## Key Concepts

<table>
<thead>
<tr>
<th>likelihood</th>
<th>prior odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>likelihood ratio</td>
<td>Bayes factor</td>
</tr>
<tr>
<td>nuisance parameter</td>
<td>noninformative</td>
</tr>
<tr>
<td>admixture proportion</td>
<td>Bayes estimate</td>
</tr>
<tr>
<td>commingling analysis</td>
<td>credible interval</td>
</tr>
<tr>
<td>frequentist</td>
<td>posterior distribution</td>
</tr>
<tr>
<td>Bayesian</td>
<td>experimentwise type I error</td>
</tr>
<tr>
<td>posterior probability of type I error</td>
<td>familywise type I error</td>
</tr>
<tr>
<td>posterior probability of type II error</td>
<td>hyperparameter</td>
</tr>
<tr>
<td>false discovery rate</td>
<td>empirical Bayes</td>
</tr>
<tr>
<td>posterior odds</td>
<td></td>
</tr>
</tbody>
</table>
Likelihood Ratios, Bayesian Methods and Multiple Hypotheses

SYMBOLS AND ABBREVIATIONS

\[ H_0 \] Null hypothesis
\[ \mu_1, \mu_2 \] Means
\[ \sigma^2 \] Variance
\[ \psi \] Admixture proportion
\[ LR \] Likelihood ratio

LIKELIHOOD RATIOS

In Chapter 6 we introduced maximum likelihood estimators as a type of estimator with many good properties. Since different statistical tests compete with each other in much the same way that different estimators do, it would be helpful to have a general approach to deriving statistical tests that have desirable properties. One such approach is to use what is known as the likelihood ratio criterion.

Recall that we ended Chapter 4 by calling the ratio of two conditional probabilities a likelihood ratio. In that instance, we were discussing the probability of a child receiving a B allele conditional on the child’s father being a man accused of being the father, relative to the probability of the same event conditional on the child’s father being a random man from a particular population. We can think of these two possibilities as two different hypotheses. In general, the likelihood ratio is simply the likelihood of one hypothesis relative to another, for an observed set of data. This is calculated as the probability of the data given the one hypothesis, often defined by a particular value of a parameter, divided by the probability of the same data given the other hypothesis.
given the second, competing hypothesis – that is, a competing value of the parameter. (Recall that we often state hypotheses as contentions about parameters.) In Chapter 6 we defined the maximum likelihood estimate of a parameter as the value of the parameter that makes ‘the probability, or likelihood, of our sample as large as possible’. It is now time to make a subtle distinction between the probability of a particular sample and its likelihood. We use the word ‘likelihood’ when we want to stress that we are interested in different values of the parameters, keeping the values of the random variables fixed at the values observed in a particular sample. We use the word ‘probability’, on the other hand, when we want to stress that we are interested in different values of the random variable, keeping the values of the parameters fixed at the population values. Thus the expression for \( P(y \text{ males}) \) on page 136 is both the probability of, and the likelihood for, the sample value \( y \), which term we use merely indicating whether we want to consider it as a mathematical function of \( y \) or of \( \pi \). Depending on which we consider fixed, the parameters or the data, we talk of the probability distribution of random variables, given the values of one or more parameters, but the likelihood distribution of one or more parameters, given a specific set of data. In this context, likelihoods are functions of one or more parameters that for estimation purposes are viewed as variables having domains (ranges of values the parameters are allowed to have) that may be restricted by hypotheses.

To a certain extent, likelihoods can be manipulated just like probabilities. Analogous to conditional and joint probabilities, we can have conditional and joint likelihoods. Like probabilities, likelihoods must be positive. Unlike probabilities, however, likelihoods need not be less than one, and they do not belong to sets of values that have to add up to one. We saw that when we are dealing with a continuous random variable, any particular sample value has (theoretically) zero probability of occurring. The likelihood for that sample value would in this case be the height of the corresponding probability density function. Probabilities, as we discussed in Chapter 4, can be interpreted as relative frequencies. Because they must add up to one, it is intuitively clear what they mean. Individual likelihoods, on the other hand do not have any similar meaning. Likelihoods are only meaningful when compared with one another – in fact, they are only defined as being proportional to the corresponding probabilities.

