Chapter 4
Hypothesis Testing and Variable Selection

4.1 Introduction

In Sects. 2.9 and 3.10, we briefly described the frequentist and Bayesian machinery for carrying out hypothesis testing. In this chapter we extend this discussion, with an emphasis on critiquing the various approaches and on hypothesis testing in a regression setting. We examine both single and multiple hypothesis testing situations; Sects. 4.2 and 4.3 consider the frequentist and Bayesian approaches, respectively. Section 4.4 describes the well-known Jeffreys–Lindley paradox that highlights the starkly different conclusions that can occur when frequentist and Bayesian hypothesis testing is carried out. This is in contrast to estimation, in which conclusions are often in agreement. In Sects. 4.5–4.7, various aspects of multiple testing are considered. The discussion includes situations in which the number of tests is known a priori and variable selection procedures in which the number of tests is driven by the data. Section 4.9 provides a discussion of the impact on inference that the careless use of variable selection can have. Section 4.10 describes a pragmatic approach to variable selection. Concluding remarks appear in Section 4.11.

4.2 Frequentist Hypothesis Testing

Early in this chapter we will consider a univariate parameter $\theta \in \mathbb{R}$. Suppose we are interested in evaluating the evidence in the data with respect to the null hypothesis:

$$H_0 : \theta = \theta_0$$

using a statistic T. By convention, large values are less likely under the null. The observed value of the test statistic is t_{obs}. As discussed in Sect. 2.9, there are various possibilities for T including squared Wald, likelihood ratio, and score

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_4,

statistics. Under regularity conditions, $T \rightarrow_d \chi^2_1$ under the null, as $n \rightarrow \infty$. If n is not large, or regularity conditions are violated, permutation or Monte Carlo tests (perhaps based on bootstrap samples, as described in Sect. 2.7) can often be performed to derive the empirical distribution of the test statistic under the null. A type I error is said to occur when we reject H_0 when it is in fact true, while a type II error is to not reject H_0 when it is false.

4.2.1 Fisherian Approach

Under the null, for continuous sample spaces, the tail-area probability $\Pr(T > t \mid H_0)$ is uniform. This is not true for discrete sample spaces, but in the following, unless stated otherwise, we will assume we are in situations in which uniformity holds. Let

$$p = \Pr(T > t_{\text{obs}} \mid H_0)$$

denote the observed p-value, the probability of observing t_{obs}, *or a more extreme value*, with repeated sampling under the null.

Fisher advocated the pure test of significance, in which the observed p-value is reported as the measure of evidence against the null (Fisher 1925a), with H_0 being rejected if p is small. Alternative hypotheses are not explicitly considered and so there is no concept of rejecting the null in favor of a specific alternative; ideally, the test statistic will be chosen to have high power under plausible alternatives, however.

4.2.2 Neyman–Pearson Approach

In contrast to the procedure of Fisher, the Neyman–Pearson approach is to specify an alternative hypothesis, H_1, with H_0 nested in H_1. The celebrated Neyman–Pearson lemma of Neyman and Pearson (1933) proved that, for fixed type I error

$$\alpha = \Pr(T > t_{\text{fix}} \mid H_0),$$

the most powerful procedure is provided by the likelihood ratio test (Sect. 2.9.5). The decision rule is to reject the null if $p < \alpha$. Due to the *fixed* threshold, this procedure controls the type I error at α.

4.2.3 Critique of the Fisherian Approach

A common explanation for seeing a "small" p-value is that *either* H_0 is not true *or* H_0 is true and we have been "unlucky." A major practical difficulty is on defining "small." Put another way, how do we decide on a *threshold* for significance?

The p-value is uniform under the null, but with a large sample size, we will be able to detect very subtle departures from the null and so will often obtain small p-values because the null is rarely "true." To rectify this a confidence interval for θ is often reported, along with the p-value, so that the scientific significance of the departure of θ from θ_0 can be determined. The ability to detect smaller and smaller differences from the null with increasing sample size suggests that the p-value threshold rule used in practice should decrease with increasing n, but there are no universally recognized rules. In a hypothesis testing context a natural definition of consistency is that the rule for rejection is such that the probability of the correct decision being made tends to 1 as the sample size increases. So the current use of p-values, in which typically 0.05 or 0.01 is used as a threshold for rejection, regardless of sample size, is *inconsistent*; by construction, the probability of rejecting the null when it is true does not decrease to zero with increasing sample size. By contrast, the type II error will typically decrease to zero with increasing sample size. A more balanced approach than placing special emphasis on the type I error would be to have both type I and type II errors decrease to zero as n increases.

There are two common misinterpretations of p-values. The most basic is to interpret a p-value as the probability of the null given the data, which is a serious misconception. Probabilities of the truth of hypotheses are only possible under a Bayesian approach. More subtly, using the *observed* value of the test statistic t_{obs} does not allow one to say that following the general procedure will result in control of the type I error at p, because the threshold is data-dependent and not fixed. The key observation is that the p-value is associated with, "observing t_{obs}, or a more extreme value," so that the tail area begins at the *observed* value of the statistic. For example, if $p = 0.013$, we cannot say that the procedure controls the type I error at 1.30%. Such control of the type I error is provided by a fixed α level procedure which is based on a fixed threshold, t_{fix} with $\alpha = \Pr(T > t_{\text{fix}} \mid H_0)$.

There is some merit in the consideration of a tail area when one wishes to control the type I error rate, but when no such control is sought, the use of a tail area seems simply of mathematically convenience. As an alternative the ordinate $p(T = t_{\text{obs}} \mid H_0)$ may be considered, which brings one closer to a Bayesian formulation (see Sect. 4.3.1), but from a frequentist perspective, it is not clear how to scale the observed statistic without an alternative hypothesis.

4.2.4 Critique of the Neyman–Pearson Approach

As with the use of p-values we need to decide on a size α for the test. The historical emphasis has been on fixing α and then evaluating power, but as with a threshold for p-values, practical guidance on how α should depend on sample size is important but lacking. With an α level that does not change with sample size, one is implicitly accepting that type II errors become more important with increasing sample size, and in a manner which is implied rather than chosen by the investigator. Pearson (1953, p. 68) expressed the desirability of a decreasing α as sample size increases:

"... the quite legitimate device of reducing α as n increases." As we have already noted, a fixed significance level with respect to n gives an inconsistent procedure.

By merely stating that $p < \alpha$, information is lost, but if we state an observed p-value, then we lose control of the type I error because control requires a fixed binary decision rule. The procedure must also be viewed in the light of both H_0 and H_1 being "wrong" since no model is a correct specification of the data-generating process.

For discrete data, the discreteness of the statistic causes difficulties, particularly for small sample sizes. To achieve exact level α tests, so-called randomization rules have been suggested. Under such rules, the same set of data may give different conclusions depending on the result of the randomization, which is clearly undesirable.

4.3 Bayesian Hypothesis Testing with Bayes Factors

4.3.1 Overview of Approaches

In the Bayesian approach, all unknowns in a model are treated as random variables, even though they relate to quantities that are in reality fixed. Therefore, the "true" hypothesis is viewed as an unknown parameter for which the posterior is derived, once the alternatives have been specified. The latter step is essential since we require a sample space of hypotheses. In the case of two hypotheses, we have the following candidate data-generating mechanisms:

$$\begin{aligned} H_0 &\Rightarrow \boldsymbol{\beta}_0 \mid H_0 \Rightarrow \boldsymbol{y} \mid \boldsymbol{\beta}_0 \\ H_1 &\Rightarrow \boldsymbol{\beta}_1 \mid H_1 \Rightarrow \boldsymbol{y} \mid \boldsymbol{\beta}_1. \end{aligned}$$

The posterior probability of H_j is, via Bayes theorem,

$$\Pr(H_j \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid H_j) \times \pi_j}{p(\boldsymbol{y})}$$

with π_j the prior probability of hypothesis H_j, $j = 0, 1$. The likelihood of the data is

$$p(\boldsymbol{y} \mid H_j) = \int p(\boldsymbol{y} \mid \boldsymbol{\beta}_j) p(\boldsymbol{\beta}_j \mid H_j)\, d\boldsymbol{\beta}_j \tag{4.1}$$

with $p(\boldsymbol{\beta}_j \mid H_j)$ the prior distribution over the parameters associated with hypothesis H_j, $j = 0, 1$, and

$$p(\boldsymbol{y}) = p(\boldsymbol{y} \mid H_0) \times \pi_0 + p(\boldsymbol{y} \mid H_1) \times \pi_1.$$

The posterior odds in favor of H_0 is therefore

$$\text{Posterior Odds} = \frac{\Pr(H_0 \mid \boldsymbol{y})}{\Pr(H_1 \mid \boldsymbol{y})} = \text{Bayes factor} \times \text{Prior Odds} \tag{4.2}$$

where the

$$\text{Bayes factor} = \frac{p(\boldsymbol{y} \mid H_0)}{p(\boldsymbol{y} \mid H_1)}, \tag{4.3}$$

and the prior odds are π_0/π_1 with $\pi_1 = 1 - \pi_0$. The Bayes factor is the ratio of the density of the data under the null to the density under the alternative and is an intuitively appealing summary of the information the data provide concerning the hypotheses. The Bayes factor was discussed previously in Sect. 3.10. From (4.2), we also see that

$$\text{Bayes Factor} = \frac{\text{Posterior Odds}}{\text{Prior Odds}},$$

which emphasizes that the Bayes factor summarizes the information in the data and does not involve the prior beliefs about the hypotheses. As can be seen in (4.1), priors on the parameters are involved in each of the numerator and denominator of the Bayes factor, since these provide the distributions over which the likelihoods are averaged.

When it comes to reporting/making decisions, various approaches based on Bayes factors are available for different contexts. Most simply, one may just report the Bayes factor. Kass and Raftery (1995), following Jeffreys (1961), present a guideline for the interpretation of Bayes factors. For example, if the negative log base 10 Bayes factor lies between 1 and 2 (so that the data are 10–100 times more likely under the alternative, as compared to the null), then there is said to be *strong* evidence against the null hypothesis. Such thresholds may be useful in some situations, but in general one would like the guidelines to be context driven. Going beyond the consideration of the Bayes factor only, one may include prior probabilities on the null and alternative, to give the posterior odds (4.2). Stating the posterior probabilities may be sufficient, but one may wish to derive a formal rule for deciding upon which of H_0 or H_1 to report.

Recall from Sect. 3.10 that, under a Bayesian *decision theory* approach to hypothesis testing, the "decision" δ is taken that minimizes the posterior expected loss. Following the notation of Table 3.3, the losses associated with type I and type II errors are L_{I} and L_{II}, respectively. Minimization of the posterior expected loss then results in the rule to choose $\delta = 1$ if

$$\frac{\Pr(H_1 \mid \boldsymbol{y})}{\Pr(H_0 \mid \boldsymbol{y})} \geq \frac{L_{\text{I}}}{L_{\text{II}}},$$

or equivalently if

$$\Pr(H_1 \mid \boldsymbol{y}) \geq \frac{1}{1 + L_{\text{II}}/L_{\text{I}}}. \tag{4.4}$$

For example, if a type I error is four times as bad as a type II error, we should report H_1 only if $\Pr(H_1 \mid \boldsymbol{y}) \geq 0.8$. In contrast, if the balance of losses is reversed, and a type II error is four times as costly as a type I error, we report H_1 if $\Pr(H_1 \mid \boldsymbol{y}) \geq 0.2$.

Discreteness of the sample space does not pose any problems for a Bayesian analysis, since one need only consider the data actually observed and not other hypothetical realizations.

4.3.2 Critique of the Bayes Factor Approach

As always with the Bayesian approach, we need to specify priors for all of the unknowns, which here correspond to each of the hypotheses and all parameters (including nuisance parameters) that are contained within the models defined under the two hypotheses. It turns out that placing improper priors upon the parameters that are the focus of the hypothesis test leads to anomalous behavior of the Bayes factor. We give an informal discussion of the fundamental difference between estimation and hypothesis testing with respect to the choice of improper priors. Suppose we have a model that depends on a univariate unknown parameter, θ with improper prior $p(\theta) = c$, for arbitrary $c > 0$. The posterior, upon which estimation is based, is

$$\frac{p(\boldsymbol{y} \mid \theta)p(\theta)}{\int p(\boldsymbol{y} \mid \theta)p(\theta)\, d\theta} \tag{4.5}$$

and so the arbitrary constant in the prior cancels in both numerator and denominator. Now suppose we are interested in comparison of the hypotheses $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$ with $\theta \in \mathbb{R}$. The Bayes factor is

$$\frac{p(\boldsymbol{y} \mid H_0)}{p(\boldsymbol{y} \mid H_1)} = \frac{p(\boldsymbol{y} \mid \theta_0)}{\int p(\boldsymbol{y} \mid \theta)p(\theta)\, d\theta},$$

so that the denominator of the Bayes factor depends, crucially, upon c. Hence, in this setting the Bayes factors with an improper prior on θ is not well defined.

