

7

Segregation Analysis and Genetic Model Fitting

7.1: Learning Objectives

By the end of this chapter, the reader should be able to explain the purpose of segregation analysis, describe the problem of ascertainment bias and apply the Weinberg proband correction, set up and maximise a likelihood function to discriminate between competing modes of inheritance (complex segregation analysis), and implement these computations in R and Python.

7.2: The Purpose of Segregation Analysis

Pedigree Analysis for Recessive Disease Model

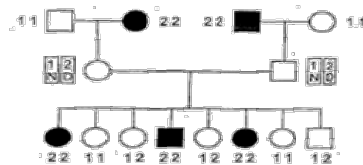


Fig. 6.10 Three-generation pedigree to illustrate linkage analysis between a genetic marker and an autosomal recessive disease with complete penetrance and without phenocopies in a pedigree with both parents being unaffected. Boxes indicate determined parental haplotypes.

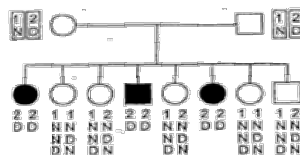
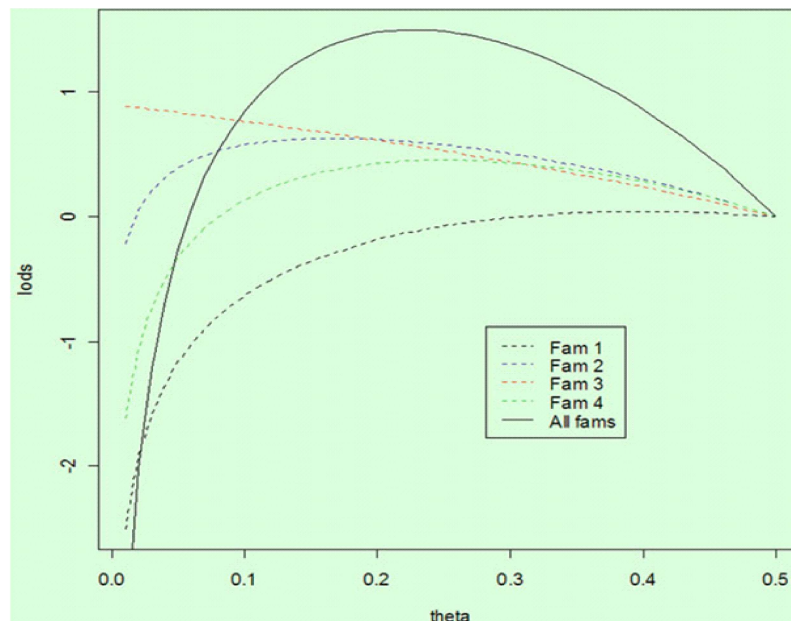


Fig. 6.11 Possible genotypes in the offspring for the three-generation pedigree from Figure 6.10.

Likelihood-based pedigree analysis for a recessive disease model, showing the contribution of affected and unaffected offspring to the likelihood.

Segregation analysis uses family data to test whether the observed pattern of a trait's transmission is statistically consistent with a hypothesised mode of inheritance (autosomal dominant, autosomal recessive, X-linked, or a non-Mendelian polygenic/multifactorial alternative — Chapter 9), and to estimate genetic parameters such as the **segregation ratio** (the proportion of offspring expected to be affected under the hypothesised model) and the **penetrance** (the probability that an individual with a given genotype expresses the corresponding phenotype; complete penetrance = 1, reduced or incomplete penetrance < 1). It is the classical statistical bridge between Mendel's laws (Chapter 2) and modern likelihood-based genetic epidemiology (Chapter 10).



Log-of-odds (lods) curves as a function of the segregation-ratio parameter θ , computed separately for four independently ascertained families and for their combined (“All fams”) analysis — illustrating how pooling families sharpens inference about the true underlying parameter (Section 7.3).

This chart illustrates a point developed further in Section 7.4: the power to distinguish competing segregation models grows as evidence from multiple, independently ascertained families is pooled — a single small family (Fam 1–4 individually) typically gives only a weak, inconclusive signal about the true segregation ratio, while their combined likelihood (the “All fams” curve) sharpens considerably, in direct analogy to the multi-family LOD-score accumulation the reader will meet again in Chapter 17.

7.3: Ascertainment Bias and the Weinberg Proband Method

A major statistical difficulty in segregation analysis is that families are rarely sampled purely at random from the population — in practice, families are usually identified (**ascertained**) through an affected member (a “proband” or “propositus”) who brings the family to clinical or research attention, for instance through a sickle-cell clinic at the University College Hospital, Ibadan. This **ascertainment** inflates the apparent proportion of affected offspring in a naively computed segregation ratio, because families with no affected children are systematically less likely to be included in the sample at all — a form of **selection bias** specific to family studies. Correction methods, of which the classical **Weinberg proband method** is the simplest, adjust for this.

