# 5 Statistical power

*Conor V. Dolan and Stèphanie M. van den Berg*

A primary concern in any study designed to detect and estimate genetic and environmental contributions to the variance of (complex) phenotypes is the probability of detecting a hypothesized effect, given that it is present. This probability is usually referred to as *statistical power* (e.g. Cohen, 1992; Kraemer, 1985). If this probability is low, one should be reticent to embark on the study. After all, why bother, if the probability of detecting the effect of interest is small? Of the various disciplines that are concerned with the etiology of individual differences in complex phenotypes, the greatest effort to evaluate and improve this probability has arguably been made in the field of statistical genetics. Almost 30 years ago this issue was addressed in the classical twin design (Martin *et al.*, 1978). Since then the subject of power has remained in the limelight (e.g. Heath and Eaves, 1985; Heath et al. 1985; Hewitt and Heath, 1988; Nance and Neale, 1989; Neale *et al.*, 1994; Posthuma and Boomsma, 2000; Rietveld *et al.*, 2003; Schmitz *et al.*, 1998). Recently, the shift in focus from the aggregated polygenic effects to the putatively small effects of individual quantitative trait loci (QTLs) has intensified the interest in power and methods to increase power. The number of recent studies devoted to power in QTL linkage and association analyses is staggering.

The aim of the present chapter is to explain statistical power and closely related concepts (type-I and type-II errors, and their probabilities) within the classical framework for statistical inference based on maximum likelihood (Azzelini, 1996; Miller and Miller, 2004). In this approach to inference, we posit a probability (distribution or density) function for the data, and cast our hypotheses in terms of the parameter values of the function. For example, we posit a normal distribution for height in the Dutch population of males and hypothesize that the average height equals 1.80 m. Statistical inference serves to actually answer the question whether a hypothesized situation (average height equals 1.80 m) or an effect

(e.g. average height is greater in Holland than in Germany) is present or not. Empirical information as provided by a sample is utilized to draw conclusions about a characteristic of the population. Statistical inference necessarily involves a degree of uncertainty. Quite simply, how can we be sure that an effect that is estimated in a finite sample is not a chance result? The short answer is that we cannot be sure. However, as explained below, we can express our uncertainty in terms of the probabilities of drawing the wrong or right conclusion. In addition, we can identify the factors that affect these probabilities, and exploit these to minimize the probability of an incorrect decision.

Below we first define the correct and incorrect decisions, and their probabilities (Section 5.1). We then consider briefly the procedure of maximum likelihood (ML) estimation (Section 5.2) and inference based on the generalized likelihood-ratio test. With this in place we return to the actual assessment of the probabilities of drawing a correct or incorrect decision. We discuss an important aspect of ML-based tests of variance components, which are important in linkage analysis. We summarize this material (Section 5.3) and present one illustrative example (Section 5.4). Although we concentrate on ML estimation and testing, we do discuss estimation based on least squares, as this method is important in regression models for QTL mapping (Section 5.5). The evaluation of probabilities of drawing a (in)correct conclusion, that is power calculations, depends critically on the feasibility of calculating so-called sufficient statistics; in certain situations one cannot avail oneself of sufficient statistics. We discuss these circumstances in Section 5.6. Finally, in Section 5.7, we discuss some limitations of the present chapter. We use the freely available R program to carry out the actual calculations (http://cran.r-project.org/; e.g. see Dalgaard, 2002; Venables *et al.*, 2001), and provide within this text all relevant R language scripts.

## 5.1 Probabilities of (In)correct Decisions

To provide a concrete context, consider a regression model (for more information on regression see Chapter 4), in which a continuous phenotypic variable X, measured in sib pairs, is regressed on an environmental variable (E), a background genetic variable (G), and a variable Q, which represents the effects of a quantitative trait locus (QTL):

$$X_{ij} = \beta_0 + \beta_q Q_{ij} + \beta_g G_{ij} + \beta_e E_{ij} + \varepsilon$$

(e.g. Ferreira, 2004; Fulker and Cherny, 1996; Posthuma *et al.*, 2003; Sham, 1998). The symbols $\beta_0$, $\beta_q$, $\beta_g$, and $\beta_e$ represent regression

coefficients ($\beta_0$, or the intercept, is the coefficient in the regression of X on a unit vector). The subscripts i and j refer to sib pair and sibling, respectively. Suppose that effects of G and E are not in doubt (i.e. $\beta_g \uparrow 0$ and $\beta_e \uparrow 0$), as is the case with many psychological variables (Turkheimer and Gottesman, 1991). Let us assume that we want to determine whether $\beta_q$ differs from zero in the population, that is whether the QTL contributes to individual differences in the phenotype X. We distinguish two hypothetical situations, namely, $\beta_q$ is or is not equal to zero in the population ($\beta_q = 0$ and $\beta_q \uparrow 0$, respectively). In addition, we may, on the basis of an estimate of $\beta_q$ calculated in the sample, conclude that $\beta_q$ does or does not deviate significantly from zero. Suppose we had posited the hypothesis that the effect is absent. We denote this hypothesis $H_0$, as traditionally it is called a null-hypothesis, that is $H_0$: $\beta_q = 0$. The alternative hypothesis, denoted $H_1$, may in principle involve any other specific value of $\beta_q$. For instance, we may state that $\beta_q$ equals a value associated with exactly five percent of the variance. This is called a *simple hypothesis*. Generally we are unable to be so precise, because we do not know the exact value under $H_1$. We therefore formulate the $H_1$ simply as $H_1$: $\beta_q \uparrow 0$, that is the parameter is not zero. This is referred to as a *composite hypothesis*. The $H_1$: $\beta_q 0$ is called two-sided, because it implies that the parameter may be greater or smaller than zero. A one-sided $H_1$ includes a direction of the effect, for example $H_1$: $\beta_q > 0$ or $H_1$: $\beta_q < 0$. For now, we simply adopt the composite $H_1$: $\beta_q \uparrow 0$. On the basis of information in the sample (i.e. the observed data), we may draw a correct or an incorrect conclusion, depending on the true value of $\beta_q$ in the population. *Table 5.1* contains the possible outcomes and their probabilities pertaining to the $H_0$ and the composite two-sided $H_1$.

We can now distinguish two types of errors. A *type-I error* amounts to rejecting $H_0$ incorrectly: $H_0$ is rejected, even though it is true (in truth $\beta_q = 0$). This probability of this error is denoted $\alpha$.

**Table 5.1 Probabilities of correct and incorrect decisions**

| | | Statistical decision | |
|---|---|---|---|
| | | Reject $H_0$ | Accept $H_0$ |
| True state of the world | $H_0$ is true $\gamma_\theta = 0$ | Incorrect decision: type – I error probability: $\alpha$ | Correct decision probability: $1-\alpha$ |
| | $H_0$ is false $\gamma_\theta \uparrow 0$ | Correct decision probability $1-\beta$(power) | Incorrect decision: type – II error probability: $\beta$ |

The probability of correctly accepting $H_0$ is then $1-\alpha$. A *type-II error* amounts to accepting $H_0$, even though it is not true (i.e. in truth $\beta_q \uparrow 0$). The probability of this type-II error is denoted $\beta$. The probability of correctly rejecting $H_0$ (i.e. in truth $\beta_q \uparrow 0$), that is $1-\beta$, is commonly referred to as the statistical power. We should note that the designation 'null-hypothesis' does not mean that the hypothesized value of the parameter should equal zero. We could just as well have posited the $H_0$ that $\beta_q$ equals a specific value x (e.g. associated with 10% of the phenotypic variance). $H_0$ usually represents the more parsimonious hypothesis, while the composite $H_1$ represents the more liberal hypothesis (often the one that the researcher wants to be true, e.g. a particular allele is related to disease). Compared to $H_1$, $H_0$ thus comprises fewer free (i.e. to be estimated) parameters.

Most of the time, researchers are interested in proving $H_1$ to be true, so that one wishes to maximize statistical power, given the choice of $\alpha$, increasing the probability that an effect is detected. Statistical power is a characteristic of the statistical test and the study design that we use to decide whether to reject a given hypothesis; a good test and a good study design are characterized by a large probability $1-\beta$. Thus, to assess the probabilities of the decisions in *Table 5.1*, we require an estimate of the parameter of interest (e.g. estimate of $\beta_q$), and a test statistic, T, upon which we base our decision to reject or accept $H_0$. In the next section, we concentrate mainly on maximum-likelihood (ML) estimation, which provides us with both optimal estimates of unknown parameters, and test statistics that follow known distributions under $H_0$ and $H_1$. With these in place, we can evaluate $\alpha$ and $1-\beta$, and examine the influence of sample size and effect size $1-\beta$.

