Subjective and Objective Bayesian Statistics: Principles, Models, and Applications
S. james Press

CHAPTER 9

Bayesian Hypothesis Testing

9.1 INTRODUCTION

In this chapter we discuss Bayesian hypothesis testing. We begin with some historical background regarding how hypothesis testing has been treated in science in the past, and show how the Bayesian approach to the subject has really provided the statistical basis for its development. We then discuss some of the problems that have plagued frequentist methods of hypothesis testing during the twentieth century. We will treat two Bayesian approaches to the subject:

1. The vague prior approach of Lindley (which is somewhat limited but easy to implement); and
2. The very general approach of Jeffreys, which is the current, commonly accepted Bayesian method of hypothesis testing, although it is somewhat more complicated to carry out.

9.2 A BRIEF HISTORY OF SCIENTIFIC HYPOTHESIS TESTING

There is considerable evidence of ad hoc tests of hypotheses that were developed to serve particular applications (especially in astronomy), as science has developed. But there had been no underlying theory that could serve as the basis for generating appropriate tests in general until Bayes' theorem was expounded. Moreover, researchers had difficulty applying the theorem even when they wanted to use it.

Karl Pearson (1892), initiated the development of a formal theory of hypothesis testing with his development of chi-squared testing for multinomial proportions. He liked the idea of applying Bayes' theorem to test hypotheses, but he could not quite figure out how to generate prior distributions to support the Bayesian approach. Moreover, he did not recognize that consideration of one or more alternative hypotheses might be relevant for testing a basic scientific hypothesis.

"Student" (William Sealy Gosset, 1908) in developing his t-test for the mean of a normal distribution, and for his work with the sample correlation coefficient, claimed that he would have preferred to use Bayes' theorem (he referred to it as "inverse probability"), but he did not know how to set his prior distribution.

Fisher (1925) developed a formal theory of hypothesis testing that would serve for a variety of scientific situations (although Fisher, like Karl Pearson, also did not consider alternative hypotheses; that modification would wait for Neyman and Pearson, 1933). Fisher attempted to develop an approach that would be "objective" in some sense, and would compare the actual observed data with how data might look if they were generated randomly. Fisher's approach to scientific hypothesis testing was totally non-Bayesian; it was based upon the thinking of his time that was dominated by the influential twentieth-century philosopher, Karl Popper, (1935, 1959). Popper advocated a theory of "falsification" or "refutability" to test scientific theories. As Popper saw it, a scientific theory should be tested by examining evidence that could, in principle, refute, disconfirm, or falsify the theory. Popper's idea for testing a scientific theory was for the scientist to set up a *strawman hypothesis* (a hypothesis opposite to what the scientist truly believes, but under consideration to see if it can be destroyed), and then show that the strawman hypothesis is indeed false. Then the theory based upon the strawman hypothesis could be discarded. Otherwise, one had to wait for additional evidence before proceeding to accept the strawman hypothesis. Fisher adopted this falsification/strawman position.

For example, Popper suggests that a scientist might believe that he/she has a strong theory about why some phenomenon takes place. The scientist might set up a hypothesis that implies that the phenomenon takes place, say, at random. The hypothesis of randomness is then tested (that's the strawman hypothesis). The scientist may then find that empirical data do not support the randomness hypothesis. So it must be rejected. But the scientist's real hypothesis that he/she believes in cannot yet be accepted as correct, if that is the alternative hypothesis. It will require more testing before it can be accepted.

This process was formalized by Fisher suggesting beginning with a null hypothesis, a hypothesis that the researcher believes *a priori* to be false, and then carrying out an experiment that will generate data that will show the null hypothesis to be false. Fisher proposed a "test of significance" in which if an appropriate test statistic exceeds some special calculated value, based upon a pre-assigned significance level, the null hypothesis is rejected. But if it turns out that the null hypothesis cannot be rejected, no conclusion is drawn. For Fisher, there was no criterion for accepting a hypothesis. In science, we never know when a theory is true in some sense; we can only show that a theory may be false because we can find contradictions or inconsistencies in its implications. Fisher also suggested that one could alternatively compute a "p-value," that is, the probability of observing the actually observed value of the test statistic, or anything more extreme, assuming the null hypothesis is true. Some frequentists also think of the p-value as a "sample significance level." (We will discuss below how p-values relate to posterior probabilities.)

To the extent that scientists should reject a hypothesis (theory) if experimental data suggest that the alternative hypothesis (theory) is more probable, this is very sensible; it is the way science proceeds. So Fisher appropriately suggested that we not accept the alternative hypothesis when we reject the null hypothesis. We should just postpone a decision until we have better information. We will see that in Bayesian hypothesis testing, we compute the weight of the experimental evidence as measured by how probable it makes the main hypothesis relative to the alternative hypothesis.

The Popper/Fisher approach to hypothesis testing was not probability based. That is, probabilities were not placed on the hypothesis being tested. For Fisher, one could not place a probability on an hypothesis because an hypothesis is not a random variable in the frequentist sense. For a Bayesian, however, there is no problem placing a probability on an hypothesis. The truthfulness of the hypothesis is unknown, so a researcher can put his/her subjective probability on it to express his/her degree of uncertainty about its truthfulness. We will see how this is done in the Jeffreys approach to Bayesian hypothesis testing.

