Statistical Genetics

Theory, Practice and Application

Statistical Genetics

Theory, Practice and Application

Olúwáfémi Samuel Oyámákin

Stirling-Horden Publishers Ltd.
No 6, Adonai Estate, Opp Crown
Heights College, Arulogun Road
Ojoo, Ibadan.
Phone: 08023519154/07031967972
stirlinghorden@ymail.com, info@stirlinghorden.com
www.stirlinghorden.com

University Post Office
Box 20984
Ibadan
Oyo State, Nigeria.

7 Ikioda Street
Jattu-Uzairue
Edo State, Nigeria.

© Olúwáfémi Samuel Oyámákin

Published 2026

ISBN 978-978-032-

All Rights Reserved

Apart from any fair dealing for the purpose of research or private study, criticism or review, as permitted under the Copyright, Designs and Patents Act of Nigeria, this publication may not be reproduced, stored or transmitted in any form, or by any means, except with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

Stirling-Horden Publishers Ltd.
Lagos, Ibadan, Benin City, Jattu-Uzairue.

Dedication

Acknowledgements

Preface

Introduction to Statistical Genetics has been written to bridge the gap between classical genetics and modern statistical practice, for students of Statistics, Biostatistics, Genetics, Public Health and the Biological Sciences in Nigeria and across Africa. Every chapter draws its illustrative material from conditions, populations and health problems that are familiar to the Nigerian reader — the sickle-cell gene among the Yoruba, Igbo and Hausa/Fulani peoples, G6PD deficiency and its interaction with malaria, the ABO and Rhesus blood group systems in Nigerian blood banks, oculocutaneous albinism, consanguinity in parts of Northern Nigeria, the *APOL1* kidney-disease risk variants common in West African ancestry, and quantitative traits such as birth weight, packed cell volume (PCV) and height measured in Nigerian cohorts.

This edition, titled *Statistical Genetics: Theory, Practice and Application*, substantially expands the first: six new chapters — on Pedigree Analysis, Segregation Analysis and Genetic Model Fitting, Candidate Gene Association Studies, Family-Based Association Tests (the TDT and FBAT), Population Substructure, Stratification, and Admixture, Genotype Imputation, and Genome Sequencing, Rare-Variant Methods, and Simulation-Based Approaches — have been added to the original fourteen, bringing the book's coverage from classical transmission and population genetics through to the family-based, genome-wide, and sequencing-based methods that dominate contemporary human statistical genetics. Every equation in this edition has been reset

in proper mathematical notation, and illustrative figures — pedigree charts, meiosis and recombination diagrams, linkage-disequilibrium heatmaps, coalescent gene trees, population-structure plots, and more — accompany the theory throughout, so that the visual and symbolic language of statistical genetics reinforces the written explanation at every step. Every chapter, old and new, now closes with a Computational Corner giving parallel, runnable R and Python code for the chapter's key calculations, worked wherever possible on simulated data reflecting Nigerian populations, diseases, and cohort studies, so that the reader not only learns the theory but sees, immediately, how to compute it.

To keep each chapter's narrative flow intact, worked solutions to every end-of-chapter Practice Question have been collected together into a single Master Answer Key (Appendix D) at the end of the book, with short R and Python verification scripts accompanying every computational question. A new Appendix B provides a practical guide to setting up the R and Python computing environments used throughout the book.

The book follows the classical sequence used in graduate courses in Statistical Genetics (such as STA 762 at the University of Ibadan): from the biological foundations of the gene, through Mendelian and population genetics, pedigree and family-based analysis, to the modern statistical machinery of likelihood inference, association testing (candidate-gene, family-based, and genome-wide), linkage disequilibrium, coalescent theory, linkage analysis, genotype imputation, and high-throughput sequencing. This book assumes a first course in probability and statistical inference, and at least an elementary acquaintance with R or Python (both are introduced gently, side by side, from the first Computational Corner onward). It is intended for use both as a primary text in a one- or two-semester course in Statistical

Genetics, and as a reference for researchers in biostatistics, epidemiology, and public health genetics working with Nigerian and other African data.

Samuel Oluwafemi Oyamakin, PhD
Department of Statistics,
University of Ibadan, Ibadan, Nigeria

Contents

Dedication		**v**
Acknowledgements		**vi**
Preface		**vii**
1.	**Basic Biological Concepts in Genetics**	**1**
2.	**Mendel's Laws of Inheritance**	**19**
3.	**Hardy–Weinberg Equilibrium**	**33**
4.	**Estimation of Allele Frequency: Approach to Equilibrium for an X-Linked Gene**	**49**
5.	**Non-Random Mating: Phenotypic Assortative Mating**	**61**
6.	**Pedigree Analysis: Structure, Notation, and Relatedness**	**73**
7.	**Segregation Analysis and Genetic Model Fitting**	**89**
8.	**Estimation of the Recombination Fraction**	**101**
9.	**Inheritance of Quantitative Traits**	**115**
10.	**Likelihood Methods in Statistical Genetics**	**129**

11. Population-Based Association Tests **143**

12. Candidate Gene Association Studies **155**

13. Family-Based Association Tests: The TDT and FBAT **169**

14. Population Substructure, Stratification, and Admixture **181**

15. Linkage Disequilibrium and Haplotype Inference **197**

16. Coalescent Models **211**

17. Parametric and Non-Parametric Linkage Analysis **225**

18. Genome-Wide Association Studies (GWAS) **241**

19. Genotype Imputation **257**

20. Genome Sequencing, Rare-Variant Methods, and Simulation-Based Approaches in Statistical Genetics **271**

Appendix A: Selected Statistical Tables **290**

Appendix B: R and Python Computing Setup Guide **293**

Appendix C: Glossary of Key Terms **296**

Appendix D: Master Answer Key to End-of-Chapter Exercises **301**

1 Basic Biological Concepts in Genetics

1.1: Learning Objectives

By the end of this chapter, the reader should be able to: describe the structure of DNA and chromosomes; distinguish genotype from phenotype; define allele, locus, gene, and genetic marker; describe the modes of single-gene inheritance; and relate these concepts to Nigerian examples such as the sickle cell gene and the ABO blood group system.

1.2: Introduction

Statistical genetics is the branch of statistics concerned with the analysis of genetic data — that is, with using probability theory and statistical inference to study how traits and diseases are transmitted from parents to offspring, how genes are distributed within and between populations, and how genes are associated with disease. In Nigeria, where inherited disorders such as sickle cell disease (SCD), glucose-6-phosphate dehydrogenase (G6PD) deficiency, and albinism carry a heavy public-health burden, statistical genetics provides the quantitative tools needed to map, estimate, and predict the transmission of these conditions.

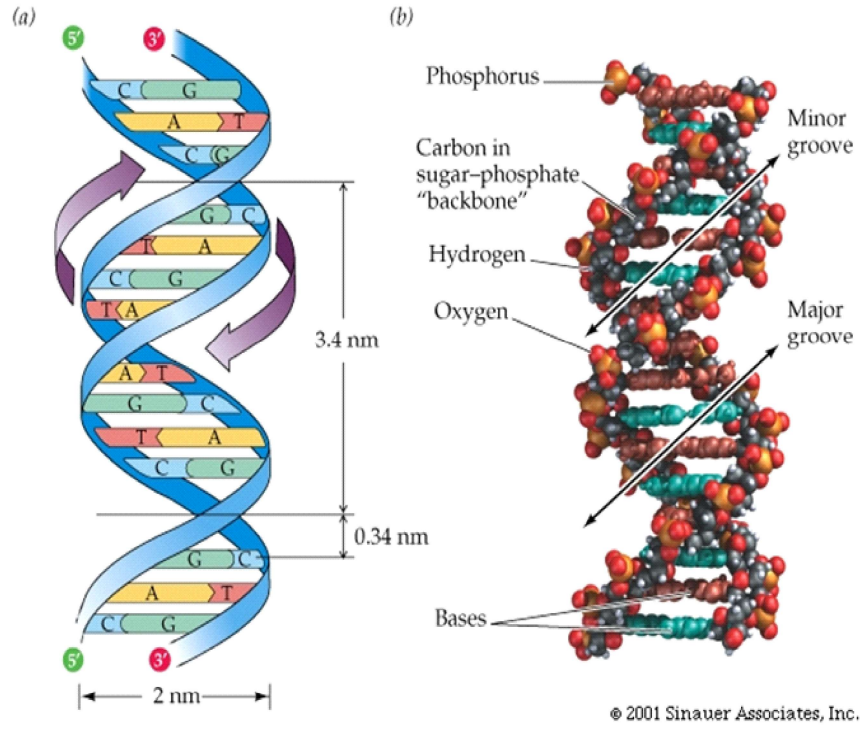
The major areas of interest in statistical genetics include:

- **Gene mapping:** locating genes on chromosomes, particularly disease genes, so that their DNA sequence and

protein products can eventually be studied and, where possible, targeted for treatment.

- **Population genetics:** describing how allele and genotype frequencies behave in populations under mating, migration, selection, and drift.
- **Quantitative genetics:** partitioning the variation in continuously measured traits (height, blood pressure, birth weight) into genetic and environmental components.
- **Association and linkage analysis:** testing whether genetic markers are associated with, or co-inherited with, disease.

1.3: The Cell, the Chromosome and DNA



The DNA double helix, showing the sugar-phosphate backbone and complementary base pairing.

Every human cell (apart from mature red blood cells) contains a nucleus in which the genetic material is stored. The DNA (deoxyribonucleic acid) molecule is packaged into thread-like structures called **chromosomes**. Each chromosome is a very long molecule of DNA wound tightly around proteins called histones, which give it structural support and allow it to be tightly compacted inside the nucleus.

