Summer Seminar: Philosophy of Statistics
Lecture Notes 9: Bayesian Inference: a brief introduction

Aris Spanos [Summer 2019]

1 Bayesian methods of inference

1.1 The Bayesian approach to statistical inference

In order to avoid any misleading impressions it is important to note that there are numerous variants of Bayesianism; more than 46656 varieties of Bayesianism according to Good (1971)! In this section we discuss some of the elements of the Bayesian approach which are shared by most variants of Bayesianism.

Bayesian inference, like frequentist inference, begins with a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, but modifies the inferential set up in two crucial respects:

(i) the unknown parameter(s) $\boldsymbol{\theta}$ are now viewed as *random variables* (not unknown constants) with their own distribution, known as the *prior distribution*:

$$\pi(.)\colon\ \Theta\rightarrow[0,1],$$

which represents the modeler's assessment of how likely the various values of $\boldsymbol{\theta}$ in Θ are *a priori*, and

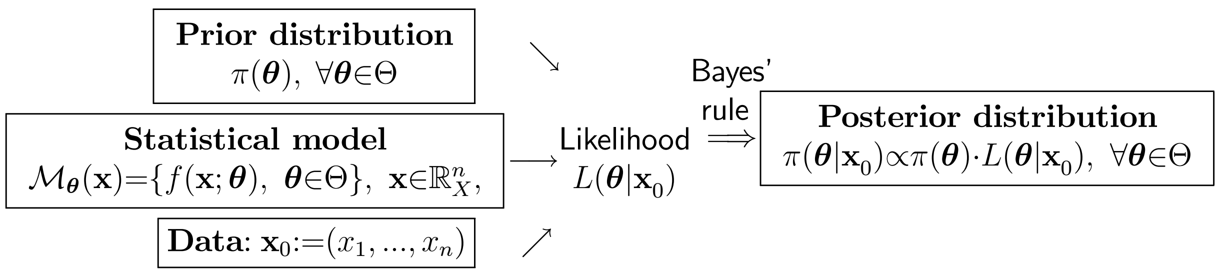
(ii) the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$ is re-interpreted by Bayesians to be defined as *conditional* on $\boldsymbol{\theta}$, and denoted by $f(\mathbf{x}|\boldsymbol{\theta})$.

Taken together these modifications imply that there exists a *joint distribution* relating the unknown parameters $\boldsymbol{\theta}$ and a sample realization $\mathbf{x}$:

$$f(\mathbf{x},\boldsymbol{\theta}){=}f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta.$$

Bayesian inference is based exclusively on the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ which is viewed as the revised (from the initial $\pi(\boldsymbol{\theta})$) *degrees of belief* for different values of $\boldsymbol{\theta}$ in light of the summary of the data by $L(\boldsymbol{\theta}|\mathbf{x}_0)$.

Table 10.8: The Bayesian approach to statistical inference


Prior distribution
$\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta$
Statistical model
$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta\},\ \mathbf{x}{\in}\mathbb{R}^n_X,$
Data: $\mathbf{x}_0{:=}(x_1,...,x_n)$
Likelihood
$L(\boldsymbol{\theta}|\mathbf{x}_0)$
Bayes' rule
Posterior distribution
$\pi(\boldsymbol{\theta}|\mathbf{x}_0){\propto}\pi(\boldsymbol{\theta}){\cdot}L(\boldsymbol{\theta}|\mathbf{x}_0),\ \forall\boldsymbol{\theta}{\in}\Theta$

Example 10.9. Consider the *simple Bernoulli model*, and let the **prior** $\pi(\theta)$ be $\mathsf{Beta}(\alpha,\beta)$ distributed with a density function:

$$\pi(\theta)=\tfrac{1}{\mathsf{B}(\alpha,\beta)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1},\ \alpha>0,\ \beta>0,\ 0<\theta<1. \tag{1}$$

Combining the likelihood with the prior yields the **posterior** distribution:

$$\begin{array}{rl}\pi(\theta|\mathbf{x}_0) & \propto \left(\tfrac{1}{\mathsf{B}(\alpha,\beta)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1}\right)\left[\theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})}\right]= \\ & =\tfrac{1}{\mathsf{B}(\alpha,\beta)}\left[\theta^{n\overline{x}+(\alpha-1)}(1-\theta)^{n(1-\overline{x})+\beta-1}\right].\end{array} \tag{2}$$

In view of the formula in (1), (2) as an 'non-normalized' density of a $\mathsf{Beta}(\alpha^{*},\beta^{*})$, where:

$$\alpha^{*}=n\overline{x}+\alpha,\quad \beta^{*}=n(1-\overline{x})+\beta. \tag{3}$$

As the reader might have suspected, the choice of the prior in this case was not arbitrary. The Beta prior in conjunction with a Binomial-type LF gives rise to a Beta posterior. This is known in Bayesian terminology as a *conjugate pair*, where $\pi(\theta)$ and $\pi(\theta|\mathbf{x}_0)$ belong to the same family of distributions.