Jerzy Neyman and Egon Pearson, in a series of papers starting in 1928 developed a theory of hypothesis testing that modified and extended Fisher's ideas in various ways (see, for example, Neyman and Pearson, 1966, where these papers are collected). Neyman and Pearson introduced the idea of alternative hypotheses. In addition, they introduced the notions of Type One and Type Two errors that could be made in testing statistical hypotheses. They defined the concept of "power of a test," and proposed that the ratio of the likelihood of the null hypothesis to the likelihood of the alternative hypothesis be used to compare a simple null hypothesis against a simple alternative hypothesis (Neyman–Pearson Lemma). But the theory was all still embedded in the frequentist falsification notions of Popper and Fisher.

Wald (1939) proposed a theory of decision making that incorporated statistical inference problems of hypothesis testing. We will be discussing decision theory in Chapter 11. This theory suggested, as did the ideas of Neyman and Pearson, that hypotheses should be tested on the basis of their consequences. Bayesian methodology shaped the development of this approach to hypothesis testing in that it was found that Bayesian-derived decision making procedures are the ones that generate the very best decision rules (see Chapter 11). When it was found to be difficult to assign prior distributions to develop the optimal procedures, given the limited development of prior distribution theory at that time, alternative *minimax* (and other) procedures were developed to generate useful decision rules. But the Bayes procedures, that is, procedures found by minimizing the expected loss, or maximizing the expected utility, of a decision rule, were seen to be the best decision rules that could be found (in contexts in which the decision maker had to make decisions in the face of uncertain outcomes of experiments that were decided by nature (this was shown in Savage, 1954). In contexts (games) in which decisions were to be made with respect to other human decision makers, minimax, or some other non-expected-loss criterion might be appropriate). There is additional discussion in Chapter 11.

Lehmann (1959) summarized the Fisher/Neyman–Pearson/Wald frequentist methods of hypothesis testing procedures in his book on hypothesis testing.

Throughout, he used the long-run frequency interpretation of probability rather than the Bayesian or subjective probability notion.

In keeping with Bayesian thinking, Good (1950, 1965) proposed and further codified the testing process, by suggesting that to compare scientific theories, scientists should examine *the weight of the evidence* favoring each of them, and he has shown that this concept is well defined in terms of a conditioning on prior information. Good (1983) developed these ideas further.

This was the state of statistical and scientific hypothesis testing until Jeffreys (1961) and Lindley (1965) proposed their Bayesian approaches that are outlined in Sections 9.4 and 9.5.

9.3 PROBLEMS WITH FREQUENTIST METHODS OF HYPOTHESIS TESTING

There are a variety of problems, difficulties, and inconsistencies associated with frequentist methods of testing hypotheses that are overcome by using Bayesian methods. Some of these problems are enumerated below.

1. *Bayesians have infrequent need to test.* To begin, hypothesis testing per se is something Bayesian scientists do only infrequently. The reason is that once the posterior distribution is found, it contains all the information usually required about an unknown quantity. The posterior distribution can and should be used to learn, and to modify or update earlier held beliefs and judgments. We do not normally need to go beyond the posterior distribution. In some situations, however, such as where the researcher is attempting to decide whether some empirical data conform to a particular theory, or when the researcher is trying to distinguish among two or more theories that might reasonably explain some empirical data, the Bayesian researcher does need to test among several hypotheses or theories. We also need to go beyond the posterior distribution in experimental design situations (see, for example, Raiffa and Schlaifer, 1961, who discuss *the value of sample information*, and *preposterior analysis*).
2. *Problems with probabilities on hypotheses.* The frequentist approach to hypothesis testing does not permit researchers to place probabilities of being correct on the competing hypotheses. This is because of the limitations on mathematical probabilities used by frequentists. For the frequentist, probabilities can only be defined for random variables, and hypotheses are not random variables (they are not observable). But (Bayesian) subjective probability is defined for all unknowns, and the truthfulness of the hypotheses is unknown. This limitation for frequentists is a real drawback because the applied researcher would really like to be able to place a degree of belief on the hypothesis. He or she would like to see how the weight of evidence modifies his/her degree of belief (probability) of the hypothesis being true. It is subjective probabilities of the competing hypotheses being true that are

compared in the *subjective* Bayesian approach. Objective Bayesians, as well as frequentists, have problems in hypothesis testing with odds ratios (as with Bayes' factors) because odds ratios are not defined when improper prior probabilities are used in both the numerator and denominator in the odds ratio.

3. *Problems with preassigned significance levels.* Frequentist methods of hypothesis testing require that a level of significance of the test (such as 5 percent) be preassigned. But that level of significance is quite arbitrary. It could just as easily be less than 1, 2, 4, 6 percent, etc. Where should the line be drawn and still have the result be significant statistically for a frequentist? The concept is not well defined. In the Bayesian approach we completely obviate the necessity of assigning such arbitrary levels of significance. If the weight of the evidence favors one hypothesis over another, that is all we need to know to decide in favor of that hypothesis.