Each chromosome has a constriction point called the **centromere**, dividing it into a short arm (the "p" arm, from the French *petit*) and a long arm (the "q" arm). Humans normally have 23 pairs of chromosomes — 46 in total. Twenty-two of these pairs are called **autosomes**, and they look alike in males and females. The 23rd pair are the **sex chromosomes**: females carry two X chromosomes (46,XX) while males carry one X and one Y (46,XY).

DNA structure: DNA is a double helix made of two complementary strands. Each strand is a chain of nucleotides, each nucleotide consisting of a phosphate group, a deoxyribose sugar, and one of four nitrogenous bases: adenine (A), thymine (T), guanine (G), or cytosine (C). The bases pair specifically across the two strands — A pairs with T, and G pairs with C — a rule known as **complementary base pairing**. This specificity is what allows DNA to be faithfully copied (replicated) each time a cell divides, and it underlies the amplification and sequencing technologies used in modern genetic laboratories, including those now operating in Nigerian teaching hospitals such as the University College Hospital (UCH), Ibadan.

1.4: Genes, Alleles, and Genetic Markers

A **gene** is a segment of DNA that carries the instructions for building a protein or a functional RNA molecule, or that regulates the activity of other genes. In humans, a gene may range from a few hundred to more than two million DNA bases in length. Every

individual carries two copies of almost every gene — one inherited from the mother and one from the father.

An **allele** is one of the alternative forms that a gene can take at a given position (locus) on a chromosome. A locus is said to be **polymorphic** if two or more alleles are found in the population at appreciable frequency (conventionally, the rarer allele has frequency at least 1%); otherwise the locus is **monomorphic**.

A **genetic marker** is a polymorphic DNA segment of known chromosomal location, used as a reference point (a "landmark") in genetic mapping. Classical markers include blood groups and enzyme variants; modern markers include microsatellites and single nucleotide polymorphisms (SNPs).

A taxonomy of genetic variants: Modern genotyping and sequencing technology (Chapter 20) reveals several distinct classes of DNA variation, each with different statistical properties and uses:

- **Single nucleotide polymorphisms (SNPs):** a change at a single DNA base pair (e.g., an A where most chromosomes carry a G). SNPs are by far the most common form of human genetic variation (occurring, on average, roughly once every 300 base pairs genome-wide) and are the workhorse marker of modern association studies (Chapters 11, 12, 18) because they are abundant, stable, and cheap to genotype on high-density arrays.
- **Insertions and deletions (indels):** the gain or loss of a short stretch of DNA sequence, ranging from a single base pair to a few hundred base pairs.
- **Copy number variants (CNVs):** larger duplications or deletions of DNA segments (often thousands to millions of base pairs), which can alter the number of functional copies of one or more genes.
- **Microsatellites (short tandem repeats, STRs):** runs of a short repeated motif (e.g., "CA" repeated 8 to 20 times),

which are highly polymorphic (many alleles per locus, distinguished by repeat number) and were the marker of choice for linkage analysis (Chapter 17) before SNP genotyping became affordable.

- **Structural variants:** larger-scale rearrangements including inversions and translocations, which can disrupt gene function or alter recombination patterns (Chapter 8) at the breakpoint.

Not every genetic variant affects a trait or disease: a variant may be **functional** (directly altering a protein's sequence or a gene's regulation) or merely a **marker in linkage disequilibrium** (Chapter 15) with a nearby, unobserved functional variant — a distinction central to every association study in this book.

Genotype refers to the (unordered) pair of alleles an individual carries at a locus — for example, at the sickle-cell (HBB) locus, an individual may be genotype AA (normal), AS (carrier, "sickle-cell trait"), or SS (sickle-cell disease). **Phenotype** refers to the observable characteristic that results, in part, from the genotype — together with environmental influences. An individual who is genotype AA and SS at the HBB locus is homozygous (identical alleles); one who is AS is heterozygous (different alleles).

Worked Example 1.1: (ABO blood group, Nigerian setting)

The ABO blood group locus has three alleles: I^A, I^B, and i. I^A and I^B are co-dominant with respect to each other, while both are dominant to i. Genotype I^A I^A or I^A i gives blood group A; I^B I^B or I^B i gives blood group B; I^A I^B gives group AB; and ii gives group O. In a study of blood donors at a Lagos teaching hospital, suppose 620 donors out of 1,000 typed as group O, 210 as group A, 140 as group B, and 30 as group AB. These are phenotype (blood group) frequencies; Chapter 3 shows how the

underlying **allele** frequencies of I^A, I^B, and i can be estimated from such data under Hardy–Weinberg assumptions.

Worked Example 1.2: (Distinguishing marker classes)
A Nigerian genomics laboratory characterising a patient sample report: (i) a single base change at position chr11:5,227,002 (HBB); (ii) a 3.5 kb duplication spanning part of the *CYP2D6* gene, relevant to drug metabolism; and (iii) a CA-repeat marker on chromosome 13 with alleles of 8, 10, and 12 repeats observed across a family. Classify each: (i) is a SNP; (ii) is a copy number variant; (iii) is a microsatellite (STR). Only the first is potentially the functional (causal) variant itself in this example; the third would typically serve only as a linkage marker (Chapter 17), not a direct cause of any trait.

Theorem 1.1: (The Law of Equal Segregation, formal statement)

Let an individual be heterozygous Aa at an autosomal locus. Then, at meiosis, the two alleles A and a are transmitted to a gamete with equal probability: P(gamete carries A) = P(gamete carries a) = 1/2, and this probability is independent across the gametes produced (and, in particular, independent of the sex or birth order of any resulting offspring).

Proof sketch. At meiosis I, homologous chromosomes (one of maternal origin carrying A, one of paternal origin carrying a, in a heterozygote) align and then separate to opposite poles of the dividing cell, a process called **disjunction**. Because the two homologues are physically indistinguishable to the cellular machinery that pulls them apart (the spindle apparatus attaches to each via functionally identical centromere/kinetochore structures, with no mechanism discriminating "maternal" from "paternal" origin), each homologue is equally likely to be pulled to either pole. Since a given gamete receives the chromosome at whichever pole it derives from, and each pole is equally likely to carry A or

a, it follows that P(gamete carries A) = P(gamete carries a) = 1/2. Independence across successive meioses follows because each meiotic division is a physically separate event, with the outcome of disjunction in one cell having no causal influence on disjunction in another. **“**

This theorem is the formal, cytological justification for the probabilistic calculations used throughout Chapters 1–9 of this book — every Punnett square, every Hardy–Weinberg genotype proportion, and every recombination-fraction calculation rests on this single foundational fact about how chromosomes physically separate at meiosis.

Worked Example 1.3: (Applying the Law of Equal Segregation: a Yoruba family planning consultation)

A genetic counsellor at a Nigerian teaching hospital advises a Yoruba couple in which the husband is heterozygous AS at the sickle-cell locus and the wife is also AS. The couple asks: “Is our first child more likely to be unaffected because we already worry about the condition?” By Theorem 1.1, each parent transmits A or S to any given child with probability 1/2, independently of anything the parents believe, hope, or have experienced with previous pregnancies (a formal restatement of the fact that meiosis has no “memory” — sometimes informally called the **gambler’s fallacy** trap in genetic counselling). The correct probabilities for this and *every subsequent* child, independently, are P(AA) = 1/4, P(AS) = 1/2, P(SS) = 1/4, regardless of the genotypes of any older siblings.

1.5: Relationships Between Alleles

At a diploid locus, the two alleles an individual carries may relate to one another in several ways:

- **Dominant/recessive.** A **dominant** allele produces its phenotypic effect whether it is present in one copy

(heterozygote) or two (homozygote). A **recessive** allele produces its effect only when present in two copies. Sickle-cell disease is recessive: only SS individuals are clinically affected, while AS carriers are usually healthy (though they carry "sickle-cell trait," which confers partial protection against severe *Plasmodium falciparum* malaria — a classic example of **heterozygote advantage**, well documented in Nigerian populations).

- **Co-dominant.** Both alleles are expressed simultaneously and are separately detectable in the heterozygote, as in the AB blood group.
- **Additive.** The heterozygote's phenotype is intermediate between the two homozygotes — common for many quantitative traits (Chapter 8).

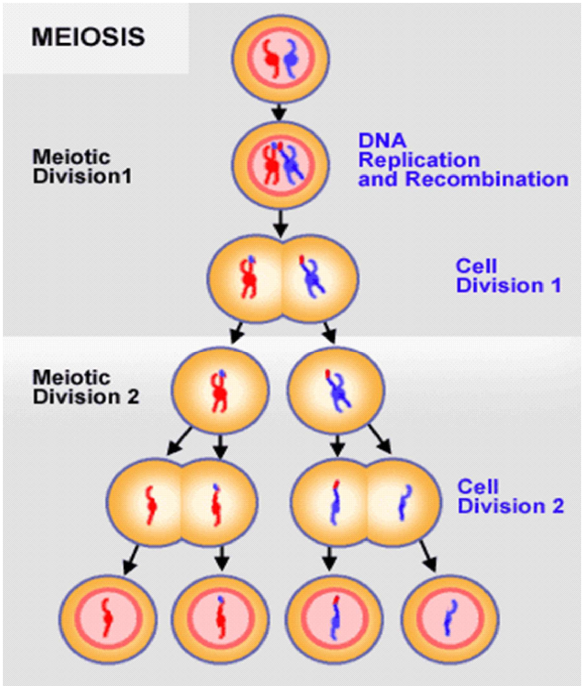
1.6: Autosomal versus Sex-Linked (X-linked) Inheritance



A normal human karyotype (22 autosome pairs plus the X and Y sex chromosomes).