In the terms of the main grounds stated above, the Bayesian approach:

[a] Adopts the degrees of belief interpretation of probability introduced via $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$.

[b] The relevant information includes both (i) the data $\mathbf{x}_0:=(x_1,x_2,...,x_n)$, and (ii) prior information. Such prior information comes in the form of a prior distribution $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$, which is assigned *a priori* and represents one's degree of belief in ranking the different values of $\boldsymbol{\theta}$ in Θ as more probable and less probable.

[c] The primary aim of the Bayesian approach is to **revise** the original *ranking* based on $\pi(\boldsymbol{\theta})$ in light of the data $\mathbf{x}_0$ by updating in the form of the *posterior distribution*:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0)=\tfrac{f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}}f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}\propto L(\boldsymbol{\theta}|\mathbf{x}_0)\cdot\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\boldsymbol{\Theta}, \tag{4}$$

where $L(\boldsymbol{\theta}|\mathbf{x}_0)\propto f(\mathbf{x}_0|\boldsymbol{\theta})$ denotes a *re-interpreted* likelihood function as being conditional on $\mathbf{x}_0$. The Bayesian approach is depicted in table 10.8. Since the denominator $m(\mathbf{x}_0)=\int_{\theta\in(0,1)}\pi(\boldsymbol{\theta})f(\mathbf{x}_0|\boldsymbol{\theta})d\boldsymbol{\theta}$, known as the **predictive** distribution, derived by integrating out $\boldsymbol{\theta}$, can be absorbed into the constant of proportionality in (4) and ignored for most practical purposes. The only exception to that is when one needs to treat $\pi(\theta|\mathbf{x}_0)$ as a proper density function which integrates to one, $m(\mathbf{x}_0)$ is needed as a normalizing constant.

Learning from data. In this context, learning from data $\mathbf{x}_0$ takes the form revising one's degree of belief for different values of $\boldsymbol{\theta}$ [i.e. different models $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta}\in\Theta$], in light of data $\mathbf{x}_0$, the learning taking the form $\pi(\boldsymbol{\theta}|\mathbf{x}_0)-\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$.. That is, the learning from data $\mathbf{x}_0$ about the phenomenon of interest takes place in the head of the modeler. In this sense, the underlying inductive reasoning is neither *factual* nor *hypothetical*, it's *all-inclusive* in nature: it pertains to *all* $\boldsymbol{\theta}$ in Θ, as ranked by $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$.

Savage (1954) summarizes Bayesian inference succinctly by:

'Inference means for us the change of opinion induced by evidence on the application of Bayes' theorem." (p. 178)

2 Bayesian estimation

A key argument used by Bayesians to taut their in favorite approach to statistics is its simplicity in the sense that all forms of inference revolve around a single function, the posterior distribution: $\pi(\boldsymbol{\theta}|\mathbf{x}_0)\propto\pi(\boldsymbol{\theta})\cdot f(\mathbf{x}_0|\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$. This, however, is only half the story. The other half is how the posterior distribution is utilized to yield 'optimal' inferences. The issue of optimality, however, is intrinsically related to what the primary objective of Bayesian inference is.

An outsider looking at Bayesian approach would surmise that its primary objective is to yield 'the probabilistic ranking' (ordering) of all values of $\boldsymbol{\theta}$ in Θ. The modeling begins with an a priori probabilistic ranking based on $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$, which is revised after observing $\mathbf{x}_0$ to derive $\pi(\boldsymbol{\theta}|\mathbf{x}_0),\ \forall\boldsymbol{\theta}\in\Theta$; hence the key role of the quantifier $\forall\boldsymbol{\theta}\in\Theta$. Indeed, O'Hagan's (1994) argues that the revised probabilistic ranking *is* the inference:

"The most usual inference question is this: After seeing the data x_0, what do we now know about the parameter θ? The only answer to this question is to present the entire posterior distribution." (p. 6).

He goes on to argue:

"Classical inference theory is very concerned with constructing good inference rules. The primary concern of Bayesian inference, ..., is entirely different. The objective is to extract information concerning θ from the posterior distribution, and to present it helpfully via effective summaries." (p. 14).

Where do these effective summaries come from? O'Hagan argues that the criteria for 'optimal' Bayesian inferences are only **parasitic** on the Bayesian approach and enter the picture via the decision theoretic perspective: "... a study of decision theory ... helps identify suitable summaries to give Bayesian answers to stylized inference questions which classical theory addresses." (p. 14).

Decision-theoretic framing of inference; initially proposed by Wald (1939; 1950). The decision-theoretic set-up has three basic components.

(i) A prespecified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

(ii) A decision space D containing all mappings $d(.)$: $\mathbb{R}^n_X \to A$, where A denotes the set of all actions available to the statistician.

(iii) A loss function $L(.,.)$: $[D \times \Theta] \to R$, representing the numerical loss if the statistician takes action $a\in A$ when the state of nature is $\theta\in\Theta$; see Ferguson (1967), Berger (1985), Wasserman (2004).