4. *Inadequacy of frequentist testing of a sharp null hypothesis.* Suppose we wish to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, for some known θ_0, and we decide to base our test on a statistic $T \equiv T(X_1, \ldots, X_n)$. It is usually the case that in addition to θ being unknown, we usually cannot be certain that $\theta = \theta_0$ precisely, even if it may be close to θ_0. (In fact, in the usual frequentist approach, we often start out believing $\theta \neq \theta_0$, which corresponds to some intervention having had an effect, but we test a null hypothesis H_0 that $\theta = \theta_0$; that is, we start out by disbelieving $\theta = \theta_0$, and then we test it.) Suppose θ is actually ε away from θ_0, for some $\varepsilon > 0$, and ε is very small. Then, by *consistency* of the testing procedure, for sufficiently large n we will reject H_0 with probability equal to one. So, depending upon whether we want to reject H_0, or whether we want to find that we cannot reject H_0, we can choose n accordingly. This is a very unsatisfactory situation. (The same argument applies to all significance testing.)

5. *Frequentist use of possible values never observed.* Jeffreys (1961, p. 385) points out that frequentist hypothesis testers have to rely upon values of observables never observed. He says:

> What the use of the (p-value) implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.

Jeffreys is reacting to the fact that the frequentist divides the sample space into a critical region for which the null hypothesis will be rejected if the test statistic fall into it, and a complementary region for which the null hypothesis will not be rejected if the test statistic falls into it. But these regions contain values of possible test statistics never actually observed. So the test depends upon values that are observable, but have never actually been observed. Tests that depend upon values that have never actually been observed violate the likelihood principle (see Chapter 3).

6. *Problems with p-values.* The p-value is generally the value reported by researchers for the statistical significance of an experiment they carried out. If p is very low ($p \leq 0.05$), the result found in the experiment is considered

statistically significant by frequentist standards. But the p-value depends upon the sample size used in the experiment. By taking a sufficiently large sample size we can generally achieve a small p-value.

Berger and Selke (1987) and Casella and Berger (1987) compared p-values with the posterior distribution for the same problem. They found for that problem that the evidence against the null hypothesis based upon the posterior distribution is generally weaker than that reflected by the p-value. That is, the p-value suggests rejecting the null hypothesis more often than the Bayesian approach would suggest. In this sense, the Bayesian hypothesis test is more conservative than is the frequentist hypothesis test. The example used by Berger and Selke (1987) is presented in the following.

Suppose the probability density function for an observable X is given by $f(x \mid \theta)$. We are interested in testing the sharp null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis that $H_1 : \theta \neq \theta_0$. Let $T(X)$ denote an appropriate test statistic. Denote the p-value:

$$p \equiv P\{T(X) \geq t \mid \theta = \theta_0\}. \tag{9.1}$$

While the result will be general, for concreteness, we will make the example very specific. Suppose we have a sample of independent and identically distributed data $X = (X_1, \ldots, X_n)$, and $(X_i \mid \theta) \sim N(\theta, \sigma_0^2)$, for known $\sigma_0^2, i = 1, \ldots, n$. The usual test statistic (sufficient) in this problem is:

$$G(X) = \frac{\sqrt{n \mid \bar{X} - \theta_0 \mid}}{\sigma_0}. \tag{9.2}$$

Define

$$g \equiv G(x) = \frac{\sqrt{n \mid \bar{x} - \theta_0 \mid}}{\sigma_0}. \tag{9.3}$$

The p-value becomes:

$$\begin{aligned} p &= P\{\mid G(X) \mid > g\} \\ &= P\{[G(X) < -g] \bigcup [G(X) > g]\} \\ &= \Phi(-g) + [1 - \Phi(g)] = 2[1 - \Phi(g)]. \end{aligned} \tag{9.4}$$

$\Phi(g)$ denotes the cdf of the standard normal distribution. Suppose, in the interest of fairness, a Bayesian scientist assigns 50 percent prior probability to H_0 and 50 percent prior probability to H_1, but he/she spreads the mass on H_1 out according to

$N(\theta_0, \sigma_0^2)$. The posterior probability on the null hypothesis is shown below to be given by:

$$P\{H_0 \mid x\} = \frac{1}{1 + \frac{1}{\sqrt{n+1}} \exp\left\{\frac{g^2}{2(1+1/n)}\right\}}. \tag{9.5}$$

Proof
By Bayes' theorem,

$$P\{H_0 \mid x\} = \frac{f(x \mid \theta_0)\pi_0}{m(x)},$$

where

$$m(x) = f(x \mid \theta_0)\pi_0 + (1 - \pi_0)m_g(x),$$

and

$$m_g(x) = \int f(x \mid \theta) g(\theta)\, d\theta, \quad g(0) \sim N(\theta_0, \sigma_0^2).$$

π_0 denotes the prior probability of H_0. Equivalently,

$$P\{H_0 \mid x\} = \frac{1}{1 + \left(\frac{1-\pi_0}{\pi_0}\right)\frac{m_g(x)}{f(x \mid \theta_0)}}.$$

Note that since

$$(\bar{X} \mid \theta) \sim N(\theta, \sigma_0^2/n), \quad \theta \sim N(\theta_0, \sigma_0^2),$$

marginally,

$$\bar{X} \sim N\left[\theta_0, \left(1 + \frac{1}{n}\right)\sigma_0^2\right].$$

Simplifying gives the result in equation (9.5).

