Chapter 1
Introduction to Bayesian Statistical Inference

Georgios P. Karagiannis

Abstract We present basic concepts of Bayesian statistical inference. We briefly introduce the Bayesian paradigm. We present the conjugate priors; a computational convenient way to quantify prior information for tractable Bayesian statistical analysis. We present tools for parametric and predictive inference, and particularly the design of point estimators, credible sets, and hypothesis tests. These concepts are presented in running examples. Supplementary material is available from GitHub.

1.1 Introduction

Statistics mainly aim at addressing two major things. First, we wish to learn or draw conclusions about an unknown quantity, $\theta \in \Theta$ called 'the parameter', which cannot be directly measured or observed, by measuring or observing a sequence of other quantities called 'observations (or data, or samples)' $x_{1:n} := (x_1, \ldots, x_n) \in \mathcal{X}^m$ whose generating mechanism is (or can be considered as) stochastically dependent on the quantity of interest θ though a probabilistic model $x_{1:n} \sim f(\cdot|\theta)$. This is an inverse problem since we wish to study the cause θ by knowing its effect $x_{1:n}$. We will refer to this as parametric inference. Second, we wish to learn the possible values of a future sequence of observations $y_{1:m} \in \mathcal{X}^m$ given $x_{1:n}$. This is a forward problem, and we will call it predictive inference. Here, we present how both inferences can be addressed in the Bayesian paradigm.[1]

Consider a sequence of observables $x_{1:n} := (x_1, \ldots, x_n)$ generated from a sampling distribution $f(\cdot|\theta)$ labeled by the unknown parameter $\theta \in \Theta$. The statistical model $\mathfrak{m}$ consists of the observations $x_{1:n}$, and their sampling distribution $f(\cdot|\theta)$; $\mathfrak{m} = (f(\cdot|\theta);\ \theta \in \Theta)$.

[1] https://github.com/georgios-stats/UTOPIAE-Bayes.

G. P. Karagiannis (✉)
Department of Mathematical Sciences, Durham University, Durham, United Kingdom
e-mail: georgios.karagiannis@durham.ac.uk

L. Aslett et al. (eds.), *Uncertainty in Engineering*,
SpringerBriefs in Statistics,
https://doi.org/10.1007/978-3-030-83640-5_1

Unlike in Frequentist statistics, in Bayesian statistics unknown/uncertain parameters are treated as random quantities and hence follow probability distributions. This is justified by adopting the subjective interpretation of probability [4], as the degree of the researcher's believe about the uncertain parameter θ. Central to the Bayesian paradigm is the specification of the so-called prior distributions $\mathrm{d}\pi(\theta)$ on the uncertain parameters θ representing the degree of believe (or state of uncertainty) of the researcher about the parameter. Different researchers may specify different prior probabilities, as this is in accordance to the subjective nature of the probability. The specification of the prior is discussed in Sect. 1.2.

The Bayesian model consists of the statistical model $f(x_{1:n}|\theta)$ containing the information about θ available from the observed data $x_{1:n}$, and the prior distribution $\pi(\theta)$ reflecting the researcher's believe about θ before the data collection. It is denoted as

$$(f(x_{1:n}|\theta), \pi(\theta)) \text{ or as } \begin{cases} x_{1:n}|\theta & \sim f(\cdot|\theta) \\ \theta & \sim \pi(\cdot) \end{cases}.$$

Bayesian parametric inference relies on the posterior distribution $\pi(\theta|x_{1:n})$ whose density or mass function (PDF or PMF) is calculated by using the Bayes theorem

$$\pi(\theta|x_{1:n}) = \frac{f(x_{1:n}|\theta)\pi(\theta)}{\int_{\Theta} f(x_{1:n}|\theta)\pi(\mathrm{d}\theta)} \tag{1.1}$$

as a tool to invert the conditioning from $x_{1:n}|\theta$ to $\theta|x_{1:n}$. Posterior distribution (1.1) quantifies the researcher's degree of believe after taking into account the observations. By using subjective probability arguments, we can see interpret (1.1) as a mechanism that updates the researcher's degree of believe from the prior $\pi(\theta)$ to the posterior $\pi(\theta|x_{1:n})$ in the light of the observations collected.

Bayesian predictive inference about a future observation y_* can be addressed based on the predictive distribution defined as

$$p(y|x_{1:n}) = \int_{\Theta} f(y|\theta)\pi(\mathrm{d}\theta|x_{1:n}) = \mathrm{E}_{\pi}(f(y|\theta)|x_{1:n}). \tag{1.2}$$

Essentially, it is the expected value of the sampling distribution averaging out the uncertain parameter θ with respect to its posterior distribution reflecting the researcher's degree of believe.

Although the posterior and predictive distributions quantify the researcher's knowledge, they are not enough to give a solid answer about the quantity to be learned. In what follows we discuss important concepts based on decision theory which are used for Bayesian inference.

