Sampling and Ascertainment in Genetic Epidemiology: A Tutorial Review

E. A. Thompson,

Department of Statistics, GN-22,
University of Washington,
Seattle, WA 98195.

January, 1993.

Fisher (1934):

> "It is a statistical commonplace that the interpretation of a body of data requires knowledge of how it was obtained"

PREFACE

This technical report consists of a preliminary version of a Chapter to be contributed to a text on Genetic Epidemiology, authored by Dr. Tim Bishop, with others. Although the chapter can stand alone, some references are made to chapters by other authors. In particular, Chapter 2 of the text provides an introduction to genetic terminology, Chapter 4 provides an introduction to likelihood analysis, Chapter 7 will discuss segregation analysis in the context of simple genetic models, and Chapter 10 describes linkage analysis. Thus this chapter, Chapter 8, assumes previous knowledge of the basics of genetics, likelihood and segregation analysis. Section 8.7 presents material that is statistically more demanding than the remainder of the chapter; the first six sections require only knowledge of the basics of probability, likelihood analysis, and maximum likelihood estimation.

In the spirit of a tutorial text, references within the body of the text are kept to a minumum. Many of the ideas presented in the chapter have, of course, been developed in research papers by many different authors. The further reading section attempts to cite the sources on which this chapter has drawn, while also providing readers with a guide to the literature and further reading. Exercises are provided, to help readers guage their understanding of the material presented; a couple of exercises follow most sections of the chapter. Answers to and notes on these exercises are provided at the end of the chapter.

The coauthors of the text have made many useful comments on earlier drafts; in particular I would like to thank Drs. Robert Elston, Tim Bishop, and John Williamson. Thanks are also due to Dr. John Hopper for the example of section 8.6, and to Dr. Anthony Edwards for bringing Huzarbazar's paper to my attention, and also, some years ago, that of Galton (1904). At the University of Washington, Drs. Ellen Wijsman and Jane Olson also made helpful comments: Ellen Wijsman suggested discussion of the example of exercise 8.4.5. Above all, I am grateful to Dr. Warren Ewens, who made many detailed comments and criticisms of an earlier draft, and entered into fruitful discussions resulting in many clarifications and corrections. Any remaining errors are, of course, my own.

This research was supported in part by NIH grant GM-46225.

CHAPTER 8: SAMPLING AND ASCERTAINMENT

CONTENTS

8.1	Sampling biases and ascertainment	4
8.2	Classical ascertainment theory; sampling sibships	7
8.3	Classical ascertainment theory applied to pedigrees	12
8.4	Further aspects of the classical approach	
	(i) Sibships of variable parental mating type	16
	(ii) Quantitative traits, population distributions and prior information	18
	(iii) Linkage analysis	21
8.5	A more general view of the sampling problem	23
8.6	Robust approaches to ascertainment; qualitative phenotypes	31
8.7	Quantitative traits and the analysis of expected log-likelihoods	38
8.8	Further reading and open problems; a brief history of ascertainment theory	42
APPENDIX:	Solutions to Exercises and Related Notes	47
	Literature cited	52

CHAPTER 8: SAMPLING AND ASCERTAINMENT

8.1. Sampling biases and ascertainment.

The likelihood is the probability of the observed data viewed as a function of the parameters (Chapter 4). However, this probability depends not only on the data observed and on the genetic model, but also on the way in which the data are sampled. In genetic epidemiology one rarely samples individuals, or even families, at random. Often the sampling will focus on families in which it is known that a trait of interest is segregating; in some cases this affects the likelihood function. In a large-scale study of the genetic epidemiology of a complex trait such as cancer, it is often hard if not impossible to state precisely the sampling criteria. The sampling frame of reference is often not clearly defined. In this chapter, we shall start with simpler examples in order to elucidate current thinking in this area and to illustrate the principles involved.

A very early discussion of the effects of the sampling method is due to Galton (1904), who was interested in the survival of names and titles, and thus in the number of kin of a given sex and of varying degree. In particular, he considered the number of sisters of a random female. In a sibship of size $2B$ he argued, correctly, that a female child would have, on average, $B - \frac{1}{2}$ brothers and $B - \frac{1}{2}$ sisters, and not, as argued by others, B brothers and $B - 1$ sisters.

An even simpler version of this problem relates to the sexes of dizygous twins. Suppose every child is independently male (M) or female (F) with probability 1/2. Thus, in the population, dizygous twin pairs have sexes MM, MF/FM, FF with probabilities 1/4, 1/2 and 1/4 respectively. Then among twin pairs containing at least one male, there are twice as many discordant (MF/FM) pairs as concordant (MM) pairs. A male twin of a randomly chosen twin pair (discarding pairs that do not contain a male twin) is twice as likely to have a sister as a brother twin. Yet if we ask a random male twin the sex of his sibling, exactly one half will have a sister, one half a brother -- naturally so, since every such sibling has (independently) a 50% chance of being M and 50% chance of being F. The discrepancy is explained by the fact that in choosing a random male twin individual, an MM pair is twice as likely to be sampled as an MF/FM pair.

A related example is that of size-biased sampling; the number of siblings of a random individual is not the same as the number of siblings of an individual in a random sibship -- sampling by parents vs sampling through offspring. Suppose the proportion of families of size x in the population is k_x, and that mean sibship size is $\mu = \sum_x x\ k_x$, and variance $V = \sum_x (x - \mu)^2 k_x$. Now a random individual derives from a sibship size x with probability

$$q_x = (x\ k_x \,/\, \sum_z z\ k_z) = x\ k_x / \mu \qquad (8.1.1)$$

since a sibship size x has a x −fold chance of being that of a randomly selected individual. Such an individual has $(x - 1)$ sibs. Thus the mean number of sibs of a random individual is

$$\sum_x q_x(x-1) = \sum_x k_x x(x-1)/\mu = ((V+\mu^2)-\mu)/\mu = (\mu-1)+(V/\mu) \qquad (8.1.2)$$

That is, the sampling procedure that weights sibships in proportion to their size, by sampling random individuals, increases the mean number of sibs by an amount V/μ.

The type of sampling bias that arises in genetic epidemiological studies is similar, although rather than being sex-biased or size-biased, it is trait-biased. In studying a rare trait, especially, it would be ineffective to sample individuals or families at random, since most would not show the trait. Instead it is very much more efficient to sample through affected individuals. If, for genetic or environmental reasons, affected individuals cluster in pedigrees, ensuring the presence of at least one affected individual greatly enriches the sample. However, this sampling through affected individuals obviously influences the numbers and patterns of affected individuals in the sampled families -- that, after all, is the intent. That is, the sampling procedure affects the likelihood.

In this chapter, we shall use the example of sampling sibships and of an incompletely penetrant recessive disease to illustrate a variety of ideas. Initially, we shall restrict attention to a simple (fully penetrant) recessive trait, and shall further assume the trait to be reproductively lethal. We shall ignore the possibility of mutation. In this simple case, there is only one parental mating type producing the affected children; both parents must be the carrier genotype B_1B_2. The problem is then the classic one of testing whether the data are in accord with the expected segregation probability of 1/4. In a sibship of size 3 arising from a $B_1B_2 \times B_1B_2$ mating, there are 3, 2, 1, and 0 affected offspring with probabilities 1/64, 9/64, 27/64 and 27/64, respectively. Thus the mean frequency of affected offspring is

$$(1\times3 + 9\times2 + 27\times1 + 27\times0)/(64\times3) = 1/4,$$

as expected. However the families with no affected offspring cannot be identified as having two heterozygous parents even when this is the case. The frequency of affected offspring among families that contain at least one affected offspring is

$$(1\times3 + 9\times2 + 27\times1)/((1+9+27)\times3) = 0.432$$

which is greater than the true segregation frequency 1/4. This is an *ascertainment bias.* It is sometimes proposed that a correct segregation probability can be found by eliminating one affected offspring from each sibship. Thus we count 2, 1, 0 "other" affected offspring among the 2 "other" offspring in each sibship, when in total there are 3, 2, and 1 affected. This introduces the idea of an "ascertainment correction" -- a method of making correct estimates under biased sampling. However, as we shall see, this particular ascertainment correction is not valid, in general. The mean value of the estimated segregation frequency is now

$$(1\times2 + 9\times1 + 27\times0)/((1+9+27)\times2) = 0.149$$

and we underestimate the segregation frequency.

More generally, we may suppose families with 3, 2, 1, and 0 affected offspring are sampled with relative frequencies p_3c_3, p_2c_2, p_1c_1 and p_0c_0, where p_3, p_2, p_1 and p_0 are the population frequencies of these family types. That is, the c_i are the weights imposed by non-random sampling of families; random sampling corresponds to $c_3 = c_2 = c_1 = c_0 = 1$. Random sampling among families containing at least one affected ("truncate selection") is $c_3 = c_2 = c_1 = 1,\ c_0 = 0$. The segregation ratio and the "delete-one-affected" segregation ratios are

$$\frac{(1 \times 3 \times c_3 + 9 \times 2 \times c_2 + 27 \times 1 \times c_1 + 27 \times 0 \times c_0)}{(1 \times c_3 + 9 \times c_2 + 27 \times c_1 + 27 \times c_0) \times 3}$$

and

$$\frac{(1 \times 2 \times c_3 + 9 \times 1 \times c_2 + 27 \times 0 \times c_1)}{(1 \times c_3 + 9 \times c_2 + 27 \times c_1) \times 2}$$

respectively.

Single ascertainment, which will be discussed more fully below, is an important special case. Here the probability of ascertainment is proportional to the number of affected offspring, or $c_i = i$; a sibship with 2 affected offspring is assumed twice as likely to enter the sample as one with 1 affected offspring. The "delete-one-affected" segregation ratio is then

$$\frac{(1 \times 2 \times 3 + 9 \times 1 \times 2 + 27 \times 0 \times 1)}{(1 \times 3 + 9 \times 2 + 27 \times 1) \times 2} = 1/4$$

the correct population value! Thus this oft quoted ascertainment correction is correct for the single-ascertainment case. However, it can also be shown that this correction is appropriate only in this case of single ascertainment.

EXERCISE 8.1.1: Suppose that sibship size has mean μ and variance V, and that each child is a boy/girl with probability $\frac{1}{2}$ independently of the sexes of other children. Find the covariance between the numbers of boys and of girls in families.

Solutions to the exercises are given at the end of the chapter, where some notes explaining the significance of the exercise or solution may also be found.

8.2. Classical ascertainment theory; sampling sibships.

The earliest formal developments of ascertainment theory introduced the idea of a *proband.* A proband is an individual who, because of his/her phenotypic characteristics and independently of any other member of the family, brings the family into the sample. Although the concept of probands had been introduced by Weinberg (1912) [see EXERCISE 8.2.1], it was Fisher (1934) who introduced likelihood estimation into questions of ascertainment, by developing a likelihood based on the numbers of probands and affecteds in ascertained families. In order to develop such a likelihood, a probability model is required both for the sampling process and for the (genetic) trait on which sampling is based.

The simplest model for determination of probands assumes a constant probability π that any affected individual becomes a proband; ascertainments of different affected individuals are assumed independent. Under this model, the probability that a sibship with r affected individuals contains at least one proband, and thus enters the sample, is

$$(1 - (1 - \pi)^r) \tag{8.2.1}$$

The ascertainment probabilities for a sibship with r affected individuals are proportional to the sampling weights c_r of section 8.1. Under the above ascertainment model, using again our example of sibships of size 3, we have

$$c_0 : c_1 : c_2 : c_3 = 0 : \pi : (1 - (1 - \pi)^2) : (1 - (1 - \pi)^3)$$

As $\pi \rightarrow 0$, $c_0 : c_1 : c_2 : c_3 \rightarrow 0 : 1 : 2 : 3$. This special case is often referred to as *single ascertainment,* since is this case there will only be a single proband in each sibship. As $\pi \rightarrow 1$, $c_0 : c_1 : c_2 : c_3 \rightarrow 0 : 1 : 1 : 1$. This special case is often referred to as *truncate ascertainment,* since the sibships with no affected individuals are truncated from the potential sample. It is also referred to as *complete ascertainment,* since every affected individual is ascertained, but this term is also used for sampling from the complete population, $c_0 : c_1 : c_2 : c_3 = 1 : 1 : 1 : 1$, so care should be taken in using this term. It is also important to recognise that, although $0 \leq \pi \leq 1$, single and truncate ascertainment are not the two extremes of ascertainment models. For example, $c_0 : c_1 : c_2 : c_3 = 0 : 1 : 4 : 9$ might be a reasonable assumption, sibships with more affected individuals being more likely to be diagnosed as having a known familial trait. Further, although it is usually natural to assume $c_0 \leq c_1 \leq c_2 \leq c_3$, there could be situations where even this is not the case; sibship with large numbers of affected members might tend to be mistakenly classified as a different syndrome, for example.

To develop the likelihood, we shall consider first the case of a reproductively-lethal simple-recessive trait. Then sibships containing affected individuals are of only one type; both parents must be carriers of the gene for the trait. In a family in which the trait is segregating, each child is affected independently with probability ζ. The plausible value of ζ is 1/4, but other values might be of interest. For example, if the trait were associated with increased early fetal death, then among live-born sibs ζ would be less than 1/4.

If each affected child becomes a proband independently with probability π, then the probability of b probands in a sibship in which r sibs are affected is the binomial probability

$$\binom{r}{b} \pi^b (1 - \pi)^{r-b}$$

and in any size s sibship with at least one proband ($b \geq 1$ and hence $r \geq 1$), the probability of r affected sibs is

$$\binom{s}{r}\zeta^r(1-\zeta)^{s-r}.$$

The ascertained sibships are, by definition, those with at least one proband, and thus among ascertained sibships of size s the probability of b probands and r affected sibs is

$$\frac{\binom{r}{b}\pi^b(1-\pi)^{r-b}\binom{s}{r}\zeta^r(1-\zeta)^{s-r}}{\sum_{r=1}^{s}\sum_{b=1}^{r}\binom{r}{b}\pi^b(1-\pi)^{r-b}\binom{s}{r}\zeta^r(1-\zeta)^{s-r}} \tag{8.2.2}$$

for $1 \le b \le r \le s$. Note that

$$\sum_{b=1}^{r}\binom{r}{b}\pi^b(1-\pi)^{r-b} = (1-(1-\pi)^r)$$

is the probability of ascertainment of a sibship with r affected individuals, and that doing both summations in the denominator of (8.2.2) gives

$$(\ 1-(1-\pi\zeta)^s) \tag{8.2.3}$$

the probability that a random at-risk sibship size s is ascertained. Note that, if each child in an at-risk family is independently affected with probability ζ, then any child is independently a proband with probability $\pi\zeta$. The probability of at least one proband among these s sibs is given by (8.2.3).

If the number of probands is not observed, the probability of r affected sibs in an ascertained sibship of size s is obtained by summing (8.2.2) over $b \ge 1$, and is

$$\frac{(1-(1-\pi)^r)\binom{s}{r}\zeta^r(1-\zeta)^{s-r}}{(1-(1-\zeta\pi)^s)} \tag{8.2.4}$$

on $1 \le r \le s$.

