Title
A method for estimating the effective number of loci affecting a quantitative character

Journal
Theoretical Population Biology, 89(C)

ISSN
0040-5809

Author
Slatkin, Montgomery

Publication Date
2013-11-01

DOI
10.1016/j.tpb.2013.08.002

Peer reviewed

Published in final edited form as:
Theor Popul Biol. 2013 November ; 0: 44–54. doi:10.1016/j.tpb.2013.08.002.

A method for estimating the effective number of loci affecting a quantitative character

Montgomery Slatkin[1]
[1] Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA

Abstract

A likelihood method is introduced that jointly estimates the number of loci and the additive effect of alleles that account for the genetic variance of a normally distributed quantitative character in a randomly mating population. The method assumes that measurements of the character are available from one or both parents and an arbitrary number of full siblings. The method uses the fact, first recognized by Karl Pearson in 1904, that the variance of a character among offspring depends on both the parental phenotypes and on the number of loci. Simulations show that the method performs well provided that data from a sufficient number of families (on the order of thousands) are available. This method assumes that loci are in Hardy-Weinberg and linkage equilibrium but does not assume anything about linkage relationships. It performs equally well if all loci are on the same non-recombining chromosome provided they are in linkage equilibrium. The method can be adapted to take account of loci already identified as being associated with the character of interest. In that case, the method estimates the number of loci not already known to be affect the character. The method applied to measurements of crown-rump length in 281 family trios in a captive colony of African green monkeys (*Chlorocebus aethiopus sabaeus*) estimates the number of loci to be 112 and the additive effect to be 0.26 cm. A parametric bootstrap analysis shows that a rough confidence interval has a lower bound of 14 loci.

Keywords

Wright-Castle method; quantitative genetics

The number of loci responsible for the variance of a quantitative character is an important part of its genetic architecture and affects the character's potential for short term and long term evolution. Existing methods for estimating the number of loci are of two types which estimate somewhat different quantities. The first type assumes that two populations that differ in the character mean are hybridized to form an F_1 population. The variance in the F_2 and backcross populations depend on the minimum number of loci responsible for the difference in the population mean. This idea was first presented by Castle (1921), using formulas derived by Sewall Wright. The method as originally proposed, often called the Wright-Castle method, assumes completely inbred lines that are homozygous at all loci. Lande (1981) showed that the assumption of complete homozygosity in the parental populations was unnecessary and that essentially the same method can be applied to

Corresponding author: Montgomery Slatkin Department of Integrative Biology University of California Berkeley, CA 94720-3140 slatkin@berkeley.edu.

populations that are initially outbred. Therefore, the Wright-Castle method is more generally applicable, and the method has become widely used. Lande noted that his method can be adapted to estimating the number of loci in a single population if divergent populations are derived from that population by the application of directional selection.

I will call the Wright-Castle method applied to outbred populations the Wright-Castle-Lande (WCL) method. Several authors further developed and tested this method (Cockerham, 1986; Otto and Jones, 2000; Zeng, 1992). As Lande emphasized, the WCL method estimates only the minimum number of loci accounting for differences in population means. If two lines carry alleles that have effects discordant with the average difference between populations (e. g. if the allele that reduces the mean happens to be in higher frequency in the population that has the larger mean), then the actual number of loci affecting the difference in the mean will be necessarily larger than the estimated number (Lande, 1981).

In contrast to the WCL method, the second type of method estimates the number of loci affecting the variance of a character in a single population. These methods are less widely known and less commonly used, in part because they are not as well developed theoretically. Pearson (1904a; 1904b) was the first to point out that, under the assumptions of Mendelism (at that time newly rediscovered), the variance of a quantitative character in offspring depends on the average of the character in the parents, i. e. the midparent value. Pearson presented this result to demonstrate the incompatibility of Mendelism with the biometrical theory of inheritance, in which the variance is independent of the midparent value. The intuition behind Pearson's result is simple. If the midparent value is near the limits of the range of variation of the character, the parents are probably homozygous for alleles that affect the character in the same way. Consequently, there will be less genetic variance among their offspring than there will be if the midparent value is near the middle of the character's range.

Although Pearson's result was known, it appears not to have been used as a way to estimate the number of loci affecting a quantitative character until 1969 (Penrose, 1969). Penrose showed that the difference in correlation structure between a parent and its offspring and between a pair of full siblings could yield an estimator of the number of loci responsible for the variance in a randomly mating population. From the data presented by Pearson and Lee (1903), Penrose inferred that human stature was governed by six loci but with a large error in the estimate. Penrose also analyzed data on finger ridge counts and concluded that the number of loci is indefinitely large. Stark (1976) and Fain (1976; 1978) elaborated on Penrose's derivation. Subsequent development of theory of this type focused on the problem of detecting the presence of a single major gene, e. g. (Ott, 1979).

Felsenstein recognized that the dependence of the offspring variance on the midparent could be used directly to estimate the number of loci and lectured on that topic at the 13th International Congress of Genetics in 1973. Although the abstract of that presentation was published (Felsenstein, 1973), he did not publish a paper describing his method (J. Felsenstein, personal communication). Karlin, Carmelli and their collaborators followed an independent but related line of research based on indices designed to distinguish between major-gene and multifactorial models (Karlin and Carmelli, 1978; Karlin et al., 1979; Karlin et al., 1983; Karlin et al., 1981)

In this paper, I introduce another method for estimating the number of loci. It is of the same type as those of Penrose and Felsenstein, but instead of using the relationship of the midparent to the offspring variance, it uses the dependence of the offspring variance on the parental phenotypes separately. That modification leads to a simple calculation of the approximate joint likelihood of the number of loci and the average additive effect of an

allelic substitution, from which maximum likelihood estimates (MLEs) of both parameters can be obtained. This method does not assume that loci are unlinked, only that they are in linkage equilibrium. The same estimate of the number of loci is obtained even if all loci are in linkage equilibrium on a single non-recombining chromosome. When allele frequencies and additive effects vary among loci, effective numbers of loci and effective additive effects can be defined. Furthermore, if some loci are already known to affect the character and the genotypes of those loci can be determined in parents and offspring, the method can be adapted to obtain the MLEs of the number and additive effect of the remaining unknown loci.

The performance of this method depends on the heritability of the character. Even the heritability is 50% and 10 loci affect the character, the sample sizes needed to obtain accurate estimates are large, on the order of thousands of parents and offspring. And the sample size needed to obtain accurate estimates increases with the number of loci affecting the character. Although it is unlikely that such large data sets would be obtained for the purpose of estimating the number of loci, such large data sets have been collected for other purposes, particularly for humans and colonies of non-human primates. To illustrate the use of this method, it is applied to data from African green monkeys.

Symmetric additive model

Assume that a quantitative character is affected by n loci and an independent environmental component. At each locus, there are two alleles denoted by + and – and the additive effect of each + allele is a. The phenotype x of an individual is

$$x=ia+e \quad (1)$$

where i is the number of + alleles and e is a normally distributed random environmental component which has mean 0 and variance σ_e^2. At each locus, the frequency of the + allele is p. Assuming both Hardy-Weinberg and linkage equilibrium, the distribution of i in a population is binomial with probability p and sample size $2n$. The population mean is $\bar{x}=2npa$, the variance is $\sigma_x^2=2np(1-p)a^2+\sigma_e^2$ and the heritability is $h^2=2np(1-p)a^2/\left[2np(1-p)a^2+\sigma_e^2\right]$.

Assume that a pair of parents with phenotypes, x_1 and x_2, have S offspring. The set of phenotypes of the offspring is represented by a vector $\mathbf{y} = \{y_1,..., y_S\}$. It is shown in Appendix 1 that, under these assumptions, the expectation, variance and covariance of the y's, given x_1 and x_2, can be calculated. For offspring j

$$E(y_j|x_1,x_2)=h^2\frac{x_1+x_2}{2}+\left(1-h^2\right)\bar{x}, \quad (2)$$

$$Var(y_j|x_1,x_2)=\frac{1}{2}\frac{2n-2}{2n-1}h^2\sigma_e^2+\frac{1}{4(2n-1)}\left[\tilde{x}_1(2na-\tilde{x}_1)+\tilde{x}_2(2na-\tilde{x}_2)\right]+\sigma_e^2, \quad (3)$$

where

$$\tilde{x}_k=h^2x_k+\left(1-h^2\right)\bar{x} \quad (4)$$

(k=1, 2), and for full siblings j and j'

$$Cov\left(y_j, y_{j'}|x_1, x_2\right) = \frac{h^2\sigma_e^2}{2} \quad j \neq j'. \quad (5)$$

Equation (2) is the standard equation for the regression of the offspring on the midparent. Equation (3) confirms the results of Pearson (1904a; 1904b) and Felsenstein (1973) and is the basis for the method developed in the next section. Note that the covariance, unlike the variance, is independent of n. That is true because there is no dominance in the symmetric additive model. With dominance, the covariance between full siblings will depends on n.