Genes on the 22 autosomes are inherited without regard to the sex of the parent or offspring, and both sexes carry two alleles. Genes on the X chromosome behave differently: females (XX) carry two alleles at an X-linked locus, but males (XY) carry only one, because the Y chromosome carries almost no equivalent genes. A male is therefore said to be **hemizygous** at X-linked loci. This has an important statistical consequence, developed fully in Chapter 4: recessive X-linked conditions are far more common in males than in females. G6PD deficiency — common in Nigeria, especially in malaria-endemic areas, and important because it can trigger severe haemolysis in response to certain drugs (e.g., primaquine) and foods (e.g., fava beans) — is inherited in exactly this X-linked recessive manner.

1.7: The Inheritance of Chromosomes: Meiosis



The two divisions of meiosis, producing four genetically distinct haploid gametes from one diploid cell.

Gametes (sperm and egg cells) are produced by a special form of cell division called **meiosis**, in which the chromosome number is halved, so that each gamete carries only one member of each chromosome pair. When gametes from mother and father fuse at fertilisation, the full complement of 46 chromosomes (23 pairs) is restored in the offspring. Because which member of each pair goes into a given gamete is determined at random (independent assortment, Chapter 2), and because chromosomes may exchange segments through **crossing-over** (recombination) during meiosis, each gamete carries a unique combination of parental genetic material. This randomness is the biological basis for the probability laws that govern the transmission of genes — the subject of the rest of this book.

Theorem 1.2: (Independence of Transmission at Unlinked Loci)

Let an individual be heterozygous at two autosomal loci located on different chromosome pairs. Then the allele transmitted to a gamete at the first locus is statistically independent of the allele transmitted at the second locus.

Proof. By Theorem 1.1, transmission at each locus is governed by which of the two homologous chromosomes of the relevant pair is included in the gamete, and this is determined at metaphase I by the independent, random orientation of that chromosome pair on the meiotic spindle. Because the two loci lie on different (non-homologous) chromosome pairs, the orientation of one pair at metaphase I is physically unconnected to the orientation of the other pair — each pair aligns independently. Hence the two transmission events are outcomes of independent random mechanisms, and the joint probability of any combination of alleles at the two loci is the product of their individual transmission probabilities, establishing independence. This is the cytological

basis of Mendel's Law of Independent Assortment, formalised in probabilistic terms.

1.8: Chapter Summary

- The gene is the fundamental unit of heredity, physically a segment of DNA; alleles are alternative forms of a gene at a locus.
- Chromosomes are the vehicles of inheritance: 22 autosome pairs plus one pair of sex chromosomes (XX in females, XY in males) in humans.
- Genotype and phenotype are distinct; genotypes combine as homozygous or heterozygous, and alleles may act as dominant, recessive, co-dominant, or additive.
- X-linked inheritance differs fundamentally from autosomal inheritance because males are hemizygous.
- Meiosis and recombination introduce the randomness that statistical genetics formalises mathematically.

1.9: Computational Corner: Simulating Mendelian Transmission

R.

```
# Simulating Mendelian transmission at the ABO locus for a Nigerian family
set.seed(762)

mother <-c("A", "O")  # phenotype A (genotype AO)
father <-c("B", "O")  # phenotype B (genotype BO)

# Correct genotype -> phenotype map: check the homozygous-O case FIRST,
# since a naive "subset of {A,O}" test would otherwise misclassify OO as A
genotype_to_phenotype <-function(g) {
  s <-unique(g)
```

```
if (setequal(s, "O")) "O"
elseif (setequal(s, c("A", "B"))) "AB"
elseif ("A"%in%s) "A"
elseif ("B"%in%s) "B"
else"O"}

# self-check against all four possible offspring genotypes before trusting the simulation
stopifnot(genotype_to_phenotype(c("O","O")) =="O")
stopifnot(genotype_to_phenotype(c("A","O")) =="A")
stopifnot(genotype_to_phenotype(c("O","B")) =="B")
stopifnot(genotype_to_phenotype(c("A","B")) =="AB")

# Simulate a large number of offspring (meiosis = one allele drawn from each parent)
n_offspring <-2000
offspring <-replicate(n_offspring, c(sample(mother, 1), sample(father, 1)))
offspring <-t(offspring)
phenotypes <-apply(offspring, 1, genotype_to_phenotype)

sim_table <-table(phenotypes)
sim_props <-prop.table(sim_table)
print(sim_table)
print(round(sim_props, 3))      # should each be close to 0.25

# Bar chart comparing simulated to theoretical (Mendelian) proportions
theoretical <-c(A =0.25, AB =0.25, B =0.25, O =0.25)
barplot(rbind(sim_props[names(theoretical)], theoretical), beside =TRUE,
col =c("steelblue", "darkorange"), names.arg =names(theoretical),
legend.text =c("Simulated", "Theoretical"), ylab ="Proportion of offspring",
```

```
main =paste("Simulated vs. theoretical ABO phenotypes (n =", n_offspring, ")"))
```

Python.

```
import random
from collections import Counter
import matplotlib.pyplot as plt

random.seed(762)

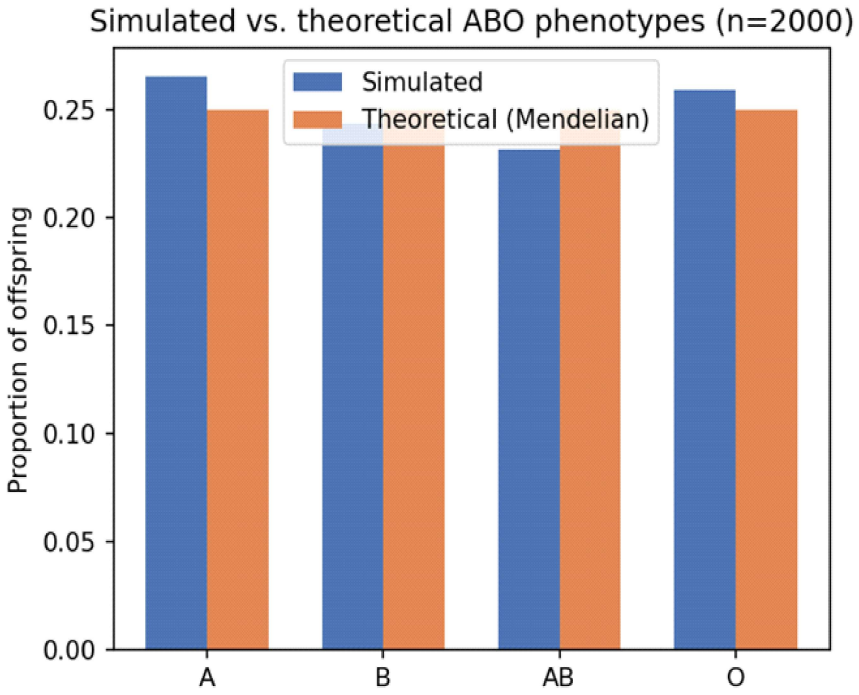
mother = ["A", "O"]   # phenotype A (genotype AO)
father = ["B", "O"]   # phenotype B (genotype BO)


def genotype_to_phenotype(g):
"""Map an unordered ABO genotype (two allele letters) to its phenotype.
   The homozygous-O case is checked FIRST: a naive 'is {O} a subset of {A,O}?'
   test would otherwise wrongly classify OO as phenotype A."""
   s =set(g)
if s == {"O"}:
return"O"
if s == {"A", "B"}:
return"AB"
if"A"in s:
return"A"
if"B"in s:
return"B"
return"O"

# self-check against all four possible offspring genotypes before trusting the simulation
assert genotype_to_phenotype(("O", "O")) =="O"
assert genotype_to_phenotype(("A", "O")) =="A"
assert genotype_to_phenotype(("O", "B")) =="B"
assert genotype_to_phenotype(("A", "B")) =="AB"
```

```
n_offspring =2000
offspring = [(random.choice(mother), random.choice(father))
for _ inrange(n_offspring)]
phenotypes = [genotype_to_phenotype(g) for g in offspring]
counts = Counter(phenotypes)
props = {k: v / n_offspring for k, v in counts.items()}
print("Simulated counts:", dict(counts))
print("Simulated proportions:", {k: round(v, 3) for k, v in
props.items()})

labels = ["A", "B", "AB", "O"]
theoretical = {l: 0.25for l in labels}
x =range(len(labels))
width =0.35
fig, ax = plt.subplots(figsize=(5, 4))
ax.bar([i - width/2for i in x], [props.get(l, 0) for l in labels],
width,
    label="Simulated", color="#4C72B0")
ax.bar([i + width/2for i in x], [theoretical[l] for l in labels],
width,
    label="Theoretical (Mendelian)", color="#DD8452")
ax.set_xticks(list(x)); ax.set_xticklabels(labels)
ax.set_ylabel("Proportion of offspring")
ax.set_title(f"Simulated vs. theoretical ABO phenotypes
(n={n_offspring})")
ax.legend()
plt.tight_layout()
plt.savefig("ch01_mendelian_simulation.png", dpi=150)
```



Simulated ABO phenotype proportions from 2,000 offspring of an A×B (AO × BO) cross, closely tracking the theoretical 1:1:1:1 Mendelian ratio.