1.2 Specification of the Prior

Prior distribution $\pi(\theta)$ needs to reflect the researcher's degree of believe about the uncertain parameter $\theta \in \Theta$. Sophisticated prior distributions often lead to ineluctable posterior or predictive probabilities, and hence Bayesian analysis. Following, we present a computationally convenient class of priors applicable to several scenarios.

1.2.1 Conjugate Priors

Conjugate priors is a mathematically convenient way to specify the prior model in certain cases. They facilitate the tractable implementation of the Bayesian statistical analysis, by leading to computationally tractable posterior distributions.

Formally, if $\mathcal{F} = \{f(\cdot|\theta); \forall\theta \in \Theta\}$ is a class of parametric models (sampling distributions), and $\mathcal{P} = \{\pi(\theta|\tau); \forall\tau\}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$\pi(\theta|x_{1:n}) \in \mathcal{P}, \quad \forall f(\cdot|\theta) \in \mathcal{F} \text{ and } \pi(\cdot) \in \mathcal{P}.$$

It is straightforward to specify a conjugate prior when the sampling distribution is member of the exponential family. Consider observation x_i generated from a sampling distribution in the exponential family

$$x_i|\theta \overset{\text{IID}}{\sim} \text{Ef}_k(u, g, h, \phi, \theta, c); \quad i = 1, \ldots, n$$

with density $\text{Ef}_k(x|u, g, h, \phi, \theta, c) = u(x)g(\theta)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{i=1}^{n} h_j(x)))$ and $g(\theta) = 1/\int u(x)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{i=1}^{n} h_j(x)))\text{d}x$. The likelihood function is equal to

$$f(x_{1:n}|\theta) = \prod_{i=1}^{n} u(x_i)g(\theta)^n \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{i=1}^{n} h_j(x_i))). \tag{1.3}$$

The conjugate prior, corresponding to likelihood (1.3), admits density of the form

$$\pi(\theta|\tau) = \frac{1}{K(\tau)} g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j) \tag{1.4}$$

where $\tau = (\tau_0, \ldots, \tau_k)$ is such that $K(\tau) = \int_\Theta g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j)\text{d}\theta < \infty$. The resulting posterior of θ has the form

$$\pi(\theta|x_{1:n},\tau) = \frac{1}{K(\tau^*)} g(\theta)^{\tau_0^*} \exp(\sum_{j=1}^{k} c_j \phi_j(\theta)\tau_j^*))$$

with $\tau^* = (\tau_0^*, \tau_1^*, \ldots, \tau_k^*)$, $\tau_0^* = \tau_0 + n$, and $\tau_j^* = \sum_{i=1}^n h_j(x_i) + \tau_j$ for $j = 1, \ldots, k$.

It is easy to see that (1.4) is conjugate to (1.3) as the posterior can be re-written as $\pi(\theta|x_{1:n},\tau) = \pi(\theta|\tau^*)$ where $\tau^* = \tau + t_n(x_{1:n})$, and $t_n(x_{1:n}) = (n, \sum_{i=1}^n h_1(x_i), \ldots, \sum_{i=1}^n h_k(x_i))$. You can check the demo in.[2]

Example: Bernoulli model (Cont.)

Consider observations $x_{1:n} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ generated from a Bernoulli distribution with success rate $\theta \in [0, 1]$; i.e., $x_i|\theta \sim \text{Br}(\theta)$, $i = 1, \ldots, n$. Interest lies in specifying a conjugate prior for θ.

The sampling distribution is member of the exponential family, with $u(x) = 1$, $g(\theta) = (1-\theta)$, $c_1 = 1$, $\phi_1(\theta) = \log(\frac{\theta}{1-\theta})$, $h_1(x) = x$, because

$$f(x|\theta) = \text{Br}(x|\theta) = \theta^x(1-\theta)^{1-x} = (1-\theta)\exp(\log(\frac{\theta}{1-\theta})x).$$

The corresponding conjugate prior has PDF such as

$$\pi(\theta|\tau) \propto (1-\theta)^{\tau_0} \exp(\log(\frac{\theta}{1-\theta})\tau_1) = \theta^{(\tau_1+1)-1}(1-\theta)^{(\tau_0-\tau_1+1)-1},$$

where we recognize Beta distribution $\pi(\theta|\tau) = \text{Be}(\theta|a, b)$, with $a = \tau_1 + 1$, $b = \tau_0 - \tau_1 + 1$. Therefore, the posterior distribution is

$$\pi(\theta|x_{1:n},\tau) = \pi(\theta|\tau_0 + n, \tau + \sum_{i=1}^n h(x_i)) \propto \theta^{(\tau_1+n\bar{x}+1)-1}(1-\theta)^{(\tau_0+n-\tau_1-n\bar{x}+1)-1}$$

which is $\text{Be}(\theta|a^*, b^*)$, with $a^* = a + n\bar{x}$, and $b^* = b + n - n\bar{x}$.