To calculate a likelihood ratio, we need to formulate two probabilities (or the heights of density functions) for the same set of data, corresponding to two hypotheses – the null, \( H_0 \), and the research hypothesis – and divide one by the other. If we multiply each probability (or density) by the same multiplier, that multiplier cancels out in the ratio. If we multiply a probability by some number, the probability changes. But if, for a given sample, we multiply the likelihoods of two different parameter values by the same constant number, there is no change in the ratio of these two likelihoods. In this sense, when we base all inferences on
Suppose, as an example, we have two random samples, one from each of two populations, and we wish to test whether the population means are equal. We assume that the random variable of interest is normally distributed with the same variance in the two populations. Under one hypothesis (the null hypothesis, $H_0$), the two samples come from identical normal distributions (same means and variances); under the other, alternative, hypothesis they come from two normal distributions with the same variances but different means. We wish to test $H_0$: $\mu_1 = \mu_2$, the means of the two samples are not different. (Of course, we already know that the two-sample $t$-test is appropriate for this situation, but we shall nevertheless use this example to illustrate how the likelihood ratio test works). Under each hypothesis the likelihood depends on the mean(s) and the variance, which are unknown parameters. To obtain the likelihood ratio, we formulate a general likelihood and maximize it twice: once over all possible values of a common mean $\mu$ and a common variance $\sigma^2$ (i.e. under $H_0$), and once over all possible values of separate means $\mu_1$ and $\mu_2$, and a common variance $\sigma^2$ (the alternative, research, hypothesis that the means are different). Notice that we maximize the likelihood to estimate the variance as well as the one or two means, and the estimate of this variance will be different in the two maximizations. When we need to estimate one or more parameters not explicitly involved in the null hypothesis, such as the variance in this example, they are called nuisance parameters.

By formulating the likelihood on the assumption that the trait is normally distributed with the same variance in the two populations, the test that the population means are equal is the two-sample $t$-test and, hence, has a $t$-distribution. We stress that the maximum likelihood estimate of the variance, a nuisance parameter, will be different in the two maximizations. It is incorrect (as has sometimes been done in the field of genetic epidemiology) to estimate a nuisance parameter from one of the two likelihoods and then fix it at that value when maximizing the other likelihood.

Consider as another example testing whether our sample data come from a single normal distribution versus a mixture of two normal distributions (this is a special case of what has been called commingling analysis in genetics). In this example, we have a single population (not two populations) but we believe it may be a mixture of two subpopulations. Once again we formulate a general likelihood and maximize it twice, under the null and alternative hypotheses. The general likelihood assumes that with (unknown) probability $\psi$ an observation in our data set comes from a normal distribution with mean $\mu_1$ and variance $\sigma^2$, and with probability $1 - \psi$ it comes from a normal distribution with mean $\mu_2$ and variance $\sigma^2$. We maximize this likelihood with and without the restriction $\mu_1 = \mu_2$ to obtain the likelihood ratio. Another example that occurs in genetics is to test whether a trait is due to segregation of a particular allele that is dominant or recessive with respect to
a specified trait. Here the two likelihoods correspond to two different modes of inheritance for a particular allele and trait phenotype. Note that if there are two phenotypes (e.g. affected and unaffected), dominance of an allele with respect to one of the phenotypes is identical to recessive inheritance with respect to the other phenotype. We detect a difference in likelihoods only if we specify, for example, that it is the less frequent allele that corresponds to the phenotype ‘affected’.

Under fairly general conditions, which apply to all the cases we have just described, it can be proved that any statistical test based on the likelihood ratio is the most powerful. To apply such a test, we need to know the null sampling distribution, that is, the distribution of this ratio when the null hypothesis is true. We leave discussion of this to Chapter 9.