We provide the values of the posterior probability $P\{H_0 \mid x\}$ in Table 9.1, from which may be seen that, for example, for $n = 50$, the frequentist researcher could reject H_0 at $p = 0.050$ (5 percent), since $g = 1.96$, whereas the posterior probability $P\{H_0 \mid x\} = 0.52$, so actually, H_0 is favored over the alternative. At $n = 50$, the frequentist approach to hypothesis testing suggests that the null hypothesis be rejected at the 5 percent level of significance, whereas from a Bayesian point of view, for $n = 50$ or more, the posterior probability says that we should not reject the

Table 9.1 Values of the Posterior Probability $P\{H_0 \mid x\}$

p-value	g	$n = 1$	$n = 5$	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 1000$
0.100	1.645	0.42	0.44	0.47	0.56	0.65	0.72	0.89
0.050	1.960	0.35	0.33	0.37	0.42	0.52	0.60	0.82
0.010	2.576	0.21	0.13	0.14	0.16	0.22	0.27	0.53
0.001	3.291	0.086	0.026	0.024	0.026	0.034	0.045	0.124

null hypothesis. So for the Bayesian hypothesis tester the evidence against the null hypothesis is weaker.

9.4 LINDLEY'S VAGUE PRIOR PROCEDURE FOR BAYESIAN HYPOTHESIS TESTING

A procedure for testing a hypothesis H_0 against an alternative hypothesis H_1 from a Bayesian point of view was suggested by Lindley, 1965, Vol. 2, p. 65. The test procedure is readily understood through a simple example.

Suppose we have the independent and identically distributed data $X_1, \ldots, X_n$, and $(X_i \mid \theta) \sim N(\theta, 1), i = 1, \ldots, n$. We wish to test $H_0 : \theta = \theta_0$, versus $H_1 : \theta \neq \theta_0$. We first recognize that $\bar{X}$ is sufficient for θ, and that $(\bar{X} \mid \theta) \sim N(\theta, 1/n)$. Adopt the vague prior density for θ, $g(\theta) \propto$ constant. The result is that the posterior density for θ is given by $(\theta \mid \bar{x}) \sim N(\bar{x}, 1/n)$. Next, develop a credibility interval for θ at level of credibility α, where say $\alpha = 5$ percent. This result is:

$$P\left\{\bar{x} - \frac{1.96}{\sqrt{n}} \leq \theta \leq \bar{x} + \frac{1.96}{\sqrt{n}} \mid \bar{x}\right\} = 95 \text{ percent.} \tag{9.6}$$

Now examine whether the interval includes the null hypothesis value $\theta = \theta_0$. If θ_0 is not included within this 95 percent credibility interval, this is considered evidence against the null hypothesis, and H_0 is rejected. Alternatively, if θ_0 is found to lie within the 95 percent credibility interval, we cannot reject the null hypothesis.

Lindley (1965) actually proposed this frequentist-like Bayesian testing procedure by using a frequentist confidence interval approach instead of the credibility interval approach outlined here. But under a vague prior density for θ the approaches are equivalent.

We note here that these ideas result from attempts to avoid totally abandoning the frequentist notions of Popper/Fisher/Neyman–Pearson. It is still required that: We preassign a level of significance for the test; we still adopt a strawman null hypothesis; and we still do not place probabilities on the competing hypotheses. The procedure also requires that we adopt a vague prior density for the unknown, θ (and for the credibility interval to be sensible, the prior density must be smooth in the vicinity of θ_0). However, suppose we have meaningful prior information about θ,

and we would like to bring that information to bear on the problem. The procedure does not afford us any way to bring the information into the problem. Even worse, if the prior information is mixed continuous and discrete, the testing approach will not be applicable. Regardless, for simple situations where vague prior densities are sensible, the Lindley hypothesis testing procedure provides a rapid, and easily applied, testing method. A more general hypothesis testing procedure is found in the Jeffreys (1961) approach described in Section 9.5.

9.4.1 The Lindley Paradox

The Bayesian approach to hypothesis testing when little prior information is available received substantial interest when Lindley (1957) called attention to the paradoxical result that a frequentist scientist could strongly reject a sharp (null) hypothesis H_0, while a Bayesian scientist could put a lump of prior probability on H_0 and then spread the remaining prior probability out over all other values in a "vague" way (uniformly) and find there are high posterior odds in favor of H_0. This paradox is equivalent to the discussion surrounding Table 9.1.

9.5 JEFFREYS' PROCEDURE FOR BAYESIAN HYPOTHESIS TESTING

Jeffreys (1961, Chapters 5 and 6) suggested a totally Bayesian approach to hypothesis testing that circumvents the inadequacies of the frequentist test procedures. This procedure is outlined below.