The likelihood classically used in segregation analysis is the product of the above posterior probabilities (8.2.2) or (8.2.4) over observed sibships. It is a joint likelihood for the parameters ζ and π; either can be estimated if the other is known, or they may be estimated jointly. Thus sibships are assumed independently ascertained, and inference is based on the distribution of probands and affecteds, conditional on the sample of observed sibships. The probability that a particular sibship, of a particular size, in a particular subpopulation, is ascertained may depend on many complex factors. The approach of conditioning on the ascertained sibships avoids consideration of demographic complexities that are likely to complicate the ascertainment process. It provides a very useful framework for tests both of genetic hypotheses and of the ascertainment probabilities.

The likelihood (8.2.2) contains more information than (8.2.4); if information on the proband count, b, is available it is foolish to discard it. In fact without such information, it may be hard to obtain accurate parameter estimates. The likelihood (8.2.2) based only on the fact of ascertainment ($b \geq 1$) may be very flat, particularly with respect to the parameter π. In fact, the value of b provides not only additional information about π, and hence ζ, but also information to test the model of independently ascertained probands; i.e. $c_r = 1 - (1-\pi)^r$. If the distribution of numbers of probands are not in accord with this hypothesis, one may want to model c_r in some alternative way, and thus correct for more complex patterns of ascertainment. It is not in fact even necessary that the ascertainment probability be only a function of the total numbers of affected sibs, r. Even for a simply determined trait, whether an individual becomes a proband may depend on other factors (age, calendar year of birth etc.) and not only on whether he is affected.

All likelihood inference is conditional on the framework of the model (Chapter 4), and provided the limitations of the model (which includes the model for the process of ascertainment) are recognised the fact that a model falls short of reality does not preclude useful inferences. A theme in the development of this chapter will be that conditioning on some aspect of observed data can facilitate valid inferences about parameters of interest. On the other hand, it should also be recognised that any conditioning results in loss of information. The ascertainment corrections, and more specifically the more robust ascertainment corrections developed and discussed in this chapter, should not be taken as advocating the view that lack of knowledge of the method of ascertainment is no problem. An acceptable balance must be found between the statistical power to make inferences, and robustness of those inferences to the model assumptions. In the worst cases, lack of knowledge of ascertainment, or failure to adhere to a proposed ascertainment scheme, may result in the need to condition on all available information; the study will be totally useless. In general, there is no substitute for developing a good sampling design in advance of a study, adhering to it, and retaining information on the proband status of all individuals.

Finally, in this section, we develop a numerical example which illustrates this last point, and also the use of data on numbers of probands and affecteds to make inferences about the genetic model and the sampling procedure. Details of the computations are left as exercises (see EXERCISE 8.2.2).

a) Suppose in a large study of sibships of size 3, 1080, 540, and 70 sibships are observed with 1, 2, and 3 affected offspring, respectively. Note that under the model of a simple lethal recessive we expect such sibships in the ratio 27:9:1 in the population, and if $\pi = 1/2$ we would see 1/2, 3/4 and 7/8 of such sibships respectively. That is, from these data we would estimate that $\pi = 1/2$ and that $\zeta = 1/4$. However, the likelihood is rather flat; there is little information about π, and if π is inferred wrongly, ζ will also be misestimated.

b) How much does it help to know how many probands there were in each sibship? Suppose we were told that in the 70 sibships with 3 affecteds, there were 10, 30, and 30, with 3, 2, and 1 proband respectively, and that in the 540 sibships with 2 affecteds there were 180 with 2 probands, and 360 with only one. This would not change our inference, for it is exactly in accord with the expectations if $\zeta = 1/4$ and $\pi = 1/2$. However, the precision of estimation of π and ζ is greatly enhanced.

c) Suppose instead we were told that of the 70 ascertained sibships with 3 affecteds, there were 40, 20, and 10 with 3, 2, 1 probands, but that in only 140 of the 540 sibships with 2 affecteds, were both probands. This not only changes our estimates of π, but provides a strong indication that no model with $c_r = (1-(1-\pi)^r)$ will fit.

d) So now we could fit a model with more flexible c_r, and/or for the number of probands in an ascertained sibship with r affected sibs. However, if we allow arbitrary c_r, or constrain c_r only so that $c_1 \le c_2 \le c_3$, there is virtually no information on the relative population proportions of sibships with r affected sibs. All information on the segregation probability ζ is lost.

EXERCISE 8.2.1: It was recognised by Weinberg (1912) that in estimating the proportion of affected offspring in sibships at risk the required weighting was to count each sibship according to the number of probands. Thus, he proposed the following estimator for the segregation probability ζ from a sample of sibships of size s:

$$\frac{\sum_r \sum_b n_{r,b}\; b\;(r-1)}{\sum_r \sum_b n_{r,b}\; b\;(s-1)} \tag{8.2.5}$$

where $n_{r,b}$ is the proportion of sibship in the sample with b probands and r affected offspring.

(a) In the case of single ascertainment, $b = 1$ for every ascertained sibship. Show that the estimator (8.2.5) is then the total number of affected siblings divided by the total number of siblings. Show that this is the "delete-one-affected" estimator of section 8.1.

(b) In the case of truncate ascertainment, $b = r$; every affected offspring individual is a proband. Show that the estimator (8.2.5) is then the proportion of affected siblings, adjusted for the "size-biased" sampling of sets of affected siblings of size r (see equation (8.1.2).

(c) Under the model of this section, show that the expected value of b, given r, is πr, and that the expected value of $r(r-1)$ is $s(s-1)\zeta^2$. Hence show that the numerator of (8.2.5) has expectation $\pi s(s-1)\zeta^2$ while the denominator has expectation $\pi s(s-1)\zeta$.
NOTE: From this it can be shown that Weinberg's estimator is a consistent estimator of ζ. However, in general it is not the maximum likelihood estimator of ζ; it is not always asymptotically efficient.

EXERCISE 8.2.2: (a) Verify the details of part (a) in the above numerical example in the text. For example, compute the log-likelihood (8.2.4) on the basis of these data, at the 9 parameter points $\zeta = (0.2, 0.25, 0.3)$, $\pi = (0.3, 0.5, 0.7)$.

(b) Verify the details of part (b) in the above numerical example in the text. For example, compute the log-likelihood (8.2.2) on the basis of these data, at the same 9 (ζ, π) parameter combinations.

Another example of the same phenomenon as here illustrated numerically by exercises 8.2.2 (a) and (b) is shown diagrammatically in section 8.6 (figure 2).

c) Verify the details of part (c) in the above numerical example in the text. Specifically, considering the log-likelihood for π alone, provided by the subdivision of the 70 sibships with 3 affecteds, one obtains as estimate $\hat{\pi} = 0.80$. For the 540 sibships with two affecteds, one obtains $\hat{\pi} = 0.41$. For the log-likelihood for both samples combined, the MLE of π is still about 0.5 (0.515 in fact), but the expectations of the numbers of sibships with given numbers of probands given by $\pi = 0.515$ are far from the observed values.

d) Verify the details of part (d) in the above example in the text. Specifically, note that if p_r is the population proportion of sibships with r affected sibs, then all that can be estimated from the numbers of sibships is that

$$p_1 c_1 : p_2 c_2 : p_3 c_3 = 108 : 54 : 7$$

This implies $p_1 \geq 2\ p_2 \geq (108/7)\ p_3$. In terms of ζ, the first inequality gives $\zeta \leq 0.33$ and the second gives $\zeta \leq 0.28$, but no lower bound on ζ can be estimated.

8.3. Classical ascertainment theory applied to pedigrees.

Unfortunately, no population consists of a set of independent sibships. Nor even is it a set of unrelated and independent pedigrees. Nonetheless the extension of the ideas of the previous section to small pedigrees has proven useful in the development and use of ascertainment theory. Thus, in this section, we extend from the sampling frame as a collection of sibships, to a collection of pedigrees. As in the case of sibships, we do not model their shapes or sizes; instead, we condition upon the shapes and sizes observed. The genetic models encompassed by the framework will also now be extended, at least to general Mendelian single-locus models for a dichotomous or quantitative trait (section 7.4). In section 8.4 we shall return to further aspects of sibship and nuclear family data in the context of these genetic models. In this and the following sections, ξ will denote the collection of parameters defining the genetic model and the ascertainment process.

The key idea is to compute the probability of the observed data conditional on the event that the pedigree of the observed structure has been ascertained, and then to form a likelihood by multiplying the probabilities obtained from each pedigree:

$$L(\xi) = \prod_{asc.ped.} P(data\ on\ pedigree\ i \mid pedigree\ i\ ascertained)$$

$$= \prod_{asc.ped} \left(\frac{P(data\ on\ i\ \&\ pedigree\ i\ has \geq 1\ proband)}{P(pedigree\ of\ structure\ of\ pedigree\ i\ has \geq 1\ proband)} \right) \tag{8.3.1}$$

If the probands are already specified as part of the data the numerator is simply the uncorrected likelihood function provided by pedigree i, the analogue of the numerator in (8.2.2). Chapter 7 developed these likelihoods in terms of the probabilities of genotypic configurations (determined by genotype probabilities of founders, and Mendelian segregation probabilities) and the probabilities of phenotype given genotype (penetrance probabilities). For a simple dichotomous trait, penetrances may be 0 or 1, but more generally they may be age or sex-dependent. For a quantitative trait, a simple penetrance function is a Gaussian probability density for the trait value, the mean of the density depending on an underlying Mendelian genotype; more generally effects of age, sex or other observed covariates may be modelled in the penetrance function. Section 7.4 shows how the "uncorrected likelihood", the probability of the observed phenotypic data on the pedigree, may be computed, even for extended pedigrees. The numerators in (8.3.1) are analogous to these "uncorrected likelihoods"; the only difference is that proband status (a dichotomous trait) is a part of the phenotype.

$$P(proband \text{ and } trait\ data\ on\ pedigree) = P(proband\ status \mid trait)P(trait\ data)$$

$$= \sum_G \prod_i P(proband_i \mid trait_i)P(trait_i \mid genotype_i)\ P(G) \qquad (8.3.2)$$

where G is a genotypic configuration on the pedigree, and summation is over all such configurations consistent with the data. The subscript i here refers to individuals; $proband_i$ is the proband status of individual i, and likewise for genotype and phenotype. The proband phenotype may be a simple stochastic function of trait phenotype. For example,

$$P(proband\ status = 1 \mid trait\ phenotype = affected) = \pi$$

$$P(proband\ status = 1 \mid trait\ phenotype = unaffected) = 0$$

or for a quantitative trait, π may be a function of the trait value. Proband status may also depend directly on age or sex, over and above any indirect effects of these covariates on trait values. Whatever the model, by considering proband status as an aspect of individual phenotype, the calculation of (8.3.2) is conceptually straightforward.

The denominator is more complex both computationally and conceptually, as there may be many different ways in which the pedigree could have contained at least one affected individual who becomes a proband. However, a pedigree is *not* ascertained only if every individual is *not* a proband. Thus, again, by regarding proband status as a "trait" dependent upon the genetic model (through trait values) as well as upon covariates, the same methods used to compute the probability of phenotypes on pedigrees permit the computation of the probability that a pedigree of given structure and covariate pattern is not ascertained.

$$P(no\ proband\ in\ pedigree) = \sum_G \prod_i P(i\ non-proband \mid genotype_i)\ P(G) \qquad (8.3.3)$$

The denominator is then one minus this probability.

The assumptions implicit in (8.3.1) should be recognised. The formulation is predicated on the concept of independently ascertained probands, whose predetermined "pedigrees" then enter the sample, There may be *multiple ascertainment* (several probands in a pedigree), but this is assumed not to lead to sampling of a more extensive pedigree structure. Despite its assumptions, the formulation (8.3.1) permits some flexibility in the sampling model for probands; it is not necessary that every affected individual be accorded the same ascertainment probability π. This is of practical importance. In many studies, families are sampled through affected individuals who reside in a given city, or who are referred to a given clinic. Thus it is possible that only individuals within a given age range, of a given sex, or living in a certain region, can become probands. More generally, even where probands are not absolutely restricted, the probability that an affected individual becomes a proband may depend on many factors. Where these factors are observable covariates pertaining only to the individual in question, permitting a covariate-dependent ascertainment probability π to be assigned to each individual, and where the assumption of independent ascertainments remains a viable approximation to the sampling procedure, the conditional likelihood (8.3.1) remains a useful approach.

It is possible to extend the above formulation to the case of *multiplex ascertainment;* in that case it is a more complex event, such as a pair of affected sibs, that is assumed to lead to ascertainment of a pedigree, but again the pedigree structure is assumed a fixed entity. The likelihood is the analogue of (8.3.1) but now an ascertainment probability would be attached to each multiplex event ("pair of affected sibs") rather than to each affected individual, and the denominator required would be that the pedigree contains an event that lead to its ascertainment. Multiplex ascertainment should be distinguished from multiple ascertainment. For multiplex ascertainment there is a single joint event that leads to study of the pedigree, whereas for multiple ascertainment a model of independently ascertained probands (or independently ascertained multiplex events) is formulated. Presence of at least one proband (or multiplex event) leads to sampling of the pedigree. However, there may be found to be several probands in a pedigree; the pedigree is then *multiply ascertained.*

As previously, the case of single ascertainment is obtained as $\pi \rightarrow 0$. Then there should be no more than one proband in any pedigree. It may even be the case that only one individual within a pedigree structure can be the proband. For example, suppose a study is undertaken of men aged 45-47, and their sibs, offspring and parents are then studied. Then, in most families, the proband will be uniquely identifiable. In this case the probability of ascertainment of the pedigree (the denominator in (8.3.1)) is simply the probability that the relevant individual is affected.

There are some cautions in extending this formulation to large pedigrees. First, whereas it may be possible to regard the population as a collection of independent small pedigrees, obviously this concept become problematic when extensive multigeneration pedigrees are sampled. A single proband will not generally lead to the decision to sample a 500-member pedigree; probands jointly cause this particular structure to be studied. The second caution is computational. Note only may the numerator (8.3.2) become difficult to compute on a large pedigree, where proband statuses and trait phenotypes must be jointly considered as data, but, since proband status is trait dependent rather than directly genotype-dependent, evaluation of terms contributing to the probability of ascertaining the pedigree may require summation over phenotypic values:

$$P(i\ non-proband \mid genotype_i) =$$

$$\sum_X P(i\ non-proband \mid phenotype_i = X) P(phenotype_i = X \mid genotype_i)$$

For some models of trait-dependent proband status, computation of the denominator of (8.3.3) can become computationally infeasible.

The situation is worse if the specification of the ascertainment is not individual-specific. As in section 8.2, there clearly is more information if the proband status of every individual is a part of the data. If this is not the case, the numerator (8.3.2) must be also be obtained by summation over all the possible proband events events, with probabilities now conditionally on observed phenotypes, leading to ascertainment (≥ 1 proband), as in equation (8.2.4). As in the simple case, there is no substitute for a predefined sampling and study design with careful specification of the proband status of every individual.