Specifying prior distributions for all of the parameters under each hypothesis can be difficult, but Sect. 3.11 describes a strategy based on test statistics which requires a prior distribution for the parameter of interest only.

In principle, one can compare non-nested models using a Bayesian approach, but in practice great care must be taken in specifying the priors under the two hypotheses, in order to not inadvertently favor one hypothesis over another. One possibility is to specify priors on functions of the parameters that are meaningful under both hypotheses; for an example of this approach, see Sect. 6.16.

As with the Neyman–Pearson approach, all of the calculations have to be conditioned upon $H_0 \cup H_1$. In a Bayesian context, we need to emphasize that we are obtaining the posterior probability of the null given one of the null or alternative is true and under the assumed likelihood and priors. Consequently,

posterior probabilities on hypotheses must be viewed in a relative, rather than an absolute, sense since the truth will rarely correspond to H_0 or H_1. Hence, the precise interpretation is that the posterior probability of H_0 is the posterior probability of H_0, *given* that one of H_0 or H_1 is true.

If one follows the decision theory route, one must also specify the ratio of losses which is usually difficult. In general, Bayes factor calculation requires analytically intractable integrals over the null and alternative parameter spaces, to give the two normalizing constants $p(\boldsymbol{y} \mid H_0)$ and $p(\boldsymbol{y} \mid H_1)$. Further, Markov chain Monte Carlo approaches do not simply supply these normalizing constants. Analytical approximations exist under certain conditions, see Sect. 3.10.

4.3.3 A Bayesian View of Frequentist Hypothesis Testing

We consider an artificial situation in which the only available data in a Bayesian analysis corresponds to knowing that the event $T > t_{\text{fix}}$ has occurred. This means that the likelihood of the data, $\Pr(\text{ data } \mid H_0)$ coincides with the α level. To obtain $\Pr(H_0 \mid \text{ data })$ we must specify the alternative hypothesis. We consider the simple case in which the model contains a single parameter θ with null $H_0 : \theta = \theta_0$ and alternative $H_1 : \theta = \theta_1$. Then

$$\Pr(H_0 \mid \text{ data }) = \frac{\Pr(\text{ data } \mid H_0) \times \pi_0}{\Pr(\text{ data } \mid H_0) \times \pi_0 + \Pr(\text{ data } \mid H_1) \times \pi_1} \tag{4.6}$$

where $\pi_j = \Pr(H_j)$, $j = 0, 1$. Dividing by $\Pr(H_1 \mid \text{ data })$ gives

$$\begin{aligned} \text{Posterior Odds} &= \frac{\Pr(\text{ data } \mid H_0)}{\Pr(\text{ data } \mid H_1)} \times \text{ Prior Odds} \\ &= \frac{\alpha}{\text{power at } \theta_1} \times \text{ Prior Odds} \end{aligned} \tag{4.7}$$

which depends, in addition to the α level, on the *prior* on H_0, π_0, and on the *power*, $\Pr(\text{ data } \mid H_1)$. Equation (4.7) implies that, for two studies that report a result as significant at the same α level, the one with the greater power will, in a Bayesian formulation, provide greater evidence against the null. The power is never explicitly considered when reporting under the Fisherian or Neyman–Pearson approaches. An important conclusion is that to make statements about the "evidence" that the data contain with respect to a hypothesis, as summarized in an α level, one would want to know the power or, as a minimum, the sample size (since this is an important component of the power).

The prior is also important which seems, as already noted, reasonable when one considers the usual interpretation of a tail area in terms of "either H_0 is true and we were unlucky or H_0 is not true." A prior on H_0 is very useful in weighing these two possibilities. A key observation is that although a particular dataset may be unlikely

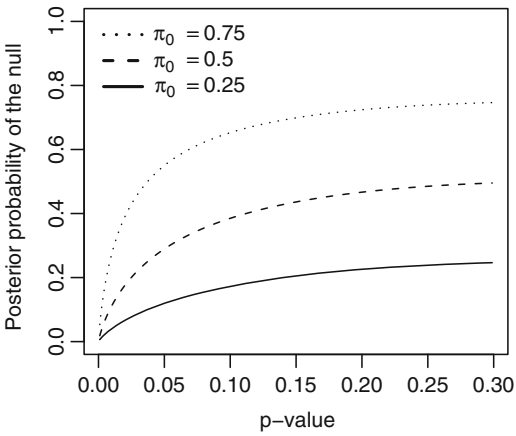

Fig. 4.1 Lower bound for $\Pr(H_0 \mid \text{data})$, under three prior specifications, as a function of the p-value

under the null, it may also be unlikely under chosen alternatives, so that there may be insufficient evidence to reject the null, at least in comparison to these alternatives.

Sellke et al. (2001) summarize a number of different arguments that lead to the following, quite remarkable, result. For a p-value $p < e^{-1} = 0.368$:

$$\Pr(H_0 \mid \text{ data }) \geq \left[1 - \left(\frac{1}{ep \log p} \times \frac{\pi_1}{\pi_0}\right)^{-1}\right]^{-1}. \tag{4.8}$$

Hence, given a p-value, one may calculate a lower bound on the posterior probability of the null. Figure 4.1 illustrates this lower bound, as a function of the p-value, for three different prior probabilities, π_0. We see, for example, that with a p-value of 0.05 and a prior probability on the null of $\pi_0 = 0.75$, we obtain $\Pr(H_0 \mid \text{ data }) \geq 0.55$.

The discussion of Sect. 4.2.3, combined with the implications of (4.7) and (4.8), might prompt one to ask why p-values are still in use today, in particular with the almost ubiquitous application of a 0.05 or 0.01 decision threshold. With these thresholds, which are often required for the publication of results, the relationship (4.8), with $\pi_0 = 0.5$, gives $\Pr(H_0 \mid \text{ data }) \geq 0.29$ and 0.11 with $p = 0.05$ and 0.01, respectively. Rejection of H_0 with such probabilities may not be unreasonable in some circumstances but the difference between the p-value and $\Pr(H_0 \mid \text{data})$ is apparent.

Small prior probabilities, π_0, were not historically the norm since, particularly in experimental situations, data would not be collected if there were little chance the alternative were true.

In some disciplines scientists may calibrate p-values to the sample sizes with which they are familiar, as no doubt Fisher did when the 0.05 rule emerged. For example, in Tables 29 and 30 of Statistical Methods for Research Workers (Fisher 1990), the sample sizes were 30 and 17, and Fisher discusses the 0.05 limit in each case, though in both cases he concentrates more on the context than on the absolute value of 0.05.

Poor calibration of p-values could be one of the reasons why so many "findings" are not reproducible, along with the other usual suspects of confounding, data dredging, multiple testing, and poorly measured covariates.

4.4 The Jeffreys–Lindley Paradox

We now discuss a famous example in which Bayesian and frequentist approaches to hypothesis testing give starkly different conclusions. The example has been considered by many authors, but Lindley (1957) and Jeffreys (1961) provide early discussions; see also Bartlett (1957). To illustrate the so-called Jeffreys–Lindley "paradox," we assume that $\overline{Y}_n \mid \theta \sim \mathrm{N}(\theta, \sigma^2/n)$ with σ^2 known and θ unknown. Suppose the null is $H_0 : \theta = 0$, with alternative $H_1 : \theta \neq 0$. Let

$$\overline{y}_n = z_{1-\alpha/2} \times \sigma/\sqrt{n}$$

where α is the level of the test and $\Pr(Z < z_{1-\alpha/2}) = 1 - \alpha/2$, with $Z \sim \mathrm{N}(0, 1)$. We define $\overline{y}_n$ in this manner, so that for different values of n the α level remains constant. For a Bayesian analysis, assume $\pi_0 = \Pr(H_0)$, and under the alternative $\theta \sim \mathrm{N}(0, \tau^2)$. In the early discussions of the paradox, a uniform prior over a finite range was assumed, but the message of the paradox is unchanged with the use of a normal prior. Then

$$\Pr(H_0 \mid \overline{y}_n) = \frac{\text{Bayes Factor} \times \text{Prior Odds}}{1 + \text{Bayes Factor} \times \text{Prior Odds}}$$

where the Bayes factor is

$$\text{Bayes Factor} = \frac{p(\overline{y}_n \mid H_0)}{p(\overline{y}_n \mid H_1)} \tag{4.9}$$

and the Prior Odds $= \pi_0/(1 - \pi_0)$. The prior predictive distributions, the ratios of whose densities give the Bayes factor (4.9), are

$$\overline{y}_n \mid H_0 \sim \mathrm{N}(0, \sigma^2/n) \tag{4.10}$$

$$\overline{y}_n \mid H_1 \sim \mathrm{N}(0, \sigma^2/n + \tau^2). \tag{4.11}$$

Figure 4.2 shows these two densities, as a function of $\overline{y}_n$, for $\sigma^2 = 1, \tau^2 = 0.2^2$, and $n = 100$. An α level of 0.05 gives $\overline{y}_n = 1.96 \times \sigma/\sqrt{n} = 0.20$, the value indicated in the figure with a dashed-dotted vertical line. For this value, the Bayes factor equals 0.48, so that the data are roughly twice as likely under the alternative as compared to the null. The Sellke et al. (2001) bound on the Bayes factor is $\mathrm{BF} \geq -ep \log p$ which for $p = 0.05$ gives $\mathrm{BF} \geq 0.41$.

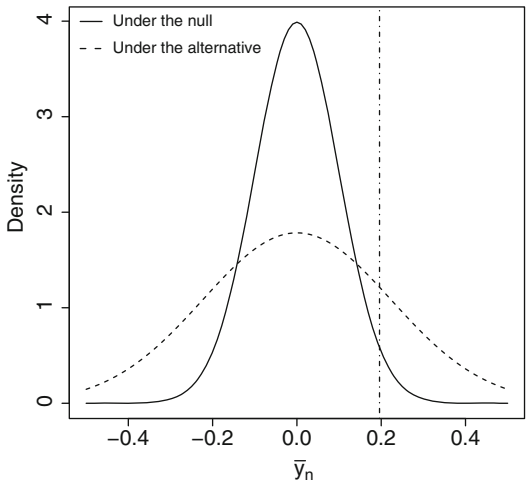

Fig. 4.2 Numerator (*solid line*) and denominator (*dashed line*) of the Bayes factor for $n = 100$. The model is $\overline{Y}_n \mid \theta \sim \mathrm{N}(\theta, \sigma^2/n)$ with $\sigma^2 = 1$. The null and alternative are $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$, and the prior under the alternative is $\theta \sim \mathrm{N}(0, \tau^2)$ with $\tau^2 = 0.2^2$. The *dashed-dotted vertical line* corresponds to $\overline{y}_n = 0.20$ which for this n gives $\alpha = 0.05$

The Bayes factor is the ratio of (4.10) and (4.11):

$$\begin{aligned}\text{Bayes Factor} &= \frac{(2\pi\sigma^2/n)^{-1/2}\exp\left[-\frac{\overline{y}_n^2}{2\sigma^2/n}\right]}{(2\pi[\sigma^2/n+\tau^2])^{-1/2}\exp\left[-\frac{\overline{y}_n^2}{2(\sigma^2/n+\tau^2)}\right]} \\ &= \sqrt{\frac{\sigma^2/n+\tau^2}{\sigma^2/n}}\exp\left[-\frac{z_{1-\alpha/2}^2}{2}\frac{\tau^2}{\tau^2+\sigma^2/n}\right]. \qquad (4.12)\end{aligned}$$

This last expression reveals that, as $n \to \infty$, the Bayes factor $\to \infty$, so that $\Pr(H_0 \mid \overline{y}_n) \to 1$. Therefore, the "paradox" is that for a level of significance α, chosen to be arbitrarily small, we can find datasets which make the posterior probability of the null arbitrarily close to 1, for some n. Hence, frequentist and Bayes procedures can, for sufficiently large sample size, come to opposite conclusions with respect to a hypothesis test.

Figure 4.3 plots the posterior probability of the null as a function of n for $\sigma^2 = 1, \tau^2 = 0.2^2, \pi_0 = 0.5, \alpha = 0.05$. From the starting position of 0.5 (the prior probability, indicated as a dashed line), the curve $\Pr(H_0 \mid \overline{y}_n)$ initially falls, reaching a minimum at around $n = 100$, and then increases towards 1, illustrating the "paradox." For large values of n, $\overline{y}_n$ is very close to the null value of 0, but there is high power to detect any difference from 0, and so an α of 0.05 is not difficult to achieve. The Bayes factor also incorporates the density under the alternative and values close to 0 are more likely under the null, as illustrated in Fig. 4.2.

