Notes accompanying BDA chapter 4

Asymptotics and connections to non-Bayesian approaches

Jeff Miller

March 4, 2016

Contents

In the (subjective) Bayesian framework, there is no way to objectively evaluate the performance of a procedure—you assume a particular model, and all your inferences are based on the assumption that the model is correct. In order to assess a Bayesian procedure, we need to step outside the Bayesian framework. This chapter discusses some of the non-Bayesian properties that are often considered when evaluating a procedure. There is also some discussion of connections between non-Bayesian and Bayesian approaches. Model checking (BDA chapter 6) and assessment of predictive performance (BDA chapter 7) also fall into the category of non-Bayesian methods of assessment (although some people might consider these Bayesian). Note: This chapter is more theoretical than most of the material in the course. A word of caution—some of the results in Appendix B of BDA seem to be missing some crucial conditions, and the proofs are very sketchy in places, so I would not recommend using Appendix B as a reference.

1 Example and counterexample

1.1 Example

First, let's look at a really simple example illustrating posterior consistency and asymptotic normality. Some of the terminology here will be explained in the sections below. Consider the following model:

$$\begin{aligned}\theta &\sim \text{Exp}(1)\\ X_1, \ldots, X_n|\theta &\sim \text{Exp}(\theta),\end{aligned}$$

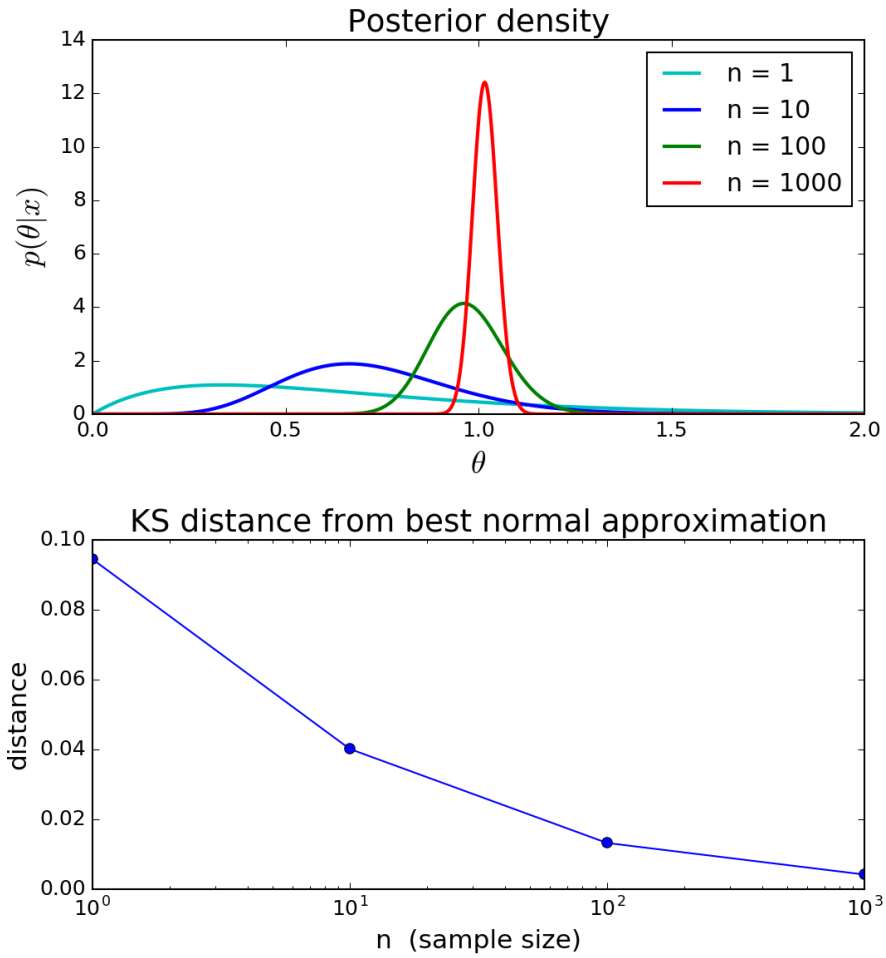
and suppose the true distribution P_0 is Exp(1). So, the model is correctly specified and the true value of the parameter is $\theta_0 = 1$. The posterior distribution is $\theta|x_{1:n} \sim \text{Gamma}(1+n, 1+\sum_i x_i)$. As illustrated in the first figure below, the posterior density appears to become more Gaussian-like as the sample size n increases, and concentrates near the true value, $\theta_0 = 1$.