EXERCISE 8.3.1: For simplicity, consider a trait that has, within a particular population studied, no genetic basis or familial pattern; the trait values of family members are independent. The study population consists of a set of predefined pedigrees; each is the set of family members living in a small city, perhaps. A particular pedigree consists of 4 males, (grand-dad, uncle, husband and son, aged 70, 36, 40 and 16), and 3 females, (mum, aunt, and daughter, aged 40, 46, and 10); "mum", "aunt" and "uncle" are full sibs. Males are affected with probability $0.0025 \times age$ and females with probability $0.0050 \times age$. Only affected individuals are potential probands.

(a) A pedigree is ascertained if it contains any affected individuals (truncate ascertainment). What is the probability the pedigree is ascertained? Hint; consider the probability that it does **not** contain any affected individuals.

(b) All affected individuals are ascertained with probability 0.6. What is the probability the pedigree is ascertained? Hint; **careful!** -- you cannot use your answer to (a) directly, but must consider separately the probability that each individual is a proband.

(c) Following (b), the pedigree is ascertained, grand-dad, mum and daughter are observed to be affected, with mum being the only proband. What is the probability of this proband and trait data configuration on the pedigree?

(d) Following (c), the pedigree is ascertained, the phenotypes observed as in (c), but the proband information is lost. How would you now compute the probability that a pedigree with this trait configuration would be ascertained? (Note this probability is required in the numerator of the likelihood; see equation (8.3.1).)

(e) Perhaps due to sampling through a particular workplace, affected men aged 20-50 are ascertained with probability 0.8, affected adult women and older men only with probability 0.2. Those under aged 20 never become probands; presumably being "affected" is not a severe condition. What is the probability the pedigree is ascertained?

(f) A pedigree is ascertained if and only if it contains at least one pair of affected sibs. How would you now compute the probability that the pedigree is ascertained?

NOTE: The set-up of this question is somewhat artificial, but it makes the point of what probabilities are required as components of the likelihood contributed by a particular ascertained pedigree. In (c) note that the probability of any specific trait and proband status configuration is likely to be small; the more specific an event the more *a*priori" less probable it is. This does not mean that the data contribute little to the comparison of alternative hypotheses; in fact, the opposite is true. The less specific an ascertainment event, the more complex the computation of the ascertainment probability. For example, for (f), one would have to sum over all trait configurations in which there is at least one pair of affected sibs; there are 20 such configurations. This computation is possible but tedious; in general, individual-specific ascertainment criteria are much easier to handle.

8.4 Further aspects of the classical approach.

(i) Sibships of variable parental mating type.

In section 8.2, only a reproductively-lethal fully-penetrant recessive trait was considered, and so there was only one possible parental mating type in ascertained families. In section 8.3, the genetic model was general, but then only a general formulation is possible. It is illuminating to consider an intermediate case, where there are several different parental mating types, and in which parental data can provide additional information. Consider, therefore, the case of a partially penetrant simple recessive trait, in which the allele for the trait has population frequency q. For simplicity we shall assume absence of selection, and Hardy-Weinberg genotypic equilibrium. Only B_1B_1 individuals are affected, and each is so with probability η. Thus this genetic model has two parameters, q and η. This model will be used to illustrate several ideas in the following sections.

Table 1; Segregation probabilities for parental mating types

Parental mating type	$B_1B_1 \times B_1B_1$	$B_1B_1 \times B_1B_2$	$B_1B_2 \times B_1B_2$	≥ 1 parent B_2B_2
Population frequency	q^4	$4q^3(1-q)$	$4q^2(1-q)^2$	$1-q^4-4q^2(1-q)$
Prob parents normal*	$(1-\eta)^2$	$(1-\eta)$	1	1
Offspring probabilities				
3 affected	η^3	$(\eta/2)^3$	$(\eta/4)^3$	0
2 affected, 1 not	$3\eta^2(1-\eta)$	$3\eta^2(1-\eta/2)/4$	$3(\eta/4)^2(1-\eta/4)$	0
1 affected, 2 not	$3\eta(1-\eta)^2$	$3\eta(1-\eta/2)^2/2$	$3\eta(1-\eta/4)^2/4$	0
all unaffected	$(1-\eta)^3$	$(1-\eta/2)^3$	$(1-\eta/4)^3$	1

*: For a reproductively lethal trait, weighting by the probability that both parents are unaffected is required.

The key components of the model are given in Table 1. First, there are three parental mating types that can give rise to an affected offspring, and these arise in the population with frequencies dependent upon q. Second, each type gives rise to a different segregation probability, the probability with which each offspring is affected, independently of other offspring. Third there may be data on the parents; only the probabilities that both parents are unaffected are given in Table 1. Finally, dependent on the segregation probability, there is a probability that r of the s offspring siblings are affected; table 8.1 shows these probabilities for $s = 3$ (see EXERCISE 8.4.1).

The general feature of this model is that there are several genotype combinations for the parents who give rise to affected offspring. The parental genotypic frequencies and the segregation probabilities depend on some parameter(s) to be estimated. The posterior probability that an ascertained family with r affected offspring is of parental mating type j is

$$\frac{P(r\ affected \mid type = j)\ P(type = j)}{\sum_l P(r\ affected \mid type = l)\ P(type = l)} \tag{8.4.1}$$

This is of course independent of any ascertainment process that depends only on r. However, the posterior probability that an ascertained family is of type j does depend on on the ascertainment weights, and is

$$\frac{\sum_r c_r\ P(r\ affected \mid type = j)\ P(type = j)}{\sum_l \sum_r c_r\ P(r\ affected \mid type = l)\ P(type = l)}$$

while the probability that an ascertained family has r affected offspring is

$$\frac{\sum_j c_r\ P(r\ affected \mid type = j)\ P(type = j)}{\sum_l \sum_r c_r\ P(r\ affected \mid type = l)\ P(type = l)} \tag{8.4.2}$$

Multiplied over ascertained families, (8.4.2) provides a likelihood that can be used to estimate the genetic parameters q and η provided the ascertainment weights c_r are known. Or, if the c_r are modelled as a function of ascertainment parameters (for example, if $c_r = (1-(1-\pi)^r)$), the genetic parameters can be estimated jointly with the ascertainment parameter π.

As in section 8.2, if we know not only the number of affected individuals in an ascertained family but also the number of probands, we have more information for the estimation of π, and may also be able to test the ascertainment assumptions. In the case of independently and uniformly ascertained probands ($c_r = (1-(1-\pi)^r)$), the posterior probability that an ascertained family of size s has r affecteds and b probands is

$$\frac{\sum_j \binom{r}{b} \pi^b (1-\pi)^{r-b}\ P(r\ affected \mid type = j)\ P(type = j)}{\sum_{r \geq 1} \sum_j (1-(1-\pi)^r)\ P(r\ affected \mid type = j)\ P(type = j)} \tag{8.4.3}$$

Using the likelihood (8.4.3) based on the proband counts is obviously more informative than (8.4.2) based only on the fact of ascertainment; if the number of probands is known it is inadvisable to discard such data. More strongly, as in section 8.2, without data on the numbers of probands, there is little information for estimation for estimation of π, nor for assessing whether the ascertainment model is correct.

EXERCISE 8.4.1: Verify the entries in Table 1.

EXERCISE 8.4.2: Suppose $q = 0.1$ and $\eta = 0.6$.
(a) What are the relative population frequencies of the three parental mating types that can produce affected offspring? What are the relative frequencies among parents with 3 offspring, all affected? What are the relative frequencies among parents of 3 offspring, one of whom is affected?
(b) Repeat (a) for the case when both the parent individuals are known to be unaffected.
(c) Assuming the ascertainment model $c_r = (1 - (1 - \pi)^r)$, and $\pi = 0.2$ what are the relative frequencies of the three parental mating types among ascertained families of size 3.

(ii) Quantitative traits, population distributions, and prior information.

So far we have considered a dichotomous trait; individuals may be affected or non-affected. Affected individuals may or may not become probands. However, often we may be studying a quantitative trait, and/or ascertain through a quantitative trait associated with some disease (e.g. cholesterol levels). In this case, a proband may be an individual whose measure for the quantitative trait exceeds a given threshold. For simplicity we consider the situation of the end of the previous section; ascertainment probabilities are small and probands are uniquely identifiable. Thus the probability that a pedigree is ascertained is the probability that the quantitative value for the potential proband is such as to bring the pedigree into the sample. The likelihood is

$$P_\xi(data) \,/\, P_\xi(X > T) \qquad (8.4.4)$$

where T is the fixed threshold, and X here is the random variable representing the proband's phenotype. As previously, ξ denotes all the parameters of the genetic model. The likelihood (8.4.4) is referred to in the literature as COT ("Conditional On Threshold"); see, for example, Ewens and Green (1988).

In the recent applied literature, the likelihood often used is the probability of the trait values conditional on the observed trait value of the proband (COP = "Conditional On Phenotype");

$$P_\xi(data) \,/\, P_\xi(X = x) \qquad (8.4.5)$$

where x ($x > T$) is the actual quantitative measurement on the proband. This is sometimes more easily computed, but it is not correct, for the event that brought the pedigree into the sample is not that the proband had his observed trait value, x, but that it exceeded a given level T. The correct ascertainment correction in this situation is COT; that is, (8.4.4).

Often it may be desirable to incorporate information on the population distribution of a quantitative trait. This is necessary when a threshold is specified as a percentile of the population distribution. For example, for a normally distributed trait, with population mean μ and variance σ^2, a threshold T specified as the $\gamma \times 100th$ percentile, must satisfy the constraint

$$T = \mu + \sigma\, \Phi^{-1}(1-\gamma) \tag{8.4.6}$$

where Φ is the cumulative normal (Gaussian) distribution function. Although (8.4.4) is the correct likelihood, and (8.4.5) is not, (8.4.5) is more robust to misspecification of the threshold value T, and thence to uncertainties about the mean and variance of the population distribution. Studies have shown that, in many practical cases, (8.4.4) and (8.4.5) are almost equivalent, both in the parameter estimates they provide in particular examples, and, more generally, in the mean square error of the estimators.

The relationship (8.4.6) is sensitive to changes in σ when γ is large, and often we may wish to take probands only from the extreme of the distribution. For example, $\gamma = 0.95$ or even $\gamma = 0.99$ would not be unusual. If there is information on the population parameters, a more natural way to include it is by direct incorporation into the likelihood function. For example, if the trait is normally distributed and information on the population mean and variance derive from a sample of n unrelated individuals, whose sample mean trait value is $\bar{x} = n^{-1} \sum_{i=1}^{i=n} x_i$ and sample sum-of-squares $S^2 = \sum_{i=1}^{i=n} (x_i - \bar{x})^2$, then these contribute a factor

$$(2\pi\sigma^2)^{-n/2} \exp(-(n(\bar{x}-\mu)^2 + S^2)/2\sigma^2) \tag{8.4.7}$$

to the likelihood.

Sometimes, there may be information on the population mean and variance, without the details of the sample that gave rise to that information. For example it might be known that the estimate of the population mean is μ_0 with standard error s_0 and population variance estimate ${\sigma_0}^2$. This information may also be expressed as a likelihood, through a putative sample, that would give rise to this same information. In this particular example, we would use (8.4.7) as a factor in the likelihood, setting $\bar{x} = \mu_0$, $n = \sqrt{S^2/{s_0}^2}$, and $S^2 = n{\sigma_0}^2$. Likelihood inference is well suited to the combination of information from different sources and studies. Any information may be expressed through a likelihood function for the relevant parameters, and likelihoods provided by different sources of information simply combined by multiplication (addition of log-likelihoods).

EXERCISE 8.4.3

Consider a trait observed in spouse pairs. Suppose the population distribution of the (age-adjusted) trait values is bivariate Gaussian (normal), with mean (μ_m, μ_f) marginal variances $({\sigma_m}^2, {\sigma_f}^2)$ and correlation ρ.

(a) Suppose husbands with a trait value K are selected, what is the distribution of trait values in the wives?

(b) Suppose husbands with a trait value $\geq K$ are selected. What is the formula for the bivariate distribution of the trait values in the selected husbands and in their wives?
(c) Suppose husbands with a trait value $\geq K$ are selected, but the individual values for these men are not recorded. Give an expression for the distribution of the trait value in the wives ?
(d) Suppose husbands with a trait value at or greater than the 95th percentile of the male age-adjusted distribution are selected. What is the formula for the distribution of the trait value in the wives ?

(iii) Linkage analysis

The basic ideas of linkage, and estimation of recombination frequencies θ, were introduced in Chapter 2. Linkage analysis will be considered more fully in Chapter 10. Although ascertainment is not generally considered a major concern in linkage analysis there are some important points to consider. Readers who have never encountered linkage analysis may prefer to defer studying this subsection until later.

In linkage analysis, when the trait model ξ is known, ascertainment is not an issue. We can condition on trait-locus (B) phenotypes, without affecting the likelihood for the linkage model θ, so any ascertainment through trait locus phenotypes is irrelevant:

$$P_{\theta,\xi}(A\ phenotypes,\ B\ phenotypes \mid ascertained\ \ pedigree)$$

$$= P_{\theta,\xi}(A - data \mid B - data)\ P_{\xi}(B - data \mid asc.\ ped.)$$

which is proportional, as a function of the linkage parameter θ to

$$P_{\theta,\xi}(A - data \mid B - data) \tag{8.4.8}$$

However, ascertainment cannot be disregarded, even where linkage is primarily of interest. First, we may want to assess (or reassess) the trait model in the light of the linked marker data and linkage analysis. Then ascertainment matters a lot; the disease locus data, whose probability distribution is affected by the ascertainment process, provides the information about ξ, and the likelihood for θ depends on the assumed value of ξ.