We now consider a Bayesian analysis of the above problem but assume that the data appear only in the form of knowing that $|\overline{Y}_n| \geq \overline{y}_n$, a censored observation. This is clearly not the usual situation since a Bayesian would condition on the *actual* value observed, but it does help to understand the paradox. The Bayes factor is

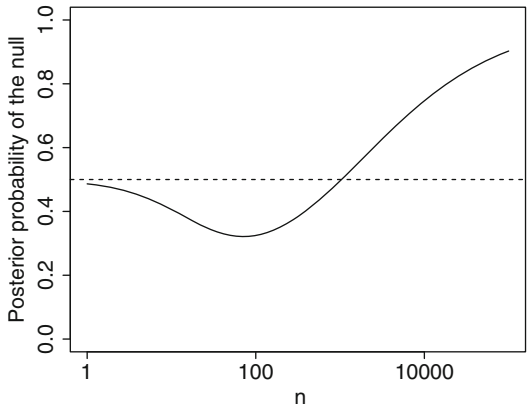

Fig. 4.3 Posterior probability of the null versus sample size, for a fixed α level of 0.05. The model is $\overline{Y}_n \mid \theta \sim \mathrm{N}(\theta, \sigma^2/n)$ with $\sigma^2 = 1$. The null and alternative are $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$, and the prior under the alternative is $\theta \sim \mathrm{N}(0, \tau^2)$ with $\tau^2 = 0.2^2$

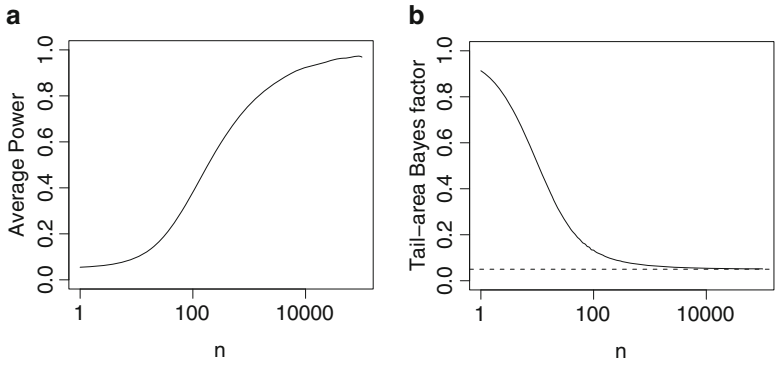

Fig. 4.4 Bayes factor based on a tail area with null and alternative of $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$: (**a**) Average power, which corresponds to the denominator of the Bayes factor, under a $\mathrm{N}(0, 0.2^2)$ prior and for a fixed α level of 0.05 and (**b**) Bayes factor based on the tail area, with $\alpha = 0.05$; the *horizontal dashed line* indicates a tail-area Bayes factor value of 0.05

$$\frac{\Pr(|\overline{Y}_n| \geq \overline{y}_n | H_0)}{\Pr(|\overline{Y}_n| \geq \overline{y}_n | H_1)} = \frac{\alpha}{\int \Pr(|\overline{Y}_n| \geq \overline{y}_n | \theta) p(\theta)\, d\theta},$$

that is, the type I error rate divided by the power averaged over the prior $p(\theta)$. Figure 4.4a gives the average power as a function of n. We see a monotonic increase with sample size towards the value 1, as we would expect with fixed α.

Since the Bayes factor is the ratio of α to the average power, we see in Fig. 4.4b that the Bayes factor based on the tail-area information is monotonic decreasing towards α as n increases (and with $\pi_0 = 0.5$, this gives the posterior probability of the null also). For our present purposes, the calculation with the tail area illustrates that when a Bayesian analysis conditions on a tail area, the conclusions are in line with a frequentist analysis.

The difference in behavior between a genuine Bayesian analysis that conditions on the actual statistic and that based on conditioning on a tail area is apparent. As noted by Lindley (1957, p. 189–190), "... the paradox arises because the significance level argument is based on the area under a curve and the Bayesian argument is based on the ordinate of the curve."

Ignoring now the comparison with tests of significance, it is informative to examine the Bayes factor for fixed $\overline{y}_n$. Upon rearrangement of (4.12),

$$\text{Bayes Factor} = \sqrt{\frac{\sigma^2 + n\tau^2}{\sigma^2}} \exp\left[-\frac{\overline{y}_n^2}{2} \frac{n/\sigma^2}{1 + \sigma^2/n\tau^2}\right].$$

As $\tau^2 \to \infty$, the Bayes Factor $\to \infty$ so that $\Pr(H_0 \mid \overline{y}_n) \to 1$, which is at first sight counter intuitive since increasing τ^2 places *less* prior mass close to $\theta = 0$. However, this behavior occurs because averaging with respect to the prior on θ with large τ^2 produces a small $\Pr(\overline{y}_n \mid H_1)$, because the prior under the alternative is spreading mass very thinly across a large range; $\tau^2 \gg 0$ suggests very little prior belief in any $\theta \neq 0$. Hence, even if the data point strongly to a particular $\theta \neq 0$, we still prefer H_0. More generally, $\tau^2 \gg 0$ should not be interpreted as "ignorance" since it supports very *big* effects. Said another way, as $\tau^2 \to 0$, the Bayes factor favors the alternative, even though as τ^2 gets smaller and smaller the prior under the alternative becomes more and more concentrated about the null.

4.5 Testing Multiple Hypotheses: General Considerations

In the following sections we examine how inference proceeds when more than a single hypothesis test is performed. There are many situations in which multiple hypothesis testing arises, but we concentrate on just two. In the first, which we refer to as a *fixed number of tests* scenario, we suppose that the number of hypotheses to be tested is known a priori, and is not data driven, which makes the task of evaluating the properties of proposed solutions (both frequentist and Bayesian) more straightforward. This case is discussed in Sect. 4.6. As an example, we will shortly introduce a running example that concerns comparing, between two populations, expression levels for $m = 1,000$ gene transcripts (during transcription, a gene is transcribed into (mutiple) RNA transcripts). In the second situation, which we refer to as *variable selection*, and which is discussed in Sect. 4.7, the number of hypotheses to be tested is random, which makes the evaluation of properties more difficult.

One of the biggest abuses of statistical techniques is the unprincipled use of model selection. Two examples of this are separately testing the significance of a large number of variables and then reporting only those that are nominally "significant" (the problem considered in Sect. 4.6), and testing multiple confounders to see which ones to control for (the problem considered in Sect. 4.7). In each of these cases, even if the exact procedure is described, unless care is exercised, interpretation is extremely difficult.

4.6 Testing Multiple Hypotheses: Fixed Number of Tests

Suppose we wish to examine the association between a response and m different covariates. In a typical epidemiological study, many potential risk factors are measured, and an exploratory, hypothesis-generating procedure may systematically examine the association between the outcome and each of the risk factors. In general, the covariates may not be independent, which complicates the analysis. Another fixed number of tests scenario is when m responses are examined with respect to a single covariate. Recently, there has been intense interest in so-called high throughput techniques in which thousands, or tens of thousands, of variables are measured, often as a screening exercise in which the aim is to see which of the variables are associated with some biological endpoint. For example, one may examine whether the expression levels of many thousands of genes are elevated or reduced in samples from cancer patients, as compared to cancer-free individuals.

When m tests are preformed, the aim is to decide which of the nulls should be rejected. Table 4.1 shows the possibilities when m tests are performed and K are flagged as requiring further attention. Here m_0 is the number of *true nulls*, B is the number of *type I errors*, and C is the number of *type II errors*, and each of these quantities is unknown. The aim is to select a rule on the basis of some criterion and this in turn will determine K. The internal cells of Table 4.1 are random variables, whose distribution depends on the rule by which K is derived.

Example: Microarray Data

To illustrate the multiple testing problem in a two-group setting, we examine a subset of microarray data presented by Kerr (2009). The data we analyze consist of expression levels on $m = 1,000$ transcripts measured in Epstein-Barr virus-transformed lymphoblastic cell line tissue, in each of two populations. Each transcript was measured on 60 individuals of European ancestry (CEU) and 45 ethnic Chinese living in Beijing (CHB). The data have been normalized, and $\log_2$ transformed, so that a one-unit difference between recorded values corresponds to a doubling of expression level.

Let $\overline{Y}_{ki}$ be the measured expression level for transcript i in population k, with $i = 1, \ldots, m$, and $k = 0/1$ representing the CEU/CHB populations. Then define

Table 4.1 Possibilities when m tests are performed and K are flagged as worthy of further attention

	Not flagged	Flagged	
H_0	A	B	m_0
H_1	C	D	m_1
	$m - K$	K	m

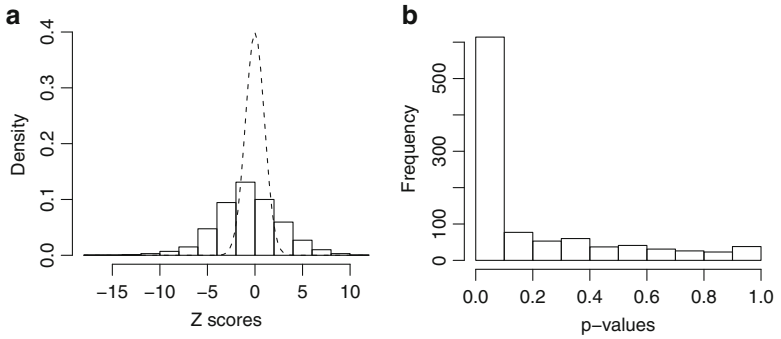

Fig. 4.5 (**a**) Z scores and (**b**) p-values, for 1,000 transcripts in the microarray data

$Y_i = \overline{Y}_{1i} - \overline{Y}_{0i}$ and let s_{ki}^2 be the sample variance in population k, for transcript i, $i = 1, \ldots, m$. We now assume

$$Y_i \mid \mu_i \sim_{iid} \mathrm{N}(\mu_i, \sigma_i^2)$$

where $\sigma_i^2 = s_{1i}^2/60 + s_{0i}^2/45$ is the sample variance, which is reliably estimated for the large sample sizes in the two populations and therefore assumed known. The null hypotheses of interest are that the difference in the average expression level between the two populations is zero. We let $H_i = 0$ correspond to the null for transcript i, that is, $\mu_i = 0$ for $i = 1, \ldots, m$. Figure 4.5a gives a histogram of the Z scores Y_i/σ_i, along with the reference $\mathrm{N}(0, 1)$ distribution. Clearly, unless there are problems with the model formulation, there are a large number of transcripts that are differentially expressed between the two populations, as confirmed by the histogram of p-values displayed in Fig. 4.5b.

4.6.1 Frequentist Analysis

In a single test situation we have seen that the historical emphasis has been on control of the type I error rate. We let $H_i = 0/1$ represent the hypotheses for the $i = 1, \ldots, m$ tests. In a multiple testing situation there are a variety of criteria that may be considered. With respect to Table 4.1, the *family-wise error rate* (FWER) is the probability of making *at least* one type I error, that is, $\Pr(B \geq 1 \mid H_1 = 0, \ldots, H_m = 0)$. Intuitively, this is a sensible criteria if one has a strong prior belief that all (or nearly all) of the null hypotheses are true, since in such a situation making at least one type I error should be penalized (this is made more concrete in Sect. 4.6.2). In contrast, if one believes that a number of the nulls are likely to be false, then one would be prepared to accept a greater number of type I errors, in exchange for discovering more true associations. As in all hypothesis testing situations, we want a method for trading off type I and type II errors.

Table 4.2 True FWER as a function of the correlation ρ between two bivariate normal test statistics

ρ	True FWER
0	0.0497
0.3	0.0484
0.5	0.0465
0.7	0.0430
0.9	0.0362

Let B_i be the event that the ith null is incorrectly rejected, so that, with respect to Table 4.1, B, the random variable representing the number of incorrectly rejected nulls, corresponds to $\cup_{i=1}^{m} B_i$. With a common level for each test $\alpha^\star$, the FWER is

$$\begin{aligned}\alpha_{\text{F}} = \Pr(B \geq 1 \mid H_1 = 0, \ldots, H_m = 0) &= \Pr\left(\cup_{i=1}^{m} B_i \mid H_1 = 0, \ldots, H_m = 0\right) \\ &\leq \sum_{i=1}^{m} \Pr(B_i \mid H_1 = 0, \ldots, H_m = 0) \\ &= m\alpha^\star. \end{aligned} \tag{4.13}$$

The *Bonferroni* method takes $\alpha^\star = \alpha_{\text{F}}/m$ to give FWER $\leq \alpha_{\text{F}}$. For example, to control the FWER at a level of $\alpha = 0.05$ with $m = 10$ tests, we would take $\alpha^\star = 0.05/10 = 0.005$. Since it controls the FWER, the Bonferroni method is stringent (i.e., conservative in the sense that the bar is set high for rejection) and so can result in a loss of power in the usual situation in which the FWER is set at a low value, for example 0.05. A little more conservatism is also introduced via the inequality, (4.13). The Sidák correction, which we describe shortly, overcomes this aspect.