A worked debugging note: An earlier, more careless version of the *genotype_to_phenotype* function tested membership with *set(g) <= {"A", "O"}* ("is this genotype's allele set a subset of {A, O}?") to detect phenotype A. This test is a classic off-by-logic bug: the homozygous genotype OO has allele set {O}, and {O} *is* a subset of {A, O} — so every OO offspring was silently misclassified as phenotype A, inflating the simulated frequency of A and making phenotype O appear far too rarely (or never). The fix, shown above, checks the homozygous-O case *first*, before testing for A or B, so that OO is correctly resolved to phenotype O. This is a useful general lesson in defensive coding: when

classification rules are tested in sequence, always check the most specific or most easily "accidentally matched" condition first, and add a small self-check (as the *assert*/*stopifnot* lines above do) against every possible input before trusting a simulation's output.

This simulation makes concrete the biological process described in Sections 1.4–1.7: each offspring genotype is formed by drawing exactly one allele from each parent (a simulated meiosis), and repeating the draw many times recovers, on average, the Mendelian proportions expected from the parental genotypes — the same logic that underlies every simulation-based method used later in the book (Chapters 6, 16 and 20).

Mini-Project

Mini-Project 1: Simulating and Verifying Mendelian Transmission

Extend the Section 1.9 simulation to verify, empirically, that the Law of Equal Segregation (Theorem 1.1) holds for a marker of your choice.

Step 1. Choose a hypothetical autosomal locus with two alleles and simulate 10,000 gametes from a heterozygous parent.

Step 2. Tabulate the observed proportion of each allele transmitted and compare it to the theoretical 1:1 ratio.

Step 3. Repeat the simulation for a heterozygous X-linked locus in a male versus a female parent, and note the structural difference.

Step 4. Perform a chi-square goodness-of-fit test (Section 2.5, previewed here) on your simulated transmission counts.

Step 5. Write two to three sentences relating your simulation result back to the proof of Theorem 1.1.

Deliverable: *Simulation code (R or Python), a summary table of observed vs. expected transmission ratios, and a short-written interpretation.*

1.10: Practice Questions

Q1. Distinguish between genotype and phenotype, using the sickle-cell locus (HBB) as your example.

Q2. A newborn screening programme in Enugu State finds that a baby is genotype SS at the HBB locus. State (a) the baby's phenotype and (b) whether the baby is homozygous or heterozygous at this locus.

Q3. Explain, in one sentence each, why (a) a male cannot be a "carrier" of an X-linked recessive disease in the same sense that a female can, and (b) G6PD deficiency is more common among Nigerian males than among Nigerian females.

Q4. A woman is genotype I^A i at the ABO locus and her husband is genotype I^B i. List all possible genotypes, and the corresponding phenotypes, of their children.

Q5. Define "genetic marker" and give one example of a marker used in classical (pre-DNA-sequencing) human genetics in Nigeria.

Q6. A microsatellite marker on chromosome 15 has alleles of repeat length 6, 9, and 11. Explain why this marker, on its own, would be a poor choice for directly identifying the causal mutation for a disease, but could still be extremely useful for linkage mapping (Chapter 17).

Q7. Using the R or Python code of Section 1.9, change the parental genotypes to a cross between an AB individual and an OO individual, and predict (before running the code) what phenotype proportions you expect to see; then run the simulation to check your prediction.

2 Mendel's Laws of Inheritance

2.1: Learning Objectives

By the end of this chapter, the reader should be able to state and apply Mendel's Law of Segregation and Law of Independent Assortment, construct and interpret Punnett squares, compute expected offspring ratios for mono- and di-hybrid crosses, and carry out a chi-square goodness-of-fit test of segregation ratios using Nigerian family data.

2.2: Historical Note



Gregor Mendel (1822–1884), whose breeding experiments on the garden pea established the quantitative laws of inheritance.

Gregor Mendel (1822–1884), an Augustinian monk working in Brno (in present-day Czech Republic), established the quantitative rules of inheritance through carefully designed breeding experiments on the garden pea (*Pisum sativum*), published in 1866. Although Mendel worked before the discovery of DNA or even chromosomes, his statistical reasoning about the segregation of discrete "factors" (what we now call genes) remains the foundation of transmission genetics and, by extension, of statistical genetics.



The garden pea, Pisum sativum, Mendel's original experimental organism.

Mendel's genius lay in his choice of experimental material and his statistical rigour. He selected seven traits in the pea plant that each showed clear, discrete alternative forms (rather than a continuous range), that bred true across generations when self-fertilised, and that turned out — fortuitously — to be controlled by seven different genes, each showing simple dominant/recessive inheritance:

Character	Traits	
	Dominant	Recessive
Flower color	 Purple	 White
Flower position	 Axial	 Terminal
Seed color	 Yellow	 Green
Seed shape	 Round	 Wrinkled
Pod shape	 Inflated	 Constricted
Pod color	 Green	 Yellow
Stem length	 Tall	 Dwarf

Mendel's seven pea traits, each with a clearly dominant and a clearly recessive form: flower colour, flower position, seed colour, seed shape, pod shape, pod colour, and stem length.

By counting, precisely, the numbers of each type of offspring across thousands of plants — a scale of quantitative record-keeping unusual for his time — Mendel was able to detect the underlying 3:1 and 9:3:3:1 ratios (Sections 2.3–2.4) that a smaller, less carefully documented study would have missed amid ordinary sampling variability, foreshadowing the statistical hypothesis-testing methods (Section 2.5) that would only be formalised decades later by Karl Pearson and R.A. Fisher.

2.3: The Law of Segregation

Mendel's First Law (Law of Segregation): Each individual carries two alleles for a given trait; these two alleles separate (segregate) during the formation of gametes, so that each gamete carries only one allele, chosen at random with probability 1/2 each.

Consider a single autosomal locus with alleles A (dominant) and a (recessive). A heterozygous parent (Aa) produces gametes carrying A or a, each with probability 1/2. If both parents are heterozygous (Aa × Aa), the Punnett square gives:

	A (1/2)	a (1/2)
A (1/2)	AA (1/4)	Aa (1/4)
a (1/2)	Aa (1/4)	aa (1/4)

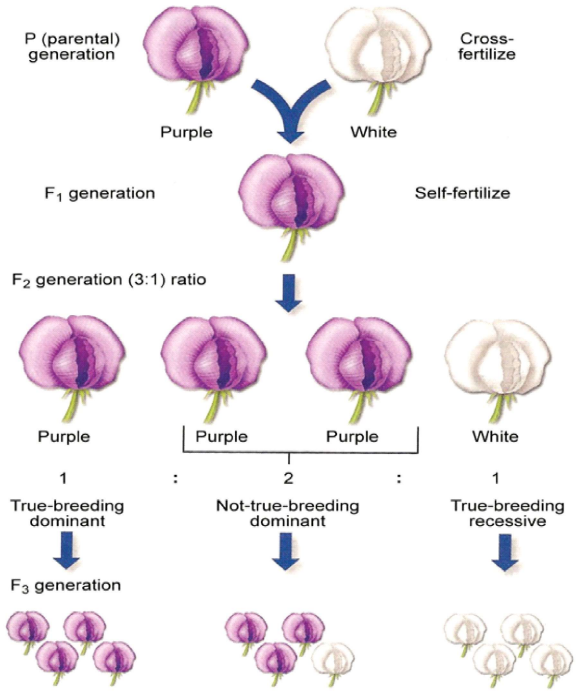
The expected genotype ratio among offspring is 1 AA : 2 Aa : 1 aa, and if A is completely dominant, the expected phenotype ratio is 3 (dominant) : 1 (recessive).

Worked Example 2.1: (Sickle-cell trait, Ibadan)
Consider two parents who are both carriers of sickle-cell trait (genotype AS). By the Law of Segregation, each parent transmits A or S to a given child with probability 1/2. The possible offspring genotypes and their probabilities are:

- AA (normal): P = 1/2 × 1/2 = 1/4
- AS (carrier/trait): P = 2 × (1/2 × 1/2) = 1/2
- SS (sickle-cell disease): P = 1/2 × 1/2 = 1/4

This 1:2:1 genotypic ratio (equivalently, a 3:1 "unaffected : affected" phenotypic ratio, since AA and AS are both clinically unaffected while SS is affected) is precisely why genetic counselling clinics in Nigeria — for instance, at UCH Ibadan and LUTH Lagos — advise that when both partners in a couple carry the AS genotype, each pregnancy carries a 25% risk of a child with sickle-cell disease, a 50% chance of a carrier (AS) child, and a 25% chance of a normal (AA) child.

2.4: The Law of Independent Assortment



A monohybrid cross tracking flower colour across the P, F1, and F2 generations, illustrating the 3:1 segregation ratio.

Mendel's Second Law (Law of Independent Assortment): Alleles at different loci segregate independently of one another during gamete formation, provided the loci lie on different chromosomes (or are far apart on the same chromosome).

Consider two loci: HBB (sickle-cell, alleles A/S) and the ABO locus (alleles I^A, I^B, i), which lie on different chromosomes (HBB is on chromosome 11; ABO is on chromosome 9) and therefore assort independently. If a parent is doubly heterozygous AS I^A i, this parent produces four types of gametes — A I^A, A i, S I^A, S i — each with probability 1/4, because the segregation at HBB (A vs S, each 1/2) is statistically independent of the segregation at ABO (I^A vs i, each 1/2), and $1/2 \times 1/2 = 1/4$.