Approximate likelihood estimation

Suppose that the data consists of F families. The first step is to estimate from the data the total variance $\left(\sigma_x^2\right)$, the heritability (h^2), and the environmental component of the variance $\left(\sigma_e^2\right)$. Then, the distribution of **y** is assumed to be multivariate normal with the mean value of each y_j given by Eq. (2), and the variance-covariance matrix is given by Eqs. (3)-(5). With that assumption and the estimated values of h^2 and σ_e^2, the likelihood of n and a for each family is given by the probability of obtaining the offspring vector (**y**) from that multivariate normal distribution. Assuming independence of families, the log-likelihoods for each family are added and the joint MLE of n and a can be obtained, $\hat{n}$ and $\hat{a}$. In the program written to implement this method, a grid search was used to obtain the MLE.

This method gives only the approximate likelihood. A full likelihood method would jointly estimate h^2, σ_e^2, n and a from all the data. The approach taken here was chosen because its computational simplicity.

Effective number of loci

In a more realistic model, the allele frequencies and additive effects at each locus will not be equal and there may be some dominance. In that case, it is possible to define an effective number of loci, n_E, and an effective additive effect, a_E, for an equivalent symmetric additive model:

$$n_E = \frac{1}{2}\left(1 + h^4 / \sum_{j=1}^{n} h_j^4\right) \quad (6)$$

and

$$a_E = \sqrt{\sum_{j=1}^{n} p_j\left(1-p_j\right)\left(a_j - \left(2p_j - 1\right)d_j\right)^2 / \left(2n_E\bar{p}\right)} \quad (7)$$

where $\bar{p}$ is the average frequency of + alleles across loci and $h_j^2 = 2p_j\left(1-p_j\right)\left(a_j - \left(2p_j-1\right)d_j\right)^2/\sigma_x^2$ is the heritability attributable to locus j. Equation (6) differs from the effective number of loci defined by Lande (1981), $n_E^{\text{Lande}} = n\left(\sum_{j=1}^{n}\sigma_j^2\right)^2 / \sum_{j=1}^{n}\sigma_j^4$ (p. 547), where σ_j^2 is the contribution of locus j to the additive genetic variance. Lande's n_E involves the square of the variances and is necessarily less than n, while n_E defined by Eq. (6) may exceed n.

Simulation tests

Symmetric model

In order to test the performance of this method, I simulated data, first under the symmetric model and then under more general assumptions. The simulation program assumes n biallelic loci. At locus j, the frequency of + is p_j, the additive effect of a + allele is a_k, and the dominance effect is d_j. In each family, the genotypes of the two parents are randomly generated under the assumption of Hardy-Weinberg and linkage equilibrium, and then the phenotype of each parent is determined by adding a random environmental component with mean 0 and variance σ_e^2. In the case of no linkage, the genotypes of haploid gametes produced by each parent are generated according to Mendel's first law applied to each locus independently. The gametes are combined into the offspring and a random environmental component is added to obtain the offspring phenotype. This process is repeated until a specified number F of families are obtained. In the case of complete linkage, one of the two chromosomes was chosen at random to be transmitted to each offspring.

For each simulated data set, the phenotypic variance $\left(\sigma_x^2\right)$ is computed and the heritability (h^2) and environmental component of the variance $\left(\sigma_e^2\right)$ are estimated by regressing the offspring on the midparent. Then the dataset, along with the estimated values of h^2 and σ_e^2, is passed to a program that calculates the log-likelihood as described above for each family and adds log-likelihoods across families for pairs of values of n and a. If, for a family, x_1, x_2, n and a, result in either the computed variance (Eq. 3) being negative or the computed covariance (Eq. 5) exceeding the computed variance (thus causing the variance-covariance matrix to not be positive definite), the log-likelihood of a and n for that family is set to $-\infty$. Hence a single such family will cause the likelihood of that n and a combination to be 0. The MLEs of n and a, $\hat{n}$ and $\hat{a}$, are obtained by performing a grid search.

The first set of simulations assumes the model under which the method was derived: each locus had the same values of a and p. The joint log-likelihood surface for one replicate is shown in Figure 1. The results shown are based on simulations of n=10 loci of additive effect a=1. At each locus, the + allele had a frequency p=0.5. The value of σ_e^2 was set to 5 which ensured that h^2 is approximately 0.5 in the simulated data. The simulated data set contained F=10,000 parent-offspring trios. From this data set, a grid search with limits 2 and 50 for n with a step size of 1 and 0.5 and 2.0 with a step size of 0.03 for a found $\hat{n}$=9 and $\hat{a}$=1.04. The shape of the log-likelihood surface is typical of other cases. There is a ridge that follows a hyperbola in the n-a plane. The hyperbola reflects the fact that in Equation (3) the variance depends primarily on the product na and only slightly on n separately.

The performance of the method, even with data simulated under the symmetric model, depends on both the sample size and the heritability. Figure 2 shows the results of applying the method to 1000 replicates with different numbers of family trios, F, in which $h^2 \approx 1/2$. The range of estimates decreases as F increases. With F=100,000, the range is small, indicating that the approximate likelihood method is roughly consistent. The performance depends on the heritability. For example in the replicates that generated the histograms in Figure 2 for F=10,000, the average $\hat{n}$ in 1000 replicates was 13.9 and the standard deviation (sd) was 11.3. In a set of 1000 replicates with the same parameter values except $h^2 \approx 0.25$, the average was 27.1 with sd=29.1. With $h^2 \approx 0.75$, the average was 11.0 with sd=2.4.

For a given F, the accuracy of the method decreases as the true number of loci simulated increases. For example, in 1000 replicates of the symmetric model with n=20, a=1, p=0.5 and F=10,000, the average and standard deviation of $\hat{n}$ were 36.4 and 37.6. If n=30 and the

other parameters remained the same, the average and standard deviation of $\hat{n}$ were 67.1 and 80.3. The upwards bias in the mean results partly from the fact that for some combinations of x_1, x_2 and y, the likelihood is 0 because the variance computed from Eq. (3) is negative for small n. Hence the lower bound of n is restricted.

If the same total number of offspring are measured, the method performs slightly better when there are more families instead of more siblings per family. The result is illustrated in Table 1, in which a total of 10,000 offspring were measured in each of 1000 replicate simulations. F is the number of families and S is the number of siblings per family, adjusted so that FS=10,000. There is slightly more bias in the mean and a slightly higher sd as S increases. The reason for this trend is that there is some covariance between the phenotypes of siblings, given the parental phenotypes (Eq. 5) and that covariance does not depend on n or a separately. Therefore, the covariance slightly reduces the information provided by each sibling about n and a.

Assuming complete linkage among the loci has no detectable effect on the results provided that the loci are in linkage equilibrium. Some results are summarized in Table 2.

The estimates of n and a do not depend on p because the offspring variance, conditional on the parental phenotypes (Eq. 3), does not contain p. The value of p is absorbed into the heritability. In simulations, however, p plays an important role. Information about the number of loci comes from differences in the offspring variance of parents with different combinations of phenotypes. For the method to work, it is necessary to have most of the range of parental phenotypes represented. Otherwise, there is insufficient variation among families for the dependence of the offspring variance on n to be detectable. Assuming p=0.5 in the symmetric model ensures that the full range of parental phenotypes will be generated. If instead p=0.2 and n=10, then the probability that any individual will have 10 or more + alleles is only 5.6×10^{-4}. Even with 10^4 family trios, no sensible estimates of n and a can be obtained if p=0.2 is used to generate the data. There is then an important limitation of the method. It is expected to perform well only if the phenotypes of the parents in a sample represent the full or nearly full range possible values.

Nonsymmetric models

If the assumptions of the symmetric additive model are not satisfied, the likelihood method can still be used to estimate n and a, but the values of $\hat{n}$ and $\hat{a}$ obtained no longer directly correspond to parameters of the simulation model. The symmetric simulation model was modified to allow the allele frequency (p_j), additive effect (a_j) and dominance effect (d_j) for each locus to be specified. If F=100,000 families are simulated, then variation in the estimates of n and a is comparable to what was found in the symmetric model (Fig. 2). I will concentrate on those results in order to focus on what the method estimates when sampling error is relatively small. The magnitude of the sampling error when fewer families are simulated is also comparable to what was found in the symmetric model.

When more variation among loci is allowed, there is more variation in the estimated parameter values even when large numbers of families are simulated. Some representative results are shown in Figures 3 and 4. In Fig. 3, only the allele frequencies varied among loci, while in Fig. 4 the additive and dominance effects were allowed to vary also. Even with very large sample sizes (F=100,000) there is considerable variation especially in $\hat{n}$. When such variation is allowed, n_E does not predict the results very well. a_E is somewhat better as a predictor. Figure 5 shows some typical results for two sets of replicates. For this figure, F=1,000,000 families were simulated to further reduce the effects of sampling. If alleles are additive in their effects (parts A and B), n_E is a good predictor of $\hat{n}$ in most cases, but the relationship is worse when the d_j are allowed to vary across loci also (parts C and D).

Application to data from African green monkeys

To illustrate the use of the method introduced in the previous sections, it were applied to measurements of crown-rump length in a captive colony of Caribbean-origin, African green monkeys (*Chlorocebus aethiopus sabaeus*). The measurements were made in the Vervet Research Colony at Wake Forest School of Medicine (Jasinska et al., 2012). All procedures resulting in the nonhuman primate data followed USDA and NIH guidelines and were approved by the Wake Forest School of Medicine IACUC. Wake Forest is fully accredited by AAALAC. The data were kindly provided by Dr. Matthew J. Jorgensen of Wake Forest University.

