysis allows for a complete description of one's beliefs about the outcomes of events. Although different people have different beliefs (priors), Robust Bayesian Analysis (RBA) 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 probability distribution, nothing can be said about which prior/posterior is the *correct* one. The point of Robust Bayesian Statistics is to convince ones audience of their result by encapsulating the viewer's belief's in their baseline. Which departs from the usual goal of statistics to efficiently as possible determine conclusion from the data. In short, the goal of Robust Bayesian Statistics is to be convincing rather than efficent. This paper goes through the derivation of the beta distribution as the conjugate prior for the binomial distribution. Then it derives formulas 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 binomial distribution. The author uses the last 30 years of the S&P 500 [Han(2020)] as an example to illustrate the methods covered in the paper.

Contents

1 Abstract **1**

2 Bayesian Inference for Bernoulli Trials **2**
- 2.1 Introduction to the Beta Distribution 2
- 2.2 Descriptive Statistics of the Beta Distribution 4
- 2.3 Example 5

3 Framework for Robust Bayesian Statistics **5**

4 RBA for Bernoulli Trials **7**

5 Analytical Computational Methods for Graphing **10**
- 5.1 Example: The S&P 500 13

6 Discussion **20**
- 6.1 Implementation 20
- 6.2 Case where $a_s = 0$ or $b_f = 0$ 21
- 6.3 Choosing Degree of Prior Certitude 21
- 6.4 $[x_f, x_s]$ Region 22
- 6.5 Conclusion 22

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

2 Bayesian Inference for Bernoulli Trials

2.1 Introduction to the Beta Distribution

Consider that someone has just started out as a vacuum salesperson going door to door trying to figure out the chance* for 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 probability density function (pdf), f, for the chance of success is $f(x) = 1$ for $x \in [0, 1]$ where x is the chance(probability) of success.

*Probability is the correct term; however, this creates esoteric sentences like the probability that the probability is 20%.

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 p_f the probability of getting 2 failures and 0 successes is $(1\text{-}p_f)^2$. The (probability that $p_s = x$)$\propto (1\text{-}x)^2$. Integrating ones gets $\displaystyle\int_0^1 (1\text{-}x)^2 dx = -\frac{1}{3}(1 - x)^3 + C\Big|_0^1 = \frac{1}{3}$ therefore $\Pr(p_s = x|2 \text{ failures})^* = 3(1\text{-}x)^2$ with the 3 having no significance other than being the normalization constant. In general, $\Pr(X|\text{Data}) = \dfrac{\Pr(\text{Data}|X)\Pr(X)}{\Pr(\text{Data})}$ (Baye's Theorem). This is the called the posterior distribution with $\Pr(p_s = x) = 1$ being the prior distribution. Now assume one has a successes and b failures with the total number of observations, $n = a+b$. 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 $g(a, b) = \displaystyle\int_0^1 x^a(1\text{-}x)^b dx$

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

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

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

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

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

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

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

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

*$\Pr(X|Y)$ means the probability of X given Y.

2.2 Descriptive Statistics of the Beta Distribution

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

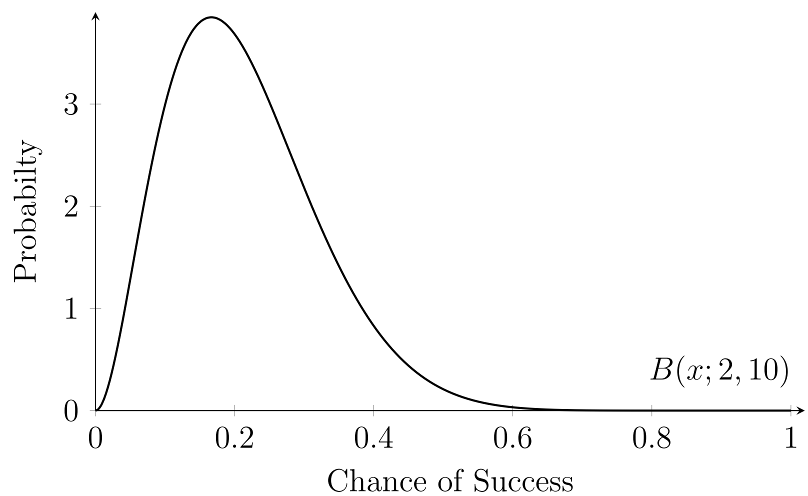
$$\frac{\partial \ln(B(x; a, b))}{\partial x} = \frac{a}{x} - \frac{b}{1 - x} = 0$$

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

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

2.3 Example

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

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, μ, and standard deviation, σ. Now suppose one person modeled $\Pr(\sigma = x) \propto \frac{1}{x}$ while a second person modeled standard deviation as $\Pr(\sigma = x) = xe^{-x}$. 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 incompatible results. There is a popular phrase in statistics, *let the data speak for itself*[*], by incorporating multiple reasonable priors and showing what these posteriors agree allows *the data to speak for itself*, because it shows how one interprets the data does not significantly affect their conclusion. This is the definition I use for robust Bayesian statistics 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 the S&P 500 will be higher at the end of the next trading day than it was at the last trading day based off the last 30 years the S&P 500. 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. For many statistics $\mathcal{R}(T) = [\inf \mathcal{R}(T), \sup \mathcal{R}(T)]$[†]. 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 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. However if years of school and experience 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*.

