1

Statistical Power and Confidence Regions

The significance level is the determining factor in the specification of the rejection region of a statistical test. Only the distribution under the null assumption of no signal plays a role in setting the level of the threshold, once the test statistic and the general form of test are decided upon. However, after setting that threshold, one can examine other statistical properties of the resulting test. A central property is *statistical power* of the test – the probability to reject the null hypothesis when a signal is present. Since this probability depends on the values of the parameters, one often speaks of the *power function* to emphasize this dependence. For a test at a single marker, this probability is obtained approximately from the normal distribution; it is a function of the noncentrality parameter given by (**??**). In this chapter we will examine the concept of power for a whole-genome scan.

The primary interest is now focused on the case where the null hypothesis is false. Statisticians define the power of an hypothesis test as the probability of concluding correctly the falsity of the null hypothesis. However, the case of a genome scan is more subtle than a simple test of hypothesis. There exists the possibility that due to random fluctuations the significance threshold is exceeded on a chromosome that does not actually contain a QTL. Unless the threshold is also crossed on a chromosome containing a QTL, one would correctly conclude that the simple null hypothesis of no QTL anywhere in the genome is false, but would identify the chromosomal location of the QTL incorrectly. We are particularly interested in the probability of exceeding the threshold on a chromosome containing a QTL, or perhaps even at some marker close to the QTL, say within 20 cM. Although any definition of power in this context is somewhat arbitrary, in this book the power to detect a particular QTL will refer to the probability of correctly identifying the chromosome inhabited by the QTL. This means that when there is more than one QTL, power refers to specific QTL and can vary from one to another, depending on the effect of the QTL on the phenotype. In the case that multiple QTL lie on the same chromosome, one might want to make more subtle distinctions.

While keeping this possibility in mind, we shall for the most part ignore it in our statistical analysis.

At one end of the spectrum, when there is no QTL or only very weak QTL on a given chromosome, the power function is essentially equal to the (chromosome specific) significance level. At the other extreme are parameter values which correspond to a signal that is so large as to make the power approximately equal to one. We will mainly be interested in interim parameter values, values for which the power function is in the range 50%–95%.

The main application of the power function is to help us choose an experimental design – especially the breeding design, marker density, and sample size. The significance level is set to be some fixed value – typically 5% – regardless of the design. The differences between designs will be reflected in their power to detect the signal. Thus, for example, a sample size which is too small, or a collection of markers which is not dense enough, may compromise the chances of successful detection of a QTL. On the other hand, it is neither economically efficient nor ethical to use more animals then needed. Moreover, since genotyping is an expensive component in a genome scan, using more markers than needed is a waste of time and money that can be used for other purposes. Careful planning of an experiment can ensure efficient distribution of resources, without a substantial reduction in power. In the body of the text we focus on issues of sample size determination and the selection of the inter-marker spacing in the context of the backcross design. The power of other breeding designs is left to exercises in the problem set at the end of the chapter.

In the first section we identify the terms that affect the power of detection in a whole-genome scan. In the second section we introduce analytic formulas for the power. These formulas, like the formulas used for the computation of the significance level, are given in the context of the Ornstein-Uhlenbeck process. They will allow us to analyze the power function and to examine the effect of changing the values of various parameters. Consequently, in the third section we apply these formulas to select a good experimental design for the detection of a QTL. In the last two sections we deal with issues related more to estimation. In the fourth section we consider the construction of confidence intervals for the location of a QTL and in the last section the construction of a lower confidence bound for the effect of the QTL.

1.1 The Power to Detect a QTL

In this section we identify the parameters that determine the statistical properties of the monitoring process in the presence of a QTL. The examination is carried out in the context of local alternatives. In the case of a single marker, which was analyzed in Chap. **??**, this corresponded to considering a shifted normal distribution. Similarly, for the multi-marker process, where the null distribution corresponded to the Gaussian Ornstein-Uhlenbeck process,

the computation under the alternative will involve the same process with a shifted mean function. Since we deal with local alternatives, the correlation structure is not affected. The power to detect a signal is the probability that the maximum absolute value of the process, with the shifted mean, exceeds the threshold.

Recall that for the single marker process the expected value under the alternative of the test statistic at a marker equals its expectation at the QTL multiplied by the correlation between the marker and the QTL (see **??**). For the backcross design and the simple model of QTL, we use the expectation at the QTL itself, which equals $\xi = (\alpha_0 + \delta_0)/(2\sigma_y)$. Here α_0 corresponds to the (local) parameter of additive effect, δ_0 is the (local) parameter of dominance effect, and σ_y is the standard deviation of the phenotype. In terms of the original parameters of the model, $\alpha_0 = n^{1/2}\alpha$ and $\delta_0 = n^{1/2}\delta$. The correlation between marker and QTL for a backcross design, under the Haldane model of recombination, is equal to $1 - 2\,\theta = \exp\{-0.02\,|t - \tau|\}$. Here $|t - \tau|$ corresponds to the distance between a QTL located τ cM from the telomere, and a marker located t cM from the telomere. This information is summarized by the formula

$$\mathrm{E}(Z_t) = \frac{\alpha_0 + \delta_0}{2\,\sigma_y}(1 - 2\,\theta) = \xi \exp(-0.02\,|t - \tau|)\ . \tag{1.1}$$

If the marker and the QTL are not on the same chromosome, then $\theta = 1/2$, the genetic distance from the QTL is defined to be infinite, and the expectation is equal to 0.

Equation (1.1) gives a complete description of the mean function of the multi-marker process under the model of a single QTL. The QTL is located on some chromosome, τ cM from the telomere. The mean function for the markers on the same chromosome is a function of their distance from the QTL. It decreases exponentially fast, on both sides of the QTL, as the distance from the QTL increases. The mean function of the multi-marker process over the other 19 chromosomes, which do not contain the QTL, is identically equal to 0. The covariance structure of the process under the alternative is identical to the covariance structure under the null assumption. Thus, the standard deviation of the test statistics Z_t is equal to one, regardless of the location of the marker and its distance from the QTL. The correlation between any pair of markers is a function of the genetic distance between them. If the two markers are located on different chromosomes, then the genetic distance between them is infinite, and the two markers are uncorrelated. If the markers are located on the same chromosome, s and t cM from the telomere, then the correlation between them equals $\exp\{-0.02\,|t - s|\}$.

