11
Null hypotheses and p-values

11.1 The null value of a parameter

With most probability models there is one particular value of the parameter which corresponds to there being *no effect*. This value is called the *null* value, or *null hypothesis*. For a parameter θ we will denote this null value by $\theta_\varnothing$. In classical statistical theory, considerable emphasis is placed on the need to disprove (or reject) the null hypothesis before claiming positive findings, and the procedures which are used to this end are called *statistical significance tests*. However, the emphasis in this theory on accepting or rejecting null hypotheses has led to widespread misunderstanding and misreporting in the medical research literature. In epidemiology, which is not an experimental science, the usefulness of the idea has been particularly questioned. Undoubtedly the idea of statistical significance testing has been overused, at the expense of the more useful procedures for *estimation* of parameters which we have discussed in previous chapters. However, it remains useful. A null hypotheses is a *simplifying* hypothesis and measuring the extent to which the data are in conflict with it remains a valuable part of scientific reasoning. In recent years there has been a trend away from a making a straight choice between accepting or rejecting the null hypothesis. Instead, the *degree* of support for the null hypothesis is measured, for example using the log likelihood ratio at the null value of the parameter.

EXAMPLE: GENETIC LINKAGE BY THE SIB PAIR METHOD

We shall illustrate the methods of this chapter with a simple statistical problem arising in the detection of *linkage* between a genetic marker and a gene which carries an increased susceptibility to a disease. At the marker locus each offspring receives one of two possible haplotypes from the mother and one of two possible haplotypes from the father. If there are many possible haplotypes we can safely assume that the mother and father together have four *different* marker haplotypes. The marker is then said to be highly *polymorphic*. If the mother has haplotypes (a,b) and the father (c,d), possible haplotype configurations for offspring are (a,c), (a,d), (b,c), and (b,d). If inheritance of the marker obeys Mendelian laws, the probability that

Table 11.1. Linkage of the HLA locus to nasopharyngeal cancer susceptibility

Haplotypes shared	Number of sib pairs	Probability (null value)
2	16	0.25
1	8	0.50
0	3	0.25

two siblings have completely different marker haplotypes (no haplotypes in common) is 0.25 and the probability that they have the same pair of haplotypes (two haplotypes in common) is also 0.25. The remaining possibility is that they have one marker haplotype in common, which has probability 0.50.

If we deliberately choose two siblings who are both affected by the disease, then these siblings will be more similar in that part of the genome surrounding the disease susceptibility gene than we would expect by chance. If the marker locus is in this vicinity, then the probabilities that two affected sibs will share 0, 1, or 2 marker haplotypes will depart from the (0.25, 0.5, 0.25) split indicated above. This way of looking for genetic linkage is called the *affected sib pair method*. If disease susceptibility is conferred by a *dominant* gene, it can be shown that the main effect of linkage is to reduce the probability of the affected sibs sharing no marker haplotypes and to increase the probability of their sharing both, while the probability of their sharing one marker haplotype is scarcely affected. A simple and reasonably efficient statistical analysis may therefore be carried out by disregarding the pairs sharing one marker haplotype.

Table 11.1 shows the frequency of shared HLA haplotypes amongst 27 pairs of sibs affected by nasopharyngeal carcinoma.* Assuming dominant inheritance of the disease susceptibility gene and ignoring the 8 sib pairs with only one marker gene in common leaves $N = 19$ pairs, 16 of which share both haplotypes, and 3 of which share no haplotypes. Let Ω be the odds that a pair shares both rather than no haplotypes. The log likelihood for Ω is

$$16\log(\Omega) - 19\log(1+\Omega).$$

The most likely value of Ω is $16/3 = 5.33$, so that the maximum value of the log likelihood is

$$16\log(5.33) - 19\log(6.33) = -8.29$$

*From Day, N.E. and Simons, J. (1976) *Tissue Antigens*, **8**, 109–119.