## 5.2 Maximum-likelihood Estimation

In maximum-likelihood (ML) estimation, we assume that the observed data are generated by a process that is characterized by a density function (continuous data) or distribution function (discrete data; Miller and Miller, 2004). The standard example is the process of flipping a coin. Let $X_j$ be the outcome of the jth flip of a coin (heads coded $X_j = 0$, tails coded $X_j = 1$). This process generates outcomes that follow a Bernoulli distribution, which we denote $\text{Bern}(X_j;\theta)$, where $\theta$ is the parameter of the distribution. $\text{Bern}(X_j;\theta) = \{\theta^{X_j}(1-\theta)^{(1-X_j)}\}$ is the probability density function associated with this process when observing the *order* in which the heads and tails occur (eg. HTTH). The probability of observing tails in a single flip is determined by the parameter $\theta$ (heads by $1-\theta$). Assuming the outcomes of repeated coin flipping are independent, the process of flipping a coin N times and observing the total number of heads

and tails (no matter in what *order*) is characterized by the binomial distribution, which we denote $Bin(X;N,\theta)=[N!/\{(N-X)!X!\}]\theta^X(1-\theta)^{(N-X)}$ (Evans *et al.*, 2000). This function assigns probabilities to the outcome of observing X tails in N flips of the coin, so with this distribution function, we can assign probabilities to outcomes. For example, if we know that $\theta ?= 0.5$, then the probability of observing three tails in 10 flips equals about 0.117.

The binomial distribution function is but one of many possible functions for discrete data. Others include the Poisson and the multinomial distribution function (Evans *et al.*, 2000; Ewens and Grant, 2001; Miller and Miller, 2004). Generally, we denote a distribution or density function with parameter vector $\boldsymbol{\theta}$, suitable for data $\mathbf{X}$, as $f_X(\mathbf{X}; \boldsymbol{\theta})$, where $\mathbf{X}$ is the $N \times p$ matrix of data (N cases and p variables). By 'suitable' we mean suitable given the process that generated the data, as above in the binomial example, or consistent with the observed distribution (e.g. bell-shaped, continuous). Given $f_X(\mathbf{X}; \boldsymbol{\theta})$ and given values for $\boldsymbol{\theta}$, we can calculate the probability of a given outcome, that is an observed dataset.

The problem of statistics is that we know the data, but we do not know the values of $\boldsymbol{\theta}$**.** In statistical analyses our hypotheses concern unknown elements of the parameter vector $\boldsymbol{\theta}$. For instance, the question whether the coin is fair, is equivalent to the question whether $\theta$ equals 0.5. ML estimation of unknown elements in the parameter vector $\boldsymbol{\theta}$ involves finding the values of $\boldsymbol{\theta}$ that maximize the probability of obtaining the data that we have, and thus maximize $f_X(\mathbf{X}; \boldsymbol{\theta})$. Since actually the data are given, another equivalent way of expressing this is that we want to maximize the *likelihood* of the parameter values, given the data, which is denoted as $L(\theta;N,X)$ (Azzelini, 1996; for a good tutorial, see Myung, 2003; for an accessible technical account, see Sorensen and Gianola, 2002). To illustrate this, suppose we observed X = 3 tails in N = 10 flips. To obtain the ML estimate of $\theta$, we regard the data (X = 3, N = 10) as fixed, and seek the value of $\theta$ that maximizes the likelihood function $L(\theta;N,X) = bin(X;N,\theta)$. More often $-2*$log-likelihood is minimized, which we denote $LogL(\theta;N,X) = -2*\log\{bin(X;N,\theta)\}$, as this is computationally easier. The value of $\theta$ that *minimizes* this function is the maximum likelihood estimate of $\theta$. We demonstrate this in a small R script (see *Panel 5.1*) by using a simple grid search, that is, we can plot the function for various values of $\theta$.

The plot, shown in *Figure 5.1*, reveals that the ML estimate of $\theta$ equals 0.3, and that the log-likelihood function at this value equals about 2.642.

As mentioned, the ML estimate of $\theta$ is the most 'likely' value of $\theta$ given the observed data $\mathbf{X}$, that is, the value that makes the data

```
X=3                                # three tails
N=10                               # total number of trails
theta=seq(.1, 9.9, by=.05)         # a vector of values of theta (.1 to .9)
logl=c()                           # a vector for the loglikelihood function
num=length(theta)                  # number of elements in theta
for (i in 1:num)
{
   logl[i]=-2*log(dbinom(x, n, theta[i]))       # estimates the logL
}
plot(theta, logl, type='1',
                xlab="Theta", ylab="-2loglikelihood"
                font.lab=2, cex.lab=1.3,las=1, lwd=2,
                cex.axis=1.2)   # plot theta by logL
minl=min(logl)                     # lowest logl
est=theta[logl==minl]              # theta at which lowest logl was observed
abline(h=minl, v=est)              # add minl and est lines to plot
grid(col="darkgray")               # add gridlines to plot
print(c(minl, est))                # print the ML estimate and logl value
```

**Panel 5.1** R script used to estimate loglikelihood of observing 3 tails in 10 flips of a coin, as a function of various trial values of $\theta$. This script generates the plot shown in Figure 2.

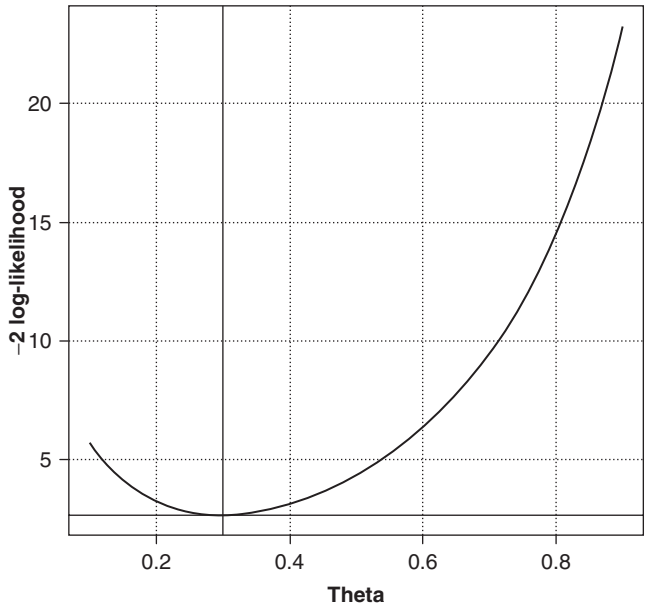


**Figure 5.1** The log-likelihood function given various trial values of $\theta$. The minimum is at $\theta = 0.3$, and the associated value of the function equals 2.642.

most probable. Given $\theta = 0.3$, and applying $Bin(X;N,\theta)$ as defined above, the probability of the data ($X = 3$, $N = 10$) is 0.266. Because there is no value of $\theta$ that results in a greater probability of observing $X = 3$ (e.g. $\theta = 0.4$ results in 0.215), the value $\theta = 0.3$ is characterized by the greatest likelihood, given $X = 3$.

This procedure is general: it can be applied to any dataset, given an appropriate choice of distribution or density function. So generally, to obtain ML estimates, we can minimize:

$$LogL(\boldsymbol{\theta};\mathbf{X}) = -2*\log\{f_X(\mathbf{X};\boldsymbol{\theta})\},$$

given the complete data matrix $\mathbf{X}$, or assuming independent cases (e.g. independendent sib pairs):

$$LogL(\boldsymbol{\theta};\mathbf{X}) = -2*\Sigma\log\{f_X(\mathbf{X}_i;\boldsymbol{\theta})\}$$

where $\mathbf{X}_i$ is the i-th case (i.e. i-th row in $\mathbf{X}$), and summation is over cases. For example, in the regression model mentioned in the Introduction, our use of ML estimation is based on the assumption that the phenotype observed in the sib pairs follows a bivariate normal distribution. We illustrate this below.

While a grid search can be convenient when the parameter space is limited, such as a proportion that necessarily lies between 0 and 1, it is very inconvenient when the parameter space ranges from minus infinity to plus infinity: how many points should be evaluated and between what values? In such cases, the minimal value for the log-likelihood can easily be missed. Generally an optimization algorithm is used to find the minimum of the log-likelihood function, rather than a grid search (Gill *et al.*, 1981; Neale and Cardon, 1992). Such algorithms minimize the log-likelihood function by finding the values of the parameters that result in zero first order derivatives ($\Delta LogL(\boldsymbol{\theta};\mathbf{X})/\Delta\boldsymbol{\theta} = 0$). In simple cases, such derivatives can be solved by hand. In the case of binomial distribution, the ML estimate is X/N (0.3 in the example above). As $(\Delta LogL(\theta;N,X)/\Delta\theta$ equals $(X - \theta N)/(\theta^2 - \theta)$, substituting X/N for $\theta$, results in $\Delta LogL(\theta;N,X)/\Delta\theta = 0$. Stated more generally, ML estimates of the unknown elements of $\boldsymbol{\theta}$ are those that solve the equation $\Delta LogL(\boldsymbol{\theta};\mathbf{X})/\Delta\boldsymbol{\theta} = 0$.

ML estimation is used extensively, because ML estimates are characterized by the following desirable properties. ML estimates are *asymptotically unbiased*, that means as sample size increases the expected value of the estimates tends towards the true value $E[\hat{\theta}] = \boldsymbol{\theta}$. Depending on the parameter, the estimate may be simply unbiased, that is independent of sample size. The estimate X/N of $\theta$ in the binomial distribution is an example of an unbiased ML estimate. The ML estimate of the variance of the normal distribution

is an example of an asymptotically unbiased estimate. Another desirable property of ML estimates is *efficiency*, that is the sample distribution of the estimate has the smallest possible variance (i.e. minimum variance). In repeated sampling we expect an estimate to vary from sample to sample, as the estimate is a function of the sample. The efficiency of the ML estimate means that this variance, var[$\hat{\theta}$], is as small as possible (i.e. it hits the so-called Cramèr-Rao bound; see Azzelini, 1996; Miller and Miller, 2004). The combination of unbiasedness and minimum variance renders the ML estimates *consistent*. This means that the ML parameter estimate converges on the true value of the parameter estimate as the sample size increases. So as N increases, E[$\hat{\theta}$] approaches $\boldsymbol{\theta}$ and var[$\hat{\theta}$] approaches 0. Finally, under fairly mild conditions, ML estimates are *asymptotically normally distributed*.