The data set consisted of measurements of crown-rump lengths of adults of 281 parent-offspring trios. When more than one measurement was made on the same individual, the average for that individual was used. Crown-rump length is sexually dimorphic. The averages for females and males are 44.19 and 50.05 cm respectively. To equalize the measurements of the two sexes, measurements in females were multiplied by the ratio, 1.132. This adjustment resulted in a roughly normal distribution of measurements in the two sexes combined (Figure 6A). The total variance is $\sigma_x^2 = 2.29 \quad \mathrm{cm}^2$, the heritability is h^2=0.725 (Fig. 6B) and the estimated environmental component of the variance is 0.553 cm^2. With these parameter values, the MLE estimates are $\hat{n}$=112 and $\hat{a}$=0.26 cm. The likelihood surface is shown in Figure 6C.

To find a rough lower bound on the estimated number of loci, I used a parametric bootstrap analysis (Efron and Tibshirani, 1993). The parametric bootstrap method seems preferable to using the asymptotic properties of the log-likelihood surface because the sample size is not large enough for the likelihood surface to be approximately bivariate normal in the neighborhood of the MLEs. I simulated 100,000 replicates of the symmetric model with n=112, a=0.26 and p=0.5. Figure 6D shows the empirical cumulative distribution of $\hat{n}$ which has a 5% quantile for n of 14, indicated by the straight line on the graph. The data are roughly consistent with variation in crown-rump in this population being governed by at least 14 loci and probably many more.

Extensions of the basic method

Data from a single parent

The above theory can be modified if only one parent per family is available. If the other parent is assumed to be drawn at random from the same population (i. e. mating is random with respect to the character of interest), then Eqs. (2) and (3) are replaced by

$$E(y|x_1) = h^2 \frac{\tilde{x}_1}{2} + \left(1 - \frac{h^2}{2}\right)\overline{x} \quad (8)$$

and

$$Var(y|x_1) = \frac{1}{2}\frac{2n-1}{2n-2}h^2\sigma_e^2 + \sigma_e^2 + \frac{\overline{x}(2na - \overline{x})}{4(2n-1)} + \frac{\tilde{x}_1(2na - \tilde{x}_1)}{4(2n-1)} \quad (9)$$

where $\tilde{x_1}$ is defined by Eq. (4). Because the covariance is created by the transmission from only one parent. Eq. (5) is replaced by

$$Cov(y_1, y_2|x_1) = \frac{h^2\sigma_e^2}{4}. \quad (10)$$

Therefore, the approximate likelihood can be obtained for families in which the phenotype of only a single parent is available.

Two populations

Assume that the symmetric model is correct for each of two populations but that the populations differ in the frequency of the + allele at each locus affecting the character. The derivation of the mean and variance, given the parental phenotypes (Eqs. 3 and 4) is still valid but with a modification of Equation (4). Assume that parent 1 is from population 1 and parent 2 from population 2. Equation (4) is replaced by

$$\tilde{x}_k = h_k^2 x_k + \left(1 - h^2\right) \bar{x}_k \quad (11)$$

where h_k^2 and $\bar{x_k}$ are the heritability and character mean in population k.

When considering individuals whose parents come from different populations, it would probably be better to analyze data from each parent separately, because the genetic variance in the two populations may be attributable to different numbers of loci. Either different loci may be polymorphic in the two populations or the same loci are polymorphic but with different allele frequencies.

Although the breeding design is the same as that of Lande (1981), this method uses patterns of variation in the F_1 population to estimate the number of loci responsible for variation in each of the parental populations. In contrast, Lande's method uses the variance in the F_2 and backcross populations to estimate the number of loci responsible for the difference in means of the parental populations. Because what is estimated by the two methods is not the same, a comparison of their results may indicate whether the loci that are responsible for variation within each population are or are not responsible for differences between their means.

Sex linkage

Another possible use of the single-parent theory is to test for sex linkage of loci affecting the character. Assume that males are the heterogametic sex and that n_A loci are on autosomes and n_X loci are on the X. The variance of the character in male offspring will be given by Eq. (9) with $n=n_A$. The X-linked loci are not transmitted to sons. The variance in female offspring is given by Eq. (9) with $n=n_A+n_X$. It is possible in principle, then, to jointly estimate n_A and n_X and to determine whether the average additive effects on autosomal and X-linked loci are different. One can anticipate that very large sample sizes will be needed to obtain accurate estimates.

Genotyping of some loci

The approximate likelihood method developed in the previous sections can be adapted to analyzing a data set in which some loci that are known to affect the quantitative character have been genotyped. I will call these loci the *known loci* and assume that their effect on the character has been estimated accurately and that the known loci have been genotyped in both parents and offspring. The known loci are assumed to be in linkage equilibrium with one another and with the remaining *unknown loci* that also affect the character.

Assume there is no interaction among any of the loci, known and unknown, but there may be dominance at the known loci. For any individual let the genotype of the known loci be denoted by $\mathbf{k} = \{k_1, k_2, ...\}$ where k_j is the number of + alleles at locus j. Let $G(\mathbf{k})$ denote the average phenotype of individuals with genotype $\mathbf{k}$ at the known loci.

Considering only the case of two parents and one offspring, assume the data consist of F sets of trios of phenotypes and net genetic effects for each of F families $[x_1, G(\mathbf{k}_1), x_2, G(\mathbf{k}_2), y, G(\mathbf{k}_y)]$ where $\mathbf{k}_1$, $\mathbf{k}_2$ and $\mathbf{k}_y$ are the genotypes of the two parents and their offspring. If we assume that the symmetric model applies to the unknown loci, then Appendix 2 shows it is possible to compute the approximate likelihood of the data as a function of n, the number of unknown loci and a the additive effect of a + allele at each unknown locus. The method is nearly the same as the one described above but with two differences. First, the net effect of the known loci has to be subtracted from each phenotype: $x_1 - G(\mathbf{k}_1)$, $x_2 - G(\mathbf{k}_2)$ and $y - G(\mathbf{k}_y)$. Second, the heritability, h^2, has to be replaced by

$$\eta^2 = \frac{\sigma^2_{A,\ \text{unknown}}}{\sigma^2_{A,\ \text{unknown}} + \sigma^2_e}, \quad (12)$$

where $\sigma^2_{A,\ \text{unknown}}$ is the additive genetic variance attributable to the unknown loci, a quantity that can be estimated from the data, as described in Appendix 2.

Discussion and Conclusions

The theory and results presented above show that it is possible to jointly estimate the number of loci and the additive effect of an allelic substitution for a normally distributed quantitative character in a randomly mating population. Although large sample sizes are needed to obtain accurate estimates, data sets with such large sample sizes are sometimes collected from humans and other species. The need for large sample sizes is not surprising given the subtle way that offspring variance depends on the number of loci.

The approximate likelihood method relies on several simplifying assumptions. It estimates n and a under a symmetric model that assumes equal allele frequency, equal additive effects and no dominance. When those assumptions are violated, it is possible to define an effective number of loci, n_E, and an effective additive effect, a_E, that reflect somewhat but not perfectly what the method would estimate if sample sizes were very large. This limitation is similar to that for the Wright-Castle-Lande method. Although Lande (1981) defined an effective number of loci estimated by that method, later work by Zeng (1992) and by Otto and Jones (2000) have shown that the estimates are sensitive to deviations from the symmetric model used in their derivation. In the method presented here, the effective additive effect, a_E, is more robust to deviations from the symmetric model than is the effective number of loci, n_E.

The method does not require the assumption that loci be unlinked, only that they be in linkage equilibrium. The assumption of linkage equilibrium is not very restrictive for outbreeding species like humans that have large genomes, but would likely not be valid for species with high selfing rates because of the extensive linkage disequilibrium found in such species.

An important assumption of the method is that environmental effects on the character are independent in each individual. That would not be true if there is a correlation created by an environmental factor shared by family members. It is difficult to intuit the effect of allowing for such correlations, and the analytic theory presented in Appendix 1 is no longer valid because the parental phenotypes are no longer independent of each other. A few sets of simulations of the symmetric model in which a common environmental effect was added to the phenotype showed that $\hat{n}$ tended to be smaller and $\hat{a}$ tended to be larger than their true values, but the effect was weak unless the environmental correlations were large.

Estimates of the number of loci affecting a quantitative character, obtained either from the Wright-Castle-Lande method or the approximate likelihood method presented here, should not be taken literally because of the uncertain relationship between the underlying genetic model and the estimates obtained. Instead, those estimates should be viewed as indications of the graininess of the genetic basis of a quantitative character, i.e. the extent of deviation from the infinitesimal model in which there is a very large number of loci each of which has a vanishingly small effect on the character. The infinitesimal model has gained support recently from the results of genome-wide association studies (GWAS) that have been performed in humans. Lango Allen et al. (2010) reported that 180 SNPs are significantly associated with differences in human height. Together, the SNPs identified account for roughly 10% of the heritability of height. Yang et al. (2010) showed that by including the many SNPs near the threshold of significance more than 50% of the heritability of height could be accounted for. Thus, the infinitesimal model is consistent with data for human height. Methods for estimating the effective number of loci can be used to determine whether that model applies to other quantitative characters as well or whether there is evidence that variation in some characters is attributable to smaller number of loci.

Software Resources

A Mathematica program that implements the approximate likelihood method described in this paper is posted at http://cteg.berkeley.edu/software.html.