The distribution of the multi-marker process can be generated under the alternative similarly to the way it is generated under the null. The basic random process in both cases is the Ornstein-Uhlenbeck process. This processes describes the stochastic element in the behavior of markers for each chromosome. The deterministic element in the behavior is the mean function. This

deterministic element is the difference between the null distribution and the distribution under the alternative. For the latter case a non-zero mean function is added for any chromosome carrying a QTL. Consequently, one can use the R function "OU.sim" that we wrote in the previous chapter in order to simulate the multi-marker process on a given chromosome. The vector of means may be added to the simulated process of marker-specific test statistics. We implement this approach in the function "add.qtl". This function takes as an input the matrix produced by "OU.sim", the location of markers, the coefficient of recombination β, and two new parameters: "q", the location on the first chromosome of the QTL (measured in cM from the telomere); and "xi", the noncentrality parameter. The mean vector is added and the altered scanning process is returned:

```
> add.qtl <- function(Z,beta,markers,q,xi)
+ {
+     d <- dim(Z)
+     if (length(markers) != d[2])
+         stop("Number of columns of simulated matrix
+         does not match the number of markers")
+     mu <- xi*exp(-beta*abs(markers - q))
+     Z <- sweep(Z,2,mu,"+")
+     return(Z)
+ }
```

The function "stop" may be used in order to stop a function in the case of a fatal error. The argument of the function is printed out if the error occurs. Similarly, a warning may be produced, in the case of a nonfatal errors, with the function "warning".

The function "sweep" returns a matrix obtained from an input matrix by sweeping the elements of a vector. The first argument is the input matrix. The second argument is the margin over which the elements of the vector should be applied. The third argument is the vector, and the fourth argument is the binary function that produces the elements of the output matrix from the application of the binary function to the element of the input matrix and the appropriate element of the vector. More generally, this function may be applied to arrays, which are higher-dimension extension of matrices.

Let us generate some paths of the resulting multi-marker process:

```
> markers <- seq(0,80,by=10)
> beta <- 0.02; q <- 40; xi <- 4
> Z <- NULL
> for (i in 1:20) Z <- cbind(Z,OU.sim(beta,markers,n.iter=5))
Loading required package: MASS
> chr1 <- 1:length(markers)
> Z[,chr1] <- add.qtl(Z,beta,markers,q,xi)
Error in add.qtl(Z, beta, markers, q, xi) :
```

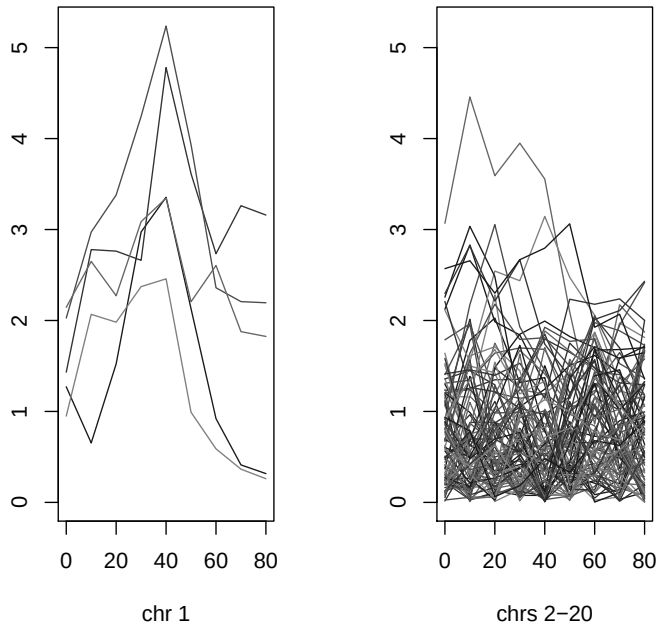

Fig. 1.1. Sample paths of the absolute value of the Ornstein-Uhlenbeck process under the alternative.

```
        Number of columns of simulated matrix
        does not match the number of markers
> Z[,chr1] <- add.qtl(Z[,chr1],beta,markers,q,xi)
```

Note an erroneous application of the function "`add.qtl`" resulted in an error message, which helped us to detect the source of the error and debug the mistake. The function "`OU.sim`" did not produce an error message since we added to its definition the expression "`require(MASS)`".

The paths of the scanning process are plotted in Fig. 1.1. The paths for chromosome 1 are shown on the left plot and the paths of chromosomes 2–20 are shown on the right plot. Observe that the values of the test statistics in the middle of chromosome 1 have levels which are consistently large. However, occasional extreme values can occur also in other chromosomes. The code that produces the two plots is:

```
> plot(c(0,80),range(abs(Z)),type="n",xlab="chr 1",ylab="")
> for(i in 1:5) lines(markers,abs(Z[i,chr1]),col=gray(i/7))
> plot(c(0,80),range(abs(Z)),type="n",xlab="chrs 2-20",
```

```
+     ylab="")
> for(j in 2:20)
+ {
+    chr = chr1+(j-1)*length(chr1)
+    for(i in 1:5) lines(markers,abs(Z[i,chr]),col=gray(i/7))
+ }
```

We detect a QTL if the maximum absolute value of the multi-marker process exceeds the significance threshold. Let us examine the distribution of this maximum, both when the signal is absent and when it is present. However, for the sake of determining the significance level we simulate these distributions for a whole-genome scan with markers at 0, 10, 20, ..., 80 cM. The power is considered only in the context of the chromosome that contains a QTL. Specifically, the QTL is located 40 cM from the telomere on chromosome 1. Note that a marker happens to be located right at that spot. The noncentrality parameter at the QTL equals 4 in our simulations.

```
> Z0 <- NULL
> for (i in 1:20) Z0 <- cbind(Z0,OU.sim(beta,markers))
> Z1 <- add.qtl(Z0[,chr1],beta,markers,q,xi)
> d0 <- density(apply(abs(Z0),1,max),from=1,to=7)
> d1 <- density(apply(abs(Z1),1,max),from=1,to=7)
> plot(d0,main="Densities of maximal statistics",
+   xlab="max |Z|")
> lines(d1,lty=2)
> legend(4.7,1,legend=c("Under H0","Under H1"),lty=1:2)
```

Examine the distributions of the test statistic under the two scenarios (Fig. 1.2). Note that although the distribution of the test statistic under the alternative tends to get higher values, still the two distributions cannot be separated perfectly. Any reasonable threshold that eliminates most of the exceedences of the test statistic under the null distribution must eliminate some occurrences of the test statistic under the alternative distribution as well. The traditional way to resolve this difficulty is to set a threshold with an acceptable proportion of the null distribution above it. This proportion is the significance level of the test – the probability of falsely rejecting the null hypothesis – and is typically set to be equal to 5%. The proportion of the distribution under the alternative that is *below* the threshold corresponds to the probability of falsely accepting the null hypothesis, and is called *the probability of a type II error.* Under the alternative the statistical power corresponds in the plot to the portion above the threshold, i.e., one minus the probability of a type II error. The larger this probability the better.