We illustrate these properties in *Figure 5.2*. This figure displays the results of estimating the binomial parameter $\theta$ 1000 times with sample size from 10 to 100 in steps of 10. The x-axis represents the sample size, the y-axis the value of $\theta$. The plots display the average

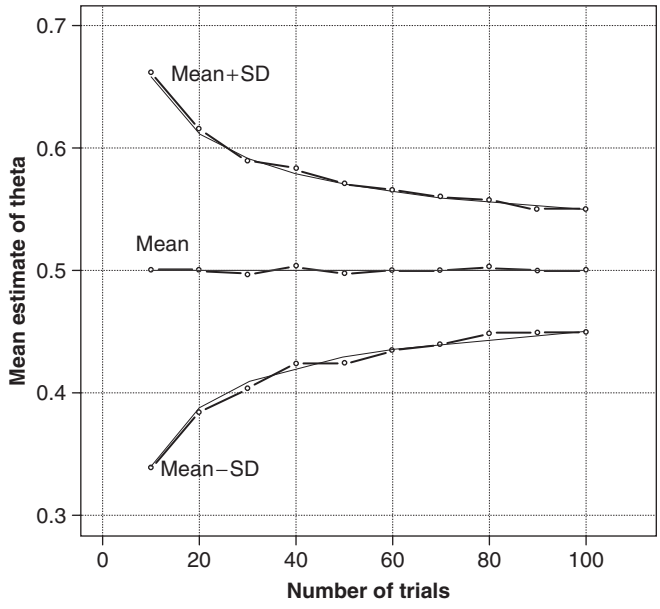


**Figure 5.2** Repeated sampling experiment. Plot of the mean estimate of $\theta$ over 1000 replications given sample sizes 10 to 100 in steps of 10 and the (middle broken line). The mean estimate closely resembles the true value (0.5; middle solid line). The lower and upper broken lines represent the mean estimate ± the observed standard deviations of the estimates. These tend towards those based on the Cramèr-Rao bound (depicted in solid lines) as N increases.

estimate based on 1000 replications, and the average estimate ± the standard deviation of the estimate. Solid lines represent the true value (0.5) and the minimal possible variance of the estimate given the sample size. It is apparent that the average estimate over 1000 replications closely resembles the true value of 0.5 (unbiasedness). In addition, we see that the standard deviations closely resemble the theoretical lower bound (the Cramèr-Rao bound). Finally as N increases, the variance of the estimate decreases.

So ML estimation yields optimal estimates of the unknown parameters in the parameter vector $\boldsymbol{\theta}$. But we still require a test statistic, T, which we can use to determine whether the parameter estimate(s) (e.g. $\hat{\theta} = 0.3$) deviate significantly from the value(s) under $H_0$ (e.g. $\theta = 0.5$).

### *Test statistics*

There are three, asymptotically equivalent, test statistics in statistical inference based on the likelihood: the generalized likelihood-ratio (or log-likelihood difference) test, the Wald test, and the score test (Azzelini, 1996; Greene, 1993; Sorensen and Gianola, 2002). All three are used in QTL analyses. The likelihood-ratio test is often applied in testing variance components (Almasy and Blangero, 1998; Eaves *et al.*, 1996; Fulker), while the Wald test and the score test are often used in sib pair regression modeling (Haseman and Elston, 1972; Putter *et al.*, 2002; Visscher and Hopper, 2001). Here we focus mainly on the likelihood-ratio test, but we do discuss the Wald test below in connection with a regression model for sib pair data. We present the general formulation of the log-likelihood difference test, and explain a slight complication in the application of the test to variance components (Carey, 2004, 2006; Dominicus *et al.*, 2006).

The log-likelihood difference test is constructed as follows. Let $\text{LogL}(\boldsymbol{\theta}_0;\mathbf{X})$ be the minimum value of the log-likelihood function under the more parsimonious $H_0$, and let $\text{LogL}(\hat{\theta}_A;\mathbf{X})$ be the minimum value of the log-likelihood function under some composite $H_1$. We assume that $H_0$ is nested under $H_1$, in the sense that the parameter vector $\boldsymbol{\theta}_0$ is a constrained version of the parameter vector $\hat{\theta}_A$(Bollen, 1989; Ewens and Grant, 2001). The test statistic is calculated as follows: $T = 2*(\text{LogL}(\boldsymbol{\theta}_0;\mathbf{X})-\text{LogL}(\hat{\theta}_A;\mathbf{X}))$. If $H_0$ is true, this test statistic asymptotically follows a $\chi^2$ distribution with degrees of freedom equal to the difference in the number of parameters estimated in the $H_0$ model and the $H_1$ model, $T \sim \chi^2(df)$. For a clear derivation of $\chi^2$ test under $H_0$ in the df = 1 case, see Ewens and Grant (2001, p. 254), and for the multiparameter case, see Sorensen and Gianola (2002, p. 171). Returning to our coin example, based on the

binomial, we have $\text{LogL}(\hat{\theta};N,X) = 2.642$. The minimum value given $\theta_0 = 0.5$ is $\text{LogL}(\theta_0;N,X) = 4.288$, so $\chi^2(1) = 1.646$.

Nesting constraints may include equality constraints or fixed parameter constraints. For instance suppose we have the parameter vector $\boldsymbol{\theta}_A = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5]$ under $H_1$. A nested model may involve the vector $\boldsymbol{\theta}_0 = [\theta_1 = \theta_2, \theta_3 = 0, \theta_4, \theta_5]$. The difference in the number of parameters, that is the df of the likelihood-ratio test, is 5–3 = 2. In the classical twin design, we can estimate a variance component due to shared environmental effects C ($\sigma_C^2$), unshared environmental effects E ($\sigma_E^2$), and additive polygenic effects A ($\sigma_A^2$). The AE model ($\boldsymbol{\theta}_0 = [\sigma_A^2, \sigma_C^2 = 0, \sigma_E^2]$) and the CE model ($\boldsymbol{\theta}_0 = [\sigma_A^2 = 0, \sigma_C^2, \sigma_E^2]$) are both nested under the ACE model ($\boldsymbol{\theta}_A = [\sigma_A^2, \sigma_C^2, \sigma_E^2]$). However, a direct comparison by means of a likelihood ratio test of the AC model and AE model is not possible, as one of the vectors of parameter $\boldsymbol{\theta}$ associated with one model is not a subset of the vector of the other model, that is, the models are not nested.

We thus have at our disposal a test statistic T that asymptotically follows a known distribution under $H_0$, the $\chi^2$ distribution. We now consider the distribution in the situation that $H_0$ is false. Strictly speaking, if $H_0$ is true, the $\chi^2$(df) statistic T follows the *central* $\chi^2$(df). The shape of this distribution depends solely on the number of degrees of freedom (df). In the event that $H_1$ is true, the test statistic T follows the *non-central* $\chi^2$(df) distribution (Saris and Satorra, 1993; Satorra and Saris, 1985). The shape of this distribution depends on the degrees of freedom and on the so-called noncentrality parameter (NCP), $\lambda$. The non-central $\chi^2$ distribution is denoted $\chi^2(df,\lambda)$. (Actually, if $H_0$ is true, the NCP equals zero, so we write $\chi^2(df,\lambda = 0)$ for central $\chi^2$ distribution). To put this distribution to use, we have to estimate the NCP $\lambda$.

To obtain a numerical estimate of NCP, $\lambda$, in the case of the $\chi^2(df,\lambda)$ distribution, we choose a sample size N, and assign specific parameter value(s) to express $H_0$ and $H_1$. That is to say, both $H_0$ and $H_1$ have to be fully specified. For instance, in terms of the regression model $X_{ij} = \beta_0 + \beta_q Q_{ij} + \beta_g G_{ij} + \beta_e E_{ij} + \varepsilon$, we could choose value of $\beta_q$ under $H_1$, while under $H_0$, we have $\beta_q = 0$. The numerical difference represents the *effect size*. It is useful to express the effect size on a scale that is readily interpretable, such as percentage of variance explained, and given this, derive the numerical value of the parameter $\beta_q$. Given this parameter value and, of course, the related values for all other parameters in the model, we calculate summary statistics under $H_1$, that is we calculate the exact population values of the means and covariance matrices associated with the choice of parameters. Finally, we fit the $H_0$ and $H_1$ models and

calculate the difference in the minima of the log-likelihood function. The NCP, λ, approximately equals this difference. The value of λ depends on the chosen effect size (5% *vs* 0% variance explained by the QTL) and the sample size N. But all other parameters in the model may also affect the value of λ. We therefore must provide a completely specified model, which includes an effect size for the parameter of immediate interest, as well as values for all other parameters. Below, we will illustrate this procedure using a concrete example on the ability to detect a violation of Hardy-Weinberg equilibrium.