Acknowledgments

This research was supported by NIH Grant R01-GM40282. I thank J. Felsenstein for the references to Pearson's papers and a copy of his published abstract, and for helpful discussions of this problem. N. G. Freimer suggested the use of data from vervets, M. J. Jorgensen provided the vervet data and information about the Vervet Research Colony. The vervet colony is supported by NIH grant RR019963/OD010965 (PI Jay R. Kaplan, Wake Forest). J. Felsensetin, M. J. Jorgensen R. Lande, S. P. Otto and R. G. Shaw made numerous helpful comments on an earlier version of this paper.

Appendix 1: Conditional means, variances and covariances

Assuming the symmetric model defined in the text, a quantitative character is determined by n loci, each of which has an additive effect a. The loci are in Hardy-Weinberg and linkage equilibrium and the frequency of the + allele at each locus is p. In an individual, the total genetic contribution to the character is ia, where i is the number of + alleles. The phenotype of this individual is

$$x=ia+e \quad \text{(A1)}$$

where e is a normally distributed random variable which has mean 0 and variance σ_e^2.

Under these assumptions, i is binomially distributed in the population with probability p and sample size $2n$. Consequently the population mean and variance of the character are

$$\overline{x}=2npa \quad \text{(A2)}$$

and

$$\sigma_x^2=2np\left(1-p\right)a^2+\sigma_e^2, \quad \text{(A3)}$$

and the heritability is

$$h^2=\frac{2np\,(1-p)\,a^2}{2np\,(1-p)\,a^2+\sigma_e^2}. \quad \text{(A4)}$$

Given the measurement of a character in an individual, the conditional distribution of i is

$$\Pr\,(i|x)=\frac{\Pr\,(x|i)\,\Pr\,(i)}{\Pr\,(x)}.. \quad \text{(A5)}$$

$\Pr(x \mid i)$ is a normal distribution with mean ai and variance σ_e^2. $\Pr(i)$ is a binomial distribution that can be approximated by a normal distribution with mean $2np$ and variance $2np(1-p)$. Hence, $\Pr(i \mid x)$ can be approximated by a normal distribution for which

$$E\,(i|x)=\frac{\tilde{x}}{a} \quad \text{(A6)}$$

where

$$\tilde{x}=h^2x+\left(1-h^2\right)\overline{x}, \quad \text{(A7)}$$

and

$$E\left(i^2|x\right)=\frac{2np\,(1-p)\,\sigma_e^2}{\sigma_e^2+2np\,(1-p)\,a^2}+\frac{\tilde{x}^2}{a^2}=h^2\frac{\sigma_e^2}{a^2}+\frac{\tilde{x}^2}{a^2}. \quad \text{(A8)}$$

The quantity $\tilde{x}$ is the breeding value of an individual with phenotype x. Note that, although the derivation leading to Eq. (A8) involves p, the result does not depend on p because it is absorbed into h^2.

Given that an individual carries i + alleles, the number j of + alleles carried by a gamete it produces has a hypergeometric distribution:

$$\Pr\,(j|i)=\frac{\left(\begin{array}{c} i \\ j \end{array}\right)\left(\begin{array}{c} 2n-i \\ n-j \end{array}\right)}{\left(\begin{array}{c} 2n \\ n \end{array}\right)}, \quad \text{(A9)}$$

for which $E(j \mid i) = i / 2$ and

$$E\left(j^2|i\right)=\frac{i^2}{4}+\frac{i\,(2n-i)}{4\,(2n-1)}. \quad \text{(A10)}$$

We can now derive the mean and variance of the character, y, in the offspring of two parents, labeled 1 and 2. The phenotypes of the parents are x_1 and x_2. Let i_1 and i_2 be the numbers of + alleles in each parent and let j_1 and j_2 be the numbers of + alleles in gametes produced by each parent. By assumption $y = a(j_1 + j_2) + e$. Therefore

$$E\,(y|x_1,x_2)=a\,[E\,(j_1|x_1)+E\,(j_2|x_2)]=a\left[\frac{E\,(i_1|x_1)}{2}+\frac{E\,(i_2|x_2)}{2}\right]=\frac{\tilde{x}_1+\tilde{x}_2}{2} \quad \text{(A11)}$$

where $\tilde{x}_k = h^2 x_k + \left(1 - h^2\right)\overline{x}$ for k=1 or 2. Equation (A11), which is equivalent to Equation (2) in the text, shows that the expectation of y is given by the standard equation for the regression of the offspring phenotype on the midparent. $(x_1+x_2)/2$, with regression coefficient h^2.

To find the variance of y, we substitute from the above expressions for the first and second moments:

$$\begin{aligned} & E\left(y^2|x_1,x_2\right) \\ & \quad = a^2 E\left(j_1^2 + 2j_1 j_2 + j_2^2 | x_1, x_2\right) \\ & \quad + \sigma_e^2 = a^2\left[E\left(j_1^2|x_1\right) + 2E\left(j_1|x_1\right)E\left(j_2|x_2\right) + E\left(j_2^2|x_2\right)\right] \\ & \quad + \sigma_e^2 = \left(\frac{\tilde{x}_1 + \tilde{x}_2}{2}\right)^2 \\ & \quad + \frac{2n-2}{2(2n-1)} h^2 \sigma_e^2 \\ & \quad + \sigma_e^2 + \frac{1}{4(2n-1)}\left[\tilde{x}_1(2na - \tilde{x}_1) + \tilde{x}_2(2na - \tilde{x}_2)\right]. \end{aligned} \quad \text{(A12)}$$

from which follows the expression for $Var(y \mid x_1, x_2)$ given by Equation (3) in the text.

We can also calculate the covariance between phenotypes in full siblings. Assume that parents with phenotypes x_1 and x_2 have two offspring which have phenotypes y_1 and y_2. The expectations and variances of y_1 and y_2 are equal and given by Eqs. (2) and (3) in the main text. To find the covariance, we assume that parent k, which has i_k + alleles, produces two gametes that carry j_k and $j_k^{'}$+alleles. Because the two gametes represent independent samples from the hypergeometric distribution (Eq. A9),

$$E\left(j_k j_k^{'}\right) = E\left(j_k\right)E\left(j_k^{'}\right) = \frac{i_k^2}{4}. \quad \text{(A13)}$$

Consequently

$$E\left(y_1 y_2 | x_1, x_2\right) = a^2 E\left[\left(j_1 + j_2\right)\left(j_1^{'} + j_2^{'}\right)|x_1, x_2\right] = E\left[\frac{i_1^2}{4} + \frac{i_1 i_2}{2} + \frac{i_2^2}{4} | x_1, x_2\right] = \frac{1}{4}\left[h^2\sigma_e^2 + \tilde{x}_1^2 + 2\tilde{x}_1\tilde{x}_2 + h^2\sigma_e^2 + \tilde{x}_2^2\right], \quad \text{(A14)}$$

and

$$Cov\left(y_1, y_2 | x_1, x_2\right) = \frac{h^2\sigma_e^2}{2}. \quad \text{(A15)}$$

Appendix 2: Multiple classes of loci and the effective number of loci

Assume there are C classes of loci. All loci within a class c have the same additive effect (a_c) and the same allele frequency (p_c). There is no dominance in any class. The method described in Appendix 1 can be generalized to obtain an expression for the offspring variance, given the parental phenotypes.

For class c, the number of loci is n_c and number of + alleles is i_c. Let **i** be a vector with elements i_c. Let j_c be the number of + alleles in a randomly generated gamete and **j** be a

vector with elements j_c. Let a_c be the additive effect of a + allele in class c and **a** be a vector with elements a_c. An offspring is generated by combining two gametes and adding an environmental component:

$$y=\mathbf{a}.\left(\mathbf{j}^{(m)}+\mathbf{j}^{(f)}\right)+e \quad \text{(A16)}$$

where the . indicate the inner product of two vectors and the superscripts (m) and (f) indicate the maternal and paternal gametes.

Given the phenotype of an individual, x, the distribution of **i** can be computed from

$$\Pr\left(\mathbf{i}|x\right)=\frac{\Pr\left(x|\mathbf{i}\right)\Pr\left(\mathbf{i}\right)}{\Pr\left(x\right)}. \quad \text{(A17)}$$

Pr(x | **i**) is a normal distribution with mean **a.i**. Pr(**i**) is, under the assumption of linkage equilibrium among the loci, the product of independent binomial distributions with sample size $2n_c$ and probability p_c. Each of these binomial distributions is approximated by a normal distribution with mean $2n_cp_c$ and variance $2n_cp_c(1-p_c)$. By completing squares in Eq. (A17), a little algebra shows that under these assumptions Pr(**i** | x) is multivariate normal with a vector of means

$$\mu_c=\frac{1}{a_c}\left[h_c^2x+\left(1-h_c^2\right)\overline{x}\right] \quad \text{(A18)}$$

and a variance-covariance matrix given by

$$\begin{aligned}\Sigma_{cc} &= \frac{\sigma_{Ac}^2\left(1-h_c^2\right)}{a_c^2\sigma_x^2}\\ \Sigma_{cc'} &= -\frac{\sigma_{Ac}^2\sigma_{Ac'}^2}{a_ca_{c'}\sigma_x^2} \quad c\neq c'\end{aligned} \quad \text{(A19)}$$

where $\sigma_{Ac}^2=2n_cp_c\left(1-p_c\right)a_c^2$ is the additive component of genetic variance attributable to class c loci and $h_c^2=\sigma_{Ac}^2/\sigma_x^2$.

