An Overview of Bayesian Analysis

Sourabh Bhattacharya

Indian Statistical Institute

Overview of the material to be covered in this lecture

- Paradoxes of the classical paradigm.
- Advantages of the Bayesian paradigm.
- Bayesian posterior calculations.
- Choice of priors.
- Credible sets and highest posterior density credible sets.
- Bayesian hypothesis testing and Bayes factors.

Fuller treatment of all these issues can be found in the books referred to at the end.

Paradoxes of the Classical Paradigm

Confidence Interval

1. Suppose $X_i \overset{iid}{\sim} N(\theta, \sigma^2)$; $i = 1, \ldots, n$, where N denotes the normal (Gaussian) distribution and iid stands for "independent and identically distributed."
2. We desire a 95% interval estimate for the population mean θ.
3. Provided n is sufficiently large (say, bigger than 30), a classical approach would use the confidence interval $\delta(\mathbf{x}) = \bar{x} \pm 1.96s/\sqrt{n}$, where $\mathbf{x} = (x_1, \ldots, x_n)$, $\bar{x}$ is the sample mean, and s is the sample standard deviation.
4. This interval has the property that on the average over repeated application, $\delta(\mathbf{x})$ will fail to capture the true mean only 5% of the time.
5. An alternative interpretation is that, before any data are collected, the probability that the interval contains the true value is 0.95.
6. This property is attractive in the sense that it holds for all true values of θ and σ^2.

Paradoxes of the Classical Paradigm

Confidence Interval (continued)

1. On the other hand, its use in any single data-analytic setting is somewhat difficult to explain and understand.
2. After collecting the data and computing $\delta(\mathbf{x})$ the interval either contains the true value or it does not; its coverage probability is not 0.95, but either 0 or 1.
3. After observing $\mathbf{x}$, a statement like, "the true θ has a 95% chance of falling in $\delta(\mathbf{x})$" is not valid, though most people (including most statisticians irrespective of their philosophical approach) interpret a confidence interval in this way.
4. Thus for the frequentist, "95%" is not a conditional coverage probability, but rather a tag associated with the interval to indicate either how it is likely to perform before we evaluate it, or how it would perform over the long haul. A 99% frequentist interval would be wider, a 90% interval narrower, but, conditional on $\mathbf{x}$, all would have coverage probability 0 or 1.

Paradoxes of the Classical Paradigm

An Example on Confidence Interval

1. Let x_1 and x_2 be two random observations from a uniform distribution on the interval $(\theta - 0.5, \theta + 0.5)$.
2. Let y_1 and y_2 be, respectively, the smaller and the larger of these two observations.
3. Then, for all θ,

$$\begin{aligned} P(y_1 < \theta < y_2 | \theta) &= P(x_1 < \theta < x_2) + P(x_2 < \theta < x_1) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \\ &= \frac{1}{2}, \end{aligned} \tag{1}$$

   so that (y_1, y_2) provides a 50% confidence interval.
4. However, the following example highlights a paradox.

Paradoxes of the Classical Paradigm

An Example on Confidence Interval (Continued)

Let the data be $(x_1, x_2) = (7.4, 8.0)$. The 50% confidence interval is then $(\min\{x_1, x_2\}, \max\{x_1, x_2\}) = (7.4, 8.0)$.

- Since we have $\theta - \frac{1}{2} < x_1, x_2 < \theta + \frac{1}{2}$, we must have $\theta + 0.5 > \max\{x_1, x_2\} = 8.0$, so that $\theta > 7.5$.
- We must also have $\theta - 0.5 < \min\{x_1, x_2\} = 7.4$, so that $\theta < 7.9$.
- So, with certainty, we must have $\theta \in (7.5, 7.9)$.
- Since $(7.5, 7.9) \subset (7.4, 8.0)$, the latter being the 50% confidence interval, this is clearly paradoxical.

Paradoxes of the Classical Paradigm

An Example on Confidence Interval (Continued)

More generally, let $y_2 - y_1 \geq 0.5$.