1.3 Point Estimation

Often interest lies in learning the ‘true’ value of the unknown parameter $\theta \in \Theta$, or the future values of a future sequence of observations $y_{1:m} \in \mathcal{X}^m$; this is performed via the Bayesian point estimator. Here, we demonstrate the theory of the Bayesian point estimator in parametric inference, and leave the extension to the predictive inference to the reader.

[2] Web-applet: https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/.

Bayes (parametric) point estimator of $\theta \in \Theta$ with respect to the loss function $\ell(\theta, \delta)$ and the posterior distribution $\pi(\theta|x_{1:n})$ is an Bayes rule δ^{π} which minimizes $\int_{\Theta} \ell(\theta, \delta)\pi(\mathrm{d}\theta|x_{1:n})$; i.e.,

$$\delta^{\pi}(x_{1:n}) = \arg\min_{\forall \delta \in \Theta} \mathrm{E}_{\pi}(\ell(\theta, \delta)|x_{1:n}) = \arg\min_{\forall \delta \in \Theta} \int_{\Theta} \ell(\theta, \delta)\pi(\mathrm{d}\theta|x_{1:n}). \tag{1.5}$$

Often the accuracy of the Bayes point estimator is represented by its standard error. A commonly accepted metric for the standard error of the j-th dimension of the estimator δ^{π} is

$$\mathrm{se}_{\pi}(\delta_j|x_{1:n}) = \sqrt{\mathrm{MSE}_{\pi}(\delta_j|x_{1:n})}$$

where $\mathrm{MSE}_{\pi}(\delta_j|x_{1:n}) = [\mathrm{E}_{\pi}\left((\theta - \delta)(\theta - \delta)^{\top}|x_{1:n}\right)]_{j,j}$ is the mean squared error of δ_j.

A number of standard Bayesian point estimates, under different loss functions, are location summary statistics of the posterior distribution (mean, median, mode, quantiles, etc.) You can check the demo in.[3]

The Bayesian estimate of θ with respect to the linear loss $\ell(\theta, \delta) = c_1(\delta - \theta)1_{\theta \leq \delta}(\delta) + c_2(\theta - \delta)1_{\{\theta \leq \delta\}^c}(\delta)$ is the $\frac{c_2}{c_1+c_2}$-th posterior quantile; i.e., $\pi(\theta \in (-\infty, \delta(x_{1:n}))|x_{1:n}) = \frac{c_2}{c_1+c_2}$. The linear loss function essentially allows the adjustment of the penalty between over-estimating and under-estimating θ, by adjusting $c_1$and c_2. In particular, for $c_1 = c_2$, we get the absolute loss $\ell(\theta, \delta) = |\theta - \delta|$ and the posterior estimator is the posterior median

$$\delta(x_{1:n}) = \mathrm{median}_{\pi}(\theta|x_{1:n}). \tag{1.6}$$

The absolute loss is more appropriate when over-estimation and under-estimation are of the same concern (as penalized the same).

The Bayes estimate $\delta^{\pi}(x_{1:n})$ of θ with respect to the quadratic loss function $\ell(\theta, \delta) = (\theta - \delta)^2$ is

$$\delta^{\pi}(x_{1:n}) = \mathrm{E}_{\pi}(\theta|x_{1:n}). \tag{1.7}$$

The posterior mean of θ as an estimator of θ essentially minimizes the estimator error $\mathrm{se}_{\pi}(\delta|x_{1:n})$, which is equal to the posterior standard error. Obviously, the standard error of the estimator (1.7) is equal to the posterior standard error. Compared to the absolute loss, the quadratic loss aims at over-penalizing large but unlikely errors. In fact, quadratic loss aims at minimizing the standard error $\mathrm{se}_{\pi}(\delta|x_{1:n})$.

Finally, the Bayesian estimate of θ with respect to the zero-one loss $\ell(\theta, \delta) = 1 - 1_{B_{\epsilon}(\delta)}(\theta)$ is the posterior mode

$$\delta(x_{1:n}) = \mathrm{mode}_{\pi}(\theta|x_{1:n}) \tag{1.8}$$

as $\epsilon \to 0$.

[3] Web-applet: https://georgios-stats-1.shinyapps.io/demo_PointEstimation/.

Example: Bernoulli model (Cont.)