If the test statistics are independent,

$$\begin{aligned}\Pr(B \geq 1) &= 1 - \Pr(B = 0) \\ &= 1 - \Pr\left(\cap_{i=1}^{m} B_i'\right) \\ &= 1 - \prod_{i=1}^{m} \Pr(B_i') \\ &= 1 - (1 - \alpha^\star)^m. \end{aligned}$$

Consequently, to achieve FWER $= \alpha_{\text{F}}$ we may take $\alpha^\star = 1 - (1 - \alpha_{\text{F}})^{1/m}$, the so-called Sidák correction (Sidák 1967).

With dependent tests, the Bonferroni approach is even more conservative; we demonstrate with $m = 2$ and bivariate normal test statistics with correlation ρ. Suppose we wish to achieve a FWER of 0.05. Table 4.2 gives the FWER achieved using Bonferroni and illustrates how the test becomes more conservative as the correlation increases. The situation becomes worse as m increases in size. The k-FWER criteria (Lehmann and Romano 2005) extends FWER to the incorrect rejection of k or more nulls (Exercise 4.2).

A simple remedy to the conservative nature of the control of FWER is to increase α_{F}. An intuitive measure to calibrate a procedure is via the expected number of false discoveries:

$$\begin{aligned}\text{EFD} &= m_0 \times \alpha^\star \\ &\leq m \times \alpha^\star\end{aligned}$$

where $\alpha^\star$ is the level for each test. If m_0 is close to m, this inequality will be practically useful. As an example, one could specify $\alpha^\star$ such that the EFD ≤ 1 (say), by choosing $\alpha^\star = 1/m$.

Recently there has been interest in a criterion that is particularly useful in multiple testing situations. We first define the false discovery proportion (FDP) as the proportion of incorrect rejections:

$$\text{FDP} = \begin{cases} \frac{B}{K} & \text{if } B > 0 \\ 0 & \text{if } B = 0. \end{cases}$$

Then the *false discovery rate* (FDR), the expected proportion of rejected nulls that are actually true, is

$$\text{FDR} = \text{E}[\,\text{FDP}\,] = \text{E}[\,B/K \mid B > 0\,]\,\Pr(B > 0).$$

Consider the following procedure for independent p-values, each of which is uniform under the null:

1. Let $P_{(1)} < \ldots < P_{(m)}$ denote the ordered p-values.
2. Define $l_i = i\alpha/m$ and $R = \max\{i : P_{(i)} < l_i\}$ where α is the value at which we would like FDR control.
3. Define the p-value threshold as $p_T = P_{(R)}$.
4. Reject all hypotheses for which $P_i \leq P_T$, that is, set $H_i = 1$ in such cases, $i = 1, \ldots, m$.

Benjamini and Hochberg (1995) show that if this procedure is applied, then regardless of how many nulls are true (m_0) and regardless of the distribution of the p-values when the null is false,

$$\text{FDR} \leq \frac{m_0}{m}\alpha < \alpha.$$

We say that the FDR is controlled at α.

Example: Hypothetical Data

We simulate data from $m = 100$ hypothetical tests in which $m_0 = 95$ tests are null, to give $m_1 = 5$ tests for which the alternative is true. Figure 4.6 displays the sorted observed $-\log_{10}(p\text{-values})$ versus the expected $-\log_{10}(p\text{-values})$, along

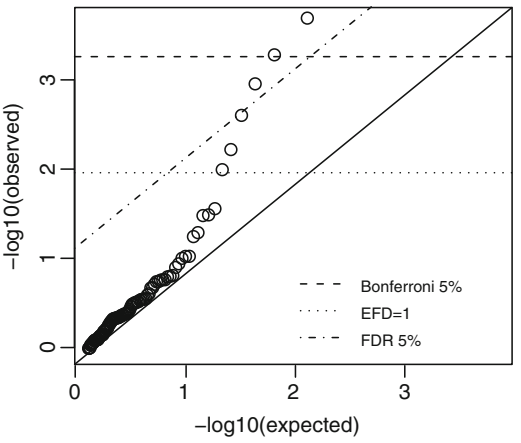

Fig. 4.6 Observed versus expected $-\log_{10}(p\text{-values})$ for a simulated set of data with 95 nulls and 5 alternatives. Three criteria for rejection, based on Bonferroni, the expected number of false discoveries (EFD), and the false discovery rate (FDR), are included on the plot

with a line of equality (solid line). Also displayed are three approaches to calling significance. The top dashed line corresponds to a Bonferroni correction at the 5% level (so that the line is at $-\log_{10}(0.05/100) = 3.30$). This criterion calls a single test as significant illustrating the conservative nature of the control of FWER at a low value. If we choose instead to control the expected number of false discoveries at 1, then the dotted line at $-\log_{10}(1/100) = 2$ results. We see that all 5 true alternatives are selected, along with a single false positive. Finally, we examine those hypotheses that would be rejected if we control the FDR at $\alpha = 0.05$, via the Benjamini–Hochberg procedure. On the log to the base 10 scale, the potential thresholds $l_i = i\alpha/m, i = 1, \ldots, m$ correspond to a line with slope 1 and intercept $-\log_{10}(\alpha)$. The dotted-dashed line gives the FDR threshold (recall the FDR is an expectation) corresponding to $\alpha = 0.05$. The use of this threshold gives three p-values as significant, for an empirical FDR of zero.

□

The algorithm of Benjamini and Hochberg (1995) begins with a desired FDR and then provides the p-value threshold. Storey (2002) proposed an alternative method by which, for any fixed rejection region, a criteria closely related to FDR, the *positive false discovery rate* $\text{pFDR} = \text{E}[B/K \mid K > 0]$, may be estimated. We assume rejection regions of the form $T > t_{\text{fix}}$ and consider the pFDR associated with regions of this form, which we write as $\text{pFDR}(t_{\text{fix}})$. We define, for $i = 1, \ldots, m$ tests, the random variables $H_i = 0/1$ corresponding to null/alternative hypotheses and test statistics T_i. Then, with $\pi_0 = \Pr(H = 0)$ and $\pi_1 = 1 - \pi_0$ independently for all tests,

$$\text{pFDR}(t_{\text{fix}}) = \frac{\Pr(T > t_{\text{fix}} \mid H = 0) \times \pi_0}{\Pr(T > t_{\text{fix}} \mid H = 0) \times \pi_0 + \Pr(T > t_{\text{fix}} \mid H = 1) \times \pi_1}.$$

Note the similarity with (4.6). Consideration of the false discovery odds:

$$\frac{\text{pFDR}(t_{\text{fix}})}{1-\text{pFDR}(t_{\text{fix}})} = \frac{\Pr(T > t_{\text{fix}} \mid H = 0)}{\Pr(T > t_{\text{fix}} \mid H = 1)} \times \frac{\pi_0}{\pi_1}$$

explicitly shows the weighted trade-off of type I and type II errors, with weights determined by the prior on the null/alternative; this expression mimics (4.7). Storey (2003) rigorously shows that

$$\text{pFDR}(t_{\text{fix}}) = \Pr(H = 0 \mid T > t_{\text{fix}}).$$

giving a Bayesian interpretation. In terms of p-values, the rejection region corresponding to $T > t_{\text{fix}}$ is of the form $[0, \gamma]$. Let P be the random p-value resulting from a test. Under the null, $P \sim \text{U}(0, 1)$, and so

$$\begin{aligned}\text{pFDR}(t_{\text{fix}}) &= \frac{\Pr(P \leq \gamma \mid H = 0) \times \pi_0}{\Pr(P \leq \gamma)} \\ &= \frac{\gamma \times \pi_0}{\Pr(P \leq \gamma)}. \end{aligned} \tag{4.14}$$

From this expression, the crucial role of π_0 is evident. Storey (2002) estimates (4.14), using uniformity of p-values under the null, to produce the estimates

$$\begin{aligned}\widehat{\pi}_0 &= \frac{\#\{p_i > \lambda\}}{m(1-\lambda)} \\ \widehat{\Pr}(P \leq \gamma) &= \frac{\#\{p_i \leq \gamma\}}{m} \end{aligned} \tag{4.15}$$

with λ chosen via the bootstrap to minimize the mean-squared error for prediction of the pFDR. The expression (4.15) calculates the empirical proportion of p-values to the right of λ and then inflates this to account for the proportion of null p-values in $[0, \lambda]$.

This method highlights the benefits of using the totality of p-values to estimate fundamental quantities of interest such as π_0. In general, information in all of the data may also be exploited, and in Sect. 4.6.2, we describe a Bayesian mixture model that uses the totality of data.

The *q-value* is the minimum FDR that can be attained when a particular test is called significant. We give a derivation of the q-value and, following Storey (2002), first define a set of nested rejection regions $\{t_\alpha\}_{\alpha=0}^{1}$ where α is such that $\Pr(T > t_\alpha \mid H = 0) = \alpha$. Then

$$p\text{-value}(t) = \inf_{t_\alpha : t \in t_\alpha} \Pr(T > t_\alpha \mid H = 0)$$

is the p-value corresponding to an observed statistic t. The q-value is defined as

$$q\text{-value}(t) = \inf_{t_\alpha : t \in t_\alpha} \Pr(H = 0 \mid T > t_\alpha). \tag{4.16}$$

Therefore, for each observed statistic t_i, there is an associated q-value. It can be shown that (Exercise 4.3)

$$\Pr(H_0 \mid T > t_{\text{obs}}) < \Pr(H_0 \mid T = t_{\text{obs}}) \tag{4.17}$$

so that the evidence for H_0 given the exact ordinate is always greater than that corresponding to the tail area.

When one decides upon a value of FDR (or pFDR) to use in practice, the sample size should again be taken into account, since for large sample size, one would not want to tolerate as large an FDR as with a small sample size. Again, we would prefer a procedure that was consistent. However, as in the single test situation, there is no prescription for deciding how the FDR should decrease with increasing sample size.

Example: Microarray Data

Returning to the microarray example, application of the Bonferroni correction to control the FWER at 0.05 produces a list of 220 significant transcripts. In this context, it is likely that there are a large proportion of non-null transcripts (Storey et al. 2007) and there are relatively large sample sizes for each test (so the power is good), and so this choice is likely to be very conservative. The procedure of Benjamini and Hochberg with FDR control at 0.05 gives 480 significant transcripts. Applying the method of Storey gives an estimate of the proportion of nulls as $\widehat{\pi}_0 = 0.33$. At a pFDR threshold of 0.05, 603 transcripts are highlighted.

4.6.2 Bayesian Analysis

In some situations, a Bayesian analysis of m tests may proceed in exactly the same fashion as with a single test, that is, one can apply the same procedure m times; see Wakefield (2007a) for an example. In this case the priors on each of the m null hypotheses will be independent. In other situations, however, one may often wish to jointly model the data so that the totality of information can be used to estimate parameters that are common to all tests.

In terms of reporting, as with a single test (as considered in Sect. 4.3), the Bayes factors

$$\text{Bayes Factor}_i = \frac{p(\boldsymbol{y}_i \mid H_i = 0)}{p(\boldsymbol{y}_i \mid H_i = 1)}, \tag{4.18}$$

$i = 1, \ldots, m$ are a starting point. These Bayes factors may then be combined with prior probabilities $\pi_{0i} = \Pr(H_i = 0)$, to give

$$\text{Posterior Odds}_i = \text{Bayes Factor}_i \times \text{Prior Odds}_i, \tag{4.19}$$

where $\text{Prior Odds}_i = \pi_{0i}/(1 - \pi_{0i})$.

Proceeding to a decision theory approach. Suppose for simplicity common losses, L_{I} and L_{II}, associated with type 1 and type 2 errors, for each test. The aim is to define a rule for deciding which of the m null hypotheses to reject. The operating characteristics, in terms of "false discovery" and "non-discovery," corresponding to this rule may then be determined. The loss associated with a particular set of decisions $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_m]$ and hypotheses $\boldsymbol{H} = [H_1, \ldots, H_m]$ is the expectation over the posterior

$$\begin{aligned}\mathrm{E}[L(\boldsymbol{\delta}, \boldsymbol{H})] &= L_{\text{I}} \sum_{i=1}^{m} \left[\delta_i \Pr(H_i = 0 \mid \boldsymbol{y}_i) + \frac{L_{\text{II}}}{L_{\text{I}}} (1 - \delta_i) \Pr(H_i = 1 \mid \boldsymbol{y}_i) \right] \\ &= L_{\text{I}} \left[\text{EFP} + \frac{L_{\text{II}}}{L_{\text{I}}} \times \text{EFN} \right]\end{aligned}$$

where EFP is the expected number of false positives and EFN is the expected number of false negatives. These characteristics of the procedure are given, respectively, by

$$\begin{aligned}\text{EFD} &= \sum_{i=1}^{m} \delta_i \Pr(H_i = 0 \mid \boldsymbol{y}_i) \\ \text{EFN} &= \sum_{i=1}^{m} (1 - \delta_i) \Pr(H_i = 1 \mid \boldsymbol{y}_i),\end{aligned}$$

where $\Pr(H_i = 0 \mid \boldsymbol{y}_i)$ and $\Pr(H_i = 1 \mid \boldsymbol{y}_i)$ are the posterior probabilities on the null and alternative. We should report test i as significant if

$$\Pr(H_i = 1 \mid \boldsymbol{y}_i) \geq \frac{1}{1 + L_{\text{II}}/L_{\text{I}}},$$

which is identical to the expression derived for a single test, (4.4).