For gamete formation, assume that the distribution of j_c is an independent hypergeometric (Eq. A9 with a subscript c added to all variables). The assumption of independence is only approximate. If the overall distribution of the total number of + alleles in a gamete is hypergeometric, then there will be a slight negative correlation in the numbers in each class.

It follows from these assumptions that

$$E\left(y|x^{(m)},x^{(f)}\right)=h^2\frac{x^{(m)}+x^{(f)}}{2}+\left(1-h^2\right)\overline{x} \quad \text{(A20)}$$

as expected. The expression for the variance is more complicated:

$$\begin{aligned} &Var\left(y^2|x^{(m)},x^{(f)}\right)\\ &=\sigma_e^2+\frac{1}{2}\sum_{c=1}^{c}\left[\frac{2n_c-2}{2n_c-1}\sigma_{A_c}^2\left(1-h_c^2\right)-h_c^2\sum_{c'\neq c}\sigma_{A_{c'}}^2\right]\\ &+\frac{1}{4}\sum_{c=1}^{C}\frac{\tilde{x}_c^{(f)}\left(2n_ca_c-\tilde{x}_c^{(f)}\right)+\tilde{x}_c^{(m)}\left(2n_ca_c-\tilde{x}_c^{(m)}\right)}{2n_c-1} \end{aligned} \quad \text{(A21)}$$

where $\tilde{x}_c^{(m,f)}=h_c^2x^{(m,f)}+\left(1-h_c^2\right)\overline{x}$.

This expression can be used to define an effective number of loci. In Eq. (A12), the variance given the parental $V(y|x^{(m)},x^{(f)})$ is the sum of quadratic functions of x_1 and x_2. Taking the second derivative with respect to either parental phenotype gives

$$\frac{d^2V}{dx_{1,2}^2}=-\frac{h^4}{2\left(2n-1\right)}. \quad \text{(A22)}$$

In Eq. (A21) the variance is also a quadratic function of the parental phenotypes, $x^{(f)}$ and $x^{(m)}$. Therefore, the effective number of loci is defined to be the n that gives the same second derivative of the variance:

$$\frac{h^4}{2n_E-1}=\sum_{c=1}^{C}\frac{h_c^4}{2n_c-1} \quad \text{(A23)}$$

or

$$n_E=\frac{1}{2}\left(1+h^4/\sum_{c=1}^{C}\frac{h_c^4}{2n_c-1}\right) \quad \text{(A23)}$$

Although (A23) was derived under the assumption that n_c is large enough that the binomial distribution of i_c can be approximated by a normal distribution, it can be used as an effective number even when the allele frequency and additive effect are different for each locus and when there is dominance. In that case, $h_j^2=2p_j\left(1-p_j\right)\left(a_j-\left(2p_j-1\right)d_j\right)^2/\sigma_x^2$

$$n_E=\frac{1}{2}\left(1+h^4/\sum_{j=1}^{n}h_j^4\right) \quad \text{(A24)}$$

where n is the true number of loci.

Given n_E, the effective additive effect of a + allele is obtained by equating the additive genetic variance in the symmetric additive model to the additive variance for the actual model:

$$2n_E\overline{p}\left(1-\overline{p}\right)a_E^2=2\sum_{c=1}^{C}n_cp_c\left(1-p_c\right)a_c^2 \quad \text{(A25)}$$

where $\overline{p}=\sum_{c=1}^{C}n_cp_c/\sum_{c=1}^{C}n_c$ is the average allele frequency in the model. Therefore

$$a_E = \sqrt{\sum_{c=1}^{C} n_c p_c (1-p_c) (a_j - (2p_j - 1) d_j)^2 / (2n_E \overline{p})}. \quad \text{(A26)}$$

Appendix 3: Genotypes of known QTNs

Of all the loci that affect the quantitative character of interest, assume that some are known, meaning that they have been identified as affecting the character, and that their average genotypic effects on the character have been determined. The remaining loci are unknown. Assume all loci affecting the character, known and unknown, are in Hardy-Weinberg and linkage equilibrium.

Let $\mathbf{k} = \{k_1, k_2, ...\}$ be the genotype of an individual at the known loci, where k_j is the number of + alleles at locus *j*, and let $G(\mathbf{k})$ be the average phenotype of an individual with genotype $\mathbf{k}$. The function $G(\mathbf{k})$ can allow for dominance at each known locus and in principle allow for interactions among the known loci. In the absence of gene interactions, $G(\mathbf{k})$ is the sum of effects of for each locus estimated separately. The additive effects and dominance effects of the + allele at each known locus are not necessarily the same.

Assume that there is no interaction between the known loci as a group and the unknown loci, and assume that the symmetric model is valid for the unknown loci. There are *n* unknown loci and the frequency of the + allele at each is *p*. The additive effect of a + allele at each unknown locus is *a* and there is no dominance. With these assumptions, the derivation in Appendix 1 can be used here with only slight modification.

Consider an individuals with phenotype *x* and genotype $\mathbf{k}$ at the known loci, and let *i* be the number of + alleles at the unknown loci. Using Bayes theorem

$$\Pr(i|x, \mathbf{k}) = \frac{\Pr(x|i, \mathbf{k}) \Pr(i|\mathbf{k})}{\Pr(x|\mathbf{k})}. \quad \text{(A27)}$$

$\Pr(x \mid i, \mathbf{k})$ is, by assumption, a normal distribution with mean $G(\mathbf{k}) + ai$ and variance σ_e^2. $\Pr(i \mid \mathbf{k}) = \Pr(i)$ because of the assumption of linkage equilibrium. As in Appendix 1, $\Pr(i)$ is a binomial distribution that can be approximated by a normal distribution that has mean $2np$ and variance $2np(1-p)$. Therefore, the numerator of Eq. (A27) is approximately a normal distribution which has expectation

$$E(i|x, \mathbf{k}) = \eta^2 \left[\frac{x - G(\mathbf{k})}{a}\right] + \left(1 - \eta^2\right) \left[\frac{\overline{x} - G(\mathbf{k})}{a}\right] = \frac{\tilde{x}(\mathbf{k})}{a} \quad \text{(A28)}$$

where

$$\eta^2 = \frac{2np(1-p)a^2}{2np(1-p)a^2 + \sigma_e^2} \quad \text{(A29)}$$

and

$$\tilde{x}(\mathbf{k}) = \left(1 - \eta^2\right)(x - G(\mathbf{k})) + \left(1 - \eta^2\right)(\overline{x} - G(\mathbf{k})). \quad \text{(A30)}$$

The second moment is

$$E\left(i^2|x,\mathbf{k}\right)=\eta^2\sigma_e^2+\left(\tilde{x}\left(\mathbf{k}\right)\right)^2. \quad \text{(A31)}$$

The generation of gametes carrying $j+$ alleles at the unknown loci is the same as in Appendix 1. The genotype $\mathbf{k}_y$ of the known loci in the offspring is assumed known. Therefore, the rest of the derivation in Appendix 1 can be used to obtain

$$\begin{aligned}
&E\left(y|\mathbf{k}_y,x_1,\mathbf{k}_1,x_2,\mathbf{k}_2\right)\\
&\quad=G\left(\mathbf{k}_y\right)+E\left(j_1|x_1,\mathbf{k}_1\right)\\
&\quad+E\left(j_2|x_2,\mathbf{k}_2\right)\\
&\quad=G\left(\mathbf{k}_y\right)+\frac{\tilde{x}\left(\mathbf{k}_1\right)+\tilde{x}\left(\mathbf{k}_2\right)}{2}\\
&\quad=G\left(\mathbf{k}_y\right)\eta^2\frac{\left(x_1-G\left(\mathbf{k}_1\right)\right)+\left(x_2-G\left(\mathbf{k}_2\right)\right)}{2}\\
&\quad+\left(1-\eta^2\right)\left(\overline{x}-\frac{G\left(\mathbf{k}_1\right)+G\left(\mathbf{k}_2\right)}{2}\right)
\end{aligned} \quad \text{(A32)}$$

and

$$Var\left(y|\mathbf{k}_y,x_1,\mathbf{k}_1,x_2,\mathbf{k}_2,\mathbf{k}_y\right)=\frac{1}{2}\frac{2n-2}{2n-1}\eta^2\sigma_e^2+\sigma_e^2+\frac{1}{4\left(2n-1\right)}\left[\tilde{x}_1\left(\mathbf{k}_1\right)\left(2La-\tilde{x}_1\left(\mathbf{k}_1\right)\right)+\tilde{x}_2\left(\mathbf{k}_1\right)\left(2La-\tilde{x}_2\left(\mathbf{k}_1\right)\right)\right]. \quad \text{(A33)}$$

The formulas are the same as Eqns. (2) and (3) in the text once $G(\mathbf{k})$ is subtracted from each phenotype and h^2 is replaced by η^2.