[*] This phrase is popular in frequentist statistics not Bayesian.
[†] Interval may include or exclude its endpoints.

Now suppose it was spinner with sides of equal area 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* 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) 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 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 31.00% it would take 32868 observations to be 95% certain that the probability of successes is within $31\% \pm .5\%$. Nonetheless, if a person had a 100 experts in a room there would be very little they could unanimously agree on†. Moreover, if someone just had a Bayesian and a frequentist they could not even agree on the correct definition of probability‡. 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.

4 RBA for Bernoulli Trials

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

*However it is not as rare as one would think with a study predicting the probability to be about 1 in 6,000 [Murray and Teare(1993)].

†Unless it was about theorems with mathematicians.

‡It is the Bayesian one.

posteriors achievable based on seeing at most one's prior degree of certitude, d, additional observations. We shall now apply this analysis to Bernoulli trials. Without loss of generality assume the baseline prior is $B(0,0)$ with one observing a_s successes, b_f failures, having prior degree of certitude, d, and $m = a_s + b_f + d$. If one wants the baseline prior to be $B(a', b')$ that is equivalent to observing $a_s + a'$ successes and $b_f + b'$ failures.

$\Pi = \{B(a_s + x, b_f + y) \mid x, y \in \mathbb{R}_{\geq 0}, x + y \leq d\}$

Denote the ends of Π as

$B_f(x) = B(x; a_s, b_f + d) = \dfrac{(m+1)!}{a_s!(b_f+d)!} x^{a_s}(1\text{-}x)^{b_f+d}$ and

$B_s(x) = B(x; a_s + d, b_f) = \dfrac{(m+1)!}{(a_s+d)!b_f!} x^{a_s+d}(1\text{-}x)^{b_f}$

$\mathcal{R}(\text{Mean}) \left[\dfrac{a_s+1}{a_s+b_f+d+1}, \dfrac{a_s+d+1}{a_s+d+b_f+1}\right]$

$\mathcal{R}(\text{Mode}) = \left[\dfrac{a_s}{a_s+b_f+d}, \dfrac{a_s+d}{a_s+d+b_f}\right]$

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

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

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

$\mu = \dfrac{a+1}{a+b+2} = 1 \text{-} \dfrac{b+1}{a+b+2} \implies \dfrac{\partial \mu}{\partial a} = \dfrac{b+1}{(a+b+2)^2}$ and $\dfrac{\partial \mu}{\partial b} = \text{-}\dfrac{a+1}{(a+b+2)^2}$

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

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

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

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

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

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

$\frac{1}{3}+\frac{1}{9n+24}\geq 1\text{-}\frac{b+1}{n+2}$
$\frac{1}{9n+24}+\frac{b+1}{n+2}\geq\frac{2}{3}$
$(9n+24)b+10n+26\geq(6n+16)(n+2)$
$6n^2+(18\text{-}9b)n+6\text{-}24b\leq 0$ carefully solving the inequality one gets
$n=a+b\leq\sqrt{(\frac{3}{4}b+\frac{7}{6})^2\text{-}\frac{1}{9}}+\frac{3}{4}b\text{-}\frac{3}{2}$
$a\leq\sqrt{(\frac{3}{4}b+\frac{7}{6})^2\text{-}\frac{1}{9}}\text{-}\frac{1}{4}b\text{-}\frac{3}{2}$
For the case $a_s<b_f$ choose d such that a_s+d is the closest possible to
$\sqrt{(\frac{3}{4}b_f+\frac{7}{6})^2\text{-}\frac{1}{9}}\text{-}\frac{1}{4}b_f\text{-}\frac{3}{2}$
For the case $b_f<a_s$ choose d such that b_f+d is the closest possible to
$\sqrt{(\frac{3}{4}a_s+\frac{7}{6})^2\text{-}\frac{1}{9}}\text{-}\frac{1}{4}a_s\text{-}\frac{3}{2}$
For the case $a_s=b_f$ variance is already at it maximum with respect to μ

Credible intervals are still needed in robust Bayesian statistics. To determine $\mathcal{R}(\text{Crl})$ one 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, $\mathcal{R}(\text{Crl}_\alpha(\theta))=\bigcup_{\pi\in\Pi}\text{Crl}_\alpha(\theta;\pi)$. The Equal-Tailed Credible Interval (ETI) 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. To see this assume there exist a $B_t\in\Pi$ with $B_t(x)=B(x;a_s+a_t,b_f+b_t)$ such that $Q_t(\alpha)<Q_f(\alpha)$ for an $\alpha\in[0,1]$. We are interested in the relationship* between B_f and B_t which switches only at points where $B_f(x)=B_t(x)$.
$B(x;a_s,b_f+d)=B(x;a_s+a_t,b_f+b_t)$
$\frac{(a_s+b_f+d+1)!}{a_s!(b_f+d)!}x^{a_s}(1\text{-}x)^{b_f+d}=\frac{(a_s+a_t+b_f+b_t+1)!}{(a_s+a_t)!(b_f+b_t)!}x^{a_s+a_t}(1\text{-}x)^{b_f+b_t}$
$\frac{(a_s+b_f+d+1)!}{a_s!(b_f+d)!}(1\text{-}x)^{d\text{-}b_t}=\frac{(a_s+a_t+b_f+b_t+1)!}{(a_s+a_t)!(b_f+b_t)!}x^{a_t}$ for $x\in(0,1)$

*$\leq$ and $\geq$.