9.5.1 Testing a Simple Null Hypothesis Against a Simple Alternative Hypothesis

First we consider the case of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$, where, θ_0 and θ_1 are preassigned constants (recall that a *simple* hypothesis is one for which there is only one possible value for the unknown). We assume that H_0 and H_1 are mutually exclusive and exhaustive hypotheses. Let $T \equiv T(X_1, \ldots, X_n)$ denote an appropriate test statistic based upon a sample of n observations. Then, by Bayes' theorem, the posterior probability of H_0, given the observed data T, is

$$P\{H_0 \mid T\} = \frac{P\{T \mid H_0\}P\{H_0\}}{P\{T \mid H_0\}P\{H_0\} + P\{T \mid H_1\}P\{H_1\}}, \tag{9.7}$$

where $P\{H_0\}$ and $P\{H_1\}$ denote the researcher's prior probabilities of H_0 and H_1. Similarly, for hypothesis H_1, we have:

$$P\{H_1 \mid T\} = \frac{P\{T \mid H_1\}P\{H_1\}}{P\{T \mid H_0\}P\{H_0\} + P\{T \mid H_1\}P\{H_1\}} \tag{9.8}$$

Note that $P\{H_0 \mid T\} + P\{H_1 \mid T\} = 1$. Equations (9.7) and (9.8) can be combined to form the ratio:

$$\frac{P\{H_0 \mid T\}}{P\{H_1 \mid T\}} = \left[\frac{P\{H_0\}}{P\{H_1\}}\right]\left[\frac{P\{T \mid H_0\}}{P\{T \mid H_1\}}\right]. \tag{9.9}$$

Recall that if two probabilities sum to one, their ratio is called the *odds* in favor of the event whose probability is in the numerator of the ratio. Therefore, Equation (9.9) may be interpreted to state that the posterior odds ratio in favor of H_0 is equal to the product of the prior odds ratio in favor of H_0 and the likelihood ratio.

Jeffreys' Hypothesis Testing Criterion

The Jeffreys' criterion for hypothesis testing becomes, in a natural way:

> If the posterior odds ratio exceeds unity, we accept H_0; otherwise, we reject H_0 in favor of H_1.

It is not necessary to specify any particular level of significance. We merely accept or reject the null hypothesis on the basis of which posterior probability is greater; equivalently, we accept or reject the null hypothesis on the basis of whether the posterior odds ratio is greater or less than one. (If the posterior odds ratio is precisely equal to one, no decision can be made without additional data or additional prior information.) Note that if there were several possible hypotheses, this approach would extend in a natural way; we would find the hypothesis with the largest posterior probability.

REMARK: When we accept the null hypothesis because the weight of the evidence shows the null hypothesis to be favored by the data over the alternative hypothesis, we should recognize that we are merely doing so on the basis that the null hypothesis is the one to be entertained until we have better information or a modified theory. We are not assuming that the null hypothesis is true, merely that with the present state of knowledge, the null hypothesis is more probable than the alternative hypothesis.

Bayes' Factors

We note from equation (9.9) that the ratio of the posterior odds ratio to the prior odds ratio, called the *Bayes' factor*, is a factor that depends only upon the sample data. The Bayes' factor reflects the extent to which the data themselves (without prior information) favor one model over another. In the case of testing a simple null hypothesis against a simple alternative hypothesis the Bayes' factor is just the likelihood ratio, which is also the frequentist test statistic for comparing two simple hypotheses (the result of the Neyman—Pearson Lemma). It becomes a bit more complicated (instead of just the simple likelihood ratio) in the case of a simple null hypothesis versus a composite alternative hypothesis. Because the prior odds ratio is often taken to be one by the objective Bayesian, the Bayes' factor acts as an objectivist Bayesian's answer to how to compare models. The subjectivist Bayesian

scientist needs only the posterior odds ratio to compare models (which may differ from the Bayes' factor depending upon the value of the prior odds ratio).

As an example, suppose $(X \mid \theta) \sim N(\theta, 1)$, and we are interested in testing whether $H_0 : \theta = 0$, versus $H_1 : \theta = 1$, and these are the only two possibilities. We take a random sample $X_1, \ldots, X_N$ and form the sufficient statistic $T = \bar{X} = (1/N)\sum_{j=1}^{N} X_j$. We note that $(T \mid H_0) \sim N(0, 1/N)$, and $(T \mid H_1) \sim N(1, 1/N)$. Assume that, *a priori*, $P\{H_0\} = P\{H_1\} = 0.5$. Then, the posterior odds ratio is given by

$$\frac{P\{H_0 \mid T\}}{P\{H_1 \mid T\}} = \left(\frac{0.5}{0.5}\right)\left[\frac{\left(\frac{N}{2\pi}\right)^{0.5} \exp\{(-0.5N)\bar{x}^2\}}{\left(\frac{N}{2\pi}\right)^{0.5} \exp\{-0.5N(\bar{x}-1)^2\}}\right] \tag{9.10}$$

$$= \exp\{-0.5N[\bar{x}^2 - (\bar{x}-1)^2]\}$$

$$= \exp\{-0.5N(2\bar{x}-1)\}. \tag{9.11}$$

Suppose our sample is of size $N = 10$; and we find $\bar{x} = 2$. Then, the posterior odds ratio becomes:

$$\frac{P\{H_0 \mid T\}}{P\{H_1 \mid T\}} = e^{-0.5N(2\bar{x}-1)} = 3.1 \times 10^{-7}. \tag{9.12}$$

Since the posterior odds ratio is so small, we must clearly reject H_0 in favor of $H_1 : \theta = 1$. Because the prior odds ratio is unity in this case, the posterior odds ratio is equal to the Bayes' factor.

