CHAPTER 5

Tests and Confidence Regions

Twenty-six more tests were going to take the rest of daylight, maybe more. Heat or no heat, the days still grew shorter as if winter really was coming on, and a failed test would take a few minutes longer that one passed, just to make certain.

Robert Jordan, *The Lord of Chaos, Book VI of the Wheel of Time.*

5.1 Introduction

Although testing theory can be perceived as a special case of Decision Theory for a restricted decision space (and even as an estimation problem), we consider testing inference in a separate chapter because there is much more ambiguity about the real inferential purpose of testing than when estimating a regular function of the parameter. In fact, this part of statistical inference is still incomplete, in the sense that many other answers have been proposed, none being entirely satisfactory. In particular, there are pronounced differences between frequentist and Bayesian testing theories. This is nonetheless a setting where a Bayesian approach is quite appealing because the notion of the probability of a hypothesis, like $\pi(\theta \in \Theta_0|x)$, can only be defined through this approach.

Some Bayesians actually think that there should be no testing, or, at least, that there should be no point-null hypothesis testing (see, for instance, Gelfand et al. (1992)), and we shall see in this chapter several philosophical reasons that somehow motivate this radical perspective. These reasons range from the reductive aspect of modeling *(no model is right, but some models are less wrong than others)*, to the unnatural modification of the prior imposed by point-null hypotheses, to the lack of problem-motivated decision-theoretic structure, to the subsequent call to rudimentary conventional $0-1$ losses and acceptance levels, to the impossibility of using improper priors in both point-null hypotheses and model choice settings (Chapter 7). But pragmatic considerations are such that the Bayesian toolbox must also include testing devices, if only because users of Statistics

have been accustomed to testing as a formulation of their problems, given their strong tendency to take problems at face value.

We first consider in Section 5.2 the usual Bayesian approach to testing, i.e., through an evaluation of decisions by $0-1$ losses, and compare the Bayesian procedures with their frequentist counterparts in Section 5.3. We then propose in Section 5.4 an alternative decision-theoretic approach through more adaptive losses which emphasize the "postdata" evaluation of testing procedures (as opposed to Neyman–Pearson procedures for which the evaluation is operated in a "predata" spirit).

This chapter exhibits a strong contrast between Bayesian and frequentist approaches under different evaluation tools; it is revealing because this shows the incompleteness of the classical modeling, which relies on artificial concepts to derive its optimal procedures. Contrary to point estimation settings, these optimal frequentist procedures are no longer limits of Bayes procedures and differ numerically from their Bayesian counterparts. However, we moderate this rejection in Section 5.3 by showing that classical answers may sometimes lead to similar conclusions than noninformative Bayes procedures. Chapter 7 deals with *model choice*, which can be seen as a particular case of testing point-null hypotheses, but also enjoys enough peculiarities to deserve a chapter on its own (besides calling almost uniformly for the computational methods presented in Chapter 6).

5.2 A first approach to testing theory

5.2.1 Decision-theoretic testing

Consider a statistical model $f(x|\theta)$ with $\theta \in \Theta$. Given a subset of interest of Θ, Θ_0, which sometimes consists of a single point $\{\theta_0\}$, the question to be answered is whether the true value of the parameter θ belongs to Θ_0, i.e., to *test* the hypothesis[1]

$$H_0: \ \theta \in \Theta_0,$$

usually called the *null hypothesis*. In linear models, Θ_0 may be a *subspace* of the vector space Θ and the testing problem is then a particular case of a *model choice* problem (see also Chapter 7).

Example 5.2.1 Consider a *logistic regression* model,

$$P_\alpha(y=1) = 1 - P_\alpha(y=0) = \exp(\alpha^t x)/(1+\exp(\alpha^t x)), \qquad \alpha, x \in \mathbb{R}^p,$$

which models the probability of developing a prostate cancer in a lifetime in terms of explanatory variables $x = (x_1, \ldots, x_p)$. Of particular interest is the dependence on work environment variables such as asbestos concentration x_{i_0} and a worker union may want to test whether the coefficient α_{i_0} corresponding to x_{i_0} is null. ‖

[1] There is a certain amount of ambiguity involved in the terminology: *Test* simultaneously denotes the question, and the procedure used to answer the question.

In the Neyman–Pearson perspective (Section 5.3), the testing problem is formalized through a decision space $\mathcal{D}$ restricted to {yes, no} or, equivalently, to $\{1, 0\}$. In fact, it makes sense to perceive testing problems as an inference about the indicator function $\mathbb{I}_{\Theta_0}(\theta)$ and therefore to propose answers in $\mathbb{I}_{\Theta_0}(\Theta) = \{0, 1\}$. Of course, the relevance of such a restriction is less obvious when considering that testing settings often occur as components (or preliminary steps) or more complex inferential structures and, in particular, that the answer to a test question has also consequences in terms of (regular) estimation errors. It may then be more interesting to propose procedures taking values in $[0, 1]$. (We discuss this approach in Section 5.4.)

In some cases, additional information is available about the support of θ, namely, that $\theta \in \Theta_0 \cup \Theta_1 \neq \Theta$. In such settings, we define the *alternative* hypothesis against which we test H_0 as

$$H_1 : \ \theta \in \Theta_1.$$

Under this formalization, every test procedure φ appears as an estimator of $\mathbb{I}_{\Theta_0}(\theta)$ and we only need a loss function $\mathrm{L}(\theta, \varphi)$ to derive the Bayes estimators. For instance, the loss function proposed by Neyman and Pearson is the $0-1$ loss

$$\mathrm{L}(\theta, \varphi) = \begin{cases} 1 & \text{if } \varphi \neq \mathbb{I}_{\Theta_0}(\theta), \\ 0 & \text{otherwise,} \end{cases}$$

introduced in Chapter 2. For this loss, the Bayesian solution is

$$\varphi^\pi(x) = \begin{cases} 1 & \text{if } P^\pi(\theta \in \Theta_0|x) > P^\pi(\theta \in \Theta_0^c|x), \\ 0 & \text{otherwise.} \end{cases}$$

This estimator is easily justified on an intuitive basis since it chooses the hypothesis with the largest posterior probability. A generalization of the above loss is to penalize differently errors when the null hypothesis is true or false. The weighted $0-1$ losses

$$\mathrm{L}(\theta, \varphi) = \begin{cases} 0 & \text{if } \varphi = \mathbb{I}_{\Theta_0}(\theta), \\ a_0 & \text{if } \theta \in \Theta_0 \text{ and } \varphi = 0, \\ a_1 & \text{if } \theta \notin \Theta_0 \text{ and } \varphi = 1, \end{cases} \tag{5.2.1}$$

are called "$a_0 - a_1$" for obvious reasons. The associated Bayes estimator is then given by the following result.

Proposition 5.2.2 *Under the loss* (5.2.1), *the Bayes estimator associated with a prior distribution* π *is*

$$\varphi^\pi(x) = \begin{cases} 1 & \textit{if } P^\pi(\theta \in \Theta_0|x) > \dfrac{a_1}{a_0 + a_1}, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. Since the posterior loss is

$$\begin{aligned} \mathrm{L}(\pi, \varphi|x) &= \int_\Theta \mathrm{L}(\theta, \varphi)\pi(\theta|x)d\theta \\ &= a_0 P^\pi(\theta \in \Theta_0|x)\mathbb{I}_{\{0\}}(\varphi) + a_1 P^\pi(\theta \notin \Theta_0|x)\mathbb{I}_{\{1\}}(\varphi), \end{aligned}$$

the Bayes estimator can be derived directly. □□

For this class of losses, the null hypothesis H_0 is rejected when the posterior probability of H_0 is too small, the *acceptance level* $a_1/(a_0+a_1)$ being determined by the choice of the loss function. Notice that φ^π only depends on a_0/a_1 and that the larger a_0/a_1 is, i.e., the more important a wrong answer under H_0 is relative to H_1, the smaller the posterior probability of H_0 needs to be for H_0 to be accepted.

Example 5.2.3 Consider $x \sim \mathcal{B}(n, p)$ and $\Theta_0 = [0, 1/2]$. Under the uniform prior distribution $\pi(p) = 1$, the posterior probability of H_0 is

$$\begin{aligned} P^\pi(p \leq 1/2|x) &= \frac{\int_0^{1/2} p^x(1-p)^{n-x}dp}{B(x+1, n-x+1)} \\ &= \frac{(1/2)^{n+1}}{B(x+1, n-x+1)} \left\{ \frac{1}{x+1} + \frac{n-x}{(x+1)(x+2)} + \ldots + \frac{(n-x)!x!}{(n+1)!} \right\} \end{aligned}$$

which can be easily computed and compared to the acceptance level. ‖

Example 5.2.4 Consider $x \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu, \tau^2)$. Then $\pi(\theta|x)$ is the normal distribution $\mathcal{N}(\mu(x), \omega^2)$ with

$$\mu(x) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2} \quad \text{and} \quad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

To test $H_0 : \theta < 0$, we compute

$$\begin{aligned} P^\pi(\theta < 0|x) &= P^\pi\left(\frac{\theta - \mu(x)}{\omega} < \frac{-\mu(x)}{\omega}\right) \\ &= \Phi\left(-\mu(x)/\omega\right). \end{aligned}$$

If z_{a_0,a_1} is the $a_1/(a_0+a_1)$ quantile, i.e., if it satisfies $\Phi(z_{a_0,a_1}) = a_1/(a_0 + a_1)$, H_0 is accepted when

$$-\mu(x) > z_{a_0,a_1}\omega,$$

the upper acceptance bound then being

$$-\frac{\sigma^2}{\tau^2}\mu - (1 + \frac{\sigma^2}{\tau^2})\omega z_{a_0,a_1}.$$

‖

Again, notice that, from a Bayesian point of view, it seems natural to base the decision upon the posterior probability that the hypothesis is true. In Section 5.4, we show that an alternative decision-theoretic approach leads to this posterior probability as the Bayesian estimator itself and thus avoids the comparison to a predetermined acceptance level. In fact, a difficulty with the losses (5.2.1) is the choice of the weights a_0 and a_1, since they are usually selected automatically rather than determined from utility considerations.