The left side of the equation is strictly decreasing and starts out greater than the right side, while the right side is strictly increasing and ends out greater than the left side. This implies their exist an unique $x_t \in (0,1)$ such that $B_f(x_t) = B_t(x_t)$ therefore,
$B_t(x) \leq B_f(x)$ for $x \in [0, x_t]$
$\implies F_t(x) \leq F_f(x)^*$ for $x \in [0, x_t]$ and
$B_f(x) \leq B_t(x)$ for $x \in [x_t, 1]$ $\therefore$

$$\int_x^1 B_f(y)dy \leq \int_x^1 B_t(y)dt \text{ for } x \in [x_t, 1]$$

$\implies 1 - F_f(x) \leq 1 - F_t(x)$ for $x \in [x_t, 1]$ $\quad\therefore$
$F_t(x) \leq F_f(x)$ for $x \in [x_t, 1]$ and for $x \in [0, x_t]$ implying
$Q_f(\alpha) = F_f^{-1}(\alpha) \leq F_t^{-1}(\alpha) = Q_t(\alpha)$ for $\alpha \in [0,1]$
mi Using the same line of reasoning it can be shown that
$Q_t(\alpha) \leq Q_s(\alpha)$ for $\alpha \in [0,1]$
This also implies that $\mathcal{R}(\text{Median}) = \left[Q_f\left(\frac{1}{2}\right), Q_s\left(\frac{1}{2}\right)\right]$

5 Analytical Computational Methods for Graphing

Unlike normal Bayesian statistics in which one plots their posterior, we plot the regions of values $\pi(\theta)^\dagger$ can take. Define the upper posterior as $\pi_u(x) = \sup \mathcal{R}(\Pr(p_s = x))$ and the lower posterior $\pi_l(x) = \inf \mathcal{R}(\Pr(p_s = x))$. π_u starts as B_f and stays B_f until x reaches the value x_f where $\frac{\partial B_f}{\partial a}\Big|_{x_f,m} = 0$ from there π_u continuously transition throughout the a values ending at the point x_s where $\frac{\partial B_s}{\partial a}\Big|_{x_s,m} = 0$. Another way to put this 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 the extrema being outside $[a_s, a_s{+}d\,]$ for $x \notin [x_f, x_s]$ and the endpoints B_f and B_s being the optimal values until $x \in [x_f, x_s]$ in which an local extrema with respect to a is in the interval $[a_s, a_s + d]$ with a_s being an extrema at $\frac{\partial B_f}{\partial a}\Big|_{x,m} = 0$ and $a_s + d$ being an extrema at $\frac{\partial B_s}{\partial a}\Big|_{x,m} = 0$. However to differentiate B_f, B_s with respect to a we need to find the derivative of $x!$. To do this we first derive the extension of factorial function into $\mathbb{R}_{\geq 0}$.

* F_t and F_f denote the cumulative distribution function of B_t and B_f respectively.
† π denotes the probability distribution of the parameter, θ.

$$\ln((M+n)!) = \ln(1\cdot 2\cdot \ldots \cdot (M-1)\cdot M\cdot (M+1)\cdot (M+2)\cdot \ldots \cdot (M+n))$$
$$= \ln(M!) + \ln((M+1)\cdot (M+2)\cdot \ldots \cdot (M+n))$$
$$= \ln(M!) + \ln\left(M^n + \frac{n(n+1)}{2}M^{n-1} + \ldots + n!\right)$$
$$= \ln(M!) + \ln\left(M^n \cdot \left(1 + \frac{n(n+1)}{2M} + \ldots + \frac{n!}{M^n}\right)\right)$$
$$\ln((M+n)!) = \ln(M!) + n\ln(M) + \ln\left(1 + \frac{n(n+1)}{2M} + \ldots + \frac{n!}{M^n}\right)$$
$$\lim_{M\to\infty} \ln\left(1 + \frac{n(n+1)}{2M} + \ldots + \frac{n!}{M^n}\right) = 0 \quad \text{therefore}$$
$\lim_{M\to\infty} \ln((M+n)!) \to \ln(M!) + n\cdot\ln(M)$ we now assume this implies that
$$\lim_{M\to\infty} \ln((M+x)!) \to \ln(M!) + x\cdot\ln(M)$$
$$\ln(x!) = \ln((M+x)!) - \big(\ln((M+x)!) - \ln(x!)\big)$$
$$\ln(x!) = \ln((M+x)!) - \big(\ln(M+x) + \ln(M+x-1) + \ldots + \ln(x+1) + \ln(x!) - \ln(x!)\big)$$
$$\ln(x!) = \ln((M+x)!) - \sum_{k=1}^{M} \ln(k+x) \quad \therefore$$
$$\lim_{M\to\infty} \ln(x!) = \lim_{M\to\infty} \left(\ln((M+x)!) - \sum_{k=1}^{M} \ln(k+x)\right)$$
$$\ln(x!) = \lim_{M\to\infty} \left(\ln(M!) + x\ln(M) - \sum_{k=1}^{M} \ln(k+x)\right) \quad \therefore$$
$$\frac{d\ln(x!)}{dx} = \lim_{M\to\infty} \left(\ln(M) - \sum_{k=1}^{M} \frac{1}{k+x}\right) = \lim_{M\to\infty} \big(\ln(M) - H_{M+x} + H_x\big) \quad \therefore$$
$$\frac{d\ln(x!)}{dx} = H_x - \gamma^* \quad \text{and} \quad \frac{dx!}{dx} = (H_x - \gamma)x!$$