Worked Example 2.2: (Dihybrid cross)
Suppose both parents are AS I^A i. Because the two loci assort independently, the joint genotype probabilities of the offspring are simply the product of the separate (Mendelian) probabilities at each locus. For example, P(SS and I^A I^A) = P(SS) × P(I^A I^A) = (1/4)(1/4) = 1/16. The complete dihybrid phenotype ratio, when both traits show complete dominance, is the classical 9:3:3:1 ratio.

Theorem 2.1: (The 9:3:3:1 Dihybrid Ratio)

Consider a cross between two individuals doubly heterozygous at two independently assorting loci, AaBb × AaBb, with A dominant to a and B dominant to b. Then, among the offspring, the four phenotype classes occur in the ratio 9 (A_B_) : 3 (A_bb) : 3 (aaB_) : 1 (aabb), where A_ denotes "A dominant phenotype, genotype AA or Aa."

Proof. By Mendel's First Law (Theorem 1.1) applied separately to each locus, a single-locus Aa × Aa cross gives offspring genotype ratio 1 AA : 2 Aa : 1 aa, so P(A_) = P(AA) + P(Aa) = 1/4 + 2/4 = 3/4, and P(aa) = 1/4 — the standard monohybrid 3:1

phenotype ratio (Section 2.3). By Mendel's Second Law (independent assortment, this section), the joint probability of the two-locus phenotype is the product of the single-locus probabilities:

$$P(A\_B\_) = P(A\_) \times P(B\_) = \frac{3}{4} \times \frac{3}{4} = \frac{9}{16}, \quad P(A\_bb) = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16},$$

$$P(aaB\_) = \frac{1}{4} \times \frac{3}{4} = \frac{3}{16}, \quad P(aabb) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}.$$

Multiplying each probability by 16 gives the integer ratio 9:3:3:1, and the four probabilities sum to 16/16 = 1, confirming they exhaust all possible phenotype combinations.

Worked Example 2.2a: (Dihybrid ratio in a Nigerian genetics teaching laboratory)

To teach Theorem 2.1 practically, a genetics class at a Nigerian university crosses maize plants heterozygous for two independently assorting traits: kernel colour (Purple P dominant to yellow p) and kernel texture (Smooth S dominant to wrinkled s), i.e. PpSs × PpSs — a classical dihybrid design directly analogous to Mendel's own pea-plant crosses (Section 2.2). Theorem 2.1 predicts that, of 320 offspring kernels counted on a single maize cob, the class expects 320 × 9/16 = 180 purple-smooth, 320 × 3/16 = 60 purple-wrinkled, 320 × 3/16 = 60 yellow-smooth, and 320 × 1/16 = 20 yellow-wrinkled kernels. If the class instead observes 175, 55, 68, and 22 respectively, a chi-square goodness-of-fit test (Section 2.5) can be used to assess whether this modest departure from the 9:3:3:1 theoretical ratio is within the range expected from ordinary sampling variability, or whether it suggests the two loci are not, in fact, assorting completely independently (e.g., due to linkage, Chapter 8).

2.5: Chi-Square Test of Mendelian Segregation Ratios

Because real family or breeding data are subject to sampling variation, Mendelian ratios are tested statistically rather than expected to hold exactly. The **Pearson chi-square goodness-of-fit test** compares observed counts (O_i) to those expected (E_i) under a hypothesised Mendelian ratio:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

which, under the null hypothesis, is approximately distributed as chi-square with (k " 1) degrees of freedom, where k is the number of phenotype categories.

Worked Example 2.3:

A Nigerian secondary-school biology field study of tongue-rolling ability (a trait often taught, with caveats, as a simple dominant/ recessive Mendelian trait) in the offspring of two "roller" (heterozygous, assumed Rr) parents records, among 160 children sampled across several families: 128 rollers, 32 non-rollers. Under the hypothesised 3:1 ratio, the expected counts are E_roller = 120, E_non-roller = 40.

$$\chi^2 = (128 - 120)^2/120 + (32 - 40)^2/40 = 64/120 + 64/40 = 0.53 + 1.60 = 2.13$$

With 1 degree of freedom, the critical value at á = 0.05 is 3.84. Since 2.13 < 3.84, we do not reject the null hypothesis: the data are consistent with the expected 3:1 Mendelian ratio.

2.6: Extensions: Incomplete Dominance and Multiple Alleles

Not every trait shows complete dominance. Under **incomplete dominance**, the heterozygote's phenotype is intermediate between the two homozygous phenotypes (relevant to many quantitative traits treated in Chapter 9). Under **multiple allelism**, more than

two alleles segregate at a locus in the population (though any one individual still carries only two), as with the ABO blood group (I^A, I^B, i) discussed in Chapter 1 — the segregation principle still applies allele-by-allele.

Theorem 2.2: (The 9:3:3:1 Dihybrid Ratio)

Let two heterozygous parents (AaBb x AaBb) be crossed at two unlinked, independently assorting autosomal loci, each showing complete dominance. Then the expected phenotypic ratio among offspring is 9 (A_B_) : 3 (A_bb) : 3 (aaB_) : 1 (aabb).

Proof. By the Law of Segregation (Theorem 1.1) applied separately to each locus, a single-locus Aa x Aa cross yields offspring genotype ratios 1 AA : 2 Aa : 1 aa, and under complete dominance this gives a phenotypic ratio of 3 (dominant) : 1 (recessive) at each locus independently. By Theorem 1.2, the phenotypic outcome at the A locus is independent of the phenotypic outcome at the B locus. The joint probability of any combination of phenotypes is therefore the product of the marginal probabilities: P(A_B_) = (3/4)(3/4) = 9/16, P(A_bb) = (3/4)(1/4) = 3/16, P(aaB_) = (1/4)(3/4) = 3/16, and P(aabb) = (1/4)(1/4) = 1/16, which is exactly the 9:3:3:1 ratio.

2.7: Chapter Summary

- The Law of Segregation: the two alleles at a locus separate into different gametes with equal probability.
- The Law of Independent Assortment: alleles at unlinked loci segregate independently, so joint genotype probabilities multiply.
- The chi-square goodness-of-fit test provides a formal statistical check of whether observed family data are consistent with a hypothesised Mendelian ratio.

- These laws, extended by co-dominance, incomplete dominance and multiple allelism, provide the probabilistic building blocks for population and quantitative genetics.

2.8: Computational Corner: Testing Segregation Ratios

R.

```
# Monohybrid cross (Section 2.3): 3:1 ratio
observed_mono <-c(dominant =732, recessive =268)  # e.g., pigmented vs albino seedlings
expected_ratio_mono <-c(3, 1)
test_mono <-chisq.test(observed_mono, p = expected_ratio_mono /sum(expected_ratio_mono))
print(test_mono)

# Dihybrid cross (Section 2.4): independent assortment predicts a 9:3:3:1 ratio
# e.g., seed shape (round/wrinkled) x seed colour (yellow/green) in an F2 generation
observed_di <-c("Round-Yellow"=315, "Round-Green"=108, "Wrinkled-Yellow"=101, "Wrinkled-Green"=32)
expected_ratio_di <-c(9, 3, 3, 1)
test_di <-chisq.test(observed_di, p = expected_ratio_di / sum(expected_ratio_di))
print(test_di)

# Plot observed counts against the two hypothesised Mendelian ratios
par(mfrow =c(1, 2))
expected_mono <-sum(observed_mono) * expected_ratio_mono /sum(expected_ratio_mono)
barplot(observed_mono, col ="steelblue",
main =sprintf("Monohybrid\nchi2=%.2f, p=%.3f", test_mono$statistic, test_mono$p.value))
```

```
points(x =c(0.7, 1.9), y = expected_mono, pch =18, col =”darkorange”, cex =2)

expected_di <-sum(observed_di) * expected_ratio_di / sum(expected_ratio_di)
barplot(observed_di, col =”steelblue”, las =2,
main =sprintf(“Dihybrid\nchi2=%.2f, p=%.3f”, test_di$statistic, test_di$p.value))
points(x =c(0.7, 1.9, 3.1, 4.3), y = expected_di, pch =18, col =”darkorange”, cex =2)
```