2.1 Optimal Bayesian rules

The decision theoretic setup provides optimal Bayesian rules the risk function $R(\theta,\widehat{\theta}){=}E_{\mathbf{X}}\left[L(\theta,\widehat{\theta}(\mathbf{X}))\right]$ to define the Bayes risk:

$$\textbf{Bayes risk}: \quad R_B(\widehat{\theta}){=}\int_{\boldsymbol{\theta}\in\Theta} R(\theta,\widehat{\theta})\pi(\theta)d\theta,$$

whose minimization with respect to all such rules $\widetilde{\theta}(\mathbf{x})$ yields:

$$\textbf{Bayes rule}: \quad \inf_{\widetilde{\theta}(\mathbf{x})} R_B(\widehat{\theta}){=}\inf_{\widetilde{\theta}(\mathbf{x})} \int_{\theta\in\Theta} R(\theta,\widehat{\theta})\pi(\theta)d\theta.$$

In light of the fact that $R_B(\widehat{\theta})$ can be expressed in the form (Bansal, 2007):

$$R_B(\widehat{\theta}) \;\; = \int_{\mathbf{x}\in\mathbb{R}^n_X} \int_{\theta\in\Theta} L(\widehat{\theta}(\mathbf{X}),\theta)\pi(\boldsymbol{\theta}|\mathbf{x})d\theta d\mathbf{x}. \tag{5}$$

where $\pi(\boldsymbol{\theta}|\mathbf{x}){\propto}f(\mathbf{x}|\theta)\pi(\theta)$. In light of (5), a Bayesian rule is 'optimal' relative to a particular loss function $L(\widehat{\theta}(\mathbf{X}),\theta)$, when it minimizes $R_B(\widehat{\theta})$. This makes it clear that what constitutes an 'optimal' Bayesian rule is primarily determined by $L(\widehat{\theta}(\mathbf{X}),\theta)$ (Schervish, 1995):

(i) when $L_2(\widehat{\theta},\theta){=}(\widehat{\theta}-\theta)^2$ the Bayes estimate $\widehat{\theta}$ is the *mean* of $\pi(\theta|\mathbf{x}_0)$,

(ii) when $L_1(\widetilde{\theta},\theta){=}|\widetilde{\theta}-\theta|$ the Bayes estimate $\widetilde{\theta}$ is the *median* of $\pi(\theta|\mathbf{x}_0)$,

(iii) when $L_{0-1}(\overline{\theta},\theta){=}\delta(\overline{\theta},\theta){=}\left\{\begin{array}{ll} 0 & \text{for } \left|\overline{\theta}{-}\theta\right| < \varepsilon \\ 1 & \text{for } \left|\overline{\theta}{-}\theta\right| \geq \varepsilon \end{array}\right.$ for $\varepsilon{>}0$, the Bayes estimate $\overline{\theta}$ is the *mode* of $\pi(\theta|\mathbf{x}_0)$.

In practice, the most widely used loss function is the square:

$$L_2(\widehat{\theta}(\mathbf{X});\theta){=}(\widehat{\theta}(\mathbf{X})-\theta)^2,\ \forall\theta{\in}\Theta,$$

whose risk function is the decision-theoretic *Mean Square Error (MSE)*:

$$R(\theta,\widehat{\theta}){=}E(\widehat{\theta}(\mathbf{X}){-}\theta)^2{=}MSE(\widehat{\theta}(\mathbf{X});\theta),\ \forall\theta{\in}\Theta. \tag{6}$$

It is important to note that (6) is the source of confusion between that and the frequentist definition of the *Mean Square Error*:

$$\mathsf{MSE}(\widehat{\theta}_n(\mathbf{X});\theta^*){=}E\{(\widehat{\theta}_n(\mathbf{X})-\theta^*)^2\}. \tag{7}$$

Example 11.36. As shown in example 10.10, for the *simple Bernoulli model* (table 11.1), with $\pi(\theta){\backsim}\mathsf{Beta}(\alpha,\beta)$, the posterior distribution is $\pi(\theta|\mathbf{x}_0){\backsim}(\alpha^*,\beta^*)$, where:

$$\alpha^*{=}n\overline{x}{+}\alpha,\quad \beta^*{=}n(1{-}\overline{x}){+}\beta.$$

Note that for $Z \backsim \text{Beta}(\alpha, \beta)$, mode$(Z){=}\frac{\alpha-1}{\alpha+\beta-2}$,

$$E(Z){=}\frac{\alpha}{\alpha+\beta},\ \ Var(Z){=}\frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}.$$

Example. The Jeffreys prior $\theta \backsim$ Beta$(.5,.5)$ (see fig. 11), in conjunction with $n\overline{x}{=}4, n{=}20$ gives rise to:

$$(\theta|\mathbf{x}_0) \backsim \text{Beta}(\alpha^*, \beta^*),\ \ \alpha^*{=}n\overline{x}{+}\alpha{=}4.5,\ \ \beta^*{=}n(1\text{-}\overline{x}) + \beta{=}16.5.$$