In order to determine the power, one must specify the threshold. In the previous chapter the value of 3.56 was suggested as a threshold for a genome scan in the backcross design with inter-marker spacing of 10 cM. This threshold was derived from an analytical formula for the significance value. Let us see the actual significance level and power via simulations.

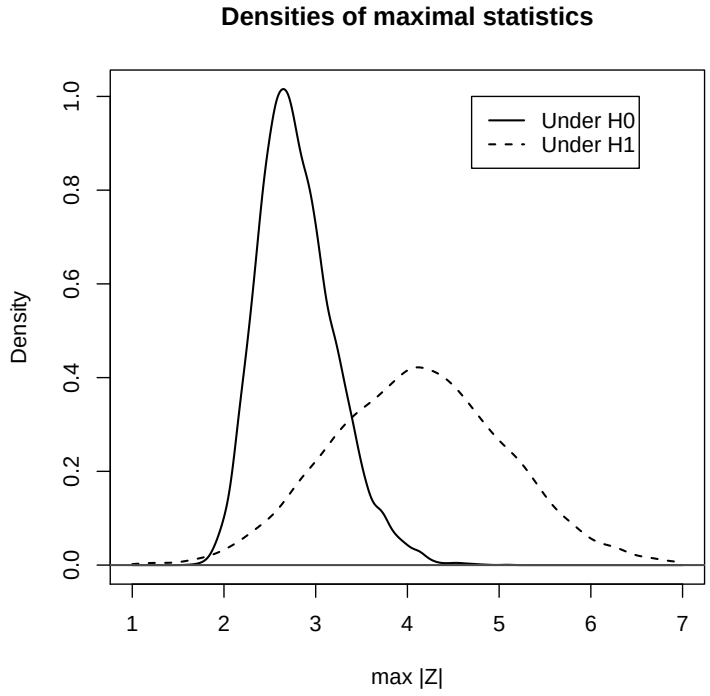

Fig. 1.2. The distribution of the test statistic under the null and under the alternative hypotheses.

```
> mean(apply(abs(Z0),1,max) >= 3.56)
[1] 0.0482
> mean(apply(abs(Z1),1,max) >= 3.56)
[1] 0.7044
```

It follows that the power to detect a QTL, located 40 cM from the telomere and with an (asymptotic) noncentrality parameter of 4, is about 70%.

What would be the power if the QTL happened to be between two markers? near the end of the chromosome, rather than near the center? for a smaller value of the noncentrality parameter? a larger value? We can obtain answers to these questions using simulations. For example:

```
> q <- 35; xi <- 4
> Z1 <- add.qtl(Z0[,chr1],beta,markers,q,xi)
> mean(apply(abs(Z1),1,max) >= 3.56)
[1] 0.6482
> q <- 5; xi <- 4
> Z1 <- add.qtl(Z0[,chr1],beta,markers,q,xi)
```

```
> mean(apply(abs(Z1),1,max) >= 3.56)
[1] 0.6455
> q <- 40; xi <- 3
> Z1 <- add.qtl(Z0[,chr1],beta,markers,q,xi)
> mean(apply(abs(Z1),1,max) >= 3.56)
[1] 0.343
> q <- 40; xi <- 5
> Z1 <- add.qtl(Z0[,chr1],beta,markers,q,xi)
> mean(apply(abs(Z1),1,max) >= 3.56)
[1] 0.9363
```

The task of exploring these question can be carried out much more efficiently once we have formulas for the statistical power, similarly to the formulas we have for the significance level. In the next section we describe such formulas.

1.2 An Analytic Approximation of the Power

As we saw, the power can be affected quite heavily by the location of the QTL on the chromosome. The probability of detecting a QTL may be substantially reduced if the QTL is located midway between markers, compared to the case where the QTL is located at the same location of a marker in the middle of the chromosome, and even more so if it is near an end of the chromosome. Since the formulas for approximating the power for the case when the QTL is between markers are substantially more complex than the formula when it is located exactly at a marker, we will present here only the formula for the latter case. For the former case, however, we do provide an `R` function, but not an explicit display of the formula. The interested reader may extract the mathematical expression from the code of the function.

When a marker is located at a QTL, detection occurs on that chromosome if either (i) the test statistic associated with the QTL/marker or (ii) the process associated with the flanking markers exceeds the threshold. The test statistic at the QTL has a normal distribution with mean ξ. Thus, the probability of the first case is simply the probability that such normal variable exceeds a threshold z. The mathematical derivation of the second case proceeds by conditioning on the value of the test statistic at the QTL, and analyzing the asymptotic conditional distribution of the process at the other markers. The resulting formula for a QTL not near either end of a chromosome is:

$$\Pr\big(\max_i |Z_{i\Delta}| \geq z\big) \approx 1 - \Phi(z-|\xi|) + \phi(z-|\xi|)\big[2\,\nu/|\xi| - \nu^2/(z+|\xi|)\big]\ , \quad (1.2)$$

where $\nu = \nu\big(z\{2\,\beta\Delta\}^{1/2}\big)$. The first term, $1-\Phi(z-|\xi|)$, is the probability that the test statistic associated with the QTL/marker exceeds the threshold

z. The second term corresponds to the probability of crossing the threshold by one or the other of the two flanking processes when the value of the statistic at the QTL is below the threshold.

Remark 1.1. Recall that for $\Delta \approx 0$, i.e., when the distribution of markers on the chromosome is very dense, the correction term ν is close to one.

Remark 1.2. When the QTL is located at the first or last marker on a chromosome, there are flanking markers only to one side. Then the approximation becomes

$$\Pr\big(\max_i |Z_{i\Delta}| \geq z\big) \approx 1 - \Phi(z - |\xi|) + \phi(z - \xi)\nu/|\xi| \,. \tag{1.3}$$

The function "`power.marker`" implements Formula (1.2):

```
> power.marker <- function(z,beta,Delta,xi)
+ {
+     nu <- Nu(z*sqrt(2*beta*Delta))
+     return(1-pnorm(z-xi) +
+         dnorm(z-xi)*(2*nu/xi-(nu/(z+xi))^2))
+ }
```

Applying this approximation we get:

```
> z <- 3.56; beta <- 0.02; Delta <- 10;
> xi <- 4
> power.marker(z,beta,Delta,xi)
[1] 0.7194996
```

Compare this to the probability of 0.7044, which was obtained via simulation.

