2

Basic Statistics

2.1 Introduction

Now that we have seen the basics of genetics, we turn to an introduction to the statistical methodologies that we will use throughout this book. Most of the statistical inferences that we will make will be based on likelihood analysis, and we will be concerned not only with constructing the appropriate likelihood function for a given model but also with methods for computing and optimizing the likelihood. We start here with some basics and work our way toward likelihood analysis of a genetic linkage model.

2.1.1 Populations and Models

The goal of a statistical analysis is to draw conclusions about a *population*, a collection of objects (typically infinite) not all of which can be measured, based on examination of a *sample*, a smaller collection of objects drawn from the population, all of which can be measured. To connect the sample to the population and have the ability to make an inference, we use a *model*.

Example 2.1 (Tomato Plant Heights). Suppose that we have the following data $\mathbf{y}$ on heights (in cm) of 12 tomato plants of a particular species:

$$\mathbf{y} = (y_1, y_2, \ldots, y_{12}) = (79, 82, 85, 87, 100, 101, 102, 103, 124, 125, 126, 127).$$

We may take the following simple model. The true mean height of the population is μ, and we observe data Y_i according to

$$Y_i = \mu + \varepsilon_i, \tag{2.1}$$

where ε_i is an error term, typically taken to have a normal distribution with mean 0 and variance σ^2.

For the model (2.1), it is often assumed that the ε_i follow a normal distribution $N(0, \sigma^2)$, where the $N(\mu, \sigma^2)$ probability density function is given by

$$\phi(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

Thus, under the normality assumption, we can also write the model (2.1) as $Y_i \sim \phi(y|\mu, \sigma)$.

As the process that we are trying to describe gets more complicated, so will the model. In this book, we will examine a range of models, with forms that are often dictated by the biology of the problem. Here are some examples of other models:

(a) *Linear Regression.* To describe a linear relationship between a dependent variable Y and an independent variable (or *covariate*) x, we could use the model

$$Y_i = a + bx_i + \varepsilon_i.$$

If we have many different covariates, we could use a *multiple regression model* $Y_i = a + \sum_j b_j x_{ij} + \varepsilon_i$.

(b) To describe a relationship between a covariate and a Bernoulli random variable (a variable Y that only takes the values 0 and 1, with $P(Y = 1) = p$), a *logistic regression model* is often used. This has the form

$$\text{logit}(p(x)) = a + bx,$$

where the *logit* is defined as $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$, and $P(Y = 1|x) = p(x)$.

(c) There are many models used to describe the growth process as a function of time. One popular model is the *logistic growth curve*, given by

$$g(t) = \frac{a}{1 + be^{-rt}},$$

where t typically is a time measurement. If we observe actual growth, we would use this model in the form $Y_i = \frac{a}{1+be^{-rt_i}} + \varepsilon_i$.

(d) The last model that we will describe here will find much use in QTL mapping (see Chapter 9). It is a *mixture model*, given by

$$Y_i = \begin{cases} \mu_1 + \varepsilon_i & \text{with probability} \quad p \\ \mu_2 + \varepsilon_i & \text{with probability} \quad 1-p \end{cases},$$

or, equivalently,

$$Y_i \sim \begin{cases} N(\mu_1, \sigma^2) & \text{with probability} \quad p \\ N(\mu_2, \sigma^2) & \text{with probability} \quad 1-p \end{cases}.$$

We can also write the mixture model as

$$Y_i = \mu_1 I(X = 1) + \mu_2 I(X = 0) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

where $P(X = 1) = p$, and I is the *indicator function*, which is equal to 1 if the argument is true and equal to 0 otherwise.

2.1.2 Samples

To begin an investigation, we would typically draw a *random sample* from the population of interest, a small collection of objects that are representative of the population. A random sample is, formally, a sample of n objects obtained in such a way that all sets of n objects have the same chance of being the sample. In practice, however, we hope to have a sample that is *independent* and *identically distributed*, or iid.

To be specific, if we are sampling from a population with probability density function $f(y|\theta)$, where θ contains the unknown parameters, and we draw an iid sample $Y_1, Y_2, \ldots, Y_n$, we want each variable to satisfy $Y_i \sim f(y|\theta)$ (identical) and for the variable to be independent. If we obtain an iid sample $y_1, y_2, \ldots, y_n$, the density function of the sample is

$$f(y_1, y_2, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta).$$

Example 2.2 (Normal Sample Density). Suppose that $Y_1, Y_2, \ldots, Y_n$ are an iid sample from an $N(\mu, \sigma^2)$ population. The sample density is

$$\begin{aligned} \phi(y_1, y_2, \ldots, y_n|\mu, \sigma^2) &= \prod_{i=1}^{n} \phi(y_i|\mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}. \end{aligned} \tag{2.2}$$

The sample density defines the relationship between the sample and the model and is thus the only means we have of estimating the unknown parameters. Actually, there are other methods, but they all lack efficiency when compared with estimation based on the sample density (see Casella and Berger 2001). Moreover, all of the information that the sample has about the parameters is contained in the sample density function. A consequence of this is that we should base our parameter estimation method on the sample density.

2.2 Likelihood Estimation

The most common estimation method, which is typically a very good method, is based on the sample density. We define the *likelihood function* as

$$L(\theta|y_1, y_2, \ldots, y_n) = f(y_1, y_2, \ldots, y_n|\theta),$$

which is merely treating the sample density f as a function of the parameter, holding the data fixed. The method of maximum likelihood estimation takes as the estimate of θ the value that maximizes the likelihood.