*Probability of type-I error: $\alpha$*

If $H_0$ is true, the test statistic T follows a central $\chi^2$(df) distribution with df equal to the difference in number of parameters under $H_0$ and $H_1$. When fitting a model, we might observe an extreme value of T, that is one greater than some predetermined critical value, and we reject $H_0$. The notion of extremeness can be related directly to the distribution of T under $H_0$, and this is where the probability α comes in. The choice of α determines the associated critical value C. For example, suppose we set $\alpha = 0.05$. Under $H_0$, and given $T \sim \chi^2(1)$, the associated critical value C equals about 3.8414. So, if T is greater than 3.8414, we reject the $H_0$. The probability of incorrectly rejecting $H_0$, that is, of committing a type-I error, is thus $\alpha = 0.05$. The incorrect rejection of $H_0$ may happen because T is a random variable, which purely by chance under $H_0$ may assume values greater than C. We can control this chance by choosing α. Critical values, and *p*-values associated with observed values of T (say, for example, 1.645) may be obtained in R using the code shown in *Table 5.2* (Aim 1).

In the example shown in *Table 5.2*, we considered for illustration that our test statistic T = 1.645. The probability of observing T = 1.645 or larger when $\alpha = 0.05$ is 0.20. This can be calculated in R using the R code in *Table 5.2* (Aim 2). Thus, had we chosen an α of 0.05 (C about 3.84), we would not reject $H_0$.

*Probability of type-II error: $\beta$*

Suppose we have chosen α, and calculated the associated critical value C under $H_0$, that is $\text{prob}[\chi^2(df) > C] = \alpha$. The probability β can be obtained by calculating the probability of observing a value of the test statistic T smaller than C, given $\chi^2(df,\lambda)$, that is $\text{prob}[\chi^2(df,\lambda) < C] = \beta$. The power of the test is then 1–β. This is illustrated in *Figure 5.3*, given the arbitrary values df = 3, $\alpha = 0.05$, and $\lambda = 4.5$. The R script in *Table 5.2* (Aim 3) can be used to calculate the power.

**Table 5.2 R code for calculating probabilities and critical values**

| Aim | R code (including illustrative values) |
|---|---|
| 1. Obtain critical value C given $\alpha^a$ and df | alpha = .05; df = 1;<br>C = qchisq(alpha, df, ncp = 0, lower.tail = F); |
| 2. Obtain probability of T or larger in the central $\chi^2$ distribution | T = 1.645; df = 1;<br>prob = pchisq(T, df, ncp = 0, lower.tail = F); |
| 3. Obtain the probability T > C in the non-central $\chi^2$ distribution given $\alpha^a$, df, and $\lambda$ (i.e. calculate the power, $1-\beta$) | lambda = 4.5; alpha = .01; df = 3;<br>C = qchisq(alpha, df, lower.tail = F);<br>power = pchisq(C, df = df, ncp = lambda, lower.tail = F);<br>beta = 1 – power; |
| 4. Obtain the probability T > C in the non-central $\chi^2$ distribution given $\alpha^*$, df, and a change in sample size from N1 to N2 | N1 = 1000; lambda1 = 1.551; alpha = .05; df = 1;<br>N2 = 4000; lambda2 = N2*(lambda1/N1);<br>C = qchisq(alpha, df, lower.tail = F);<br>power1 = pchisq(C, df = df, ncp = lambda1,<br>lower.tail = F);<br>power2 = pchisq(C, df = df, ncp = lambda2,<br>lower.tail = F) |

[a]As explained below, if the test concerns a variance component, which is subject to a boundary constraint, the alpha should be doubled.

*Figure 5.3* illustrates the fact that $\alpha$ and $\beta$ are not independent. All things being equal, a decrease in $\alpha$ (type-I error probability) results in an increase in $\beta$ (type-II error probability). For instance, if the $\alpha$ is chosen to be $\alpha = 0.025$ instead of $\alpha = 0.05$, we set C = 9.35 instead of C = 7.81 (*Figure 5.3A* vs *C*). In *Figure 5.3B*, the light gray area ($\beta$) goes from 0.60 to 0.71, and so the power, $1-\beta$, decreases, from about 0.40 to 0.29. So a smaller $\alpha$ results in a larger C for which prob[$\chi^2$(df)>C] = $\alpha$ holds, and a larger C results in a larger probability $\beta$, prob[$\chi^2$(df,$\lambda$)<C] = $\beta$, and thus smaller power, 1-$\beta$.

*Testing variance components*

The validity of the log-likelihood difference test is based on certain conditions that relate to the admissible parameter space ('regularity conditions'; Azzelini, 1996). One highly relevant condition is that the parameters of interest in $H_0$ may *not* be on the boundary of the parameter space. A well-known instance in which this condition is violated, is in tests of variance components that under $H_0$ are placed on the boundary of the admissible parameter space, namely zero (variance cannot assume negative values). Variance components are commonly tested in genetic modeling, so we have to consider the effects of this violation on the $\chi^2$ test. This issue

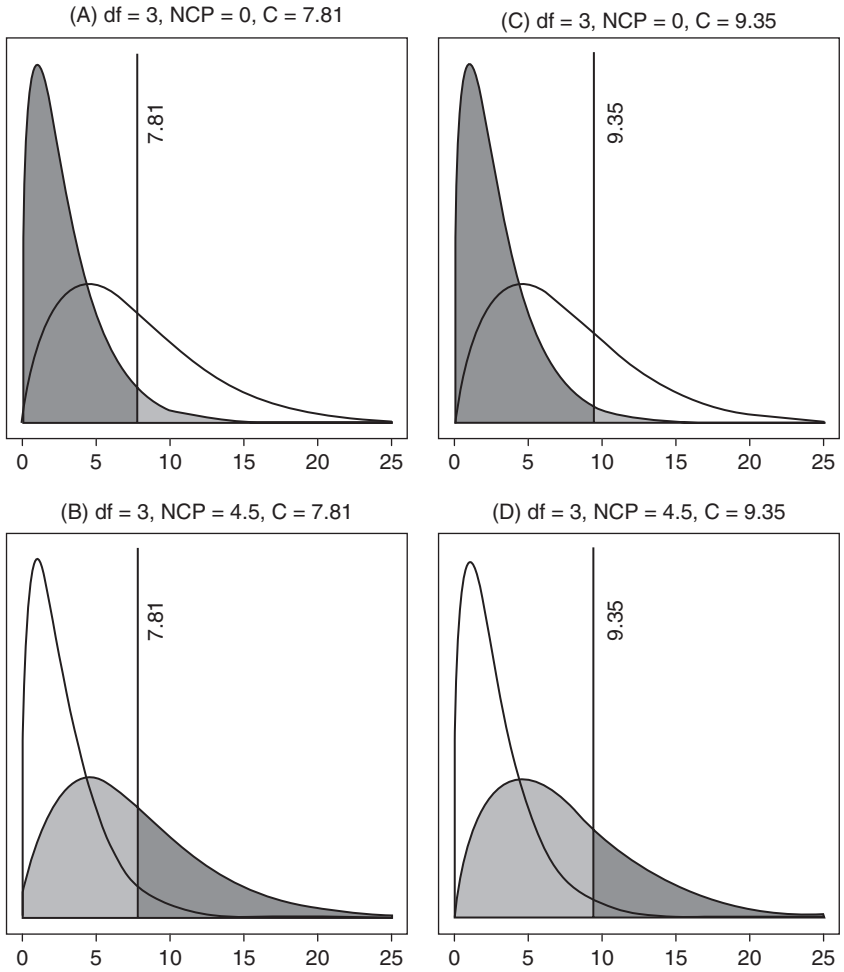


**Figure 5.3** (A) $\chi^2$(df = 3, $\lambda$ = 0). The critical value C = 7.81 is associated with the $\alpha$ of 0.05 (the light gray region). Thus if H–0 is correct, the probability of rejecting it is 0.05. (B) $\chi^2$(df =3, $\lambda$ = 4.5). The light gray region represents $\beta$, the dark gray region represents 1–$\beta$, that is the power. In this case, 1–$\beta$ equals 0.40. Thus if H–A is correct, the probability of rejecting H–0 in favor of H–A is 0.40. (C) The critical value C = 9.35 is associated with $\alpha$ = 0.025 (power = 0.29). The relationship between $\alpha$ and $\beta$ is revealed in this figure: the smaller the light gray area in (A) and (C) ($\alpha$), the greater the light grey area in (B) and (D) ($\beta$) and so the smaller 1–$\beta$. The values df = 3, a = 0.05 (0.025), and $\lambda$ = 4.5 were chosen for illustrative purposes.

was discussed by Hopper and Matthews (1982) and, more recently, by Dominicus *et al.* (2006) and Carey (2004, 2006) in genetic covariance structure modeling, and by Williams and Blangero (1999) in connection with variance component modeling of QTLs.

Suppose we are interested in establishing whether a variance component $\sigma_q^2$, due, say, to a putative QTL, is greater than zero.