The worst case scenario is to have a QTL midway between markers. The formula corresponding to (1.2) is much more complex since it involves conditioning on the values of the process $Z_{i\Delta}$ at both flanking markers. The expression is omitted, but we use the function `power.midway` in order to approximate the power in this case.

```
> power.midway <- function(z,beta,Delta,xi)
+ {
+     ul <- 5
+     nu <- Nu(z*sqrt(2*beta*Delta))
+     zz <- z - xi*exp(-beta*Delta/2)
+     cc <- sqrt(1 - exp(-2*beta*Delta))
+     fun1 <- function(x,beta,Delta,zz,cc) dnorm(zz-x)*
+         pnorm((zz-exp(-beta*Delta)*(zz-x))/cc)
+     term1 <- integrate(fun1,0,ul,beta=beta,
+         Delta=Delta,zz=zz,cc=cc)
+     fun2 <- function(x,z,beta,Delta,zz,cc) exp(-z*x)*
```

```
+         dnorm(zz-x)*pnorm((zz-exp(-beta*Delta)*(zz-x))/cc)
+     term2 <- integrate(fun2,0,ul,z=z,beta=beta,
+         Delta=Delta,zz=zz,cc=cc)
+     fun3 <- function(x,z,beta,Delta,zz,cc) dnorm(zz-x)*
+         exp(-z*x-z*(zz-exp(-beta*Delta)*(zz-x))+z^2*cc/2)*
+         pnorm((zz-exp(-beta*Delta)*(zz-x))/cc-z*cc)
+     term3 <- integrate(fun3,0,ul,z=z,beta=beta,
+         Delta=Delta,zz=zz,cc=cc)
+     return(1-term1$value+2*nu*term2$value-nu^2*term3$value)
+ }
```

The analytical expression involves an integral. Numerical integrals of functions with respect to their first argument can be computed with the function "`integrate`". The output is a list, with the component "`value`" containing the result of the integration.

In the simulations we obtained a power of 0.6482 when $\xi = 4$, $\Delta = 10$, and the QTL is located halfway between markers. Compare this probability to the analytical approximation:

```
> power.midway(z,beta,Delta,xi)
[1] 0.6498196
```

The *power function* involves the evaluation of the statistical power over the range of parameters under the alternative distribution. In the case of a whole-genome scan using the backcross and a given set of markers, these parameters are the location of the QTL and the strength of the signal, i.e., the noncentrality parameter ξ. Let us evaluate the analytical approximations over the range of the power function.

We start with the case of a QTL, which is located next to a marker in the middle of a chromosome. We consider here the case of the backcross design ($\beta = 0.02$), and an inter-marker spacing of 10 cM:

```
> q <- 40
> xi <- seq(0,6,by=0.1)
> n <- length(xi);
> ap.marker <- p.marker <- vector(mode="numeric")
> for (i in 1:n)
+ {
+     Z1 <- add.qtl(Z0[,chr1],beta,markers,q,xi[i])
+     p.marker[i] <- mean(apply(abs(Z1),1,max)>=z)
+     ap.marker[i] <- power.marker(z,beta,Delta,xi[i])
+ }
> plot(c(0,6),c(0,1),type="n",xlab="xi",ylab="Power")
> lines(xi,p.marker)
> lines(xi,ap.marker,lty=2)
```

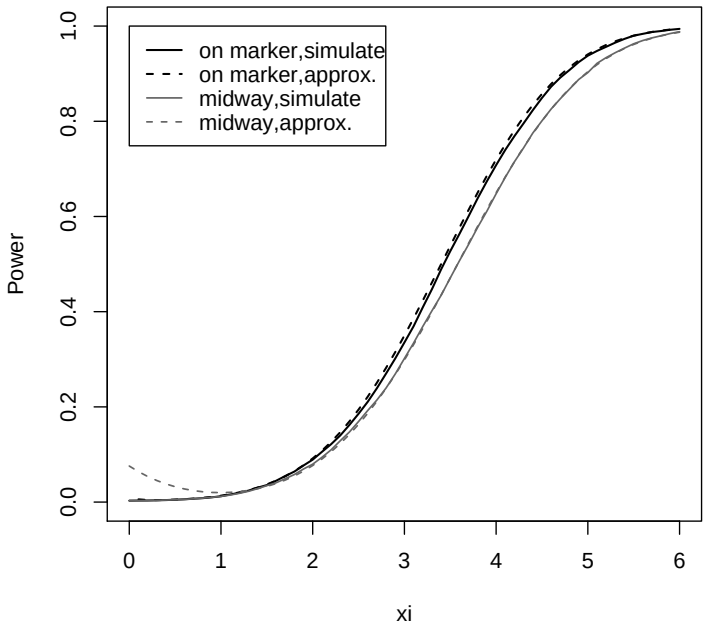

Fig. 1.3. The power function when the QTL is next to a marker and when it is midway between markers.

Next, let us consider the case of a QTL midway between markers:

```
> q <- 35
> ap.midway <- p.midway <- vector(mode="numeric")
> for (i in 1:n)
+ {
+     Z1 <- add.qtl(Z0[,chr1],beta,markers,q,xi[i])
+     p.midway[i] <- mean(apply(abs(Z1),1,max)>=z)
+     ap.midway[i] <- power.midway(z,beta,Delta,xi[i])
+ }
> lines(xi,p.midway,col=gray(0.5))
> lines(xi,ap.midway,lty=2,col=gray(0.5))
> legend(0,1,legend=c("on marker,simulate",
+    "on marker,approx.","midway,simulate","midway,approx."),
+    lty=c(1,2,1,2),col=gray(c(0,0,0.5,0.5)))
```

The resulting plot is displayed in Fig. 1.3. Note the reduction in power when the QTL is not perfectly linked to a marker. Observe good agreement

between the analytical approximation and the simulated value. This agreement is destroyed when the QTL is between markers for values of ξ less than one. Luckily, the exact evaluation of the power for such low values of the noncentrality parameter is of little practical interest.

1.3 Designing an Experiment

Experiments aimed at the dissection of the genetic component of traits in mice require substantial investment. It is unadvisable, therefore, to start such an effort, unless one is likely to obtain a successful outcome. The careful planning of the experiment is key in this regard. It ensures, on the one hand, that sufficient resources are devoted for the task. On the other hand, the optimal distribution of these resources lowers the chance of wasting both time and money.