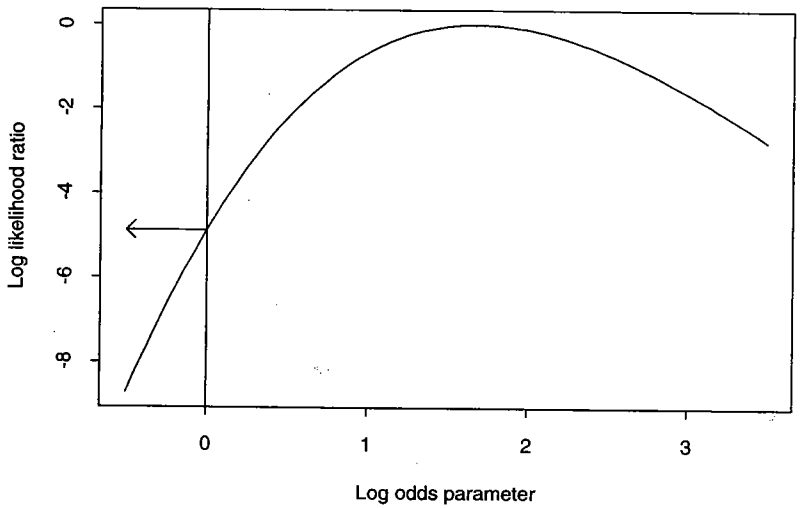

Fig. 11.1. Log likelihood ratio for HLA linkage.

and the log likelihood ratio for any other value of Ω is

$$16\log(\Omega) - 19\log(1+\Omega) - (-8.29).$$

Fig. 11.1 shows the log likelihood ratio plotted against $\log(\Omega)$.

Under the null hypothesis that there is no linkage, the two outcomes are equally probable, so the null value of Ω is 1.0 and the null value for $\log(\Omega)$ is 0. This is indicated in Fig. 11.1 by the vertical line. The log likelihood ratio for $\Omega = 1$ is

$$16\log(1) - 19\log(2) - (-8.29) = -4.88$$

(indicated on the graph with an arrow). The null value of Ω does not fall within the range which we have regarded as supported.

Whether the mode of inheritance of disease susceptibility is dominant or recessive must be established in studies of extended families. If it is dominant, the likelihood ratio test described above provides an efficient test of linkage. However, if the disease susceptibility gene is *recessive*, the probability that affected sibs will share one marker haplotype in common is also reduced and a more efficient test for linkage examines the 16:11 split between 2 and < 2 shared haplotypes. In this case the null value of the odds parameter Ω is $0.25/0.75 = 0.333$.

Exercise 11.1. If the evidence for Ω is based on the 16:11 split of sib pairs, find the log likelihood ratio for $\Omega = 0.333$.

11.2 Log likelihood ratios and p-values

As with the supported range for a parameter a general need is felt to measure support for the null hypothesis on the more familiar scale of probability. The way this is done in frequentist statistical theory is very similar to the way in which coverage probabilities are calculated for confidence intervals (see Chapter 10). We imagine a large number of repetitions of the study with the parameter equal to its null value and define the *p-value* as the proportion of these studies which provide less support for the null value than the data actually observed. If the p-value is small the data are at odds with the null hypothesis and the finding is said to be *statistically significant*. If the p-value is large, the finding is said to be *not statistically significant*. Traditionally the value $p = 0.05$ has been used to divide significant from non-significant results, but the modern practice is to report the actual p-value, particularly when it lies in the range 0.001 to 0.10. Outside this range it is enough to give the p-value as $p < 0.001$ or $p > 0.10$.

The argument which defines the p-value closely follows that used to define the coverage probability of a supported range in Chapter 10. As in that case, we shall start with the problem of drawing conclusions about the value of the Gaussian mean, μ, on the basis of a single observation, x. In this case, the value of the log likelihood ratio for a null value $\mu_\varnothing$ is equal to

$$-\frac{1}{2}\left(\frac{x-\mu_\varnothing}{\sigma}\right)^2.$$

Exercise 11.2. You observe a value $x = 116$ and wish to test the hypothesis that it was obtained from a Gaussian distribution with mean $\mu = 100$ (the null value). Assuming that σ is known to take the value 10, what is the value of the log likelihood ratio at the null value?