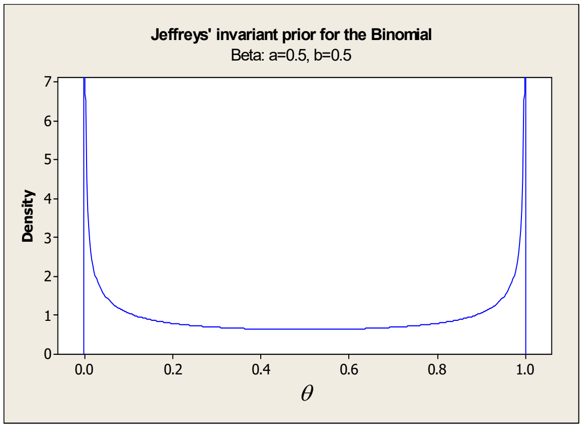


Fig. 11: Jeffreys prior $\pi(\theta){=}\frac{1}{B(.5,.5)}\theta^{-.5}(1{-}\theta)^{-.5}$

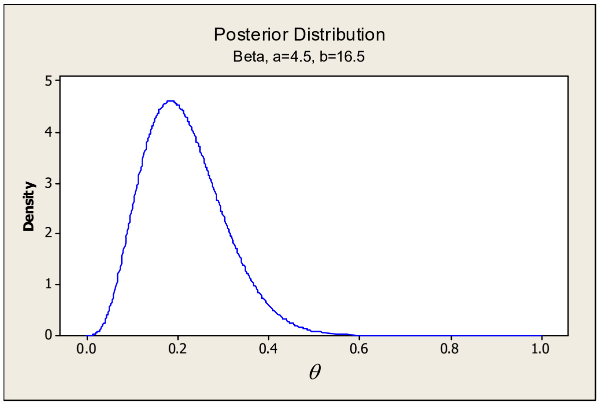


Fig. 12: $\pi(\theta|\mathbf{x}_0) \backsim$ Beta(4.5,16.5)

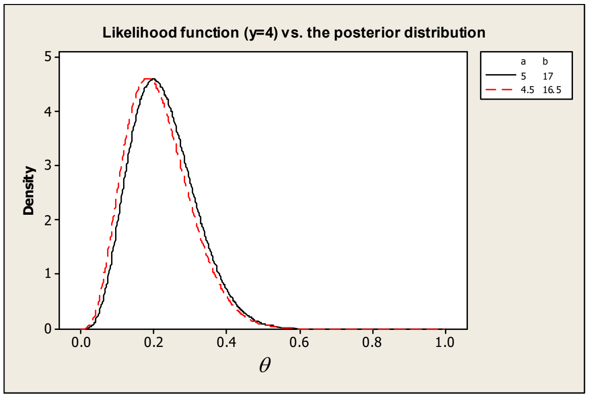


Fig. 13: $\pi(\theta|\mathbf{x}_0) \backsim$ Beta(4.5,16.5)
vs. the Likelihood Function $(y = 4)$

(a) When the relevant loss function is $L_{0-1}(\overline{\theta}, \theta)$, the optimal Bayesian rule is the *mode* of $\pi(\theta|\mathbf{x}_0){\backsim}(\alpha^*, \beta^*)$, which takes the form:

$$\widetilde{\theta}_B{=}\tfrac{\alpha^*-1}{\alpha^*+\beta^*-2}{=}\tfrac{(n\overline{X}+\alpha-1)}{(n+\alpha+\beta-2)}. \qquad (8)$$

(b) When the relevant loss function is $L_2(\widehat{\theta}, \theta){=}(\widehat{\theta}-\theta)^2$, the optimal Bayesian rule is the *mode* of $\pi(\theta|\mathbf{x}_0){\backsim}\text{Beta}(\alpha^*, \beta^*)$, which takes the form:

$$\widehat{\theta}_B{=}\tfrac{\alpha^*}{\alpha^*+\beta^*}{=}\tfrac{(n\overline{X}+\alpha)}{(n+\alpha+\beta)}. \qquad (9)$$

For a Jeffreys prior $\pi(\theta)\backsim$Beta$(.5,.5)$, $n\overline{x}$=4, n=20, α^{*}=$n\overline{x}$+α=4.5, β^{*}=n(1-$\overline{x}$)+β=16.5:

$$\widetilde{\theta}_{B}=\frac{3.5}{21-2}=.184, \quad \widehat{\theta}_{B}=\frac{4.5}{4.5+16.5}=.214. \tag{10}$$

2.2 Where do loss functions come from?

A closer scrutiny of the Bayesian set up reveals that the loss function needs to invoke '**information from sources other than the data**', which is usually not readily available. Indeed, such information is available in very restrictive situations, such as acceptance sampling in quality control. In light of that, a proper understanding of the intended scope of statistical inference calls for distinguishing the special cases where the loss function is part and parcel of the available substantive information from those that no such information is either relevant or available.