To evaluate this at non-integer points we need to derive the continuation of the Harmonic Numbers.
H_n denotes the nth Harmonic number with $H_n = \sum_{k=1}^{n} \frac{1}{k} = \frac{1}{n} + H_{n-1}$
$$H_{M+n} = \sum_{k=1}^{M+n} \frac{1}{k} = \sum_{k=1}^{M} \frac{1}{k} + \sum_{k=1}^{n} \frac{1}{M+k} = H_M + \frac{n}{M} + \sum_{k=1}^{n} \frac{1}{M+k} - \frac{1}{M}$$

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

$$\lim_{M\to\infty}\sum_{k=1}^{n}\frac{1}{M+k}-\frac{1}{M}=0 \quad \therefore \quad \lim_{M\to\infty} H_{M+n}\to H_M+\frac{n}{M}$$

We assume this implies $\quad \lim_{M\to\infty} H_{M+x}\to H_M+\frac{x}{M}$

We also assume that $H_{x+1}=\frac{1}{x+1}+H_x \Leftrightarrow H_{x+n}=H_x+\sum_{k=1}^{n}\frac{1}{k+x}$

$$H_x=H_{M+x}-(H_{M+x}-H_x)=H_{M+x}-\sum_{k=1}^{M}\frac{1}{k+x}$$

$$\lim_{M\to\infty} H_x=\lim_{M\to\infty}\left(H_{m+x}-\sum_{k=1}^{M}\frac{1}{k+x}\right)=\lim_{M\to\infty}\left(H_M+\frac{x}{M}-\sum_{k=1}^{M}\frac{1}{x+k}\right)$$

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

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

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

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

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

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

Now given the constraint $a+b=m$ using Lagrange multipliers to find the constrained extrema of $\ln(B)$ one gets $\frac{\partial\ln(B)}{\partial a}=\lambda$ and $\frac{\partial\ln(B)}{\partial b}=\lambda$ therefore $\frac{\partial\ln(B)}{\partial a}=\frac{\partial\ln(B)}{\partial b}$ and we can replace b with $m-a$ getting the equation

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

$\ln(B(x;a,m-a))$ is an extrema when $x=\frac{1}{e^{H_{m-a}-H_a}+1}=1-\frac{1}{1+e^{H_a-H_{m-a}}}$

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

and B_s is the maximum for $x_s=1-\frac{1}{1+e^{H_{a_s+d}-H_{b_f}}}\le x\le 1$

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 $B_f(x_c) = B_s(x_c)$

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

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

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

If $x_c \notin (x_f, x_s)$ then there would exist $x \in [0, 1]$ where $\pi_u(x) = \pi_l(x)$

From this one sees there are 4 sections*:

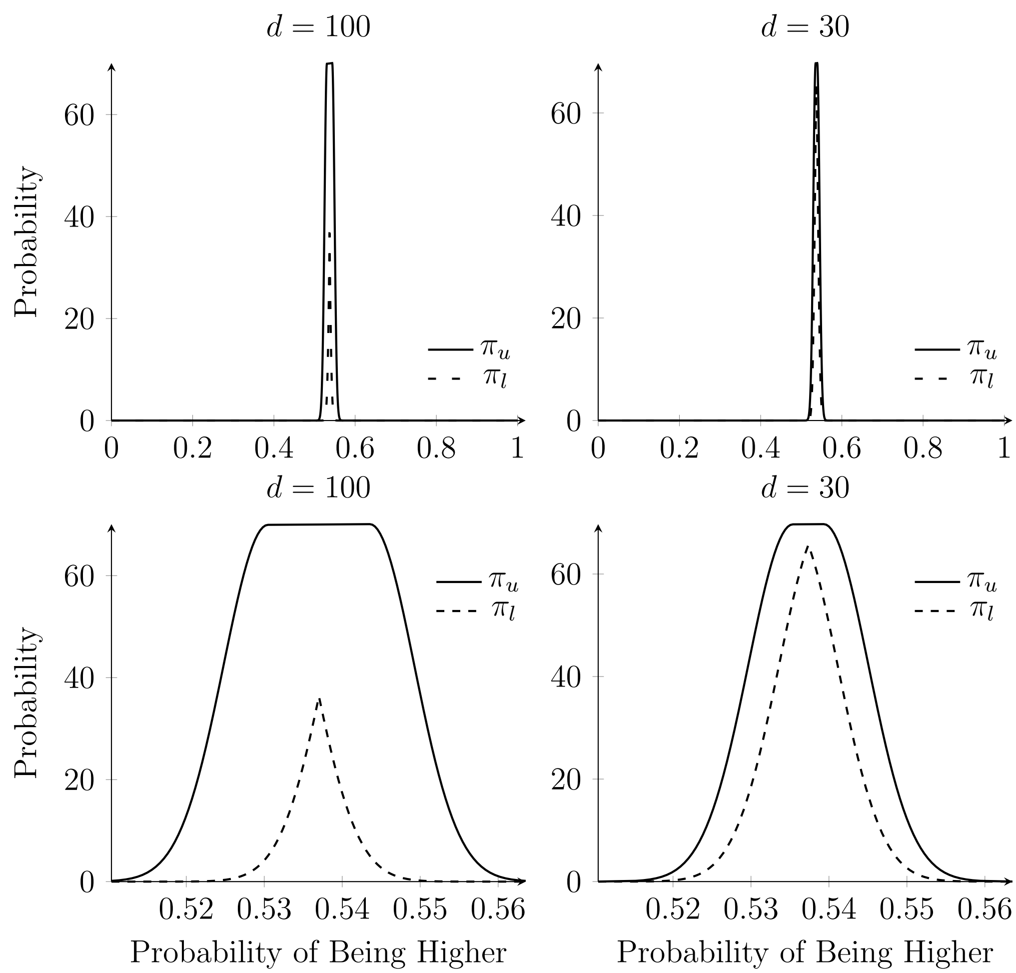
$0 \le x \le x_f$ $\quad \pi_u = B_f$ and $\pi_l = B_s$
$x_f \le x \le x_c$ $\quad \pi_u(x(a)) = B(x(a); a, m\text{-}a)$ and $\pi_l = B_s$
$x_c \le x \le x_s$ $\quad \pi_u(x(a)) = B(x(a); a, m\text{-}a)$ and $\pi_l = B_f$
$x_s \le x \le 1$ $\quad \pi_u = B_s$ and $\pi_l = B_f$