The role of statistical experimental design is to identify the minimal requirements needed in order to be able to extract scientifically significant signals in the presence of background noise. It helps to use separate terminology in order to distinguish between *statistical significance* and *scientific significance*. Statistical significance is a formal term associated with the properties of the random mechanism underlying the background noise. It contrasts the strength of the observed signal in light of what could have been produced in a scenario where no real signal is present. The statistical significance is computed in the context of a null hypothesis, which assumes the absence of any signal. An observed signal can turn out to be statistically significant even if the underlying true signal is very weak. This can occur if the level of the background noise is low relative the amount of data gathered. The scientific significance, on the other hand, is not determined by statistical considerations. It reflects the specifics of the particular scientific discipline, and is given in terms of the strength of the underlying signal. Thus, in experimental genetics we may aim at detecting QTLs which have a strong enough effect on the phenotypic variance. This corresponds to large enough values of the locus specific heritability. The experiment is designed to discover genetic terms that have an effect above a given threshold.

To be more specific, let us consider an experiment using the backcross design. The strength of the genetic effect is given in terms of the locus specific heritability (see Chap. **??**):

$$h^2 = \frac{(\alpha + \delta)^2/4}{\sigma_y^2} \, .$$

In order to design the experiment, we may set a minimal level of this quantity. Based on this level, the specifications of the trials can be determined. In this section we will describe the computations which identify the density of the genotyped markers and the sample size required in order to ensure a

reasonably large chance to detect this minimal level of signal. We will carry these computations backwards. First, we will determine, for each inter-marker spacing, the appropriate noncentrality parameter which ensures the minimal statistical power. Second, we will determine the sample size associated with this noncentrality parameter. Finally, we will select the design which minimizes the overall cost.

Determining the Noncentrality Parameter

Thousands of polymorphic markers, scattered throughout the mouse genome, are available for use. Hundreds of thousands, and even millions more, are expected with the identification of more and more SNP markers. Although not all markers are polymorphic between a given pair of inbred strains, the availability of genetic markers is typically not a limiting factor. Consequently, in principle, we can consider any density of markers. However, in order to simplify the computations, we will analyze here only four different possibilities of inter-marker spacings: $\Delta = 20, 10, 5$, or 1 cM.

In order to have a fair comparison, we will require that all cases possess the same significance level – 5%. Consequently, the thresholds will vary with the spacing. From the computations we made when dealing with the significance level, we found that these thresholds were approximately equal to $z = 3.46, 3.56, 3.65$, and 3.78, respectively.

Let us use the root finder "`uniroot`" and apply it to the function "`power.midway`" in order to identify the value of the noncentrality parameter that produces a power of 85% for each one of the designs. Note that the power is computed for a QTL between markers. This makes the conditions less favorable for designs with larger inter-marker spacings:

```
> delta <- c(20,10,5,1)
> z <- c(3.46,3.56,3.65,3.78)
> xi <- vector(mode="numeric")
> ap <- function(xi,z,beta,Delta,p=0)
+           power.midway(z,beta,Delta,xi)-p
> for (i in 1:length(z))
+ {
+     xi[i] <- uniroot(ap,interval=c(4,6),z=z[i],beta=beta,
+         Delta=delta[i],p=0.85)$root
+ }
> names(xi) <- delta
> round(xi,2)
  20   10    5    1
4.97 4.71 4.64 4.65
```

The last row gives target values for the noncentrality parameter for each of the indicated inter-marker spacings.

Determining the Sample Size

Next we turn to the determination of the sample sizes. Recall the definition of the noncentrality parameter:

$$\xi = \frac{n^{1/2}(\alpha + \delta)/2}{\sigma_y^2} \, .$$

Simple algebra can be used in order to express n in terms of the other parameters (ξ, α, δ, and σ_y):

$$n = \frac{\xi^2 \sigma_y^2}{(\alpha + \delta)^2/4} \, .$$

Thus, one can easily compute the required sample size for each combination of parameters. In order to give an example, let $\alpha = \delta = 0.5$, and $\sigma_y^2 = 1.25$. Then

```
> a <- 0.5; d <- 0.5; ss <- 1.25
> n <- xi^2*ss^2/((a+d)^2/4)
> round(n)
 20  10   5   1
154 139 135 135
```

The increase in the required sample size in going from $\Delta = 1$ to $\Delta = 20$ is not as extreme as one might have suspected.

Selecting the Design

One would like to select the best of the alternative designs. However, a design that involves the use of fewer animals will require more extensive genotyping, and vice versa. A way to evaluate the balance between these competing resources is via analysis of cost effectiveness. For the sake of the comparison we assume here that the expense per animal is \$30, including purchase, maintenance, and phenotyping. For the genotyping one must order a set of primers for each marker. The actual genotyping is then carried out for each marker, across all the animals. We take the price of the set of primers to be \$70, and the price of a genotype reaction to be \$2. This leads to the figures:

```
> Pheno <- 30; Primers <- 70; Geno <- 2;
> cost <- n*Pheno + (1600/delta)*(Primers + n*Geno)
> round(cost)
    20     10      5      1
 34925  59695 112732 549005
```

According to these calculations, if judged only by the power to detect a QTL, the sparsest design is the most efficient. In the following sections we discuss problems of estimating the location and effect of a QTL, for which denser markers promise greater accuracy.

1.4 Confidence Sets

Statistical power refers to the probability of detecting a QTL anywhere on a chromosome. It is also important to estimate the genomic location of the detected QTL as precisely as possible. A confidence interval is a range of parameter values that depend on the data and contain the unknown actual value of the parameter with high probability. In this section we will discuss briefly the issue of the construction of confidence interval for the genomic location of the QTL. We will assume that a QTL has been detected on some chromosome. We will also assume that additional markers have been typed in the region suspected to contain the QTL, so the inter-marker spacing is small, say 1 cM.

For markers in the neighborhood of a detected QTL, the values of $|Z_t|$ will be substantially higher than the values for markers in other regions of the genome. Indeed, if we seek to guess the actual location of a QTL, the marker, say $\hat{t}$, where $|Z_t|$ assumes its maximum value is a reasonable choice. Because of random fluctuations, however, this guess will probably not be exactly correct. Consequently, we may want to incorporate the possibility that the QTL is somewhere in the neighborhood of $\hat{t}$. It seems intuitively clear that if $|Z_t|$ is almost as large at nearby markers as at $\hat{t}$, then the QTL might also be close to those markers, while markers where $|Z_t|$ is much smaller are unlikely to be close to the QTL.