5.2.2 *The Bayes factor*

While from a decision-theoretic point of view the *Bayes factor* is only a one-to-one transform of the posterior probability, this notion came out to be considered on its own ground in Bayesian testing.

Definition 5.2.5 *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, i.e.,*

$$B_{01}^{\pi}(x) = \frac{P(\theta \in \Theta_0 \mid x)}{P(\theta \in \Theta_1 \mid x)} \bigg/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}.$$

This ratio evaluates the modification of the odds of Θ_0 against Θ_1 due to the observation and can naturally be compared to 1, although an exact comparison scale can only be based upon a loss function. In the particular case where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, the Bayes factor simplifies to the usual *likelihood ratio*

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

In general, the Bayes factor depends on prior information, but is still proposed as an "objective" Bayesian answer, since it partly eliminates the influence of the prior modeling and emphasizes the role of the observations. Actually, it can be perceived as a Bayesian likelihood ratio since, if π_0 is the prior distribution under H_0 and π_1 the prior distribution under H_1, $B_{01}^{\pi}(x)$ can be written as

$$B_{01}^{\pi}(x) = \frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)\, d\theta}{\int_{\Theta_1} f(x|\theta_1)\pi_1(\theta)\, d\theta} = \frac{m_0(x)}{m_1(x)}, \tag{5.2.2}$$

thus replacing the likelihoods with the marginals under both hypotheses. Alternatively, if $\hat{\theta}_0$ is the maximum likelihood estimator on Θ_0 and $\hat{\theta}_1$ the maximum likelihood estimator on Θ_1, the likelihood ratio

$$R(x) = \frac{f(x|\hat{\theta}_0)}{f(x|\hat{\theta}_1)} = \frac{\sup_{\Theta_0} f(x|\theta)}{\sup_{\Theta_1} f(x|\theta)}$$

appears as a particular case of $B_{01}^{\pi}(x)$ when π_0 and π_1 are Dirac masses at $\hat{\theta}_0$ and $\hat{\theta}_1$. This does not legitimize the use of $R(x)$ in the least, though, because π_0 and π_1 both depend on x.

As indicated above, the Bayes factor is, from a decision-theoretic point of view, completely equivalent to the posterior probability of the null hypothesis as, under (5.2.1), H_0 is accepted when

$$B_{01}^{\pi}(x) > \frac{a_1}{a_0} \Big/ \frac{\varrho_0}{\varrho_1} = \frac{a_1 \varrho_1}{a_0 \varrho_0}, \tag{5.2.3}$$

where

$$\varrho_0 = \pi(\theta \in \Theta_0) \quad \text{and}$$
$$\varrho_1 = \pi(\theta \in \Theta_1) = 1 - \varrho_0. \tag{5.2.4}$$

This alternative version of Proposition 5.2.2 thus provides an illustration of the duality existing between loss and prior distribution, already mentioned in Chapter 2. Indeed, (5.2.3) shows that it is equivalent to weight both hypotheses equally, $\varrho_0 = \varrho_1 = 1/2$, and to modify the error penalties into $a_i' = a_i \varrho_i$ $(i = 0, 1)$ or to penalize similarly both types of errors ($a_1 = a_0 = 1$), when the prior distribution incorporates the actual weights in the weighted prior probabilities,

$$\varrho_0' = \frac{a_0 \varrho_0}{a_0 \varrho_0 + a_1 \varrho_1}, \quad \varrho_1' = \frac{a_1 \varrho_1}{a_0 \varrho_0 + a_1 \varrho_1}.$$

Following Good (1958) and Jeffreys (1961), many Bayesians now consider the Bayes factor on its own ground (see, e.g., Kass and Raftery (1995) for a detailed review). In particular, Jeffreys (1961) developed a scale to judge the evidence in favor of or against H_0 brought by the data, *outside a true decision-theoretic setting*. The scale goes as follows:

- if $\log_{10}(B_{10}^{\pi})$ 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*.

Obviously, this scaling of the Bayes factor gives some indication of the strength of the evidence, but the precise bounds separating one strength from another are a matter of convention and they can be arbitrarily changed, as shown in Kass and Raftery (1995). This is a consequence of the lack of decision-theoretic backup via a loss function. (But the same criticism applies to the conventional α levels of 0.05 or 0.01 used for $a_0/(a_0 + a_1)$ in (5.2.1).)

Chapter 6 will give precisions on methods used to approximate Bayes factors when the integral in (5.2.2) cannot be computed analytically, which is often the case.

Example 5.2.6 (Kass and Raftery (1995)) The "hot hand" in basketball is the belief that players have good and bad days, rather than a fixed probability to win a shoot. For a given player, the model under the null hypothesis *(no hot hand)* is then $H_0 : y_i \sim \mathcal{B}(n_i, p)$ $(i = 1, \ldots, G)$, where G denotes the number of games and n_i (y_i) the number of shoots (of good shoots) during the ith game. The model under the general alternative is $H_1 : y_i \sim \mathcal{B}(n_i, p_i)$, the probability p_i varying from game to game. Under a conjugate prior $p_i \sim \mathcal{B}(\xi/\omega, (1-\xi)/\omega)$, the average $\mathbb{E}[p_i|\xi,\omega] = \xi$ is distributed from a uniform prior $\mathcal{U}([0,1])$, as is p under H_0, and ω is fixed. The Bayes factor is then

$$B_{10} = \frac{\int_0^1 \prod_{i=1}^G \int_0^1 p_i^{y_i}(1-p_i)^{n_i-y_i} p_i^{\alpha-1}(1-p_i)^{\beta-1}\, d\,p_i \quad \Gamma(1/\omega)/[\Gamma(\xi/\omega)\Gamma((1-\xi)/\omega)]d\xi}{\int_0^1 p^{\sum_i y_i}(1-p)^{\sum_i (n_i-y_i)}\, d\,p}$$

$$= \frac{\int_0^1 \prod_{i=1}^G [\Gamma(y_i+\xi/\omega)\Gamma(n_i-y_i+(1-\xi)/\omega)/\Gamma(n_i+1/\omega)]\ \Gamma(1/\omega)/[\Gamma(\xi/\omega)\Gamma((1-\xi)/\omega)]d\xi}{\Gamma(\sum_i y_i+1)\Gamma(\sum_i(n_i-y_i)+1)/\Gamma(\sum_i n_i+2)},$$

where $\alpha=\xi/\omega$ and $\beta=(1-\xi)/\omega$. Formally, the numerator can be computed exactly, despite the gamma functions, because of the simplification

$$\Gamma(y_i+\xi/\omega)/\Gamma(\xi/\omega) = \prod_{j=1}^{y_i}(j-1+\xi/\omega),$$

$$\Gamma(n_i-y_i+(1-\xi)/\omega)/\Gamma((1-\xi)/\omega)] = \prod_{j=1}^{n-i-y_i}(j-1+(1-\xi)/\omega)\,,$$

but the function of ξ to be integrated is then a high-degree polynomial. The resolution of the integral thus calls for a software like `Maple` or `Mathematica`. For a given player, the value of B_{10} is 0.16 for $\omega=0.005$ and $G=138$, which does not indicate any conclusive evidence in favor of the hot hand hypothesis. ||

5.2.3 Modification of the prior

The notion of Bayes factor is also instrumental in pointing out an important aspect of Bayesian testing. In fact, this factor is only defined when $\varrho_0\neq 0$ and $\varrho_1\neq 0$. This implies that, if H_0 or H_1 are a priori impossible, the observations will not modify this absolute information: Null probabilities are absorbing states! Therefore, a point-null hypothesis $H_0:\theta=\theta_0$ cannot be tested under a *continuous* prior distribution. More generally, model choice is incompatible with prior distributions that are absolutely continuous with respect to the Lebesgue measure on the largest space.

Testing of point-null hypotheses and the like thus impose a drastic modification of the prior distribution, since it requires to derive a prior distribution on both subsets Θ_0 and Θ_1, for instance, the distributions π_0 and π_1 with densities

$$g_0(\theta)\propto\pi(\theta)\mathbb{I}_{\Theta_0}(\theta),\quad g_1(\theta)\propto\pi(\theta)\mathbb{I}_{\Theta_1}(\theta),$$

(with respect to the natural measures on Θ_0 and Θ_1) although this definition is not always free of ambiguity (see Exercise 5.5). Joined with the prior probabilities ϱ_0 and ϱ_1 of Θ_0 and Θ_1 given by (5.2.4), π_0 and π_1 define the prior π. In other words,

$$\pi(\theta)=\varrho_0\pi_0(\theta)+\varrho_1\pi_1(\theta).$$

(When $\Theta_0=\{\theta_0\}$, the prior distribution on Θ_0 is just the Dirac mass at θ_0.)

This modification of the prior is puzzling from a measure-theoretic point of view, since it puts some weight on a set previously of measure 0. It

also highlights the dichotomy imposed by the usual approach to testing for which the null hypothesis is either right or wrong. However, unless the decision maker is adamant about the prior distribution π, in which case H_0 should be rejected if π does not give any weight to Θ_0, the testing problem can be considered as providing some additional (although vague) information about θ. Indeed, to test for $\theta\in\Theta_0$ implies that there is a chance that θ truly belongs to Θ_0 and therefore that some possibly ill-defined indication has been provided about this fact.