### BAYESIAN METHODS

The statistical methods for estimation and hypothesis testing that we have described so far in this book are known as frequentist methods, based on probabilities of sample data that can be calculated when the parameters are known. In Chapter 7 we defined the significance level \( \alpha \) as the probability of making a type I error, the probability of making an error when \( H_0 \) is true. We also defined the power \( 1 - \beta \) as the probability of not making a type II error, the probability of rejecting the null hypothesis when in fact it is false. But what most investigators really want to know, after performing a statistical test, is the probability that \( H_0 \) is true or equivalently the probability that it is false. These two probabilities are equivalent in the sense that once we know one we know the other, because the two probabilities must add up to 1. Suppose we could assign to the letters \( a, b, c \) and \( d \) in the following \( 2 \times 2 \) table numbers that represent the frequency with which each possibility actually occurs:

<table>
<thead>
<tr>
<th>True state of nature</th>
<th>( H_0 ) is true</th>
<th>( H_0 ) is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision made</td>
<td>( a )</td>
<td>( b )</td>
</tr>
<tr>
<td>Accept ( H_0 )</td>
<td>( c )</td>
<td>( d )</td>
</tr>
</tbody>
</table>

Then \( \alpha \), the probability of making a type I error, is equal to \( c/(a + c) \), and \( \beta \), the probability of making a type II error, is equal to \( b/(b + d) \). These probabilities should be contrasted with the probabilities \( b/(a + b) \) and \( c/(c + d) \), which are of more scientific interest. If we knew them, we would also know their complements, \( a/(a + b) \) and \( d/(c + d) \), the probabilities of making the right decision. The error probabilities \( b/(a + b) \) and \( c/(c + d) \) are sometimes called the posterior probabilities of type II and type I error, respectively. The posterior probability of type I error is also called the false discovery rate – the probability of identifying false research
hypotheses as being true (but the term ‘false discovery rate’ is often used for a particular method of calculating a quantity that could be larger than the true posterior probability of type I error). Why, you may ask, if these posterior error probabilities are the real probabilities we should be controlling, is this ignored by the frequentist methods? The answer lies in the fact that they depend on information that is usually unknown; but sometimes – especially in genetics – the required information is known. We need to know, prior to any data collection, the probability that $H_0$ is true (or false). With that information, we can use Bayes’ theorem to obtain the answers we want.

Recall that in case 2 of the paternity example in Chapter 4, we used Bayes’ theorem with $S$ denoting the result of the paternity test (the child received a B allele from the true father), $D_1$ being the event that the alleged father is the true father, and $D_2$ the event that a random man from a specified population is the true father. Think of $D_2$ as $H_0$ and $D_1$ as the alternative hypothesis – $H_0$ is false and so the research hypothesis is true. In the paternity example, we assumed we knew the prior probabilities, before the paternity test was performed, that the accused father is the true father or not: $P(D_1) = 0.65$, $P(D_2) = 0.35$. We also calculated $P(S|D_1) = 0.5$ and $P(S|D_2) = 0.06$, with the result that the eventual (posterior) probability of paternity was found to be $P(D_1|S) = 0.94$. Now rewrite the equation we had before,

$$P(D_1|S) = \frac{0.65 \times 0.5}{0.65 \times 0.5 + 0.35 \times 0.06} = 0.94.$$  

as

$$P(H_0 \text{ is false | the data}) = \frac{P(H_0 \text{ false})P(\text{the data | } H_0 \text{ false})}{P(H_0 \text{ false})P(\text{the data | } H_0 \text{ false}) + P(H_0 \text{ true})P(\text{the data | } H_0 \text{ true})}.$$  

Because $P(H_0 \text{ is false | the data})$ is the same as $P(\text{the research hypothesis is true | the data})$, this gives the researcher exactly what is wanted, once the data have been collected and analyzed.