- Since we have $\theta - \frac{1}{2} < x_1, x_2 < \theta + \frac{1}{2}$, we must have $\theta + 0.5 > \max\{x_1, x_2\} = y_2$, so that $\theta > y_2 - 0.5$. Since $y_2 \geq y_1 + 0.5$, this implies $\theta > y_1$.
- We must also have $\theta - 0.5 < \min\{x_1, x_2\} = y_1$, so that $\theta < y_1 + 0.5$. Hence, $\theta < y_2$ (since $y_1 \leq y_2 - 0.5$).
- So, with certainty, we must have $\theta \in (y_1, y_2)$.

Paradoxes of the Classical Paradigm

An Example on Hypothesis Testing

1. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$.
2. Consider $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$; to reject H_0 if $\sqrt{n}|\bar{x}| > 1.96$, where the size of the test is $\alpha = 0.05$.
3. Say, for instance, true $\theta = 10^{-10}$, and sample size $n = 10^{24}$.
4. Then, with extremely high probability, $\bar{x}$ will be within 10^{-11} of true $\theta = 10^{-10}$ (the standard deviation of $\bar{X}$ is only 10^{-12}).
5. Then clearly, $10^{12}|\bar{x}| > 1.96$.
6. So, the classical test is virtually certain to reject H_0 even though the true mean is negligibly different from zero.

This example shows that a "statistically significant" difference between the true parameter and the null hypothesis, in practice, can be an unimportant difference. Similarly, a statistically insignificant difference can be practically important.

Another Example on Hypothesis Testing

1. The effectiveness of a drug is measured by $X \sim N(\theta, 3^2)$; significantly positive value of θ would indicate that the drug is effective.
2. Here we test $H_0 : \theta \leq 0$ vs $H_1 : \theta > 0$.
3. A sample of 9 observations yield $\bar{x} = 1$, say.
4. This is not significant (for a one-tailed test), at say, the 0.05 significance level ($T(\mathbf{x}) = \sqrt{n}\bar{x}/\sigma = \sqrt{9} \times 1/3 = 1 < 1.64 = \tau_{0.05} \Rightarrow$ accept H_0).
5. However, it is significant at $\alpha = 0.16$ significance level (since $T(\mathbf{x}) = 1 > 0.99 = \tau_{0.16} \Rightarrow$ reject H_0), which is moderately convincing.
6. Based solely on the data, we would probably decide that the drug was effective.

Paradoxes of the Classical Paradigm

	C	E	Total
Tumor	0	3	3
No Tumor	50	47	97
Total	50	50	100

Consider a carcinogen bioassay where a control group (C) and an exposed group (E) with 50 rodents in each (see above table) are compared.

Let p_1 denote the proportion of rodents in group C affected by tumors and let p_2 denote the same for group E. We wish to test

$$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 < p_2$$

In the control group, 0 tumors are found; in the exposed group, there are 3.

Paradoxes of the Classical Paradigm (continued)

Let $X =$ the number of tumors in group E. Then

$$[X = x] = \frac{\binom{50}{x}\binom{50}{3-x}}{\binom{100}{3}}.$$

The P-value is given by

$$[X \geq 3] = [X = 3] = \frac{\binom{50}{3}\binom{50}{0}}{\binom{100}{3}} = 0.1212.$$

So, the one-sided Fisher's exact test yields a non-significant P-value of 0.1212.

Paradoxes of the Classical Paradigm

However, a biologist may claim that three tumors in 50 rodents is definitely biologically significant.

This belief may be based on information from other experiments in the same lab in the previous year in which the tumor has never shown up in control rodents. For example, if there were 400 historical controls in addition to the 50 concurrent controls, none with a tumor, then the contingency table takes the following form.

	C	E	Total
Tumor	0	3	3
No Tumor	450	47	497
Total	450	50	500

Now the one-sided P-value becomes 0.0009. Statistical and biological significance are now compatible.

However, it is generally inappropriate simply to pool historical and concurrent information.