To make this idea more precise, we assume that there are several markers in the neighborhood of $\hat{t}$, so to a rough approximation it is reasonable to assume that the QTL coincides with some marker locus. The procedure we propose is to determine a suitable constant c^2, so that the QTL might reasonably be located at any marker locus τ (on the same chromosome as $\hat{t}$) such that

$$\max_t Z_t^2 - Z_\tau^2 < c^2 \ . \tag{1.4}$$

Suppose we can choose c^2 to have the property that

$$\mathrm{Pr}_{\tau,\xi}\big(\max_t Z_t^2 - Z_\tau^2 \geq c^2\big) \leq p \ , \tag{1.5}$$

where $\mathrm{Pr}_{\tau,\xi}$ denotes that the probability is calculated under the condition that the true QTL is at the marker locus τ, and ξ is the noncentrality parameter at that locus. The traditional values for p are 0.1, 0.05, or some other small probability. The set of all marker loci τ satisfying (1.4) where c^2 satisfies (1.5), say Q, is called a $(1-p)$-*confidence set* for the QTL. Note that Q is a random set, since it is a function of $\max_t Z_t^2$, Z_τ^2, and c^2. It can be shown to have the property that for the true QTL q itself $\mathrm{Pr}_q\{q \in Q\} \geq 1-p$, hence the name "confidence set".

Unfortunately, there is a technical complication to this approach: the probability on the left-hand side of (1.5) depends both on τ and on the value of the noncentrality parameter ξ, which is unknown. Hence we cannot compute

the left-hand side of (1.5), even in principle. However, this probability is fairly constant over a reasonably wide range of plausible values of ξ, so knowing ξ exactly is not of critical importance. One can select a representative value for the noncentrality parameter, and use that value in the computation.

Let us demonstrate the approach in an example. Consider an inter-marker spacing of 1 cM. Let us select a critical value c^2 that satisfies (1.5) for $p = 0.1$ and $\xi = 6$:

```
> markers <- 0:80
> Z <- OU.sim(beta,markers)
> q <- 40; xi <- 6
> ZZ <- (add.qtl(Z,beta,markers,q,xi))^2
> ZZ.max <- apply(ZZ,1,max)
> d <- dim(ZZ)
> ZZ.dif <- matrix(ZZ.max,nrow=d[1],ncol=d[2])-ZZ
> conf.level <- function(cc,q,ZZ.dif,cl=0)
+     mean(ZZ.dif[,q+1] < cc)-cl
> cc <- uniroot(conf.level, interval=c(0,10),
+         q=q, ZZ.dif=ZZ.dif, cl=0.9)$root
> cc
[1] 4.678781
> 1-conf.level(cc,q,ZZ.dif,cl=0)
[1] 0.1
```

Next we turn to the investigation of the change in the confidence level of the confidence set as a function of the noncentrality parameter. While we are at it, let us compute also the expected lenght of the confidence set:

```
> xi <- 3:8
> cs.length <- cs.level <- vector(mode="numeric")
> for (i in 1:length(xi))
+ {
+     ZZ <- (add.qtl(Z,beta,markers,q,xi[i]))^2
+     ZZ.max <- apply(ZZ,1,max)
+     ZZ.dif <- matrix(ZZ.max,nrow=d[1],ncol=d[2])-ZZ
+     cs.level[i] <- mean(ZZ.dif[,q+1] < cc)
+     cs.length[i] <- mean(apply(ZZ.dif < cc,1,sum))
+ }
> names(cs.level) <- xi
> cs.level
     3      4      5      6      7      8
0.8593 0.8727 0.8864 0.9000 0.9109 0.9216
> names(cs.length) <- xi
> cs.length
      3       4       5       6       7       8
31.1014 18.0115 11.2120  7.6695  5.6781  4.3922
```

It can be seen from the results of the simulations that for values of ξ ranging from 3 to 7 the probability that $\max_t Z_t^2 - Z_\tau^2 < c^2$ is not too far from its target value of 0.9. Observe, also, that the expected length of the confidence set is quickly reduced when the noncentrality parameter increases. In essence: the stronger the signal the more accurately it can be located.

Let us propose a theoretical explanation to the phenomena we have just observed. Suppose that markers are equally spaced at inter-marker distance Δ. It can be shown by an argument similar to that used to derive (1.2) that when the QTL τ coincides with a marker locus not near either end of the chromosome,

$$\Pr_{\tau,\xi}\{\max_i Z_{i\Delta}^2 - Z_\tau^2 \geq c^2\} \approx 2\,\nu\big([2\,\beta\Delta(c^2+\xi^2)]^{1/2}\big)(1+c^2/\xi^2)^{1/2}\exp\left(-c^2/2\right). \tag{1.6}$$

The probability (1.6) is determined primarily by the factor $2\exp(-c^2/2)$ and only to a lesser degree by the factors depending on ξ. For a numerical example, suppose $\Delta = 1\,\text{cM}$ and $c^2 = 4.6$, so $2\exp(-c^2/2) = 0.2$. For ξ increasing from 4 to 7, which would be reasonable values for a detectable but not overwhelming QTL, the right-hand side of (1.6) decreases from 0.134 to 0.090, so under these conditions we have approximately a 0.90 confidence set.

We can compare the simulated confidence probabilities to the approximations proposed in (1.6):

```
> Delta <- 1
> cs.level <- rbind(cs.level,1-2*Nu(sqrt(2*beta*Delta*
+                 (cc+xi^2)))*sqrt(1+cc/xi^2)*exp(-cc/2))
> rownames(cs.level) <- c("simulated","analytical")
> round(cs.level,3)
               3     4     5     6     7     8
simulated  0.859 0.873 0.886 0.900 0.911 0.922
analytical 0.848 0.873 0.891 0.904 0.915 0.924
```

Equation (1.1) tells us how fast the process Z_t is expected to drop away from its maximum as we move away from the true QTL. Hence, it also gives us a rough idea of the length of the confidence set: If we assume that ξ approximately equals $\max_t Z_t$ (cf. Prob. 1.6) and $\xi\exp(-0.02\,|\tau - t|)$ equals Z_t, then solving the equation $\xi^2 - \xi^2\exp(-0.04\,|\tau - t|) = c^2$, for $2|\tau - t|$, should approximately produce the expected length of the confidence set. Solution to the equation yields the formula $50\,\log[\xi^2/(\xi^2 - c^2)]$ cM as a rough approximation of the expected length of the confidence set. Compare it to the results of the simulations:

```
> cs.length <- rbind(cs.length,50*log(xi^2/(xi^2-cc))/Delta)
> rownames(cs.length) <- c("simulated","approximation")
> round(cs.length,3)
                    3      4      5      6      7      8
```

```
simulated      31.101 18.012 11.212 7.670 5.678 4.392
approximation  36.684 17.295 10.361 6.961 5.018 3.796
```

Note that even under favorable conditions the confidence set is more than 5 cM in width, which is still sizable if measured by the number of genes it can contain. Under less favorable conditions, the interval is much wider. As we saw in Chap. **??**, the recombination fraction between two marker loci for a RI design is larger than for a BC or IC design. This means that the rate of decay of the value of Z_t for a RI design from its maximum value near a QTL is more rapid. On the one hand, this has a negative effect on the power to detect linkage, since it means that the noncentrality parameter at markers located at a relatively short distance from a QTL can be substantially smaller than at the QTL itself. On the other hand, it permits the construction of more accurate confidence regions – roughly four times as accurate for a RI strain produced by sib mating and twice as accurate for a RI strain produced by selfing. For the example discussed above with $\xi = 5$, the expected size of the confidence set based on a BC design is about 10 cM. The same argument applied to $\theta_{\text{RI}} = 4\theta/(1 + 6\theta)$ for a RI strain produced by sib mating suggests a confidence set of about 2.5 cM, which would be a substantial improvement in accuracy.