We imagine a large number of repetitions of the study when the null hypothesis is true. The p-value is the proportion of such repetitions with log likelihood ratios less than this observed value. One way that the p-value can be calculated is by computer simulation of such repetitions of the study.

Exercise 11.3. Such a simulation is envisaged in Exercise 10.1. Of the first four values generated, what proportion have log likelihood ratios at the null value less than that observed?

This is a very inaccurate estimate of the p-value. An accurate estimate would, of course, require several thousand repetitions to be generated.

The method of generating a p-value by computer simulation is known as a *Monte Carlo* test and it is quite widely used. However, in this case we do not need to resort to the computer as we can work out the p-value theoretically. If X represents the value obtained in such a repetition, the p-value is defined as the probability that this yields a smaller log likelihood

ratio than that observed, that is,

$$\Pr\left[-\frac{1}{2}\left(\frac{X-\mu_\oslash}{\sigma}\right)^2 < \text{Observed log likelihood ratio}\right].$$

This is the same as

$$\Pr\left[\left(\frac{X-\mu_\oslash}{\sigma}\right)^2 > -2 \times (\text{Observed log likelihood ratio})\right],$$

and since we are assuming that the null hypothesis is true in such repetitions, the above probability is obtained by referring

$$-2 \times (\text{Observed log likelihood ratio})$$

to the chi-squared distribution on one degree of freedom.

Exercise 11.4. Use the table of the chi-squared distribution in Appendix D to find the p-value for the example of Exercise 11.2

For N observations from a Gaussian distribution, the same rule for obtaining the p-value holds, the value of minus twice the log likelihood ratio now being

$$\left(\frac{M-\mu_\oslash}{S}\right)^2$$

where M is the mean of the N observations and $S = \sigma/\sqrt{N}$.

This relationship between the log likelihood ratio and the p-value holds *approximately* for non-Gaussian log likelihoods. The approximation will be adequate providing there is a sufficient amount of data to ensure that the log likelihood curve is approximately quadratic.

In our example of testing for genetic linkage, using the method most appropriate for dominant inheritance, the log likelihood ratio at the null parameter value is -4.88 so that

$$-2 \times (\text{log likelihood ratio}) = 9.76.$$

The probability of this being exceeded in a chi-squared distribution on one degree of freedom is 0.0018, so that the p-value is approximately 0.002. This is an example of a *log likelihood ratio test.*

Exercise 11.5. Use tables of the chi-squared distribution to find the p-value corresponding to the log likelihood ratio calculated in Exercise 11.1.

There are two other approximate methods of obtaining p-values which are widely used. These are called *Wald* tests and *score* tests, and both involve

quadratic approximations to the log likelihood curve. The problem of calculating exact p-values when these approximate methods cannot be used will be discussed in Chapter 12.

11.3 Wald tests

The first quadratic approximation we shall consider is the same as that used for approximate confidence intervals in Chapter 9. For a parameter, θ, the log likelihood is approximated by the quadratic curve

$$-\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2$$

where M is the most likely value of the parameter and S is the standard deviation of the Gaussian approximation, calculated from the curvature of log likelihood at its peak. This provides the closest possible approximation *in the region of the most likely value.* Using this approximation, the approximate value of minus twice the log likelihood ratio at the null value, $\theta_\oslash$, is

$$\left(\frac{M-\theta_\oslash}{S}\right)^2$$

For the log odds parameter of the Bernoulli likelihood, Ω, the values of M and S are

$$\begin{aligned} M &= \log\left(\frac{D}{N-D}\right) \\ S &= \sqrt{\frac{1}{D}+\frac{1}{N-D}}. \end{aligned}$$

For the log likelihood shown in Fig. 11.1,

$$\begin{aligned} M &= \log\left(\frac{16}{3}\right) = 1.674 \\ S &= \sqrt{\frac{1}{16}+\frac{1}{3}} = 0.629. \end{aligned}$$