Define $K = \sum_{i=1}^{m} \delta_i$ as the number of rejected tests. Then dividing EFD by K gives an estimate, based on the posterior, of the proportion of false discoveries, and dividing EFN by $m - K$ gives a posterior estimate of the proportion of false non-discoveries. Hence, for a given ratio of losses, we can determine the expected number of false discoveries and false non-discoveries, and the FDR and FNR. As n_i, the sample size associated with test i, increases, under correct specification of the model, the power for each test increases, and so EFD/K and $\text{EFN}/(m-K)$ will tend to zero (assuming the model is correct). This is in contrast to the frequentist approach in which a fixed (independent of sample size) FDR rule is used so that the false non-discovery rate does not decrease to zero (even when the model is true).

Notice that the use of Bayes factors does not depend on the number of tests, m, so that, for example, we could analyze the data in the same way regardless of whether m is 1 or 1,000,000. Similarly, for the assumed independent priors, the posterior probabilities do not depend on m, and for the loss structure considered, the decision

does not depend on m. Hence, the Bayes procedure gives thresholds that depend on n (since the Bayes factor will depend on sample size, see Exercise 4.1 for an example) but not on m, while the contrary is true for many frequentist procedures such as Bonferroni.

There is a prior that results in a Bayesian Bonferroni-type correction. If the prior probabilities of each of the nulls are independent with $\pi_{0i} = \pi_0$ for $i = 1, \ldots, m$. Then the prior probability that all nulls are true is

$$\Pi_0 = \Pr(H_1 = 0, \ldots, H_m = 0) = \pi_0^m$$

which we refer to as prior P_1. For example, if $\pi_0 = 0.5$ and $m = 10$, $\Pi_0 = 0.00098$, which may not reflect the required prior belief. Suppose instead that we wish to fix the prior probability that all of the nulls are true at Π_0. A simple way of achieving this is to take $\pi_{0i} = \Pi_0^{1/m}$, a prior specification we call P_2. Westfall et al. (1995) show that for independent tests

$$\alpha_{\text{B}} = \Pr(H_i = 0 \mid \boldsymbol{y}_i, P_2) \approx m \times \Pr(H_i = 0 \mid \boldsymbol{y}_i, P_1) = m \times \alpha_{\text{B}}^{\star}$$

so that a Bayesian version of Bonferroni is recovered.

An alternative approach is to specify a full model for the totality of data. These data can then be exploited to estimate common parameters. In particular, the proportion of null tests π_0 can be estimated, which is crucial for inference since posterior odds and decisions are (unsurprisingly) highly sensitive to the value of π_0. The decision is still based on the posterior, and there continues to be a trade-off between false positive and false negatives depending on the decision threshold used. We illustrate using the microarray data.

Example: Microarray Data

Recall that we assume $Y_i \mid \mu_i \sim_{ind} \text{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, m$ where $m = 1{,}000$. We first describe a Bayesian analysis in which the m transcripts are analyzed separately. We assume under the null that $\mu_i = 0$, while under the alternative $\mu_i \sim_{iid} \text{N}(0, \tau^2)$ with τ^2 fixed. For illustration, we assume that for non-null genes, a fold change in the mean greater than 10%, that is, $\log_2 \mu_i > 0.138$, only occurs with probability 0.025. Given

$$\Pr\left(-\infty < \frac{\mu_i}{\tau} < \frac{\log_2(1.1)}{\tau}\right) = 0.975$$

we can solve for τ to give

$$\tau = \frac{\log_2(1.1)}{\Phi^{-1}(0.975)} = 0.070,$$

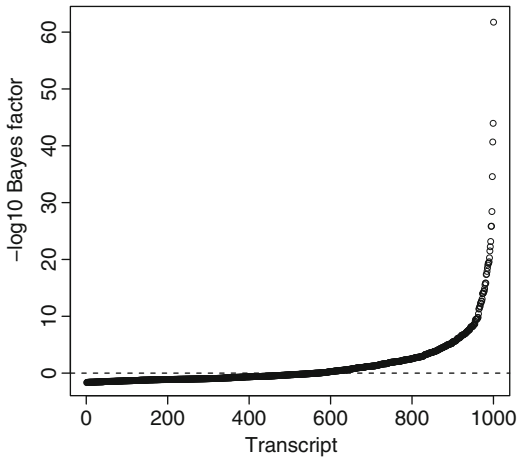

Fig. 4.7 Ordered $-\log_{10}$(Bayes factors) for the microarray data. The *dashed line* at 0 is for reference

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable. The prior on μ_i is therefore

$$\mu_i = \begin{cases} 0 & \text{with probability } \pi_0 \\ \mathrm{N}(0, 0.138^2) & \text{with probability } \pi_1 = 1 - \pi_0 \end{cases}.$$

The Bayes factor for the ith transcript is

$$\text{Bayes Factor}_i = \sqrt{\frac{\sigma_i^2 + \tau^2}{\tau^2}} \exp\left[-\frac{Z_i^2}{2}\frac{\tau^2}{\sigma_i^2 + \tau^2}\right] \tag{4.20}$$

where $Z_i = Y_i/\sigma_i$ is the Z score for the ith transcript. Therefore, we see that the Bayes factor depends on the power through σ_i^2 (which itself depends on the sample sizes), as well as on the Z-score, while the p-value depends on the latter only. In Fig. 4.7, we plot the ordered $-\log_{10}($ Bayes factors $)$ (so that high values correspond to evidence against the null). A reference line of 0 is indicated and, using this reference, for 487 transcripts the data are more likely under the alternative than under the null.

To obtain the posterior odds, we need to specify a prior for the null. We assume $\pi_0 = \Pr(H_i = 0)$ so that the prior is the same for all transcripts. The posterior odds are the product of the Bayes factor and the prior odds and are highly sensitive to the choice of π_0. For illustration, suppose the decision rule is to call a transcript significant if the posterior odds of $H = 0$ are less than 1 (which corresponds to a ratio of losses, $L_{\text{II}}/L_{\text{I}} = 1$). Figure 4.8 plots the number of such significant transcripts under this rule, as a function of the prior, π_0. The sensitivity to the choice of π_0 is evident. To overcome this problem, we now describe a joint model for the data on all $m = 1{,}000$ transcripts that allows estimation of parameters that are common across transcripts, including π_0. Notice that for virtually the complete range of π_0 more transcripts would be called as significant under the Bayes rule than under the FWER.

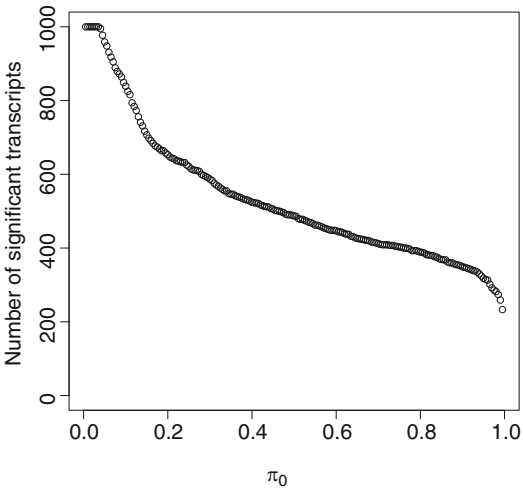

Fig. 4.8 Number of significant transcripts in the microarray data, as measured by the posterior probability on the null $\Pr(H_i = 0 \mid \boldsymbol{y}) < 0.5$, $i = 1, \ldots, 1,000$, as a function of the common prior probability on the null, $\pi_0 = \Pr(H_i = 0)$

We specify a mixture model for the collection $[\mu_1, \ldots, \mu_m]$, with

$$\mu_i = \begin{cases} 0 & \text{with probability } \pi_0 \\ \mathrm{N}(\delta, \tau^2) & \text{with probability } \pi_1 = 1 - \pi_0 \end{cases}.$$

We use mixture component indicators $H_i = 0/1$ to denote the zero/normal membership model for transcript i. Collapsing over μ_i gives the three-stage model:

Stage One:

$$Y_i \mid H_i, \delta, \tau, \pi_0 \sim_{ind} \begin{cases} \mathrm{N}(0, \sigma_i^2) & \text{if } H_i = 0 \\ \mathrm{N}(\delta, \sigma_i^2 + \tau^2) & \text{if } H_i = 1. \end{cases}$$

Stage Two: $H_i \mid \pi_1 \sim_{iid} \text{Bernoulli}(\pi_1)$, $i = 1, \ldots, m$.

Stage Three: Independent priors on the common parameters:

$$p(\delta, \tau, \pi_0) = p(\delta)p(\tau)p(\pi_0).$$

We illustrate the use of this model with

$$\begin{aligned} p(\delta) &\propto 1, \\ p(\tau) &\propto 1/\tau \\ p(\pi_0) &= 1, \end{aligned}$$

so that we have improper priors for δ and τ^2. The latter choice still produces a proper posterior because we have fixed variances at the first stage of the model (see Sect. 8.6.2 for further discussion). Implementation is via a Markov chain Monte Carlo algorithm (see Sect. 3.8). Exercise 4.4 derives details of the algorithm.

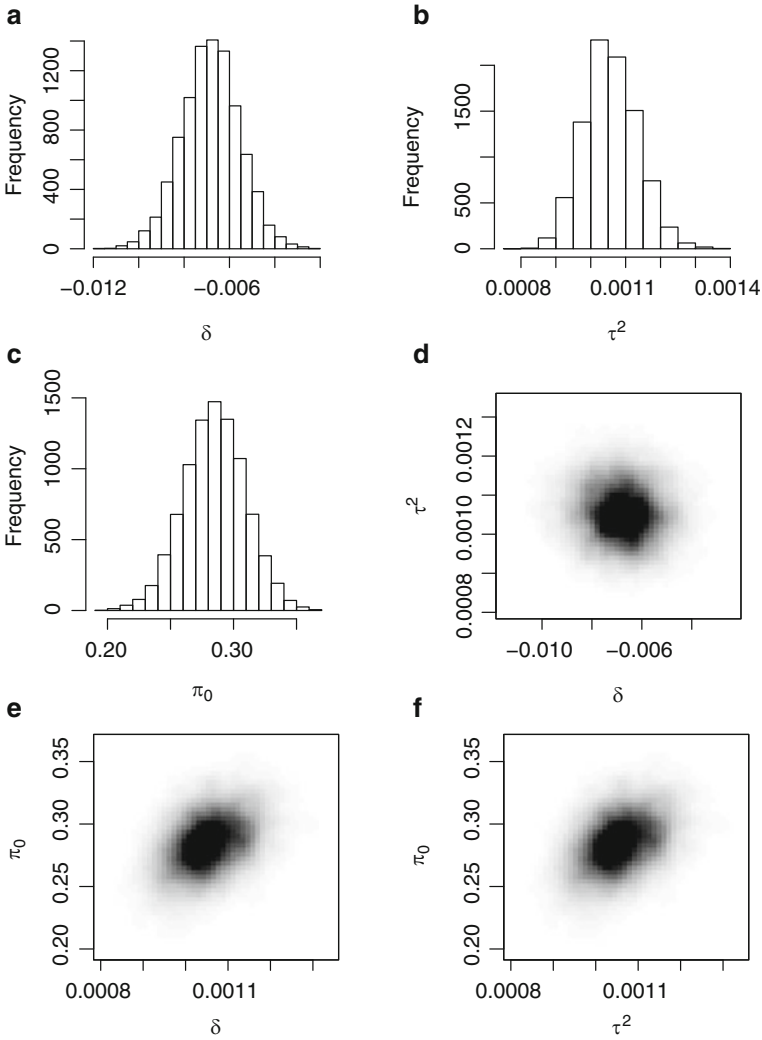

Fig. 4.9 Posterior distributions for selected parameters of the mixture model, for the microarray data: **(a)** $p(\delta \mid \boldsymbol{y})$, **(b)** $p(\tau^2 \mid \boldsymbol{y})$, **(c)** $p(\pi_0 \mid \boldsymbol{y})$, **(d)** $p(\delta, \tau^2 \mid \boldsymbol{y})$, **(e)** $p(\delta, \pi_0 \mid \boldsymbol{y})$, **(f)** $p(\tau^2, \pi_0 \mid \boldsymbol{y})$

The posterior median and 95% interval for δ ($\times 10^{-3}$) is $-6.8\,[-9.4, -0.40]$, while for τ^2 ($\times 10^{-3}$), we have 1.1 [0.92,1.2]. Of more interest are the posterior summaries for π_0: 0.29 [0.24,0.33], giving a range that is consistent with the pFDR estimate of 0.33. Figure 4.9 displays univariate and bivariate posterior distributions. The distributions resemble normal distributions, reflecting the large samples within populations and the number of transcripts.