To consider testing settings as sources of information is even more convincing if the final decision is not the answer to the test but the estimation of a function of θ, that is, if the test appears as the choice of a submodel. A preliminary test about the vague information may then improve the estimation step. Moreover, keeping this model choice perspective as the real purpose of the analysis, it also makes sense to build up a separate prior distribution for each subspace because only one of the two Θ_i will be considered after the testing step. For instance, given a ***point-null hypothesis***, $H_0:\theta=\theta_0$, the noninformative distribution $\pi(\theta)=1$ cannot be considered as an acceptable prior on Θ because the particular value θ_0 has been singled out as a possible value for θ. (In Chapter 7, we shall further defend the perspective that similar parameters appearing in two different models must be considered as separate entities.) In general, to consider that the testing problem occurs because of (unavailable) additional observations may help in the derivation of a noninformative prior, even though there is no consensus on noninformative Bayes modeling for tests (see Section 5.3.5).

5.2.4 Point-null hypotheses

A usual criticism of point-null hypothesis settings is that they are not *realistic* (see, e.g., Casella and Berger[2] (1987)). For instance, as pointed out by Good (1980), it does not actually make sense to test whether the probability of rain for tomorrow is 0.7163891256 . . .[3] However, some statistical problems definitely call for point-null hypothesis testing. For instance, in mixture estimation (see Section 1.1 and Section 6.4), it may be important to know whether a mixture distribution has two or three components, so one may test whether one of the component weights is 0. Similarly, in linear regression, tests on the nullity of the regression coefficients are useful for the elimination of useless exogenous variates, as in Example 5.2.1. Even more to the point, testing whether the Universe is expanding, contracting or stable is akin to testing whether a certain constant is larger, smaller or equal to a specific value.

More generally, two-sided hypotheses such as $H_0:\ \theta\in\Theta_0=[\theta_0-\epsilon,\theta_0+$

[2] Roger Berger, not James Berger!

[3] But it would still make sense to test whether the prediction of 75% given by the local weather forecaster is exact, that is, whether the probability of rain for the given day is 0.75, or another of the probabilities announced by the forecaster (see Example 2.3.4).

$\epsilon]$ can be approximated by $H_0 : \theta = \theta_0$, with hardly any modification of the posterior probabilities when ϵ is small enough. For instance, this happens when the likelihood is constant in a neighborhood of θ_0 (see Berger (1985a) and Berger and Delampady (1987)). Point-null hypotheses are also quite important in practice; for instance, while it makes sense to determine whether a medical treatment has a positive or negative effect, the first issue may be to decide whether it has any effect at all.

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 $\pi_0(\theta) = \varrho_0 \mathbb{I}_{\Theta_0}(\theta) + (1-\varrho_0) g_1(\theta)$ and the posterior probability of H_0 is given by

$$\pi(\Theta_0|x) = \frac{f(x|\theta_0)\varrho_0}{\int f(x|\theta)\pi(\theta)\, d\theta} = \frac{f(x|\theta_0)\varrho_0}{f(x|\theta_0)\varrho_0 + (1-\varrho_0) m_1(x)},$$

the marginal distribution on H_1 being

$$m_1(x) = \int_{\Theta_1} f(x|\theta) g_1(\theta)\, d\theta.$$

This posterior probability can also be written as

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\varrho_0}{\varrho_0} \frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

Similarly, the Bayes factor is

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)\varrho_0}{m_1(x)(1-\varrho_0)} \Big/ \frac{\varrho_0}{1-\varrho_0} = \frac{f(x|\theta_0)}{m_1(x)}$$

and we derive the following general relation between the two quantities:

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\varrho_0}{\varrho_0} \frac{1}{B_{01}^{\pi}(x)}\right]^{-1}.$$

Example 5.2.7 (Example 5.2.3 continued) Consider the test of $H_0 : p = 1/2$ against $p \neq 1/2$. For $g_1(p) = 1$, the posterior probability is then given by

$$\begin{aligned}\pi(\Theta_0|x) &= \left[1 + \frac{1-\varrho_0}{\varrho_0} 2^n B(x+1, n-x+1)\right]^{-1} \\ &= \left[1 + \frac{1-\varrho_0}{\varrho_0} \frac{x!(n-x)!}{(n-1)!} 2^n\right]^{-1},\end{aligned}$$

since $m(x) = \binom{n}{x} B(x+1, n-x+1)$. For instance, if $n = 5$, $x = 3$, and $\varrho_0 = 1/2$, the posterior probability is

$$\left(1 + \frac{2}{120} 2^5\right)^{-1} = \frac{15}{23}$$

and the corresponding Bayes factor is $15/8$, close to 2. So, in the most supporting cases, the posterior probabilities tend to favor H_0. When the

Table 5.2.1. *Posterior probabilities of $p = 1/2$ when $n = 10$.*

x	0	1	2	3	4	5
$P(p = 1/2 \mid x)$	0.0055	0.0953	0.3737	0.6416	0.7688	0.8025

Table 5.2.2. *Posterior probabilities of $\theta = 0$ for different values of $z = x/\sigma$ and for $\tau = \sigma$.*

z	0	0.68	1.28	1.96
$\pi(\theta = 0 \mid z)$	0.586	0.557	0.484	0.351

sample size increases, the range of the possible answers also widens. For instance, if $\pi(p)$ is $\mathcal{B}e(1/2, 1/2)$ and $n = 10$, the posterior probabilities are given in Table 5.2.1 and support H_0 for x close to 5, even though the prior distribution is rather biased against the null hypothesis (since it heavily weights the extreme values, 0 and 1). $\|$

Example 5.2.8 (Example 5.2.4 continued) Consider the test of $H_0 : \theta = 0$. It seems reasonable to choose π_1 as $\mathcal{N}(\mu, \tau^2)$ and $\mu = 0$, if no additional information is available. Then

$$\begin{aligned}\frac{m_1(x)}{f(x|0)} &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}} \frac{e^{-x^2/2(\sigma^2+\tau^2)}}{e^{-x^2/2\sigma^2}} \\ &= \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left\{\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right\},\end{aligned}$$

and the posterior probability can be derived as

$$\pi(\theta = 0|x) = \left[1 + \frac{1-\varrho_0}{\varrho_0} \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left(\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right]^{-1}.$$

In the special case when $\varrho_0 = 1/2$ and $\tau = \sigma$, Table 5.2.2 gives the posterior probabilities in terms of $z = x/\sigma$. $\|$

Consider now the alternative case $\tau^2 = 10\sigma^2$; it is supposed to indicate a more diffuse prior information on θ. The posterior probabilities of H_0 are then modified as shown in Table 5.2.3.

5.2.5 Improper priors

The recourse to noninformative prior distributions for testing hypotheses is rather delicate, and DeGroot (1973) states that improper priors should not be used *at all* in tests. In fact, as noticed previously, the testing setting

Table 5.2.3. *Posterior probabilities of $\theta = 0$ for $\tau^2 = 10\sigma^2$ and $z = x/\sigma$.*

z	0	0.68	1.28	1.96
$\pi(\theta = 0\|x)$	0.768	0.729	0.612	0.366

Table 5.2.4. *Posterior probabilities of $|\theta| < 1$.*

x	0.0	0.5	1.0	1.5	2.0
$\pi(\|\theta\| \leq 1\|x)$	0.683	0.625	0.477	0.302	0.157

is not coherent with an absolute lack of information, since it implies at least a division of the parameter space into two subsets, out of which at least one is of measure zero under the conventional Jeffreys prior. However, the inconvenience with the use of improper prior distributions goes deeper, since they are incompatible with most tests of point-null hypotheses.

We illustrate this difficulty in the normal setting, $x \sim \mathcal{N}(\theta, 1)$, for the point-null hypothesis $H_0 : \theta = 0$ to test against $H_1 : \theta \neq 0$. If we use the improper prior $\pi(\theta) = 1$ on $\{\theta \neq 0\}$, i.e., if π is

$$\pi(\theta) = \frac{1}{2}\mathbb{I}_0(\theta) + \frac{1}{2} \cdot 1,$$

the posterior probability of H_0 is

$$\pi(\theta = 0|x) = \frac{e^{-x^2/2}}{e^{-x^2/2} + \int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\, d\theta} = \frac{1}{1 + \sqrt{2\pi}e^{x^2/2}}.$$

(The particular choice of the constant 1 in the prior is crucial for the following discussion, while being arbitrary, as seen below.) Therefore, this posterior probability of H_0 is bounded from above by $1/(1 + \sqrt{2\pi}) = 0.285$. This implies that the posterior distribution is rather biased against H_0, even in the most favorable case. Unless the scale of comparison, that is, the loss, is modified to account for these low values, the null hypothesis will quite often be rejected . A similar phenomenon occurs when Θ_0 is compact. For instance, the test of $H_0 :\ |\theta| \leq 1$ versus $H_1 :\ |\theta| > 1$ leads to the following posterior probability:

$$\begin{aligned}\pi(|\theta| \leq 1|x) &= \frac{\int_{-1}^{1} e^{-(x-\theta)^2/2}\, d\theta}{\int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\, d\theta} \\ &= \Phi(1 - x) - \Phi(-1 - x) \\ &= \Phi(x + 1) - \Phi(x - 1),\end{aligned}$$

whose numerical values are given in Table 5.2.4. Therefore, the maximal support of H_0, 0.683, is still moderate.

Table 5.2.5. *Posterior probabilities of $\theta = 0$ for the Jeffreys prior $\pi(\theta) = 1$.*

x	0.0	1.0	1.65	1.96	2.58
$\pi(\theta = 0\|x)$	0.285	0.195	0.089	0.055	0.014

An interesting feature of the Lebesgue prior distribution can be exhibited for the point-null hypothesis $H_0 : \theta = 0$. The resulting procedure agrees with the corresponding classical answer, as shown by Table 5.2.5. The posterior probability $\pi(\theta = 0|x)$ is indeed quite close to the classical significance levels 0.10, 0.05, and 0.01 when x is 1.65, 1.96, or 2.58 (it will be demonstrated in Section 5.3.4 that this comparison is meaningful). This coincidence does not hold for all values of x but shows that, for usual significance levels (and testing purposes), the classical answer could be considered as a noninformative Bayes answer, even though it corresponds to a hardly defendable prior.