The approximate log likelihood ratio curve corresponding to these values is shown in Fig. 11.2 (broken lines). The arrow indicates the approximate log likelihood ratio at the null value, $\log(\Omega) = 0.0$,

$$-\frac{1}{2}\left(\frac{1.674-0.0}{0.629}\right)^2 = -3.54$$

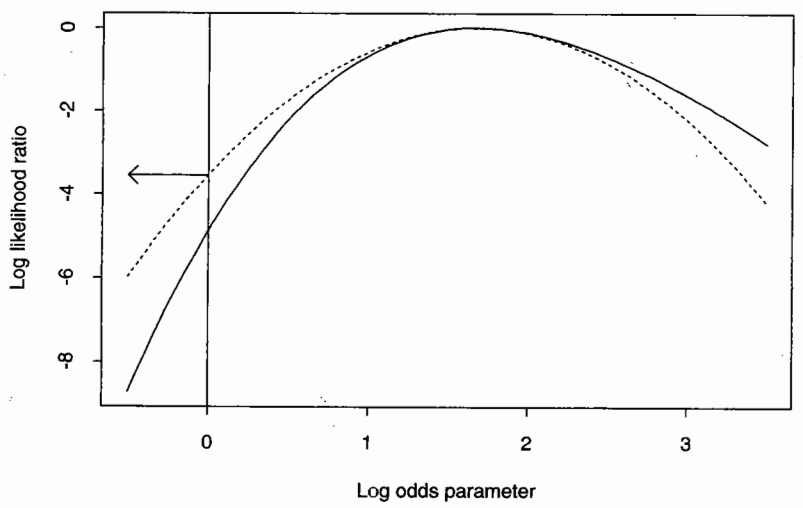

Fig. 11.2. The Wald test.

The approximate value of minus twice the log likelihood ratio is

$$\left(\frac{1.674 - 0.0}{0.629}\right)^2 = 7.08$$

and referring this value to the chi-squared distribution yields an approximate p-value of 0.008. This method of obtaining an approximate p-value is called the *Wald test.*

Exercise 11.6. Carry out the Wald test which approximates the log likelihood ratio of Exercise 11.1.

11.4 Score tests

The second quadratic approximation to the log likelihood ratio which we consider is based on the gradient and curvature of the log likelihood curve at the null value of the parameter. This is the most accurate quadratic approximation *in the region of the null value.* This approximation to the log likelihood ratio of Fig. 11.1 is shown in Fig. 11.3. Here we have displaced the true log likelihood ratio curve upwards in order to demonstrate that the true and approximate curves are the same shape in the region of the null value.

If U is the gradient of the log likelihood at the null value of the parameter, $\theta_\varnothing$, and V is minus the curvature (also at the null value), then this

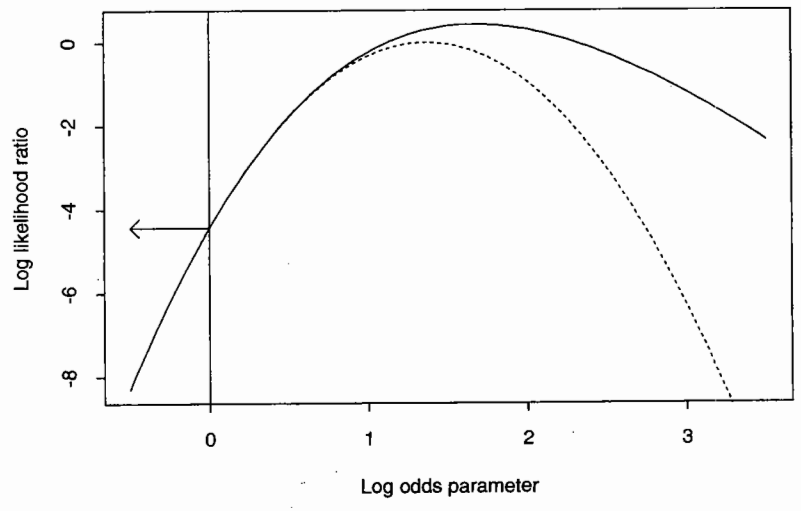

Fig. 11.3. The score test.

approximation to the log likelihood ratio is given by the formula

$$-\frac{V(\theta - \theta_\varnothing - U/V)^2}{2}.$$

This approximate curve has its maximum value at $\theta_0 + U/V$ and minus twice the log likelihood ratio at $\theta = \theta_\varnothing$ is

$$\frac{(U)^2}{V}.$$

The gradient, U, is called the *score* and we shall call V the *score variance.* The approximate score test is carried out by comparing $(U)^2/V$ with the chi-squared distribution with one degree of freedom.†