Interest lies in calculating the Bayesian point estimator under the absolute loss function. This is the Maximum A posteriori Estimator (the posterior mode). It is

$$\log(\pi(\theta|x_{1:n})) \propto (n\bar{x}+a-1)\log(\theta)+(n-n\bar{x}+b-1)\log(1-\theta).$$

For $a>0$, $b>0$, $\frac{\mathrm{d}}{\mathrm{d}\theta}\log(\pi(\theta|x_{1:n}))|_{p=\delta(x)}=0$ implies $\delta(x)=\frac{n\bar{x}+a-1}{n+a+b-2}$. Note that (a.) if $a\to 1$, $b\to 1$ (aka $\pi(\theta|a,b)\propto 1$), then $\delta^{\pi}(x)=\bar{x}$ similar to frequentists stats; (b.) if $a\to 0$, $b\to 0$ (aka $\pi(\theta|a,b)\propto\theta^{-1}(1-\theta)^{-1}$), then $\delta(x)=\frac{n\bar{x}-1}{n-2}$; if $a\to 1/2, b\to 1/2$ (aka $\pi(\theta|a,b)\propto\theta^{-1/2}(1-\theta)^{-1/2}$), then $\delta(x)=\frac{n\bar{x}-1/2}{n-1}$; if $n\to\infty$, $a>0$, $b>0$, then $\delta(x)=\bar{x}$.

1.4 Credible Sets

Instead of just reporting parametric (or predictive) point estimates for θ (or $y_{1:m}$), it is often desirable and more useful to report a subset of values $C_a\subseteq\Theta$ (or $C_a\subseteq\mathcal{X}^m$) where the posterior (or predictive) probability that $\theta\in C_a$ (or $y_{1:m}\in C_a$) is equal to a certain value a reflecting one's degree of believe.

The definition below describes the credible set [1, 5].

Definition 1.1 (*Posterior Credible Set*) A set $C_a\subseteq\Theta$ such that

$$\pi(\theta\in C_a|x_{1:n})=\int_{C_a}\pi(\mathrm{d}\theta|x_{1:n})\geq 1-a$$

is called '$100(1-a)\%$' posterior credible set for θ, with respect to the posterior distribution $\pi(\mathrm{d}\theta|x_{1:n})$.

In contrast to the frequentist stats, in Bayesian stats we can speak meaningfully of the probability that θ is in C_a, because probability $1-a$ reflects one's degree of believe that $\theta\in C_a$.

Among all the credible sets C_a in Definition 1.1, we are often interested in those that have the minimum volume. It can be proved [2] that the highest probability density (HPD) sets have this property. HPD consider those values of θ corresponding to the highest posterior pdf/pmf (aka the most likely values of θ).

Definition 1.2 (*Posterior highest probability density (HPD) set*) The $100(1-a)\%$ highest probability density set for $\theta\in\Theta$ with respect to the posterior distribution $\pi(\theta|x_{1:n})$ is the subset C_a of Θ of the form

$$C_a=\{\theta\in\Theta:\pi(\theta|x_{1:n})\geq k_a\} \tag{1.9}$$

where k_a is the largest constant such that

$$\pi(\theta \in C_a | x_{1:n}) \geq 1 - a. \tag{1.10}$$

From the decision theory perspective, HPD set C_a is the Bayes estimate of C_a the credible interval under the loss function $\ell(C_a, \theta) = k|C_a| - 1_{C_a}(\theta)$, for $k > 0$ which penalizes sets with larger volumes. The proof is available in [2].

Example: Multivariate Normal model

Consider observations $x_1, \ldots, x_n$ independently drawn from a q-dimensional normal $\mathrm{N}_q(\mu, \Sigma)$ with unknown $\mu \in \mathbb{R}^q$, $q \geq 1$, and known Σ, μ_0, Σ_0. Assume prior $\mu \sim \mathrm{N}_q(\mu_0, \Sigma_0)$. Interest lies in calculating the C_a parametric HPD credible interval for μ.

The posterior PDF of μ is

$$\begin{aligned}\pi(\mu|x_{1:n}) \propto f(x_{1:n}|\mu)\pi(\mu) &= \prod_{i=1}^{n} \mathrm{N}_q(x_i|\mu, \Sigma)\mathrm{N}_q(\mu|\mu_0, \Sigma_0) \\ &\propto \exp(-\frac{1}{2}(\mu - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n)) \propto \mathrm{N}_q(\mu|\hat{\mu}_n, \hat{\Sigma}_n)\end{aligned}$$

where $\Sigma_n = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}$, and $\hat{\mu}_n = \hat{\Sigma}_n(n\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0)$. So $\mu|x_{1:n} \sim \mathrm{N}_q(\hat{\mu}_n, \hat{\Sigma}_n)$.