Another illustration of the delicate issue of improper priors in testing settings is provided by the *Jeffreys–Lindley paradox*. In fact, limiting arguments are not valid in testing settings and prevent an alternative derivation of noninformative answers. For instance, considering the conjugate prior distributions introduced in Example 5.2.4, the posterior probabilities are

$$\pi(\theta = 0|x) = \left\{1 + \frac{1 - \varrho_0}{\varrho_0}\sqrt{\frac{\sigma^2}{\sigma^2 + \tau^2}} \exp\left[\frac{\tau^2 x^2}{2\sigma^2(\sigma^2 + \tau^2)}\right]\right\}^{-1},$$

which converge to 1 when the prior variance τ goes to $+\infty$, for every x. This limit differs from the "noninformative" answer derived previously $[1 + \sqrt{2\pi}\exp(x^2/2)]^{-1}$ and, more importantly, is totally useless. This phenomenon can also be observed by comparing Tables 5.2.2 and 5.2.3, since the probability is larger when $\tau^2 = 10\sigma^2$ than when $\tau = \sigma$ for all the values of z considered in the tables. See Aitkin (1991) and Robert (1993b) for recent discussions on this paradox.

Paradoxes associated with improper priors like the Jeffreys–Lindley example are actually because of a weighting indeterminacy that does not occur for point estimation, or even for one-sided tests.

Example 5.2.9 Consider $x \sim \mathcal{N}(\theta, 1)$ and $H_0 :\ \theta \leq 0$ to test versus $H_1 : \theta > 0$. For the diffuse distribution $\pi(\theta) = 1$,

$$\begin{aligned}\pi(\theta \leq 0|x) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} e^{-(x-\theta)^2/2}\, d\theta \\ &= \Phi(-x).\end{aligned}$$

In this case, the generalized Bayes answer is also the classical procedure, called the *p-value* (see Section 5.3.4). ||

For two-sided problems, if g_0 and g_1 are σ-finite measures corresponding

Table 5.2.6. *Posterior probabilities of $\theta = 0$ for the Jeffreys prior $\pi(\theta) = 10$.*

x	0.0	1.0	1.65	1.96	2.58
$\pi(\theta = 0\|x)$	0.0384	0.0236	0.0101	0.00581	0.00143

to truncated noninformative priors on the subspaces Θ_0 and Θ_1, the choice of the normalizing constants will influence the Bayesian estimator. In fact, if g_i is replaced by $c_i g_i$ $(i = 0, 1)$, the Bayes factor is multiplied by c_0/c_1. For instance, if the Jeffreys prior is uniform and $g_0 = c_0$, $g_1 = c_1$, the posterior probability is

$$\begin{aligned}\pi(\theta \in \Theta_0|x) &= \frac{\varrho_0 c_0 \int_{\Theta_0} f(x|\theta)\, d\theta}{\varrho_0 c_0 \int_{\Theta_0} f(x|\theta)\, d\theta + (1-\varrho_0) c_1 \int_{\Theta_1} f(x|\theta)\, d\theta} \\ &= \frac{\varrho_0 \int_{\Theta_0} f(x|\theta)\, d\theta}{\varrho_0 \int_{\Theta_0} f(x|\theta)\, d\theta + (1-\varrho_0)[c_1/c_0] \int_{\Theta_1} f(x|\theta)\, d\theta},\end{aligned}$$

which does depend on the ratio c_1/c_0.

It is therefore necessary to extend the noninformative perspective to these testing settings by developing a technique able to derive the weights c_i in a noninformative and acceptable way. Bernardo (1980), Spiegelhalter and Smith (1980), Smith and Spiegelhalter (1982), Aitkin (1991), Pettit (1992), Robert (1993b) and Berger and Pericchi (1996a,b) have made proposals in this direction, as detailed in Section 5.2.6. Notice that Jeffreys (1961) proposed instead to use proper priors in such settings, like $\mathcal{C}(0, \sigma^2)$ or $\mathcal{N}(0, 10\sigma^2)$ in the case of $x \sim \mathcal{N}(\theta, \sigma^2)$ and $H_0 : \theta = 0$. The problem is then that the choice of the proper prior distribution will influence the answer to the test. For instance, the equivalent of Table 5.2.5 for $\pi(\theta) = 10$ are given in Table 5.2.6, with important differences for most values of x, since they vary by an order of magnitude.

Before introducing in Section 5.2.6 some of the recent developments linked to the use of improper priors, let us make the following point: it still does not feel right to use improper priors, like the Jeffreys priors, for two-sided tests because they seem to lead to too much arbitrariness, in the sense that many competing solutions abound, with about the same theoretical background and with different numerical values, in contradiction with the Likelihood Principle. In other words, although the solutions proposed in the next section are interesting and overcome, as working principles, the difficulties related to the use of improper priors, they are not, strictly speaking, pertaining to the Bayesian paradigm. We consider in Section 5.3 an alternative approach which defines a least favorable Bayesian answer as a lower bound on the (proper) Bayes estimators.

The difficulties encountered with noninformative priors in testing settings also point out that a testing problem cannot be treated in a coherent way

if no prior information is available, that is, that the information brought by the observations alone is usually not enough to infer about the truth of a hypothesis in a categorical fashion (*yes/no*). This obviously reinforces the motivation for a Bayesian treatment of such testing problems, since it is the only coherent approach taking advantage of the residual information.

5.2.6 Pseudo-Bayes factors

Most[4] of the solutions proposed to overcome the difficulties related to the use of improper priors (see Section 5.2.5) are based on the idea to use part of the data/information to transform the priors into proper distributions, or to call to imaginary observations to obtain the same effect.

Definition 5.2.10 *Given an improper prior π, a sample $(x_1, \ldots, x_n)$ is a training sample if the corresponding posterior $\pi(|x_1, \ldots, x_n)$ is proper and is a minimal training sample if no subsample is a training sample.*

Example 5.2.11 For the $\mathcal{N}(\mu, \sigma^2)$ model, the minimal training sample size associated with the improper prior $\pi_0(\mu, \sigma^2) = 1/\sigma^2$ is 2 since

$$\begin{aligned}&\int e^{-\{(x_1-\mu)^2+(x_2-\mu)^2\}/2\sigma^2} \sigma^{-4} d\mu\, d\sigma^2 \\ &= \int_0^\infty \sigma^{-3} e^{-s^2/2\sigma^2} d\sigma^2 \\ &= \int_0^\infty \omega^{3/2-2} e^{-s^2\omega/2} d\omega,\end{aligned}$$

while

$$\int e^{-(x_1-\mu)^2/2\sigma^2} \sigma^{-3} d\mu\, d\sigma^2 = \infty .$$

If we now consider the prior $\pi_1(\mu, \sigma^2) = 1/\sigma$, the minimal training sample size is 3 since

$$\begin{aligned}&\int e^{-\{(x_1-\mu)^2+(x_2-\mu)^2\}/2\sigma^2} \sigma^{-3} d\mu\, d\sigma^2 \\ &= \int_0^\infty \sigma^{-2} e^{-s^2/2\sigma^2} d\sigma^2 \\ &= \int_0^\infty \omega^{-1} e^{-s^2\omega/2} d\omega = \infty,\end{aligned}$$

which is a good argument in favor of π_0 against π_1. ||

The idea is then to use a minimal training sample, $x_{(\ell)}$ say, to "properize" the improper prior π into $\pi(\cdot|x_{(\ell)})$, and then to use this posterior distribution *as if* it were a regular proper prior for the remainder of the sample,

[4] This section, which may be skipped on a first reading, contains more advanced material on the so-called intrinsic priors developed by Berger and Pericchi (1996a,b). It will not be used in the rest of the book, except in Chapter 7. See Berger and Pericchi (2001) for a much more detailed review, upon which this section is based.

$x_{(-\ell)}$ say, in order to avoid the data twice, as in the proposal of Aitkin (1991). When facing a hypothesis H_0 with a prior distribution π_0, with a broader alternative H_1 with a prior distribution π_1, if the minimal training sample under H_1 is such that $\pi_0(\cdot|x_{(\ell)})$ is also proper, the *pseudo-Bayes factor*

$$B_{10}^{(\ell)} = \frac{\int_{\Theta_1} f_1(x_{(-\ell)}|\theta_1)\pi_1(\theta_1|x_{(\ell)})d\theta_1}{\int_{\Theta_0} f_0(x_{(-\ell)}|\theta_0)\pi_0(\theta_0|x_{(\ell)})d\theta_0} \tag{5.2.5}$$

is then independent from the normalizing constants used in both π_0 and π_1. To see this more clearly, consider the following representation proposed in Berger and Pericchi (2001).

Lemma 5.2.12 *In the case of independent distributions, the pseudo-Bayes factor can be written as*

$$B_{10}^{(\ell)} = B_{10}(x) \times B_{01}(x_{(\ell)}) , \tag{5.2.6}$$

with

$$B_{10}(x) = \frac{\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\Theta_0} f_0(x|\theta_0)\pi_0(\theta_0)d\theta_0}$$

and

$$B_{01}(x_{(\ell)}) = \frac{\int_{\Theta_0} f_0(x_{(\ell)}|\theta_0)\pi_0(\theta_0)d\theta_0}{\int_{\Theta_1} f_1(x_{(\ell)}|\theta_1)\pi_1(\theta_1)d\theta_1} .$$

In this decomposition, $B_{10}(x)$ and $B_{01}(x_{(\ell)})$ are the Bayes factors computed for the unormalized priors π_1 and π_0, for the whole sample x and the training sample $x_{(\ell)}$, respectively, as if these were regular priors. It is then straightforward to see that multiplying π_0 by c_0 and π_1 by c_1 has no influence on $B_{10}^{(\ell)}$ because these constants cancel. Notice the interesting inversion from $B_{10}(x)$ to $B_{01}(x_{(\ell)})$: the training sample effect is removed from the Bayes factor $B_{10}(x)$.