Note that comparing the posterior odds ratio with unity is equivalent to choosing the larger of the two posterior probabilities of the hypotheses. If we could assign losses to the two possible incorrect decisions, we would choose the hypothesis with the smaller expected loss. (See Chapter 11 for the role of loss functions in decision making)

9.5.2 Testing a Simple Null Hypothesis Against a Composite Alternative Hypothesis

Next we consider the more common case of testing a simple hypothesis H_0 against a composite hypothesis H_1. Suppose there is a parameter θ (possibly vector valued) indexing the distribution of the test statistic $T = T(X_1, \ldots, X_N)$. Then, the ratio of the posterior density of H_0 compared with that of H_1 is:

$$\frac{P\{H_0 \mid T\}}{P\{H_1 \mid T\}} = \frac{P\{T \mid H_0\}P\{H_0\}}{P\{T \mid H_1\}P\{H_1\}} = \left[\frac{P\{H_0\}}{P\{H_1\}}\right] \frac{P\{T \mid H_0, \theta\}}{\int P\{T \mid H_1, \theta\} g(\theta)\, d\theta}, \tag{9.13}$$

where: $g(\theta)$ denotes the prior density for θ under H_1. Thus, the posterior odds ratio, in the case of a composite alternative hypothesis, is the product of the prior odds ratio times the ratio of the averaged or marginal likelihoods under H_0 and H_1. (Note that under H_0, because it is a simple hypothesis, the likelihood has only one value, so its average is that value. If the null hypothesis were also composite, we would need to use an integral average in that case as well.) We assume, of course, that these integrals converge. (In the event $g(\theta)$ is an improper density, the integrals will not always exist.) Note also that in this case the Bayes' factor is the ratio of the likelihood under H_0 to the averaged likelihood under H_1.

We have assumed there are no additional parameters in the problem. If there are, we deal with them by integrating them out with respect to an appropriate prior distribution.

For example, suppose $(X \mid \theta, \sigma^2) \sim N(\theta, \sigma^2)$, and we are interested in testing the hypothesis $H_0 : \{\theta = 0, \sigma^2 > 0\}$, versus the alternative hypothesis $H_1 : \{\theta \neq 0, \sigma^2 > 0\}$. If $X_1, \ldots, X_n$ are i.i.d., $(\bar{X}, s^2)$ is sufficient for (θ, σ^2), where s^2 is the sample variance. Then, the posterior odds ratio for testing H_0 versus H_1 is:

$$\frac{P\{H_0 \mid \bar{x}, s^2\}}{P\{H_1 \mid \bar{x}, s^2\}} = \left[\frac{P\{H_0\}}{P\{H_1\}}\right] \times \left[\frac{\int f_1(\bar{x} \mid \theta = 0) f_2(s^2 \mid \sigma^2) g_2(\sigma^2)\, d(\sigma^2)}{\iint f_1(\bar{x} \mid \theta) f_2(s^2 \mid \sigma^2) g_1(\theta) g_2(\sigma^2)\, d(\sigma^2)\, d\theta}\right], \quad (9.14)$$

for appropriate prior densities, $g_1(\theta)$ and $g_2(\sigma^2)$.

As an example of a simple versus composite hypothesis testing problem in which there are no additional parameters, suppose $(X \mid \theta) \sim N(\theta, 1)$, and we are interested in testing $H_0 : \theta = 0$, versus $H_1 : \theta \neq 0$. We take a random sample $X_1, \ldots, X_{10}$, of size $N = 10$, and form the sufficient statistic $T = \bar{X}$, and assume $\bar{X} = 2$. Assume $P\{H_0\} = P\{H_1 = 0.5$. We note that $(\bar{X} \mid \theta) \sim N(\theta, 1/N)$, and so

$$P\{T \mid H_0, \theta\} = \left(\frac{N}{2\pi}\right)^{0.5} e^{-(0.5N)\bar{x}^2}, \quad (9.15)$$

and

$$P\{T \mid H_1, \theta\} = \left(\frac{N}{2\pi}\right)^{0.5} e^{-(0.5N)(\bar{x}-\theta)^2}. \quad (9.16)$$

As a prior distribution for θ under H_1 we take $\theta \sim N(1, 1)$. Then,

$$g(\theta) = \frac{1}{\sqrt{2\pi}} e^{-0.5(\theta-1)^2}.$$

The posterior odds ratio becomes:

$$\frac{P\{H_0 \mid T\}}{P\{H_1 \mid T\}} = \frac{\left(\frac{N}{2\pi}\right)^{0.5} e^{-0.5N\bar{x}^2}}{\int \left(\frac{N}{2\pi}\right)^{0.5} e^{-0.5N(\bar{x}-\theta)^2} \times \frac{1}{\sqrt{2\pi}} e^{-0.5(\theta-1)^2} d\theta}$$

$$= \frac{\sqrt{2\pi} e^{-(0.5N)\bar{x}^2}}{\int e^{-0.5[(\theta-1)^2+N(\theta-\bar{x})^2]} d\theta}$$

$$= \sqrt{(N+1)} \exp\left\{(-0.5)\left[\frac{(N\bar{x}+1)^2}{(N+1)} - 1\right]\right\}.$$