Remark 1.3. From the argument given above and from the approximation $\log[\xi^2/(\xi^2 - c^2)] = -\log(1 - c^2/\xi^2) \approx c^2/\xi^2$, we see that the length of the confidence interval is roughly inversely proportional to ξ^2. Since ξ^2 is itself proportional to the sample size n, the length of the confidence interval is inversely proportional to n. It is common in statistical problems that confidence intervals are inversely proportional to $n^{1/2}$. The difference reflects special features of gene mapping as a statistical problem. Apart from this theoretical curiosity, we have the more practical implication that the confidence interval can be made half as long by doubling the sample size, whereas one would ordinarily expect that the sample size must be quadrupled in order to halve the confidence interval. Nevertheless, it remains essentially impossible to identify a QTL uniquely by using only methods of gene mapping. We return to this issue briefly in Chap. **??**.

1.5 Confidence Bounds for the Genetic Effect of a QTL

A confidence region for the genomic location of a QTL gives us an idea of where to concentrate a systematic search for the QTL. A confidence region for the genetic effect of a QTL, as measured, for example, in a backcross by the noncentrality parameter $\xi = \mathrm{E}(Z_\tau) = n^{1/2}(\alpha+\delta)/\sigma_y$, provides information about the importance of the QTL in determining the phenotype, indicates the range of outcomes we might expect to see if we replicated the experiment, and provides a basis for comparing the QTL in different strains of mice. If we could observe Z_τ itself (and were aware that the marker we are observing is indeed

the QTL τ), it would be a simple matter to use the normal distribution to find an approximate confidence interval for ξ. However, since we must search for τ, the surrogate observation $\max_t Z_t$ will often not equal Z_τ, especially if the spacing between markers is small, so a large value at one marker may be accompanied by a large value at nearby markers. When τ itself is a marker or close to a marker, $\max_t Z_t$ may exceed Z_τ, hence giving a biased estimator of ξ. (See Prob. 1.6 for an exploration of this issue by simulations.)

To deal with this problem we find in this section a lower confidence bound for ξ, which deals with the problem of multiple comparisons and indicates that the genetic effect is at least of some minimal size. The same methods can be used to find an upper confidence bound (hence also a confidence interval), but in view of the applications and the problem of bias, the lower bound seems scientifically more interesting.

Consider $\max_t Z_t$, where the maximum is taken over all chromosomes and all marker loci. In testing the null hypothesis of no genetic effect we ask whether the observed value of $\max_t Z_t$ is incompatible with the null hypothesis in the sense that the probability when the null hypothesis is true of a still larger value is so small that it renders the null hypothesis untenable. A lower confidence bound for ξ is found by asking a very similar question in a slightly different way. We consider a trial value of ξ, say ξ_0. In the test of hypothesis this was 0. Now, instead of asking whether ξ_0 is unreasonable, we ask how large ξ_0 must be so that is at least minimally reasonable. This can be formalized by regarding the observed maximum $z_{\max} = \max_t Z_t$ as a constant, then finding the value of ξ_0 that satisfies

$$\Pr_{\xi_0}\big(\max_t Z_t \geq z_{\max}\big) = p \tag{1.7}$$

(cf. (6.5)), where p is some (usually small) probability. The value of ξ_0, which is a function of $z_{\max}$, say $\xi_0(z_{\max})$, is called a $(1-p)$-*lower confidence bound* for ξ, because it can be shown to satisfy

$$\Pr\big(\xi_0(\max_t Z_t) \leq \xi\big) = 1 - p\,. \tag{1.8}$$

In particular, if $\max_t Z_t$ were to equal exactly the 0.05 significance threshold, then the 0.95 lower confidence bound would be $\xi_0 = 0$. The case $p = 0.5$ is also of interest. In this case the confidence bound can be regarded as a point estimator of ξ, called a *median unbiased estimator* because by (1.8) the median of its cumulative distribution function is ξ. As a point estimator it does not have the bias inherent in using $\max_t Z_t$.

To compute the probability (1.7), we assume that there is only one QTL contributing to the trait. The complementary event, $\{\max_t Z_t < z_{\max}\}$, consists of the intersection of the events that the indicated inequality holds for all t on the chromosome where the QTL is located and on all other (unlinked) chromosomes. These two events are independent, so the probability of their intersection is the product of their probabilities. Moreover, the first probability is just one minus the power, which is given approximately in (1.2) in the

special case that the QTL coincides with a marker locus. Let $Q_1(z, \xi_0)$ denote the power. The second probability is just one minus the type I error probability associated with a genome containing one less chromosome (namely the linked chromosome), which is given approximately by (**??**). If we denote the type I error probability by $Q_0(z, C, L)$, then the probability on the left-hand side of (1.7) is

$$Q_1(z_{\max}, \xi_0) + Q_0(z_{\max}, C, L) - Q_1(z_{\max}, \xi_0) Q_0(z_{\max}, C, L) \ ,$$

which can be used, as indicated above, to give a confidence bound for ξ.

For a numerical example, suppose that $z_{\max} = \max_t Z_t = 3.9$ with a marker spacing of $\Delta = 1$ cM. Recall that the 0.05 significance level with this marker spacing was about 3.78, so the assumed value of z is only marginally significant. If this were the outcome of directly observing Z_τ, which is normally distributed with mean ξ and variance one, a point estimator of ξ would be 3.9 itself while a 0.95-lower confidence bound would be $3.9 - 1.65 = 2.25$. The method described above to account for the genome scan gives a 0.95-lower confidence bound of 1.35 if the QTL is assumed to coincide with a marker and we use the approximations (1.2) and (**??**) for Q_1 and Q_2. If we assume that the QTL is midway between markers, we require a more complicated calculation of $Q_1(z, \xi)$ (using, for example, the program given at the end of Sect. 1.2). The result would be a lower confidence bound of 1.45. Assuming the QTL is located exactly at a marker gives the most conservative result (smallest lower confidence bound).