While the normalizing constant problem disappears, a remaining difficulty is that the solution $B_{10}^{(\ell)}$ is not strictly Bayesian. Moreover, besides sequential sampling, there is no obvious choice for $x_{(\ell)}$, while this choice of the training sample influences the resulting value of $B_{10}^{(\ell)}$ (and thus violates the Likelihood Principle).

Example 5.2.13 (Example 5.2.11 continued) If $H_0 : \mu = 0$, with $\pi_0(\sigma^2) = 1/\sigma^2$, and $H_1 : \mu \neq 0$, with $\pi_1(\mu, \sigma^2) = 1/\sigma^2$, the minimal training sample of size 2 under H_1. Then

$$\pi_1(\mu, \sigma^2|x_1, x_2) = \frac{1}{\sigma}\exp\{-2(\mu - \bar{x}_1)^2/2\sigma^2\}s_1^5\sigma^{-3}e^{-s_1^2/2\sigma^2}$$

and

$$\pi_0(\sigma^2|x_1, x_2) = \frac{s_0^6}{\sigma^4}e^{-s_0^2/2\sigma^2} ,$$

with the following notations:

$$\bar{x}_1 = \frac{x_1 + x_2}{2}, \quad s_1^2 = \frac{(x_1 - x_2)^2}{2}, \quad s_0^2 = x_1^2 + x_2^2 .$$

Then

$$B_{10}^{(2)} = \frac{s_1^5 \int e^{-\{(n-2)(\bar{x}_2-\mu)^2-2(\mu-\bar{x}_1)^2-s_2^2-s_1^2\}/2\sigma^2}\sigma^{-n-2}d\mu d\sigma^2}{s_0^6 \int_0^\infty e^{-\{-s_3^2-s_0^2\}/2\sigma^2}\sigma^{-n-2}d\sigma^2}$$

depends on the choice of (x_1, x_2) through $(\bar{x}_1 - \bar{x}_2)^2$, s_1^2, s_0^2 (see Exercise 5.14). ||

A way to remove this dependence on the training sample is to average the different pseudo-Bayes factors (5.2.6) over all the possible training samples $x_{(\ell)}$. The next difficulty is to decide what kind of averaging should be used here. For instance, Berger and Pericchi (1996a, 1998, 2001) list

- the *arithmetic intrinsic Bayes factor*,

  $$B_{10}^A = \frac{1}{L}\sum_{x_{(\ell)}} B_{10}^{(\ell)} = B_{10}(x)\frac{1}{L}\sum_{x_{(\ell)}} B_{01}(x_{(\ell)}) , \tag{5.2.7}$$

  where L is the number of different training samples;
- the *geometric intrinsic Bayes factor*,

  $$B_{10}^G = \exp\frac{1}{L}\sum_{x_{(\ell)}} \log B_{10}^{(\ell)} = B_{10}(x)\exp\frac{1}{L}\sum_{x_{(\ell)}} \log B_{01}(x_{(\ell)}) ; \tag{5.2.8}$$

  and
- the *median intrinsic Bayes factor*,

  $$B_{10}^M = \text{med}\, B_{10}^{(\ell)} = B_{10}(x)\text{med}\, B_{01}(x_{(\ell)}) , \tag{5.2.9}$$

  where med $B_{10}^{(\ell)}$ denotes the median of the $B_{10}^{(\ell)}$ over the different training samples.

Although these solutions all come close to a Bayesian answer, in particular by using the data only once (Exercise 5.15) by separating the part used to "properize" the improper prior and the part used to run the test, none of the above is a true Bayesian answer. We shall discuss below more serious drawbacks to these different intrinsic Bayes factors, but it appears that they sometimes correspond to genuine Bayes factors under different prior distributions, called *intrinsic priors* in Berger and Pericchi (1996a, 1998).[5] (This phenomenon will also occur in Section 5.3.5 with Berger and Sellke's (1987) lower bounds.)

Example 5.2.14 (Berger and Pericchi (1998)) In the $\mathcal{N}(\theta, 1)$ case, when $H_0 : \theta = 0$ and $\pi_1(\theta) = 1$, for a sample $(x_1, \ldots, x_n)$, the arithmetic intrinsic Bayes factor,

$$B_{10}^A = B_{10}(x)\frac{1}{\sqrt{2\pi}}\frac{1}{n}\sum_{i=1}^n e^{-x_i^2/2},$$

[5] The name *intrinsic* associated with the Bayes factor and the corresponding prior tries to convey an idea of quantities solely derived from the distribution, but the diversity of possible answers shows this is somehow a misnomer!

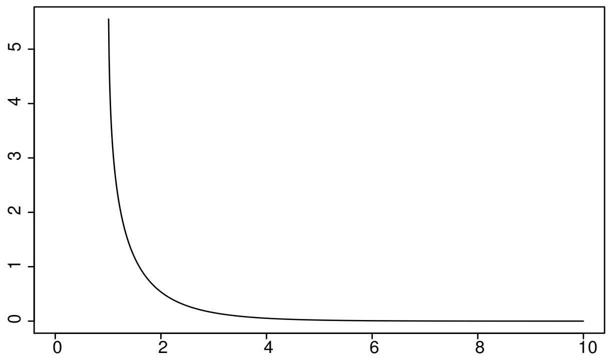

Figure 5.2.1. *Graph of the intrinsic prior associated with the exponential testing, when* $\theta_0 = 1$.

is almost identical to the regular Bayes factor associated with a normal $\mathcal{N}(0,2)$ prior under H_1. ||

Example 5.2.15 (Berger and Pericchi (1998)) For $x_1, \ldots, x_n$ i.i.d. observations from a translated exponential distribution, with density $\exp(\theta - x)\mathbb{I}_{x\geq\theta}$, if $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$, with $\pi_1(\theta) = 1$,

$$B_{10}^A = B_{10}(x)\frac{1}{n}\sum_{i=1}^n \left[e^{x_i-\theta_0} - 1\right]^{-1},$$

which corresponds to the regular Bayes factor associated with the proper prior

$$\pi_2(\theta) = e^{\theta_0-\theta}\left\{1 - \log\left(1 - e^{\theta_0-\theta}\right)\right\},$$

which behaves as indicated on Figure 5.2.1. ||

O'Hagan (1995) introduces an alternative to intrinsic Bayes factors that avoids the selection of—and the subsequent averaging over—training samples. His idea is to use a *fraction b* of the likelihood to properize the prior, that is, to take $0 < b < 1$ such that

$$\int_{\Theta_0} f_0(x|\theta_0)^b \pi_0(\theta_0) d\theta_0 < \infty$$

and

$$\int_{\Theta_1} f_1(x|\theta_1)^b \pi_1(\theta_1) d\theta_1 < \infty .$$

The remaining fraction $(1-b)$ of the likelihood is then used to run the test, as in the intrinsic Bayes factor case. The *fractional Bayes factor* is thus defined as

$$B_{10}^F = \frac{\int_{\Theta_1} f_1(x|\theta_1)^{1-b}\pi_1^b(\theta_1|x)d\theta_1}{\int_{\Theta_0} f_0(x|\theta_0)^{1-b}\pi_0^b(\theta_0|x)d\theta_0}$$

$$= B_{10}(x)\frac{\int_{\Theta_0} f_0(x|\theta_0)^b\pi_0(\theta_0)d\theta_0}{\int_{\Theta_1} f_1(x|\theta_1)^b\pi_1(\theta_1)d\theta_1}, \tag{5.2.10}$$

where $\pi_0^b(\theta_0|x)$ and $\pi_1^b(\theta_1|x)$ denote the pseudo-posteriors associated with $f_0(x|\theta_0)^b$ and $f_1(x|\theta_1)^b$, respectively. For exponential families, the fraction b clearly corresponds to a training sample size, since for n observations from an exponential family with sufficient statistic T,

$$\exp\{\theta \cdot n\,T(x) - n\Psi(\theta)\}^b = \exp\{\theta \cdot [bn]\,T(x) - [bn]\Psi(\theta)\} .$$

For other distributions, the fraction b must be determined through a more ad-hoc approach (see O'Hagan (1995, 1997)).

As in the intrinsic Bayes factor case, there exist occurrences where this solution is equal to a regular Bayes factor with a corresponding "intrinsic" prior.

Example 5.2.16 (Example 5.2.14 continued) For any $0 < b < 1$,

$$B_{10}^F = \frac{\int e^{-n(1-b)(\bar{x}-\theta)^2/2}\sqrt{b}e^{-nb((\bar{x}-\theta)^2/2}d\theta}{\sqrt{2\pi}e^{-n(1-b)\bar{x}^2/2}}$$

$$= \sqrt{b}e^{n(1-b)\bar{x}^2/2}, \tag{5.2.11}$$

which is equal to the Bayes factor associated with the proper $\theta \sim \mathcal{N}(0,(1-b)/nb)$ under H_1. ||

There are, however, enough difficulties with these pseudo-Bayes factors to make us question their use in testing and model choice problems:

(i) Bayes factors, when associated with proper priors, do satisfy some *coherence* properties such as

$$B_{12} = B_{10}B_{02} \qquad \text{and} \qquad B_{01} = 1/B_{10} .$$

Most pseudo-Bayes factors do not, even though the fractional Bayes factor satisfies $B_{01}^F = 1/B_{10}^F$.

(ii) When the pseudo-Bayes factors can be expressed as true Bayes factors, the corresponding intrinsic priors are not necessarily appealing, as shown in Example 5.2.14 for the arithmetic Bayes factor and in Example 5.2.16 for the fractional Bayes factor, and these priors do depend on the choice of the improper reference priors π_0 and π_1, hence are hardly intrinsic.