Example 2.3. (**Finding a Maximum Likelihood Estimator (MLE)**). To find the MLEs in Example 2.2, we need to find the values of μ and σ^2 that maximize equation (2.2). Since likelihood functions from iid samples are products (which are nasty to maximize), it is often easier to work with the logarithm of the likelihood (known as the log-likelihood). The maxima are the same as the original function, and the calculations are quite a bit easier.

From (2.2), the normal log-likelihood is

$$logL(\mu, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Differentiating and setting equal to zero

$$\frac{\partial}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0,$$
$$\frac{\partial}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(y_i - \mu)^2 = 0,$$

gives the MLEs $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$. For the data of Example 2.1, we have $\hat{\mu} = 103.41$ and $\hat{\sigma}^2 = 303.24$.

The rationale behind maximum likelihood estimation is the following. If the sample $y_1, y_2, \ldots, y_n$ is representative, then the values of y_i should come from the regions of f that have high probability; that is, the sample should "sit" under the mode of f. We do not know where this mode is because we do not know the value of θ. However, for the given sample, we can find the value of θ that makes $f(y_1, y_2, \ldots, y_n|\theta)$, or equivalently $L(\theta|y_1, y_2, \ldots, y_n)$, the highest. This puts the sample in the highest probability region, and the resulting estimator is the maximum likelihood estimator (MLE).

Example 2.4. (**Tomato Data Revisited**). As a slightly more realistic example, we return to the tomato heights of Example 2.1, but we now ask if there may be evidence of genetic control of height. Specifically, if a "height" gene is segregating in an F_2 progeny according to Mendel's first law, then we would expect to see the genotype AA:Aa:aa segregating in the ratio 1:2:1. We hypothesize that there is a gene associated with height, and it is segregating in the ratio p:q:$1-p-q$ for genotypes AA:Aa:aa. As the values of p and q are unknown, and as we do not know the genotype of the plant that we observed, the model of Example 2.1 becomes the mixture model

$$Y_i = \begin{cases} \mu_{AA} + \varepsilon_i & \text{with probability } p \\ \mu_{Aa} + \varepsilon_i & \text{with probability } q \\ \mu_{aa} + \varepsilon_i & \text{with probability } 1-p-q \end{cases},$$

where $\varepsilon_i \sim N(0, \sigma^2)$. If we let ϕ_{AA} denote a normal density with mean μ_{AA} and variance σ^2, and if we define ϕ_{Aa} and ϕ_{aa} similarly, then, writing $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the likelihood and log-likelihood functions are

$$L(\mu_{AA}, \mu_{Aa}, \mu_{aa}, \sigma^2|\mathbf{y}) = \prod_{i=1}^{n} \left[p\phi_{AA}(y_i) + q\phi_{Aa}(y_i) + (1-p-q)\phi_{aa}(y_i) \right],$$

$$\log L(\mu_{AA}, \mu_{Aa}, \mu_{aa}, \sigma^2|\mathbf{y}) = \sum_{i=1}^{n} \log \left[p\phi_{AA}(y_i) + q\phi_{Aa}(y_i) + (1-p-q)\phi_{aa}(y_i) \right].$$

If the gene were segregating in the ratio 1:2:1 then we would have an Mendelian F_2 population.

We can find the MLEs through differentiation. Although we will see that there is no explicit solution, the equations will lead to a nice iterative solution. This solution will have more general applicability, so to address that we will solve the likelihood equations for a general mixture model and then return to the example.

Given iid observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ from the general mixture model

$$Y_i \sim \sum_{j=1}^{k} p_j f(y_i|\theta_j), \quad \sum_j p_j = 1,$$

the log likelihood is

$$\log L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} p_j f(y_i|\theta_j) \right),$$

where we write $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

For now, we will assume that p and q are known, but we will return to this case later. Differentiating with respect to $\theta_{j'}$ gives

$$\frac{\partial}{\partial \theta_{j'}} \log L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} p_j \frac{\partial}{\partial \theta_{j'}} f(y_i|\theta_j)}{\sum_{j=1}^{k} p_j f(y_i|\theta_j)}.$$

Notice that if $\theta_{j'}$ does not contain any of the same components as θ_j, the derivative in the numerator will be zero unless $j = j'$. We will see this in detail in Example 2.5.

For convenience, we now write

$$\frac{\partial}{\partial \theta_{j'}} f(y_i|\theta_j) = f(y_i|\theta_j) \frac{\partial}{\partial \theta_{j'}} \log \left(f(y_i|\theta_j) \right)$$

$$P_j(y_i) = \frac{p_j f(y_i|\theta_j)}{\sum_{j=1}^{k} p_j f(y_i|\theta_j)},$$

and then we have

$$\frac{\partial}{\partial \theta_{j'}} \log L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} P_j(y_i) \frac{\partial}{\partial \theta_{j'}} \log \left(f(y_i|\theta_j) \right). \tag{2.3}$$

We can solve for the MLEs with the following iterative algorithm. Start with initial values $\theta_j^{(0)}, P_j^{(0)}(y_i)$ for $j = 1, \ldots, k$. For $t = 0, 1, \ldots$:

1. For $j = 1, \ldots, k$, set $P_j^{(t+1)}(y_i) = \frac{p_j f(y_i|\theta_j^{(t)})}{\sum_{j=1}^k p_j f(y_i|\theta_j^{(t)})}$.
2. For $j' = 1, \ldots, k$, solve for $\theta_{j'}^{(t+1)}$ in

$$\sum_{i=1}^{n} \sum_{j=1}^{k} P_j^{(t+1)}(y_i) \frac{\partial}{\partial \theta_{j'}} \log\left(f(y_i|\theta_j)\right) = 0.$$