For the Bernoulli log likelihood in terms of the log odds parameter, $\log(\Omega)$, the score and score variance at the null value $\Omega_\varnothing$ are most easily expressed in terms of the null value of the probability parameter,

$$\pi_\varnothing = \frac{\Omega_\varnothing}{1 + \Omega_\varnothing}.$$

†The score test is usually carried out using the *expected value* of V (worked out assuming the null hypothesis to be true). In the applications discussed in this book this is not usually possible, and we have defined the score test in terms of the observed value of V.

They are

$$\begin{aligned} U &= D - N\pi_\varnothing, \\ V &= N\pi_\varnothing(1-\pi_\varnothing). \end{aligned}$$

In our example, $D = 16$, $N = 19$, and $\pi_\varnothing = 0.5$ so that

$$\begin{aligned} U &= 16 - 9.5 = 6.5 \\ V &= 19 \times 0.5 \times 0.5 = 4.75 \end{aligned}$$

The score test is $(6.5)^2/4.75 = 8.89$ and the probability that chi-squared exceeds this value is 0.003.

Exercise 11.7. Carry out the score test which approximates the log likelihood ratio of Exercise 11.1.

11.5 Which method is best?

The methods for calculating p-values given in this chapter are approximate except for the special case of a Gaussian likelihood with known standard deviation σ, when the three methods coincide and yield exact p-values. In other cases, where the log likelihood is roughly quadratic, the approximations to the p-value are good and the three methods give similar answers. When the three methods give seriously different answers this means that the quadratic approximations are not sufficiently close to the true log likelihood curve over the region stretching from the null value of the parameter to the most likely value. Of course, if the most likely value and the null value are very far apart, the curve is very difficult to approximate. In this situation, all three methods will give very small p-values and although these may differ substantially from on another, the choice of statistical method would not affect our scientific conclusions. This is the case in our example in which the three methods gave p-values of 0.002, 0.008, and 0.003.

The log likelihood ratio test is the only one of the three tests which remains the same when the parameter is transformed, and is to be preferred in general. The approximate equivalence of the other two tests to the log likelihood ratio test depends on the quadratic approximation, and will be improved by choosing an appropriate scale for the parameter. In particular, for parameters such as the *odds*, or the *rate*, which can take only positive values, it is better to calculate Wald and score tests in terms of the log parameter. If the three methods differ seriously, even after choosing an appropriate scale for the parameter, it is usual to advise the use of exact p-values. Methods for calculating these will be discussed in Chapter 12, but these are not without their difficulties.

11.6 One-sided p-values ⋆

We have defined the p-value as the probability that, when the null hypothesis is true, a repeated study will provide less support for the null value of the parameter than did the study actually observed. We have measured support for the null value of the parameter as the difference between the log likelihood at the null value and the log likelihood at the most likely value. This is satisfactory when the model allows the parameter to take any value within its natural range, but needs to be redefined if the model allows the parameter to vary only within a restricted range. In our HLA linkage example, if Ω is the odds that a sib pair shares both haplotypes rather than neither, the null value is $\Omega = 1$ and linkage is indicated by values in the range $\Omega > 1$. Values in the range $\Omega < 1$ are not allowed in a model for genetic linkage. In these circumstances, the value of Ω which is best supported by a study in which 5 sib pairs are found to share both haplotypes and 10 sib pairs to share neither is no longer 5/10, since this parameter value is not allowed by the model. The best supported value amongst *allowable* values is $\Omega = 1$. Thus only studies in which the split is in the expected direction would be regarded as providing evidence against the null hypothesis. The p-value calculated from this viewpoint is called a *one-sided* p-value, while the more usual p-value appropriate when the model allows the parameter to take values to both sides of the null value is called a *two-sided* p-value.