From Definition 1.2, the credible set has the form

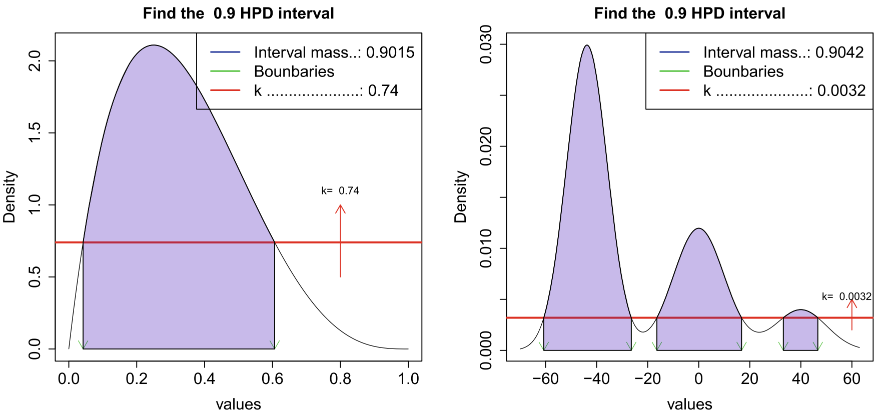
$$\begin{aligned}C_a &= \{\mu \in \mathbb{R}^q : \pi(\mu|x_{1:n}) \geq k_a\} \\ &= \{\mu \in \mathbb{R}^q : (\mu - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n) \leq -\log(2\pi \det(\hat{\Sigma}_n)))k_a = \tilde{k}_a\}\end{aligned}$$

where k_a is the greatest value satisfying

$$\begin{aligned}&\pi_{\mathrm{N}_q(\hat{\mu}_n, \hat{\Sigma}_n)}(\mu \in C_a|x_{1:n}) \geq 1 - a \Longleftrightarrow \\ &\pi_{\chi^2_q}((\mu - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n) \leq \tilde{k}_a) \geq 1 - a.\end{aligned} \tag{1.11}$$

Here, $(\mu - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n) \sim \chi^2_q$ as a sum of squares of independent standard normal random variables, and hence $\tilde{k}_a$ is the $1 - a$-th quantile of the χ^2_q distribution; i.e., $\tilde{k}_a = \chi^2_{q,1-a}$. Therefore, C_a parametric HPD credible set for μ is

$$C_a = \{\mu \in \mathbb{R}^q : (\mu - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n) \leq \chi^2_{q,1-a}\}$$



(a) Uni-modal case

(b) Multi-modal case

Fig. 1.1 Schematic of 1D HPD set

In real applications, the calculation of the credible interval might be intractable, due to the inversion in (1.9) or integration in (1.10). Below, we present a Naive algorithm [1] that can be implemented in a computer.[4]

- Create a routine which computes all solutions θ^* to the equation $\pi(\theta|x_{1:n}) = k_a$, for a given k_a. Typically, $C_a = \{\theta \in \Theta : \pi(\theta|x_{1:n}) \geq k_a\}$ can be constructed from those solutions.
- Create a routine which computes $\pi(\theta \in C_a|x_{1:n}) = \int_{\theta \in C_a} \pi(\theta|x_{1:n})\mathrm{d}\theta$
- Numerically solve the equation $\pi(\theta \in C_a|x_{1:n}) = 1 - a$ as k_a varies.

Figure 1.1 demonstrates the above procedure in 1D unimodal and tri-modal cases. Specifically, the red horizontal bar denotes k_a moves upwards, and intersects the density at locations which are the potential boundaries of C_a. The bar stops to move when the total density above regions of the parametric space is equal to $1 - \alpha$. The HPD credible set results as the union of these sub-regions. You can check the demo in.[5]

Theorem 1.1 suggests a computationally convenient way to calculate HPD credible intervals in 1D, and unimodal cases. The proof is available in [3].

Theorem 1.1 *Let θ follows a distribution with unimodal density $\pi(\theta|x_{1:n})$. If the interval $C_a = [L, U]$ satisfies*

1. $\int_L^U \pi(\theta|x_{1:n})d\theta = 1 - a$,
2. $\pi(U) = \pi(L) > 0$, *and*
3. $\theta_{mode} \in (L, U)$, *where* θ_{mode} is the mode of $\pi(\theta|x_{1:n})$,

then interval $C_a = [L, U]$ is the HPD interval of θ with respect to $\pi(\theta|x_{1:n})$.

[4] Web-applet: https://georgios-stats-1.shinyapps.io/demo_crediblesets/.

[5] Web-applet: https://georgios-stats-1.shinyapps.io/demo_crediblesets/.

Example: Bernoulli model (Cont.)

Interest lies in calculating the 2-sides 95% HPD interval for θ, given a sample with $n = 30$, and $\sum_{i=1}^{30} x_i = 15$, and prior hyper-parameters $a = b = 2$.