3. Increment t and return to 1. Repeat until convergence.

Example 2.5. (**Normal Mixture**). If $f(y|\theta_j)$ is $N(\mu_j, \sigma^2)$, when differentiating with respect to $\mu_{j'}$, the numerator in equation (2.3) contains only one term, the one with $j = j'$. However, the differentiation with respect to σ^2 contains the entire sum, as the parameter σ^2 is common to all densities in the mixture. We have

$$\begin{aligned}
\sum_{i=1}^{n} P_j^{(t+1)}(y_i) \frac{\partial}{\partial \mu_j} \log\left(f(y_i|\mu_j, \sigma^2)\right) &= \sum_{i=1}^{n} P_j^{(t+1)}(y_i) \frac{1}{2\sigma^2}(y_i - \mu_j), \\
\sum_{i=1}^{n} \sum_{j=1}^{k} P_j^{(t+1)}(y_i) \frac{\partial}{\partial \sigma^2} \log\left(f(y_i|\mu_j, \sigma^2)\right) &= \sum_{i=1}^{n} \sum_{j=1}^{k} P_j^{(t+1)}(y_i) \\
&\quad \times \left[-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \mu_j)^2\right].
\end{aligned}$$

In setting these equations equal to zero and solving, we can treat each μ_j separately and get

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n y_i P_j^{(t+1)}(y_i)}{\sum_{i=1}^n P_j^{(t+1)}(y_i)}.$$

For σ^2, after substituting $\mu_j^{(t+1)}$ for μ_j,we have

$$\sigma^{2(t+1)} = \frac{\sum_{i=1}^n \sum_{j=1}^k P_j^{(t+1)}(y_i) \left(y_i - \mu_j^{(t+1)}\right)^2}{n \sum_{j=1}^k P_j^{(t+1)}(y_i)}.$$

Example 2.6. (**Tomato Data Revisited, Completed**). If the gene were segregating in the ratio 1:2:1 then we would have a Mendelian F_2 population. We now estimate the normal parameters under this scenario.

Using the iterations described above, we estimate

$$\hat{\mu}_{AA} = 125.5, \quad \hat{\mu}_{Aa} = 101.5, \quad \hat{\mu}_{aa} = 83.25, \quad \hat{\sigma}^2 = 0.324.$$

The sequence of iterations is shown in Fig. 2.1, where the rapid convergence of the algorithm can be seen. An `R` program to reproduce Figure 2.1 is given in Appendix B.2.

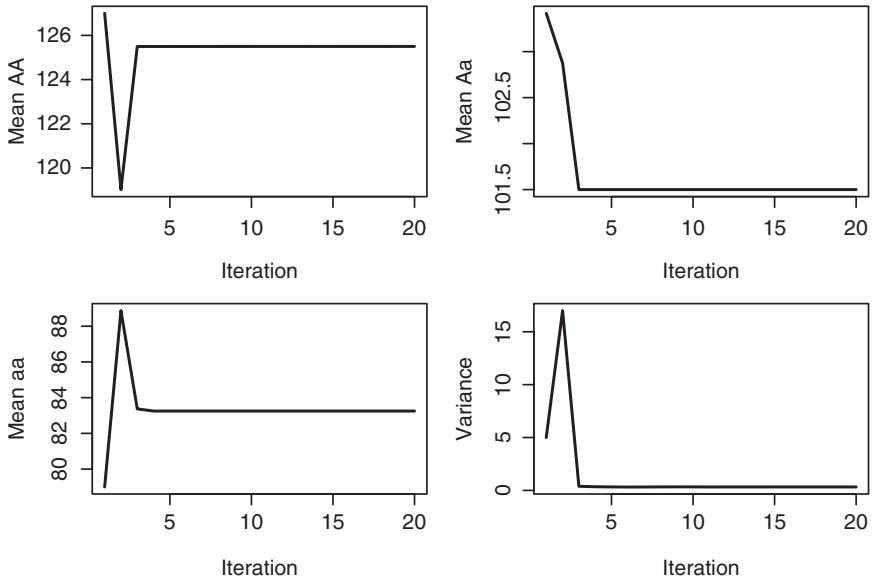

Fig. 2.1. Graphs of the iterations for the MLE of Example 2.6. The plots show the convergence of the estimates of the three means and the variance.

2.3 Hypothesis Testing

A major activity in statistical analysis is the testing of hypotheses. Here we review a number of approaches, both classical and more recent.

2.3.1 The Pearson Chi-Squared Test

We first look at a situation where there are two classes of genotypes. Suppose that there are N individuals with probability p of being in the first class. If the individuals are independent, the probability of observing n_1 individuals in the first class (and $n_2 = N - n_1$ in the second class) is given by the binomial distribution

$$P(n_1, n_2) = \frac{N!}{n_1! n_2!} p^{n_1} (1-p)^{n_2}.$$

The mean of the distribution is Np and the variance is $Np(1-p)$. In large samples, it is typical to approximate the binomial distribution by a normal distribution with the same mean and variance. Thus

$$Z = \frac{n_1 - Np}{(Np(1-p))^{1/2}} \tag{2.4}$$

is approximately a standard normal random variable, and hence Z^2 is approximately a chi-squared random variable with one degree of freedom. It can be shown that

$$Z^2 = \chi_1^2 = \frac{(n_1 - Np)^2}{Np} + \frac{(n_2 - N(1-p))^2}{N(1-p)}, \tag{2.5}$$

which is the usual formula for the *Pearson chi-squared statistic*; that is, the sum of the squares of the differences between observed (O) and expected counts (E) divided by expected counts:

$$\chi^2 = \sum \frac{(O-E)^2}{E}. \tag{2.6}$$

Note that in testing for the segregation ratio 1:1, equation (2.6) can be written $(n_1 - n_2)^2/N$. In fact, the segregation test can be based directly on the standard normal test using equation (2.4), which would produce the same result as the Pearson chi-squared test of equation (2.5). However, the Pearson chi-squared test has an advantage in that the test statistic (2.6) can be generalized to situations involving more than two categories of genotypes. In such cases, the multinomial distribution is used to model the observations, and the chi-squared approximation is given by equation (2.6).