Indeed, the second figure indicates that the Kolmogorov–Smirnov (KS) distance[1] between the posterior and the "best" normal approximation to it (defined here as the normal distribution with the same mean and variance as the posterior) appears to be going to zero as n increases. In other words, the posterior is becoming more normal/Gaussian as n increases. The fact that the KS distance is going to zero is not merely due to the fact that the posterior is concentrating. The KS distance is invariant under affine transformations of θ, i.e.,

$$\sup_{\theta\in\mathbb{R}} |F(\theta) - G(\theta)| = \sup_{\theta\in\mathbb{R}} |F(a\theta + b) - G(a\theta + b)|$$

for any $a \neq 0$, $b \in \mathbb{R}$. So in particular, KS is not dependent on the scale.

[1] The Kolmogorov–Smirnov distance between univariate distributions with CDFs F and G is $KS(F, G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|$.

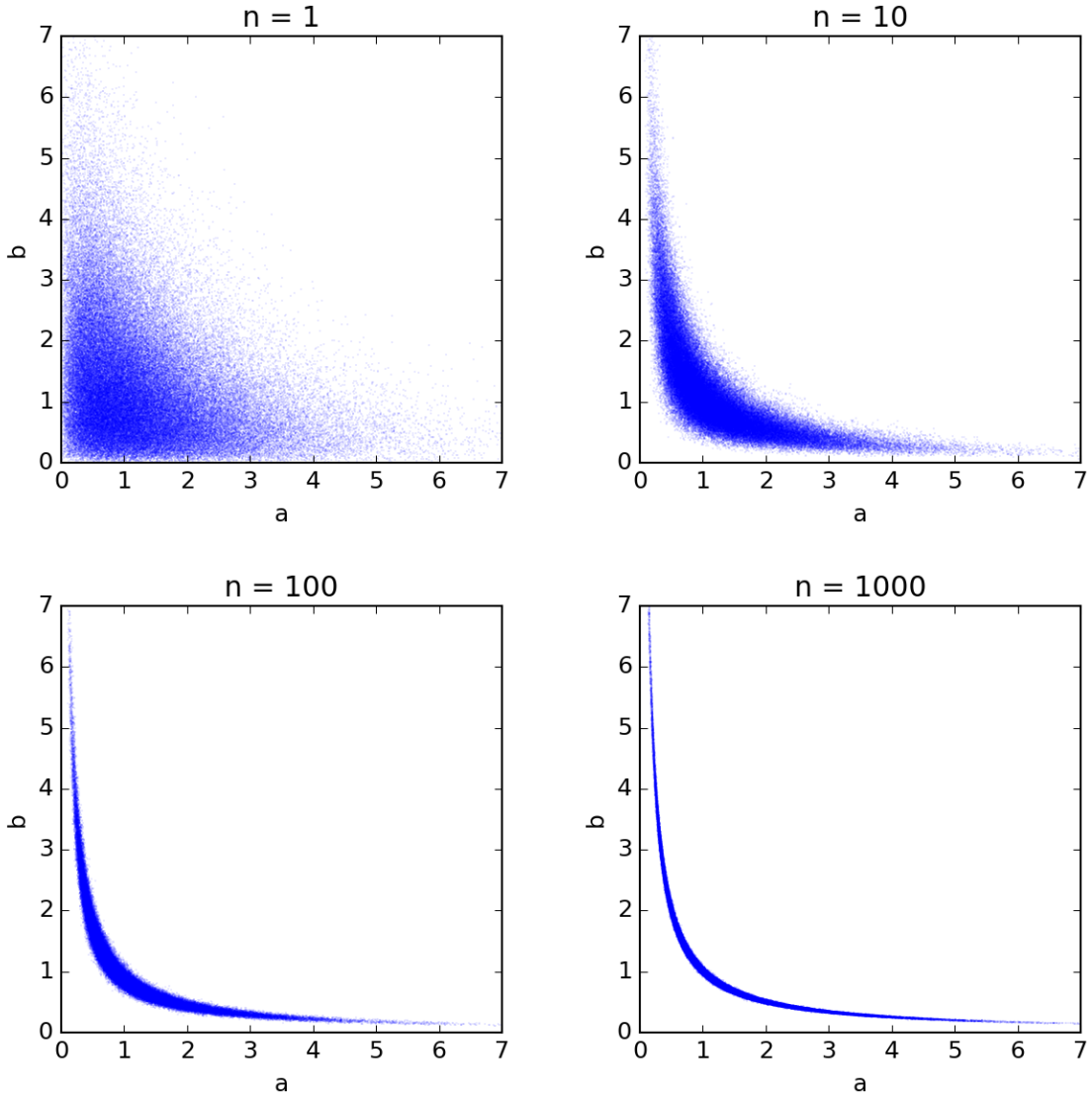


1.2 Counterexample

Now, let's look at an example in which posterior consistency and asymptotic normality do not hold. There are a number of ways in which this can happen, and they do occur in practice, so one needs to be careful. One of the basic requirements is that the parameter be identifiable—that is, that if we knew P_θ exactly, we could uniquely recover θ (i.e., if $\theta \neq \theta'$ then $P_\theta \neq P_{\theta'}$). To illustrate this with a very simple example, consider the model:

$$\begin{aligned} a, b &\sim \text{Exp}(1) \\ X_1, \ldots, X_n | a, b &\sim \text{Exp}(ab), \end{aligned}$$

and suppose the true values of the parameters are $a_0 = 1$ and $b_0 = 1$, so that the true distribution P_0 is Exp(1). Note that any values of a and b such that $ab = 1$ will give rise to a distribution that matches P_0, so if we define $\theta = (a, b)$, then θ is not identifiable. To see what happens if we sample from the posterior, the figures below show 10^5 Gibbs samples, for each $n \in \{1, 10, 100, 1000\}$.



Note that the posterior is concentrating on the curve satisfying the equation $ab = 1$. However, it is not concentrating at any particular point on the curve, and it is far from normal/Gaussian. Incidentally, this also serves as an example in which Gibbs sampling mixes poorly when n is large. (Why?)

2 Consistency

In the frequentist setting, consistency and asymptotic normality are two basic properties of estimators. In the Bayesian setting, we also consider consistency and asymptotic normality of posterior distributions, which, while similar, are slightly different than the corresponding properties of estimators.

2.1 Consistency of estimators