With algebraic manipulation, as shown in the Appendix, this can be written:

posterior odds of the research hypothesis = prior odds of the research hypothesis $\times$ LR,

where ‘posterior’ refers to probabilities after the data have been collected, ‘prior’ refers to probabilities before any data are collected, and LR is the likelihood ratio, defined as

$$LR = \frac{P(\text{the data | } H_0 \text{ is false})}{P(\text{the data | } H_0 \text{ is true})}.$$
Thus, to answer the question that is often of most interest (i.e. now that I have obtained a result, what is the probability that $H_0$ is false?), we need to know, before seeing the results of the study, the relative prior probabilities $P(H_0 \text{ is false})$ and $P(H_0 \text{ is true})$. Note that if we assume that the null hypothesis $H_0$ and the research hypothesis have equal prior probabilities, the prior odds of the event ‘$H_0$ is false’ (the research hypothesis is true) will equal 1, and then the posterior odds is the same as the likelihood ratio.

When we test a hypothesis, which is the frequentist approach, the more plausible the research hypothesis is before we conduct the study, the less stringent the test need be before we reject $H_0$. If the research hypothesis has low prior probability of being true, then the study should be required to attain high statistical significance (i.e. a very small $p$-value) before any confidence is placed in it; if it has relatively high prior probability of being true, less significance (a larger $p$-value) could suffice.

If we use Bayesian methods, on the other hand, the choice of an appropriate significance level is bypassed in favor of determining the relative probabilities of the two hypotheses using both the data and all the prior information available. We simply assign probabilities to hypotheses and parameters, making no essential distinction between random variables and parameters in this regard. The end result is that we can derive probabilities for the hypotheses of interest and, instead of testing hypotheses, we summarize the evidence for or against a research hypothesis as a Bayes factor. Furthermore, estimates are obtained from the posterior distribution of the parameters and ‘credible intervals’ replace confidence intervals.

**BAYES’ FACTORS**

In Chapter 4, when we first introduced the likelihood ratio, specifically pointing out that the paternity index is a likelihood ratio, we stated that this particular likelihood ratio is also a special case of what is known as a *Bayes factor*. It is a Bayes factor that assumes equal prior probabilities that the two competing hypotheses are true, or *noninformative* prior probabilities, and in which there are no nuisance parameters. Bayes factors, just like likelihood ratios, measure the strength of the evidence for one hypothesis versus another. The essential difference is that the Bayes factor incorporates any prior knowledge we may have. Table 8.1 suggests how Bayes factors should be interpreted.

Thus using Bayes factors is very similar to significance testing, where we simply end up with a $p$-value, rather than hypothesis testing, where we make a decision to accept one hypothesis or the other. Unlike $p$-values, however, we cannot attach frequentist probabilities to them. But likelihood ratios provide the most powerful way of testing one hypothesis versus another, without the need of making any assumptions about their relative prior probabilities.
Table 8.1 Suggested interpretation of Bayes factors

<table>
<thead>
<tr>
<th>Bayes factor</th>
<th>Strength of the evidence for the research hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1</td>
<td>Negative evidence</td>
</tr>
<tr>
<td>2</td>
<td>Hardly worth mentioning</td>
</tr>
<tr>
<td>6</td>
<td>Substantial</td>
</tr>
<tr>
<td>20</td>
<td>Fairly strong</td>
</tr>
<tr>
<td>60</td>
<td>Very strong</td>
</tr>
<tr>
<td>100</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