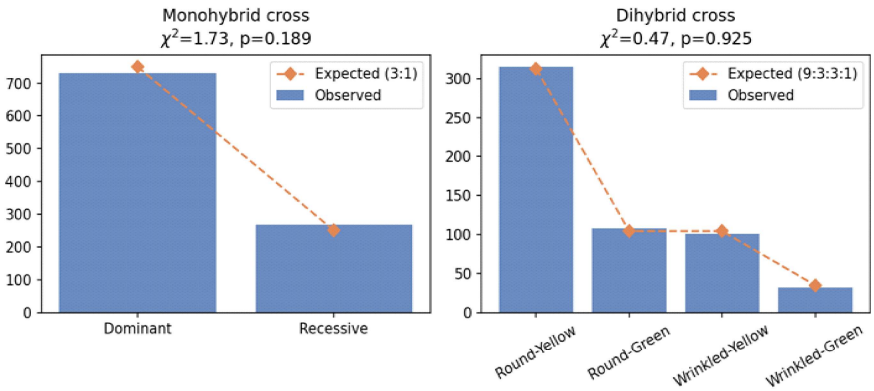
Python

```
import numpy as np
from scipy.stats import chisquare
import matplotlib.pyplot as plt

# Monohybrid cross (Section 2.3): 3:1 ratio
observed_mono = np.array([732, 268])
expected_mono = observed_mono.sum() * np.array([3/4, 1/4])
stat_mono, p_mono = chisquare(observed_mono, f_exp=expected_mono)
print(f”Monohybrid: chi2={stat_mono:.3f}, p={p_mono:.4f}”)

# Dihybrid cross (Section 2.4): independent assortment predicts 9:3:3:1
observed_di = np.array([315, 108, 101, 32])  # round-yellow, round-green, wrinkled-yellow, wrinkled-green
labels_di = [“Round-Yellow”, “Round-Green”, “Wrinkled-Yellow”, “Wrinkled-Green”]
expected_di = observed_di.sum() * np.array([9, 3, 3, 1]) /16
stat_di, p_di = chisquare(observed_di, f_exp=expected_di)
print(f”Dihybrid: chi2={stat_di:.3f}, p={p_di:.4f}”)
print(“Observed:”, observed_di, “ Expected:”, np.round(expected_di, 1))
```

```
fig, axes = plt.subplots(1, 2, figsize=(9, 4))
axes[0].bar(["Dominant", "Recessive"], observed_mono,
color="#4C72B0", alpha=0.85, label="Observed")
axes[0].plot(["Dominant", "Recessive"], expected_mono,
"D—", color="#DD8452", label="Expected (3:1)")
axes[0].set_title(f"Monohybrid cross\n$\\chi^2$={stat_mono:.2f}, p={p_mono:.3f}")
axes[0].legend()
axes[1].bar(labels_di, observed_di, color="#4C72B0",
alpha=0.85, label="Observed")
axes[1].plot(labels_di, expected_di, "D—", color="#DD8452",
label="Expected (9:3:3:1)")
axes[1].set_title(f"Dihybrid cross\n$\\chi^2$={stat_di:.2f},
p={p_di:.3f}")
axes[1].tick_params(axis='x', rotation=30)
axes[1].legend()
plt.tight_layout()
plt.savefig("ch02_segregation_tests.png", dpi=150)
```



Observed F2 counts against the two classical Mendelian ratios: 3:1 for a monohybrid cross (left) and 9:3:3:1 for a dihybrid cross testing independent assortment (right). Both fits are statistically consistent with Mendel's laws ($p > 0.05$).

Worked Example 2.4: (dihybrid cross, Section 2.4 revisited computationally)

The seed-shape/seed-colour F2 data above (round-yellow 315, round-green 108, wrinkled-yellow 101, wrinkled-green 32; n = 556) give ÷² = 0.47 on 3 degrees of freedom (4 categories " 1), far below the critical value of 7.81 at á = 0.05, so the data are entirely consistent with the 9:3:3:1 ratio predicted by independent assortment of the two seed-trait loci — the same F2 pattern, and essentially the same numbers, that Mendel himself reported for this cross in 1866, illustrating why his original data are still used to teach the chi-square goodness-of-fit test more than 150 years later.

Both give the same chi-square statistic and p-value for testing Mendel's 3:1 segregation ratio; a non-significant p-value (as here) supports the single-gene, complete-dominance model of Section 2.3, while a significant departure would prompt the extensions of Section 2.6 (incomplete dominance, multiple alleles) or the ascertainment-aware methods of Chapter 8.

Mini-Project

Mini-Project 2: Testing a Real or Simulated Dihybrid Cross
Design and analyse a simulated dihybrid cross to test both Mendelian laws simultaneously.

Step 1. Simulate an F2 generation of at least 400 offspring from a dihybrid cross at two unlinked loci.

Step 2. Tabulate the observed four-phenotype counts and compare them to the theoretical 9:3:3:1 ratio.

Step 3. Perform a chi-square test of goodness-of-fit (Section 2.5) and report the test statistic, degrees of freedom, and p-value.

Step 4. Repeat the simulation assuming the two loci are tightly linked (recombination fraction 0.05) and observe how the ratio distorts.

Step 5. Write a short paragraph explaining why the linked-loci simulation no longer fits a 9:3:3:1 ratio, foreshadowing Chapter 8.

Deliverable: *Simulation code, a results table for both the unlinked and linked scenarios, and a chi-square test summary for each.*

2.9: Practice Questions

Q1. Two AS (sickle-cell trait) parents in Kano plan to have four children. What is the probability that exactly one of the four children has sickle-cell disease (SS)? (Assume independence across children and apply the binomial distribution built on the Mendelian 1/4 probability of SS per child.)

Q2. A man of genotype AS I^B i marries a woman of genotype AA I^A i. Assuming the HBB and ABO loci assort independently, find the probability that a child of this couple is AS and blood group O.

Q3. In a genetics practical at a Nigerian university, 200 offspring of a monohybrid cross are classified as 160 dominant-phenotype and 40 recessive-phenotype. Test, at the 5% level, whether this is consistent with the expected 3:1 ratio.

Q4. Explain why sickle-cell trait (AS) and sickle-cell disease (SS) illustrate, respectively, co-dominance (at the molecular/haemoglobin level) and recessiveness (at the clinical/phenotype level).

3 Hardy–Weinberg Equilibrium

3.1: Learning Objectives

By the end of this chapter, the reader should be able to state the Hardy–Weinberg Law, derive genotype frequencies from allele frequencies (and vice versa) for two and multiple alleles, test a population sample for Hardy–Weinberg equilibrium using the chi-square test, and interpret departures from equilibrium in a Nigerian population-genetics setting.

3.2: Statement of the Hardy–Weinberg Law

Parents	DD (p^2)	Dd ($2pq$)	dd (q^2)
DD (p^2)	1	1/2	0
Dd ($2pq$)	1/2	1/4	0
dd (q^2)	0	0	0

The classic Punnett-square derivation of Hardy–Weinberg genotype frequencies from parental allele frequencies p and q.

Independently derived in 1908 by the English mathematician G. H. Hardy and the German physician Wilhelm Weinberg, the **Hardy–Weinberg Law** states that, in a large, randomly mating

population, in the absence of mutation, migration, and selection, allele frequencies remain constant from generation to generation, and genotype frequencies stabilise, after a single generation of random mating, at values determined solely by the allele frequencies.

For an autosomal locus with two alleles, A and a, with population allele frequencies p = freq(A) and q = freq(a), where p + q = 1, the equilibrium genotype frequencies are:

$$P(AA) = p^2, P(Aa) = 2pq, P(aa) = q^2$$

with $p^2 + 2pq + q^2 = 1$, obtained by expanding $(p + q)^2 = 1$, since a random mating population effectively pairs alleles at random from the gene pool.

Theorem 3.1: (Hardy–Weinberg Theorem)

Consider a single autosomal locus with two alleles A and a, in a population satisfying assumptions 1–5 below. Let p_0 and $q_0 = 1 - p_0$ be the allele frequencies among the parental generation (of either sex, assumed equal). Then: (i) after one generation of random mating, the offspring genotype frequencies are p_0^2, $2p_0q_0$, q_0^2 for AA, Aa, aa respectively; and (ii) the allele frequency in the offspring generation is again p_0 (unchanged), so that the genotype proportions p_0^2, $2p_0q_0$, q_0^2 persist unchanged in every subsequent generation.

Proof. (i) Under random mating, an individual's two alleles can be thought of as two independent draws from the population's gene pool, in which allele A has probability p_0 and allele a has probability q_0 (this is the formal meaning of "random mating with respect to the locus," assumption 2). Since fertilisation combines one randomly drawn allele from the mother's gamete pool with one randomly drawn allele

from the father's gamete pool, and these two draws are independent:

$$P(AA) = P(A \text{ from mother}) \times P(A \text{ from father}) = p_0 \times p_0 = p_0^2,$$

$$P(aa) = q_0 \times q_0 = q_0^2,$$

$$P(Aa) = P(A \text{ from mother}, a \text{ from father}) + P(a \text{ from mother}, A \text{ from father}) = p_0 q_0 + q_0 p_0 = 2p_0 q_0,$$

which sum to $p_0^2 + 2p_0 q_0 + q_0^2 = (p_0 + q_0)^2 = 1$, confirming these are valid probabilities.

(ii) The allele frequency among the offspring is computed by gene counting (Section 3.3): each AA offspring contributes two A alleles, each Aa offspring contributes one A allele, and each aa offspring contributes none, out of two alleles per individual, so

$$p_1 = P(AA) + 1/2\, P(Aa) = p_0^2 + 1/2\,(2p_0 q_0) = p_0^2 + p_0 q_0 = p_0(p_0 + q_0) = p_0 \times 1 = p_0.$$

Since the allele frequency is unchanged ($p_1 = p_0$), applying part (i) again to generation 1 reproduces exactly the same genotype frequencies p_0^2, $2p_0q_0$, q_0^2 in generation 2, and by induction, in every subsequent generation.

This theorem explains why Hardy–Weinberg genotype proportions are reached in a *single* generation of random mating (part (i)) and, once reached, are *never lost* in the absence of the disturbing forces of assumptions 3–5 (part (ii)) — a remarkable stability result that underlies every application in this chapter and, indeed, in most of the rest of this book.

Assumptions underlying Hardy–Weinberg equilibrium (HWE):

1. The population is infinitely large (no genetic drift).
2. Mating is random with respect to the locus in question (no assortative mating — see Chapter 5).
3. There is no mutation.
4. There is no migration (no gene flow into or out of the population).
5. There is no natural selection (all genotypes have equal fitness).