Tiao and Box (1975), p. 624, argued:

"Now it is undoubtedly true that on the one hand that situations exist where the loss function is at least approximately known (for example certain problems in business) and sampling inspection are of this sort. ... On the other hand, a vast number of inferential problems occur, particularly in the analysis of scientific data, where there is no way of knowing in advance to what use the results of research will subsequently be put."

Lehmann (1984) warned statisticians about the perils of **arbitrary loss functions**: "It is argued that the choice of a loss function, while less crucial than that of the model, exerts an important influence on the nature of the solution of a statistical decision problem, and that an arbitrary choice such as squared error may be baldly misleading as to the relative desirability of the competing procedures." (p. 425).

Tukey (1960) went even further arguing that the decision-theoretic framing distorts frequentist testing by replacing error probabilities with losses and costs: "Wald's decision theory ... has given up fixed probability of errors of the first kind, and has focused on gains, losses or regrets." (p. 433).

He went on to echo Fisher's (1955) view by contrasting decisions vs. inferences:

"Conclusions are established with careful regard to evidence, but without regard to consequences of specific actions in specific circumstances." (p. 425).

Hacking (1965) brought out the key difference even more clearly:

"... to conclude that an hypothesis is best supported is, apparently, to decide that the hypothesis in question is best supported. Hence it is a decision like any other. But this inference is fallacious. **Deciding that something is** the case differs from **deciding to do something**. ... Hence deciding to do something falls squarely in the province of decision theory, but deciding that something is the case does not." (p. 31).

Another important aspect of using loss functions in inference is that in practice the same statistical inference problem can give rise to very different decisions/actions depending on one's loss function. To illustrate that consider an example from Chatterjee (2002):

"... consider the case of a new drug whose effects are studied by a research scientist attached to the laboratory of a pharmaceutical company. The conclusion of the study

may have different bearings on the action to be taken by (a) the scientist whose line of further investigation would depend on it, (b) the company whose business decisions would determined by it, and (c) the Government whose policies as to health care, drug control, etc. would take shape on that basis." (p. 72)
In practice, each one of these different agents is likely to have a very different loss function, but their inferences should have a common denominator: the scientific evidence pertaining to θ^*, the true θ, that stems solely from the observed data.

2.3 Bayesian Credible Intervals

A Bayesian $(1-\alpha)$ credible interval for θ is constructed by finding the area between the $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ percentiles of the posterior distribution, say a and b, respectively:

$$\Pi(a \leq \theta < b){=}1-\alpha, \quad \int_a^1 \pi(\theta|\mathbf{x}_0)d\theta{=}(1-\tfrac{\alpha}{2}), \; \int_b^1 \pi(\theta|\mathbf{x}_0)d\theta{=}\tfrac{\alpha}{2}, \tag{11}$$

where $\Pi(.)$ denotes probabilistic assignments based on the *posterior* distribution $\pi(\theta|\mathbf{x}_0)$, $\forall\theta{\in}\Theta$. In practice one can define an infinity of $(1-\alpha)$ credible intervals using the same posterior $\pi(\theta|\mathbf{x}_0)$. To avoid this indeterminancy one needs to impose additional restrictions like the interval with the *shortest length* or one with *equal tails*; see Robert (2007).

Example 11.37. In the case of the simple Bernoulli model (table 11.1), the end points of an equal-tail credible interval can be evaluated by transforming the Beta distribution into the F distribution via:

$$Z \backsim \mathsf{Beta}(\alpha^*, \beta^*) \Rightarrow \quad \tfrac{\beta^* Z}{\alpha^*(1-Z)} \backsim \mathsf{F}(2\alpha^*, 2\beta^*). \tag{12}$$

Denoting the $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ percentiles of the $\mathsf{F}(2\alpha^*, 2\beta^*)$ distribution, by $\mathsf{f}(\frac{\alpha}{2})$ and $\mathsf{f}(1-\frac{\alpha}{2})$, respectively, the Bayesian $(1-\alpha)$ credible interval for θ is:

$$\left(1 + \tfrac{\beta^*}{\alpha^*\mathsf{f}(1-\frac{\alpha}{2})}\right)^{-1} \leq \theta \leq \left(1 + \tfrac{\beta^*}{\alpha^*\mathsf{f}(\frac{\alpha}{2})}\right)^{-1}. \tag{13}$$

Example 11.38. For the simple Bernoulli model (table 11.1), with Jeffreys prior $\pi_J(\theta) \backsim \mathsf{Beta}(.5, .5)$, $n\overline{x}{=}2$, $n{=}20$, $\alpha{=}.05$:

$$\alpha^*{=}n\overline{x} + \alpha{=}2.5, \;\; \beta^*{=}n(1-\overline{x}) + \beta{=}18.5, \;\; \mathsf{f}(1-\tfrac{\alpha}{2}){=}.163, \;\; \mathsf{f}(\tfrac{\alpha}{2}){=}2.93,$$

$$\left(1 + \tfrac{18.5}{2.5(.163)}\right)^{-1} \leq \theta \leq \left(1 + \tfrac{18.5}{2.5(2.93)}\right)^{-1} \Leftrightarrow \;\; (.0216 \leq \theta \leq .284). \tag{14}$$

It important to contrast the frequentist Confidence Interval (CI) with the Bayesian Credible Interval to bring out their key differences.