Since $N = 10$, and $\bar{x} = 2$, we have

$$\frac{P\{H_0 \mid T\}}{P\{H_1 \mid T\}} = 1.1 \times 10^{-8}. \tag{9.18}$$

Thus, we reject $H_0 : \theta = 0$ in favor of $H_1 : \theta \neq 0$. That is, the evidence strongly favors the alternative hypothesis H_1.

9.5.3 Problems With Bayesian Hypothesis Testing with Vague Prior Information

Comparing models or testing hypotheses when the prior information about the unknowns is weak or vague presents some difficulties. Note from Equation (9.14) that as long as $g_1(\theta)$ and $g_2(\sigma^2)$ are proper prior densities, the integrals and the Bayes' factor are well defined. The subjectivist Bayesian scientist who uses subjective information to assess his/her prior distributions will have no difficulty adopting Jeffreys' method of testing hypotheses or comparing models and theories. But the objectivist Bayesian has considerable problems, as will be seen from the following discussion.

Suppose, for example, that $g_2(\sigma^2)$ is a vague prior density so that:

$$g_2(\sigma^2) \propto \frac{1}{\sigma^2}.$$

The proportionality constant is arbitrary. So, in this situation, in the ratio of integrals in Equation (9.14), there results an arbitrary ratio of constants, rendering the criterion for decision arbitrary.

A solution to this problem was proposed by Lempers (1971, Section 5.3). He suggested that in such situations, the data could be divided into two parts, the first of which is used as a *training sample*, and the remaining part for hypothesis testing. A two-step procedure results. For the first step, the training sample is used with a vague

prior density for the unknowns and a posterior density is developed in the usual way. This posterior distribution is not used for comparing models. In the second step, the posterior density developed in the first step is used as the (proper) prior density for model comparison with the remaining part of the data. Now there are no hypothesis testing problems. The resulting Bayes' factor is now based upon only part of the data (the remaining part of the data after the training data portion is extracted) and accordingly is called a *partial Bayes' factor.*

A remaining problem is how to subdivide the data into two parts. Berger and Pericchi (1996) suggested that the training data portion be determined as the smallest possible data set that would generate a proper posterior distribution for the unknowns (a proper posterior distribution is what is required to make the Lempers proposal operational). There are, of course, many ways to generate such a minimal training data set. Each such training data set would generate a feasible Bayes' factor. Berger and Pericchi call the average of these feasible Bayes' factors the *intrinsic Bayes' factor.* If the partial Bayes' factor is robust with respect to which training data set is used (so that resulting posterior probabilities do not vary much) using the intrinsic Bayes' factor is very reasonable. If the partial Bayes' factor is not robust in this sense, there can still be problems.

O'Hagen (1993) considers the case in which there are very large data sets so that asymptotic behavior can be used. For such situations he defines a *fractional Bayes' factor* that depends upon the fraction "b" of the total sample of data that has not been used for model comparison. It is not clear that the use of fractional Bayes' factors will improve the situation in small or moderate size samples.

The Bayesian (Jeffreys) approach is now the preferred method of comparing scientific theories. For example, in the book by Mathews and Walker (1965, pp. 361–370), in which in the Preface the authors explain that the book was an outgrowth of lectures by Richard Feynman at Cornell University, Feynman suggests that to compare contending theories (in physics) one should use the Bayesian approach. (This fact was called to my attention by Dr. Carlo Brumat.)

SUMMARY

This chapter has presented the Bayesian approach to hypothesis testing and model comparison. We traced the development of scientific hypothesis testing from the approach of Karl Pearson to that of Harold Jeffreys. We showed how the Bayesian approach differs from the frequentist approach, and why there are problems with the frequentist methodology. We introduced both the Lindley vague prior approach to hypothesis testing as well as the Jeffreys general prior approach to testing and model comparison. We examined the testing of simple null hypotheses against simple alternative hypotheses as well as the testing of simple versus composite hypotheses. We discussed Bayes' factors, partial Bayes' factors, intrinsic Bayes' factors, and fractional Bayes' factors.

EXERCISES

9.1 Suppose that $X_1, \ldots, X_{10}$ are independent and identically distributed as $N(\theta, 4)$. Test the hypothesis that $H_0 : \theta = 3$ versus the alternative hypothesis $H_0 : \theta \neq 3$. Assume that you have observed $\bar{X} = 5$, and that your prior probabilities are $P\{H_0\} = 0.6$, and $P\{H_1\} = 0.4$. Assume that your prior probability for θ follows the law $N(1,1)$. Use the Jeffreys' testing procedure.