(iii) Following the above point, the pseudo-Bayes factors may also exhibit a bias in favor of one of the hypotheses, in the sense that they can be expressed as a true Bayes factor multiplied by a certain factor.

Example 5.2.17 (Example 5.2.15 continued) For the median intrinsic Bayes factor,

$$B_{10}^M = B_{10}(x)\left[e^{\mathrm{med}(x_i)} - \theta_0\right]^{-1}$$
$$= 0.69\tilde{B}_{10}(x) \tag{5.2.12}$$

where $\tilde{B}_{10}(x)$ is the Bayes factor associated with the prior $\pi_3(\theta) \propto (2\exp\{\theta-\theta_0\}-1)^{-1}$, which, while being similar to π_2, does not provide exactly the same coverage of regions near 1. ||

In such cases, the pseudo-Bayes factors can be seen as modifying the probabilities of both hypotheses from the reference value 1/2, a feature we shall also encounter for least favorables bounds in Section 2.5.2.

(iv) Most often, however, the pseudo-Bayes factors do not correspond to any true Bayes factor, and they may give strongly biased solutions. For instance, Berger and Pericchi (2001) state that arithmetic intrinsic Bayes factors do not have intrinsic priors for most one-sided testing problems.

Example 5.2.18 (Example 5.2.15 continued) The fractional Bayes factor,

$$B_{10}^F = B_{10}(x)bn\left\{e^{-bn(x_{(1)}-\theta_0)} - 1\right\}^{-1}, \tag{5.2.13}$$

is always larger than 1, thus, always favors the alternative hypothesis, according to Jeffreys' scale. This paradoxical behavior can be related to the fact that the fraction b does not modify the indicator function. ||

(v) Pseudo-Bayes factors may simply not exist for a whole class of models.

Example 5.2.19 Mixtures of normal distributions

$$p\mathcal{N}(\mu_1,\sigma_1^2) + (1-p)\mathcal{N}(\mu_2,\sigma_2^2)$$

were presented in Example 1.1.6. As shown in Exercise 1.57, improper priors of the form $\pi_1(\mu_1,\sigma_1)\pi_2(\mu_2,\sigma_2)$ cannot be used in this setting, whatever the sample size n. (The fundamental reason for this is that there is a probability $(1-p)^n$ that no observation is associated with the first component $\mathcal{N}(\mu_1,\sigma_1^2)$.) Therefore, standard noninformative priors do not allow for training samples and intrinsic Bayes factors cannot be constructed. The same applies to fractional Bayes factors (see Exercise 5.21). ||

(vi) As shown in this section, there are many ways of defining pseudo-Bayes factors and, while most are arguably logical, there is no coherent way of ordering them. Pseudo-Bayes factors, as defined here, do agree with the Likelihood Principle, but the multiplication of possible answers, even if those are close, is not a good signal to users.[6] Similarly, there is no clear-cut procedure for the choice of the fraction b in the fractional Bayes factors, since the minimal training sample size is not always clearly defined.

(vii) The issue of computing pseudo-Bayes factors has not been mentioned so far, for lack of appropriate tools, which will be introduced in Chapters 6 and 7. But each Bayes factor $B_{10}^{(\ell)}$ may be a complex integral and the derivation of the averaged intrinsic Bayes factor may involve $\binom{m}{n}$ integrals of this kind, if m is the minimal training sample size. Fractional Bayes factors are easier to compute in exponential settings, but other distributions are much more difficult to handle (Exercise 5.22).

5.3 Comparisons with the classical approach

5.3.1 UMP and UMPU tests

The classical approach to testing theory is the theory of Neyman–Pearson, as presented, for instance, in Lehmann (1986). For the $0-1$ loss, denoted L below, the frequentist notion of optimality is based on the power of a test, defined as follows:

Definition 5.3.1 *The power of a testing procedure φ is the probability of rejecting H_0 under the alternative hypothesis, that is, $\beta(\theta) = 1 - \mathbb{E}_\theta[\varphi(x)]$ when $\theta \in \Theta_1$. The quantity $1-\beta(\theta)$ is called type–two error, while the type–one error is $\mathbb{E}_\theta[\varphi(x)]$ when $\theta \in \Theta_0$.*

Optimal frequentist tests are then those that minimize the risk $\mathbb{E}_\theta[\mathrm{L}(\theta, \varphi(x))]$ *only under* H_1:

Definition 5.3.2 *If $\alpha \in]0,1[$ and $\mathcal{C}_\alpha$ is the class of the procedures φ satisfying the following constraint on the type I error:*

$$\sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\mathrm{L}(\theta,\varphi(x))] = \sup_{\theta\in\Theta_0} P_\theta(\varphi(x)=0) \le \alpha,$$

a test procedure φ is said to be uniformly most powerful at level α (UMP) if it minimizes the risk $\mathbb{E}_\theta[\mathrm{L}(\theta,\varphi(x))]$ uniformly on Θ_1 in $\mathcal{C}_\alpha$.

[6] Berger and Pericchi (2001) argue that the multiplicity of intrinsic Bayes factors is no more worrying than the multiplicity of possible default priors. The parallel is slightly flawed, though, because *every* default prior induces a multiplicity of intrinsic Bayes factors!

a null hypothesis—for instance, the elliptical nature of planet trajectories—is always rejected from a frequentist point of view for a sample size large enough, i.e., when enough observations have accumulated. ||

Another major difficulty with the Neyman–Pearson approach is that the selection of the level α should be equivalent to the selection of the weights a_0 and a_1 in the loss function, and thus should be based on utility considerations. Instead, the practice is to completely bypass this selection and, following a suggestion by Fisher (1956), it became a formal rule to adapt "classical" α-levels of 5% or 1%, no matter what the problem, the sample size, or the type II error were. Since the Neyman–Pearson approach is quite predominant nowadays, this dogmatic attitude created a publication bias, since results of experiments which are not "*significant at level* 5%" are most often rejected by editors or even censored by the authors themselves in many fields, including biology, medicine, and the social sciences.

5.3.4 The p-values

Frequentists (and practitioners) tried to compensate for the drawbacks of the Neyman–Pearson approach by removing the significance level α and proposing an answer taking values in $[0,1]$, and more importantly depending on the observations on a more adaptive way than comparing $T(x)$ with a given threshold separating acceptance and rejection. The following notion was first introduced by Fisher (1956).

Definition 5.3.11 *The p-value associated with a test is the smallest significance level α for which the null hypothesis is rejected.*

A general definition for point-null hypotheses (see Thompson (1989)) considers that a p-value is any statistic with a uniform distribution under the null hypothesis, but it leads to the difficult problem of selecting one of these statistics, even though the same can be said about the *test* introduced in the above definition. Actually, if a test with rejection region R_α is available for every significance level α and if these regions are nested (that is, $R_\alpha \subset R_\beta$ if $\beta > \alpha$), the procedure

$$p(x) = \inf\{\alpha;\, x \in R_\alpha\}$$

is uniformly distributed if $\mathbb{E}_{\theta_0}[\mathbb{I}_{R_\alpha}(x)] = \alpha$ (see Goutis et al. (1996)). In the event of several contending tests, we suggest using the distribution of the likelihood ratio under the null hypothesis, if it is a point-null hypothesis.

Example 5.3.12 (Example 5.3.6 continued) Since the critical region (i.e., the rejection region for H_0) of the UMPU test is $\{|x| > k\}$, an usual p-value is

$$p(x) = \inf\{\alpha;\, |x| > k_\alpha\}$$

$$= P^X(|X| > |x|), \qquad X \sim \mathcal{N}(0,1)$$
$$= 1 - \Phi(|x|) + \Phi(|x|) = 2[1 - \Phi(|x|)].$$

Therefore, if $x = 1.68$, $p(x) = 0.10$ and, if $x = 1.96$, $p(x) = 0.05$. ||

Example 5.3.13 Consider $x \sim \mathcal{B}(n,p)$, when the hypothesis to be tested is $H_0:\ p = 1/2$ and $H_1:\ p \neq 1/2$. The p-value associated with the likelihood ratio

$$\frac{f(x|1/2)}{\sup_p f(x|p)} = \frac{(1/2)^n}{\left(\frac{x}{n}\right)^x \left(1 - \frac{x}{n}\right)^{n-x}} \propto x^{-x}(n-x)^{-(n-x)}$$

is the function

$$\tilde{p}(x) = P_{1/2}\left(X^X (n-X)^{(n-X)} \leq x^x (n-x)^{(n-x)}\right),$$

where $X \sim \mathcal{B}(n, 1/2)$. ||

P-values are thus adaptive procedures that can be acceptable from a frequentist point of view and that, furthermore, meet the requirements of Kiefer (1977) and Robinson (1979) for a *conditional frequentist approach.* However, they are still exposed to criticisms because

(i) p-values also evaluate the wrong quantity, namely, the probability of *overpassing* the observed value of the test statistic. They therefore contradict the Likelihood Principle by involving the whole distribution of the observation.

(ii) Even if derived from optimal test procedures, p-values have no intrinsic optimality because they are not evaluated under a loss function. In fact, as shown in Section 5.4, they may even be suboptimal.

(iii) The new decision space, $\mathcal{D} = [0,1]$, lacks a decision-theoretic foundation and thus the use of p-values is not made explicit. In particular, the p-values are often perceived as providing a frequentist approximation to $P(\theta \in \Theta_0|x)$, even though this expression is meaningless in a non-Bayesian setting.

(iv) From a classical perspective, p-values do not summarize all the information about the testing problem, since they should be compared with *type–two errors*, which are usually omitted from the analysis. Berger and Wolpert (1988) illustrate the danger of using only p-values by the following example. If $x \sim \mathcal{N}(\theta, 1/2)$, to test $\theta = -1$ versus $\theta = 1$ when $x = 0$ leads to an (UMP) p-value of 0.072, seemingly indicating a strong rejection of the null hypothesis, although the corresponding p-value for the test reversing H_0 and H_1 takes exactly the same value. In fact, while a rejection of H_0 should not always imply acceptance of H_1, practitioners often consider p-values as *the* testing procedure and they assume that it encompasses all the information about the testing problem, thus ending with this implication.