In general, when there are m observed counts, the Pearson chi-squared statistic that arises from the calculation of m "expected counts" is asymptotically chi-squared with $m-k$ degrees of freedom. Here m is one less than the number of observed counts (cells) and k is the number of parameters to be estimated in the calculation of the expected counts (see Section 2.3.2).

For the backcross, in which there are two genotype classes, $m = 1$. This is because only one class can be filled arbitrarily, as once a number is assigned to the first class the number in the second class is determined. For the F_2, which has a total of three genotype classes at a codominant marker, $m = 2$.

If N is not very large, the binomial distribution may not be well-approximated by a normal. In such cases, it may be best to calculate an exact p-value (see Section 2.3.3). We can also use a continuity corrected χ^2 statistic:

$$\chi^2 = \sum \frac{(O-E)^2 - |O-E| + \frac{1}{4}}{E}, \tag{2.7}$$

which trades ease of use for some accuracy in the test.

Example 2.7. (**F_1 Hybrid Population**). We use an example from Yin et al. (2001), who constructed a genetic linkage map using molecular markers in an F_1 interspecific hybrid population between two different poplar species, Chinese quaking poplar and white poplar. The mapping population was comprised of 103 hybrid trees, each genotyped with RAPD markers. Some markers are heterozygous in one parent but null in the other, whereas other markers have an inverse pattern. Since these markers are segregating in the same pattern as two-way backcross markers do, such an F_1 population is called a *two-way pseudo-test backcross* (Grattapaglia and Sederoff 1994). In a two-way pseudo-test backcross population, two parent-specific maps can be constructed. In this example, six markers are chosen that formed linkage group 16 in the white poplar genetic linkage map (Yin et al. 2001).

Table 2.1. Pearson test statistic for testing Mendelian segregation 1:1 using RAPD markers in an interspecific poplar hybrid population.

			Pearson χ^2	
Marker	n_1	n_2	χ^2	p-value
I18_1090	44	59	2.184	0.139
W2_1050	45	58	1.641	0.200
AK12_2700	51	52	0.010	0.920
N18_1605	49	54	0.243	0.622
AK17_1200	40	63	5.135	0.023
I13_1080	41	62	4.282	0.038

Three different methods were used to test whether these six testcross markers follow the Mendelian segregation 1:1 in the mapping population. According to the Pearson chi-squared test (2.6), assuming a large sample size, markers I18_1090, W2_1050, AK12_2700, and N18_1605 follow the 1:1 ratio ($p > .05$), but markers AK17_1200 and I13_1080 do not (Table 2.1). This suggests that the latter two markers display possible distorted segregation.

The sample size here is actually rather large, and the continuity-corrected statistic (2.7) gave exactly the same answers. We also note that when doing multiple tests, there is the danger of rejection by chance alone. Thus, although we found two significant markers, it is best to consider this only evidence of significance, and further investigation should be done.

2.3.2 Likelihood Ratio Tests

The testing of hypotheses can be carried out very efficiently using likelihood methodology. We first consider a general setup where we observe iid observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from a population with density function $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ could be a vector of parameters. For example, $\boldsymbol{\theta} = (\mu, \sigma^2)$ if we are modeling normals, or $\boldsymbol{\theta} = (p_1, \ldots, p_k)$, $\sum_j p_j = 1$, if we are modeling probabilities.

Given a sample $\mathbf{y} = (y_1, \ldots, y_n)$, the likelihood function for these data is $L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^n f(y_i|\boldsymbol{\theta})$. To test a hypothesis of the form

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ vs. } H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0,$$

we evaluate the ratio of the maxima of the likelihood functions under both hypotheses:

$$\lambda(\mathbf{y}) = \frac{\max_{\boldsymbol{\theta}:\boldsymbol{\theta}=\boldsymbol{\theta}_0} L(\boldsymbol{\theta}|\mathbf{y})}{\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y})}.$$

The value of λ is less than 1 by construction since the denominator is calculating the maximum over a larger set and hence must be a bigger number. Moreover, values of λ close to 1 provide support for H_0, as then the restricted likelihood function is close to the unrestricted one, which in turn should be close to the truth. On the other hand, small values of λ lead to rejection of H_0.

To actually carry out the test, we could calculate a p-value as

$$P(\lambda(\mathbf{Y}) < \lambda(\mathbf{y})|H_0 \text{ is true}) = p(\mathbf{y}) \tag{2.8}$$

and reject H_0 for small values of $p(\mathbf{y})$, say $p(\mathbf{y}) < .01$. In general it is not an easy job to figure out the exact distribution of the probability in equation (2.8), but there is a famous approximate result that is often helpful (see Appendix A.1).

It is typical to transform λ to $-2\log\lambda$, and in this form we would now reject H_0 if $-2\log\lambda$ is large, the usual scenario. However, there is a second, more important consequence, described in the following result (see Appendix A.1 for technical details).