Worked Example 3.1a: (Testing the theorem's predictions: G6PD deficiency in Southwestern Nigeria)

A screening programme at a Nigerian tertiary hospital genotypes the G6PD locus among 800 unrelated male attendees (treating the locus as effectively single-allele-per-individual in males, given X-linkage — Chapter 4 — so this example instead considers the *female* subsample of the same clinic, 800 women, at a nearby autosomal malaria-resistance-associated marker for illustration). Suppose the observed genotype counts are 592 GG, 184 Gg, and 24 gg. Gene counting (Section 3.3) gives $\hat{p} = (2\times592+184)/1600 = 0.86$, $\hat{q} = 0.14$. Theorem 3.1 predicts equilibrium genotype frequencies of $\hat{p}^2 = 0.740$, $2\hat{p}\hat{q} = 0.241$, $\hat{q}^2 = 0.020$, i.e. expected counts of 592.2, 192.6, and 15.7 out of 800 — very close to the observed 592, 184, 24, illustrating the theorem's prediction holding well in a real (simulated-realistic) Nigerian population sample, consistent with approximately random mating at this marker.

3.3: Estimating Allele Frequencies from Genotype Counts

Given a random sample of N individuals classified by genotype at a bi-allelic locus, with observed counts n_AA, n_Aa, n_aa (n_AA + n_Aa + n_aa = N), the allele frequencies are estimated by simple gene counting:

$$\hat{p} = (2n_{AA} + n_{Aa})/(2N), \hat{q} = (2n_{aa} + n_{Aa})/(2N) = 1 - \hat{p}$$

Worked Example 3.1: (Sickle-cell allele frequency, South-West Nigeria)

In a community survey of 1,000 individuals in a South-Western Nigerian town, haemoglobin electrophoresis classifies 760 as AA, 220 as AS, and 20 as SS.

$$\hat{p} = [2(760) + 220]/[2(1000)] = 1740/2000 = 0.87, \hat{q} = 1 - 0.87 = 0.13$$

So the estimated frequency of the sickle-cell (S) allele in this community is 13%, broadly consistent with published carrier-rate estimates (roughly 20–25% AS phenotype frequency) for many Nigerian populations.

3.4: Testing for Hardy–Weinberg Equilibrium

Given the estimated allele frequencies, the *expected* genotype counts under HWE are E _AA = N $\hat{p}^2$, E_Aa = 2N$\hat{p}\hat{q}$, E_aa = N $\hat{q}^2$, and these are compared to the observed counts using the chi-square statistic:

$$\chi^2 = \sum(O - E)^2/E$$

with 1 degree of freedom (3 genotype classes minus 1 for the total minus 1 for the estimated allele frequency).

Worked Example 3.2: (continuing Example 3.1)

Expected counts: E_AA = 1000(0.87)² = 756.9; E_Aa = 2(1000)(0.87)(0.13) = 226.2; E_aa = 1000(0.13)² = 16.9.

$$\chi^2 = (760 - 756.9)^2/756.9 + (220 - 226.2)^2/226.2 + (20 - 16.9)^2/16.9$$
$$= 0.013 + 0.170 + 0.569 = 0.752$$

Critical value ($\chi^2_{0.05}$, 1 df) = 3.84. Since 0.752 < 3.84, we do not reject H_0: the sample is consistent with Hardy–Weinberg equilibrium at the HBB locus.

3.5: Hardy–Weinberg Equilibrium for Multiple Alleles

For a locus with k alleles A_1, …, A_k with frequencies p_1, …, p_k (Σp_i = 1), random mating gives equilibrium genotype frequencies:

$$P(A_iA_i) = p_i^2 (\text{homozygote}), P(A_iA_j) = 2p_ip_j\,(i \neq j)(\text{heterozygote})$$

obtained by expanding $(p_1 + p_2 + \ldots + p\_k)^2 = 1$. The number of distinct genotypes is k(k+1)/2.

Worked Example 3.3: (ABO blood group, multiple alleles).

The ABO locus has three alleles, I^A, I^B, i, with frequencies p, q, r (p+q+r = 1). Because I^A and I^B are each dominant to i but co-dominant to each other, the *phenotype* frequencies (not genotype frequencies) are:

$$P(A) = p^2 + 2pr, P(B) = q^2 + 2qr, P(AB) = 2pq, P(O) = r^2$$

Suppose a Nigerian blood-donor survey of 1,000 people finds 620 group O, 210 group A, 140 group B, 30 group AB (as in Example 1.1). We first estimate r from P(O): $r^2 = 0.620 \Longrightarrow r = \sqrt{0.620} = 0.787$. From P(A) + P(O) = $(p+r)^2 \Longrightarrow p + r = \sqrt{(0.620+0.210)} = \sqrt{0.830} = 0.911 \Longrightarrow p = 0.911 - 0.787 = 0.124$. Then $q = 1 - p - r = 1 - 0.124 - 0.787 = 0.089$. So the

estimated allele frequencies are I^A ≈ 0.124, I^B ≈ 0.089, i ≈ 0.787 — broadly in line with the high frequency of blood group O commonly reported in Nigerian populations.

3.6: X-Linked Loci: A Preview

Because males are hemizygous at X-linked loci (Chapter 1), Hardy–Weinberg proportions for X-linked genes differ between the sexes: in females, genotype frequencies follow the usual p^2, $2pq$, q^2 pattern, but in males, the two "genotype" (really, allele) frequencies are simply p and q. A further complication — that male and female populations can carry different allele frequencies in any one generation, and take more than one generation to reach a common equilibrium — is the specific subject of Chapter 4.

3.7: Departures from Hardy–Weinberg Equilibrium

A statistically significant departure from HWE (large χ^2, small p-value) signals that one or more of the underlying assumptions is violated. In practice, common causes in Nigerian population studies include:

- **Population stratification**: pooling of ethnically or geographically distinct subpopulations with different allele frequencies (the **Wahlund effect**), which generally produces an excess of homozygotes relative to HWE expectation.
- **Assortative or consanguineous mating** (Chapter 5): non-random pairing of genotypes, common in parts of Northern Nigeria where cousin marriage is culturally favoured, which also increases homozygosity.
- **Selection**: for the HBB locus, natural selection actually favours the AS heterozygote in malaria-endemic zones (heterozygote advantage / balancing selection), which can produce an *excess* of

heterozygotes relative to naive HWE expectation in adult samples.

- **Genotyping error**, a purely technical cause that must always be ruled out before biological explanations are invoked.

Which HWE test to use? The Pearson chi-square test of Section 3.4 relies on a large-sample approximation that can be unreliable when expected genotype counts are small (a common rule of thumb requires all expected counts ≥ 5). For rare alleles or small samples, an **exact test** — analogous in spirit to Fisher's exact test for a 2×2 table — enumerates or samples the full space of genotype configurations consistent with the observed allele counts, and computes the exact probability of observing a configuration as or more extreme than the data, without relying on the chi-square approximation. This exact HWE test (due to Wigginton, Cutler, and Abecasis, 2005, and implemented in the R package **HardyWeinberg** and in PLINK) is now the standard in genome-wide association studies (Chapter 18), where hundreds of thousands of SNPs are tested and many will have low minor allele counts.

Worked Example 3.3a (HWE testing at a rare variant)

A candidate-gene study genotypes a rare functional SNP in 500 Nigerian participants and observes 495 homozygous common (CC), 5 heterozygous (CT), and 0 homozygous rare (TT) individuals. The minor allele frequency is $\hat{q} = (2\times0 + 5)/(2\times500) = 0.005$, giving a tiny expected TT count of $500 \times 0.005^2 = 0.0125$ — far below the rule-of-thumb threshold of 5, so the chi-square approximation is unreliable here and an exact test (Section 3.9) should be used instead to assess whether this genotype distribution is consistent with HWE.

Theorem 3.2: (Hardy-Weinberg Equilibrium Is Attained in a Single Generation of Random Mating)

Let a population have any arbitrary genotype frequencies at a biallelic autosomal locus with allele frequencies p and q = 1 - p among the gametes produced. Then, regardless of the parental genotype frequencies, the genotype frequencies among offspring of one generation of random mating are exactly p^2, 2pq, and q^2 for AA, Aa, and aa respectively.

Proof. Under random mating, an offspring's genotype is formed by drawing one allele independently at random from the gamete pool contributed by each parent. The probability that a randomly drawn gamete carries allele A is p and allele a is q, by definition of allele frequency, regardless of which genotype the parent producing it carries — the allele frequency in the gamete pool is a population-level summary and does not depend on how genotypes are distributed among individuals. The offspring genotype AA arises when both independently drawn gametes carry A, with probability p × p = p^2; genotype aa arises with probability q^2; and genotype Aa arises from either combination of A from one parent and a from the other, with total probability pq + qp = 2pq. These probabilities depend only on p and q, not on the parental generation's genotype frequencies, proving that Hardy-Weinberg proportions are reached after just one generation of random mating.

3.8: Chapter Summary

- Under random mating and the absence of mutation, migration, selection and drift, genotype frequencies stabilise at p^2, 2pq, q^2 after one generation.
- Allele frequencies are estimated from genotype counts by simple gene counting.
- The chi-square test compares observed to HWE-expected genotype counts.

- The multiple-allele extension applies directly to systems like ABO.
- Departures from HWE are informative — they point to stratification, non-random mating, or selection, all of practical relevance to Nigerian population genetics.