5.3.5 Least favorable Bayesian answers

The problem of evaluating p-values under an adapted loss is considered in Section 5.4. We conclude this section with a comparison of p-values with their Bayesian counterparts, the posterior probabilities. To consider the lower posterior probability on a class of prior distributions provides a Bayesian *least favorable answer* with respect to the null hypothesis. This lower bound cannot be considered as a noninformative procedure, since it enhances the prior most opposed to the null hypothesis, and is both biased against H_0 and dependent on the observation. It should be interpreted as an indicator of the *range* of the posterior probabilities, the most favorable answer being 1. An extensive literature is now available on this approach and we refer to Berger and Sellke (1987), Berger and Delampady (1987) and Berger and Mortera (1991) for additional references. Note 5.7.4 presents a different perspective of Berger, Boukai and Wang (1997) that reconciles frequentist and Bayesian testing by modifying the decision-theoretic framework.

Berger and Sellke (1987) and Berger and Delampady (1987) consider the case of a point-null hypothesis, $H_0: \ \theta = \theta_0$, against the alternative hypothesis $H_1: \ \theta \neq \theta_0$. For a family G of prior distributions on the alternative hypothesis, the evaluation measures of the veracity of H_0 are given by the lower bounds

$$\underline{B}(x, G) = \inf_{g \in G} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta) g(\theta)\, d\theta},$$

$$\underline{P}(x, G) = \inf_{g \in G} \frac{f(x|\theta_0)}{f(x|\theta_0) + \int_\Theta f(x|\theta) g(\theta)\, d\theta}$$

on the Bayes factors and posterior probabilities (for $\varrho_0 = 1/2$, considered to give equal weights to both hypotheses). These bounds can also be written as

$$\underline{B}(x, G) = \frac{f(x|\theta_0)}{\sup_{g\in G} \int_\Theta f(x|\theta) g(\theta) d\theta}, \quad \underline{P}(x, G) = \left[1 + \frac{1}{\underline{B}(x, G)}\right]^{-1}.$$

They obviously vary, depending on the class G considered. In the more general case, i.e., when G is G_A, the set of all prior distributions, the following result is straightforward.

Lemma 5.3.14 *If there exists a maximum likelihood estimator of θ, $\hat\theta(x)$, the lower bounds on the Bayes factors and posterior probabilities of H_0 are, respectively,*

$$\underline{B}(x, G_A) = \frac{f(x|\theta_0)}{f(x|\hat\theta(x))}, \quad \underline{P}(x, G_A) = \left[1 + \frac{f(x|\hat\theta(x))}{f(x|\theta_0)}\right]^{-1}.$$

A consequence of Lemma 5.3.14 is that the Bayesian answer will never

Table 5.3.1. *Comparison between p-values and Bayesian answers in the normal case. (Source: Berger and Sellke (1987).)*

p-value	0.10	0.05	0.01	0.001
$\underline{P}$	0.205	0.128	0.035	0.004
$\underline{B}$	0.256	0.146	0.036	0.004

strongly favor the null hypothesis, since

$$\underline{B}(x, G_A) \le 1, \qquad \underline{P}(x, G_A) \le \frac{1}{2}.$$

This behavior is not particularly surprising because the lower bounds correspond to the worst possible choice of g with respect to H_0. A more surprising phenomenon is that the decrease of these bounds when $|x|$ increases is much slower than for the p-values, as shown by the following example.

Example 5.3.15 (Example 5.3.6 continued) In the normal case, the lower bounds associated with $H_0: \theta_0 = 0$ are

$$\underline{B}(x, G_A) = e^{-x^2/2} \quad \text{and} \quad \underline{P}(x, G_A) = \left(1 + e^{x^2/2}\right)^{-1},$$

leading to Table 5.3.1, which compares the p-values with the Bayesian least favorable answers. ‖

Therefore, the difference with the frequentist answers is quite important. P-values are smaller for the significance levels of interest, thus reject the null hypothesis H_0 "too often." Of course, for smaller values of x, the p-values are larger than the lower bounds, but what matters is that, in the range of values of x where the decision is the most difficult to take, i.e., for significance levels between 0.01 and 0.1, there is such a discrepancy between the Bayesian and frequentist answers.

Results such as those above are quite surprising because classical procedures usually belong to the range of Bayesian answers. Moreover, the class G_A is rather unreasonable, including point masses leading to the lower bound. The only justification for this class of priors relates to the minimax principle and the corresponding least favorable distribution. The above example shows that p-values are not minimax in this sense. Obviously, the discrepancy is more important for smaller classes of distributions. For instance, if G is G_S, the set of distributions which are symmetric around θ_0, the equivalent of Lemma 5.3.14 is:

Lemma 5.3.16 *The smallest Bayes factor when $g \in G_S$ is*

$$\underline{B}(x, G_S) = \frac{f(x|\theta_0)}{\sup_\xi \frac{1}{2}[f(x|\theta_0 - \xi) + f(x|\theta_0 + \xi)]},$$

which leads to the corresponding lower bound on posterior probabilities.

on θ, i.e., subsets C_x of the parameter space Θ where θ should be with high probability (in the frequentist or in the Bayesian sense). This notion also extends to non-bijective transforms of θ. It is also of considerable interest in *forecasting* and prediction settings.

Example 5.5.1 Consider the IBM stock prices of Example 4.5.3, represented in Figure 4.5.1. If the series (x_t) has been observed till time T, the value at time $T+1$, x_{T+1}, is obviously of considerable interest and it is important to present the investor not only with the most likely value of x_{T+1} given the past series, but also with a range of possible values of x_{T+1} so that she can take a decision against the corresponding possible gains. $\|$

Once again, the Bayesian formulation that θ has a given probability to belong to a fixed region C_x is more appealing than the frequentist interpretation that a random region C_x has a given probability to contain the unknown parameter θ.

5.5.1 Credible intervals

As in the testing setting, the Bayesian paradigm proposes a notion of confidence regions that is more natural that its frequentist counterpart since, again, the notation $P(\theta \in C_x)$ is meaningful even conditional upon x.

Definition 5.5.2 *For a prior distribution π, a set C_x is said to be an α-credible set if*

$$P^\pi(\theta \in C_x|x) \geq 1-\alpha.$$

This region is called an HPD α-credible region (for highest posterior density) if it can be written under the form[10]

$$\{\theta;\ \pi(\theta|x) > k_\alpha\} \subset C_x^\pi \subset \{\theta;\ \pi(\theta|x) \geq k_\alpha\},$$

where k_α is the largest bound such that

$$P^\pi(\theta \in C_x^\alpha|x) \geq 1-\alpha.$$

To consider only HPD regions is motivated by the fact that they *minimize the volume among α-credible regions* and, therefore, can be envisioned as optimal solutions in a decision setting.

Example 5.5.3 If $\theta \sim \mathcal{N}(0,\tau^2)$, the posterior distribution of θ is $\mathcal{N}(\mu(x), \omega^{-2})$ with $\omega^2 = \tau^{-2}+\sigma^{-2}$ and $\mu(x) = \tau^2 x/(\tau^2+\sigma^2)$. Then

$$C_\alpha^\pi = \left[\mu(x) - k_\alpha \omega^{-1}, \mu(x) + k_\alpha\omega^{-1}\right],$$

where k_α is the $\alpha/2$-quantile of $\mathcal{N}(0,1)$. In particular, if τ goes to $+\infty$, $\pi(\theta)$ converges to the Lebesgue measure on $\mathbb{R}$ and gives

$$C_\alpha = [x - k_\alpha\sigma, x + k_\alpha\sigma],$$

[10] This formulation allows for coverage of the special case when $\{\theta; \pi(\theta|x) = k_\alpha\}$ is not empty.

Table 5.5.1. *Confidence intervals for the binomial distribution.*

x	0	1	2
$\alpha = 5\%$	$[0.000, 0.38]$	$[0.022, 0.621]$	$[0.094, 0.791]$
$\alpha = 10\%$	$[0.000, 0.308]$	$[0.036, 0.523]$	$[0.128, 0.74]$

i.e., the usual confidence interval, as a generalized Bayes estimator. $\|$

Example 5.5.4 Consider $x \sim B(n,p)$ and the noninformative distribution $p \sim \mathcal{B}e(1/2,1/2)$. Then $p|x \sim \mathcal{B}e(x+1/2, n-x+1/2)$ and confidence intervals on p can be derived from the c.d.f. of the beta distribution. Table 5.5.1 gives these intervals for $n=5$ and $\alpha = 5\%$, 10%. $\|$

Notice the significant advantage of using a Bayesian approach in this setting of *discrete distributions*, as compared with a classical approach. In fact, the usual confidence intervals involve a *randomization* step to attain nominal confidence levels (see Blyth and Hutchinson (1961) for an illustration in the binomial case). Prior modeling avoids this addition of random noise and, on the contrary, takes advantage of the available prior information. Notice also that *improper priors* can be used in this setting, and do not encounter the same difficulties as when testing point-null hypotheses. In fact, posterior credible regions can be derived provided the posterior distribution is defined. Some classical confidence regions can be expressed as credible regions associated with generalized distributions.