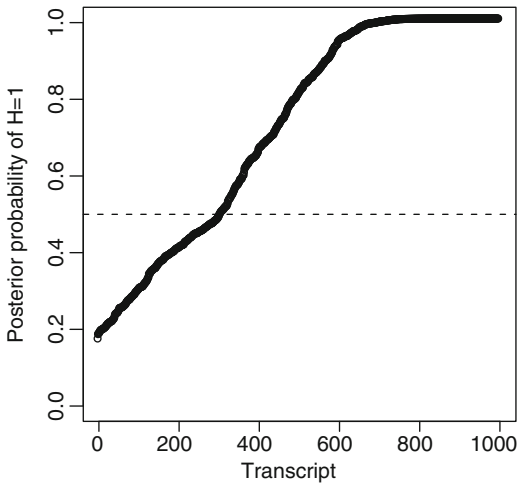

Fig. 4.10 Posterior probabilities $\Pr(H_i = 1 \mid \boldsymbol{y})$, from the mixture model for the microarray data, for each of the $i = 1, \ldots, 1{,}000$ transcripts, ordered in terms of increasing posterior probability on the alternative

For transcript i, we may evaluate the posterior probabilities of the alternative

$$\begin{aligned}\Pr(H_i = 1 \mid y_i) &= \mathrm{E}[H_i \mid \boldsymbol{y}] \\ &= \mathrm{E}_{\delta,\tau^2,\pi_0|y}\left[\Pr(H_i \mid \delta, \tau^2, \pi_0)\right] \\ &= \mathrm{E}_{\delta,\tau^2,\pi_0|y}\left[\Pr(H_i = 1 \mid \boldsymbol{y}, \delta, \tau^2, \pi_0)\right] \\ &= \mathrm{E}_{\delta,\tau^2,\pi_0|y}\left[\frac{p(\boldsymbol{y} \mid H_i = 1, \delta, \tau^2) \times \pi_1}{p(\boldsymbol{y} \mid H_i = 1, \delta, \tau^2) \times \pi_1 + p(\boldsymbol{y} \mid H_i = 0) \times \pi_0}\right]\end{aligned} \tag{4.21}$$

where

$$p(\boldsymbol{y} \mid H_i = 1, \delta, \tau^2, \pi_0) = [2\pi(\sigma_i^2 + \tau^2)]^{-1/2} \exp\left[-\frac{(y_i - \delta)^2}{2(\sigma_i^2 + \tau^2)}\right]$$

$$p(\boldsymbol{y} \mid H_i = 0, \delta, \tau^2, \pi_0) = [2\pi\sigma_i^2]^{-1/2} \exp\left[-\frac{y_i^2}{2\sigma_i^2}\right].$$

Expression (4.21) averages $\Pr(H_i = 1 \mid \boldsymbol{y}, \delta, \tau^2, \pi_0)$ with respect to the posterior $p(\delta, \tau^2, \pi_0 \mid \boldsymbol{y})$ and may be simply evaluated via

$$\frac{1}{T}\sum_{t=1}^{T} \frac{p(\boldsymbol{y} \mid H_i = 1, \delta^{(t)}, \tau^{2(t)})\pi_1^{(t)}}{p(\boldsymbol{y} \mid H_i = 1, \delta^{(t)}, \tau^{2(t)}, \pi_0^{(t)})\pi_1^{(t)} + p(\boldsymbol{y} \mid H_i = 0)\pi_0^{(t)}}$$

given samples $\delta^{(t)}, \tau^{2(t)}, \pi_0^{(t)}$, $t = 1, \ldots, T$, from the Markov chain.

Figure 4.10 displays the ordered posterior probabilities, $\Pr(H_i = 1 \mid \boldsymbol{y})$, $i = 1, \ldots, m$, along with a reference line of 0.5. Using this line as a threshold, 689 transcripts are flagged as "significant," and the posterior estimate of the proportion

of false discoveries is 0.12. Interestingly, the posterior estimate of the proportion of false negatives (i.e., non-discoveries) is 0.35. The latter figure is rarely reported but is a useful summary. Previously, using a pFDR threshold of 0.05, there were 603 significant transcripts. Interestingly, using a rule that picked the 603 transcripts whose posterior probability on the alternative was highest yielded an estimate of the posterior probability of the proportion of false discoveries as 0.07, which is not very different from the pFDR estimate. This is reassuring for both the Bayesian and the pFDR approaches.

For this example, sensitivity analyses might relax the independence between transcripts and, more importantly, the normality assumption for the random effects μ_i.

The Bayes factor, (4.20), was derived under the assumption of a normal sampling likelihood. In general, if we have large sample sizes, we may take as likelihood the sampling distribution of an estimator and combine this with a normal prior, to give a closed-form estimator. The latter is an approximation to a Bayesian analysis with weakly informative priors on the nuisance parameters and was described in Sect. 3.11, with Bayes factor (3.45).

4.7 Testing Multiple Hypotheses: Variable Selection

A ubiquitous issue in regression modeling is deciding upon which covariates to include in the model. It is useful to distinguish three scenarios:

1. *Confirmatory:* In which a summary of the strength of association between a response and covariates is required. We include in this category the situation in which an a priori hypothesis concerning a particular response/covariate relationship is of interest; additional variables have been measured and we wish, for example, to know which to adjust for in order to reduce confounding.
2. *Exploration:* In which the aim is to gain clues about structure in the data. A particular example is when one wishes to gain leads as to which covariates are associated with a response, perhaps to guide future study design.
3. *Prediction:* In which we are not explicitly concerned with association but merely with predicting a response based on a set of covariates. In this case, we are not interested in the numerical values of parameters but rather in the ability to predict new outcomes. Chapters 10–12 examines prediction in detail, including the assessment of predictive accuracy.

For exploration, formal inference is not required and so we will concentrate on the confirmatory scenario. As we will expand upon in Sect. 5.9, a trade-off must be made when deciding on variables for inclusion and it is often not desirable to fit the full model. To summarize the discussion, as we include more covariates in the model, bias in estimates is reduced, but variability may be increased, depending on how strong a predictor the covariate is and on its association with other covariates.

Example: Prostate Cancer

To illustrate a number of the methods available for variable selection, we consider a dataset originally presented by Stamey et al. (1989) and introduced in Sect. 1.3.1. The data were collected on $n = 97$ men before radical prostatectomony. We take as response the log of prostate-specific antigen (PSA) which was being forwarded in the paper as a preoperative marker, that is, a predictor of the clinical stage of cancer. The authors examined log PSA as a function of eight covariates: log(can vol); log(weight) (where weight is prostate weight); age; log(BPH); SVI; log(cap pen); the Gleason score, referred to as gleason; and percentage Gleason score 4 or 5, referred to as PGS45.

Figure 1.1 shows the relationships between the response and each of the covariates and indicates what look like a number of strong associations, while Fig. 1.2 gives some idea of the dependencies among the more strongly associated covariates. After Sect. 4.9, we will return to this example, after describing a number of methods for selecting variables in Sect. 4.8 and discussing model uncertainty in Sect. 4.9.

4.8 Approaches to Variable Selection and Modeling

We now review a number of approaches to variable selection. Let k be the number of covariates, and for ease of exposition, assume each covariate is either binary or continuous, so that the association is summarized by a univariate parameter. We also exclude interactions so that the largest model contains $k + 1$ regression coefficients. Allowing for the inclusion/exclusion of each covariate only, there are 2^k possible models, a number which increases rapidly with k. For example, with $k = 20$ there are $1,048,576$ possible models. The number of models increases even more rapidly with the number of covariates, if we allow variables with more than two levels and/or interactions.

The hierarchy principle states that if an interaction term is included in the model, then the constituent main effects should be included also. If we do not apply the hierarchy principle, there are 2^{2^k-1} *interaction* models (i.e., models that include main effects and/or interactions), where k is the number of variables. For example, $k = 2$ leads to 8 models. Denoting the variables by A and B, these models are

$$1,\ A,\ B,\ A+B,\ A+B+A.B,\ A+A.B,\ B+A.B,\ A.B.$$

The class of hierarchical models includes all models that obey the hierarchy principle. Applying the hierarchy principle in the $k = 2$ case reduces the number from 8 to 5, as we lose the last three models in the above list. With $k = 5$ variables, there are 2,147,483,648 interaction models, illustrating the sharp increase in the number of models with k. There is no general rule for counting the number

of models that satisfy the hierarchy principle for a given dimension. For some discussion, see Darroch et al. (1980, Sect. 6). The latter include a list of the number of hierarchical models for $k = 1, \ldots, 5$; for $k = 5$, the number of hierarchical models is 7,580.

We begin by illustrating the problems of variable selection with a simple example.

Example: Confounder Adjustment

Suppose the true model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{4.22}$$

with $\epsilon_i \mid \sigma^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, n$. We take x_1 as the covariate of interest, so that estimation of β_1 is the focus. However, we decide to "control" for the possibility of $\beta_2 \neq 0$ via a test. For simplicity, we assume that σ^2 is known and assess significance by examining whether a 95% confidence interval for β_2 contains zero (which is equivalent to a two-sided hypothesis test with $\alpha = 0.05$). If the interval contains zero, then the model,

$$\mathrm{E}[Y_i \mid x_{1i}, x_{2i}] = \beta_0^* + \beta_1^* x_{1i},$$

is fitted; otherwise, we fit (4.22). We illustrate the effects of this procedure through a simulation in which we take $\beta_0 = \beta_1 = \beta_2 = 1$, $\sigma^2 = 3^2$, and $n = 10$. The covariates x_1 and x_2 are simulated from a bivariate normal with means zero, variances one and correlation 0.7.

In Fig. 4.11a, we display the sampling distribution of $\widehat{\beta}_1$ given the fitting of model (4.22). The mean and standard deviation of the distribution of $\widehat{\beta}_1$ are 1.00 and 1.23, respectively. Unbiasedness follows directly from least squares/likelihood theory (Sect. 5.6).

Figure 4.11b displays the sampling distribution of the *reported* estimator when we allow for the possibility of adjustment according to a test of $\beta_2 \neq 0$. The mean and standard deviation of the distribution of the reported estimator of β_1 are 1.23 and 1.01, respectively, showing positive bias and a reduced variance. This distribution is a mixture of the sampling distribution of $\widehat{\beta}_1$ (the estimator obtained from the full model), and the sampling distribution of $\widehat{\beta}_1^\star$, with the mixing weight on the latter corresponding to one minus the power of the test of $\beta_2 = 0$. The sampling distribution of $\widehat{\beta}_1^\star$ is shifted because the effects of both x_1 and x_2, are being included in the estimate and the distribution is shifted to the right because x_1 and x_2 are positively correlated. Using the conditional mean of a bivariate normal (given as (D.1) in Appendix D) we have

$$\begin{aligned} \mathrm{E}[Y \mid x_1] &= \beta_0 + \beta_1 x_1 + \beta_2 \mathrm{E}[X_2 \mid x_1] \\ &= \beta_0 + (\beta_1 + 0.7) \times x_1 \\ &= \beta_0^\star + \beta_1^\star x_1 \end{aligned}$$

illustrating the bias,

$$\mathrm{E}[\widehat{\beta}_1^\star] - \beta_1 = 0.7 \tag{4.23}$$

when the reduced model is fitted. Allowing for the possibility of adjustment gives an estimator with a less extreme bias, since sometimes the full model is fitted (if the null is rejected). The reason for the lower *reported* variance in the potentially adjusted analysis is the bias-variance trade-off intrinsic to variable selection. In model (4.22), the information concerning β_1 and β_2 is entangled because of the correlation between x_1 and x_2, which results in a higher variance. Section 5.9 provides further discussion. The reported variance is not appropriate, however, since it does not acknowledge the model building process, an issue we examine in Sect. 4.9. As $n \to \infty$, the power of the test to reject $\beta_2 = 0$ tends to 1, and we recover an unbiased estimator with an appropriate variance.

□

4.8.1 Stepwise Methods

A number of methods have been proposed that proceed in a stepwise fashion, adding or removing variables from a current model. We describe three of the most historically popular approaches.