If $Y_1, \ldots, Y_n$ is a random sample from a density $f(x|\theta)$, then an approximate level α test of the hypothesis

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \notin \Theta_0$$

is to reject H_0 if

$$-2\log\lambda(\mathbf{X}) > \chi^2_{\nu,\alpha},$$

where $\chi^2_{\nu,\alpha}$ is the upper α cutoff from a chi-squared distribution with ν degrees of freedom, equal to the difference between the number of free parameters specified by H_0 and the number of free parameters specified by H_1.

So, for example, if f is $N(\mu, \sigma^2)$, the test of $H_0 : \mu = \mu_0, \quad \sigma^2 = \sigma_0^2$ vs. $H_0 : \mu \neq \mu 0, \sigma^2 \neq \sigma_0^2$ has two degrees of freedom, while the test of $H_0 : \mu = \mu_0$ vs. $H_0 : \mu \neq \mu 0$ since σ^2 is free in both hypotheses.

Example 2.8. (**First Testing Example**). As a first simple example, return to the situation of Example 2.2, where we have $Y_1, Y_2, \ldots, Y_n$ iid from an $N(\mu, \sigma^2)$ population. To test $H_0 : \mu = \mu_0$ vs. $H_0 : \mu \neq \mu 0$ we calculate

$$\begin{aligned}\lambda(\mathbf{y}) &= \frac{\max_{\mu=\mu_0} L(\mu, \sigma^2|\mathbf{y})}{\max_{\mu,\sigma^2} L(\mu, \sigma^2|\mathbf{y})} \\ &= \frac{\max_{\mu=\mu_0} L(\mu, \sigma^2|\mathbf{y})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{y})},\end{aligned}$$

where we see that the denominator maximum is attained by substituting the MLEs for the parameter values (see Example 2.3). In maximizing the numerator, we set $\mu = \mu_0$ and maximize in σ^2, giving

$$\max_{\mu=\mu_0} L(\mu, \sigma^2|\mathbf{y}) = L(\mu_0, \hat{\sigma}_0^2|\mathbf{y}),$$

where $\hat{\sigma}_0^2 = (1/n)\sum_{i=1}^n (y_i - \mu_0)^2$. This gives the LR statistic

$$\begin{aligned}\lambda(\mathbf{y}) &= \frac{L(\mu_0, \hat{\sigma}_0^2|\mathbf{y})}{L(\hat{\mu}, \hat{\sigma}^2|\mathbf{y})} \\ &= \frac{\left(\frac{1}{2\pi\hat{\sigma}_0^2}\right)^{n/2} \exp\left\{\frac{1}{2\hat{\sigma}_0^2}\sum_{i=1}^n (y_i - \mu_0)^2\right\}}{\left(\frac{1}{2\pi\hat{\sigma}^2}\right)^{n/2} \exp\left\{\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (y_i - \hat{\mu})^2\right\}} \\ &= \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2},\end{aligned}$$

which is, in fact, equivalent to the usual t–test.

We next look at another model and likelihood ratio test, based on the *multinomial distribution*. A classic experiment can be analyzed with this distribution and method.

Example 2.9. (**Morgan (1909) Data**). In 1909, Morgan experimented on fruit flies, crossing two inbred lines with the following genotypic traits:

Eye color	A:red	a:purple
Wing length	B:normal	b:vestigial

He then obtained 2839 crosses of $AABB \times aabb$ and observed the four genotypes $AaBb$, $Aabb$, $aaBb$, and $aabb$. (Note that the middle two genotypes are recombinants.)

A model for the Morgan experiment can be based on the multinomial distribution, a discrete distribution that is used to model frequencies. Suppose that a random variable Y can take on one of k values, the integers $1, 2, \ldots, k$, each with probability $p_1, p_2, \ldots, p_k$. More precisely,

$$P(Y = j) = p_j, j = 1, \ldots k.$$

Note that if $k = 2$ we have the binomial distribution.

If we now have an iid sample $Y_1, \ldots, Y_n$, and we let $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, where $\sum_j p_j = 1$, then the sample density (the likelihood function) is

$$L(\mathbf{p}|\mathbf{y}) = \prod_{i=1}^n f(y_i|\mathbf{p}) = p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

where n_j = number of $y_1, \ldots, y_n$ equal to j.

Example 2.10. (**Morgan (1909) Data Continued**). For the Morgan experiment, we have $k = 4$ categories, the four genotypes $AaBb$, $Aabb$, $aaBb$, and $aabb$. There are $n = 2839$ observations, which were observed to be

$AaBb$	$Aabb$	$aaBb$	$aabb$
1339	151	154	1195

so $n_1 = 1339$, etc.

Typical null hypotheses specify patterns in the cell probabilities p_j, with the hypothesis of *equal* cell probabilities being a popular one. That is, test

$$H_0 : p_1 = p_2 = \cdots = p_k \qquad \text{versus} \qquad H_1 : H_0 \text{is not true.}$$

Of course, under this H_0, all of the $p_j's$ equal $1/k$. As a slightly more interesting example, and one that is applicable to the Morgan experiment, suppose that $k = 4$ and we want to test

$$H_0 : p_1 = p_4, \quad p_2 = p_3 \text{ vs. } H_1 : H_0 \text{ is nottrue.} \tag{2.9}$$

using a likelihood ratio test. We proceed as before and calculate

$$\lambda(\mathbf{y}) = \frac{\max_{p_1=p_4,\quad p_2=p_3} L(\mathbf{p}|\mathbf{y})}{\max_{\mathbf{p}} L(\mathbf{p}|\mathbf{y})} = \frac{\max_{p_1=p_4,\quad p_2=p_3} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}}{\max_{\mathbf{p}} p_1^{n_1} p_2^{n_2} p_3^{n_3} p_4^{n_4}}. \tag{2.10}$$