5.1 Example: The S&P 500

Looking at the last 30 years of the S&P 500 in order to determine the likelihood of the S&P 500 being higher the next trade day than the trade day before. I analyzed 2 cases using $d = 100$ and $d = 30$. For the case where $d = 100$ what one can say about this probability is still true given one saw up to 100 additional days of S&P 500 daily increases through decreases. For $d = 30$ what one can say about this probability is still true given one saw up to 30 additional days of S&P 500 daily increases through decreases. Lastly, the baseline prior is the initially belief that are probabilities are equally likely $B(x; 0, 0)$. Here is the robust Bayesian statistical analysis of the S&P 500.

*The sections all use non-strict inequalities because at those points both definitions give the same answer.

S&P 500	$d = 30$	$d = 100$
Successes	4063	4063
Failures	3496	3496
Total	7559	7559
$\mathcal{R}$(Mean)	$53.735\% \pm .395\%$	$53.701\% \pm 1.305\%$
$\mathcal{R}$(Mode)	$53.736\% \pm .395\%$	$53.702\% \pm 1.306\%$
$\mathcal{R}$(90% ETI)	$52.60\% - 54.83\%$	$52.11\% - 55.29\%$
$\mathcal{R}$(95% ETI)	$52.41\% - 55.05\%$	$51.93\% - 55.47\%$
$\mathcal{R}$(99% ETI)	$52.06\% - 55.40\%$	$51.58\% - 55.82\%$



From this even with ±100 days of increases through decreases the probability of the S&P 500 being higher the next trading day than the current trading day is between 51.6% and 55.8%. Showing that in the long-run the S&P 500 is only slightly biased towards increasing 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 density, and the rest of the points it samples will have almost zero probability density.

I implemented robust Bayesian analysis of Bernoulli trials in python using 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 would be the linear function that they wanted plotted. 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 next x point where the area between the linear and quadratic approximation reaches some threshold, ε, and sample from there. Mathematically, $f(x_0 + x) \approx f(x_0) + f'(x_0)x + \dfrac{f''(x_0)}{2}x^2$ and we want to find $\Delta x > 0$ such that $\displaystyle\int_0^{\Delta x} |f(x_0 + t) - (f(x_0) + f'(x_0)t)|dt = \varepsilon$ using the second order approximation we get $\displaystyle\int_0^{\Delta x} |f(x_0) + f'(x_0)t + \frac{f''(x_0)}{2}t^2 - (f(x_0) + f'(x_0)t)|dt = \varepsilon$ so $\displaystyle\int_0^{\Delta x} \frac{|f''(x_0)|}{2}t^2 dt = \varepsilon$ or $\dfrac{|f''(x_0)|}{6}\Delta x^3 = \varepsilon$ so $\Delta x = \sqrt[3]{\dfrac{6\varepsilon}{|f''(x_0)|}}$.
Given f, $\Delta x_{\text{avg}} \propto \varepsilon^{1/3}$ therefore, $\varepsilon^{-1/3} \propto \frac{1}{\Delta x_{\text{avg}}} \propto n_s$ where n_s is the number of points sampled Using regression from different values of d, n, and p_s, the asymptotical number of points sampled for π_u, n_a, is approximately $\sqrt[3]{6.61889/\varepsilon}$ giving us $\varepsilon = 6.61889/n_a^3$. 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&P 500 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. In order to implement said algorithm, we will derive $\dfrac{d^2\pi_u}{dx^2}, \dfrac{d^2\pi_l}{dx^2}$.