Roughly speaking, an estimator $\hat{\beta}_n$ of a quantity of interest β is said to be *consistent* if $\hat{\beta}_n \to \beta$ as $n \to \infty$, in other words, if it is guaranteed to converge to the true value. What exactly does this mean? From the frequentist perspective, the observed data $X_{1:n}$ are viewed as random variables generated from some unknown "true" distribution P_0. Since $\hat{\beta}_n$ is a function of the observed data $X_{1:n}$—often denoted by writing $\hat{\beta}_n = \hat{\beta}_n(X_{1:n})$—it follows that $\hat{\beta}_n$ is a random variable as well, from this perspective. Further, let's consider β to be some property of P_0—that is, $\beta = \beta(P_0)$—for example, β might be the mean of some statistic of interest, or the mean and covariance matrix, or some other vector of properties. Then, more precisely, (almost sure) consistency occurs when $\hat{\beta}_n(X_{1:n}) \to \beta(P_0)$ with probability 1, for all P_0 in some relevant class. The phrase "almost surely", often abbreviated a.s., means "with probability 1". Consistency in terms of weaker modes of convergence, such as convergence in probability, can also be useful.[2]

2.2 Consistency of posterior distributions

So, now we know what it means for an estimator to be consistent. What does it mean for a posterior distribution to be consistent? Roughly speaking, the posterior is consistent for β if the posterior distribution on β concentrates in neighborhoods of the true value $\beta(P_0)$. To make this more precise, we first need to have a model $p(x|\theta)$ and a prior $p(\theta)$. Let's use P_θ to denote the distribution with density $p(x|\theta)$. (Sometimes, the quantity of interest β will be θ itself, but this requires one to assume the model is "correctly specified"—see definition below). Formally, the posterior is said to be (almost surely) *consistent for* β if for any $\varepsilon > 0$,

$$\mathbb{P}\big(|\beta(P_\theta) - \beta(P_0)| > \varepsilon \;\big|\; X_{1:n}\big) \to 0$$

with probability 1, as $n \to \infty$, for all P_0 in some relevant class. To understand what this is saying, it is crucial to note that we are dealing with two different probability models here: the true distribution, P_0, which is the distribution of $X_{1:n}$, and the assumed model family, P_θ, which is used to compute the posterior. To clarify further, for any given $x_{1:n}$, we can write the conditional probability above as

$$\mathbb{P}\big(|\beta(P_\theta) - \beta(P_0)| > \varepsilon \;\big|\; x_{1:n}\big) = \int \mathbb{1}\big(|\beta(P_\theta) - \beta(P_0)| > \varepsilon\big) p(\theta|x_{1:n}) d\theta.$$

(Here, $\mathbb{1}(E)$ is the indicator function, which equals 1 if E is true, and equals 0 otherwise.) When $x_{1:n}$ are replaced by random variables $X_{1:n}$ with distribution P_0, this conditional probability becomes a random variable, and consistency occurs if it converges to 0 (a.s., or in probability, etc.) for any $\varepsilon > 0$.