Table 2: Marker allele sharing in affected sib pairs ($\psi = \theta^2 + (1-\theta)^2$)

Parental mating type	$B_1B_1 \times B_1B_1$	$B_1B_1 \times B_1B_2$	$B_1B_2 \times B_1B_2$
2 marker alleles shared	1/4	$\psi/2$	ψ^2
1 marker allele shared	1/2	1/2	$2\psi(1-\psi)$
0 marker alleles shared	1/4	$(1-\psi)/2$	$(1-\psi)^2$

Secondly, we may want to use a linkage analysis robust to trait model assumptions, such as haplotypes sharing in sibs (see section 10.7). To be specific, consider again the case of an incompletely penetrant recessive, in sibships of size 3, now with a very highly polymorphic linked marker. Here $\xi = (q, \eta)$. Consider in particular, the likelihood provided by the observation that a pair of affected sibs share two marker alleles. Table 2 shows the marker allele sharing probabilities for two affected sibs, when the recombination fraction between marker and disease gene is θ. (The assumption that the marker is highly polymorphic means that alleles are shared only if they derive from the same parental marker allele.) Now if the posterior probabilities of the three parental mating types are $Q_{11,11}$, $Q_{11,12}$ and $Q_{12,12}$ the probability of the observation is, from Table 2,

$$0.25 \times Q_{11,11} + (\psi/2)Q_{11,12} + \psi^2 Q_{12,12}$$

In families with exactly two affected ($c_0 = c_1 = c_3 = 0$)

$$Q_{11,11} : Q_{11,12} : Q_{12,12} =$$

$$3\eta^2(1-\eta)q^4 : (3\eta^2(1-\eta/2)/4)4q^3(1-q) : (3(\eta/4)^2(1-\eta/4))4q^2(1-q)^2 =$$

$$3(1-\eta)q^2 : 3(1-\eta/2)q(1-q) : 3(1-\eta/4)(1-q)^2/4$$

whereas for families with at least two affected ($c_0 = c_1 = 0$, $c_2 = c_3 = 1$;

$$Q_{11,11} : Q_{11,12} : Q_{12,12} =$$

$$(3\eta^2(1-\eta)+\eta^3)q^4 : (3\eta^2(1-\eta/2)/4 + \eta^3/8)4q^3(1-q) : (3(\eta/4)^2(1-\eta/4) + \eta^3/64)4q^2(1-q)^2 =$$

$$(3-2\eta)q^2 : (3-\eta)q(1-q) : (3-\eta/2)(1-q)^2/4$$

Thus we get a different likelihood if we sample families with exactly two affecteds, or a random pair of affecteds in families with at least 2. In the case where trait and marker are linked, knowledge of the number of unaffected sibs also affects the probabilities of marker allele sharing among the affected ones, again through affecting the posterior probabilities of the parental mating types (see also section 6.6)..

A third reason for caution regarding ascertainment, even when estimating linkage parameters, lies in the practice of selecting pedigrees and parts of pedigrees that are most informative for linkage (see EXERCISES 8.4.4, 8.4.5 below). Where there is reason to suspect an association between a marker and trait, this approach can result in a form of ascertainment bias.

EXERCISE 8.4.4

In a very early discussion of linkage analysis, Fisher (1934b) was concerned to demonstrate that pedigrees exhibiting a dominant disorder selected for linkage analysis should not be subject to ascertainment bias. He wished to select families with certain marker type configurations that, in the days before computers, were particularly amenable to simple analysis. Show the converse of equation (8.4.8), that by conditioning on the selected (biased) marker configurations a valid linkage analysis is possible, provided the pedigrees are not selected with respect to trait phenotypes. Given the choice of biasing selection either of trait phenotypes, or of marker phenotypes, (but not both), without invalidating linkage analysis, which would you prefer, and why?

EXERCISE 8.4.5

A recent linkage analysis of familial combined hyperlipidemia (FCHL) by Wojciechowski et al. (1991) selected families for analysis as follows. (The description is simplified here for easier presentation.) Probands were sampled through elevated lipid levels, though definition of a familial case required at least one other affected first degree relative; in practice therefore the ascertained pedigrees contained an affected sib of the designated "proband". The 16 probands showed a high frequency of a given DNA marker allele, as compared with a control group. Seven of the eight pedigrees in which the proband carried this allele were therefore selected for linkage analysis, to assess linkage between the disease and this same DNA marker locus. The eighth family was omitted because the proband had a hyperlipidemic spouse. Discuss this sampling strategy, with particular reference to any possible bias in the linkage analysis. (I am indebted to Dr. Ellen Wijsman for bringing this example to my attention.)

8.5 A more general view of the sampling problem.

So far we have considered the "classical" set-up of a population of distinct sampling units. In real populations, only individuals and sibships can form distinct units. Nuclear families and more general pedigree structures overlap; an individual is potentially in two nuclear families. The sampling units need not be independent, but each has a probability of ascertainment, dependent upon the number (and maybe pattern) of affected individuals within it. We have obtained the likelihood by considering the probability distribution for the data (phenotypes) within the ascertained units (pedigrees).

Historically, in the development of this theory, it has come to be recognised that the concept of independent probands with equal probabilities of ascertainment does not meet all cases. First, (for example, Elston (1973)) extensions were made to accommodate varying individual probabilities of becoming a proband, but the concept of individual probands remained. Second (for example, Stene (1977)), models were developed that allowed for other forms of the probability of ascertaining a family, as functions of the phenotypic data within it. More recently, rather different views of the sampling procedure have been developed. One of these is the "ascertainment assumption free" approach, which permits the factors leading to ascertainment of a family to be specified only in the broadest terms (Ewens and Shute, 1986; Ewens and Green, 1988). An alternative approach that also permits one to escape from the constraints of "probands" who are "ascertained with some probability" is the sequential sampling view (Cannings and Thompson, 1977; Thompson and Cannings, 1979). In this section, we develop the sequential sampling view. In the following section, we consider the robust "ascertainment assumption free" approach.

It is important to distinguish between a sampling procedure that conditions on or reweights subsets of the population, and random sampling within a defined subset. The "irrelevance of the sampling rule", which is important tenet of likelihood (and Bayesian) inference, refers only to the latter. To take a simple classic example, tossing a coin n times and observing k heads provides the same likelihood for $\zeta = P(head)$ as does tossing it until the k th. head appears and observing that this took n tosses. The former likelihood arises from the binomial distribution for k for given n, and the latter from the negative binomial distribution for n for given k, but in each case the likelihood is proportional to

$$\zeta^k \, (1-\zeta)^{n-k}$$

On the other hand, if we were to restrict attention to coins that had, in independent testing, come up heads 10 times out of 10, we would be weighting the sample in favour of two-headed coins, bent coins, or otherwise heavily biased coins. Inferences about $P(head)$ for such coins can be made as for a randomly chosen coin, but such inferences do not relate to the original population. In order to make inferences about $P(head)$ in the overall population of coins we would have to model the distribution of $P(head)$ and to adjust for the way the coins were chosen.

To relate this to sequential sampling of pedigrees, consider sampling, in an infinitely large population, in a way where the decision to sample relatives depends on phenotypes already observed. For example, we might select random individuals, and if these are affected observe their sibs, and then the offspring of those sibs that are affected, and so on. There is here no "ascertainment" and no conditioning of the likelihood is necessary. Because we have assumed an infinite population, all our pedigrees will have been independently observed, and the likelihood is

$$\prod_i P_\xi(data \ on \ pedigree \ i) \tag{8.5.1}$$

where again ξ denotes all the parameters of the genetic model. Under the particular example scheme above many of our "pedigrees" will consist of unaffected unrelated individuals, each contributing a factor $P_\xi(U)$ to the likelihood, where U here denotes "unaffected". In practice this would be inefficient -- these individuals essentially provide only information on the population frequency of the trait, and this is often known, at least approximately. They provide no additional information of the segregation of the trait, nor for linkage analysis.

Thus we shall, at least, normally wish to initiate sampling of a pedigree from an affected individual, or an individual whose quantitative trait value exceeds a given threshold. In this case we are restricting the sampling frame of reference; it is necessary to condition on the event that initiates study of each pedigree. The required likelihood is now

$$\prod_i \frac{P_\xi(data\ on\ pedigree\ i)}{P_\xi(ascertainment\ event\ \text{for}\ pedigree\ i)} \tag{8.5.2}$$

If sampling of each pedigree is initiated from an affected individual, the "event" denominator, for each term of the product, is the probability that an individual be affected. If it is that individuals are first selected with a quantitative trait value greater than a threshold T, then the denominator is the probability that a phenotype exceeds T. If the event is that one must first find a father and two sons, all affected, then this is the event whose probability is required.

One way to consider the likelihood (8.5.2) is as a version of (8.5.1) where some of the data have been discarded. Suppose we observe random individuals, and initiate sampling the family only where a sampled individual is affected, as above. Suppose now we lose the information on the unrelated unaffected individuals; that is, to obtain the likelihood, we must sum over all the factors $P_\xi(U)$ in (8.5.1). Now

$$\sum_k P_\xi(U)^k = 1/(1 - P_\xi(U)) = 1/P_\xi(affected).$$

Thus, summing over all the sets of unrelated unaffected individuals that intervene before each affected one whose family is then sampled, we obtain

$$\prod_i \frac{P_\xi(data\ on\ pedigree\ i)}{P_\xi(affected)} \tag{8.5.3}$$

a particular case of (8.5.2). Of course, if we had the information on k it would be foolish to discard it, but if we did so we could adjust the likelihood by conditioning on the affected individual who initiates each remaining pedigree, and still have a valid likelihood. In practice, one would be more likely to obtain affected individuals directly from clinical records (for example); the same adjustment of conditioning on that event is required.

As a simple explicit example of a sequential sampling rule, consider again our example of sibships of size 3, and the following rule. An individual is sampled. If the individual is affected, a sib of the individual is sampled. If, and only if, this sib is **not** affected, the third member of the sibship is sampled. This might be appropriate rule if we wished to obtain affected sib pairs for typing for linkage analysis, as described above; in those sibships where two affected sibs had been obtained by the rule, these could be typed, and their haplotype sharing analyzed. There are four possible observations on a sibship under this sampling rule; these, and their probabilities, are given in Table 3. Note that the relative frequencies of these four occurrences are precisely the unconditioned probabilities of the ordered observation on the sibship. Note also that this example exemplifies the two critical rules of sequential sampling. First, we cannot "look ahead" and sample the third sibs only in those cases that will provide a second affected sib. To do so might be more efficient with a view to

Table 3; Probabilities of observations under sequential sampling

Parental mating type	$B_1B_1 \times B_1B_1$	$B_1B_1 \times B_1B_2$	$B_1B_2 \times B_1B_2$	≥ 1 parent B_2B_2
Population frequency	q^4	$4q^3(1-q)$	$4q^2(1-q)^2$	$1-q^4-4q^2(1-q)$
Observation;				
1 normal only	$1-\eta$	$1-(\eta/2)$	$1-(\eta/4)$	1
2 affected only	η^2	$\eta^2/4$	$(\eta/4)^2$	0
affected, normal, affected	$\eta^2(1-\eta)$	$\eta^2(1-\eta/2)/4$	$\eta^2(1-\eta/4)/16$	0
affected, normal, normal	$\eta(1-\eta)^2$	$\eta(1-\eta/2)^2/2$	$\eta(1-\eta/4)^2/4$	0

Combined probabilities:
P(1 normal only) = $1-q^2\eta$ = 1 - P(affected)
P(2 affected) = $\eta^2q^2(1+q)^2/4$ = P(pair of sibs affected)
P(affected, normal, affected) = P(2 affected in sibship size 3)/3
P(affected, normal, normal) = P(1 affected in sibship size 3)/3

the analysis of haplotype sharing, but elimination of the fourth possibility would mean that the probabilities of the observed configurations of affecteds would no longer be the unconditioned probabilities. Second, we cannot "discard" any observations that we have made without adjusting the likelihood for having done so (as in (8.5.3) above). The second and third observation possibilities each provides two affected sibs, but the act of observation of a normal sib in the third case affects not only the likelihood for the parameters q and η, but also the probabilities of haplotype sharing for the affected sibs.

"No look-ahead" and "no discards" are essential rules if the likelihood is to be valid. Clearly any individual included or not included because of his/her own phenotypic status, or non-incorporated information related to that phenotypic status, will alter the probability of the observed data configuration under the model; i.e. will alter the likelihood.

It is also important to recognise that although the likelihoods may be the same under sequential sampling as under a single-ascertainment correction for the initial proband, this does not mean that the probability distributions are the same -- only that they are proportional as a function of the parameters. For example, in the simplest coin-tossing example, the number of head in n tosses is binomial, but the number of tosses up to and including the k th, head is negative binomial. In the former case k is $B(n,\zeta)$; in the latter n is $NB(k,\zeta)$. Thus variances of estimators, and power and size of tests, may differ. Where tests are based on the presumed asymptotic properties of maximum likelihood estimators and log-likelihood differences, then (provided the presumption is accurate) the same tests, estimates, standard errors, and P-values will result from either the sequential or single-ascertainment sampling models, but any exact finite-sample based test will differ.

In the above examples, we have discussed sequential sampling of a pedigree as starting from a single affected individual, or an individual of given phenotypic characteristics such as having an above-threshold value of a quantitative trait. In (8.5.2), the denominator is simply the probability that a random individual has the required phenotype, or exceeds the required threshold (COT). That is, (8.5.2) is the same likelihood as would be obtained under a classical single ascertainment assumption. (Recall that likelihood functions are defined only up to constants of proportionality. There may be several affected individuals in the pedigree who could be the single proband under a single-ascertainment model, but this does not affect the likelihood; see the factors of 3 in the footnotes to Table 3.) The derivation of (8.5.2) does assume an infinite population; thus every pedigree will be initiated from a single event. In this sense, it is "singly ascertained". However, the argument leading to (8.5.2) is not that of the classical single ascertainment model. There is no concept of probands being randomly ascertained.

We could assume a complex multiplex ascertainment event. For example, a physician might report "an affected grandparent, uncle and two offspring". An extensive pedigree might then be sampled sequentially because of this trait information. The sequential sampling adjustment would be the same as the single-ascertainment correction for this multiplex ascertainment event; the denominator would be the probability of this event. However, it would be an unusual study that demanded, in advance, "an affected grandparent, uncle and two offspring" as an event that would cause ascertainment of a family; disease registries do not normally come in this form!

Although the sequential sampling approach is sometimes considered a version of classical single ascertainment, it should be recognised that this is only true in the sense that the assumption of an infinite population ensures that each pedigree sequentially sampled is singly initiated -- pedigrees cannot "join up" in sampling. It differs from classical (single) ascertainment in that there is no concept of "probands coming to attention with probability π"; the process by which the initiating individual (or set of individuals) comes to attention need not be modelled. Under the classical model, the number of probands, and hence the number of families sampled, is random -- in the simplest case binomial $B(N, \pi)$, where N is the number of affected individuals in the population. Were this model valid, there would be information in the number of families, jointly about π and the genetic parameters which determine N. However, normally this information is disregarded; wisely, since on a population level the simple ascertainment model can seldom be valid. The classical likelihood thus conditions on the families ascertained, and this conditioning makes the single ascertainment likelihood identical to the sequential sampling likelihood adjusted for a single initiating event. This classical example of conditioning, implicit in the likelihoods of section 8.2, is of interest in the context of the recent development of ascertainment-robust conditional likelihoods that will be discussed in the next section.

Unfortunately, real populations are finite, and with regard to individuals affected with a rare trait may in fact be quite small. Thus in practice, pedigrees can be multiply independently ascertained, and pedigrees sampled sequentially from separate instances of an initiating event may be found to join into a single pedigree. Under a model for the structures, sizes and numbers of pedigrees on the population, a genetic model for the trait, and an ascertainment model, the joining (or failure to join) of pedigrees sampled sequentially, conveys information. However, it is not practicable to construct the valid "hypermodel" that would allow this information to be extracted. As for the classical proband-models of ascertainment, it is not practicable to do other than condition on the pedigree structures sampled. However, what then is the initiating event for which the likelihood of the combined pedigree should be adjusted? Recent practice has been to condition on the joint phenotypes and relationships of the combined initiating events. For example, if sampling is initiated from affected father-son pairs, and the pedigrees join because the fathers are first cousins, the combined event will be of two first-cousin affected males, each with an affected son. In the next section, we shall see that it is important to condition on at least the phenotypic information that is responsible for the sampling of the pedigree. Thus recent standard practice seems appropriate, but the optimal procedure remains an open question.