We can fit the model without ($H_0$: $\sigma_q^2$=0) and with the effect ($H_1$: $\sigma_q^2$>0), and calculate the test statistic T on the basis of the values of the two log-likelihood functions. As mentioned above, an ML parameter estimate of a given parameter is expected to be asymptotically normally distributed, with the mean value equal to the true value in the population. If $H_0$ is correct, the true value of the parameter is zero ($\sigma_q^2 = 0$), and we expect the parameter, in repeated sampling, to vary about this value. In fitting $H_1$, the parameter is freely estimated, but subject to bound. The boundary condition can be specified explicitly, by the imposition of an actual boundary constraint, or implicitly, by estimating $\sigma_q$ instead of $\sigma_q^2$. So, if the true value is zero, we expect the parameter, in a repeated sample scenario, to hit the lower bound of zero in 50% of the analyses. In each case that the parameter hits the lower bound, the value of the log-likelihood under $H_1$ equals that under $H_0$, and the log-likelihood difference, the $\chi^2$, will equal zero. In the other 50% of the cases, the parameter will assume a value greater than zero, so that the log-likelihood difference will be greater than zero, that is, follow the expected $\chi^2(1)$ distribution. The implication of this is that the distribution of the test statistic T will follow a 50%:50% mixture of a central $\chi^2(1)$ and a $\chi^2(0)$ distribution (where $\chi^2(0)$ is a point mass or spike at zero), rather than the usual central $\chi^2(1)$ distribution. In determining the critical value given the choice of $\alpha$, we have to refer to this mixture distribution, rather than the central $\chi^2(1)$. In this simple case, we can obtain the correct value by doubling the value of $\alpha$. For instance, given $\alpha = 0.05$, the critical value C is 2.7055, rather than the usual 3.8414. We refer the reader to Dominicus *et al.* (2006) for a detailed discussion of this in the case of a single parameter. Carey (2004, 2006) discusses this issue in the multiparameter case, namely the estimation of a covariance matrix by means of the Cholesky decomposition (rather than a single variance component). A multiparameter situation also arises in testing the additive genetic and dominance variance components of a QTL. In this case the distribution of the test statistic T follows a mixture of $\chi^2(0)$, $\chi^2(1)$, and $\chi^2(2)$ under $H_0$. The mixing proportions depend on the actual parameter values. Both analytical methods (Self and Liang, 1987; Stram and Lee, 1994) and numerical methods are available to obtain these (Dardanoni and Forcina, 1998). Box and Tiao (1973) discuss a solution to the problem of testing variance components for Bayesian statistical tradition.

Here we limit our attention to the situation involving a single variance component using the ML approach. As the calculation of power depends on the critical value C, it is important to take into account the distribution of the likelihood-ratio test under H = 0.

For example, suppose that $\alpha = 0.05$, $\lambda = 4.15$, df = 1. Using the incorrect critical value of 3.8414, we obtain $1-\beta$ equal to 0.5308, while the correct value of $1-\beta$, based on C = 2.705, is 0.653.

## 5.3 Summary

To summarize the concepts we have discussed thus far, in likelihood-based inference we distinguish four variables that affect the probability of drawing a correct or incorrect conclusion in comparing $H_0$ and $H_A$: the type-I probability $\alpha$, sample size N, effect size, and power ($1-\beta$, or $\beta$, the type-II error probability). If we fix any three of these, we can calculate the fourth. Usually, a fixed $\alpha$ is chosen, and power is calculated given various choices of N and the effect size that together determine the value of the noncentrality parameter $\lambda$. *Table 5.3* extends *Table 5.1* with the relevant test statistic T and its distribution based on the likelihood ratio in the case that the parameter of interest is not on the boundary under $H_0$. *Table 5.4* is the same table for the situation in which a single variance component is tested (parameter on boundary, i.e. fixed to zero under $H_0$).

## 5.4 Example

We now present an illustrative example of likelihood-ratio testing, which concerns the ability to detect a violation of Hardy-Weinberg equilibrium given disruptive selection. A random process that generates K outcomes, with fixed probabilities $\theta_1,\ldots, \theta_K$, is characterized by the multinomial distribution function (e.g. six possible

**Table 5.3 Probabilities of correct and incorrect decisions concerning the parameter $\theta$, given critical value C, and associated test statistic T**

| | | Statistical decision | |
|---|---|---|---|
| | | Reject $H_0$ | Accept $H_0$ |
| True state of the world | $H_0$ is true $\gamma_\theta = 0$ | Type – I error<br>$T \sim \chi^2(df, \lambda = 0)$, $\alpha = prob(T > C)$<br>(R script 1 from *Table 4.2*) | Correct decision<br>$T \sim \chi^2(df, \lambda = 0)$, $1-\alpha = prob(T < C)$<br>(R script 1 from *Table 4.2*) |
| | $H_0$ is false $\gamma_\theta \uparrow 0$ | Correct decision<br>$T \sim \chi^2(df, \lambda > 0)$, $1-\beta = prob(T > C)$<br>(R script 2 from *Table 4.2*) | Type – II error<br>$T \sim \chi^2(df, \lambda > 0)$, $\beta = prob(T < C)$<br>(R script 2 from *Table 4.2*) |

The test statistic T is the log-likelihood difference. The parameter $\theta$ is not subject to a boundary constraint under H-A.

**Table 5.4 Probabilities of correct and incorrect decisions concerning the parameter θ, given critical value C, and associated test statistics**

| | | Statistical decision | |
|---|---|---|---|
| | | Reject $H_0$ | Accept $H_0$ |
| True state of the world | $H_0$ is true $\gamma_q=0$ | Type – I error | Correct decision |
| | | $T\sim.5*\chi^2(df=1, \lambda=0)+.5*\chi^2(df=0, \lambda=0)$ | $T\sim.5*\chi^2(df=1, \lambda=0)+.5*\chi^2(df=0, \lambda=0)$ |
| | | $\alpha = prob(T>C)$ | $1-\alpha = prob(T<C)$ |
| | $H_0$ is false $\gamma_q\neq0$ | Correct decision | Type – II error |
| | | $T\sim\chi^2(df, \lambda>0)$ | $T\sim\chi^2(df, \lambda>0)$ |
| | | $1-\beta = prob(T>C)$ | $\beta = prob(T<C)$ |
| | | (R script 2 from *Table 4.2*) | (R script 2 from *Table 4.2*) |

The test statistic T is the log-likelihood difference. The parameter θ is subject to the boundary constraint $\theta>0$ under H-A.

outcomes of rolling a six-sided dice). The multinomial distribution Mult(**X**;N,**θ**):

$$\frac{N!}{X_1!X_2!\ldots X_K!}\theta_1^{X_1}\theta_2^{X_2}\ldots\theta_K^{X_K},$$

where N is the total number of trials (e.g. rolls of a dice), $X_i$ is the number of times outcome I is observed (e.g. 4), and $\theta_i$ is the probability of outcome $X_i$ (presumably 1/6). Furthermore $\mathbf{X} = [X_1, X_2,\ldots,X_K]$, $\boldsymbol{\theta} = [\theta_1, \theta_2,\ldots, \theta_K]$.

The multinomial distribution function can be used to model genotype frequencies (Sham, 1998, p. 42; Sorensen and Gianola, 2002, p. 190). An important question is whether the genotype frequencies in the population are in HWE. We consider a biallelic codominant locus, with allele frequencies $p$ and $q = 1-p$, and alleles $A$ and $a$. Let $X_1$, $X_2$, and $X_3$, denote the number of times genotypes $AA$, $Aa$, and $aa$ are observed, respectively, in a random sample of size $N = \Sigma X_i$. The genotype frequencies depend on the allele probabilities $p$ and $q$, and on a parameter r as follows: $f_1 = p^2/t$, $f_2 = (2pq-r)/t$, and $f_3 = (q^2)/t$, where $t = p^2+(2pq-r)+q^2$, and $f_1$, $f_2$, and $f_3$ are the relative frequencies of $AA$, $Aa$, and $aa$, respectively. The genotype frequencies are in HWE, if r = 0. Disequilibrium is introduced by any deviation of r from 0, which in the present setup mimics disruptive selection. Under $H_1$, the log-likelihood function equals

$$\text{LogLA}(\boldsymbol{\theta}_A;\mathbf{X},N) = -2*(\log(N!/(X_1!X_2!X_3!)) + \log(\theta_{A1})^{X1} + \log(\theta_{A2})^{X2} + \log(\theta_{A3})^{X3}),$$
$$= -2*(c + X_1\log(\theta_{A1}) + X_2\log(\theta_{A2}) + X_3\log(\theta_{A3})),$$

where c is a constant function of the observed data (not of the parameters that we want to estimate). Under $H_1$ we do not constrain the genotype frequencies, so $\theta_{Ai} = X_i/N$. Under $H_1$, we estimate two free parameters, say, $\theta_{A1}$ and $\theta_{A2}$ (the third is constrained $\theta_{A3} = 1-\theta_{A1}-\theta_{A2}$).Under the more parsimonious $H_0$, we specify frequencies consistent with HWE: $p_0 = [(2X_1+X_2)/2N]$, $q_0 = 1-p_0$ and $\theta_{01} = p_0^2$, $\theta_{02} = 2p_0q_0$, $\theta_{03} = q_0^2$, that is we estimate only the allele frequency $p_0$. The log-likelihood function under the more parsimonious $H_0$ is:

$$\text{LogL0}(\boldsymbol{\theta}_0;\mathbf{X},N) = -2*(c+X_1\log(\theta_{01})+X_2\log(\theta_{02})+X_3\log(\theta_{03})),$$

so the test statistic $T = (\text{LogL0}(\boldsymbol{\theta}_0;\mathbf{X},N)-\text{LogLA}(\boldsymbol{\theta};\mathbf{X},N))$ is $\chi^2(1,\lambda)$, with $\lambda = 0$, when $r = 0$, and $\lambda>0$, if $r \uparrow 0$. We want to know whether the allele probability *p* affects the power to reject the equilibrium model. We choose $N = 500$, and calculate the power, given $\alpha = 0.05$, for $p = 0.1$ and $p = 0.5$, and the effect sizes of $r = 0.01$, $r = 0.05$, and $r = 0.10$. The results, shown in *Table 5.5*, suggest that the allele probability *p* has little effect on the power to detect the violation of HWE. Of course this conclusion does not necessarily generalize to other violations of the HWE, for example directional or stabilizing selection. However, the power to detect other types of selection can be determined quite easily by adapting the R script, and running the analyses.