Given a set of parent-offspring trios, the data for each family consist of genotypes at known loci and phenotypes, $\left\{x_1^{(f)},\mathbf{k}_1^{(f)},x_2^{(f)},\mathbf{k}_2^{(f)},y^{(f)},\mathbf{k}_y^{(f)}\right\}$ f= 1,2,...,F. The first step is to estimate σ_x^2, h^2, $\sigma_A^2=h^2\sigma_x^2$, and $\sigma_e^2=\left(1-h^2\right)\sigma_x^2$. from the phenotypes of the parents. The next step is to estimate from the genotypes of the parents the additive genetic variance attributable to known loci $\sigma_{A,\text{known}}^2$, from which the additive variance attributable to the unknown loci can be estimated: $\sigma_{A,\text{unknown}}^2=\sigma_A^2-\sigma_{A,\text{known}}^2$. From that, we can estimate η^2:

$$\eta^2=\frac{\sigma_{A,\text{unknown}}^2}{\sigma_{A,\text{unknown}}^2+\sigma_e^2}. \quad \text{(A34)}$$

The final step is to assume $y^{(f)}-G(\mathbf{k}_y)$ is normally distributed with mean and variance given by Eqns. (A32) and (A33), from which the approximate likelihood can be computed for each family.

Literature Cited

Castle WE. An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. Science. 1921; 54:223. [PubMed: 17792870]

Cockerham CC. Modifications in estimating the number of genes for a quantitative character. Genetics. 1986; 114:659–664. [PubMed: 3770473]

Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall; New York: 1993.

Fain, PR. Ph. D. thesis. University of Colorado; Boulder: 1976. Major gene analysis: An alternative approach to the study of the genetics of human behavior..

Fain PR. Characteristics of simple sibship variance tests for detection of major loci and application to height, weight and spatial performance. Annals of Human Genetics. 1978; 42:109–120. [PubMed: 686678]

Felsenstein J. Estimation of number of loci controlling variation in a quantitative character. Genetics. 1973; 74:s79.

Jasinska AJ, et al. A non-human primate system for large-scale genetic studies of complex traits. Human Molecular Genetics. 2012; 21:3307–3316. [PubMed: 22556363]

Karlin S, Carmelli D. Evolutionary aspects and sensitivity studies of some major gene models. Journal of Theoretical Biology. 1978; 75:197–222. [PubMed: 758023]

Karlin S, et al. Index measures for assessing the mode of inheritance of continuously distributed traits: 1. Theory and justifications. Theoretical Population Biology. 1979; 16:81–106. [PubMed: 531766]

Karlin S, et al. Structured exploratory data-analysis (SEDA) of finger ridge-count inheritance: 1. Major gene index, midparental correlation, and offspring-between-parents function in 125 South Indian families. American Journal of Physical Anthropology. 1983; 62:377–396. [PubMed: 6666769]

Karlin S, et al. Structured exploratory data-analysis (SEDA) for determining mode of inheritance of quantitative traits. 1. Simulation studies on the effect of background distributions. American Journal of Human Genetics. 1981; 33:262–281. [PubMed: 7211841]

Lande R. The minimum number of genes contributing to quantitative variation between and within populations. Genetics. 1981; 99:541–553. [PubMed: 7343418]

Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838. [PubMed: 20881960]

Ott J. Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. American Journal of Human Genetics. 1979; 31:161–175. [PubMed: 453201]

Otto SP, Jones CD. Detecting the undetected: estimating the total number of loci underlying a quantitative trait. Genetics. 2000; 156:2093–2107. [PubMed: 11102398]

Pearson K. Mathematical contributions to the theory of evolution. XII. On a generalised theory of alternative inheritance, with special reference to Mendel's laws. Philosophical Transactions of the Royal Society of London. 1904a; 203:53–86.Series A.

Pearson K. On a criterion which may serve to test various theories of inheritance. Proceedings of the Royal Society of London. 1904b; 73:262–280.

Pearson K, Lee A. On the laws of inheritance in man. I. Inheritance of physical characters. Biometrika. 1903; 2:357–462.

Penrose LS. Effects of additive genes at many loci compared with those of a set of alleles at one locus in parent-child and sib correlations. Annals of Human Genetics. 1969; 33:15–21. [PubMed: 5821315]

Stark AE. Method of Penrose of estimating number of effective factors contributing to a character. Annals of Human Genetics. 1976; 39:465–470. [PubMed: 952489]

Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010; 42:565–569. [PubMed: 20562875]

Zeng ZB. Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character: a further improved method. Genetics. 1992; 131:987–1001. [PubMed: 1325390]

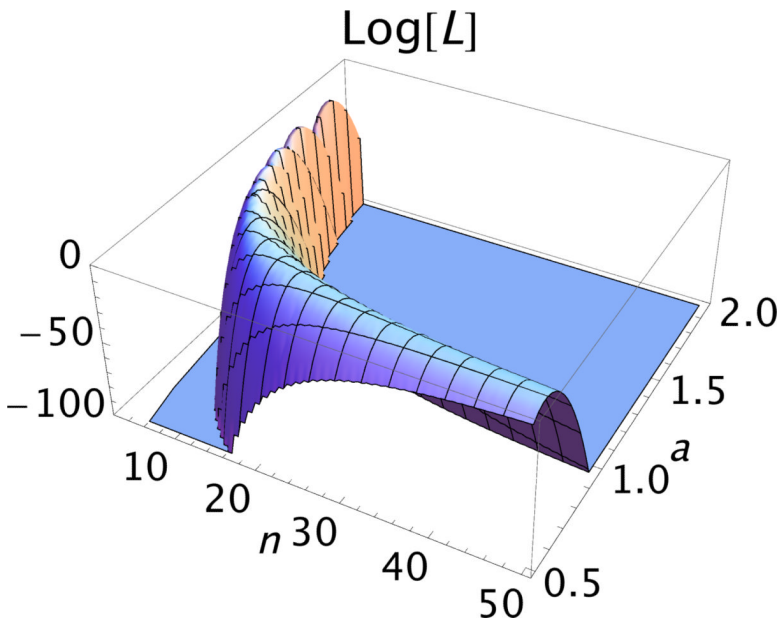

Figure 1.
Log-likelihood surface for one replicate of the symmetric model with p=0.5, a=1, $\sigma_e^2=5$ $\left(h^2 \approx 0.5\right)$ and F=10,000 parent-offspring trios. $\hat{n}$=9, $\hat{a}$=1.04. In the grid search $2 \leq n \leq 50$ with grid size 1 and $0.5 \leq a \leq 2$ with grid size 0.03. The surface represents log(L)–max[log(L)], where max[log(L)]=–24983.8.

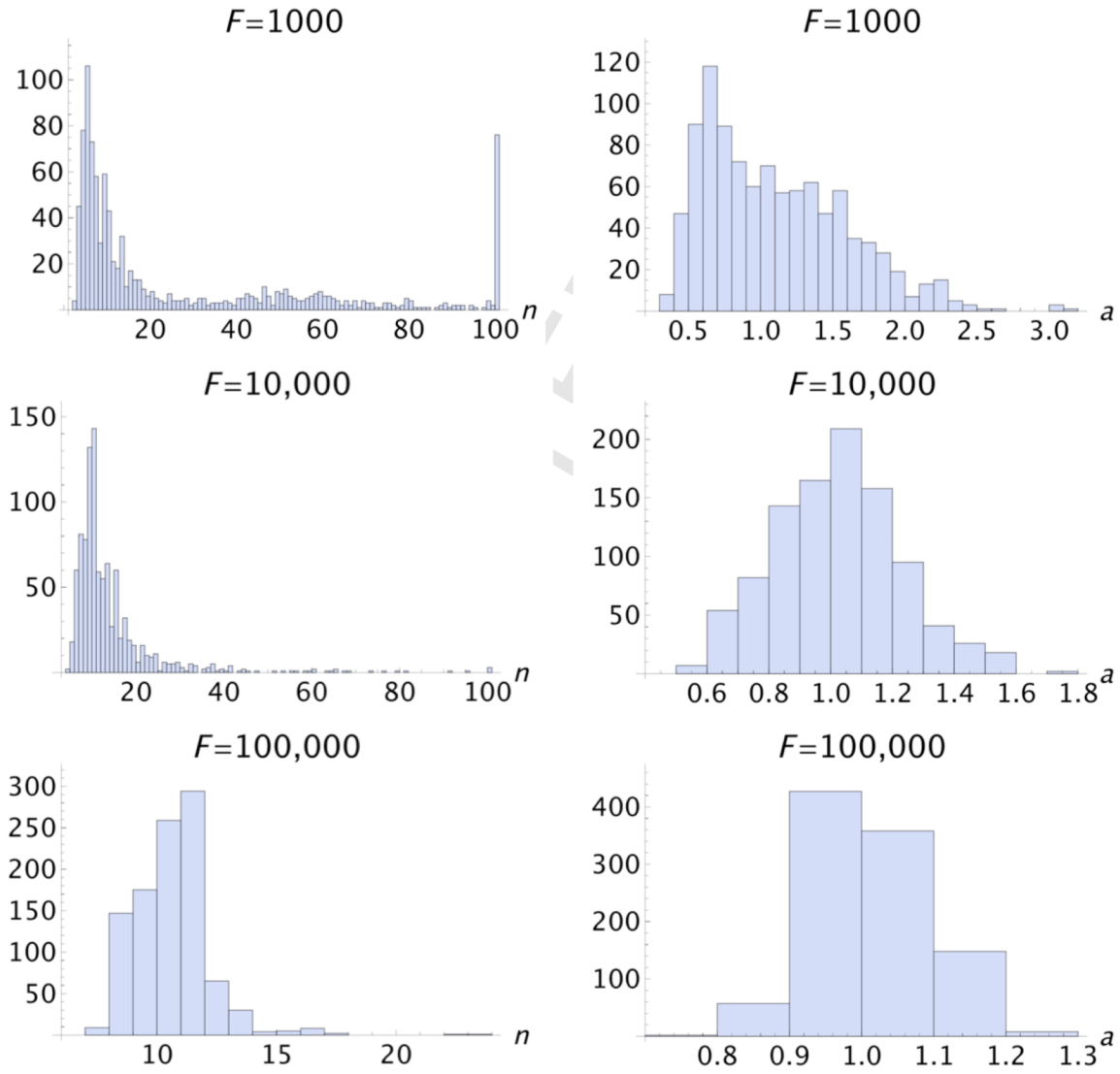

Figure 2.