Paradoxes of the Classical Paradigm

The Likelihood Principle with Illustration

Definition of the Likelihood Principle: In making inferences or decisions about θ after $\mathbf{x}$ is observed, all relevant experimental information is contained in the likelihood function of the observed $\mathbf{x}$. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ).

Illustration of the Likelihood Principle

1. Suppose in 12 independent tosses of a coin, I observe 9 heads and 3 tails.
2. I wish to test the null hypothesis $H_0 : \theta = \frac{1}{2}$ vs $H_1 : \theta > \frac{1}{2}$ is the true probability of heads.
3. Given only this much information, two choices for the sampling distribution emerge as candidates...

Paradoxes of the Classical Paradigm

Illustration of the Likelihood Principle (continued)

1. Binomial: The number $n = 12$ tosses was fixed beforehand, and the random quantity X was the number of heads observed in the n tosses. Then $X \sim \text{Binomial}(12, \theta)$ and the likelihood function is given by

$$L_1(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{12}{9} \theta^9 (1-\theta)^3.$$

2. Negative binomial: Data collection involved flipping the coin until the third tail appeared. Here, the random quantity X is the number of heads required to complete the experiment, so that $X \sim \text{Negative-Binomial}(r = 3, \theta)$, with likelihood function given by

$$L_2(\theta) = \binom{r+x-1}{x} \theta^x (1-\theta)^r = \binom{11}{9} \theta^9 (1-\theta)^3.$$

Paradoxes of the Classical Paradigm

Illustration of the Likelihood Principle (continued)

1. Under either of these two alternatives, we can compute the P-value corresponding to the rejection region "Reject H_0 if $X \geq c$."
2. Doing so using the Binomial likelihood we obtain

$$\alpha_1 = P_{\theta=\frac{1}{2}}(X \geq 9) = 0.075.$$

3. For the Negative-Binomial likelihood, we obtain

$$\alpha_2 = P_{\theta=\frac{1}{2}}(X \geq 9) = 0.0325.$$

4. Using the usual Type I error level $\alpha = 0.05$ we would reject H_0 if $X \sim$ Negative-Binomial, but not if $X \sim$ Binomial, thus violating the likelihood principle.

Advantages of the Bayesian Paradigm

1. Bayesian methods allow one to formally incorporate prior information.
2. All inferences are conditional on the actual data.
3. Follows the likelihood principle.
4. Bayesian results are rationally and simply interpretable.
5. All Bayesian analyses follow directly from the posterior–no need for separate theories of estimation, hypothesis testing, multiple hypotheses testing, etc.
6. Any question can be directly answered through Bayesian analysis.
7. Bayesian methods possess numerous optimality properties.
8. Does not rely on asymptotic theory at all.
9. Offers complete freedom for realistically modelling any real data or physical phenomenon.
10. However complex the Bayesian model is, arbitrarily accurate inferences can always be obtained via simulation-based methods.

Posterior Distribution as Combination of Likelihood and Prior

Likelihood specification:

$$Y|\theta \sim f(y|\theta).$$

Prior specification:

$$\theta \sim \pi(\theta),$$

where either Y or θ can be vectors.

Posterior distribution: The posterior distribution of θ is given by

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{m(y)},$$

where

$$m(y) = \int \pi(u)f(y|u)du$$

is the marginal density of the data y.

Posterior Calculation: Examples

Example 1:

$$\theta \sim N(\mu, \tau^2)$$
$$Y|\theta \sim N(\theta, \sigma^2)$$

The posterior distribution $\pi(\theta|Y)$ is normal with mean and variance

$$E(\theta|Y) = B\mu + (1 - B)Y \quad \text{and}$$
$$V(\theta|Y) = (1 - B)\sigma^2,$$

where $B = \sigma^2/(\sigma^2 + \tau^2)$.