The **first important difference** is that the basis of a CI is a pivot $q(\mathbf{X}, \theta)$, whose sampling distribution is evaluated under $\theta{=}\theta^*$, with the probability firmly attached to $\mathbf{x}\in\mathbb{R}_X^n$. In contrast, a $(1-\alpha)$ Credible Interval represents **the highest posterior density interval**.

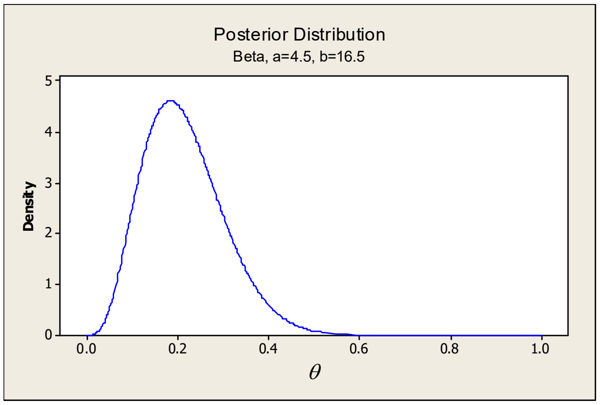


Fig. 12: $\pi(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(4.5{,}16.5)$

That is, it is simply the shortest interval whose area (integral) under the posterior density function $\pi(\theta|\mathbf{x}_0)$ has value $(1-\alpha)$, where the probabilities are firmly attached to $\theta\in\Theta$. Hence, any comparison of tail areas amounts to likening apples to oranges.

The **second key difference** is that the CI:

$$\left(\widehat{\theta}_L(\mathbf{X}) \leq \theta \leq \widehat{\theta}_U(\mathbf{X});\ \theta{=}\theta^*\right),$$

is random and its primary purpose is to cover θ^* with probability $(1-\alpha)$. There is **nothing random** about a $(1-\alpha)$ Credible Interval, and nothing connects that interval to θ^*. Bayesians would like to think that it does, since it covers a large part of the posterior, but there is nothing in the above derivation or its underlying reasoning that ensures that. Indeed, the very idea of treating θ **as a random variable** runs afoul any notion of true value θ^* that could have generated data $\mathbf{x}_0$. In the case of a CI, the evaluation of the sampling distribution of $q(\mathbf{X},\theta)$ under $\theta{=}\theta^*$ secures exactly that.

2.4 Bayesian testing using the Bayes factor

The Bayesian testing of the hypotheses:

$$H_0\colon\ \theta\in\Theta_0 \text{ vs. } H_1\colon\ \theta\in\Theta_1,$$

in the context of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}){=}\{f(\mathbf{x};\theta),\ \theta\in\Theta\},\ \mathbf{x}\in\mathbb{R}_X^n$, is naturally based on the ratio of the posterior distribution under the two hypotheses. Using the notation in chapter 10, the posterior is defined by:

$$\pi(\theta|\mathbf{x}_0) \;=\tfrac{f(\mathbf{x}_0|\theta)\cdot\pi(\theta)}{m(\mathbf{x}_0)}\propto L(\theta|\mathbf{x}_0)\cdot\pi(\theta),\ \forall\theta\in\boldsymbol{\Theta},$$

where $\pi(\theta)$ denotes the prior, $L(\theta|\mathbf{x}_0)$ the likelihood and $m(\mathbf{x}_0)=\int_{\boldsymbol{\theta}} f(\mathbf{x}_0|\theta)\cdot\pi(\theta)d\theta$. Assuming that $\pi_i=\pi(\theta\in\Theta_i)$, $i=0,1$, are the prior probabilities for H_0 and H_1, the 'renormalized' priors are:

$$\overline{\pi}_i(\theta)=(\pi(\theta)/\pi_i),\ \ \theta\in\Theta_i,\ \ \int_{\theta\in\Theta_i}\overline{\pi}_i(\theta)d\theta=1,\ \ i=0,1.$$

Hence, the posterior probabilities for H_0 and H_1 are defined by:

$$\pi(\theta\in\Theta_i|\mathbf{x}_0)=\int_{\theta\in\Theta_i}\pi(\theta)L(\theta|\mathbf{x}_0)d\theta=\pi_i\int_{\theta\in\Theta_i}\overline{\pi}_i(\theta)L(\theta|\mathbf{x}_0)d\theta,\ \ i=0,1.$$