---

[2] Definition: $Z_n \to Z$ a.s. if $\mathbb{P}(\lim_n Z_n = Z) = 1$.
Definition: $Z_n \to Z$ in probability if $\mathbb{P}(|Z_n - Z| > \varepsilon) \to 0$ for all $\varepsilon > 0$.

2.3 Consistency guarantees for Bayesian models

First, let's assume that the model is "correctly specified", that is, $P_0 = P_{\theta_0}$ for some θ_0. If θ is finite-dimensional, then under quite general conditions, the posterior mean of $\beta(P_\theta)$ will be a consistent estimator—that is, $\mathbb{E}(\beta(P_\theta)|X_{1:n}) \to \beta(P_0)$—and the posterior will be consistent for β. When θ is infinite-dimensional, things are quite a bit more subtle, and the study of asymptotic properties such as consistency in infinite-dimensional cases has been an area of research in recent years.

Usually, consistency theorems require several regularity conditions, but there is a remarkable result called Doob's theorem that is relatively easy to understand and applies very generally. Roughly, Doob's theorem says that with probability 1, if the true parameter θ_0 is drawn from the prior, and $X_1, \ldots, X_n$ are drawn i.i.d. from P_{θ_0}, then

1. the posterior mean of $\beta(P_\theta)$ is a consistent estimator, and
2. the posterior distribution of $\beta(P_\theta)$ is consistent.

Doob's theorem is very general[3], but if you think carefully about the statement of the theorem, you will see that it has one big weakness—it only guarantees consistency on a set of θ_0's that has probability 1 under the prior. In the finite-dimensional setting, this is not such a big deal, but in the infinite-dimensional setting, it is a significant limitation.

So far, we've been assuming that the model is correctly specified. What if the model is misspecified, i.e., what if there is no θ_0 such that $P_0 = P_{\theta_0}$? Typically, what happens in this case is that the posterior concentrates at a point θ^* minimizing the Kullback–Leibler divergence from P_0, that is, at $\theta^* = \operatorname{argmin}_\theta D(p_0 \| p_\theta)$, where

$$D(p_0 \| p_\theta) = \int p_0(x) \log \frac{p_0(x)}{p_\theta(x)} dx,$$

assuming P_0 and P_θ have densities p_0 and p_θ, respectively. This makes intuitive sense, since

$$\begin{aligned} \operatorname*{argmin}_\theta D(p_0 \| p_\theta) &= \operatorname*{argmax}_\theta \int p_0(x) \log p_\theta(x) dx \approx \operatorname*{argmax}_\theta \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) \\ &= \operatorname*{argmax}_\theta \prod_{i=1}^n p_\theta(X_i), \end{aligned}$$

which is the maximum likelihood estimator (MLE).

[3]The main assumptions are identifiability of θ and certain measurability conditions; also for (1), the prior mean of $\beta(P_\theta)$ needs to exist. The proof is an elegant application of martingale theory, which Doob himself developed, and which is now a cornerstone of advanced probability.

3 Asymptotic normality

What is the point of establishing asymptotic normality of the posterior? How is it useful? Having a simple interpretation of the asymptotic behavior of the posterior is useful for a variety of purposes. For example, when appropriate, normal approximations to the posterior can significantly reduce computation (recall that Gaussians are particularly nice to work with). Additionally, having a good intuition for the asymptotic behavior of a model can be very helpful when determining what modeling assumptions are appropriate for a given problem. Further, asymptotic normality can be used to ensure that the posterior is correctly calibrated in terms of frequentist coverage.