Forward selection begins with the null model, $\mathrm{E}[Y \mid \boldsymbol{x}] = \beta_0$, and then fits each of the models

$$\mathrm{E}[Y \mid \boldsymbol{x}] = \beta_0 + \beta_j x_j,$$

$j = 1, \ldots, k$. Subject to a minimal requirement (i.e., a particular p-value threshold), the model that contains the covariate that provides the greatest "improvement" in fit is then carried forward. This procedure is then iterated until no covariates meet the minimal requirement (i.e., all the p-values are greater than the threshold), or all the variables are in the model.

Backward elimination has the same flavor but begins with the full model, and then removes, at each stage, the covariate that is contributing least to the fit. For example, the variable with the largest p-value, so long as it is bigger than some prespecified value, is removed from the model.

Each of these approaches can miss important models. For example, in forward selection, x_1 may be the "best" single variable, but x_1 and any other variable may be "worse" than x_2 and x_3 together (say), but the latter combination will never be considered. Related problems can occur with backward elimination. Such considerations lead to *Efroymson's algorithm* (Efroymson 1960) in which forward selection is followed by backward elimination. The initial steps are identical to

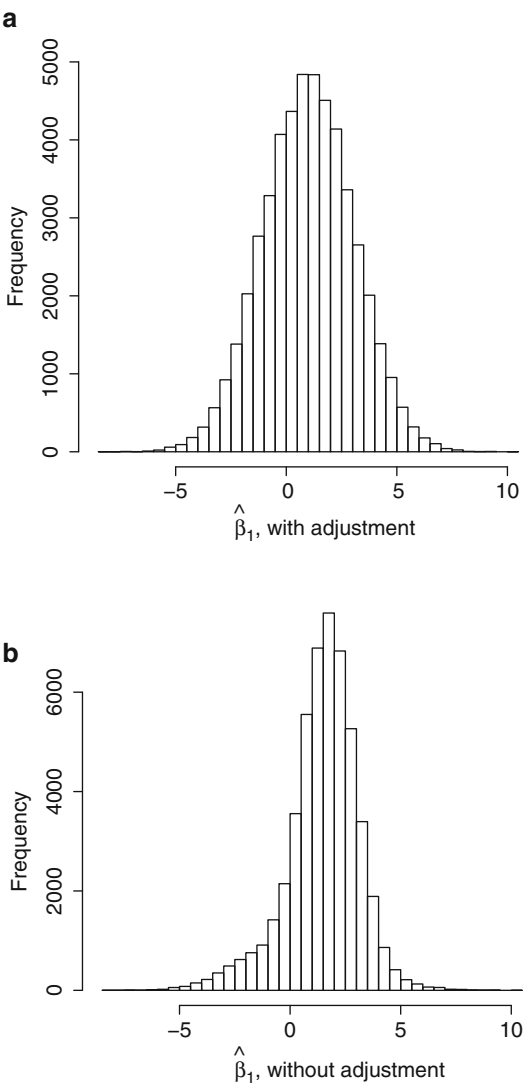

Fig. 4.11 (**a**) Sampling distribution of $\widehat{\beta}_1$, controlling for x_2, and (**b**) sampling distribution of $\widehat{\beta}_1$, given the possibility of controlling for x_2

forward selection, but with three or more variables in the model, the loss of fit of each of the variables (excluding the last one added) is examined, in order to avoid the scenario just described, since in this case if the order of variables being added was x_1, x_2, x_3, it would then be possible for x_1 to be removed. The "p-value to enter" value (i.e., the threshold for forward selection) is chosen to be smaller than the "p-value to remove" value (i.e., the threshold for backward elimination), to prevent cycling in which a variable is continually added and then removed. The choice of inclusion/exclusion values is contentious for forward selection, backward elimination and Efroymson's algorithm.

The Efroymson procedure, although overcoming some of the deficiencies of forward selection and backwards elimination, can still miss important models. The overall frequentist properties of any subset selection approach are difficult to determine, as we discuss in Sect. 4.9.

Each of the stepwise approaches may miss important models. A popular alternative is to examine all possible models and to then select the "best" model. We next provide a short summary of some of the criteria that have been suggested for this selection.

4.8.2 All Possible Subsets

We first consider linear models and again suppose there are k potential regressors, with the full model of the form

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.24}$$

with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \mathrm{I}_n$, and where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{x}$ is $n \times (k+1)$, and $\boldsymbol{\beta}$ is $(k+1) \times 1$.

The R^2 measure of variance explained is

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{CTSS}}$$

where the residual and corrected total sum of squares are given, respectively, by

$$\begin{aligned} \mathrm{RSS} &= (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}) \\ \mathrm{CTSS} &= (\boldsymbol{y} - \mathbf{1}\overline{y})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{1}\overline{y}). \end{aligned}$$

Consequently, R^2 can be interpreted as measuring the closeness of the fit to the data, with $R^2 = 1$ for a perfect fit ($\mathrm{RSS} = 0$) and $R^2 = 0$ if the model does not improve upon the intercept only model. In terms of a comparison of nested models, the R^2 measure is nondecreasing in the number of variables, and so picking the model with the smallest R^2 will always produce the full model.

Let P represent a model constructed from covariates whose indices are a subset of $\{1, 2, \ldots, k\}$, with $p = |P| + 1$ regression coefficients in this model. The number of parameters p accounts for the inclusion of an intercept so that in the full model $p = k + 1$. Suppose the fit of model P yields estimator $\widehat{\boldsymbol{\beta}}_P$ and residual sum of squares RSS_P. For model comparison, a more useful measure than R^2 is the adjusted R^2 which is defined as

$$\begin{aligned} R^2_a &= 1 - \frac{\mathrm{RSS}_P/(n-p)}{\mathrm{CTSS}/(n-1)} \\ &= 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right). \end{aligned}$$

Maximization of R_a^2 leads to the model that produces the smallest estimate of σ^2 across models.

A widely used statistic, known as Mallows C_P, was introduced by Mallows (Mallows 1973).[1] For the model associated with the subset P

$$C_P = \frac{\text{RSS}_P}{\widehat{\sigma}^2} - (n - 2p) \tag{4.25}$$

where $\text{RSS}_P = (\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_P)^{\text{T}}(\boldsymbol{y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_P)$ is the residual sum of squares and $\widehat{\sigma}^2 = \text{RSS}_k/(n-k-1)$ is the error variance estimate from the full model that contains all k covariates. This criteria may be derived via consideration of the prediction error that results from choosing the model under consideration (as we show in Sect. 10.6.1). It is usual to plot C_P versus p and for a good model C_P will be close to, or below, p, since $\text{E}[\text{RSS}_P] = (n - p)\sigma^2$ and so $\text{E}[C_P] = p$ for a good model.

Lindley (1968) showed that Mallows C_P can also be derived from a Bayesian decision approach to multiple regression in which, among other assumptions, the aim is prediction and the X's are random and multivariate normal.

We now turn to more general models than (4.24). Consideration of the likelihood alone is not useful since the likelihood increases as parameters are added to the model, as we saw with the residual sum of squares in linear models. A number of *penalized likelihood* statistics have been proposed that penalize models for their complexity. A large number of statistics have been proposed, but we concentrate on just two, AIC and BIC. *An Information Criteria* (AIC, Akaike 1973) is a generalization of Mallows C_P and is defined as

$$\text{AIC} = -2l(\widehat{\boldsymbol{\beta}}_P) + 2p \tag{4.26}$$

where $l(\widehat{\boldsymbol{\beta}}_P)$ denotes the maximized log-likelihood of, and p the number of parameters in, model P. A derivation of AIC is presented in Sect. 10.6.5. We have already encountered the Bayesian information criterion (BIC) in Sect. 3.10 as an approximation to a Bayes factor. The BIC is given by

$$\text{BIC} = -2l(\widehat{\boldsymbol{\beta}}_P) + p \log n.$$

For the purposes of model selection, one approach is to choose between models by selecting the one with the minimum AIC or BIC. In general, BIC penalizes larger models more heavily than AIC, so that in practice AIC tends to pick models that are more complicated. As an indication, for a single parameter ($p = 1$ in (4.26)), the significance level is $\alpha = 0.157$ corresponding to $\Pr(\chi_1^2 < 2)$, which is a very liberal threshold. Given regularity conditions, BIC is *consistent* (Haughton 1988, 1989; Rao and Wu 1989), meaning if the correct model is in the set being considered, it will be picked with a probability that approaches 1 with increasing sample size,

[1] Named in honor of Cuthbert Daniel with whom Mallows initially discussed the use of the C_P statistic.

while AIC is not. The appearance of n in the penalty term of BIC is not surprising, since this is required for consistency.

4.8.3 Bayesian Model Averaging

Rather than select a single model, Bayesian model averaging (BMA) places priors over the candidate models, and then inference for a function of interest is carried out by averaging over the posterior model probabilities. Section 3.6 described this approach in detail, and we will shortly demonstrate its use with the prostate cancer data.

4.8.4 Shrinkage Methods

An alternative approach to selecting a model is to consider the full model but to allow shrinkage of the least squares estimates. Ridge regression and the lasso fit within this class of approaches and are considered in detail in Sects. 10.5.1 and 10.5.2, respectively. Such methods are often used in situations in which the data are sparse (in the sense of k being large relative to n).

4.9 Model Building Uncertainty

If a single model is selected on the basis of a stepwise method or via a search over all models, then bias will typically result. Interval estimates, whether they be based on Bayesian or frequentist approaches, will tend to be too narrow since they are produced by conditioning on the final model and hence do not reflect the mechanism by which the model was selected; see Chatfield (1995) and the accompanying discussion.

To be more explicit, let P denote the procedure by which a final model M is selected, and suppose it is of interest to examine the properties of an estimator $\widehat{\phi}$ of a univariate parameter ϕ, for example, a regression coefficient associated with a covariate of interest. The usual frequentist unbiasedness results concern the expectation of an estimator within a fixed model. We saw an example of bias following variable selection, with the bias given by (4.23). In general, the estimator obtained from a selection procedure will not be unbiased with respect to the final model chosen, that is,

$$\mathrm{E}[\widehat{\phi} \mid P] = \mathrm{E}_{M|P}[\,\mathrm{E}(\widehat{\phi} \mid M)\,] \tag{4.27}$$

$$\neq \mathrm{E}(\widehat{\phi} \mid \widehat{M}), \tag{4.28}$$

where $\widehat{M}$ is the final model chosen. In addition,

$$\text{var}(\widehat{\phi} \mid P) = \text{E}_{M|P}[\text{var}(\widehat{\phi} \mid M)] + \text{var}_{M|P}(\text{E}[\widehat{\phi} \mid M]) \tag{4.29}$$

$$\neq \text{var}(\widehat{\phi} \mid \widehat{M}) \tag{4.30}$$

where the latter approximates the first term of (4.29) only. Hence, in general, the reported variance conditional on a chosen model will be an underestimate. The bias and variance problems arise because the procedure by which $\widehat{M}$ was chosen is not being acknowledged.

From a Bayesian standpoint, the same problem exists because the posterior distribution should reflect all sources of uncertainty and a priori all possible models that may be entertained should be explicitly stated, with prior distributions being placed upon different models and the parameters of these models. Model averaging should then be carried out across the different possibilities, a process which is fraught with difficulties not least in placing "comparable" priors over what may be fundamentally different objects (see Sect. 6.16.3 for an approach to rectifying this problem). Suppose there are m potential models and that $p_j = \Pr(M_j \mid \boldsymbol{y})$ is the posterior probability of model j, $j = 1, \ldots, m$. Then

$$\text{E}[\phi \mid \boldsymbol{y}] = \sum_{j=1}^{k} \text{E}[\phi \mid M_j, \boldsymbol{y}] \times p_j$$

$$\neq \text{E}[\phi \mid \widehat{M}, \boldsymbol{y}], \tag{4.31}$$

where the latter is that which would be reported, based on a single model $\widehat{M}$. The "bias" is $\text{E}[\phi \mid \widehat{M}, \boldsymbol{y}] - \text{E}[\phi \mid \boldsymbol{y}]$. In addition,

$$\text{var}(\phi \mid \boldsymbol{y}) = \sum_{j=1}^{m} \text{var}(\phi \mid M_j, \boldsymbol{y}) \times p_j + \sum_{j=1}^{m} \left(\text{E}[\phi \mid M_j, \boldsymbol{y}] - \text{E}[\phi \mid \boldsymbol{y}]\right)^2 \times p_j \tag{4.32}$$

$$\neq \text{var}(\phi \mid \widehat{M}, \boldsymbol{y}), \tag{4.33}$$

so that the variance in the posterior acknowledges both the weighted average of the within-model variances, via the first term in (4.32), and the weighted contributions to the between-model variability, via the second term. Note the analogies between the frequentist and Bayesian biases, (4.28) and (4.31), and the reported variances, (4.30) and (4.33).