The posterior distribution of θ is $\text{Be}(a + n\bar{x} = 17, b + n - n\bar{x} = 17)$, which is 1D and unimodal; hence we use Theorem 1.1. It is

$$1 - a = \int_L^U \text{Be}(\theta|17, 17)\text{d}\theta = \text{Be}(\theta < U|17, 17) - \text{Be}(p < L|17, 17).$$

Note that Beta PDF is symmetric around 0.5 when $a^* = b^*$, and so is here where $\text{Be}(17, 17)$. Then,

$$1 - a = \text{Be}(\theta < U|17, 17) - (1 - \text{Be}(\theta < U|17, 17)) = 2\text{Be}(\theta < U|17, 17) - 1$$

so $\text{Be}(\theta < U|17, 17) = 1 - a/2$ and $L = 1 - U$. For $a = 0.95$, the 95% posterior credible interval for θ is $[L, U] = [0.36, 0.64]$.

Remark 1.1 Predictive credible sets for a future sequence of observations $y_{1:m}$, are defined and constructed as parametric ones by replacing θ with $y_{1:m}$ and $\pi(x_{1:n}|\theta)$ with $p(y_{1:m}|x_{1:n})$ in Definitions 1.1 and 1.2, and their consequences in this section. It is left as an Exercise.

1.5 Hypothesis Test

Often there is interest in reducing the overall parametric space Θ (aka the set of possible values of that the uncertain parameter θ can take) to a smaller subset. For instance; whether the proportion of Brexiters is larger than 0.5 ($p > 0.5$) or not ($p \leq 0.5$).

Such a decision can be formulated as a hypothesis test [1], namely the decision procedure of choosing between two non-overlapping hypotheses

$$\text{H}_0 : \theta \in \Theta_0 \quad \text{vs} \quad \text{H}_1 : \theta \in \Theta_1 \tag{1.12}$$

where $\{\Theta_0, \Theta_1\}$ partitions the space Θ. Typically, hypotheses, $\{\text{H}_k\}$, are categorized in three categories. Single hypothesis for θ is called the hypothesis where $\Theta_j = \{\theta_j\}$ contains a single element. Composite hypothesis for θ is called the hypothesis where $\Theta_j \subseteq \Theta$ contains many elements. General alternative hypothesis for θ is called the composite hypothesis where $\Theta_1 = \Theta - \{\theta_0\}$ when it is compared against a single hypothesis $\text{H}_0 : \theta = \theta_0$. It is denoted as $\text{H}_1 : \theta \neq \theta_0$.

Based on the partitioning implied by (1.12), the overall prior π can be expressed as $\pi(\theta) = \pi_0 \times \pi_0(\theta) + \pi_1 \times \pi_1(\theta)$ where $\pi_k = \int_{\Theta_k} \pi(\text{d}\theta)$, and $\pi_k(\theta) = \frac{\pi(\theta)1_{\Theta_k}(\theta)}{\int_{\Theta_k} \pi(\text{d}\theta)}$.

Here, π_0, and π_1 describe the prior probabilities on H_0 and H_1, respectively, while $\pi_0(\theta)$ and $\pi_1(\theta)$ describe how the prior mass is spread out over the hypotheses H_0 and H_1, respectively.

We could see the hypothesis testing (1.12) as parametric point inference about the indicator function

$$1_{\Theta_1}(\theta) = \begin{cases} 0 & , \theta \in \Theta_0 \\ 1 & , \theta \in \Theta_1 \end{cases} . \tag{1.13}$$

To estimate (1.13), a reasonable loss function $\ell(\theta, \delta)$ would be the $c_{\mathrm{I}} - c_{\mathrm{II}}$ loss function

$$\ell(\theta, \delta) = \begin{cases} 0 & , \text{if } \theta \in \Theta_0,\ \delta = 0 \\ 0 & , \text{if } \theta \notin \Theta_0,\ \delta = 1 \\ c_{\mathrm{II}} & , \text{if } \theta \notin \Theta_0,\ \delta = 0 \\ c_{\mathrm{I}} & , \text{if } \theta \in \Theta_0,\ \delta = 1 \end{cases} \tag{1.14}$$

where $c_{\mathrm{I}} > 0$ and $c_{\mathrm{II}} > 0$ are specified by the researcher. Here, $c_{\mathrm{I}} > 0$ (and $c_{\mathrm{II}} > 0$) denote the loss if we decide to accept H_0 (and H_1) while the correct answer would be to choose H_1 (H_0). According to (1.5), under (1.14), the Bayes estimator of (1.13) is

$$\delta(x_{1:n}) = \begin{cases} 0 & , \text{if } \pi(\theta \in \Theta_0 | x_{1:n}) > \frac{c_{\mathrm{II}}}{c_{\mathrm{II}} + c_{\mathrm{I}}} \\ 1 & , \text{otherwise} \end{cases} \tag{1.15}$$

where $\pi(\theta \in \Theta_0 | x_{1:n}) = \int_{\Theta_0} \pi(\mathrm{d}\theta | x_{1:n})$. In other words, hypothesis H_1 is accepted if $\frac{\pi(\theta \in \Theta_0 | x_{1:n})}{\pi(\theta \in \Theta_1 | x_{1:n})} < \frac{c_{\mathrm{II}}}{c_{\mathrm{I}}}$.