The posterior odds ratio is defined by:

$$\frac{\pi(\theta\in\Theta_0|\mathbf{x}_0)}{\pi(\theta\in\Theta_1|\mathbf{x}_0)}=\frac{\pi_0\int_{\theta\in\Theta_0}\overline{\pi}_0(\theta)L(\theta|\mathbf{x}_0)d\theta}{\pi_1\int_{\theta\in\Theta_1}\overline{\pi}_1(\theta)L(\theta|\mathbf{x}_0)d\theta}.$$

To avoid the charge that prior probabilities render such testing susceptible to ad hoc manipulation, Bayesians often prefer to partially eliminate them from this ratio by defining the *Bayes Factor* (BF):

$$BF(\mathbf{x}_0)=\frac{\pi(\theta\in\Theta_0|\mathbf{x}_0)}{\pi(\theta\in\Theta_1|\mathbf{x}_0)}\left(\frac{\pi_1}{\pi_0}\right)\ =\frac{\int_{\theta\in\Theta_0}\overline{\pi}_0(\theta)L(\theta|\mathbf{x}_0)d\theta}{\int_{\theta\in\Theta_1}\overline{\pi}_1(\theta)L(\theta|\mathbf{x}_0)d\theta}. \qquad (15)$$

A close look at BF indicates that this is analogous to the frequentist likelihood ratio:

$$\lambda_n(\mathbf{X})=\frac{\max_{\boldsymbol{\theta}\in\Theta}L(\boldsymbol{\theta};\mathbf{X})}{\max_{\boldsymbol{\theta}\in\Theta_0}L(\boldsymbol{\theta};\mathbf{X})}=\frac{L(\widehat{\boldsymbol{\theta}};\mathbf{X})}{L(\widetilde{\boldsymbol{\theta}};\mathbf{X})}, \qquad (16)$$

but instead of maximizing the values of θ over Θ and Θ_0, the **parameters are integrated out** of the likelihood function over the two subsets using the respective (renormalized) priors as weights. By choosing the priors strategically this ratio can be simplified enough to involve only the likelihood functions.

Example 13.24. For a simple Bernoulli model, consider the hypotheses:

$$H_0\text{: }\theta=\theta_0 \text{ vs. } \ H_1\text{: }\theta\neq\theta_0,$$

and choose the priors $\pi_i=.5$, $i=0,1$. In this case $BF(\mathbf{x}_0)=\frac{L(\theta_0|\mathbf{x}_0)d\theta}{\int_0^1 L(\theta|\mathbf{x}_0)d\theta}$.

Bayesian testing is based on $\log_{10} BF(\mathbf{x}_0)$ in conjunction with certain thresholds concerning the *strength* of the degree of belief *against* H_0 (Robert, 2007, p. 228):

(i) $0\leq-\log_{10} BF(\mathbf{x}_0)\leq .5$, the degree of belief against H_0 is *poor*,

(ii) $.5<-\log_{10} BF(\mathbf{x}_0)\leq 1$, the degree of belief against H_0 is *substantial*,

(iii) $1<-\log_{10} BF(\mathbf{x}_0)\leq 2$, the degree of belief against H_0 is *strong*, and

(iv) $-\log_{10} BF(\mathbf{x}_0)>2$, the degree of belief against H_0 is *decisive*.

These 'rules of thumb', however, have been questioned by Kass and Raftery (1995) as largely ad hoc. There is no principled argument based on sampling distributions as in the case of the frequentist likelihood ratio test in (16).

2.4.1 Interval hypotheses

Consider the following hypotheses:

$$H_0\colon\ \theta \leq \theta_0 \quad \text{vs.} \quad H_1\colon\ \theta > \theta_0,\ \ \theta_0{=}.5, \tag{17}$$

in the context of the simple Bernoulli case with a Jeffreys invariant prior, with data $n\overline{x}{=}12$, $n{=}20$.

An obvious way to evaluate the posterior odds for these two interval hypotheses is as follows:

$$\pi(\theta{\leq}\theta_0|\mathbf{x}_0){=}\tfrac{\Gamma(21)}{\Gamma(12.5)\Gamma(8.5)}\int_0^{.5}\left(\theta^{11.5}(1\text{-}\theta)^{7.5}\right)d\theta{=}.186,\quad \pi(\theta{>}\theta_0|\mathbf{x}_0){=}1\text{-}\pi(\theta \leq \theta_0|\mathbf{x}_0){=}.814$$

One can then employ the posterior odds criterion:

$$\tfrac{\pi(\theta\leq\theta_0|\mathbf{x}_0)}{\pi(\theta>\theta_0|\mathbf{x}_0)}{=}\tfrac{.186}{.814}{=}.229,$$

which in terms of $-\log_{10} BF(\mathbf{x}_0){=}.64$ indicates that the degree of belief against H_0 is *substantial*.

2.4.2 Point null hypotheses

It is interesting to note that in their attempt to deflect attention from their technical difficulty in assigning posterior probabilities to point hypotheses, Bayesians criticized their use in frequentist testing as non-sensical because they can never be exactly true! Of course, this is a non-sensical argument because in statistical inference the notion of exactly true, in a deterministic sense, has no place.