The fundamental message here is that carrying out model selection leads to estimators whose frequency properties are not those of the estimators without any tests being performed (Miller 1990; Breiman and Spector 1992) and Bayesian single model summaries are similarly misleading. This problem is not unique to

Table 4.3 Parameter estimates, standard errors, and T statistics for the prostate cancer data. The full model and models chosen by stepwise/BIC and C_P/AIC are reported

		Full model			Stepwise/BIC model			C_P/AIC model		
	Variable	Est.	(Std. err.)	T stat.	Est.	Std. err.	T stat.	Est.	Std. err.	T stat.
1	log(can vol)	0.59	(0.088)	6.7	0.55	(0.075)	7.4	0.57	(0.075)	7.6
2	log(weight)	0.46	(0.17)	2.7	0.51	(0.15)	3.9	0.42	(0.17)	2.5
3	age	−0.020	(0.011)	−1.8	–	–	–	−0.015	(0.011)	−1.4
4	log(BPH)	0.11	(0.058)	1.8	–	–	–	0.11	(0.058)	1.9
5	SVI	0.77	(0.24)	3.1	0.67	(0.21)	3.2	0.72	(0.21)	3.5
6	log(cap pen)	−0.11	(0.091)	−1.2	–	–	–	–	–	–
7	gleason	0.045	(0.16)	0.29	–	–	–	–	–	–
8	PGS45	0.0045	(0.0044)	1.0	–	–	–	–	–	–
	σ	0.78	–	–	0.72	–	–	0.71	–	–

variable selection. Similar problems occur when other forms of model refinement are entertained, such as transformations of y and/or x, or experimenting with a variety of variance models and error distributions.

Example: Prostate Cancer

We begin by fitting the full model containing all eight variables. Table 4.3 gives the coefficients, standard errors, and T statistics. For this example, the forward selection and backward elimination stepwise procedures all lead to the same model containing the three variables log(can vol), log(weight), and SVI. The p-value thresholds were chosen to be 0.05. The standard errors associated with the significant variables all decrease for the reduced model when compared to the full model. This behavior reflects the bias-variance trade-off whereby a reduced model may have increased precision because of the fewer competing explanations for the data (for more discussion, see Sect. 5.9). We emphasize, however, that uncertainty in the model search is not acknowledged in the estimates of standard error. We see that the estimated standard deviation is also smaller in the reduced model.

Turning now to methods that evaluate all subsets, Figure 4.12 plots the C_P statistic versus the number of parameters in the model. For clarity, we do not include models with less than four parameters in the plot, since these were not competitive. Recall that we would like models with a small number of parameters whose C_P value is close to or less than the line of equality. The variable plotting labels are given in Table 4.3. For these data, we pick out the model with variables labeled 1, 2, 3, 4, and 5 since this corresponds to a model that is close to the line in Fig. 4.12 and has relatively few parameters. The five variables are log(can vol), log(weight), age, log(BPH), and SVI, so that age and log(BPH) are added to the stepwise model.

Carrying out an exhaustive search over all main effects models, using the adjusted R^2 to pick the best model (which recall is equivalent to picking that model with

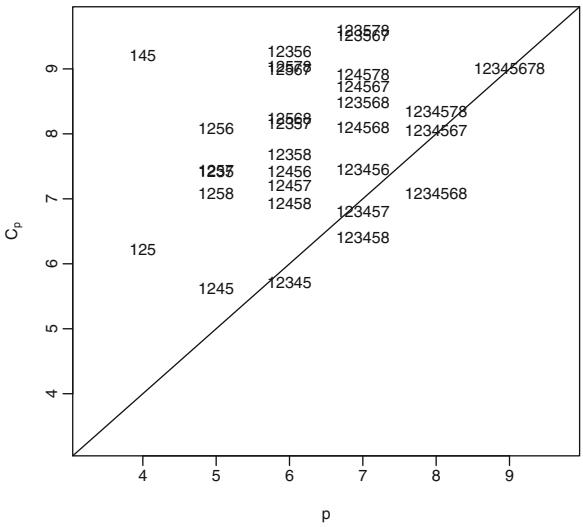

Fig. 4.12 Mallows' C_P statistic plotted versus p, where $p-1$ is the number of covariates in the model, for the prostate cancer data. The line of equality is indicated, for a good model $\mathrm{E}[C_P] \approx p$, where the expectation is over repeated sampling. The variable labels are given in Table 4.3

the smallest $\widehat{\sigma}^2$), gives a model with seven variables (gleason is the variable not included). The estimate of the error variance is $\widehat{\sigma} = 0.70$. The minimum BIC model was the same model as picked by the stepwise procedures.

We used Bayesian model averaging with, for illustration, equal weights on each of the 2^8 models and weakly informative priors. The most probable model has posterior probability 0.20 and contains log(can vol), log(weight), and SVI, while the second replaces log(weight) with log(BPH) and has posterior probability 0.09. The third most probable model adds log(BPH) to the most probable model and has probability 0.037. Cumulatively across models, the posterior probability that log(can vol) is in the model is close to 1, with the equivalent posterior probabilities for SVI, log(weight), and log(BPH) being 0.69, 0.66, and 0.27, respectively. A more detailed practical examination of BMA is presented at the end of Chap. 10.

4.10 A Pragmatic Compromise to Variable Selection

One solution to deciding upon which variables for inclusion in a regression model is to never refine the model for a given dataset. This approach is philosophically pure but pragmatically dubious (unless one is in the context of, say, a randomized experiment) since we may obtain appropriate inference for a model that is a very poor description of the phenomenon under study. It is hard to state general strategies, but on some occasions, it may be safest, and the most informative, to report multiple models.

We consider situations that are not completely confirmatory and not completely exploratory. Rather we would like to obtain a good description of the phenomena

under study and also have some faith in reported interval estimates. The philosophy suggested here is to think as carefully as possible about the model before the analysis proceeds. In particular, context-specific models should be initially posited. Hopefully the initial model provides a good description, but after fitting the model, model checking should be carried out and the model may be refined in the face of *clear* model inadequacy, with refinement ideally being carried out within distinct a priori known classes. A key requirement is to describe the procedure followed when the results are reported.

If a model is chosen because it is clearly superior to the alternatives then, roughly speaking, inference may proceed as if the final model were the one that was chosen initially. This is clearly a subjective procedure but can be informally justified via either frequentist or Bayesian approaches. From a frequentist viewpoint, it may be practically reasonable to assume, with respect to (4.28), that $\mathrm{E}[\phi \mid P] \approx \mathrm{E}[\phi \mid \widehat{M}]$ because $\widehat{M}$ would be almost always chosen in repeated sampling under these circumstances. In a similar vein, under a Bayesian approach, the above procedure is consistent in which model averaging in which the posterior model weight on the chosen model is close to 1 (since alternative models are only rejected on the basis of clear inadequacy), that is, with reference to (4.31), $\mathrm{E}[\phi \mid \boldsymbol{y}] \approx \mathrm{E}[\phi \mid \widehat{M}, \boldsymbol{y}]$, because $\Pr(\widehat{M} \mid \boldsymbol{y}) \approx 1$. The aim is to provide probability statements, from either philosophical standpoints that are "honest" representations of uncertainty.

The same heuristic applies more broadly to examination of model choice, beyond which variables to put in the mean model. As an example of when the above procedure should not be applied, examining quantile–quantile plots of residuals for different Student's t distributions and picking the one that produces the straightest line would not be a good idea.

4.11 Concluding Comments

In this chapter, we have discussed frequentist and Bayesian approaches to hypothesis testing. With respect to variable selection, we make the following tentative conclusions. For pure confirmatory studies, one should not carry out model selection and use instead background context to specify the model. Prediction is a totally different enterprise and is the subject of Chaps. 10–12. In exploratory studies, stepwise and all subsets may point to important models, but attaching (frequentist or Bayesian) probabilistic statements to interval estimates is difficult. For studies somewhere between pure confirmation and exploratory, one should attempt to minimize model selection, as described in Sect. 4.10.

From a Bayesian or a frequentist perspective, regardless of the criteria used in a multiple hypothesis testing situation, it is essential to report the exact procedure followed, to allow critical interpretation of the results.

We have seen that when a point null, such as $H_0 : \theta = 0$, is tested, then frequentist and Bayesian procedures may well differ considerably in their conclusions. This is in contrast to the testing of a one-sided null such as $H_0 : \theta \leq 0$; see Casella

and Berger (1987) for discussion. We conclude that hypothesis testing is difficult regardless of the frequentist or Bayesian persuasion of the analysis. A particular difficulty is how to calibrate the decision rule; many would agree that the Bayesian approach is the most natural since it directly estimates $\Pr(H = 0 \mid \boldsymbol{y})$, but this estimate depends on the choices for the alternative hypotheses (so is a relative rather than an absolute measure) and on all of the prior specifications. The practical interpretation of the p-value depends crucially on the power (sample size and observed covariate distribution in a regression setting) and reporting point and interval estimates alongside a p-value or an α level is strongly recommended.

Model choice is a fundamentally more difficult endeavor than estimation since we rarely, if ever, specify an exactly true model. In contrast, estimation is concerned with parameters (such as averages or linear associations with respect to a population) and these quantities are well defined (even if the models within which they are embedded are mere approximations).

4.12 Bibliographic Notes

There is a vast literature contrasting Bayesian and frequentist approaches to hypothesis testing, and we mention just a few references. Berger (2003) summarizes and contrasts the Fisherian (p-values), Neyman (α levels), and Jeffreys (Bayes factors) approaches to hypothesis testing, and Goodman (1993) provides a very readable, nontechnical commentary. Loss functions more complex than those considered in Sect. 4.3 are discussed in, for example, Inoue and Parmigiani (2009).

The running multiple hypothesis testing example concerned the analysis of multiple transcripts from a microarray experiment. The analysis of such data has received a huge amount of attention; see, for example, Kerr (2009) and Efron (2008).

4.13 Exercises

4.1 Consider the simple situation in which $Y_i \mid \theta \sim_{iid} \mathrm{N}(\theta, \sigma^2)$ with σ^2 known. The MLE $\widehat{\theta} = \overline{Y} \sim \mathrm{N}(\theta, V)$ with $V = \sigma^2/n$. The null and alternative hypotheses are $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$, and under the alternative, assume $\theta \sim \mathrm{N}(0, W)$. Consider the case $W = \sigma^2$:

(a) Derive the Bayes factor for this situation.

(b) Suppose that the prior odds are $\mathrm{PO} = \pi_0/(1 - \pi_0)$, with π_0 the prior on the null, and let $R = L_{\mathrm{II}}/L_{\mathrm{I}}$ be the ratio of losses of type II to type I errors. Show that this setup leads to a decision rule to reject H_0 of the form

$$\sqrt{1+n} \times \exp\left(-\frac{Z^2}{2}\frac{n}{1+n}\right) \times \mathrm{PO} < R \tag{4.34}$$

where $Z = \widehat{\theta}/\sqrt{V}$ is the usual Z-statistic.

(c) Rearrangement of (4.34) gives a Wald statistic threshold of

$$Z^2 > \frac{2(1+n)}{n} \log\left(\frac{\text{PO}}{\text{R}}\sqrt{1+n}\right).$$

Form a table of the p-values corresponding to this threshold, as a function of π_0 and n and with $R = 1$. Hence, comment on the use of 0.05 as a threshold.

4.2 The k-FWER criteria controls the probability of rejecting k or more true null hypotheses, with $k = 1$ giving the usual FWER criteria. Show that the procedure that rejects only the null hypotheses H_i, $i = 1, \ldots, m$ for those p-values with $p_i \leq k\alpha/m$, controls the k-FWER at level α.

4.3 Prove expression (4.17).

4.4 In this question, an MCMC algorithm for the Bayesian mixture model described in Sect. 4.6.2 will be derived and applied to "pseudo" gene expression data that is available on the book website.

The three-stage model is:

Stage One:

$$Y_i \mid H_i, \delta, \tau, \pi_0 \sim_{ind} \begin{cases} \text{N}(0, \sigma_i^2) & \text{if } H_i = 0 \\ \text{N}(\delta, \sigma_i^2 + \tau^2) & \text{if } H_i = 1 \end{cases}.$$

Stage Two: $H_i \mid \pi_1 \sim_{iid} \text{Bernoulli}(\pi_1)$.

Stage Three: Independent priors on the common parameters:

$$p(\delta, \tau, \pi_0) \propto 1/\tau.$$

Derive the form of the conditional distributions

$$\begin{aligned} \delta &\mid \tau^2, \pi_0, \boldsymbol{H} \\ \tau^2 &\mid \delta, \pi_0, \boldsymbol{H} \\ \pi_0 &\mid \tau^2, \delta, \boldsymbol{H} \\ H_i &\mid \delta, \tau^2, \pi_0, H_i, \quad i = 1, \ldots, m, \end{aligned}$$

where $\boldsymbol{H} = [H_1, \ldots, H_m]$. The form for τ^2 requires a Metropolis–Hastings step (as described in Sect. 3.8.2).

Implement this algorithm for the gene expression data on the book website.