3.1 Approximate normality of the posterior

We will focus on the intuition, without going into rigorous details. Assume θ is finite-dimensional. Given any (sufficiently smooth) function $f(\theta)$, and a point $\hat{\theta}$ in the interior of its domain, we can approximate $f(\theta)$ near $\hat{\theta}$ using a second-order Taylor approximation:

$$f(\theta) \approx f(\hat{\theta}) + f'(\hat{\theta})^\intercal(\theta - \hat{\theta}) + \tfrac{1}{2}(\theta - \hat{\theta})^\intercal f''(\hat{\theta})(\theta - \hat{\theta}),$$

where $f'(\theta)$ is the gradient and $f''(\theta)$ is the Hessian matrix, i.e., $f'(\theta)_i = \frac{\partial f}{\partial \theta_i}(\theta)$ and $f''(\theta)_{ij} = \frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\theta)$. If we choose $f(\theta) = \log p_\theta(x_{1:n})$, and let $\hat{\theta} = \hat{\theta}_n(x_{1:n})$ be the MLE, then the second term vanishes and we have

$$\log p_\theta(x_{1:n}) \approx \log p_{\hat{\theta}}(x_{1:n}) - \tfrac{1}{2}(\theta - \hat{\theta})^\intercal I(\hat{\theta}; x_{1:n})(\theta - \hat{\theta}),$$

where $I(\theta; x_{1:n})$ is the *observed information matrix*, defined[4] as the matrix in which entry (i, j) is $I(\theta; x_{1:n})_{ij} = -\sum_{k=1}^n \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_\theta(x_k)$. Exponentiating both sides yields

$$p_\theta(x_{1:n}) \mathrel{\tilde{\propto}_\theta} \exp\big(-\tfrac{1}{2}(\theta - \hat{\theta})^\intercal I(\hat{\theta}; x_{1:n})(\theta - \hat{\theta})\big) \propto_\theta \mathcal{N}\big(\theta \mid \hat{\theta}, I(\hat{\theta}; x_{1:n})^{-1}\big),$$

where $\mathrel{\tilde{\propto}_\theta}$ means "approximately proportional to, as a function of θ". So, the likelihood is approximately proportional to a normal distribution with mean equal to the MLE, and precision equal to the observed information. Note that the observed information grows with n, and thus, the likelihood becomes more and more concentrated around the MLE as n increases. In particular, if the prior density $p(\theta)$ is continuous at $\hat{\theta}$ and $p(\hat{\theta}) > 0$, then $p(\theta)$ will be approximately constant over all θ near $\hat{\theta}$ when n is sufficiently large, and therefore, the posterior will behave similarly to the likelihood:

$$p(\theta | x_{1:n}) \propto p_\theta(x_{1:n}) p(\theta) \mathrel{\tilde{\propto}} p_\theta(x_{1:n}) \mathrel{\tilde{\propto}} \mathcal{N}\big(\theta \mid \hat{\theta}, I(\hat{\theta}; x_{1:n})^{-1}\big). \tag{3.1}$$

Thus, the posterior is approximately normal when n is large.

[4] BDA defines it to include the prior, but the definition here is more standard.

3.2 An interesting symmetry

Further insight into the asymptotic normality of the posterior can be obtained from the following thought experiment. Suppose $\hat{\theta}|\theta \sim \mathcal{N}(\theta, C)$ and θ is given a prior that is very diffuse relative to C. Then

$$p(\theta|\hat{\theta}) \propto_\theta p(\hat{\theta}|\theta)p(\theta) \lesssim_\theta p(\hat{\theta}|\theta) = \mathcal{N}(\hat{\theta} \mid \theta, C) = \mathcal{N}(\theta \mid \hat{\theta}, C).$$

Thus, we would have both

$$\hat{\theta}|\theta \sim \mathcal{N}(\theta, C) \quad \text{and} \quad \theta|\hat{\theta} \approx \mathcal{N}(\hat{\theta}, C).$$

It turns out that this thought experiment fairly accurately represents what happens in many models (subject to some regularity conditions, of course) when the sample size is sufficiently large and $\hat{\theta}$ is the MLE. In exponential families, $p(\theta|x_{1:n}) = p(\theta|\hat{\theta})$ (the technical term here is that $\hat{\theta}$ is a *sufficient statistic*). In fact, in many other models $p(\theta|x_{1:n}) \approx p(\theta|\hat{\theta})$ — basically, once we know $\hat{\theta}$, knowing $x_{1:n}$ doesn't tell us much more about θ. So, combined with equation 3.1, this explains the $\theta|\hat{\theta}$ part:

$$\theta|\hat{\theta} \approx \theta|x_{1:n} \approx \mathcal{N}\big(\hat{\theta}, I(\hat{\theta}; x_{1:n})^{-1}\big).$$