Approximate one-sided p-values can be obtained in most circumstances by simply halving the corresponding two-sided p-value. This follows from the fact that approximately half of the hypothetical repetitions of the study under the null hypothesis would lead to results in the wrong direction and, in a one-sided test, these would not be treated as evidence against the null value. In our example, the log likelihood ratio test for linkage gave $p \approx 0.0018$ and the approximate one-sided p-value is 0.0009.

The assumption that the probability model only allows its parameter to take on values to one side of the null value is a strong one and rarely justified in practice. Thus, one-sided p-values should only be used in exceptional circumstances. The genetic linkage example is one of these.

11.7 Tests for the rate parameter

We have described the three methods for obtaining approximate p-values using a null hypothesis which concerns the parameter of a simple binary probability model. These methods were all based on the Bernoulli likelihood. In this section we shall describe the corresponding methods for the rate parameter, λ, for a cohort study. Here the log likelihood takes the Poisson form:

$$D\log(\lambda) - \lambda Y,$$

where D is the number of failures observed and Y is the person-years observation.

The log likelihood ratio test for the null value $\lambda = \lambda_\varnothing$ compares the log likelihood at $\lambda_\varnothing$ with the log likelihood at $\lambda = D/Y$, the most likely value. The log likelihood ratio is, therefore,

$$[D\log(\lambda_\varnothing) - \lambda_\varnothing Y] - \left[D\log\left(\frac{D}{Y}\right) - \frac{D}{Y}Y\right]$$

which simplifies to

$$-D\log\left(\frac{D}{E}\right) + (D-E),$$

where $E = \lambda_\varnothing Y$ is the 'expected' number of failures obtained by multiplying the null value of the rate parameter by the person-years observation in the study. Minus twice this value can be compared with the chi-squared distribution with one degree of freedom.

The Wald test is based on the best Gaussian approximation to the log likelihood in the region of the most likely value. It is best carried out on the $\log(\lambda)$ scale, where $M = \log(D/Y)$ and $S = \sqrt{1/D}$.

Finally, the score test is based on the best Gaussian approximation to the log likelihood in the region of $\lambda_\varnothing$. Some simple calculus shows that the score and score variance (on the $\log(\lambda)$ scale) are given by

$$U = D - E, \qquad V = E,$$

so that the score test is $(D-E)^2/E$.

The null hypothesis most frequently of interest is that the rate in the cohort is no different from the rate in a *reference population*. Typically this reference rate is based on official statistics for a whole country and is estimated from so many events that it can be assumed to be a known constant. In practice the expected number of failures is usually calculated separately for different age bands and summed and E refers to the total expected number added over age bands. In Chapter 15 we show that the theory described above extends without change to this situation.

Exercise 11.8. In the vicinity of a nuclear reprocessing plant, 4 cases of childhood leukaemia were observed over a certain period while, from national registration rates, we would have expected only 0.25. Compare the log likelihood ratio and score tests of the null hypothesis that the incidence rates of leukaemia in the area do not differ from the national rates.*

In this case the two methods differ considerably, although both suggest a very small p-value. This reflects the fact that D is very small and the

*These data are discussed in detail by Gardner, M.J. (1989) *Journal of the Royal statistical Society, Series A*, **152**, 307–326.

Gaussian approximations are unreliable. We shall discuss methods for use in such situations in Chapter 12.

11.8 Misinterpretation of p-values

Reporting of p-values has come into disfavour because they have been widely misinterpreted. Although the same is true of confidence intervals, the nature of the misinterpretation of these is much less serious.

Most scientists interpret the 90% confidence interval as a range within which there is a 90% *probability* that the parameter value lies. We saw in Chapter 10 that, in the frequentist view of statistics, this is not correct — such an interpretation requires probability to be interpreted in terms of subjective degree of belief. In practice, however, it is not a serious error and does not usually lead to serious scientific misjudgement. The corresponding misinterpretation of the p-value, as the probability that the null hypothesis is *true*, is a much more serious error. Small studies which should be quite unconvincing are quoted as strongly negative findings because they have large p-values. The fact that this error is still widespread is the main reason why many authors currently discourage the use of p-values.