The 0.5-confidence bound, i.e., the median unbiased estimator for the hypothetical data of the preceding paragraph when the QTL coincides with a marker locus is 3.57, somewhat less than the naïve biased estimate of 3.9.

Consider now the example from Sen and Churchill [**?**] mentioned in the preceding chapter, where $\Delta \approx 10$ cM. For the value $z_{\max} = 5.8$ on chromosome 4, the 95% lower confidence bound would be approximately 4.10. For the value $z_{\max} = 3.7$ on chromosome 15 it would be approximately 1.08. In the first case the confidence bound is very close to the naïve lower confidence bound of $z_{\max} - 1.645 \approx 4.15$; but in the second case the problem of bias is more serious and hence the difference from the naïve confidence bound is substantial.

Remark 1.4. The approximation for $Q_1(z, \xi)$ given in (1.2), which involves division by ξ, is a poor approximation for very small values of ξ. Since the lower confidence bound involves evaluation of $Q_1(z, \xi)$, for small ξ, before putting too much faith in the approximation, it is advisable to check its accuracy by simulation. (Consider Prob. 1.7 and see also Fig. 1.3.)

1.6 Bibliographical Remarks

The analytic approximations to the power are found in Feingold, Brown and Siegmund [**?**]. The discussion of confidence regions is taken from [**?**] and [**?**].

Problems

1.1. The two parameters that determine the power for a given cross and a given inter-marker spacing are the proximity of the QTL to the markers and the noncentrality parameter (ξ).

(a) Use the programs in the text to simulate the power function, for the backcross design with 10 cM, 5 cM, and 1 cM inter-marker spacings, over a grid of values of the noncentrality parameter.
(b) Plot the power function and interpret the resulting plots.

1.2. Given a budget of \$50,000 and other costs as described in the text, what would be the smallest effect that can still be detected with a power of 85% for the backcross design?

1.3. Write a function that simulates the test statistic for the intercross when the QTL is present. Use this function in order to examine the power function for detecting a QTL using this design. Consider both the cases that the trait is additive ($\delta = 0$) and the trait is either dominant or recessive ($\delta = \pm\alpha$). (Remember that the test statistic for an IC has two degrees of freedom and hence requires a different threshold from a BC – Probs. **??** and **??**.) Assuming σ_e^2 is the same for an IC and for a BC, discuss the comparative advantages and disadvantages of a BC and an IC design.

1.4. In this problem we use the notation of Prob. **??**. A mathematical approximation for the power of an intercross design for a QTL at 0 recombination distance from the nearest marker is given by

$$\Pr\big(\max_i U_{i\Delta} \geq u\big) \approx$$
$$1 - \Phi\big(u^{1/2} - \xi\big) + \phi\big(u^{1/2} - \xi\big)\Big[\frac{1}{2\,\xi} + \Big(\frac{u^{1/2}}{\xi}\Big)^{1/2}\Big\{\frac{2\,\nu}{\xi} - \frac{\nu^2}{u^{1/2} + \xi}\Big\}\Big] \,. \quad (1.9)$$

In this equation $\xi = (\xi_1^2 + \xi_2^2)^{1/2}$, where $\xi_1 = (n/2)^{1/2}\alpha/\sigma_y$ and $\xi_2 = n^{1/2}\delta^2/2\,\sigma_y$ are noncentrality parameters for the additive and dominance effects, respectively, and $\nu = \nu\big((2\bar{\beta}\Delta u)^{1/2}\big)$, where $\bar{\beta} = (\xi_1^2\beta_1 + \xi_2^2\beta_2)/\xi^2$. Compare the numerical results you obtained by simulation in the preceding problem with those obtained from this approximation.

1.5. Use the approximate power formula given in (1.3) to compare the power of a BC and a RI design based on repeated sib mating. Assume that the trait is additive and that σ_e^2 is the same for both designs. (This latter assumption may not be satisfied if there is more than one major QTL. See Prob. **??**.) Recall that the appropriate value of β for the RI design is 0.08 (Prob. **??**), so the RI design will have both a higher threshold and a steeper drop-off in power when markers are relatively widely spaced and the QTL is located between markers.

1.6. A parameter of interest is the locus-specific heritability, defined in a backcross by $h^2 = (\alpha + \delta)^2/4\,\sigma_y^2$. This parameter gives the proportion of the total phenotypic variance that is attributable to the QTL. Suppose we would like to use the data from a genome scan to get some idea of the magnitude of h. We found in Chap. **??** that $\mathrm{E}(Z_\tau) \approx n^{1/2}h$. Hence if we had a single marker at the QTL itself, we could use our test statistic to obtain an unbiased estimate of h. In a genome scan we could use the statistic $\max_t Z_t/n^{1/2}$, as a surrogate for $Z_\tau/n^{1/2}$. Simulate the expected value of this statistic for different values of h ranging from 3.0 to 6.0 and Δ ranging from 1 to 10 cM, and for different positions of τ with respect to the nearest flanking markers. Does it seem reasonable from your simulations to conclude that $\max_t Z_t/n^{1/2}$ provides an acceptable estimator of h? Discuss.

1.7. Calculate lower confidence bounds for values of $\max_t Z_t$ ranging from 3.85 to 6.0. How does the "bias adjustment" for multiple comparisons change as $\max_t Z_t$ increases? Use simulations to approximate $Q_1(z, \xi)$ for some of the smaller values of ξ.

1.8. A strategy to minimize the cost of genotyping, while maximizing the amount of marker information in the neighborhood of a QTL, is to select markers to genotype in two stages. In the first stage one uses fairly widely spaced markers, say at 20 or 40 cM intervals. Then in regions where there is some evidence of linkage, say a Z value in the range of 1 to 2, one adds more markers. For a specific example, suppose that markers are originally genotyped at 20 cM intervals. If $Z_t > z_1$, where z_1 is a parameter to be chosen, then in each of the 20 cM intervals flanking t, additional markers are placed at 5 cM intervals. Linkage is detected if for the final collection of markers, $\max_t Z_t > z_2$. Simulate this experiment with different values of z_1, and determine z_2 so that the significance level is 0.05. Find the power of this procedure for several values of the noncentrality parameter. How do the significance threshold and power compare with the case of a single stage of 5 cM genome scan? On average, how many fewer markers are genotyped per mouse? Repeat the same experiment with the modified first stage rule: if for two consecutive markers $(Z_t + Z_{t+\Delta})/2 > z_3$, then add new markers between the loci t and $t+\Delta$. Does one of these first stage procedures seem much better than the other?