What about $\hat{\theta}|\theta$? It is a classical result that the MLE is asymptotically normally distributed[5]:

$$\hat{\theta}|\theta \approx \mathcal{N}(\theta, I(\hat{\theta}; x_{1:n})^{-1}).$$

3.3 Discussion

Theorems proving asymptotic normality of the posterior (usually in a stronger sense than what the discussion above would suggest) are often called Bernstein–von Mises results. Asymptotic normality results for the posterior distribution of a function of θ (for example, $\beta(P_\theta)|X_{1:n}$) can be derived from the asymptotic normality of $\theta|X_{1:n}$ using a technique called the delta method.

4 Frequentist coverage

Having posterior consistency gives us a guarantee that the posterior will concentrate near the true parameter value θ_0, however, it is also important that the posterior be appropriately calibrated in terms of how concentrated it is. Roughly speaking, θ_0 should be "well-supported" under the posterior, on average. If the posterior is too concentrated, then θ_0 might fall outside the range of well-supported values, while if the posterior is not concentrated enough, then it will be indicating a greater amount of uncertainty than necessary.

[5]The precise statement is that $I(\hat{\theta}_n; X_{1:n})^{1/2}(\hat{\theta}_n - \theta)$ converges in distribution to $\mathcal{N}(0, I)$.

From the subjective Bayesian perspective, the posterior is always correctly calibrated with respect to the assumed prior beliefs. Sensitivity analysis can be used to see how much the posterior depends on the particular prior assumed, however, some set of priors must still be chosen. Is there a more objective method of evaluation?

From the frequentist perspective, uncertainty about parameters is usually communicated using confidence intervals (or more generally, confidence regions) rather than posterior distributions. A confidence region $C(x_{1:n})$ is a subset of parameter values that depends on the data $x_{1:n}$. The coverage probability of $C(x_{1:n})$ is the probability that $C(X_{1:n})$ will contain the true parameter θ_0 when the data $X_{1:n}$ is generated according to θ_0:

$$\mathbb{P}\big(\theta_0 \in C(X_{1:n}) \;\big|\; \theta_0\big). \tag{4.1}$$

Confidence regions are usually constructed with the intent of providing coverage as close as possible to some user-specified level for all θ_0. For instance, ideally, a 95% confidence interval would have coverage equal to 0.95 for all θ_0.

How can we use this concept to evaluate the calibration of a posterior? Well, we could use the posterior to construct a credible region with posterior probability equal to the desired coverage, and see how well it attains that coverage. For example, a 95% credible region is a subset $C(x_{1:n})$ of parameter values with the property that

$$\mathbb{P}\big(\theta \in C(x_{1:n}) \;\big|\; x_{1:n}\big) = 0.95.$$

Note that in this expression, θ is a random variable and $x_{1:n}$ is fixed, while in equation 4.1, $X_{1:n}$ is a random variable and θ_0 is fixed. (This is essentially *the* difference between Bayesianism and frequentism.) Interestingly, although they are not specifically designed to do so, Bayesian credible intervals often have very good frequentist coverage—sometimes even better than standard frequentist confidence intervals.

To illustrate, consider the following model:

$$\begin{aligned} p &\sim \text{Beta}(1,1) \\ X_1, \ldots, X_n | p &\sim \text{Bernoulli}(p). \end{aligned}$$

A classical frequentist approach to constructing a 95% confidence interval for p is the Wald-type interval:

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

where $\hat{p}$ is the sample mean, $\bar{x}$. A standard Bayesian approach to constructing a 95% credible interval for p is the equal-tailed interval $[a(x_{1:n}), b(x_{1:n})]$ where

$$\mathbb{P}(\theta < a(x_{1:n})|x_{1:n}) = \mathbb{P}(\theta > b(x_{1:n})|x_{1:n}) = 0.025.$$

The figures below show typical intervals for these two methods, as well as their coverage probabilities, for increasing sample sizes n, when the true distribution is Bernoulli(0.1) (i.e., when the true value of p is 0.1). There are a few salient points to note. First,

and most importantly, the credible interval has significantly better coverage than the Wald-type confidence interval when $n < 100$. Second, when $n = 1$, the Wald confidence interval is always degenerate at either 0 or 1, so it has coverage equal to zero. Third, the Wald interval may contain values outside $[0, 1]$, which is unnecessary, of course. Finally, note that both intervals have coverage tending to 0.95 as n increases.