To maximize the numerator, recall that $\sum_j p_j = 1$ implies that $p_k = 1 - \sum_{j=1}^{k-1} p_j$ (so there are really only $k-1$ parameters in the general problem). If $p_1 = p_4 = p/2$, then $p_2 = p_3 = (1-p)/2$, where p is the only unknown parameter. The numerator of equation (2.10) becomes

$$\max_p \left(\frac{p}{2}\right)^{n_1} \left(\frac{1-p}{2}\right)^{n_2} \left(\frac{1-p}{2}\right)^{n_3} \left(\frac{p}{2}\right)^{n_4} = \max_p \left(\frac{p}{2}\right)^{n_1+n_4} \left(\frac{1-p}{2}\right)^{n_2+n_3}.$$

This is a binomial likelihood, and taking logs and differentiating will show that the MLE of p under H_0 is $\hat{p}_0 = (n_1 + n_4)/(n_1 + n_2 + n_3 + n_4) = (n_1 + n_4)/n$. For the denominator, write $p_4 = 1 - p_1 - p_2 - p_3$ and take logs to get

$$L(\mathbf{p}|\mathbf{y}) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log(1 - p_1 - p_2 - p_3),$$

and differentiating with respect to p_1, p_2, p_3 shows that $\hat{p}_j = n_j/n, j = 1, \ldots, 4$. The likelihood test statistic then becomes

$$\lambda(\mathbf{y}) = \frac{\hat{p}_0^{n_1+n_4}(1-\hat{p}_0)^{n_2+n_3}}{\hat{p}_1^{n_1}\hat{p}_2^{n_2}\hat{p}_3^{n_3}\hat{p}_4^{n_4}} = \frac{\left(\frac{n_1+n_4}{2}\right)^{n_1+n_4}\left(\frac{n_2+n_3}{2}\right)^{n_2+n_3}}{n_1^{n_1} n_2^{n_2} n_3^{n_3} n_4^{n_4}}.$$

To perform the hypothesis test, we can use the approximation that $-2\log\lambda(\mathbf{y})$ has a chi-squared distribution. But first we must get the degrees of freedom correct.

Under H_1, there is no restriction on the parameters other than that they must sum to one, so there are three free parameters. Under H_0, we have placed an additional two restrictions, so there is one free parameter under H_0, and the chi square has $3 - 1 = 2$ degrees of freedom.

Example 2.11. (**Morgan (1909) Data–First Conclusion**). The four genotypes are nonrecombinant (*AaBb* and *aabb*) and recombinant (*Aabb* and *aaBb*), and the hypothesis (2.9) specifies that the nonrecombinant genotypes segregate with a common parameter, and the recombinant genotypes segregate with a common parameter, but these two common parameters need not be equal. The alternative hypothesis merely says that this is not so.

The observed value of $\lambda(\mathbf{y})$ is $\lambda(\mathbf{y}) = 0.0164$ with $-2\log\lambda(\mathbf{y}) = 8.217$, to be compared with a chi-squared distribution with two degrees of freedom. The .05 cutoff, $\chi^2_{2,.05} = 5.99$, leads us to reject the null hypothesis.

2.3.3 Simulation-Based Approach

The chi-squared approximation used in Section 2.3.2 relies on an asymptotic approximation – its validity is dependent on having large cell sizes. Moreover, the discrepancy in the size of the cells can also have an effect on the adequacy of the approximation.

If there is reason to suspect the adequacy of the approximation, or if the evidence in the data is difficult to interpret, it may be reasonable to try another approach to assessing the evidence against H_0. (Actually, this is a good idea in most cases.)

We describe in this section a simulation technique that is sometimes known as the *parametric bootstrap* (Efron and Tibshirani 1993); it is an all-purpose simulation-based technique. We first describe it in general, then apply it to the Morgan data.

Simulation-Based Hypothesis Assessment

Given an iid sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and a density $f(x|\theta)$, we assess the hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \notin \Theta_0$ as follows.

(1) Estimate θ with the MLE $\hat{\theta}$, and calculate the observed likelihood statistic $-2\log\lambda(\mathbf{y})$.
(2) Generate $t = 1, \ldots, M$ new iid samples $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)$, where $Y_i^* \sim f(x|\hat{\theta})$, and calculate $-2\log\lambda(\mathbf{y}_t^*)$.
(3) A p-value for the test can be calculated as

$$\hat{p}(\mathbf{y}) = \frac{1}{M}\sum_{t=1}^{M} I\left(\lambda(\mathbf{y}_t^*) > \lambda(\mathbf{y})\right),$$

where $I(\cdot)$ is the indicator function, which is equal to 1 if the argument is true and 0 otherwise. A histogram of the $\lambda(\mathbf{y}_t^*)$ can also be drawn.

Example 2.12. (**Morgan (1909) Data–Second Conclusion**). To do a simulation-based assessment of the hypotheses (2.9) we would simulate random variables Y_1^*, Y_2^*, ... from a multinomial distribution with $n = 2839$ and probability vector

$$\mathbf{p} = \left(\frac{1339}{2839}, \frac{151}{2839}, \frac{154}{2839}, \frac{1195}{2839}\right) = (0.471, 0.054, 0.053, 0.421).$$

Figure 2.2 shows the results of the simulation of 10000 values of $-2\log\lambda(\mathbf{y}^*)$, and they are quite different from the results of the chi-square test of Example 2.11. The observed value of $-2\log\lambda(\mathbf{y}) = 8.217$ is now right in the middle of the distribution, with an estimated p-value of .5718, meaning that 5718 of the 10000 random variables simulated were larger than the observed value of the test statistic. This puts the statistic right in the middle of the distribution and thus leads us to accept the null hypothesis.