Posterior Calculation: Examples (continued)

Observations on the posterior calculations in Example 1

1. Since $0 < B < 1$, the posterior mean is a weighted average of the prior mean μ and the direct estimate Y; the direct estimate is pulled back (or shrunk) toward the prior mean.
2. The weight B on the prior mean depends on the relative variability of the prior distribution and the likelihood. If σ^2 is large relative to τ^2 (i.e., our prior knowledge is more precise than the data information), then B is close to 1, producing substantial shrinkage.
3. If σ^2 is small (i.e., our prior knowledge is imprecise relative to the data information), B is close to 0 and the direct estimate is moved very little toward the prior mean.
4. This shrinkage provides an effective tradeoff between variance and bias, with beneficial effects on the resulting mean squared error (since $\text{MSE} = \text{Var} + \text{Bias}^2$).

Posterior Calculation: Examples (continued)

Example 2:

Let Y be the number of events in n independent trials and let θ be the event probability.

Then the likelihood is given by

$$P(Y = y|\theta) = f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Let the prior on θ be the following:

$$\theta \sim \text{Beta}(a, b).$$

Then the posterior of θ is Beta with mean and variance given by

$$\hat{\theta} = E(\theta|Y) = \frac{M}{M+n}\mu + \frac{n}{M+n}\frac{Y}{n} \quad \text{and}$$
$$V(\theta|Y) = \hat{\theta}(1-\hat{\theta})/(M+n),$$

where $M = a + b$ and $\mu = a/(a+b)$.

Choice of Prior

Conjugate Prior

1. In choosing a prior belonging to a specific distributional family $\pi(\theta|\eta)$ some choices may be more computationally convenient than others.
2. In particular, it may be possible to select a member of that family which is conjugate to the likelihood $f(y|\theta)$, that is, one that leads to a posterior distribution $\pi(\theta|y)$ belonging to the same distributional family as the prior.
3. The exponential family of distributions, from which we typically draw our likelihood functions, has conjugate priors.

Example of Conjugate Prior

Suppose that X is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month. We assume a Poisson likelihood, given by

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, \quad x \in \{0, 1, 2, \ldots\}, \quad \theta > 0.$$

Let

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad \theta > 0, \alpha > 0, \beta > 0,$$

that is, $\theta \sim \mathcal{G}(\alpha, \beta)$. The gamma distribution has mean $\alpha\beta$ and variance $\alpha\beta^2$.

The posterior of θ is given by

$$\begin{aligned}\pi(\theta|x) &\propto \pi(\theta)f(x|\theta) \\ &\propto \left(\theta^{\alpha-1}e^{-\theta/\beta}\right)\left(e^{-\theta}\theta^x\right) \\ &= \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)}.\end{aligned}$$

That is, $\theta|x \sim \mathcal{G}(\alpha^*, \beta^*)$, where $\alpha^* = \alpha + x$ and $\beta^* = (1 + 1/\beta)^{-1}$.

Noninformative Priors

Uniform Distribution

1. Suppose the parameter space $\Theta = \{\theta_1, \ldots, \theta_n\}$. Then

$$\pi(\theta_i) = \frac{1}{n}; \quad i = 1, \ldots, n,$$

   is a non-informative prior for θ.

2. Suppose $\Theta = [a, b]$, where $-\infty < a < b < \infty$. Then

$$\pi(\theta) = \frac{1}{b-a}; \quad \theta \in [a, b],$$

   is (arguably) a non-informative prior for θ.

3. Suppose $\Theta = (-\infty, \infty)$. Then, for any $c > 0$,

$$\pi(\theta) = c; \quad \theta \in (-\infty, \infty),$$

   is (arguably) a non-informative (but improper) prior for θ.

Noninformative Priors

Criticism of the Uniform Prior

The uniform prior does not always qualify as non-informative since it is not invariant under one-to-one transformations.

To illustrate, let $\pi(\theta) = 1$; $\theta > 0$. Consider the reparameterization $\gamma = g(\theta) = \log(\theta)$. Then the prior on γ is given by

$$\pi^*(\gamma) = |J|\pi(e^\gamma),$$

where $J = d\theta/d\gamma = e^\gamma$, the Jacobian of the inverse transformation. Hence

$$\pi^*(\gamma) = e^\gamma; \quad -\infty < \gamma < \infty,$$

which is clearly non-uniform.