If we wish, as a way of hypothesis testing, we can make a firm decision on the basis of a Bayes factor, such as deciding to reject $H_0$ if it is larger than 20 and to accept it otherwise. But then, to know the power of such a procedure, we need to determine the null distribution of the Bayes factor. In certain cases, as we shall see in Chapter 9, we can very easily determine the null distribution of the likelihood ratio, whereas this is typically always difficult for a Bayes factor. But, except in the special case that the Bayes factor is identical to a likelihood ratio, Bayes factors and likelihood ratios are essentially different. In order to calculate Bayes factors, first we need to have a prior probability for each of the possible hypotheses. Separately in the numerator and denominator, we must multiply each of the probabilities in the likelihood ratio by the corresponding prior probability and add together these products. To take a very simple example, suppose we have just two discrete hypotheses, that with respect to the phenotype ‘disease’ a particular allele is either dominant or recessive. Then the Bayes factor in favor of dominant inheritance would be calculated as

$$
\frac{P(\text{dominant inheritance})P(\text{data} | \text{dominant inheritance})}{P(\text{dominant inheritance})P(\text{data} | \text{dominant inheritance}) + P(\text{recessive inheritance})P(\text{data} | \text{recessive inheritance})}
$$

Second, not only do Bayes factors require the assumption of prior odds, there is also a difference in the way they treat nuisance parameters. In likelihood ratios, nuisance parameters are simply replaced by their maximum likelihood estimates. In the calculation of Bayes factors, an averaging is performed. When the hypotheses are stated in terms of parameter values on an interval scale so that they are continuous (such as in the case of the mean or variance of a distribution), their probabilities are densities and the calculus operation of integration replaces the summation. This integration has to be over all possible values the parameters, including any nuisance parameters, can take on, separately in the numerator and denominator. The net result of this is that each nuisance parameter is averaged (over its prior distribution) rather than replaced by a single estimate. These ‘average’ values of the nuisance parameters tend to be more stable in small samples, and this is an advantage of Bayes factors. But the averages depend on the prior parameter distributions assumed, and this a major disadvantage if these distributions are unknown and influence the result unduly. Whenever Bayes factors are used, it is important to investigate how
sensitive they are to the parameter distributions assumed. However, if sufficiently large samples are used, the likelihood ratio calculated from the data alone will be the overwhelmingly dominating factor in determining a numerical value for the right-hand side of the equation,

\[
\text{posterior odds of the research hypothesis} = \text{prior odds of the research hypothesis} \times LR.
\]

Thus asymptotically (if the sample size is infinite), there will be no difference between Bayes factors and likelihood ratios; and the estimates based on them, to which we now turn, will be identical.

**BAYESIAN ESTIMATES AND CREDIBLE INTERVALS**

Just as maximum likelihood estimates are based on the likelihood distribution – we choose the parameter value that makes this distribution a maximum – so Bayesian estimates are based on the posterior parameter distribution given the data. We can choose the parameter value that maximizes this distribution, the modal value, or base the estimates on any other characteristic of the posterior distribution, such as its mean or median. But usually we are more interested in estimating an interval in which the parameter probably lies. For this purpose we choose an interval that contains a given percentage of the posterior distribution, and this is then called a *credible interval*. Figure 7.1 depicts the sampling distribution of a standardized sample average. If we had not standardized the sample average, Figure 7.1 would have been similar to Figure 8.1. The percentiles indicated along the x-axis of this figure are the same as the confidence interval limits calculated in Chapter 7 for the mean weight of men on a weight-reduction program. Now let us suppose, as indicated in the legend of Figure 8.1, that what is depicted is the posterior

![Figure 8.1](image-url)
distribution of the mean weight. Then the interval from 165.0 to 203.0, which now contains 95% of the posterior distribution of the (unknown parameter) mean weight, would be a 95% credible interval, and the interval from 168.6 to 199.4, which contains 90% of the posterior distribution, would be a 90% credible interval. Credible intervals are analogous to confidence intervals, but their interpretation is quite different. Go back to Chapter 7 and be sure you understand the meaning of a confidence interval. Credible intervals give an interval that actually includes the unknown parameter with a specified probability, which is much more meaningful to most researchers. But they depend on the prior distributions that were assumed for all the parameters, and different researchers might disagree on what the appropriate distributions should be. If the sample size is sufficiently large, however, credible intervals and confidence intervals will be almost identical.