The results from *Table 5.5* were obtained using the following R script (*Panel 5.2*):

## 5.5 Least-squares Estimation

So far, we have only considered ML estimation. An alternative method of estimation is least-squares estimation. This bears mentioning as it is often used in regression modeling of QTLs in sib pair data (Fulker and Cardon, 1994; Fulker *et al.*, 1995; Haseman and Elston, 1972; Visscher and Hopper, 2001). Under certain distributional assumptions, least-squares estimation produces

**Table 5.5 Power to detect the effect of disruptive selection on HWE in N=500, with $\alpha$=0.05**

| Effect size: | r = 0.01 | | r = 0.05 | | r = 0.10 | |
|---|---|---|---|---|---|---|
| Freq. p | $\lambda$ | power | $\lambda$ | power | $\lambda$ | power |
| p=0.1 | 0.051 | 0.056 | 1.49 | 0.231 | 7.54 | 0.784 |
| p=0.5 | 0.051 | 0.056 | 1.38 | 0.218 | 6.18 | 0.701 |

```
N=500                                                  # sample size
p=.1                                                   # parameters, p, and r,
distortion
r=.1
q=1-p
fr=c(p^2, 2*p*q-r, q^2)
fr=fr/sum(fr)                                          # normalize
onfr=round(N*fr)                                       # "observed" data
pe=(onfr[1]*2+onfr[2])/(2*N)                           # estimate p
qe=1-pe                                                # corresponding estimate for q
efr=c(pe^2, 2*pe*qe, qe^2)                             # expected freq under H-W
LoglA=0
for (1 in 1:3)
{
  logLA=logLA+onfr[i]*log(fr[i])
}
logL0=0
for (1 in 1:3)
{
  logL0=logL0+onfr[i]*log(efr[i])
}
logLA=-2*logLA
logL0=-2*logLA
lambda=logL0-logLA
alpha=0.5
df=1
c=qchisq(alpha, df, lower.tail=F)                      # the critical value
power=pchisq(c, df=df, ncp=lambda, lower.tail=F)       # 1-beta, power
beta=1-power                                           # beta
print(c(N, alpha, lambda, power))
```

**Panel 5.2** R script used to estimate the power to detect the effect of disruptive selection on HWE in N=500, with $\alpha$=0.05. Results shown in Table 4.

exactly the same results as ML estimation. Specifically, the least-squares estimates, when plugged into the log-likelihood function, satisfy $\Delta \text{LogL}(\boldsymbol{\theta};\mathbf{X})/\Delta\boldsymbol{\theta} = 0$. In addition, least-squares estimation is known to be quite robust to violations of the assumptions. The results thus retain their utility in ML-based statistical inference. The robustness to violations of distributional assumptions renders this method highly attractive (Feingold, 2002).

Although regression models are amenable to log-likelihood difference testing, often the Wald test is used to determine whether the QTL effect is significant (e.g. Visscher and Hopper, 2001). To explain the gist of it, we first return to the binomial example presented above. In that example, we observed an ML estimate of

$\hat{\theta} = 0.3$. We do not expect this to be necessarily the true value, as it is based on a sample of just 10 flips. Rather we expect the estimate to display sampling fluctuation, that is, variation in the estimate from one sample to the next. This is illustrated in *Figure 5.3*, where given an N of, say, 20 and the $\theta$ of 0.5, the estimate varies quite considerably. The question thus arises whether the observed value of 0.3 deviates in a statistically significant sense from some hypothesized value, such as 0.5, given that the estimate is subject to sampling variation. The *standard error* of an ML estimate reflects this sampling variation. As shown in *Figure 5.3*, the standard error can be interpreted as the expected standard deviation of the ML estimate obtained in repeated sampling. Technically, the standard error is calculated by taking the square root of the inverse of the second order derivative of the log-likelihood function, with respect to the parameter, $\text{var}[\hat{\theta}]^{1/2} = \text{se}(\hat{\theta}) = [((^2\text{LogL}(\theta;\text{N},\text{X})/(^2\theta]^{-1/2}$. The standard error of the estimate can be obtained by substituting the ML estimate for $\theta$, that is X/N. In the case of the binomial, this equals $\text{se}(\hat{\theta}) = ((\text{X}*\text{N}-\text{X}^2)/\text{N}^3)$. If we observe three tails in 10 flips, the estimate of $\theta$ equals 0.3, and the standard error equals 0.145. If $H_0$ is correct, we expect the estimate of $\theta$, upon repeated sampling, to be approximately normally distributed $\hat{\theta} \sim \text{N}[\theta_0, \sqrt{((\text{X}*\text{N}-\text{X}^2)/\text{N}^3)}]$.

One way to test whether an ML estimate $\hat{\theta}$ is equal to the $H_0$ value $\theta$ is based on the standard error. Letting $\theta_0$ denote the value under the null hypothesis, the test statistic $\text{T} = (\hat{\theta} - \theta_0)/\text{se}(\hat{\theta})$. In the binomial example, T equals (0.3-0.5)/0.145 = –1.38. This is known as the Wald statistic. Under $H_0$, T follows a Student t distribution, with df = N–1, and, asymptotically, a standard normal distribution. More generally, the standard error of an ML estimate is calculated as follows. Let $\mathbf{I}[\hat{\theta}]$ denote the matrix of second order partial derivatives $\partial^2\text{LogL}(\hat{\theta};\mathbf{X})/\partial\hat{\theta}\,\partial\hat{\theta}^{\text{t}}$, that is the so-called information matrix (Azzelini, 1996). The standard error of the i-th element in $\hat{\theta}$ is the square root of the i-th diagonal element of the inverse of this matrix, $\mathbf{I}[\hat{\theta}]^{-1}$. We are concerned here with a univariate test, but the Wald test procedure has a straightforward multivariate extension (Sorensen and Gianola, 2002, p. 179). The t-test and the $\chi^2$ test are related, as $\text{t}([\text{N}-1])^2 = \chi^2(1)$, asymptotically.

The power calculations for the Wald test proceed along exactly the same lines as the log-likelihood differences test. In fact, because asymptotically $\text{t}([\text{N}-1])^2 = \chi^2(1)$, the NCP $\lambda$ in the Wald test equals the square root of the NCP $\lambda$ of the log-likelihood difference test. To illustrate this, when say $\lambda = 6$ and $\alpha = 0.05$, a sample size of N = 4000 confers a power of about 0.79 using the log-likelihood differences test. Here is the R code to calculate the power for both the log-likelihood and the Wald test (*Panel 5.3*).

```
alpha=.05
df1=1
N=4000
lambda1=6
C1=qchisq(alpha*2, df1, lower.tail=F)
power1=pchisq(C1, df=df1, NCP=lambda1, lower.tail=F)
lambda2=sqrt(lambda1)
df2=N-1
C2=qt(alpha, df2, lower.tail=F)
power2=pt(C2, df=df2, NCP=lambda2, lower.tail=F)
print(c(power1, power2))
```

**Panel 5.3** R code to calculate the power for the log-likelihood and Wald tests.

As mentioned above, the third test procedure in inference based on the likelihood, the score test, is also applied in regression modeling of sib pair data. We refer the reader to Azzelini (1996), Sorensen and Gianola (2003) for a general discussion of this test, and to Putter *et al.* (2002), and Feingold (2002) for a discussion of the application in QTL analysis.

## 5.6 Sufficient Statistics

Above we calculated the NCP $\lambda$ by analyzing so-called summary statistics. In so doing we assume that these statistics are *sufficient* in the sense that they retain all the information in the data that is relevant to the log-likelihood. For instance, if the observed data are normally distributed, the sample mean and covariance matrix contain all the information, and are thus sufficient. If the case of N = 1000 cases and two variables, one can analyze 2000 elements in the complete data matrix **X**, or just the $2 \times 2$ covariance matrix and two means (a total of only 3+2 = 5 observed statistics). As demonstrated above, because the covariance matrix is sufficient, we can base our power calculations on the population value of the matrix under $H_A$, and obtain the NCP $\lambda$ by fitting H0. Similarly, in the multinomial distribution, the genotype counts ($X_1$, $X_2$, $X_3$) are sufficient statistics. So if N = 500 cases, we do not need the complete vector of 500 outcomes, we only require the numbers $X_1$, $X_2$, and $X_3$. The availability of sufficient statistics greatly facilitates numerical power calculations, such as those presented above. The availability of sufficient statistics allows one to derive analytic expressions, where the NCP $\lambda$ is expressed as an explicit function of the parameters. For instance, Sham *et al.* (2000; see also Chen and Abecasis, 2006; Rijsdijk *et al.*, 2001; Williams and Blangero,

1999; Yu *et al.*, 2004) exploited the availability of summary statistics to obtain analytic expressions for the expected values of the NCP in a variety of models for QTL analysis, including the QTL linkage model presented above. These expressions for the NCP $\lambda$ form the basis for the genetic power calculator of Purcell *et al.* (2003).