$$\frac{\partial B(x;a,b)}{\partial x} = \frac{\partial}{\partial x}\left(\frac{(a+b+1)!}{a!b!}x^a(1-x)^b\right)$$

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

$$\frac{\partial B(x;a,b)}{\partial x} = \left(\frac{a}{x}-\frac{b}{1-x}\right)B(x;a,b)$$

$$\frac{\partial^2 B(x;a,b)}{\partial x^2} = \frac{\partial}{\partial x}\left(\left(\frac{a}{x}-\frac{b}{1-x}\right)B(x;a,b)\right)$$

$$= \left(-\frac{a}{x^2}-\frac{b}{(1-x)^2}\right)B(x;a,b) + \left(\frac{a}{x}-\frac{b}{1-x}\right)^2 B(x;a,b)$$

$$\frac{\partial^2 B(x;a,b)}{\partial x^2} = \left[\left(\frac{a}{x}-\frac{b}{1-x}\right)^2-\left(\frac{a}{x^2}+\frac{b}{(1-x)^2}\right)\right]B(x;a,b)$$

For $x \in [x_f, x_s]$ $\quad x(a) = \dfrac{e^{H_a-H_{m-a}}}{1+e^{H_a-H_{m-a}}}$ but to compute $\dfrac{dx}{da}$

we first need to be able to compute $\dfrac{dH_t}{dt}$ recall $H_t = \displaystyle\sum_{k=1}^{\infty}\frac{1}{k}-\frac{1}{k+t}$

$$\frac{dH_t}{dt} = \sum_{k=1}^{\infty}\frac{1}{(k+t)^2} = \psi_1(t)^{*} \text{ In general, } \frac{d^nH_t}{dt^n} = \psi_n(t)^{\dagger} = \sum_{k=1}^{\infty}\frac{(-1)^{n+1}n!}{(k+t)^{n+1}} \quad \therefore$$

$$\frac{dx}{da} = \frac{e^{H_a-H_{m-a}}}{(1+e^{H_a-H_{m-a}})^2}(\psi_1(a)+\psi_1(m-a)) = \frac{\psi_1(a)+\psi_1(m-a)}{1+e^{H_a-H_{m-a}}}x$$

$$\frac{d^2x}{da^2} = \frac{\psi_2(a)-\psi_2(m-a)}{1+e^{H_a-H_{m-a}}}x - \frac{e^{H_a-H_{m-a}}(\psi_1(a)+\psi_1(m-a))^2}{(1+e^{H_a-H_{m-a}})^2}x + \frac{(\psi_1(a)+\psi_1(m-a))^2}{(1+e^{H_a-H_{m-a}})^2}x$$

$$1-2x = 1-\frac{2e^{H_a-H_{m-a}}}{1+e^{H_a-H_{m-a}}} = \frac{1-e^{H_a-H_{m-a}}}{1+e^{H_a-H_{m-a}}} \quad \therefore$$

---

*This is the trigamma function.

†Using nonstandard notation of k starting at 1 and not at 0 for the polygamma functions,ψ_n.

$$\frac{d^2x}{da^2} = \frac{x}{1+e^{H_a\text{-}H_{m\text{-}a}}}\left(\psi_2(a)\text{-}\psi_2(m\text{-}a) + (1\text{-}2x)\big(\psi_1(a)+\psi_1(m\text{-}a)\big)^2\right)$$

For $x \in [x_f, x_s]$ $\quad \pi_u(a) = B(x(a); a; m\text{-}a) = \dfrac{(m+1)!}{a!(m\text{-}a)!}\dfrac{e^{a(H_a\text{-}H_{m\text{-}a})}}{(1+e^{H_a\text{-}H_{m\text{-}a}})^m}$

$$\ln\pi_u(a) = \ln((m+1)!)\text{-}\ln(a!)\text{-}\ln((m\text{-}a)!) + a(H_a\text{-}H_{m\text{-}a})\text{-}m\ln\big(1+e^{H_a\text{-}H_{m\text{-}a}}\big)$$

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

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

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

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

Surprisingly $\dfrac{d^2f}{dx^2} \not\Leftrightarrow 1/\dfrac{d^2x}{df^2}$ $\quad$ Proof: $f(x)=x^2 \quad \dfrac{d^2f}{dx^2}=2 \quad 1/\dfrac{d^2x}{df^2} = -4x^3$

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

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

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

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

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

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

$$\frac{d^2\pi_u}{dx^2} = \pi_u\left(\frac{da}{dx}\right)^2\left(\frac{d^2\ln\pi_u}{da^2} + \frac{d\ln\pi_u}{da}\left(\frac{d\ln\pi_u}{da}\text{-}\frac{da}{dx}\frac{d^2x}{da^2}\right)\right) \text{ for } x\in[x_f, x_s]$$

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

$\varepsilon \leftarrow \dfrac{6.61889}{n_a^3}$

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

$x \leftarrow x_f$

$\pi_u, \pi_l = B_f, B_s$

While $x \geq 0$:

    Sample π_u, π_l

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

$a \leftarrow a_s$

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

While $a \leq a_s + d$:

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

    Sample π_u

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

$x \leftarrow x_c$

While $x \geq x_f$:

    Sample π_l

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

$x \leftarrow x_c$

$\pi_u, \pi_l = B_s, B_f$

While $x \leq x_s$:

    Sample π_l

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

While $x \leq 1$:

    Sample π_u, π_l

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

The author's implementation of the code can be found at *github.com/RaleighPriour/Robust-Bayesian-Analysis [Priour(2025)]*

The last thing we will show is that the highest probability density value of π_u is in the distribution with its mode furthest from $\frac{1}{2}$ which is $B_f(\frac{a_s}{m})$ when $2a_s \leq n$ and $B_s(\frac{a_s+d}{m})$ when $2a_s \geq n$. As well as, Mode $B_f \leq x_f$ and $x_s \leq$ Mode B_s. We shall do this by looking at the structure of the beta family.