EXERCISE 8.5.1: Individuals are affected independently, each with probability ζ.
(a) Three individuals are sampled at random, are are found to be Not affected, Not affected and Affected, respectively. What is the likelihood for ζ, what is the MLE of ζ, and what is the observed information (the value of the negative second derivative of the log-likelihood at the MLE)?
(b) Individuals are sampled sequentially until an affected individual is sampled; three individuals are observed which must therefore be Not affected, Not affected and Affected, respectively. What is the likelihood for ζ, what is the MLE of ζ, and what is the observed information (the value of the negative second derivative of the log-likelihood at the MLE)?
(c) For the experiment of (a), what is the expected (Fisher) information as a function of ζ?
(d) For the experiment of (b), what is the expected (Fisher) information as a function of ζ? For what true values of the parameter ζ would you prefer this experiment to that of (a)?
NOTE: In a real example, one would sample many more than three individuals if sampling binomially, and until many more than one affected had been found if sampling sequentially. The example here is used only to illustrate the general points made in this section.

EXERCISE 8.5.2: Consider sampling within sibships of size 3, using the same scheme and model considered in this section (Table 3). Verify the statements made following Table 3:
(a) that the probability of the configuration consisting of a single unaffected individual is, under this sampling scheme, the same as the probability that a randomly chosen member of the population is unaffected,
(b) and that the probability of the configuration consisting of a pair of affected sibs is, under this sampling scheme, the same as the probability that a randomly chosen pair of sibs chosen from the

population are both affected.

EXERCISE 8.5.3: For simplicity assume now a model in which half the sibships of the population contain no affected individuals, and in the other half sibs are independently affected with probability ζ.

(a) 128 sibships are sampled; 91 are observed to have no affected, 27 one affected, 9 two affected and 1 three affected. Plot the log-likelihood as a function of ζ. Note the maximum is at $\zeta = 0.25$; the sample is precisely the expected sample for this ζ value. For easier comparison of sampling strategies, plot the difference between the log-likelihood and its value at $\hat{\zeta} = 0.25$. Plot also this log-likelihood difference divided by the total number of individuals observed -- 128×3.

(b) 128 sibships are sampled, using the sequential scheme of Table 3. 112 of the 128 observations consist of a single unaffected individual (N), there are 4 pairs of affected sibs (AA), 3 consist of the ANA configuration, and 9 of the ANN configuration. Plot the log-likelihood as a function of ζ. Note the maximum is at $\zeta = 0.25$; the sample is precisely the expected sample for this ζ value. On the same graphs as (a). plot the difference between the log-likelihood and its value at $\hat{\zeta} = 0.25$, and also this log-likelihood difference divided by the total number of individuals observed -- $112 + 8 + 9 + 27 = 156$.

(c) What do you conclude about the relative performance of these two sampling schemes?

NOTE: To assist you, the graphs requested in (a) and (b) are shown in figure 1. Figure 1a shows the log-likelihoods for values of ζ from 0.05 to 0.75; the random sampling scheme is shown as points with lines, and the sequential scheme as a smooth curve. Figure 1b shows the two log-likelihoods divided by the sample sizes.

Note also that performance can be assessed by the curvature of the log-likelihoods. The more information provided by the sample, the sharper the log-likelihood curve. Finally note that the samples assumed in this exercise are the counts expected under the model. In section 8.6 we will address more formally the use of the expected log-likelihood curve to assess information, and the fact that in many examples (such as this one), consideration of the expected log-likelihood is equivalent to consideration of the log-likelihood provided by the expected sample.

8.6. Robust approaches to ascertainment; qualitative phenotypes.

In practice, it is often impossible to specify retrospectively the ascertainment process. If the form of ascertainment can be defined up to values of unknown parameters (e.g. π) then those parameters can, in principle, be estimated jointly with the parameters of the genetic model. However, there may be questions even as to the appropriate form of ascertainment model. Thus, it is desirable to have methods of analysis that are robust to this aspect of a study. This can be achieved by conditioning on aspects of the data that may have have caused ascertainment of a particular pedigree or set of pedigrees. Any conditioning decision is a balance between efficiency and robustness.

To condition reduces information, in the technical statistical sense, and so increases the standard errors of estimators and reduces the power of statistical tests. However, a test or estimator based on a conditional likelihood is robust to misspecification of the model for those data observations that are conditioned upon. In order to obtain asymptotically unbiased, $n^{\frac{1}{2}}$-consistent, asymptotically normal maximum likelihood estimators (Chapter 4) only validity of the conditional probability distributions is required. Further, these MLEs based on conditional likelihoods will satisfy the other general asymptotic optimality properties of MLEs (Chapter 4). In particular, they are efficient, in the sense that they extract all information in the conditional probability distributions; obviously they cannot extract the information lost by conditioning.

The information about a model that may be contained in data observations can derive not only from the phenotypes of individuals, but also from the structures, and even number, of pedigrees ascertained. However, use of this information requires a valid model for the ascertainment of these pedigree structures, and may require estimation of nuisance parameters, such as parameters of population structure. It also requires assumptions that the existence and availability of relatives are not influenced by the trait itself; for many diseases the assumption of population levels of survival and fertility for affected individuals may be unrealistic. If any aspect of the ascertainment model is incorrect, then estimates of the genetic parameters may be biased, even asymptotically. Even where the model is correct, the information about genetic parameters available from some aspects of the data may be slight, while inclusion of the part of the likelihood based on those aspects of the data may involve estimation of many nuisance parameters. Thus it is often better to condition on parts of the data.

Thus, the conditioning in the classical likelihood formulation on the number, sizes and structures of pedigrees observed can be understood as an implicit recognition that these are data affected by ascertainment in uncertain ways, although, within the context of a valid ascertainment model for probands, there is information in the sizes and structures of ascertained families. For example, the probability that a family of given size is ascertained is a function of the genetic and ascertainment parameters. Thus, given a prior family size distribution in the population (perhaps up to unknown parameters also to be estimated) the sizes of ascertained families can provide information on ascertainment probabilities, and on genetic parameters where these are jointly estimated. Likewise, the number of unaffected sibs can provide information in affected sibling methods for linkage or segregation analysis. Usually, the additional information provided by such aspects as pedigree structure and family size is slight, and considerations of robustness suggest that conditioning on the observed pedigrees is the wiser practical approach.

Caution is necessary in the extent of conditioning. One naive criticism that has sometimes been made of the whole process of likelihood inference is that the maximum likelihood model is necessarily the one that accords probability one to the observed data. This criticism denies the whole framework of parametric models, of experiments, of studies, and of samples. It would maintain that no matter what is observed in any situation has no relevance to any other. While the process of scientific inference dismisses such an extreme view, arguments in the ascertainment area sometimes approach it. If indeed a pedigree comes to attention because of the specific pattern of phenotypes of all the individuals within it, then clearly that pattern conveys no information on causal genetic mechanisms. In that case, the relevant ascertainment model would indeed specify that all ascertained pedigrees must have that pattern of phenotypes; the probability of that phenotypic pattern in any ascertained pedigree would be one. Such a pedigree could be used for the study of other traits, or for linkage analysis if a trait model were known, but no trait model inferences would be possible. Of course, we seldom believe that phenotypes must be *precisely* as observed, in order for a pedigree to have come to attention. Indeed a study will normally encompass several pedigrees that do not exhibit *precisely* the same patterns. Nonetheless, given the practical difficulties in defining ascertainment criteria, it is advisable to keep this *reductio ad absurdum* in mind, when tempted to condition too extensively on precisely the information that was designed to address the scientific questions of a genetic epidemiological study.

The conditional likelihood approach recently developed to provide robust approaches to the estimation of genetic parameters avoids the necessity to consider the proband status of individuals. Instead the ascertainment process is formulated as follows. The data X on the family or pedigree are divided into two sets; $X = (X_1, X_2)$. The part X_2 must be irrelevant to the ascertainment of the family, while the part X_1 must include at least all the phenotypic information relevant to ascertainment of the family. However, the ascertainment process need not be specified in detail; the ascertainment probability can be an arbitrary function of the pedigree structure and size and of X_1. The conditional likelihood considered is then

$$L_C(\xi) = \prod_{pedigrees} P_\xi(X_2 \mid X_1 \,\&\, asc.) = \prod_{pedigrees} \frac{P_\xi(X_2 \,\&\, X_1 \,\&\, asc.)}{P_\xi(X_1 \,\&\, asc.)} \qquad (8.6.1)$$

$$= \prod_{pedigrees} \frac{P_\xi(asc. \mid X_2 \,\&\, X_1)\, P_\xi(X_2 \,\&\, X_1)}{P_\xi(asc \mid X_1)\, P_\xi(X_1)} = \prod_{pedigrees} \frac{P_\xi(X_2 \,\&\, X_1)}{P_\xi(X_1)} = \prod_{pedigrees} P_\xi(X_2 \mid X_1)$$

The fourth equality holds provided $P_\xi(asc. \mid X_2 \,\&\, X_1) = P_\xi(asc. \mid X_1)$; that is, only the variables X_1 are relevant to ascertainment. Thus the likelihood is based on the data on the ascertained pedigrees, conditioning on any data that may have been relevant to their ascertainment. No specific ascertainment model is required. This conditional likelihood will provide consistent MLEs, efficient in that all the information on ξ in X_2 is utilised; the estimators will have asymptotically minimum variance among estimators based on the subset of the data X_2.

While the above formulation is general, quantitative traits with a threshold form of ascertainment lead to different specific formulae from the case of qualitative traits. For simplicity we consider in this section only dichotomous traits (affected/not affected), and defer quantitative traits and more general considerations to the next section. The examples most extensively considered are nuclear family analyses, where ascertainment is known to be through the offspring. Then X_1 is taken to be the offspring trait phenotypes, or, equivalently the number of affected (and hence also of unaffected) offspring.

More generally, if ascertainment is known to be dependent on the number of affected children within a given age range, X_1 can be restricted to the counts for the children in that age range. Or, in linkage examples, one can condition on *all* the trait information, and consider X_2 to be only the marker data (see equation (8.4.8)), but of course all information on the trait segregation is then lost. It is important to condition on at least the information relevant to ascertainment. Conditioning on parental phenotypes has sometimes been recommended as a way to correct for trait-based fertility biases. However this is only correct if ascertainment is through the parents. If ascertainment is through the offspring, conditioning on parental phenotypes (and not on the numbers of affected and unaffected offspring) can increase asymptotic biases.

Table 4: Frequency and ascertainment probability of the phenotypic configurations of Table 1, when only the two elder sibs can be ascertained.

Offspring phenotypes (a, b); see note 2	Population probability (see note 1)	Ascertainment probability
(2,1)	$q^2\eta^3(4q(1-\eta)+(1-q))^2/16$	$2\pi-\pi^2$
(2,0)	$q^2\eta^2(16(1-\eta)^3q^2+8(1-\eta)(2-\eta)q(1-q)+(4-\eta)(1-q)^2)/16$	$2\pi-\pi^2$
(1,1)	$q^2\eta^2(16(1-\eta)^3q^2+8(1-\eta)(2-\eta)q(1-q)+(4-\eta)(1-q)^2)/8$	π
(1,0)	$q^2\eta(16(1-\eta)^4q^2+8(1-\eta)(2-\eta)^2q(1-q)+(4-\eta)^2(1-q)^2)/8$	π
(0,1)	$q^2\eta(16(1-\eta)^4q^2+8(1-\eta)(2-\eta)^2q(1-q)+(4-\eta)^2(1-q)^2)/16$	0
(0,0)	$q^2(16(1-\eta)^5q^2+8(1-\eta)(2-\eta)^3q(1-q)+(4-\eta)^3(1-q)^2)/16$ $+(1-q^4-4q^2(1-q)$	0

Note 1: The trait is assumed reproductively lethal (i.e. the parents have Normal phenotype).

Note 2: In the notation (a, b) for the offspring trait configuration, $a = 0, 1, 2$ is the number of affected sibs among the two elder sibs, and $b = 0, 1$ is the number of affected among the youngest of the three sibs.

As an explicit simple example of a robust approach to ascertainment corrections, consider again the families of three sibs. In order to have some flexibility for conditioning, assume first that the sibs are ordered, and that only the two elder sibs can become probands; the third sib has ascertainment probability 0. The simplest fully specified ascertainment model might then be that each of the two elder sibs, if affected, becomes a proband with probability $\pi > 0$, and the sibship enters the study if it contains at least one proband. Thus the family types are as in Table 1, but a family with

Table 5: Expected proportions of given combinations of probands and affected in ascertained families, for a lethal trait, under the model of Table 1 and the ascertainment of Table 4, when $\pi = 0.2$, $q = 0.1$ and $\eta = 0.6$. $P_\xi(a, b)$ is the population probability of the configuration (a, b) under the model $\xi = (q, \eta)$ of Table 4.

	Population probabilities of ascertainable configurations			Population probabilities	Probabilities in ascertained families	
	total	2 probands	1 proband	total	2 probands	1 proband
(2,1)	$P_\xi(2,1)$	$P_\xi(2,1)\pi^2$	$2P_\xi(2,1)\pi(1-\pi)$	0.0001517	0.0029	0.0233
(2,0)	$P_\xi(2,0)$	$P_\xi(2,0)\pi^2$	$2P_\xi(2,0)\pi(1-\pi)$	0.0007127	0.0137	0.1093
(1,1)	$P_\xi(1,1)$		$P_\xi(1,1)\pi$	0.0014253		0.1367
(1,0)	$P_\xi(1,0)$		$P_\xi(1,0)\pi$	0.0074491		0.7142

three affected offspring only has probability $1-(1-\pi)^2 = 2\pi - \pi^2$ of ascertainment. The ascertainment probabilities for the families with 2 and 1 affected are likewise modified, and a distinction is now necessary between the phenotypic status of the youngest sib and that of his elders (Table 4). In this example, we shall assume the trait is reproductively lethal, so the parents are known to be unaffected and the probabilities in Table 4 include the parental penetrance probabilities (see Table 1). As a specific example, we assume, as previously, that the true value of π is 0.2, and that $q = 0.1$, $\eta = 0.6$. Then Table 5 shows the form of the probabilities, and their numerical values, for the population probabilities of ascertainable configurations and of the conditional probabilities of ascertained families with the given patterns of affecteds and probands.