It should be mentioned that the overall conclusion is not crystal clear, but the evidence is certainly pointing toward the conclusion that H_0 is a tenable hypothesis, and the asymptotic approximation of the chi-square distribution is not the best in this case.

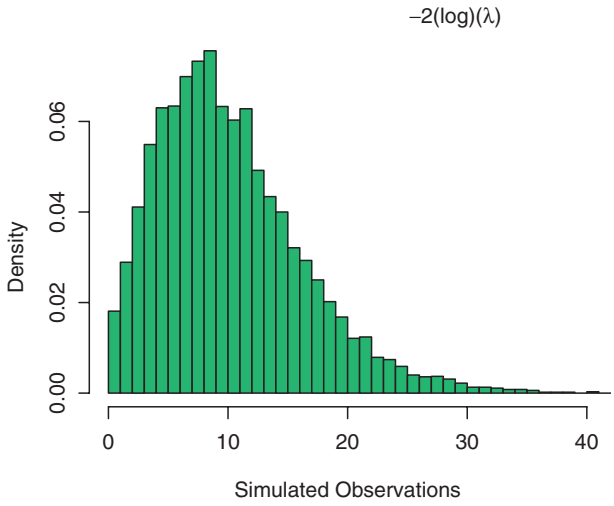

Fig. 2.2. Histogram of 10000 values of the likelihood ratio statistic $-2\log\lambda$ for the Morgan data.

2.3.4 Bayesian Estimation

Throughout this chapter, we have been describing the classical approach to statistics, where we base our evidence on a repeated-trials assessment of error. There is an alternative approach, the Bayesian approach, which is fundamentally different from the classical approach. Here the assessment is based on an experimenter's prior belief and how that belief is altered by the data.

It is not constructive to view the two approaches in opposition. Rather, it is better to examine each problem at hand, and choose the approach that will give the most useful answer.

In the classical approach, the parameter, say θ, is thought to be an unknown, but fixed, quantity. A random sample $X_1, \ldots, X_n$ is drawn from a population indexed by θ and, based on the observed values in the sample, knowledge about the value of θ is obtained. In the Bayesian approach, θ is considered to be a quantity whose variation can be described by a probability distribution (called the *prior distribution*). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen. A sample is then taken from a population indexed by θ, and the prior distribution is updated with this sample information. The updated prior distribution is called the *posterior distribution*.

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\mathbf{x}|\theta)$, then the posterior distribution, the conditional distribution of θ given the sample, $\mathbf{x}$, is calculated using Bayes' Rule as

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)d\,\theta} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}, \tag{2.11}$$

where $m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$ is the marginal distribution of $\mathbf{X}$.

Once $\pi(\theta|\mathbf{x})$ is obtained, it contains all of the information about the parameter θ. We can plot this distribution to see the shape, and perhaps where the modes are, and whether it is symmetric or not. It is also typical to calculate the posterior mean and variance to get a point estimator and a measure of spread. These are given by

$$\mathrm{E}(\theta|\mathbf{x}) = \int \theta\pi(\theta|\mathbf{x})d\theta, \quad \mathrm{Var}(\theta|\mathbf{x}) = \int [\theta - \mathrm{E}(\theta|\mathbf{x})]^2\pi(\theta|\mathbf{x})d\theta.$$

We illustrate Bayesian estimation with some examples.

Example 2.13. (**Bayes Estimation in the Normal Distribution**). Let $X \sim n(\theta, \sigma^2)$, and suppose that the prior distribution on θ is $n(\mu, \tau^2)$. (Here we assume that σ^2, μ, and τ^2 are all known.) The posterior distribution of θ is

$$\begin{aligned}\pi(\theta|x) &\propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\theta-\mu)^2} \\ &\propto \left[\frac{\sqrt{\sigma^2+\tau^2}}{\sqrt{2\pi\sigma^2\tau^2}} e^{-\frac{\sigma^2+\tau^2}{2\sigma^2\tau^2}(\theta-\mathrm{E}(\theta|x))^2}\right] \left[\frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} e^{-\frac{1}{2\sigma^2\tau^2}(x-\mu)^2}\right],\end{aligned}$$

where the distribution in the first square brackets is the posterior and the second distribution is the marginal. The calculation is somewhat long and tedious, and depends on completing the square in the exponent.

Upon inspection, we see that the posterior distribution of θ is also normal, with mean and variance given by

$$\begin{aligned}\mathrm{E}(\theta|x) &= \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \\ \mathrm{Var}\,(\theta|x) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.\end{aligned} \tag{2.12}$$

The Bayes estimator is a linear combination of the prior and sample means. Notice also that as τ^2, the prior variance, is allowed to tend to infinity, the Bayes estimator tends toward the sample mean. We can interpret this as saying that as the prior information becomes more vague, the Bayes estimator tends to give more weight to the sample information. On the other hand, if the prior information is good, so that $\sigma^2 > \tau^2$, then more weight is given to the prior mean.