By definition $\max\{B(x; a, b) \mid x \in [0, 1]\} = B(\text{Mode } B(a, b); a, b)$.
Taking the gradient of the beta family with the parameterization of the mode denoted as ν and n one gets

$$B(\nu, n\nu, n(1 - \nu)) = \frac{(n+1)!}{(n\nu)!(n(1 - \nu))!}\nu^{n\nu}(1 - \nu)^{n(1-\nu)} \quad \therefore$$

$$\frac{\partial \ln B}{\partial \nu} = n(H_{n\nu} - \gamma) - n(H_{n(1-\nu)} - \gamma) + n + n\ln(\nu) - n - n\ln(1 - \nu)$$

$$\frac{\partial \ln B}{\partial \nu} = n(H_{n\nu} - H_{n(1-\nu)}) + n\ln\Big(\frac{\nu}{1-\nu}\Big)$$

Now notice that $\dfrac{\partial \ln B}{\partial \nu}$ is 0 and odd* about the point $\nu = \dfrac{1}{2}$ and is increasing Therefore with respect to ν the beta distribution achieves a minimum at $\nu = \frac{1}{2}$ and symmetrically increases from $\nu = \frac{1}{2}$.

Now let us look at the beta distribution's behavior with respect to n

$$\frac{\partial \ln B}{\partial n} = H_{n+1} - \gamma - \nu(H_{n\nu} - \gamma) - (1-\nu)(H_{n(1-\nu)} - \gamma) + \nu\ln(\nu) + (1-\nu)\ln(1-\nu)$$

$$\frac{\partial \ln B}{\partial n} = H_{n+1} - \nu H_{n\nu} - (1-\nu)H_{n(1-\nu)} + \nu\ln(\nu) + (1-\nu)\ln(1-\nu)$$

It is obvious that $\dfrac{\partial \ln B}{\partial n}$ is non-decreasing with respect to n.

Evaluating $\dfrac{\partial \ln B}{\partial n}$ at $n = 0$ one gets $\dfrac{\partial \ln B}{\partial n}\Big|_{n=0} = 1 + \nu\ln(\nu) + (1-\nu)\ln(1-\nu)$

Which has a minimum of 1 - ln(2) at $\nu = \frac{1}{2}$ which one can easily verify. Therefore the probability density at any beta distribution's modal value increases with respect to n, which makes intuitive sense the more data one collects the more informed their posterior distribution should be. From the gradient of the beta distribution we see that the highest probability density value of π_u is in the distribution with its mode furthest from $\frac{1}{2}$ which is $B_f(\frac{a_s}{m})$ when $2a_s \leq n$ and $B_s(\frac{a_s+d}{m})$ when $2a_s \geq n$

Finally, we will show Mode $B_f \leq x_f$ and $x_s \leq$ Mode B_s. By showing the following inequalities:

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

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

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

Recall $x(a) = 1 - \dfrac{1}{1 + e^{H_a - H_{m-a}}}$ and $\ln\Big(\dfrac{x(a)}{1 - x(a)}\Big) = H_a - H_{m-a}$

$$H_a - H_{m-a} = \sum_{k=1}^{\infty}\frac{1}{k} - \frac{1}{k+a} - \sum_{k=1}^{\infty}\frac{1}{k} - \frac{1}{k+m-a} = \sum_{k=1}^{\infty}\frac{1}{k+m-a} - \frac{1}{k+a}$$

assume $a > \dfrac{m}{2}$ so the sum is positive and therefore

* $\frac{\partial \ln B}{\partial \nu}(\nu = \frac{1}{2} - t) = -\frac{\partial \ln B}{\partial \nu}(\nu = \frac{1}{2} + t)$

$$\ln\left(\frac{x(a)}{1 - x(a)}\right) = \sum_{k=1}^{\infty} \frac{1}{k+m-a} - \frac{1}{k+a} < \int_0^{\infty} \frac{1}{k+m-a} - \frac{1}{k+a}\, d\boldsymbol{k}$$

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

$$\frac{x(a)}{1-x(a)} < \frac{a}{m-a} \Leftrightarrow x(a) < \frac{a}{m-a} - \frac{a}{m-a}x(a) \Leftrightarrow \frac{m}{m-a}x(a) < \frac{a}{m-a}$$

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

If $a = \frac{m}{2}$ then $x(a) = 1 - \frac{1}{1+e^{H_{m/2} - H_{m-m/2}}} = 1 - \frac{1}{2} = \frac{1}{2}$ and Mode $B(\frac{m}{2}, m) = \frac{1}{2}$