Suppose, however, that we are unwilling to assume independence and equal probability for the ascertainments of the two elder sibs. We might nonetheless know, from the structure of the study, that ascertainment was dependent only on the number of affected among those two sibs. The most robust approach would then be to condition on this count. No precise ascertainment process need now be modelled, but only the third sib provides information about the genetic model. In the population at large, the probability that the third sib is affected given that the first two are so is

$$P_\xi(2,1)/(P_\xi(2,1) + P_\xi(2,0)) \quad or \quad 1517/(1517+7127) = 0.1755$$

in the numerical example, while the probability that the third sib is affected given that only one of the first two is so is

$$P_\xi(1,1)/(P_\xi(1,1) + P_\xi(1,0)) \quad or \quad 14253/(14253+74491) = 0.1606$$

in the numerical example. These are precisely the same as the probabilities within the ascertained sample, despite the fact that these are differentially weighted by functions of the ascertainment probabilities π. For example; where the elder two sibs are affected

$$(0.0029 + 0.0233)/(0.0029 + 0.0233 + 0.0137 + .1093) = 0.1756$$

(i.e. 0.1755 to within rounding error of the conditional probability computations). The same will be true regardless of the form of the ascertainment probability, provided ascertainment is dependent only on the phenotypes of the two eldest sibs; the mathematically inclined reader may verify this by inserting an arbitrary ascertainment probability function $\pi(a)$ for the elder-sibs phenotypic configuration a into the formulation of Tables 4 and 5.

Although the correct conditional probabilities are obtained, information is lost by considering only the type of the third sib, conditional upon the elder two. Suppose we have a sample of 10,000 ascertained families and observe them in accordance with expectation under the ascertainment model; that is, we observe 29, 233, 137, 1093, 1367 and 7142 in each of the six categories (see Table 5). For simplicity, we take $q = 0.1$ as known, and consider the log-likelihoods as a function of η for $\pi = 0.2$ (figure 2a) and as a function of π for $\eta = 0.6$ (figure 2b). In practice, of course joint estimation of these parameters (and q also) would be required. In figure 2a each of the three curves has its maximum at $\eta = 0.6$. The flatter curve corresponds to the conditional approach, conditioning on the types of the elder sibs. The other two curves are identical to each other; one corresponds to the case where proband information is discarded (recall section 8.2), and the third to full use of the six observations, assuming the model to be true. We see that discarding the proband count is irrelevant to this particular estimation when the correct values of π is assumed known, but that conditioning on the phenotypes of the two eldest sibs loses much of the information about η. In figure 2b, each curve has its maximum at $\pi = 0.2$. The conditional likelihood is flat in this dimension; this likelihood does not depend on π. Note that in this dimension the log-likelihood which discards information on the proband counts is also almost flat; as discussed in EXERCISE 8.2.2 almost all the information on π is lost if these data are discarded.

Table 6: Sibship configuration probabilities in sibships of size 2 when sibs are independently affected, but have different probabilities of being affected.

Configuration	Probability	Number affected	Conditional probability
1 1	$\zeta_1\zeta_2$	2	1.0
0 1	$(1-\zeta_1)\zeta_2$	1	$\dfrac{(1-\zeta_1)\zeta_2}{(\zeta_1+\zeta_2-2\zeta_1\zeta_2)}$
1 0	$\zeta_1(1-\zeta_2)$	1	$\dfrac{(1-\zeta_2)\zeta_1}{(\zeta_1+\zeta_2-2\zeta_1\zeta_2)}$
0 0	$(1-\zeta_1)(1-\zeta_2)$	0	1.0

A similar example arises where one is willing to assume that the ascertainment probability of a sibship is a function c_k of the number of affected sibs k (see section 8.1) but unwilling to assign a specific functional form (such as $(1-(1-\pi)^k)$) to c_k. In this case, a robust approach is to condition on the number k of affected sibs. If only the sibs themselves are observed, and if all sibs have equal probability of being affected, all information is lost by this conditioning. However, if there are covariate effects on disease status and these covariates are known, information remains in the data on precisely which of the sibs are affected. In this case, the likelihood contributed by a sibship is

$$\frac{P(\ observed\ trait\ status\ of\ each\ sib \mid covariate\ information)}{\sum_{C_k} P(C_k\ trait\ status\ of\ each\ sib \mid covariate\ information)} \tag{8.6.2}$$

where summation is over the total set of trait configurations C_k in which precisely k sibs are affected. This points to a possible additional practical problem with this approach; as for the classical pedigree ascertainment correction (8.3.1), determination of the denominator can be computationally intensive or even infeasible.

We consider, therefore, only the simplest possible example. Suppose all sibships are of size two, and suppose that the elder sib is affected with probability ζ_1 while the two younger one is affected with probability ζ_2. The probabilities of the four data configurations, and the probabilities conditional on the number of affected sibs, are given in Table 6. Sibships with no affected sibs, or with two, provide no information; the trait configuration cannot be separated from the uncertain ascertainment probabilities c_0 and c_1. Sibships with two one affected member still provide information; the information loss due to the conditioning will be considered further below, as an application of a more general formulation.

8.7. Quantitative traits and the analysis of expected log-likelihoods.

It is possible to consider more formally the biases due to incorrect modelling of ascertainment, and the information loss due to unnecessary conditioning. As for the general properties of likelihood estimators (Chapter 4), the key lies in the expected log-likelihood function and its derivatives. Suppose that the family of distributions used to model the data is $\{f(X;\xi)\}$, while the true probability distribution giving rise to the data is $g(X)$. Then the expected log-likelihood function is

$$M_g(\xi) = \mathbb{E}_g(log\ f(X;\xi)) \tag{8.7.1}$$

If $g = f(.\,;\xi_0)$ then (8.7.1) is maximised w.r.t. ξ at the true parameter value ξ_0. However, if the family $f(.\,;\xi)$ does not include the true distribution, (8.7.1) will be maximised at some ξ_1. Asymptotically, as more data are collected under the model g, the MLE under the assumed family of models will converge to ξ_1 with probability one. By evaluating (8.7.1), and obtaining the value of ξ_1 for the genetic parameters of interest, we can examine the asymptotic bias caused by model misspecification.

We can also now see why conditioning on at least those variables that lead to ascertainment of a sibship is sufficient to ensure consistency of estimates, and hence absence of asymptotic bias.

Using again the notation of equation (8.6.1), for given values of the relevant ascertainment variables X_2, the relative probabilities of data configurations $(X_1, X_{2)}$ conditional upon ascertainment are precisely the population probabilities $f(X_1|X_2;\xi)$. The conditional expected log-likelihood

$$\mathbb{E}_{\xi_0}(\log f(X_1|X_2;\xi) \mid X_2) \tag{8.7.2}$$

is maximised at $\xi = \xi_0$ for every X_2. Taking expectations over X_2, the overall expected log-likelihood

$$\mathbb{E}_{\xi_0}(\log f(X_1|X_2;\xi))$$

is maximised at $\xi = \xi_0$ regardless of the distribution of X_2. Thus, regardless of how the ascertainment process has distorted distribution of X_2 in the sample from that in the population, consistent estimates will be obtained.

The question of information can be considered similarly. Recall (Chapter 4) that the asymptotic variance-covariance matrix of the MLE is the inverse of the Fisher information matrix -- the matrix of expected negative second derivatives of the log-likelihood. In this chapter, rather than computing variances of estimators, we have preferred to consider the shape of the likelihood surface. The (observed) information matrix is the negative second derivatives of the log-likelihood at the MLE; that is, the curvature of the likelihood surface at its maximum. A highly peaked likelihood has high curvature at its maximum and provides precise estimates. A flat likelihood has low curvature; estimates of the parameters have higher variance. Thus, in this section, we shall evaluate expected information rather than asymptotic standard errors; that is, we consider the Fisher information matrix.

Suppose that the family of models contains the true probability distribution, $f(.\,;\xi_0)$, then the Fisher information matrix for the vector of parameters ξ is

$$\mathbb{E}_{\xi_0}\left(-\frac{\delta^2 \log f(X;\xi)}{\delta \xi^2} \right) = \mathbb{E}_{\xi_0}\left(\left(\frac{\delta \log f(X;\xi)}{\delta \xi} \right) \left(\frac{\delta \log f(X;\xi)}{\delta \xi} \right)^T \right) \tag{8.7.3}$$

A conditional likelihood, conditioning on data event C, has the form

$$L_C(\xi) = f(X \mid C;\xi) = \frac{f(X;\xi)}{P_\xi(C)}$$

Obviously the observed data, X, must satisfy the condition C. Thus,

$$\frac{\delta^2 \log f(X;\xi)}{\delta \xi^2} = \frac{\delta^2 \log L_C(\xi)}{\delta \xi^2} - \frac{\delta^2 \log P_\xi(C)}{\delta \xi^2} \tag{8.7.4}$$

Taking expectations w.r.t. the true probability distribution, the loss of information due to conditioning becomes

$$- \mathbb{E}_{\xi_0} \frac{\delta^2 \log P_\xi(C)}{\delta \xi^2}$$

which can be evaluated in any specific example. Of course, if the probability of the event C does not depend on ξ then the statistics in the conditional distribution are sufficient for ξ, the event C is ancillary, and conditioning on it does not lose any information.

Another debate that can now be better understood in the context of conditioning is that of the COP and COT ((8.4.4) and (8.4.5)) alternatives for single ascertainment in the context of a quantitative trait value exceeding a threshold, T. Recall the two likelihoods:

$$L_{COT}(\xi) = \frac{P_\xi(data)}{P_\xi(X > T)}$$

and

$$L_{COP}(\xi) = \frac{P_\xi(data)}{P_\xi(X = x > T)}$$

where X is the trait value random variable and x is the value for the single identified proband. L_{COT} conditions on only the minimum information for ascertainment; i.e. that X exceeds the threshold. L_{COP} conditions on more specific data; i.e. that $X = x$. The ratio of the two likelihoods is the probability density of the proband's trait value, conditional upon its exceeding the threshold. The situation is not exactly as developed in (8.6.1) and (8.7.2); both the COP and COT likelihoods condition on some aspect of the proband trait value X rather than COP conditioning on additional variables. Nonetheless, COP involves an additional conditioning, providing an extra term in the (expected) log-likelihood. This term, the expectation of the log of the density of the proband's trait value conditional on exceeding the threshold, is also maximised at the true parameter values (ξ_0 in the above notation). Thus both COP and COT provide consistent estimators, but COP loses information. Following (8.7.4), we see that the information lost is precisely the information in this conditional density of proband trait values. Thus the information lost is normally very slight, and COP is robust to the specification of T. Note that this derivation is specific to the single ascertainment case. It applies also to single multiple ascertainment; that is, conditioning either on the proband values or on their exceeding a threshold will provide consistent estimates. However, for multiple ascertainment the ratio of the two likelihoods is no longer a simple probability density; Ewens and Green (1988) give a discussion of this case.

Many of the likelihoods arising in genetic epidemiology are multinomial likelihoods; one observes certain numbers of families with certain numbers or patterns of affecteds, the probability of the *i th* such pattern being some complex function $p_i(\xi)$ of the vector of genetic and ascertainment parameters, ξ. The theoretical properties of such likelihoods have, in the genetic epidemiological ascertainment context, been studied by assuming so-called "deterministic data" or "perfect data"

-- perhaps better called "expected data". The form of the likelihood is $\sum_i n_i \log p_i(\xi)$, and the expected number of families of data category i is $\mathbb{E}_{\xi_0}(n_i) = n\ p_i(\xi_0)$. Thus in assuming data $n_i = n\ p_i(\xi_0)$ and studying the resultant log-likelihood one is analysing precisely the expected log-likelihood function (8.7.1). In the examples of figure 1 (EXERCISE 8.5.3) and figure 2, note that the assumed data were indeed the expected counts under the assumed models; the curves plotted were precisely the expected log-likelihood curves for those models and sampling procedures.

An alternative approach is to consider the Fisher Information directly. The second derivative of the log-likelihood is

$$\sum_i \left(\frac{n_i}{p_i(\xi)} \frac{\delta^2 p_i(\xi)}{\delta \xi^2} - \frac{n_i}{p_i(\xi)^2} \left(\frac{\delta p_i(\xi)}{\delta \xi} \right)^2 \right)$$

Taking expectations, n_i is replaced by $n\ p_i(\xi)$, the first term becomes 0, and the Fisher information is

$$n \sum_i \left(\frac{1}{p_i(\xi)} \left(\frac{\delta p_i(\xi)}{\delta \xi} \right)^2 \right) \tag{8.7.5}$$

To illustrate the use of equations (8.7.4) and (8.7.5) we return to the example of Table 6. In the unconditional formulation, application of (8.7.5) to the first column of probabilities in Table 6 gives the Fisher information matrix as

$$\begin{pmatrix} \zeta_2/\zeta_1 & 1 \\ 1 & \zeta_1/\zeta_2 \end{pmatrix} + \begin{pmatrix} \zeta_2/(1-\zeta_1) & -1 \\ -1 & (1-\zeta_1)/\zeta_2 \end{pmatrix} + \begin{pmatrix} (1-\zeta_2)/\zeta_1 & -1 \\ -1 & \zeta_1/(1-\zeta_2) \end{pmatrix}$$

$$+ \begin{pmatrix} (1-\zeta_2)/(1-\zeta_1) & 1 \\ 1 & (1-\zeta_1)/(1-\zeta_2) \end{pmatrix} \tag{8.7.6}$$

This reduces to

$$\begin{pmatrix} \dfrac{1}{\zeta_1(1-\zeta_1)} & 0 \\ 0, & \dfrac{1}{\zeta_2(1-\zeta_2)} \end{pmatrix}$$

which from binomial theory is the correct result for observation of two independent sibs, affected or unaffected with probability ζ_1 or ζ_2. The count of affected sibs takes values 2, 1, and 0 with probabilities $p_2 = \zeta_1\zeta_2$, $p_1 = (\zeta_1 + \zeta_2 - 2\zeta_1\zeta_2)$, and $p_0 = (1-\zeta_1)(1-\zeta_2)$, and the information matrix for this distribution can be evaluated again using (8.7.5), and is found to be

$$\begin{pmatrix} \zeta_2/\zeta_1 & 1 \\ 1 & \zeta_1/\zeta_2 \end{pmatrix} + \frac{1}{\zeta_1 + \zeta_2 - 2\zeta_1\zeta_2} \begin{pmatrix} (1-2\zeta_2)^2 & (1-2\zeta_1)(1-2\zeta_2) \\ (1-2\zeta_1)(1-2\zeta_2) & (1-2\zeta_1)^2 \end{pmatrix}$$

$$+\begin{pmatrix} (1-\zeta_2)/(1-\zeta_{1)} & 1 \\ 1 & (1-\zeta_1)/(1-\zeta_2) \end{pmatrix} \tag{8.7.7}$$

In the conditional distribution, all information from sibships with 2 or 0 affected sibs is lost; the first and last terms of (8.6.9) are precisely those of (8.6.8). Consider now the case of 1 affected sib, which fall into two categories, according as the elder or younger is affected. Let

$$P_\xi(0,1) = p_{10} \text{ and } P_\xi(1,0) = p_{11}$$

with $p_{11} + p_{10} = p_1$, the probability that 1 sib is affected. The conditional probabilities of the two subcategories are p_{10}/p_1 and p_{11}/p_1. Now messy but straightforward algebra shows that

$$p_1\left(\frac{1}{(p_{11}/p_1)}\left(\frac{\delta}{\delta\xi}\left(\frac{p_{11}}{p_1}\right)\right)^2 + \frac{1}{(p_{10}/p_1)}\left(\frac{\delta}{\delta\xi}\left(\frac{p_{10}}{p_1}\right)\right)^2 \right)$$

reduces to

$$\frac{1}{p_{11}}\left(\frac{\delta p_{11}}{\delta\xi}\right)^2 + \frac{1}{p_{10}}\left(\frac{\delta p_{10}}{\delta\xi}\right)^2 - \frac{1}{p_1}\left(\frac{\delta p_1}{\delta\xi}\right)^2$$

or that the Fisher information in this conditional distribution is precisely the difference between (8.7.6) and (8.7.7). For a numerical example, see EXERCISE 8.7.3 below.