### THE MULTIPLE TESTING PROBLEM

The multiple testing problem, sometimes referred to as the multiple comparisons problem, occurs whenever we try to interpret the results of multiple hypothesis tests. It is now, as a consequence of cheaper and cheaper high-throughput molecular technology, an acute problem in interpreting the results of many genetic studies. Although it is a general problem whatever statistical method is used, we discuss it here, in the same chapter as Bayesian methods, because these methods are now being developed specifically to tackle the enormous number of tests that are currently being performed with micro-array SNP-chip data. (After this chapter we shall discuss only standard frequentist statistical methods. Because there is no general agreement about what prior distributions should be assumed, more than one Bayesian method is possible for each of the standard frequentist methods available, depending on the particular situation to which it is applied).

To understand the multiple testing problem, consider the fact that in the 1970s just about every disease imaginable was reported to be associated with one or other HLA allele. This polymorphism had been recently discovered and shown to be genetically determined by segregation at the HLA system, a set of loci in close proximity on chromosome 6 that could be investigated as a single locus with many alleles. It was a simple matter to type a series of random patients with a particular disease and a series of random controls without disease for, say, 40 alleles; then 40 tests would be performed, testing whether each allele frequency was the same in the two populations, patients and controls. (For our present purpose it is unimportant to know the details of the statistical test performed.) Now suppose we validly perform each test of $H_0$, that the frequencies are the same at a pre-specified significance level $\alpha$, and that the 40 tests are independent. Then, if $H_0$ is true, the probability we would not reject any particular one $H_0$ is $1 - \alpha$, and the probability we would
not reject any of the 40 $H_0$s is $(1 - \alpha)^{40}$. It follows that the probability of rejecting at least one of the 40 null hypotheses would be $1 - (1 - \alpha)^{40}$. If we put $\alpha = 0.05$ in this expression, we find $1 - (1 - \alpha)^{40} \approx 1 - 0.12851 \approx 0.87$. In other words, even if there were no association whatsoever, we had an 87% chance of finding an allele significantly associated at the 5% level! This explains why so many diseases were originally reported to be associated with HLA.

Note carefully that each of the 40 tests by itself would be valid, and there would have been no harm in reporting 40 $p$-values. The problem arose because only those associations that were significant were published. Of course, if all 40 results were reported and, before any replication study was performed, the reader inadvertently gave special attention to those alleles with $p$-values less than 0.05, the problem would still be there. Nowadays it is not difficult to perform hundreds, thousands, or even hundreds of thousands of tests, with no possibility of reporting all the $p$-values – though they can be plotted as points against position along the genome, with many of the points overlapping and a few, presumably those indicating real associations, standing out from the rest. So we need to have a way of sorting out which results are significant at, for example, the 5% level, after ‘adjustment’ for the number of tests performed. Before the adjustment, the type I error is called the experimentwise type I error and the $p$-values are called ‘nominal’. After allowing for the multiple tests, the type I error is called the familywise type I error and the $p$-values are called ‘adjusted’ (the expression ‘familywise’ comes from the fact that a ‘family’ of tests has been performed – it has nothing to do with family data). Phrases such as ‘adjusted to be genomewide significant at the 5% level’, or ‘genomewide significance at the 5% level’ are found in the literature.