Example 2.14. (**Bayes Estimation in the Binomial**). Let $X_1, \ldots, X_n$ be iid Bernoulli (p). Then $Y = \sum X_i$ is binomial(n, p). We take the prior distribution on p to be

$$\mathrm{beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

The posterior distribution of p is

$$\pi(p|y) \propto \left[\binom{n}{y} p^y (1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]$$
$$\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

which is again a beta distribution, now with parameters $y+\alpha$ and $n-y+\beta$. We can estimate p with the mean of the posterior distribution

$$\mathrm{E}(p|y) = \frac{y+\alpha}{\alpha+\beta+n}.$$

Consider how the Bayes estimate of p is formed. The prior distribution has mean $\alpha/(\alpha+\beta)$, which would be our best estimate of p without having seen the data. Ignoring the prior information, we would probably use the MLE $p = y/n$ as our estimate of p. The Bayes estimate of p combines all of this information. The manner in which this information is combined is made clear if we write $\mathrm{E}(p|y)$ as

$$\mathrm{E}(p|y) = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right).$$

Thus p_{B} is a linear combination of the prior mean and the sample mean, with the weights being determined by α, β, and n.

Example 2.15. (**Bayes Estimation in the Binomial**). We revisit the data of Example 2.7 and now estimate p, the true proportion, using a Bayes estimator. Consider marker I18_1090 with counts of 44 and 59 in the two genotype classes. The MLE of the true proportion in the first genotype is $44/(44+59) = .427$.

Suppose we estimate p using a Bayes estimator. If the experimenter believes that the genes are segregating independently, then he believes that $p = 1/2$. We can reflect this belief with a beta distribution that is symmetric around $1/2$. We choose a beta $(2,2)$, which does not give much weight to the prior distribution. Under this prior distribution the Bayes estimator of p is $(44+2)/(44+59+2+2) = .430$, which is very similar to the MLE. Thus, for this choice of prior distribution, the information in the data is very strong and the Bayes estimator is virtually the same as the MLE.

If the experimenter is very certain that p is close to $1/2$, this can be reflected in the prior distribution by increasing the parameter values. If we choose $\alpha = \beta = 100$, the prior mean remains $1/2$ but the prior variance is decreased. The resulting Bayes estimator is $(44+100)/(44+59+200) = .475$, and the posterior distribution is now quite symmetric. This is illustrated in Fig. 2.3, where we see the symmetry of the prior distributions, but one is more concentrated. The resulting posterior distributions are close to the likelihood function and skewed when $\alpha = \beta = 2$ and symmetric and far from the likelihood function when $\alpha = \beta = 100$.

2.4 Exercises

2.1 (a) Illustrate the binomial approximation to the normal as described in Section 2.3.1. For $N = 5, 15, 50$ and $p = .25, .5$, draw the binomial histogram and the normal density. Calculate the .05 and .01 tail area in each case.

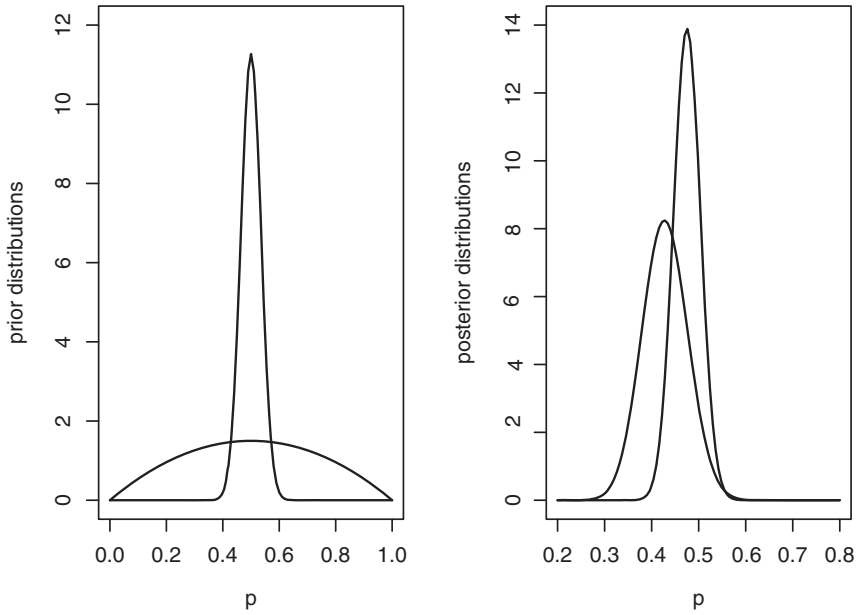

Fig. 2.3. Graphs of prior distributions (left panel) and posterior distributions (right panel) of Example 2.15. The beta $(2, 2)$ prior is very flat and results in a skewed posterior that is indistinguishable from the likelihood function. The beta $(100, 100)$ prior is very peaked and results in a peaked and symmetric posterior distribution.

(b) Show that if $p = \frac{1}{2}$, equation (2.6) can be simplified to $(n_1 - n_2)^2/N$.

2.2 For the situation of Example 2.8:

(a) Verify the formula for $\hat{\sigma}_0^2$.

(b) Show that

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \mu_0)^2} = \frac{1}{1 + \frac{(\bar{y}-\mu_0)^2}{\hat{\sigma}^2}}.$$

(For the second equality, use the identity $\sum_{i=1}^n (y_i - \mu_0)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu_0)^2$.

(c) Verify that, in this last Form, we will reject H_0 if $|\bar{y} - \mu + 0|/\hat{\sigma}$ is large, making this test equivalent to the t–test.

2.3 (a) Verify the MLEs of equation (2.10).

Springer

http://www.springer.com/978-0-387-20334-8

Statistical Genetics of Quantitative Traits
Linkage, Maps and QTL
Wu, R.; Ma, C.; Casella, G.
2007, XVI, 368 p., Hardcover
ISBN: 978-0-387-20334-8