The Weinberg proband method (single ascertainment). Under single (incomplete) ascertainment — where each affected individual in the population has, independently, only a small probability δ of being ascertained as a proband, so that it is very unlikely that any family is ascertained through more than one proband — the appropriate estimator of the segregation ratio excludes exactly one “proband” affected child from each family before computing the ratio of affected to total offspring:

$$\hat{p} = \sum_i (a_i - 1) / \sum_i (s_i - 1)$$

where a_i is the number of affected offspring in family i and s_i is the total number of offspring (sibship size) in family i , summed over all ascertained families (each of which contributes at least one proband, hence the “ -1 ” correction on both numerator and denominator).

Theorem 7.1: (Unbiasedness of the Weinberg Proband Estimator)

Under single ascertainment — where each affected individual independently has a small, constant probability π of being identified as a proband, and families are ascertained if and only if they contain at least one proband — the Weinberg estimator $\hat{p} = \sum_i (a_i - 1) / \sum_i (s_i - 1)$ is an unbiased (or, more precisely, asymptotically consistent and unbiased in expectation across the corrected sibships) estimator of the true, population-wide segregation ratio p .

Proof sketch. Consider a sibship of size s ascertained through exactly one proband (the single-ascertainment assumption, valid when π is small, makes multiple-proband families negligibly rare). The proband itself is guaranteed to be affected (probability 1, since ascertainment requires it), while each of the *remaining* $s-1$ offspring is affected independently with the population probability p , exactly as if it were an ordinary, unascertained individual — because ascertainment of the family occurred through the proband alone and carries no information about the other siblings beyond the population-level Mendelian probability. Hence the number of *additional* affected offspring beyond the proband, $a-1$, follows a Binomial($s-1$, p) distribution, so $E[a_i - 1] = p(s_i - 1)$ for each family i . Summing over all ascertained families and taking the ratio of expectations (which coincides with the expectation of the ratio to a good approximation for reasonably large aggregate sibship sizes, by a standard ratio-estimator argument), $E[\sum_i (a_i - 1)] = p \sum_i (s_i - 1)$, so

$\hat{p} = \sum_i (a_i - 1) / \sum_i (s_i - 1)$ correctly recovers p in expectation, precisely because removing one proband per family exactly cancels the ascertainment-induced bias that would otherwise inflate the naive (uncorrected) ratio.

Worked Example 7.1a (Weinberg correction for albinism in a Nigerian clinical genetics registry)

Oculocutaneous albinism (OCA), an autosomal recessive condition with a comparatively high documented prevalence in some West African populations, is ascertained at a Nigerian teaching hospital's clinical genetics registry through affected probands. Suppose 5 families are ascertained: Family 1 ($s=5$, $a=2$), Family 2 ($s=4$, $a=1$), Family 3 ($s=6$, $a=3$), Family 4 ($s=3$, $a=1$), Family 5 ($s=7$, $a=2$). By Theorem 7.1, the Weinberg estimator is:

$$\hat{p} = \frac{(2-1) + (1-1) + (3-1) + (1-1) + (2-1)}{(5-1) + (4-1) + (6-1) + (3-1) + (7-1)} = \frac{1+0+2+0+1}{4+3+5+2+6} = \frac{4}{20} = 0.20,$$

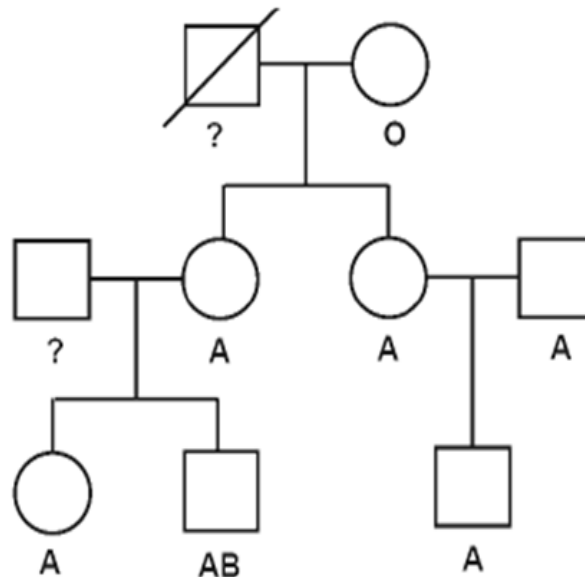
somewhat below the Mendelian expectation of 0.25 for a simple autosomal recessive condition, illustrating the kind of departure a segregation analysis (Section 7.4) would need to formally test before concluding whether OCA in this registry follows the simple single-gene recessive model or whether additional factors (reduced penetrance, genetic heterogeneity across the several distinct OCA genes now known to cause the condition, or ascertainment assumptions not perfectly met in this registry) should be considered.