Summary statistics, however, are not always available. In the case of some distributions, such as the Cauchy distribution, sufficient statistics do not exist at all (e.g. Box and Tiao, 1973, p. 64). Happily, they do exist for the most commonly applied distributions. Even so, summary statistics are not always available. If one expects data to be missing, the nature of the 'missingness' may be such that summary statistics no longer retain all the information in the data that is relevant to ML estimation of the unknown parameters. Similarly, if a parameter is expected to be continuous (e.g. the proportion of alleles shared IBD), sufficient statistics may not be available. Generally, if sufficient statistics are not available one may resort to a simulation study involving the analysis of a large number of simulated datasets. Simulation studies provide empirical estimates of $\lambda$, and so of the power.

## 5.7 Conclusions and Limitations

The aim of the present chapter was to explain the workings of statistical inference based on the likelihood. In an attempt to produce a reasonably self-contained text, we included brief accounts of ML estimation, and the most current likelihood-based test statistics (the Wald test and the likelihood-ratio test). Finally, we emphasized computational aspects of carrying out power calculations, which are perfectly tractable provided one can avail oneself of sufficient statistics (population values of the summary statistics according to $H_A$), and one has at one's disposal software to integrate the distributions of the tests statistics under $H_0$ and $H_A$. As we have seen, the R program is a great resource in this respect (see *Table 5.2*). The genetic power calculator (Purcell *et al.*, 2003) can be used to evaluate power in a number of standard QTL linkage and association designs.

The present chapter is limited in many respects. We have focused on ML estimation and inference, as this is the dominant method in QTL analysis. We have not considered Bayesian estimation and testing (Sorensen and Gianola, 2002), even though this is attracting a good deal of attention due to advances in statistical computing, and because of its flexibility in model specification (Eaves and Erkanli, 2003; Eaves *et al.*, 2005; van den Berg *et al.*, 2006a, 2006b). Within the ML framework we have limited ourselves to the standard asymptotic tests. Computationally intensive

methods provide important alternatives to asymptotic tests, such as permutation testing. For instance, Churchill and Doerge (1994) discuss the use of permutation testing to determine empirical critical values associated with overall and single test $\alpha$s. This method was used by Posthuma *et al.* (2005) in a linkage analysis of intelligence data. While computer intensive methods are important and useful, calculations based on standard asymptotic tests remain an important point of departure in assessing power.

Power calculations primarily serve the purpose of establishing that $1-\beta$ is large enough to justify the time, effort, and expense of a given study. However, power calculations are a useful source of information in their own right. Power calculations provide useful insight into the role of peripheral variables (e.g. background variance) in a given design, and may suggest ways of improving power.

## Further reading

Allison, D.S. and Faith, M.S. (2000) The concept of genome-wide power and a consideration of its potential use in mapping polygenic traits: the example of sib-pairs. In: *Advances in Twin and Sib-pair Analysis* (eds T.D. Spector, H. Snieder and A.J. Macgregor). Greenwich Medical Media, London, pp. 181–188.

Allison D.B., Thiel B., St. Jean P., Elston R.C., Infante M.C. and Schork N.J. (1998) Multiple phenotype modeling in gene mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* **63**: 1190–1201.

Allison, D.B., Neale, M.C., Zannolli, R., Schork, N J., Amos, C.I. and Blangero, J. (1999) Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.* **65**: 531–544.

Allison, D.B., Fernandez, J., Heo, M. and Beasley, M. (2000) Testing the robustness of the new Haseman-Elston quantitative trait loci (QTL) mapping procedure. *Am. J. Hum. Genet.* **67**: 249–252.

Boomsma D.I. (1996) Using multivariate genetic modeling to detect pleiotropic quantitative trait loci. *Behav. Genet.* **26**: 161–166.

Boomsma D.I. and Dolan, C.V. (1998) A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factors scores. *Behav. Genet.* **28**: 5.

Boomsma, D.I. and Dolan, C.V. (2000) Multivariate QTL analysis using structural equation modeling: a look at power under simple conditions. In: *Advances in Twin and Sib-pair Analysis* (eds T.D. Spector, H. Snieder and A.J. Macgregor). Greenwich Medical Media. London, pp. 203–218.

Cardon, L.R. and Fulker, D.W. (1994). The power of interval mapping of quantitiative trait loci, using selected sib pairs. *Am. J. Hum. Genet.* **55**: 825–833.

Carey, G. and Williamson, J. (1991) Linkage analysis of quantitative traits: increased power by using selected samples. *Am. J. Hum. Genet.* **49**: 786–796.

Crainiceanu, C.M. and Ruppert, D. (2004) Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B* **66**: 165–185.

Derks, E.M., Dolan, C.V. and Boomsma, D.I. (2004) Effects of censoring on parameter estimates and power in genetic modeling. *Twin Res.* **7**: 659–669.

Derks E.M., Dolan C.V. and Boomsma D.I. (2006) Statistical power to detect genetic and environmental influences in the presence of missing data. *Twin Res. Hum. Genet* (submitted).

Devlin, B, Daniels, M, and Roeder, K. (1997) Heritability of IQ. *Nature* **388**: 468–471.

Diao, G. and Lin, D.Y. (2006) Improving the power of association tests for quantitative traits in family studies. *Genet. Epidemiol.* **30**: 301–313.

Dolan, C.V., and Boomsma, D.I. (1999) Optimal selection of sib pairs from random samples for linkage analysis of a QTL using the EDAC test. *Behav. Genet.* **28**: 3.

Dolan, C.V., Boomsma, D.I. and Neale, M.C. (1999) A note on the power provided by sibships of sizes 2, 3, and 4 in genetic covariance modeling of a codominant QTL. *Behav. Genet.* **29**: 3.

Dolan, C.V., van der Sluis, S. and Grasman, R. (2005) A note on normal theory power calculation in structural equation modeling with data missing completely at random. *Struct. Equation Model.* **12**: 245–262.

Eaves, L. and Meyer, J. (1994) Locating human quantitative trait loci – guidelines for the selection of sibling pairs for genotyping. *Behav. Genet.* **24***:* 443–455.

Gelman, A, Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*, 2nd Edn. Chapman and Hall, London.

Gillespie, N.A. and Neale, M.C. (2006) A finite mixture model for genotype and environment interactions: detecting latent population heterogeneity. *Twin Res. Hum. Genet.* **9**: 412–423.

Gu, C., Todorov, A. and Rao, D.C. (1996) Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genet. Epidemiol.* **13**: 513–533.

Lander, E. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**: 241–247.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical analysis with missing data*, 2nd Edn. Wiley and Sons, New York, NY.

Lubke, G.H., Neale, M.C. and Dolan, C.V. (2004) Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. *Twin Res.* **7**: 292–298.

Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

Markon, K.E. (2006) Semiparametric maximum likelihood variance component estimation using mixture moment structure models. *Twin Res. Hum. Genet.* **9**: 360–366.

Muthèn, B, Asparouhov, T. and Rebollo, I. (2006) Advances in behavioral genetics modeling using Mplus: applications of factor mixture modeling to twin data. *Twin Res. Hum. Genet.* **9**: 313–328.

Muthèn, L.K. and Muthèn, B.O. (2002) How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equation Model.* **9**: 599–620.

Neale, M.C. and Miller, M.B. (1997) The use of likelihood-based confidence intervals in genetic models. *Behav. Genet.* **27**: 113–120.

Neale, M.C, Lubke, G.H., Aggen, S.H. and Dolan, C.V. (2005) Problems with using sum scores for estimating variance components: contamination and measurement non-invariance. *Twin Res. Hum. Genet.* **16**: 553–568.

Purcell, S. (2002) Variance components models for gene-environment interaction in twin data. *Twin Res.* **5**: 554–571.

Purcell, S., Cherny, S.S., Hewitt, J.K. and Sham, P. (2001) Optimal sibship selection for genotyping in quantitative trait locus linkage analysis. *Hum. Hered.* **52**: 1–13.

Putter, H., Lebrec, J. and van Houwelingen, J.C. (2003) Selection strategies for linkage studies using twins. *Twin Res.* **6**: 377–382.

Risch, N. and Zhang, H.P. (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**: 1584–1589.

Schafer, J. L. and Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychol. Methods* **7**: 147–177.

Schork, N.J., Allison, D.B. and Thiel, B. (1996) Mixture distributions in human genetics research. *Stat. Methods Med. Res.* **5**: 155–178.

Sham, P. C., and Purcell, S. (2001) Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.* **68**: 1527–1532.

Sham, P.C., Purcell, S., Cherny, S.S. and Abecasis, G.R. (2002) Powerful regression-based quantitative trait linkage analysis for general pedigrees. *Am. J. Hum. Genet.* **71**: 238–253.