A common way to adjust nominal $p$-values for multiple testing is to use what is known as Bonferroni’s inequality. Suppose, for example, that a sample of patients with a particular disease is compared with a sample of controls with respect to a panel of $c$ different HLA antigens. Then, if the reported nominal $p$-value is $p^*$, we simply multiply $p^*$ by $c$, that is to say, the Bonferroni-adjusted $p$-value, which cannot be smaller than the $p$-value we want, is $c \times p^*$. If the tests are independent, we can calculate the adjusted $p$-value as $1 - (1 - p^*)^c$ (this is known as Šidák’s method). When $p^*$ is small, this is only very slightly smaller than $c \times p^*$. If the tests are not independent, either of these methods is conservative, resulting in an upper bound for the true familywise $p$-value. In Chapter 11, when we consider the analysis of variance, we discuss a way of pooling the data to obtain a more accurate estimate of the variance while at the same time controlling type I error when multiple comparisons are made. In Chapter 12 we discuss multivariate methods, in which multiple outcome responses are measured on each study subject and we make use of the dependences among the responses to obtain tests that can be more powerful while at the same time controlling type I error. The gain in power comes from particular assumptions about the dependencies in the data, and so these
assumptions must always be carefully scrutinized. In any case, these particular tests can be performed only if the number of independent study subjects in the sample is much larger than the number of response outcomes being studied. It is now not uncommon to study the genetic expression of tens of thousands of loci with a sample of only a few hundred individuals, or hundreds of thousands of SNPs on a few hundred or thousand cases and controls. In such situations Bayesian methods are starting to be used and appear to provide very useful results.

The typical Bayesian approach is to make some assumptions about the relative probabilities of all the alternative hypotheses and then control the posterior probability of Type I error. The end result is often quoted as a $q$-value, the Bayesian analogue of a $p$-value. It is in fact possible to estimate an upper bound for this posterior probability without making the prior assumptions necessary for a fully Bayesian approach, and this is often the method referred to in the literature as controlling the false discovery rate. But here we wish to indicate the kinds of assumptions that are being made – assumptions that, if true, allow us to calculate the actual probability of a particular Bayesian method ending up making a posterior type I or type II error.

Consider, as one example, determining the difference in genetic expression between tumor and nontumor cells at 20,000 loci. One might expect no difference in expression for about half the loci and a flat uniform distribution of differences for the other half. This is the 'spike and slab' model – the spike being a discrete probability of one half at no difference and the slab being a uniform distribution of differences for the other half of the loci. As a second example, if one has a million diallelic SNPs to be compared in a case–control study and we give each person a score equal to the number of minor alleles he or she has, that score must be 0, 1 or 2 (the minor allele is the allele that has the smaller population frequency). We might then expect the million case–control differences in scores to be symmetrically distributed about 0 with a tall spike at 0 (reflecting no case–control difference in the number of minor alleles a person has), smaller probabilities for score differences of $-1$ and $+1$, and even smaller probabilities for score differences of $-2$ and $+2$. Thus we would assume five prior probabilities that add up to 1 for the five possible score differences. We might assume this prior distribution can be expressed as a probability function that depends on only one or two parameters. Any parameters that determine a prior distribution for the parameters that have to be specified in order to perform a Bayesian analysis are called hyperparameters. These hyperparameters are then estimated in addition to the parameters of interest – the tumor–control mean difference in expression between tumor cells and non-tumor cells of each of the 20,000 loci in the first example, the mean case–control score difference for each of the million SNPs in the second example. The hyperparameters are treated as nuisance parameters in the analysis. In an empirical Bayes analysis, all the parameters, including the hyperparameters, are estimated by maximum likelihood. The net
effect of this is that an *empirical Bayes* analysis is nothing more than a maximum likelihood analysis that uses a more comprehensive statistical model. We take into account the fact that we expect some kind of distribution of test results but assume nothing more than the form of this distribution. We do not need to assume, as in a true Bayesian analysis, that both random variables and parameters have distributions. However, because an empirical Bayes analysis is often formulated as though parameters have distributions, it is rarely recognized as being essentially nothing more than a maximum likelihood analysis that uses a more comprehensive model than would normally be used in a frequentist maximum likelihood approach.

**SUMMARY**

1. The likelihood ratio is the likelihood of one hypothesis relative to another for an observed set of data. It is calculated as the probability of the data given the one hypothesis divided by the probability of the same data given the other hypothesis. It can be proved that under fairly general conditions any statistical test based on the likelihood ratio is the most powerful.

2. Nuisance parameters are parameters that need to be estimated but are not explicitly stated as part of the null hypothesis.

3. Bayes factors incorporate prior information. When there are no nuisance parameters and the prior probabilities of the two hypotheses are equal (the prior odds = 1), a Bayes factor is the same as a likelihood ratio.