11.9 Lod scores and p-values ★

Our example in this chapter concerns genetic linkage and geneticists have taken a rather different approach to measuring the amount of evidence against the null hypothesis. Typically the result of a linkage analysis is presented as a *lod score* defined in terms of the log (base 10) likelihood for a parameter, θ, where this is defined as one minus the probability that two genes are passed from parent to offspring together. This probability is 0.5 when the two loci are unlinked but greater than 0.5 when there is linkage. Thus the null value of θ, which is called the *recombination fraction*, is 0.5 and linkage is represented by $\theta < 0.5$. The lod score for any specified value of θ compares the log likelihood with its value at $\theta = 0.5$. It is conventional to consider linkage to have been demonstrated if the most likely value of θ is less than 0.5 and gives a lod score greater than 3.0.

Using the relationship between the different systems of logarithms explained in Appendix A, a lod score of 3.0 corresponds to

$$-2 \times (\text{log likelihood ratio}) = 13.82$$

and, referring this to the chi-squared distribution on one degree of freedom shows this to be approximately equivalent to a p-value of 0.0002. However, since we are only interested in values of θ less than 0.5, the test is one-sided and this value must be halved to yield $p \approx 0.0001$. This is much smaller than we would require p-values to be in other areas of research, and it would appear that geneticists are much more difficult to dissuade from the null

hypothesis than other scientists. This is usually justified on the grounds that the human genome is immense and, *a priori*, it is very unlikely that any one marker locus is linked to a disease susceptibility gene. This argument has considerable force when searching a large number of markers in a 'blind fishing expedition', but would not hold if there were good *a priori* reasons to suspect linkage in a specified region. The interpretation of lod scores, like that of p-values, must take account of the scientific context and rigid criteria should be avoided.

Solutions to the exercises

11.1 At the most likely value, $\Omega = 16/11 = 1.455$, the log likelihood is

$$16\log(1.455) - 27\log(2.455) = -18.249$$

while at the null value $\Omega = 0.333$, the log likelihood is

$$16\log(0.333) - 27\log(1.333) = -25.354.$$

The log likelihood ratio at the null value is therefore

$$-25.354 - (-18.249) = -7.105.$$

11.2 The value of the log likelihood ratio at $\mu = 100$ is

$$-\frac{1}{2}\left(\frac{116-100}{10}\right)^2 = -1.28.$$

11.3 The first four observations of the computer simulation were 104, 115, 82 and 92 and the solution to Exercise 10.1 showed that the corresponding values of the log likelihood ratio at $\mu = 100$ are -0.08, -1.125, -1.62 and -0.32. Only 1 of these is less than the observed log likelihood ratio — a proportion of 0.25.

11.4 The value of minus twice the observed log likelihood ratio is 2.56 and referring this to the table of the chi-squared distribution in Appendix D shows the p-value to be a little over 0.10.

11.5 Minus twice the log likelihood ratio is 14.21. This corresponds to a very small p-value, 0.00016. Such results are usually reported as $p < 0.001$.

11.6 The most likely value of the log odds parameter is

$$M = \log(16/11) = 0.375,$$

and the standard deviation of the Gaussian approximation to the log likelihood around M is

$$S = \sqrt{\frac{1}{16} + \frac{1}{11}} = 0.392.$$

The null value of the log odds is $\log(0.333) = -1.100$ so that the Wald test is

$$\left(\frac{0.375 - (-1.100)}{0.392}\right)^2 = 14.16.$$

This is very close to minus twice the log likelihood ratio and the approximate p-value is 0.00017.

11.7 The null value for the probability parameter is $\pi_\varnothing = 0.25$ so that

$$\begin{aligned} U &= 16 - 27 \times 0.25 = 9.25, \\ V &= 27 \times 0.25 \times 0.75 = 5.0625. \end{aligned}$$

The score test is

$$\frac{(9.25)^2}{5.0625} = 16.90$$

and p-value is less than 0.001.

11.8 The log likelihood ratio chi-squared value is

$$-2 \times \left[-4\log\left(\frac{4}{0.25}\right) + (4 - 0.25)\right] = 14.681.$$

The score test is

$$\frac{(4-0.25)^2}{0.25} = 56.250.$$

Both give $p < 0.001$.