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

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

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

6 Discussion

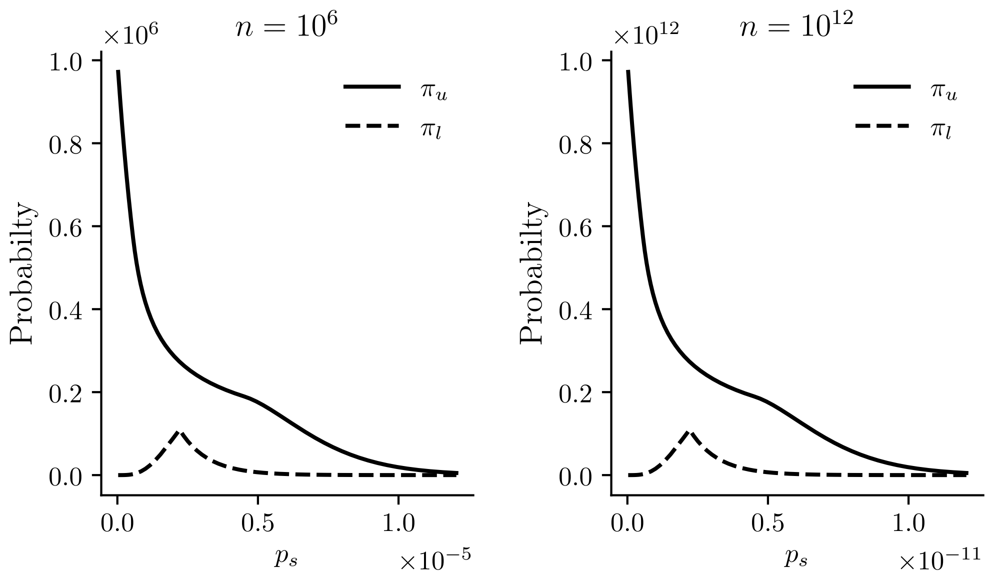
6.1 Implementation

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 matplotlib and other libraries use of 64 bit numbers the graphs outputted become unstable; however, the actually statistics work for up to $n \approx 10^{150}$ after that python is unable to handle the numerical values. I also tried the procedure of calculating the next sample point where the difference between the linear and quadratic approximation reach some threshold. Along with a method where the angle between the points plotted reached some threshold from 180°. Both methods worked significantly worse than the original cumulative error method.

*https://github.com/RaleighPriour/Robust-Bayesian-Analysis

6.2 Case where $a_s = 0$ or $b_f = 0$

Unlike the cases where $\lim_{n\to\infty} a_s, b_f \to \infty$ where the beta distribution approaches the normal distribution. In these cases, π_u and π_l approach the same 2 relative curves that are most certainly not the normal distribution. Here is an example with $d = 4$



6.3 Choosing Degree of Prior Certitude

This is a very interesting question. Considering the S&P 500 for our discussion, 30 years is a long time with the dataset containing 7559 trade days. An initial thought is to first start with the baseline prior and then calculate one's posterior having $d = \lambda\,\mathbb{E}[\sigma] = \lambda\sqrt{n}\,\mathbb{E}[\sqrt{p_s(1\text{-}p_s)}]$. However, by doing this ones is changing the hyperparameter, d, based on the data and cannot define d before the data is collected. What one can do instead is compute the $\lambda\sqrt{n}\,\mathbb{E}[\sqrt{p_s(1\text{-}p_s)}]$ of their baseline prior. For the S&P 500 analysis $d = 100$, $\lambda \approx 2.93$ and for $d = 30$ is $\lambda \approx 0.88$. However, this again is not the point of robust Bayesian analysis since the amount of reasonable priors do not increase proportionally to the $\sqrt{n}$. Moreover, the point of robust Bayesian statistics is not to be a worse way of computing confidence intervals. Using confidence intervals in robust Bayesian statistics allows at least γ credibility* that the parameter of interest, θ is in $\mathcal{R}(\mathrm{Crl}_\gamma(\theta))$. So when using credible intervals scaling d by $\sqrt{n}$ is unneeded and just allows one's credible intervals for their parameters to be excessively vague.

6.4 $[x_f, x_s]$ Region

From the Analysis of the S&P 500 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.

6.5 Conclusion

This paper has derived the basic descriptive statistics for robust Bayesian analysis of the beta distribution and computed all the steps to create an very effective graphing algorithm for π_l and π_u. Additionally, it showed how robust Bayesian analysis can be used in the real world by determining the likelihood that the S&P 500 will be higher at the end of the next trading day than it was at the last trading day is between $52\%-55\%$.

References

[Han(2020)] Henry Han. S&P 500 Historical Data. *Kaggle*, 2020. URL `https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data`.

[Murray and Teare(1993)] Daniel B. Murray and Scott W. Teare. Probability of a tossed coin landing on edge. *Phys. Rev. E*, 48:2547–2552, Oct 1993. doi: 10.1103/PhysRevE.48.2547. URL `https://link.aps.org/doi/10.1103/PhysRevE.48.2547`.

[Priour(2025)] Raleigh Priour. Robust-bayesian-analysis. *GitHub*, 2025. URL `https://github.com/RaleighPriour/Robust-Bayesian-Analysis`.

*$\mathrm{Crl}_\gamma(\theta; \pi \in \Pi) \in \mathcal{R}(\mathrm{Crl}_\gamma(\theta))$