9.2 Give the Bayes' factor for Exercise 9.1.

9.3 Suppose the probability mass function for an observable variable X is given by: $f(x \mid \lambda) = (e^{-\lambda}\lambda^x)/x!, x = 0, 1, \ldots, \lambda > 0$. Your prior density for λ is given by: $g(\lambda) = 2e^{-2\lambda}$. You observe $X = 3$. Test the hypothesis $H_0 : \lambda = 1$, versus the alternative hypothesis $H_1 : \lambda \neq 1$. Assume that your prior probabilities on the hypotheses are: $P\{H_1\} = P\{H_0\} = 1/2$.

9.4 Explain the use of the *intrinsic Bayes' factor.*

9.5 Explain the difference between the Lindley and Jeffreys methods of Bayesian hypothesis testing.

9.6* What is meant by "Lindley's paradox"?

9.7 Explain some of the problems associated with the use of p-values and significance testing.

9.8 Explain how frequentist hypothesis testing violates the likelihood principle.

9.9 Suppose $X_1, \ldots, X_n, n = 50$, are i.i.d. observations from $N(\theta, 7)$. You observe $\bar{X} = 2$. Suppose your prior distribution for θ is vague. Use the Lindley hypothesis testing procedure to test $H_0 : \theta = 5$, versus $H_1 : \theta \neq 5$.

9.10 Suppose that $X_1, \ldots, X_n$ are i.i.d. following the law $N(\theta, \sigma^2)$. We assume that σ^2 is unknown. Form the sample mean and variance: $\bar{X} = 1/n \sum_1^n X_i$ and $s^2 = 1/n \sum_1^n (X_i - \bar{X})^2$. You observe $\bar{x} = 5$, $s^2 = 37$, for $n = 50$. You adopt the prior distributions: $\theta \sim N(1, 1)$ and $g(1/\sigma^2) \propto (\sigma^2)^4 e^{-2\sigma^2}$, with θ and σ^2 a *priori* independent. Assume that $P\{H_0\} = 0.75, P\{H_1\} = 0.25$. Test the hypothesis $H_0 : \theta = 3$, versus $H_1 : \theta \neq 3$.

9.11 Find the Bayes' factor for the hypothesis testing problem in Exercise 9.10.

9.12* Suppose r denotes the number of successes in n trials, and r follows a binomial distribution with parameter p. Carry out a Bayesian test of the hypothesis $H : p = 0.2$, versus the alternative $A : p = 0.8$, where these are the only two possibilities. Assume that $r = 3$, and $n = 10$, and that the prior probabilities of H and A are equal.

FURTHER READING

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction," *J. Am. Statist. Assoc.*, **91**, 109–122.

*Solutions for asterisked exercises may be found in Appendix 7.

Berger, J. O. and Selke, T. (1987). "Testing A Point Null Hypothesis: The Irreconcilability of p-Values and Evidence", *Jr. Am. Statist. Assoc.*, **82**(397), 112–122.

Casella, G. and Berger, R. L. (1987). "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," *Jr. Am. Statist. Assoc.*, **82**(397), 106–111.

Fisher, R. A. (1925) 1970. *Statistical Methods for Research Workers*, 14th Edition, New York; Hafner; Edinburgh, Oliver and Boyd.

Good, I. J. (1950). *Probability and the Weighting of Evidence*, London, Charles Griffin and Co., Ltd.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Research Monograph #30, Cambridge, MA, The MIT Press.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*, Minneapolis, University of Minnesota Press.

Jeffreys, H. (1939), (1948), (1961). *Theory of Probability*, 3rd Edition, Oxford, The Clarendon Press.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*, New York, John Wiley and Sons, Inc.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*, Rotterdam, University Press.

Lindley, D. V. (1957). "A Statistical Paradox," *Biometrika*, **44**, 187–192.

Lindley, D. V. (1965). *Introduction to Probability and Statistics (Part 1—Probability, and Part 2—Inference)*, Cambridge, Cambridge University Press.

Mathews, J. and Walker, R. L. (1965). *Mathematical Methods of Physics*, New York, W. A. Benjamin, Inc.

Neyman, J. and Pearson, E. S. (1933). "On the Testing of Statistical Hypotheses in Relation to Probability A Priori," *Proc. Of the Cambridge Phil. Soc.*, **29**, 492–510.

Neyman, J. and Pearson, E. S. (1966). *Joint Statistical Papers of J. Neyman and E. S. Pearson*, Berkeley, CA, University of California Press (10 papers).

O'Hagen, A. (1993). "Fractional Bayes Factors for Model Comparison," Statistical Research Report 93-6, University of Nottingham.

Pearson, K. (1892). *The Grammar of Science*, London, Adam and Charles Black.

Popper, K. (1935), (1959). *The Logic of Scientific Discovery*, New York, Basic Books; London, Hutchinson.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*, Boston, Graduate School of Business Administration, Harvard University.

Savage, L. J. (1954). *The Foundation of Statsistics*, New York; John Wiley & Sons Inc.

"Student" (William Sealy Gosset) (1908). *Biometrika*, **6**, 1–25. (Paper on the Student t-distribution.)

Wald, A. (1939). "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *Ann. Math. Statist.*, **10**, 299–326.