EXERCISE 8.7.1: Verify the outlined derivation of equation (8.7.5).

EXERCISE 8.7.2; Review your answer to EXERCISE 8.4.3, parts (a) and (b). See how the theory of this section provides that if if the true ascertainment is COT (b), then use of the COP likelihood (a) loses info, but the estimates are still consistent.

EXERCISE 8.7.3: Consider the example of the pair of sibs. Evaluate the Fisher information matrix (8.7.6) and the loss in information due to conditioning (8.7.7). For example, take $(\zeta_1, \zeta_2) =$ (0.5, 0.25), (0.2, 0.8), and (0.25, 0.25).
Note: Not only is very substantial information is lost, but that in all cases the conditional information is singular; ζ_1 and ζ_2 cannot be separately estimated. However, if the sibs are not a random sample the use of the unconditioned distribution will give incorrect answers, even asymptotically. Of course, if ascertainment does not depend only on the number of affected sibs, so will the conditional approach.

8.8 Further reading and open problems: a brief history of ascertainment theory.

Although Galton (1904) seems to be the earliest reference to problems of ascertainment in a demographic context (section 8.1), it was Bateson (1909) who recognised that the method by which the data are sampled would affect the appropriate analysis of genetic data. Alpert (1914) ascribed an (apparently) unexpectedly high segregation ratio for albinism to the way in which the data were collected, but did not attempt any ascertainment correction. It was Weinberg (1912) who first introduced the concept of *probands,* and developed methods for adjusting for sampling families through records of affected individuals.

Early genetic analysis focussed primarily on validating Mendel's Laws, and establishing the Mendelian nature of traits. Thus early work in ascertainment theory focussed almost entirely on segregation ratios, especially for recessive traits whose mode of inheritance in families was less "obvious" that that of dominant traits. Hogben (1931) and Haldane (1932) were both interested in testing segregation frequencies, and both therefore were drawn into the question of ascertainment corrections. However, it was Fisher (1934) who developed the classical likelihood, based on Weinberg's proband sampling model, showed that Weinberg's estimator was not in general the MLE, and thus not fully efficient, and considered more formally questions of consistency, efficiency and information. Morton (1959) provides a clear state-of-the art exposition of the likelihood estimation developments in ascertainment theory to that date, defining the terminology of single, complete and truncate ascertainment, and the loss of information involved in failure to specify probands (section 8.2). He also considered the effects of mixing different types of sibship, with different parental mating types and hence different segregation ratios (section 8.4.1). The text by Elandt-Johnson (1971) contains a detailed treatment of much of this theory.

However, new questions for more complex traits, and the development of linkage analysis, soon dictated the development of segregation analysis for more complex traits, and the analysis of data extending beyond nuclear families (section 8.3). Ascertainment theory for these wider problems was therefore developed, remaining in the context of Fisher's likelihood approach and Weinberg's proband sampling model. Elston (1973) and Elston and Yelverton (1975) considered problems of traits of variable age of onset, and developed methods for determining the probability that a nuclear family will contain at least one proband. Elston and Sobel (1979) generalise these considerations to individuals with varying probabilities of becoming probands, and showed how to determine the probability that a pedigree of given structure, containing individuals of specified variates relevant to ascertainment (age, sex, place of residence, trait value) would contain at least one proband. Thus they extended Fisher's likelihood based on the distribution of probands and affecteds within ascertained families to general pedigrees, and more general genetic analysis questions.

With increasing interest in quantitative traits, ascertainment theory needed to accommodate this case. However, where potential probands could be defined as individuals whose phenotypes exceeded a given threshold, or where a probability of becoming a proband could be defined as a function of a quantitative phenotype, no radically new ideas were involved. However, quantitative traits do introduce complexities into computation of likelihoods (for example, for the mixed model) and thence into the computation of ascertainment corrections. For convenience, likelihoods tended to be conditioned on proband phenotypic values, rather than on the event of exceeding a threshold.

Discussion of COP and COT and the constraints implicit in a quantitative threshold is given by Rao and Wette (1987). Young et al. (1988) provide further discussion, and recommend the inclusion of information via a likelihood, rather than through equality constraints on the parameters.

Recent developments in ascertainment theory have been founded on the recognition that the identification of "probands" is often impossible, and that Weinberg's original model cannot cover all situations. Elston and Sobel (1979) generalised the form of individual ascertainment probabilities, allowing their dependence on observable covariates such as age, place of residence. Stene (1977) developed more general forms of events (including events dependent on phenotypes of several individuals) to dictate the probability that a family be ascertained; thus probands were no longer assumed independent. However, the concept of families becoming ascertained with probabilities dictated by the phenotypes of its members was retained. The developments of Cannings and Thompson, and of Ewens and Shute, can be regarded as alternative approaches to avoiding this form of specification of the ascertainment process.

One objective of Cannings and Thompson (1977) in developing the sequential sampling approach was to avoid the concept of probands arising as a result of some probabilistic process requiring to be modelled, and instead conditioning on the specific event that initiated sampling of a pedigree. Thompson and Cannings (1979) investigate the foundations of the alternative approaches in more detail, pointing out that *if* the model of randomly ascertained probands is to be believed, and if there is a model for the numbers, sizes and structures of pedigrees in the population, then the numbers, sizes and structures of ascertained pedigrees contains information on genetic parameters. The classical likelihood formulation of Fisher (1934) which conditions on the sizes and structures of the ascertained families is a conditional likelihood.

In fact, the question of conditional and unconditional likelihoods was first discussed by Bailey (1951), who considered the unconditional likelihood, and pointed out that it did not provide identical results to the conditional one, although the differences were marginal. Fisher did not deny that in some sense there was information in the number of families ascertained, but indicated a clear preference for his conditional likelihood. He does not seem to have indicated whether his preference was based on robustness, on practical concerns about the excessive numbers of nuisance parameters, or on a recognition of the very limited amounts of information in the sizes and numbers of ascertained families. Various other special cases of the information available from the structures of ascertained families have been considered in the literature. For example, Hodge (1975) and Ewens et al. (1986) consider the effect of including family size data in their likelihood, in the presence and absence of knowledge of the population family-size distribution.

The work of Shute and Ewens can be regarded as the reverse side of the Thompson and Cannings (1979) development. Whereas Thompson and Cannings developed unconditional likelihoods that could extract additional information *if* the ascertainment model were to be believed, Ewens and

Shute start from the premise that the usual proband models are seldom fully valid, and thus that approaches robust to ascertainment assumptions should be developed. The excellent thesis by Shute (1988) gives a thorough review of classical ascertainment ideas, and develops the theory of robust approaches to fitting and testing genetic models through conditioning on the at least the data responsible for ascertainment. Some parts of this work have been published (Ewens and Shute, 1986; Shute and Ewens, 1988 a,b) This approach of conditioning at least on the relevant ascertainment information has been called "Ascertainment Assumption Free" (AAF) by Shute and Ewens. The example in section 8.6, of conditioning on the number of affected sibs when the sibs have covariate-dependent probabilities of being affected, is taken from a more complex example discussed by Hopper et al. (1990). Following a similar line of development, Hodge (1988) has also discussed the appropriate level of conditioning in the context of ascertainment.

Rao et al. (1988) placed the comparison of COP and COT in the context of the work of Ewens and Shute, showing that both provide consistent estimators when the true ascertainment is COT. Ewens and Green (1988) give a far more general discussion and clarification of the effects of various levels and types of conditioning in the case of quantitative trait data. Sawyer (1990) also considered quantitative traits. He places the questions of model misspecification and conditioning in a more general framework, and considers both the asymptotic bias of model misspecification and the loss of information in unnecessary conditioning, in the context of several ascertainment examples.

In this connection, one should note the use of "expected data" counts to analyse properties of log-likelihoods. In genetic epidemiology, this technique apparently dates to only to Morton (1984), although it has been used by several subsequent authors. In the more general statistical context of full exponential families, it seems that Huzurbazar (1949) first recognised the usefulness of the fact that for these distributions the log-likelihood is linear in the sufficient statistics, and hence the expected log-likelihood is (up to an additive constant) the log-likelihood based on any data providing the expected values of these statistics (Edwards, 1972).

Ascertainment has been not much considered in the context of linkage analysis, and particularly not in the context of affected sib and affected relative methods for linkage analysis, where a pair of affected individuals is the usual conditioning event (Bishop and Williamson, 1990; Risch 1990). However, the question was addressed by Risch, (1983), and becomes more pertinent in the case of sets of affected relatives in a sibship, nuclear family or pedigree, where there may be trait and family size information available for other than the affected sibs, and where alternative ascertainment schemes can be considered. The thesis of Tierney (1992) considers a number of different possible levels of conditioning in the context of affected sib methods, and the loss of power of likelihood ratio tests caused by a higher level of conditioning than the true ascertainment method requires.

The reopening of the ascertainment debate in recent years, the development and implementation of sequential studies in epidemiological studies, the use of more extended pedigrees both in segregation and linkage analysis, and the development of conditional likelihoods in a variety of genetic epidemiological contexts, have all lead to a greater understanding of the issues involved in ascertainment corrections. However, questions of ascertainment in genetic epidemiological studies are not closed; there are few definitive answers.

Too often, studies are analyzed retrospectively; the sampling strategy is seldom designed prospectively, and when it is, it may not be possible to follow it. While it is always (in principle) possible to condition on "at least the data variables relevant to ascertainment of a pedigree", this may result in loss of all the information about the model for the trait. Alternatively, we can try to fit the problem into the classical framework, by identifying probands; but are they correctly identified, and more importantly can they be assumed independent? Where extended pedigrees are studied, some form of sequential sampling often has, in fact, been implemented, and determination of (at least) the multiplex event that initiated sampling of a pedigree is often possible. However, it is difficult to be sure that the requirements of the sequential sampling likelihood, adjusting only for the initiating event, have been met. Where the assumption of "no look-ahead" and "no discards" are violated, valid ascertainment correction becomes almost impossible. Moreover, in many studies of real traits in actual populations, multiple ascertainment, or joining of sequentially sampled pedigrees, has non-negligible probability and often occurs. The joining of pedigrees is an event whose probability is dependent both on the genetic model for the trait and on population structure. The most efficient way to use information on the fact of multiple ascertainment or joining of pedigrees remains an open question.

In general, even when a sufficient conditioning is both clear and computationally feasible, it must be recognised that extensive conditioning results in substantial loss of power. There is no substitute for a well designed sampling procedure determined in advance of the study, and adhered to throughout the sampling process, with the proband status of each individual clearly recorded.

APPENDIX; Solutions to Exercises and Related Notes.

EXERCISE 8.1.1: Let X be the number of boys in a randomly chosen sibship, and N the sibship size. Then

$$P(N = n) = k_n, \quad \text{and } P(X = x \mid N = n) = \binom{n}{x} 2^{-n},$$

so that

$$\mathbb{E}(X(N - X)) \sum_n \sum_x x\,(n - x) \binom{n}{x} 2^{-n} k_n = \sum_n \frac{n(n-1)}{4}\, k_n = (V + \mu^2 - \mu)/4$$

while $E(X) = \mu/2$ so that $cov(X, n - X) = (V - \mu)/4$.

NOTE: If V is sufficiently large the numbers of boys and of girls is positively correlated -- an individual with more brothers, has, on average, also more sisters. If V is small, the reverse is true. For a Poisson distribution of sibship size, $V = \mu$, and in this case $cov(X, n - X) = 0$. In fact, more is true; for a Poisson sibship size distribution, the number of boys is independent of the number of girls. mis-estimated.

EXERCISE 8.2.1

(a) If $b = 1$ for every ascertained sibship, $\sum_r n_r(r - 1)$ is the number of affected sibs of the proband (not counting the proband himself), and the denominator is likewise the total number of sibs (not counting the probands). This is the "delete-one-affected" estimator; the proband is eliminated from the counts in each sibship.

(b) The weighting by r is now precisely as the weighting by x in (8.1.1) and (8.1.2).

(c) Given r, b is binomial with index r and parameter π. Given s, r is binomial with index s and parameter ζ. The results follow, conditioning first on r in the case of the numerator.

NOTE: From this it can be shown that Weinberg's estimator is a consistent estimator of ζ. However, in general it is not the maximum likelihood estimator of ζ; it is not always asymptotically efficient.

EXERCISE 8.2.2

(a) Log-likelihoods are; -3110, -3120, -3147, -3106, -3102, -3107, -3142, -3123, -3106.

(b) Log-likelihoods are;
-3557, -3518, -3586, -3553, -3497, -3547, -3590, -3514, -3546.)

Another example of the same phenomenon as here illustrated numerically by exercises 8.2.2 (a) and (b) is shown diagrammatically in section 8.6 (figure 2).

(c) Specifically, considering the log-likelihood for π alone, provided by the subdivision of the 70 sibships with 3 affecteds, one obtains as estimate $\hat{\pi} = 0.80$. For the 540 sibships with two affecteds, one obtains $\hat{\pi} = 0.41$. For the log-likelihood for both samples combined, the MLE of π is still about 0.5 (0.515 in fact), but the expectations of the numbers of sibships with given numbers of probands given by $\pi = 0.515$ are far from the observed values.

(d) Specifically, note that if p_r is the population proportion of sibships with r affected sibs, then all that can be estimated from the numbers of sibships is that

$$p_1 c_1 : p_2 c_2 : p_3 c_3 = 108 : 54 : 7$$

This implies $p_1 \geq 2\ p_2 \geq (108/7)\ p_3$. In terms of ζ, the first inequality gives $\zeta \leq 0.33$ and the second gives $\zeta \leq 0.28$, but no lower bound on ζ can be estimated.

EXERCISE 8.3.1: The set-up of this question is somewhat artificial, but it makes the point of what probabilities are required as components of the likelihood contributed by a particular ascertained pedigree. Since the individuals are independent, probabilities can be multiplied over individuals, but the full trait/proband configuration must be considered in each case. The probabilities the individuals are affected are

$$0.18,\ 0.09,\ 0.10,\ 0.04,\quad 0.20,\ 0.23\ \text{ and }\ 0.05$$

for the individuals in the order listed in the question.

(a) The probabilities that the individuals are not affected are, respectively,

$$0.82,\ 0.91\ 0.90,\ 0.96,\quad 0.80,\ 0.77\ \text{ and }\ 0.95$$

Multiplying these, we have a probability 0.377 that none is affected, or a probability 1-0.377 = 0.633 that the pedigree is ascertained.