Example 5.5.5 Consider $x_1,\ldots,x_n$ i.i.d. $\mathcal{N}(\theta,\sigma^2)$. The prior distribution is the noninformative prior

$$\pi(\theta,\sigma^2) = \frac{1}{\sigma^2}.$$

We showed in Section 4.4.2 that the marginal posterior distribution for $1/\sigma^2$ is a gamma distribution $\mathcal{G}\left((n-1)/2, s^2/2\right)$ with $s^2 = \sum(x_i-\bar{x})^2$. Therefore

$$\frac{s^2}{\sigma^2}|\bar{x}, s^2 \sim \chi^2_{n-1},$$

and we get the same confidence interval as in the classical approach, but it is now justified conditional upon s^2. $\|$

Example 5.5.6 Consider $x \sim \mathcal{B}(n,p)$ and $p \sim \mathcal{B}e(\alpha,\beta)$. In this case, $\pi(p|x)$ is the beta distribution $\mathcal{B}e(\alpha+x, \beta+n-x)$. Depending on the values of α, β, n, and x, the confidence regions are of four types:

(i) $0 \leq p \leq K(x)$;

(ii) $K(x) \leq p \leq 1$;

(iii) $K_1(x) \le p \le K_2(x)$; and

(iv) $0 \le p \le K_1(x)$ or $K_2(x) \le p \le 1$.

The last region is quite artificial and rather useless. Notice that it corresponds to the case

$$\alpha + x < 1 \quad \text{and} \quad \beta + n - x < 1,$$

thus implying that α and β must be sufficiently negative since $\alpha+\beta < 2-n$. Hence, this feature disappears for n large enough, unless α and β depend on n, which is not desirable from a Bayesian point of view. Notice that the limiting case $\alpha = \beta = 0$, which corresponds to Haldane's (1931) distribution

$$\pi(p) = [p(1-p)]^{-1},$$

already leads to regions of types (i)–(iii), although the posterior distribution is not defined for all x's (Example 1.5.3). ||

When phenomena such as case (iv) of Example 5.5.6 occur, that is, when the confidence region is not connected (see also Exercise 5.5), the usual solution is to replace the HPD α-credible region by an interval with equal tails, i.e., $[C_1(x), C_2(x)]$ such that

$$P^\pi(\theta < C_1(x)|x) = P^\pi(\theta > C_2(x)|x) = \alpha/2.$$

Berger (1985a) remarks that the occurrence of nonconnected HPD regions also points out a discrepancy between the prior distribution and the observations and that this phenomenon should question the choice of the prior or of the sampling distribution. It may also exhibit a non-identifiabile structure which is responsible for the multimodality of the posterior distribution.

If, conceptually, the determination of credible sets is rather straightforward, the practical derivation of these regions can be quite involved, especially when the dimension of Θ is large or when the posterior distribution is not available in a closed form. A first solution is to use numerical methods similar to those developed in Chapter 6, the problem being to assess the impending error (which can be much larger than the point estimation approximation errors). (Notice that equal tail credible regions are usually easier to approximate than HPD regions; see Eberly and Casella (1999).) A second solution, used in Berger (1980a, 1985a), is to build up a normal approximation, i.e., to consider that the posterior distribution of θ is roughly $\mathcal{N}_p(\mathbb{E}^\pi(\theta|x), \mathrm{Var}^\pi(\theta|x))$, and to derive from this approximation the confidence region

$$C_\alpha = \left\{\theta; (\theta - \mathbb{E}^\pi(\theta \mid x))^t \, \mathrm{Var}^\pi(\theta|x)^{-1} (\theta - \mathbb{E}^\pi(\theta|x)) \le k_\alpha^2 \right\},$$

where k_α^2 is the α-quantile of χ_p^2. This approximation is only justified for a large sample size (see Hartigan (1983)), but it still provides fast and rather efficient confidence regions.

(see also Exercise 5.54). Such phenomena led Kiefer (1977) to suggest a partition of the sample space $\mathcal{X}$ into subsets, and to allocate each of these subsets with a different confidence level (see also Brown (1978)). Following Fisher's analysis, he suggested that these subsets should be indexed by ancillary statistics. For instance, the adequate ancillary statistic for Example 2.3.1 is $x_1 - x_2$. Unfortunately, in most settings, the choice of the ancillary statistic modifies the confidence report, and Berger and Wolpert (1988) give an example in which different ancillary statistics lead to different confidence statements, a setting incompatible with the Likelihood Principle. We consider that, fundamentally, the problem exhibited by the existence of relevant biased sets is not related to the confidence region C_x itself but rather with the associated confidence level α, which should be replaced by a more adaptive (or more conditional) confidence statement $\alpha(x)$ (see Section 4.2). In fact, the existence of relevant betting procedures is equivalent to the domination of the constant confidence report under quadratic loss (Robinson (1979)).

5.7.4 Reconciliation perspective

While Section 5.3 has shown that the frequentist answers, that is, the p-values, were intrinsically *and* numerically different from their Bayesian counterparts (see also Note 5.7.1), a modification of the decision setting, brought up by Berger, Brown and Wolpert (1994), allows for a partial reconciliation of both perspectives. While, from a Bayesian point of view, reconciliation is not an important feature—the basic purpose of a procedure is to act in an optimal way when faced with a decision problem, rather than to enjoy a long run stability—it offers several advantages from a practical point of view: firstly, it makes statisticians more comfortable with using a Bayesian procedure if this procedure simultaneously enjoys frequentist properties. Secondly, it eliminates the difficulty with the interpretation of a p-value as a posterior probability.

The modification is to introduce a "no-decision" alternative to the "accept" and "reject" answers used in classical tests. Even though this additional possibility may seem ludicrous from a decision-making point of view, it certainly makes sense from a statistical point of view: there are indeed cases where the data is too unconclusive to answer satisfactorily about a given hypothesis H_0 and when we want to ask the client for more observations, or for more precise prior information. In fact, this perspective has been around for sequential testing since Wald's *sequential probability ratio test* (see Lehmann (1986)). (Notice, however, that the procedure of Berger et al. (1994) does not account for repeated testing, which affects the acceptance levels. See also Example 1.3.6.)

In the case of two simple hypotheses,

$$H_0: \; x \sim m_0(x) \quad \text{versus} \quad H_1: \; x \sim m_1(x),$$

where m_0 and m_1 are known densities, the Bayes factor B_{10} is equal to the likelihood ratio $m_1(x)/m_0(x)$. If -1 denotes the no-decision answer, the modified Bayesian test of Berger et al. (1994) is of the form

$$\varphi(x) = \begin{cases} 1 & \text{if } B_{10}(x) \ge a, \\ 0 & \text{if } B_{10}(x) \le r, \\ -1 & \text{if } r < B_{10}(x) < a, \end{cases} \tag{5.7.1}$$

with an associated probability error report given by

$$\gamma(x) = \begin{cases} 1/(1+B_{10}(x)) & \text{if } B_{10}(x) \geq a, \\ B_{10}(x)/(1+B_{10}(x)) & \text{if } B_{10}(x) \leq r. \end{cases}$$

Notice that $\gamma(x)$ is the posterior probability of the rejected hypothesis and is thus optimal under squared error loss. (But the procedure φ is arguably not a decision procedure. See Exercise 5.58.)

Now, if we denote F_0 and F_1 the c.d.f.'s of $B_{10}(x)$ associated with m_0 and m_1 respectively, and if we define $\Psi(b) = F_0^{-1}(1 - F_1(b))$, then $\Psi^{-1}(b) = F_1^{-1}(1 - F_0(b))$, Berger et al. (1994) take

$$\begin{aligned} r = 1 \text{ and } a = \Psi(1) \qquad & \text{if } \Psi(1) > 1 \\ r = \Psi^{-1}(1) \text{ and } a = 1 \qquad & \text{if } \Psi(1) < 1\,. \end{aligned}$$

They show that the report $\gamma(x)$ is then correct from a conditional frequentist perspective: conditional on

$$S(x) = \min\{B_{10}(x), \Psi^{-1}(B_{10}(x))\}\,, \tag{5.7.2}$$

the procedure (φ, γ) satisfies

$$P_0(B_{10}(x) \geq a | S(x) = s) = \gamma(s)\,, \qquad P_1(B_{10}(x) \leq r | S(x) = s) = \gamma(s)\,,$$

where $\gamma(x)$ only depends on s (Exercise 5.56). Notice however that $S(x)$ is not an ancillary statistic, except in special cases (Exercise 5.59).

The extension of this result to composite hypotheses,

$$H_0 : \ \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \ \theta \in \Theta_1$$

follows by representing H_1 as in Section 5.3.5, that is, with

$$H_1 : \ x \sim m_1(x) = \int_{\Theta_1} f(x|\theta)\pi_1(\theta) d\theta\,.$$

Then Berger, Boukai and Wang (1997) show that the conditional frequentist evaluation under H_0 still coincides with the Bayesian report, in a weaker sense that, while being satisfactory from a Bayesian perspective, is less compelling from a frequentist perspective (Hinkley (1997), Louis (1997)).

Example 5.7.4 (Berger et al. (1997)) For $x_1, \ldots, x_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$, with σ known, consider the test of $H_0 : \ \theta = \theta_0$ under the conjugate prior $\theta \sim \mathcal{N}(\mu, k\sigma^2)$. Then, if $z = \sqrt{n}(\bar{x}_n - \theta_0)/\sigma$,

$$m_0(z) = \frac{1}{\sqrt{2\pi}} \exp\{-z^2/2\}$$

and

$$m_1(z) = \frac{1}{\sqrt{2\pi}\sqrt{1+kn}} \exp\left\{\frac{-(z+\sqrt{kn}\Delta)^2}{2(1+kn)}\right\}\,,$$

where $\Delta = (\theta_0 - \mu)/\sqrt{k}\sigma$. Then the Bayes factor is given by

$$B_{10}(x) = \sqrt{1+kn} \exp\left\{-\frac{kn}{2(1+kn)}\left[z - \frac{\Delta}{\sqrt{kn}}\right]^2 + \frac{\Delta^2}{2}\right\}\,,$$

$\Psi(1) > 1$, $r = 1$ and $a = F_0^{-1}(1 - F_1(1))$. ‖