Histograms of $\hat{n}$ and $\hat{a}$ in 1000 replicates of data simulated under the symmetric model with p=0.5, n=10, a=1 and σ_e^2=5. In each replicate, a grid search over the ranges $2 \leq n \leq 100$ and $0.2 \leq a \leq 4$ was performed to find the MLEs. F is the number of family trios in each simulated data set.

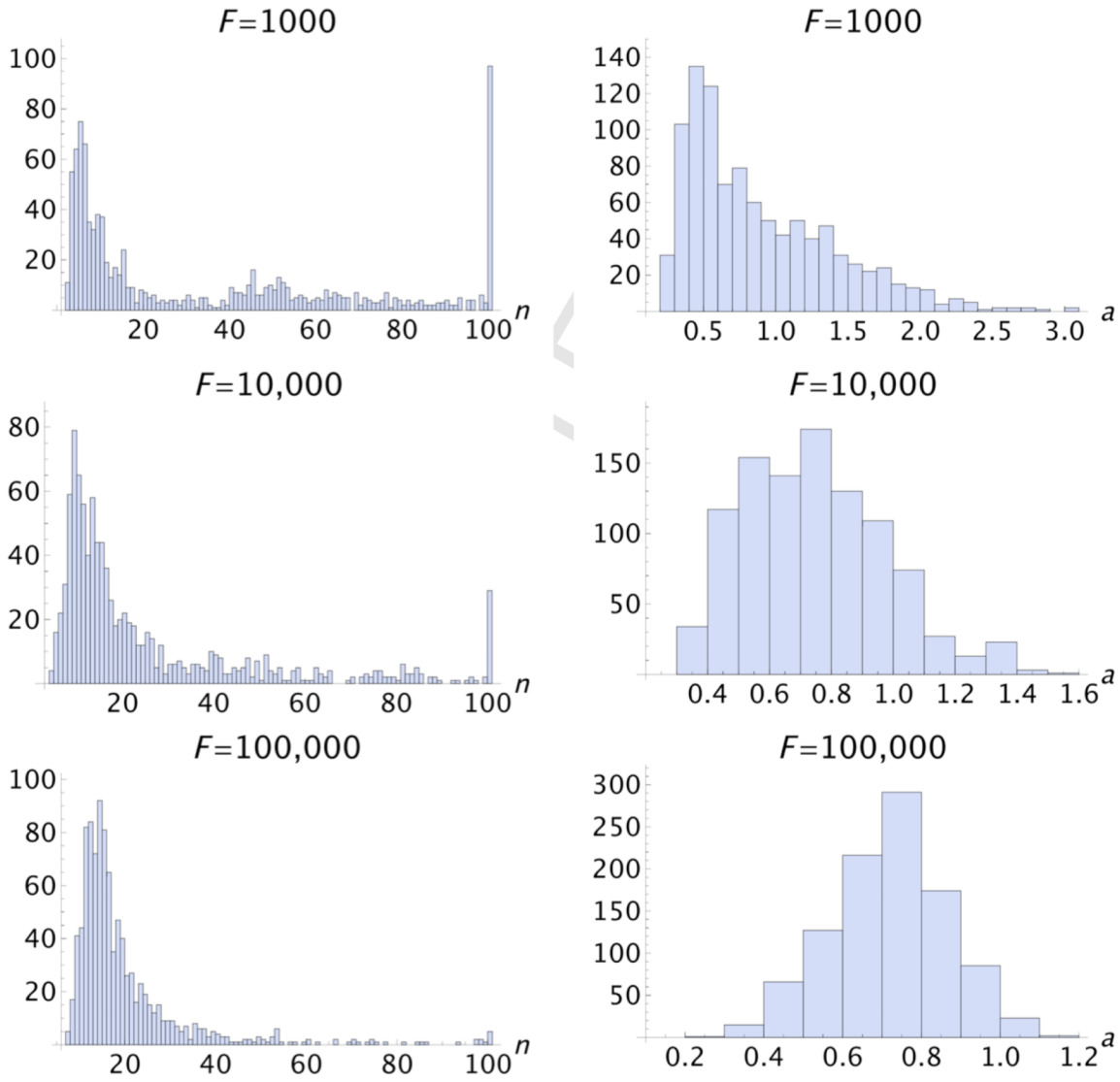

Figure 3.

Histograms of $\hat{n}$ and $\hat{a}$ in 1000 replicates of data simulated under the assumption that n=10 and a=1 but with allele frequencies that differed across loci. In each replicate, p_j (the frequency of the + allele at locus j) was drawn independently from a beta distribution with mean 0.5 and variance 0.05. In each replicate, σ_e^2 was adjusted so that $h^2 \approx 0.5$. In each replicate a grid search over the ranges $2 \le n \le 100$ and $0.2 \le a \le 4$ was performed to find the MLEs. F is the number of family trios in each simulated data set.

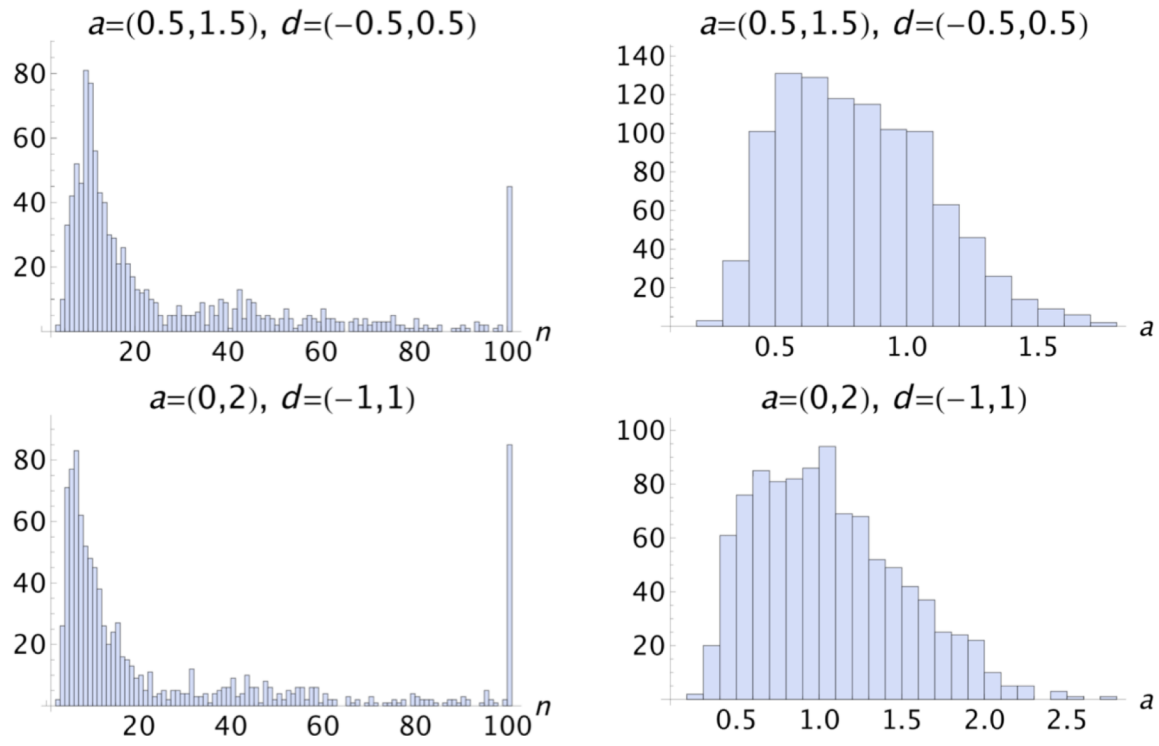

Figure 4.

Histograms of $\hat{n}$ and $\hat{a}$ in 1000 replicates of data from F=10,000 parent offspring trios simulated under the assumption that n=10 but with allele frequencies, additive effects and dominance deviations that differed across loci. In each replicate, p_j was drawn independently from a beta distribution with mean 0.5 and variance 0.05, a_j and d_j were drawn from uniform distributions with limits specified in each histogram. In each replicate, σ_e^2 was adjusted so that $h^2 \approx 0.5$. In each replicate a grid search over the ranges $2 \leq n \leq 100$ and $0.2 \leq a \leq 4$ was performed to find the MLEs.