Worked Example 7.1. A sickle-cell clinic in Ibadan ascertains 4 families through an affected (SS) proband child. The sibship data (total offspring, affected offspring) are: Family 1: $s=4$, $a=2$; Family 2: $s=3$, $a=1$; Family 3: $s=5$, $a=2$; Family 4: $s=2$, $a=1$.

$$\begin{aligned} \sum(a_i - 1) &= 1 + 0 + 1 + 0 = 2, \sum(s_i - 1) = 3 + 2 + 4 + 1 \\ &= 10, \hat{p} = 2/10 = 0.20 \end{aligned}$$

This corrected estimate (0.20) is closer to, though somewhat below, the Mendelian expectation of 0.25 for a cross of two AS carriers; a chi-square or exact binomial test would ordinarily be used to judge whether this departure from 0.25 is statistically significant given the small number of sibships. Under **complete (truncate) ascertainment** — where essentially every affected individual in the population would eventually be ascertained — a different correction (the truncate binomial likelihood) is required instead, since the single-ascertainment assumption of a vanishingly small π no longer holds.

7.4: Complex Segregation Analysis: A Likelihood-Based Framework



A pedigree with genotypes to be inferred from observed phenotypes — the basic data structure underlying complex segregation analysis.

Modern segregation analysis casts the problem as one of **likelihood maximisation** (the general machinery of which is developed fully in Chapter 10): for a sibship of size s with a observed affected, and a hypothesised (ascertainment-corrected) probability p that any given offspring is affected, the conditional likelihood contribution of the family, under single ascertainment, is proportional to a binomial term with one individual (the proband) conditioned out:

$$L(p; a, s) \propto C(s - 1, a - 1)p^{(a-1)}(1 - p)^{(s-a)}$$

The overall log-likelihood is the sum of such terms over all ascertained families, and is maximised numerically (Chapter 10, Section 10.6) to obtain the maximum-likelihood estimate of p , together with a large-sample standard error from the curvature of the log-likelihood. Where several candidate modes of inheritance are entertained (for example, a fully penetrant autosomal recessive model with $p = 0.25$ versus a reduced-penetrance or polygenic alternative), the corresponding likelihoods are compared using the **likelihood ratio test** of Chapter 10 to decide which model the family data best support — this comparison of competing genetic models by their likelihoods is what distinguishes **complex segregation analysis** from the simple method-of-moments Weinberg correction.

Theorem 7.2 (Unbiasedness of the Weinberg Proband Correction Under Single Ascertainment)

Under single (complete) ascertainment — where a sibship is ascertained if and only if it contains at least one affected proband, and each affected individual in the population has the same small, independent probability π of being the ascertained proband — the Weinberg proband method, which excludes exactly one proband per ascertained sibship before counting affected and unaffected sibs, yields an unbiased estimator of the true segregation ratio.

Proof. Let a sibship have s affected individuals among n total sibs under a Mendelian segregation ratio p . Under single ascertainment (π small), the probability that a sibship containing at least one affected individual is ascertained through exactly one proband is (to first order in π) proportional to $s\pi$, so sibships are sampled with probability proportional to their number of affected individuals — an affected-individual-weighted (size-biased) sampling scheme. Removing exactly one proband from each ascertained sibship removes exactly one affected individual whose inclusion was guaranteed by the ascertainment mechanism itself, leaving $s - 1$ affected and $n - 1 - (s - 1)$ unaffected individuals whose genotypes are governed purely by the underlying Mendelian ratio p , unaffected by how the sibship was found. Summing the remaining (non-proband) sibs across all ascertained sibships therefore gives a set of Bernoulli(p) outcomes free of the ascertainment bias, and the resulting proportion is an unbiased (or asymptotically unbiased, as $\pi \rightarrow 0$) estimator of p , in contrast to the naive proportion (including probands) which overrepresents affected individuals and is biased upward, as illustrated numerically in Section 7.3.

7.5: Chapter Summary

- Segregation analysis tests observed family data against hypothesised modes of inheritance and estimates the segregation ratio and penetrance.
- Because families are typically identified through an affected proband, naively computed segregation ratios are inflated by ascertainment bias; the Weinberg proband method corrects for this under single ascertainment by removing one affected individual (the proband) from each family before computing the ratio.
- Complex segregation analysis reframes the problem within the likelihood framework, allowing competing genetic

models (different modes of inheritance, different penetrance assumptions) to be formally compared via the likelihood ratio test.