Hypothesis tests in Bayesian statistics can also be addressed with the aid of Bayes factors. Bayes factor $\mathrm{B}_{01}(x_{1:n})$ is the ratio of the posterior probabilities of H_0 and H_1 over the ratio of the prior probabilities of H_0 and H_1

$$\mathrm{B}_{01}(x_{1:n}) = \frac{\pi(\theta \in \Theta_0 | x_{1:n}) / \pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1 | x_{1:n}) / \pi(\theta \in \Theta_1)} \tag{1.16}$$

$$= \begin{cases} \frac{f(x_{1:n}|\theta_0)}{f(x_{1:n}|\theta_1)} & ; \mathrm{H}_0 : \text{single vs } \mathrm{H}_1 : \text{single} \\ \frac{\int_{\Theta_0} f(x_{1:n}|\theta) \pi_0(\mathrm{d}\theta)}{\int_{\Theta_1} f(x_{1:n}|\theta) \pi_1(\mathrm{d}\theta)} & ; \mathrm{H}_0 : \text{composite vs } \mathrm{H}_1 : \text{composite} \\ \frac{f(x_{1:n}|\theta_0)}{\int_{\Theta_1} f(x_{1:n}|\theta) \pi_1(\mathrm{d}\theta)} & ; \mathrm{H}_0 : \text{single vs } \mathrm{H}_1 : \text{composite} \end{cases} . \tag{1.17}$$

Under the $c_{\mathrm{I}} - c_{\mathrm{II}}$ loss function, (1.15) implies that one would accept H_0 if $B_{01}(x_{1:n}) > \frac{c_{\mathrm{II}}}{c_{\mathrm{I}}} \frac{\pi_1}{\pi_0}$, and accept H_1 if otherwise. Alternatively, Jeffreys [6] developed a scale rule (Table 1.1) to judge the strength of evidence in favor of H_0 or against H_0 brought by the data. Although Jeffreys' rule avoids the need to specify c_I and c_{II}, it is a heuristic rule-of-thumb guide, not based on decision theory concepts, and hence many researchers argue against its use.

Table 1.1 Jeffreys' scale rule [6]

B_{01}	$\log_{10}(B_{01})$	Strength of evidence
$(1, +\infty)$	$(0, +\infty)$	H_0 is supported
$(10^{-1/2}, 1)$	$(-1/2, 0)$	Evidence against H_0: not worth more than a bare
$(10^{-1}, 10^{-1/2})$	$(-1, -1/2)$	Evidence against H_0: substantial
$(10^{-3/2}, 10^{-1})$	$(-3/2, -1)$	Evidence against H_0: strong
$(10^{-2}, 10^{-3/2})$	$(-2, -3/2)$	Evidence against H_0: very strong
$(0, 10^{-2})$	$(-\infty, -2)$	Evidence against H_0: decisive

Example: Bernoulli model (Cont.)

We are interested in testing the hypotheses $\mathrm{H}_0 : \theta = 0.5$ and $\mathrm{H}_1 : \theta \neq 0.5$, given that $\pi_0 = 1/2$, and using the $c_{\mathrm{I}} - c_{\mathrm{II}}$ loss function with $c_{\mathrm{I}} = c_{\mathrm{II}}$. Here, $\Theta_0 = \{0.5\}$ and $\Theta_1 = [0, 0.5) \cup (0.5, 1]$. The overall prior is $\pi(\theta) = \pi_0 1_{\theta_0}(\theta) + (1 - \pi_0)\mathrm{Be}(\theta|a, b)$. The Bayes factor is

$$\mathrm{B}_{01}(x_{1:n}) = \frac{\prod_{i=1}^{n} \mathrm{Br}(x_i|\theta_0)}{\int_{(0,1)} \prod_{i=1}^{n} \mathrm{Br}(x_i|\theta)\mathrm{Be}(\theta|a, b)\mathrm{d}\theta} = \frac{\theta_0^{x_*}(1-\theta_0)^{n-x_*}}{\mathrm{B}(n\bar{x}+a, n-n\bar{x}+b)/\mathrm{B}(a, b)}.$$

Given $a = b = 2$, $n = 30$, and $\sum_{i=1}^{30} x_i = 15$, it is $\mathrm{B}_{01}(x_{1:n}) = 18.47 > c_{\mathrm{II}}/c_{\mathrm{I}} = 1$. Hence, we accept H_1.

1.5.1 Model Selection

Often the researcher is uncertain which statistical model (sampling distribution) can better represent the real data generating process. There is a set $\mathcal{M} = \{\mathfrak{m}_1, \mathfrak{m}_2, \ldots\}$ of candidate statistical models $\mathfrak{m}_k = \{f_k(\cdot|\varphi_k);\ \varphi_k \in \Phi_k\}$, where $f_k(\cdot|\varphi_k)$ denotes the sampling distribution, and φ_k denotes the unknown parameters for $k = 1, 2, \ldots$ Let $\pi_k = \pi(\mathfrak{m}_k)$ denote the marginal model prior and $\pi_k(\varphi_k) = \pi(\varphi_k|\mathfrak{m}_k)$ denote the prior of the unknown parameters φ_k of given model $\mathfrak{m}_k$.