Noninformative Priors

Jeffreys' Prior

Jeffreys' prior achieves invariance with respect to one-to-one transformations. In the univariate case, Jeffreys' prior is given by

$$\pi(\theta) \propto [I(\theta)]^{1/2},$$

where $I(\theta)$ is the expected Fisher information in the model, given by

$$I(\theta) = -E_{X|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

Note that Jeffreys' prior depends upon the form of the likelihood but not on the actual observed data.

Jeffreys' Prior (continued)

With respect to the previous example on the log-transformation of the uniform prior on $\Theta = (0, \infty)$, note that

$$[I(\gamma)]^{1/2} = [I(\theta)]^{1/2} \left| \frac{d\theta}{d\gamma} \right| .$$

Hence, directly computing Jeffreys' prior for γ produces the same answer as computing the Jeffreys prior for θ and subsequently performing the usual Jacobian transformation to the γ scale.

In the multiparameter case, Jeffreys' prior is given by

$$\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2} ,$$

where $|\cdot|$ denotes determinant and $I(\boldsymbol{\theta})$ is the Fisher's information matrix, with (i,j)-th element

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x|\boldsymbol{\theta}) \right] .$$

Jeffreys' Prior–Special Cases

Suppose that $f(x|\theta, \sigma)$ is normal with mean θ and standard deviation σ.

1. Suppose that σ is known; then the Jeffreys' prior for θ is given by
$$\pi(\theta) = 1; \quad -\infty < \theta < \infty.$$
2. Suppose that θ is known; then the Jeffreys' prior for σ is given by
$$\pi(\sigma) = \frac{1}{\sigma}; \quad \sigma > 0.$$
3. Suppose that both θ and σ are unknown; then the joint Jeffreys' prior for (θ, σ) is
$$\pi(\theta, \sigma) = \frac{1}{\sigma^2}; \quad -\infty < \theta < \infty; \quad \sigma > 0.$$
4. The common approach, however, is to use the following prior for (θ, σ):
$$\pi(\theta, \sigma) = \frac{1}{\sigma}; \quad -\infty < \theta < \infty; \quad \sigma > 0.$$

Jeffreys' Prior–Special Cases (continued)

Consider the Binomial likelihood

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Here the Jeffreys' prior for θ is the Beta$(\frac{1}{2}, \frac{1}{2})$ distribution.

Credible Sets and Highest Posterior Density Credible Sets

A $100(1-\alpha)\%$ credible set for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ satisfying

$$1-\alpha \leq \pi(C|\mathbf{y}) = \int_C \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}.$$

The highest posterior density (HPD) credible set is defined as the set

$$C = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|\mathbf{y}) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant satisfying

$$\pi(C|\mathbf{y}) \geq 1-\alpha.$$

Hypothesis Testing and Bayes Factors

Suppose we have two candidate models $\mathcal{M}_1$ and $\mathcal{M}_2$ for data $\mathbf{Y}$ with respective parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

Under prior densities $\pi_i(\boldsymbol{\theta}_i), i = 1, 2$, the marginal denisties of $\mathbf{Y}$ are given by

$$m(\mathbf{y}|\mathcal{M}_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i; \quad i = 1, 2.$$

The Bayes Factor (BF) is the ratio of posterior odds of $\mathcal{M}_1$ to the prior odds of $\mathcal{M}_1$, and is given by

$$\begin{aligned} BF &= \frac{\pi(\mathcal{M}_1|\mathbf{y})/\pi(\mathcal{M}_2|\mathbf{y})}{\pi(\mathcal{M}_1)/\pi(\mathcal{M}_2)} \\ &= \frac{\frac{m(\mathbf{y}|\mathcal{M}_1)\pi(\mathcal{M}_1)}{m(\mathbf{y})} \Big/ \frac{m(\mathbf{y}|\mathcal{M}_2)\pi(\mathcal{M}_2)}{m(\mathbf{y})}}{\pi(\mathcal{M}_1)/\pi(\mathcal{M}_2)} \\ &= \frac{m(\mathbf{y}|\mathcal{M}_1)}{m(\mathbf{y}|\mathcal{M}_2)}, \end{aligned}$$

the ratio of the observed marginal densities for the two models.