Turkheimer, E., Haley, A., Waldron, M., D'Onofio, B. and Gottesman, I.I. (2003) Socioeconomic status modifies heritability of IQ in young children. *Psychol. Sci.* **14**: 623–628.

van den Berg, S.M., Glas, C.A.W. and Boomsma, D.I. (2006) Variance decomposition using an IRT measurement model. Submitted.

van den Oord, E.J.C.G., Simonoff, E., Eaves, L. J., Pickles, A., Silberg, J. and Maes, H. (2000) An evaluation of different approaches for behavior genetic analyses with psychiatric symptom scores. *Behav. Genet.* **30**: 1–18.

Visscher, P.M. and Duffy, D.L. (2006) The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. *Genet. Epidemiol.*, **30**: 30–36.

Zhang, H. and Risch, N. (1996) Mapping quantitative trait loci in humans by use of extreme concordant sib pairs: selected sampling by parental phenotypes. *Am. J. Hum. Genet.* **59**: 951–957.

## References

Almasy, L., and Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**: 1198–1121.

Azzelini, A. (1996) *Statistical Inference Based on the Likelihood.* Chapman and Hall, London.

Bollen, K.A. (1989) *Structural Equations with Latent Variables*. Wiley, New York, NY.

Box, G.E.P. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

Carey, G.C. (2004) Cholesky problems. *Behav. Genet.* **34***:* 633.

Carey, G.C. (2006) Cholesky problems. *Behav. Genet.* (in press).

Chen, W.-M. and Abecasis, G.R. (2006) Estimating the power of variance component linkage analysis in large pedigrees. *Genet. Epidemiol.* **30**: 1–14.

Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.

Cohen, J. (1992) A power primer. *Psychol. Bull.* **112**: 155–159.

Dalgaard, P. (2002) *Introductory Statistics with R*. Springer, New York, NY.

Dardanoni, V. and Forcina, A. (1998) A unified approach to likelihood inference on stochastic ordering in a nonparametric context. *J. Am. Stat. Assoc.* **93**: 1112–1123.

Dominicus, A., Skorndal, A., Gjessing, H.K., Pedersen, N.L. and Palmgren, J. (2006) Likelihood ratio tests in behavioral genetics: problems and solutions. *Behav. Genet.* **36**: 331–340.

Eaves, L.J. and Erkanli, A. (2003) Markov Chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and G×E interaction. *Behav. Genet.* **33**: 279–299.

Eaves, L.J., Neale, M.C. and Maes, H.H. (1996) Multivariate multipoint linkage analysis of quantitative trait loci. *Behav. Genet.* **26**: 519–526.

Eaves, L.J., Erkanli, A., Silberg, J., Angold, A., Maes, H.H. and Foley, D. (2005) Application of Bayesian inference using Gibbs sampling to item-response theory modeling of multi-symptom genetic data. *Behav. Genet.* **35**: 765–780.

Evans, M., Hastings, N. and Peacock, B. (2000) *Statistical Distributions*, 3rd Edn. Wiley, New York, NY.

Ewens, W.J. and Grant, G.R. (2001) *Statistical Methods in Bioinformatics*. Springer, New York, NY.

Ferreira, M.A.R. (2004) Linkage analysis: principles and methods for the analysis of human quantitative traits. *Twin Res.* **7**: 513–530.

Fulker, D.W. and Cardon, L.R. (1994) A sib-pair approach to interval mapping of quantitative trait loci. *Am. J. Hum. Genet.* **54**: 1092–1103.

Fulker, D.W. and Cherny, S.S. (1996) An improved multipoint sib-pair analysis of quantitative traits. *Behav. Genet.* **26**: 527–532.

Gill, P.E., Murray, W. and Wright, M.H. (1981) *Practical Optimization*. Academic Press, London.

Greene, W.H. (1993) *Econometric Analysis*, 2nd Edn. Macmillan, New York, NY.

Haseman, J.K. and Elston, R.C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**: 3–19.

Heath, A.C. and Eaves L.J. (1985) Resolving the effects of phenotype and social background on mate selection. *Behav. Genet.* **15**: 15–30.

Heath, A.C., Kendler K.S., Eaves L.J. and Markell, D. (1985) The resolution of cultural and biological inheritance: informativeness of different relationships. *Behav. Genet.* **15**: 439–465.

Hewitt, J.K. and Heath, A.C. (1988) A note on computing the chi-square non-centrality parameter for power analysis. *Behav. Genet.* **18**: 105–108.

Hopper, J.L. and Matthews, J.D. (1982) Extensions to multivariate normal models for pedigree analysis. *Ann, Hum. Genet.* **46**: 373–383.

Kraemer, H.C. (1985) A strategy to teach the concept and application of power of statistical tests. *J. Educat. Stat.* **10**: 173–195.

Martin, N.G., Eaves, L.J., Kersey, M.J. and Davies, P. (1978) The power of the classical twin study. *Heredity* **40**: 97–116.

Miller, I. and Miller M. (2004) *John E. Freund's Mathematical Statistics with Applications*, 7th Edn. Pearson Education International, Upper Saddle River, NJ.

Myung, I.J. (2003) Tutorial on maximum likelihood estimation. *J. Math. Psychol.* **47**: 90–100.

Nance, W.E. and Neale, M.C. (1989) Partitioned twin analysis: a power study. *Behav. Genet.* **19**: 143–150.

Neale, M.C. and Cardon, L.R. (1992) *Methodology for Genetic Studies of Twin and Families*. Kluwer Academic, Dordrecht.

Neale, M.C., Eaves, L.J. and Kendler, K.S. (1994) The power of the classical twin study to resolve variation in threshold traits. *Behav. Genet.* **24**: 239–258.

Posthuma, D. and Boomsma, D.I. (2000) A note on the statistical power in extended twin designs. *Behav. Genet.* **30**: 147–158.

Posthuma, D., Beem, L.A., de Geus, E.J.C., van Baal, G.C.M., von Hjelmborg, J.B. Iachine, I. and Boomsma, D.I. (2003) Theory and practice in quantitative genetics, *Twin Res.* **6**: 361–376.

Posthuma, D., Luciano, M., de Geus, E.J.C., Wright, M.J., Slagboom, P.E., Montgomery, G.W., Boomsma, D.I. and Martin, N.G. (2005) Genome-wide scan for intelligence identifies quantitative trait loci on 2q and 6p. *Am. J. Hum. Genet.* **77**: 318–326.

Purcell, S. Cherny, S.S. and Sham, P.C. (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**: 149–150.

Putter, H., Sandkuijl, L.A. and van Houwelingen, J.C. (2002) Score test for detecting linkage to quantitative traits. *Genet. Epidemiol.* **22**: 345–355.

Rietveld, M.J.H., Posthuma, D., Dolan, C.V. and Boomsma, D.I. (2003) ADHD: sibling interaction or dominance: an evaluation of statistical power. *Behav. Genet.* **33**: 247–255.

Rijsdijk, F.V., Hewitt, J.K. and Sham, P.C. (2001) Analytic power calculation for QTL linkage analysis of small pedigrees. *Eur. J. Hum. Genet.* **9**: 335–340.

Saris,W.E. and Satorra, A. (1993) Power evaluations in structural equation models. In: *Testing Structural Equation Models* (eds K.A. Bollen and J.S. Long). Sage, Newbury Park, CA,pp. 181–204.

Satorra, A. and Saris, W.E. (1985) The power of the likelihood ratio test in covariance structure analysis. *Psychometrika* **50**: 83–90.

Schmitz, S. Cherny, S.S. and Fulker, D.W. (1998) Increase in power through multivariate analyses. *Behav. Genet.* **28**: 357–364.

Self, S.G. and Liang, K.Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**: 605–610.

Sham, P.C. (1998) *Statistics in Human Genetics*. Arnold, London.

Sham, P.C., Cherny, S.S., Purcell, S. and Hewitt, J.K. (2000) Power of linakge versus association analysis of quantitative traits by use of variance components models, for sibship data. *Am. J. Hum. Genet.* **66**: 1616–1630.

Sorensen, D. and Gianola, D. (2002) *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer, New York, NY.

Stram, D.O. and Lee, J.W. (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**: 1171–1177.

Turkheimer, E. and Gottesman, I.I. (1991) Is $H^2 = 0$ a null hypothesis anymore? *Behav. Brain Sci.* **14**: 410–411.

van den Berg, S.M., Setiawan, A., Bartels, M., Polderman, T.J.C., Van der Vaart, A.W. and Boomsma, D.I. (2006a) Individual differences in puberty onset in girls: Bayesian estimation of heritabilities and genetic correlations. *Behav. Genet.* [Epub ahead of print].

van den Berg, S.M., Beem, L. and Boomsma, D.I. (2006b) Fitting genetic models using Markov Chain Monte Carlo algorithms with BUGS. *Twin Res. Hum. Genet.* **9**: 334–342.

Venables, W.N., Smith, D.M. and the R Development Core Team (2001) *An Introduction to R.* Network Theory Limited, Bristol.

Visscher, P.M. and Hopper, J.L. (2001) Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. *Ann. Hum. Genet.* **65**: 583–601.

Williams, J.T. and Blangero, J. (1999) Power of variance component linkage analysis to detect quantitative trait loci. *Ann. Hum. Genet.* **63**: 545–563.

Yu, X., Knott, S.A. and Visscher, P.M. (2004) Theoretical and empirical power of regression and maximum likelihood methods to map quantitative trait loci in general pedigrees. *Am. J. Hum. Genet.* **75**: 17–26.