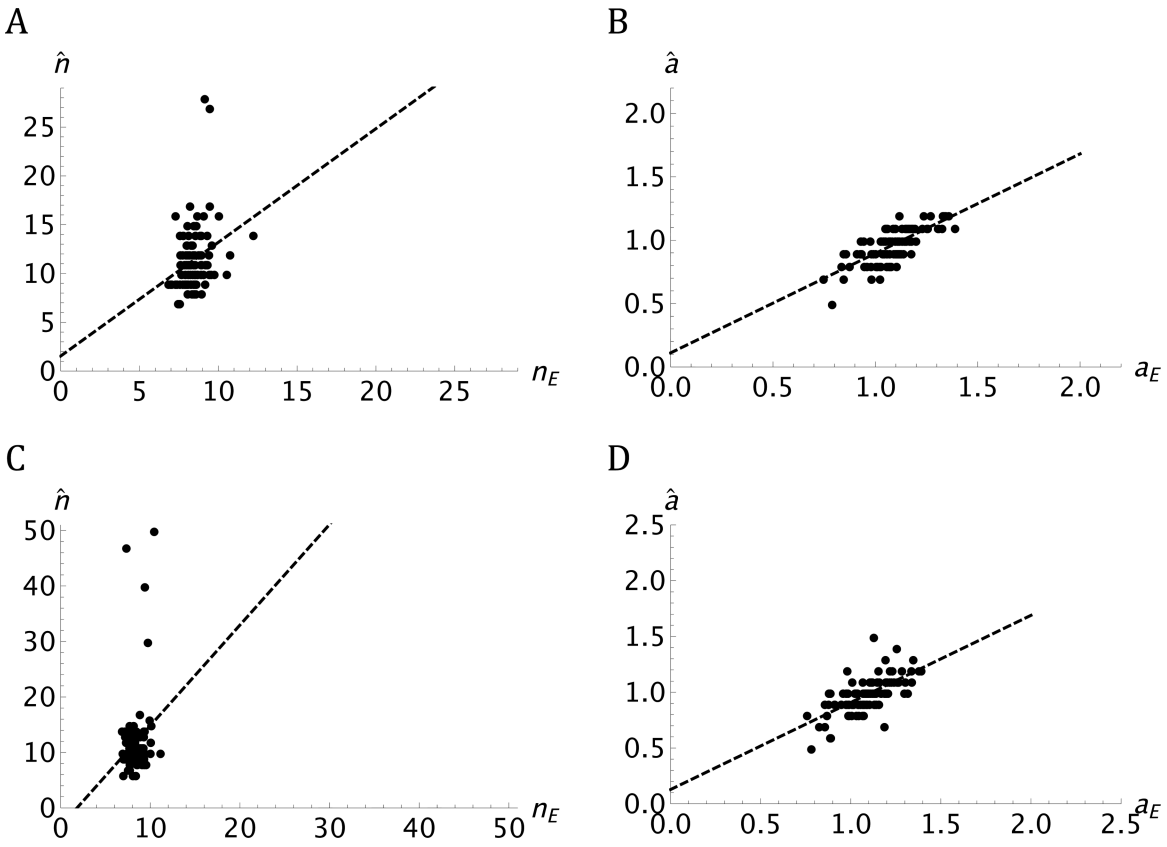

Figure 5.

Comparison of effective and estimated numbers of loci (A and C) and additive effects (B and D) in two sets of simulated data. In both sets, 100 replicate simulations of F=1,000,000 families were run for n=10 loci, the p_j were drawn from a beta distribution with mean 0.5 and variance 0.03, and a_j were drawn from a uniform distribution on (0.5, 1.5). In parts A and B, d_j=0, and in parts C and D, d_j was drawn from a uniform distribution on (–0.5, 0.5). n_E and a_E were computed for each replicate from Eqs. (10) and (11). The dashed lines are the regression lines fitted to the points. The regression equations for each part are (A) $1.54+1.16x$ (B) $0.11+0.78x$ (C) $-3.30+1.81x$ (D) $0.13+0.78x$.

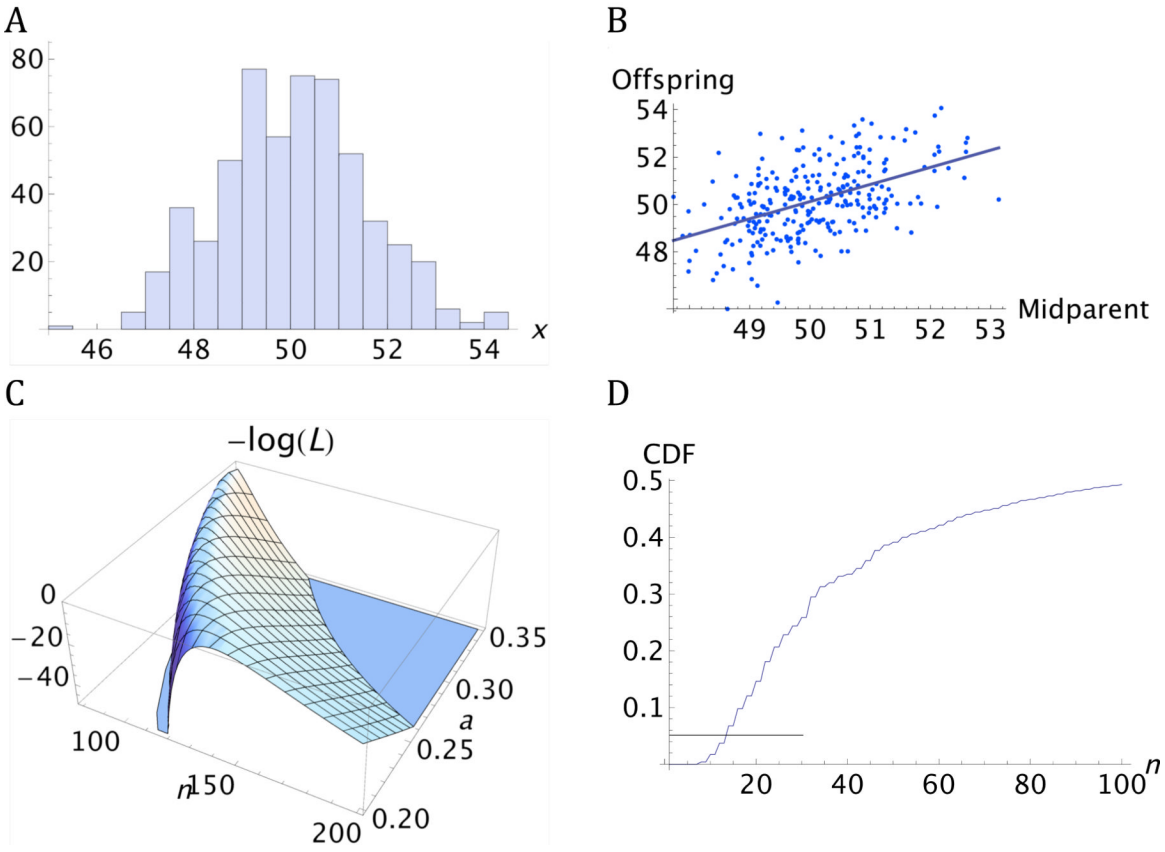

Figure 6.

Results from the analysis of crown-rump length in 281 parent-offspring trios of vervet monkeys (*Chlorocebus aethiopus sabaeus*). All measurements in females were multiplied by 1.132 to equalize the means of males and females. Part A shows the roughly normal histogram of adjusted measurements in males and females combined. Part B shows the regression of offspring on the midparent value. The regression equation is 13.88+0.725*x*. Part C shows the difference between the maximum log-likelihood and the computed log-likelihood for ranges of *a* and *n*. The MLEs are $\hat{n}$=112 and $\hat{a}$=0.26 cm. Part D shows the cumulative distribution function (CDF) of the distribution of $\hat{n}$ in 100,000 replicate simulations of the symmetric model in which *n*=112, *a*=0.26, and *F*=281. The 5% quantile at *n*=14 is indicated by the straight line.

Table 1

Dependence of the performance of the likelihood method on the number of siblings per family measured.

	$\hat{n}$	sd	$\hat{a}$	sd
F=10,000, S=1	6.0	2.8	0.96	0.15
F=5000, S=2	6.0	3.0	0.95	0.21
F=2500, S=4	6.3	4.0	0.96	0.18
F=1000, S=10	6.5	4.6	0.95	0.21

In all cases, the symmetric model with n=5, p=0.5 and a=1 was assumed. F is the number of independent families and S is the number of full siblings per family. $\hat{n}$ and $\hat{a}$ are the averages over 1000 replicate simulations and the sd values are the standard deviations

Table 2

Dependence of performance of the likelihood method on whether loci are linked or unlinked. The results are from 1000 replicates of the symmetric model with n=10, a=1 and p=0.5. The averages and standard deviations for unlinked loci are calculated from the results shown in Figure 1.

	Unlinked loci		Completely linked loci	
F	$\hat{n}$ **(sd)**	$\hat{a}$ **(sd)**	$\hat{n}$ **(sd)**	$\hat{a}$ **(sd)**
1000	27.7 (30.7)	1.06 (0.49)	27.3 (31.5)	1.07 (0.50)
10,000	13.9 (11.3)	0.97 (0.21)	14.1 (12.7)	0.98 (0.21)
100,000	10.7 (1.6)	0.98 (0.08)	10.7 (1.6)	0.97 (0.07)