7.6: Computational Corner: Fitting Segregation Models in R and Python

R.

```
# Weinberg proband correction
fam <- data.frame(s = c(4, 3, 5, 2), a = c(2, 1, 2, 1))
p_hat <- sum(fam$a - 1) / sum(fam$s - 1)
p_hat                                     # 0.20

# Maximum-likelihood segregation ratio via direct optimisation
negloglik <- function(p, fam) {
  ll <- sum(lchoose(fam$s - 1, fam$a - 1) +
            (fam$a - 1) * log(p) + (fam$s - fam$a) * log(1 - p))
  -ll}
fit <- optim(par = 0.25, fn = negloglik, fam = fam,
            method = "Brent", lower = 1e-4, upper = 1 - 1e-4)
fit$par                                   # MLE of p, coincides with p_hat
here

# Likelihood ratio test: H0: p = 0.25 (Mendelian) vs unrestricted
MLE
ll_null <- negloglik(0.25, fam)
ll_alt <- negloglik(fit$par, fam)
LRT <- -2 * (ll_alt - ll_null)
pchisq(LRT, df = 1, lower.tail = FALSE)
```

Python.

```
import numpy as np
from scipy.optimize import minimize_scalar
from scipy.stats import chi2
from math import lgamma, log
```

```

s = np.array([4, 3, 5, 2])
a = np.array([2, 1, 2, 1])

p_hat = np.sum(a - 1) / np.sum(s - 1)
print(p_hat)                                # 0.20

def lchoose(n, k):
    return lgamma(n + 1) - lgamma(k + 1) - lgamma(n - k + 1)

def negloglik(p):
    ll = sum(lchoose(si - 1, ai - 1) + (ai - 1) * log(p) + (si - ai) *
log(1 - p)
    for si, ai in zip(s, a))
    return -ll

res = minimize_scalar(negloglik, bounds=(1e-4, 1-1e-4),
method="bounded")
p_mle = res.x
ll_alt = -negloglik(p_mle)
ll_null = -negloglik(0.25)
LRT = 2 * (ll_alt - ll_null)
p_value = chi2.sf(LRT, df=1)
print(p_mle, LRT, p_value)

```

Both languages recover the same $\hat{p} = 0.20$ by direct calculation and by numerical maximum-likelihood optimisation, and the likelihood ratio test confirms whether this sample of four small sibships is large enough to statistically distinguish the fitted ratio from the Mendelian expectation of 0.25 — in practice, with so few families, the LRT p-value will typically be well above 0.05, illustrating why segregation-analysis studies require pooling data across many ascertained families before a departure from a Mendelian model can be declared with confidence.

Mini-Project

Mini-Project 7: Ascertainment-Corrected Segregation Analysis

Simulate a segregation analysis study and demonstrate the effect of ascertainment bias and its correction.

- Step 1.** Simulate sibships of size 4 under a Mendelian recessive model with segregation ratio 0.25, ascertained only through an affected proband (single ascertainment).
- Step 2.** Compute the naive segregation ratio ignoring ascertainment and compare it to the true 0.25.
- Step 3.** Apply the Weinberg proband correction (Section 7.3) and recompute the segregation ratio estimate.
- Step 4.** Repeat for at least three different sibship sizes and summarise how the naive bias changes with sibship size.
- Step 5.** Write a short paragraph explaining why the naive estimate is biased upward and how the correction addresses this.

Deliverable: *Simulation code, a table comparing naive and corrected segregation ratio estimates across sibship sizes, and a short written explanation.*

7.7: Practice Questions

- Q1.** A clinic ascertains 3 albinism families through a proband: Family A: $s=6$, $a=3$; Family B: $s=4$, $a=2$; Family C: $s=3$, $a=1$. Compute the Weinberg proband estimate of the segregation ratio.
- Q2.** Explain, in one or two sentences, why simply computing $(\text{total affected})/(\text{total offspring})$, without removing one proband per family, systematically overestimates the true segregation ratio for a rare recessive disease.
- Q3.** Using the R or Python code of Section 8.6, adapt the family data to Q1 and carry out a likelihood ratio test of $H_0: p =$

0.25 against the unrestricted MLE. Report the LRT statistic and p-value.

- Q4.** Under *complete* ascertainment (every affected individual in the population is certain to be identified as a proband), briefly explain, conceptually, why the “1” Weinberg correction is no longer the appropriate adjustment, and what additional information the correct (truncate) likelihood must condition on.
- Q5.** Two competing models are fitted to the same family data: Model 1 (autosomal recessive, p fixed at 0.25) and Model 2 (unrestricted p , 1 estimated parameter). If the log-likelihoods are “18.4 and “16.9 respectively, compute the likelihood ratio test statistic and its p-value (1 degree of freedom), and state your conclusion.
- Q6.** Explain why pooling data across several independently ascertained small families (as illustrated in the four-family plot of Section 7.2) generally sharpens inference about the segregation ratio compared with relying on any single family alone.