(b) The probabilities that the individuals are probands are 0.6 times the probability that each is affected:

$$0.108,\ 0.054,\ 0.06,\ 0.024,\quad 0.12,\ 0.138\ \text{ and }\ 0.03$$

The probabilities that the individuals are not probands are

$$0.892,\ 0.946,\ 0.94,\ 0.976,\quad 0.88,\ 0.862\ \text{ and }\ 0.97.$$

Hence the probability that there is no proband is 0.570, and the probability the pedigree is ascertained is 0.430.

(c) The probability of the stated trait configuration is

$$0.18 \times 0.91 \times 0.90 \times 0.96 \times 0.2 \times .77 \times .05 = 0.001090$$

and of the proband configurations, given the trait information,

$$0.4 \times 1 \times 1 \times 1 \times 0.6 \times 1 \times 0.4 = 0.096$$

Combining these gives the result 0.0001046. Note that the probability of any specific trait and proband status configuration is likely to be small; the more detailed an event, the less *a priori* probable it is. This does not mean that the data contribute little to the comparison of alternative hypotheses; in fact, the opposite is true.

(d) We have to sum over the proband configurations giving rise to at least one proband, given the trait configuration in (c). In this case this is not hard since all affected individuals have the same probability of becoming probands; the probability of at least one proband is $(1-(1-0.6)^3) = 0.936$ giving overall result $0.936 \times 0.00109 = 0.00102$. However, this additional summation over proband configurations is required; in other situations, such as those of (e) for example, this is more tedious.

(e) The individual ascertainment probabilities for affected individuals are now

$$0.2,\ \ ,0.8,\ 0.8,\ 0.0,\quad 0.2,\ 0.2,\ \text{and}\ 0.0$$

and hence that each individual is a proband is

$$0.036,\ \ ,0.072,\ 0.08,\ 0.0,\quad 0.04,\ 0.046, \text{and}\ 0.0.$$

The analogous computation to (b), gives a probability 0.754 that none is a proband, or 0.246 that the pedigree is ascertained.

(f) One would have to sum over all trait configurations in which there is at least one pair of affected sibs. Since this small 7-member pedigree contains two sibships, one of size 3, the other of size 2, there are 20 such configurations. This computation is possible but tedious; in general, individual-specific ascertainment criteria are much easier to handle.

EXERCISE 8.4.1: Within each parental mating type, the offspring are independently affected, each with probability η time the probability the mating produces a $B_1 B_1$ child.

EXERCISE 8.4.2: (a) 0.0001, 0.0036, 0.0324 (or, relatively, 0.25%, 10% and 89.75%);
With 3 affected offspring; 9.0%, 40.4% and 50.6%;
With 1 affected offspring; 0.24%, 13.1% and 86.7%;
(b) If the parents are known to be unaffected:
With 3 affected offspring; 1.6%, 43.7% and 54.7%
With 1 affected offspring; 0.04%, 5.6% and 94.3%.
(c) In ascertained families: 0.9%, 17.5% and 81.6%.
Note the variation with offspring information, with parental data, and with fact of ascertainment.

EXERCISE 8.4.3:

(a) Bivariate normal theory gives the answer as normal, with mean $\mu_f + \rho \dfrac{\sigma_f}{\sigma_m}(K - \mu_m)$ and variance ${\sigma_f}^2(1 - \rho^2)$.
(b) There is no simple form. The answer is the bivariate density divided by the tail probability $P(X_m \geq K) = \Phi((\mu_f - K)/\sigma_f)$.
(c) The numerator in (b) must now be integrated over $X_m \geq K$. There is no explicit formula for the result of this integration.
(d) Now also in (c), K must be replaced by the function $\mu_m + \sigma_m \Phi^{-1}(0.95)$ of the parameters.
Note the successively more complex functions of the parameters, and consequent greater complexity of likelihood estimation of these parameters from observed data on the wives. Of course, to implement (d) prior information on some of these parameters would be necessary.

EXERCISE 8.4.4:

The proof is immediate, reversing the roles of trait and marker. Generally, despite Fisher's argument, is is very unlikely that an extended pedigree segregating a disorder of interest would not be subject to selection bias of some kind. Even supposing this could be so, for a rare dominant trait it would be preferable to select on the trait rather than marker. Only the double-heterozygotes provide information for linkage, and to have sufficient information for an effective linkage analysis it is necessary to have large numbers of individuals carrying the dominant trait allele.

EXERCISE 8.4.5:

The selection criteria through lipid levels, involving also those of first degree relatives and a spouse, would make any ascertainment correction for a segregation analysis almost impossible, but would not affect a linkage analysis. However, selection of pedigrees exhibiting a particular marker in probands, for linkage analysis to that same marker, could cause problems. This is particularly so in this case, as the existence of an affected first degree relative, who is thus (whether or not the loci are linked) likely to be carrying the same allele, could give a false impression of linkage. (I am indebted to Dr. Ellen Wijsman for bringing this example to my attention.)

EXERCISE 8.5.1:

(a) The likelihood is proportional to $(1-\zeta)^2\zeta$, the MLE is $\hat{\zeta} = 1/3$ and the observed information is $\dfrac{2}{1-\zeta} + \dfrac{1}{\zeta^2}$, or 13.5 at the MLE.

(b) The answers here are identical to those of (a), as the likelihood is the same.

(c) For this binomial experiment, the data are $B(3,\zeta)$ and the Fisher information is $\dfrac{3}{\zeta(1-\zeta)}$ (which is also 13.5 at $\zeta = 1/3$).

(d) For this negative binomial experiment, the data are geometric or $NB(1,\zeta)$. If K individuals are observed in total, there are $K-1$ unaffected followed by 1 affected. The negative second derivative of the log-likelihood is $\dfrac{K-1}{(1-\zeta)^2} + \dfrac{1}{\zeta^2}$, and since the expectation of K is $\dfrac{1}{\zeta}$, the Fisher information reduces to $\dfrac{1}{\zeta^2(1-\zeta)}$. This is also 13.5 at $\zeta = 1/3$. For smaller ζ values this is larger than the Fisher information in (c) -- this sequential sampling would be more informative. For larger values of ζ the binomial sampling would be preferred.

NOTE: In a real example, one would sample many more than three individuals if sampling binomially, and until many more than one affected had been found if sampling sequentially. The example here is used only to illustrate the general points made in this section.

EXERCISE 8.5.2:

(a) Sum the probabilities of each parental mating type multiplied by the conditional probability of a single unaffected individual:

$$q^4(1-\eta) + 4q^3(1-q)(1-(\eta/2)) + 4q^2(1-q)^2(1-(\eta/4)) + (1-q^4-4q^2(1-q)).\,1$$

which reduces to the required result, $1-q^2\eta$.

(b) Similarly to (a), from Table 3, for the case of two affected offspring only.

EXERCISE 8.5.3: Figure 1a shows the log-likelihoods for values of ζ from 0.05 to 0.75; the random sampling scheme is shown as points with lines, and the sequential scheme as a smooth curve. The random expected sample gives more information (a sharper log-likelihood) than the expected sample under the sequential scheme. However, the sequential scheme requires examination of fewer individuals, and when adjusted for this fact (figure 1b) the information per observation is greater for the sequential expected sample than for the sample of complete sibships.

EXERCISE 8.7.2

The ratio of the two likelihoods is just the conditional distribution of proband trait value conditional on exceeding the threshold. The expected log of this conditional density maximises at the true value, as shown in the text. The information lost is the information in this conditional density.

EXERCISE 8.7.3

For the suggested ζ-values the unconditional information matrices are, respectively,

$$\begin{pmatrix} 4 & 0 \\ 0 & 5.33 \end{pmatrix}, \begin{pmatrix} 6.25 & 0 \\ 0 & 6.25 \end{pmatrix}, \begin{pmatrix} 5.33 & 0 \\ 0 & 5.33 \end{pmatrix},$$

and the information losses are

$$\begin{pmatrix} 2.5 & 2 \\ 2 & 2.67 \end{pmatrix}, \begin{pmatrix} 4.78 & 1.47 \\ 1.47 & 4.78 \end{pmatrix}, \begin{pmatrix} 2.67 & 2.67 \\ 2.67 & 2.67 \end{pmatrix},$$

and the information in the conditional distribution is

$$\begin{pmatrix} 1.5 & -2 \\ -2 & 2.67 \end{pmatrix}, \begin{pmatrix} 1.47 & -1.47 \\ -1.47 & 1.47 \end{pmatrix}, \begin{pmatrix} 2.67 & -2.67 \\ -2.67 & 2.67 \end{pmatrix},$$

Note that not only is very substantial information is lost, but that in all cases the conditional information is singular; ζ_1 and ζ_2 cannot be separately estimated. However, if the sibs are not a random sample the use of the unconditioned distribution will give incorrect answers, even asymptotically. Of course, if ascertainment does not depend only on the number of affected sibs, so will the conditional approach.

Literature cited

Alpert, E. 1914. The laws of Naudin-Mendel. *Journal of Heredity* **5:** 492-497.

Bailey, N. J. T. 1951. Estimation of frequencies of recessives with incomplete multiple selection. *Annals of Eugenics* **16:** 215-222.

Bateson, W. 1909. *Mendel's Principles of Heredity.* (2 nd. ed.) Cambridge University Press.

Bishop, D. T. and J. A. Williamson 1990. The power of identity by state methods for linkage analysis. *American Journal of Human Genetics* **46:** 254-265.

Cannings, C. and Thompson E. A. 1977. Ascertainment in the sequential sampling of pedigrees. *Clinical Genetics* **12:** 208-212.

Edwards, A. W. F. 1972. *Likelihood.* Cambridge University Press; Cambridge, UK.

Elandt-Johnson, R. C. 1971. *Probability Models and Statistical Methods in Genetics.* Wiley; New York.

Elston, R. C. 1973. Ascertainment and age of onset in pedigree analysis. *Human Heredity* **23:** 105-112.

Elston, R. C. and K. C. Yelverton. 1975. General models for segregation analysis. *American Journal of Human Genetics,* **27:** 31-45.

Elston, R. C. and E. Sobel. 1979. Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics* **31:** 62-69.

Ewens, W. J., S. E. Hodge, and F. H. Ping. 1986. The effects of known family size distribution on the estimation of genetic parameters. *American Journal of Human Genetics* **38:** 555-566.

Ewens W. J. and N. C. E. Shute. 1986. A resolution of the ascertainment sampling problem. I. Theory. *Theoretical Population Biology* **30:** 388-412.

Ewens W. J. and R. Green, 1988. A resolution of the ascertainment problem. IV. Continuous phenotypes. *Genetic Epidemiology* **5:** 433-444.

Fisher R. A. 1934. The effects of methods of ascertainment on the estimation of frequencies. *Annals of Human Genetics* **6:** 13-25.

Galton, F. 1904. Average number of kinfolk in each degree. *Nature* **70:** 529 and 626.

Haldane, J. B. S. 1932. A method for investigating recessive characters in man. *Journal of Genetics* **25:** 251-255.

Hogben, L. T. 1931. The genetic analysis of familial traits. I. Single gene substitutions. *Journal of Genetics* **25:** 97-112.

Hodge, S. E. 1985. Family-size distribution and Ewens' equivalence theorem. *American Journal of Human Genetics* **37:** 166-177.

Hodge, S. E. 1988. Conditioning on subsets of the data: Applications to ascertainment and to other genetic problems. *American Journal of Human Genetics* **43:** 364-373.

Hopper, J. L., F. K. Judd, P. L. Derrick, G. T. Macaskill, and G. D. Burrows. 1990. A family study of panic disorder: Reanalysis using a regressive logistic model that incorporates a sibship environment. *Genetic Epidemiology* **7:** 151-161.

Huzurbazar, V. S. 1949. On a property of distributions admitting sufficient statistics. *Biometrika,* **36:** 71-74.

Morton, N. E. 1959. Genetic tests under incomplete ascertainment. *American Journal of Human Genetics* **11:** 1-16.

Morton, N. E. 1984. Trials of segregation analysis by deterministic and macro simulation. *Human Population Genetics: The Pittsburg Symposium.* A. Chakravarti (ed.). Van Nostrand Reinhold, New York.

Rao, D.C., R. Wette and W. J. Ewens. 1988. Multifactorial analysis of family data ascertained through truncation: a comparative evaluation of two methods of statistical inference. *American Journal of Human Genetics* **42:** 506-515.

Rao, D.C. and R. Wette. 1987. Non-random sampling in genetic epidemiology: maximum likelihood methods for multifactorial analysis of quantitative data ascertained through truncation. *Genetic Epidemiology* **4:** 357-376.

Risch, N. 1983. The effects of reduced fertility, method of ascertainment, and a second unlinked locus on affected sib-pair marker allele sharing. *American Journal of Medical Genetics* **16:** 243-259.

Risch, N. 1990. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *American Journal of Human Genetics* **46:** 229-241.

Sawyer, S. A. 1990. Maximum likelihood estimators for incorrect models, with an application to ascertainment bias for continuous characters. *Theoretical Population Biology* **38:** 351-366.

Shute, N. C. E. 1988. *Statistical and Ascertainment Problems in Human Genetics* Ph.D. Thesis; Monash University.

Shute N. C. E. and W. J. Ewens. 1988a A resolution of the ascertainment sampling problem. II. Generalisations and numerical results. *American Journal of Human Genetics* **43:** 374-386.

Shute N. C. E. and W. J. Ewens. 1988b A resolution of the ascertainment sampling problem. III. Pedigrees. *American Journal of Human Genetics* **43:** 387-395.

Stene, J. 1977. Assumptions for different ascertainment models in human genetics. *Biometrics* **33:** 523-527.

Thompson, E. A. and C. Cannings 1979 Sampling schemes and ascertainment. In *Genetic Analysis of Common Diseases* (C.F. Sing and M. Skolnick, eds.). Alan R. Liss Inc., New York (pp. 363-382).

Tierney, C. 1992. *Power of Affected Sibling Method Tests for Linkage* Ph.D. Thesis; University of Washington.

Weinberg, W. 1912. Zur Verebung der Anlage der Bluterkrankheit mit methodol. Erganzungen meiner Geschwistermethode. *Arch. Rass. u. GesBiol.* **9:** 694-709.

Wojciechowski. A. P., M. Farral, P. Cullen, T.M.E. Wilson, J. D. Bayliss, B. Farren, B. A. Griffin, M. J. Caslake, C. J. Packard, J. Shepherd., R. Thakker and J. Scott. Familial combined hyperlipidemia linked to the apolipoprotein AI-CIII-AIV gene cluster on chromosome 11q23-q24. *Nature,* **349:** 161-164.

Young, M. R., M. Boehnke and P. P. Moll. (1988) Correcting for single ascertainment by truncation for a quantitative trait. *American Journal of Human Genetics* **43:** 705-708.

Corrections and addenda to June 1995

Added: Fisher (1934) quote on header page.

Corrected: Equation (8.3.3)

Deleted: Final sentence of paragraph containing (8.4.6)

Corrected: 2 minor typos in Table 3.