Hypothesis Testing and Bayes Factors (continued)

Consider testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. Then, the Bayes factor is the ratio of the posterior probabilities of the null and the alternative hypotheses over the ratio of the prior probabilities of the null and the alternative hypotheses, that is,

$$\begin{aligned} BF_{01} &= \frac{\pi(\theta \in \Theta_0|\mathbf{y})/\pi(\theta \in \Theta_1|\mathbf{y})}{\pi(\theta \in \Theta_0)/\pi(\theta \in \Theta_1)} \\ &= \frac{m_0(\mathbf{y})}{m_1(\mathbf{y})}, \end{aligned}$$

where

$$m_i(x) = \int_{\Theta_i} f(\mathbf{y}|\theta)\pi_i(\theta)d\theta; \quad i = 0, 1$$

Hypothesis Testing and Bayes Factors (continued)

Rule of thumb:

- if $\log_{10}(BF_{10})$ varies between 0 and 0.5, the evidence against H_0 is poor,
- if it is between 0.5 and 1, it is is substantial,
- if it is between 1 and 2, it is strong, and
- if it is above 2 it is decisive.

Hypothesis Testing and Bayes Factors (continued)

Considering the point-null hypothesis $H_0 : \theta = \theta_0$, we denote by ρ_0 the prior probability that $\theta = \theta_0$ and by g_1 the prior density under the alternative. The prior distribution is then

$$\pi(\theta) = \rho_0 I_{\theta_0}(\theta) + (1 - \rho_0) g_1(\theta),$$

and the posterior probability of H_0 is given by

$$\begin{aligned} \pi(\Theta_0|x) &= \frac{f(x|\theta_0)\rho_0}{\int f(x|\theta)\pi(\theta)d\theta} \\ &= \frac{f(x|\theta_0)\rho_0}{f(x|\theta_0)\rho_0 + m_1(x)}, \end{aligned}$$

where

$$m_1(x) = \int f(x|\theta) g_1(\theta) d\theta$$

is the marginal distribution on H_1.

Hypothesis Testing and Bayes Factors (continued)

This posterior probability can also be written as

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}$$

Similarly, the Bayes Factor is

$$\begin{aligned} B_{01}(x) &= \frac{f(x|\theta_0)\rho_0}{m_1(x)(1-\rho_0)} \Bigg\backslash \frac{\rho_0}{1-\rho_0} \\ &= \frac{f(x|\theta_0)}{m_1(x)}. \end{aligned}$$

Hence,

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{B_{01}(x)}\right]^{-1}.$$

Hypothesis Testing and Bayes Factors (continued)

Let $X \sim \text{Binomial}(n, p)$ and consider testing $H_0 : p = 1/2$ vs $H_1 : p \neq 1/2$. For $g_1(p) = 1$, the posterior probability is then given by

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0} 2^n B(x+1, n-x+1)\right]^{-1}$$
$$= \left[1 + \frac{1-\rho_0}{\rho_0} 2^n \frac{x!(n-x)!}{(n+1)!}\right]^{-1}$$

For instance, if $n = 5$, $x = 3$ and $\rho_0 = 1/2$, then the posterior probability is $15/19$ and the corresponding Bayes Factor B_{01} is $15/8$, which is close to 2.

References

Berger, J. O. (1985). **Statistical Decision Theory and Bayesian Analysis**. *Springer, New York*.

Bernardo, J. M. and A. F. M. Smith (2000). **Bayesian Theory**. *J. Wiley and Sons, New York*.

Carlin, B. P. and T. A. Louis (2000). **Bayes and Empirical Bayes Methods for Data Analysis**. *Chapman and Hall/CRC.*

Robert, C. P. (2007). **The Bayesian Choice**. *Springer, New York*.