Pretending that point hypotheses are small intervals. A 'pragmatic' way to handle point hypotheses in Bayesian inference is to sidestep the technical difficulty in handling hypotheses of the form:

$$H_0\colon\ \theta{=}\theta_0 \quad \text{vs.} \quad H_1\colon\ \theta \neq \theta_0, \tag{18}$$

by *pretending* that the point hypothesis:

$$H_0\colon\ \theta{=}\theta_0,$$

is actually a small interval (calling it a more realistic formulation) of the form:

$$H_0\colon\ \theta \in \Theta_0{:=}(\theta_0 - \epsilon,\ \theta_0 + \epsilon),$$

and attaching a *spiked prior* of the form:

$$\pi(\theta{=}\theta_0){=}p_0, \quad p_1{=}\int_0^1 \pi(\theta{\neq}\theta_0)d\theta{=}1-p_0.$$

That is, attach a prior of value p_0 to the point null, and then distribute the rest $1-p_0$ to all the other values of θ; see Berger (1985).

Using Credible Intervals as surrogates for tests. **Lindley** (1965) suggested an adaptation of a frequentist procedure of using the duality between the acceptance region and Confidence Intervals as surrogates for tests, by replacing the latter with Credible Intervals. His Bayesian adaptation to handle point null hypotheses, say:

$$H_0\text{: } \theta{=}.8 \quad \text{vs.} \quad H_1\text{: } \theta \neq .8, \tag{19}$$

is to construct a $(1-\alpha)$ Credible Interval using an "uninformative" prior and reject H_0 if it lies outside that interval.

Example. For $n\overline{x}{=}12$, $n{=}20$, the likelihood function is:

$$L(\theta;\mathbf{x}_0) \propto \theta^{12}(1-\theta)^8, \;\; \theta \in (0,1),$$

and the posterior density resulting from a *uniform* prior: $\pi(\theta) \backsim \mathsf{Beta}(1,1)$ is:

$$\pi(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(13,9), \;\; \theta{\in}(0,1).$$

A .95 credible interval for θ is:

$$\Pi(.384 \leq \theta < .7817){=}.95, \tag{20}$$

$$\tfrac{\Gamma(22)}{\Gamma(13)\Gamma(9)} \int_{.384}^{1} \theta^{12}(1{-}\theta)^8 d\theta{=}0.975, \quad \tfrac{\Gamma(22)}{\Gamma(13)\Gamma(9)} \int_{.7817}^{1} \theta^{12}(1{-}\theta)^8 d\theta{=}.025.$$

This suggests that the null $\theta_0{=}.8$ should be rejected.

2.5 Bayesian Prediction

The best Bayesian predictor for X_{n+1} is based on its posterior predictive density, which is defined by:

$$f(x_{n+1}|\mathbf{x}_0){=} \int_0^1 f(x_{n+1}|\mathbf{x_0};\theta) f(\mathbf{x}_0|\theta)\pi(\theta)d\theta.$$

Note that:

$$f(x_{n+1},\mathbf{x}_0,\boldsymbol{\theta}){=}\left[f(x_{n+1}|\mathbf{x_0};\theta){\cdot}f(\mathbf{x}_0|\theta){\cdot}\pi(\theta)\right]$$

defines the joint distribution of $(x_{n+1},\mathbf{x}_0,\boldsymbol{\theta})$.

Given that $X_{n+1} \backsim \mathsf{Ber}(\theta(1-\theta))$, integrating out θ yields:

$$f(x_{n+1}|\mathbf{x}_0){=}\begin{cases} \frac{(n\overline{x}+\alpha)}{(n+\alpha+\beta)}, & \text{if } x_{n+1}{=}1, \\ \frac{(n(1-\overline{x})+\beta)}{(n+\alpha+\beta)}, & \text{if } x_{n+1}{=}0, \end{cases}$$

which is a Bernoulli density with $\theta^*{=}\frac{(n\overline{x}+\alpha)}{(n+\alpha+\beta)}$. The Bayesian predictor, based on the mode of $f(x_{n+1}|\mathbf{x})$ is:

$$\widetilde{X}_{n+1}{=}\begin{cases} 1, & \text{if } \max\left(\theta^*,[1-\theta^*]\right){=}\theta^*, \\ 0, & \text{if } \max\left(\theta^*,[1-\theta^*]\right){=}[1-\theta^*]. \end{cases} \tag{21}$$

The posterior expectation predictor is given by:

$$E(X_{n+1}|\mathbf{x}_0) \quad = \tfrac{(n\overline{x}+\alpha)}{(n+\alpha+\beta)},$$

which in the case where $\alpha{=}\beta{=}1,\ \pi(\theta) \backsim \mathsf{Beta}(1,1){=}U(0,1)$, the posterior expectation predictor gives rise to **Laplace's law of succession:**

$$E(X_{n+1}|\mathbf{x}_0) \quad = \tfrac{(n\overline{x}+1)}{(n+2)}. \tag{22}$$