4. Bayes factors, like likelihood ratios, measure the strength of the evidence for one hypothesis versus another. The essential difference is that a Bayes factor incorporates any prior knowledge we may have. In the calculation of the likelihood ratio, nuisance parameters are replaced by their maximum likelihood estimates. In the calculation of Bayes factors, they are averaged over an assumed prior distribution.

5. Bayesian estimates are based on the posterior parameter distribution given the data. The intervals obtained from this distribution are called credible intervals. If the sample size is sufficiently large, credible intervals and confidence intervals will be almost identical.

6. The probability of identifying a false research hypotheses as being true is the posterior probability of type I error, also called the false discovery rate. A Bayesian analysis typically controls this error rather than the probability of making an error when the null hypothesis is true.

7. Whenever we highlight the most significant of multiple hypothesis tests we encounter the multiple testing, or the multiple comparisons, problem. Nominal
*p*-values reflect the experiment-wise type I error, adjusted *p*-values reflect familywise type I error. If *c* tests are performed, the Bonferroni method multiplies the nominal *p*-value *p* by *c*; if the tests are independent, the adjusted *p*-value is $1 - (1 - p^*)^c$.

8. Any parameters that determine a prior distribution for the parameters that have to be specified in order to perform a Bayesian analysis are called hyperparameters. These are estimated in addition to the parameters of interest, being treated as nuisance parameters in the analysis. In an empirical Bayes analysis, all the parameters, including the hyperparameters, are estimated by maximum likelihood.

**PROBLEMS**

1. In hypothesis testing, the posterior probability of type I error is also called
   A. Bonferroni coefficient
   B. False discovery rate
   C. Odds the hypothesis is false
   D. Nuisance estimator
   E. Binomial error

2. The paternity index is a likelihood ratio that is also a
   A. Dominant phenotype
   B. Critical value
   C. Nomogram
   D. Commingling scale
   E. Bayes factor

3. An interval that actually includes the unknown parameter being estimated with a specified probability is called a
   A. Credible interval
   B. Empirical interval
   C. Hyperbolic interval
   D. Confidence interval
   E. Nominal interval

4. A maximum likelihood analysis that uses a more comprehensive model than would be used in a frequentist maximum likelihood approach to estimation is called a
   A. Biased analysis
   B. Sensitivity analysis
   C. Empirical Bayes analysis
D. Paternity analysis
E. Spike and slab analysis

5. A researcher investigated a set of 100 SNPs at a gene-locus by typing a series of random patients with a disease and a series of random controls without the disease. Then 100 statistical tests were performed to assess whether the allele frequencies at each SNP are the same in the diseased and control populations. Suppose the researcher validly performs each test (of $H_0$ that the frequencies are the same in the two populations) at a pre-specified level 0.05 and the 100 tests are independent. The probability the researcher rejects at least one of the 100 tests is

A. 0.05
B. $(0.05)^{100}$
C. $(0.95)^{100}$
D. $1 - (0.05)^{100}$
E. $1 - (0.95)^{100}$

6. In the multiple testing problem, $p$-values are typically adjusted to provide a known familywise type I error. If this adjustment is not made, and each of a series of statistical tests is performed at the 0.05 level, the nominal $p$-values may lead to an inflated

A. experimentwise type I error
B. commingling of errors
C. number of falsely accepted hypotheses
D. estimate of bias
E. posterior likelihood ratio

7. In a statistical test that employs a Bayesian approach, the end result is often quoted as a $q$-value which is a Bayesian analogue of a

A. Bonferroni adjusted confidence interval
B. paternity index
C. confidence coefficient
D. quotient factor
E. frequentist $p$-value

8. A Bayes factor that assumes equal prior probabilities and in which there are no nuisance parameters is a

A. Šidák adjusted $p$-value
B. biased estimate of relative risk
C. likelihood ratio
D. hypothetical parameter
E. admixture proportion