Selection of the 'best' model from a set of available candidate models can be addressed via hypothesis testing. For simplicity, we consider there are only two models $\mathfrak{m}_0$ and $\mathfrak{m}_1$ with unknown parameters $\vartheta_0 \in \Phi_0$ and $\vartheta_1 \in \Phi_1$. Then, model selection is performed as a hypothesis test

$$\mathrm{H}_0 : (\mathfrak{m}, \varphi) \in \Theta_0 \quad \text{vs} \quad \mathrm{H}_1 : (\mathfrak{m}, \varphi) \in \Theta_1 \tag{1.18}$$

where $\Theta_k = \{\mathfrak{m}_k\} \times \Phi_k, \Theta = \cup_k \Theta_k$. The overall joint prior is specified as $\pi(\mathfrak{m}, \varphi) = \pi_0 \times \pi_0(\varphi_0) + \pi_1 \times \pi_1(\varphi_1)$ on $(\mathfrak{m}, \varphi) \in \Theta$ where $\Theta = \cup_k \Theta_k$, where $\pi_k(\varphi_k) = \frac{\pi(\mathfrak{m},\varphi) 1_{\mathfrak{m}_k}(\mathfrak{m})}{\int_{\Phi_k} \pi(\mathfrak{m}, \mathrm{d}\varphi)}$ on $\varphi_k \in \Phi_k$, and $\pi_k = \int_{\Theta_k} \pi(\mathfrak{m}_k, \mathrm{d}\varphi_k)$. Now the model selection problem has been translated into a hypothesis test.

Example: Negative binomial vs. Poisson model [2]

We are interested in testing the hypotheses

$$\mathrm{H}_0 : x_i|\phi \sim \mathrm{Nb}(\phi, 1), \quad \phi > 0, \quad \text{vs.} \quad \mathrm{H}_1 : x_i|\lambda \sim \mathrm{Pn}(\lambda), \quad \lambda > 0$$

by using the $c_{\mathrm{I}} - c_{\mathrm{II}}$ loss function with $c_{\mathrm{I}} = c_{\mathrm{II}}$. Consider two observations $x_1 = x_2 = 2$ are available. Consider overall prior $\pi(\theta)$ with density $\pi(\theta) = \pi_0 \mathrm{Be}(\phi|a_0, b_0) + \pi_1 \mathrm{Ga}(\lambda|a_1, b_1)$ with $\pi_0 = \pi_1 = 0.5$.

This is a composite vs. composite hypothesis test. It is

$$\begin{aligned}\int_{\Theta_0} f(x_{1:n}|\varphi_0)\pi_0(\mathrm{d}\varphi_0) &= \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} \int_0^1 \phi^{n+a_0-1}(1-\phi)^{n\bar{x}+b_0-1}\mathrm{d}\phi \\ &= \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} \frac{\Gamma(n+a_0)\Gamma(n\bar{x}+b_0)}{\Gamma(n+n\bar{x}+a_0+b_0)}\end{aligned}$$

$$\begin{aligned}\int_{\Theta_1} f(x_{1:n}|\varphi_1)\pi_1(\mathrm{d}\varphi_1) &= \frac{b_1^{a_1}}{\Gamma(a_1)(n+b_1)^{n\bar{x}+a_1}} \int_0^\infty \lambda^{n\bar{x}+a_1-1}\exp(-(n+b_1)\lambda)\mathrm{d}\lambda \\ &= \frac{\Gamma(n\bar{x}+a_1)}{\Gamma(a_1)(n+b_1)^{n\bar{x}+a_1}} \frac{1}{\prod_{i=1}^n x_i!}\end{aligned}$$

and hence $B_{01}(x_{1:n}) = \frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)} \frac{\Gamma(n+a_0)\Gamma(n\bar{x}+b_0)}{\Gamma(n+n\bar{x}+a_0+b_0)} \frac{\Gamma(a_1)(n+b_1)^{n\bar{x}+a_1}}{\Gamma(n\bar{x}+a_1)} \prod_{i=1}^n x_i!$. It is $B_{01}(x_{1:n}) = 0.29 > 1$, and hence I accept H_1 and the Poisson model.

References

1. James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.
2. José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.
3. George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.
4. Bruno De Finetti. *Theory of probability: a critical introductory treatment*, volume 6. John Wiley & Sons, 2017.

5. Morris H DeGroot. *Optimal statistical decisions*, volume 82. John Wiley & Sons, 2005.
6. Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.