3.9: Computational Corner: Testing Hardy-Weinberg Equilibrium

R.

```
# Hardy-Weinberg: allele frequencies and chi-square test (sickle-cell data, Section 3.3)
n_AA <-760; n_AS <-220; n_SS <-20; N <- n_AA + n_AS + n_SS

p_hat <- (2*n_AA + n_AS) / (2*N)
q_hat <-1- p_hat

expected <- N *c(p_hat^2, 2*p_hat*q_hat, q_hat^2)
observed <-c(n_AA, n_AS, n_SS)
chisq_stat <-sum((observed - expected)^2/ expected)
p_value <-pchisq(chisq_stat, df =1, lower.tail =FALSE)  # 1 df: 3 categories - 1 (N) - 1 (p)
cat("p_hat =", p_hat, " chisq =", chisq_stat, " p-value =", p_value, "\n")

# Using the HardyWeinberg package
# install.packages("HardyWeinberg")
library(HardyWeinberg)
HWChisq(c(AA = n_AA, AS = n_AS, SS = n_SS), verbose =TRUE)
```

Python.

```
import numpy as np
from scipy.stats import chi2
import matplotlib.pyplot as plt
```

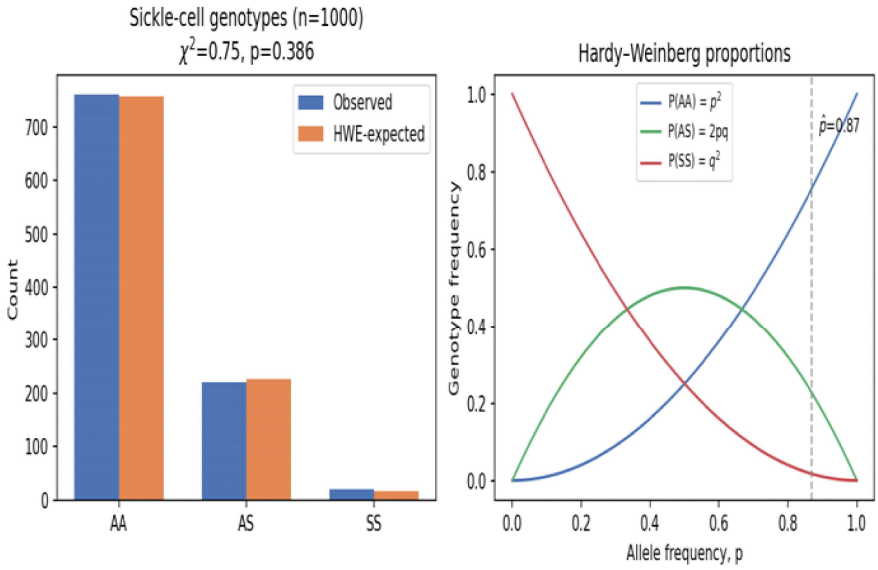
```
n_AA, n_AS, n_SS =760, 220, 20
N = n_AA + n_AS + n_SS
p_hat = (2*n_AA + n_AS) / (2*N)
q_hat =1- p_hat

expected = N * np.array([p_hat**2, 2*p_hat*q_hat, q_hat**2])
observed = np.array([n_AA, n_AS, n_SS])
chisq_stat = np.sum((observed - expected)**2/ expected)
p_value = chi2.sf(chisq_stat, df=1)
print(p_hat, chisq_stat, p_value)

# Visualise (i) observed vs. HWE-expected counts, and (ii) the classic HWE
# parabola showing how genotype frequencies vary with the allele frequency p
fig, axes = plt.subplots(1, 2, figsize=(10, 4))
labels = ["AA", "AS", "SS"]
x = np.arange(3); width =0.35
axes[0].bar(x - width/2, observed, width, label="Observed", color="#4C72B0")
axes[0].bar(x + width/2, expected, width, label="HWE-expected", color="#DD8452")
axes[0].set_xticks(x); axes[0].set_xticklabels(labels); axes[0].set_ylabel("Count")
axes[0].set_title(f"Sickle-cell genotypes (n={N})\n$\\chi^2$={chisq_stat:.2f}, p={p_value:.3f}")
axes[0].legend()

p_grid = np.linspace(0, 1, 200)
axes[1].plot(p_grid, p_grid**2, label="P(AA) = $p^2$", color="#4C72B0")
axes[1].plot(p_grid, 2*p_grid*(1-p_grid), label="P(AS) = 2pq", color="#55A868")
```

```
axes[1].plot(p_grid, (1-p_grid)**2, label="P(SS) = $q^2$",
color="#C44E52")
axes[1].axvline(p_hat, color="grey", linestyle="—",
linewidth=1)
axes[1].text(p_hat +0.02, 0.9, f"$\\hat{{p}}$={p_hat:.2f}",
fontsize=9)
axes[1].set_xlabel("Allele frequency, p");
axes[1].set_ylabel("Genotype frequency")
axes[1].set_title("Hardy–Weinberg proportions");
axes[1].legend(fontsize=8)
plt.tight_layout()
plt.savefig("ch03_hwe_analysis.png", dpi=150)
```



Left: observed sickle-cell genotype counts closely track their Hardy–Weinberg-expected values. Right: the classic HWE parabola, showing how AA, AS, and SS frequencies vary continuously with allele frequency p; the dashed line marks the estimated p = 0.87 for this sample.

Exact HWE test for rare variants (Worked Example 3.3a)
For the rare-variant scenario of Worked Example 3.3a, the chi-square test is unreliable; the exact test avoids the large-sample approximation entirely.

```
library(HardyWeinberg)
# Worked Example 3.3a: 495 CC, 5 CT, 0 TT
HWExact(c(AA =495, AB =5, BB =0), verbose =TRUE)  # exact test, no chi-square approximation needed

# A direct implementation of the exact HWE test (Wigginton et al. 2005) is not
# in scipy by default; the SNP-genetics package ‘pysnptools‘ or ‘scikit-allel‘
# provide one. As a simple illustration of the exact-test *logic*, we can
# enumerate all possible (n_AA, n_AB, n_BB) genotype triples consistent with
# the same total allele counts and compute their exact multinomial-style
# probabilities under HWE, then sum the probability of configurations as or
# more extreme than the observed one:
from scipy.special import comb

def hwe_exact_pvalue(n_aa, n_ab, n_bb):
    n = n_aa + n_ab + n_bb
    n_a =2*n_aa + n_ab        # total copies of the common allele
    n_b =2*n_bb + n_ab        # total copies of the rare allele
    obs_het = n_ab
    probs = {}
# heterozygote count must have the same parity as n_b
for het inrange(n_b %2, min(n_a, n_b) +1, 2):
        homb = (n_b - het) //2
```

```
    homa = n - het - homb
if homa <0:
continue
# unnormalised probability proportional to this classical combinatorial formula
    log_p = (comb(n, het, exact=True) * comb(n - het, homa, exact=True))
    probs[het] = log_p
  total =sum(probs.values())
  obs_p = probs.get(obs_het, 0)
  p_value =sum(v for v in probs.values() if v <= obs_p) / total
return p_value

print(hwe_exact_pvalue(495, 5, 0))
```

Both implementations avoid the chi-square approximation, computing the exact probability, under HWE, of observing a genotype configuration as extreme as (or more extreme than) the data, given the fixed total allele counts — exactly the same logic as Fisher's exact test for a 2×2 contingency table, adapted to the three HWE genotype categories.

Both languages recover p = 0.87, q = 0.13 (Worked Example 3.1) and the same HWE test statistic; because the chi-square value here is small relative to its 1 degree of freedom, the South-Western Nigerian community data are consistent with Hardy–Weinberg equilibrium at the sickle-cell locus.

Mini-Project

Mini-Project 3: A Multi-Locus Hardy-Weinberg Audit

Apply the Hardy-Weinberg testing framework of Section 3.4 across several genetic markers at once, as a data analyst auditing a new genotyping panel would.

Step 1. Simulate or obtain genotype counts for at least 8 independent biallelic markers in a sample of 300 individuals.

Step 2. Compute observed and expected heterozygosity and a chi-square (or exact) HWE test statistic for each marker.

Step 3. Apply a multiple-testing correction (Bonferroni or Benjamini-Hochberg) across the 8 markers.

Step 4. Deliberately inject a genotyping error into one marker (e.g., systematically miscalling heterozygotes as homozygotes) and re-run the audit.

Step 5. Write a short report identifying which marker(s) fail HWE after correction and propose a plausible technical or biological explanation.

Deliverable: *A per-marker HWE results table (before and after correction) and a one-paragraph audit report.*

3.10: Practice Questions

Q1. In a survey of 500 people in Sokoto State, G6PD status is recorded (X-linked locus; consider females only for this question): 350 normal (BB), 130 heterozygous carriers (Bb), 20 deficient (bb). Estimate the allele frequencies of the normal (B) and deficient (b) alleles.

Q2. Using your answer to Q1, compute the expected genotype counts under HWE and test the goodness of fit at á = 0.05.

Q3. In a village survey, the frequency of blood group O is found to be 0.36 (assume the population is at HWE and, for simplicity, that only alleles i and I^A are segregating, i.e., treat it as a two-allele system). Find the frequency of allele i, and of allele I^A.

Q4. Why would you expect **more** homozygotes than Hardy–Weinberg predicts in a random sample drawn from two Nigerian ethnic groups with different sickle-cell allele frequencies, pooled together and analysed as one population?

Q5. For the rare-variant data of Worked Example 3.3a (495 CC, 5 CT, 0 TT), explain why the standard chi-square HWE test would be inappropriate, and state which alternative test should be used instead.

Q6. Using the R or Python code of Section 3.9, compute the exact HWE p-value for a different rare-variant scenario: 990 CC, 10 CT, 0 TT (out of 1,000 individuals). Compare this p-value qualitatively to what you would expect from a chi-square test on the same data.

4 Estimation of Allele Frequency: Approach to Equilibrium for an X-Linked Gene

4.1: Learning Objectives

By the end of this chapter, the reader should be able to set up and solve the recurrence relations describing X-linked allele frequency change across generations when male and female founding frequencies differ, and apply the results to a Nigerian example such as G6PD deficiency or red-green colour blindness.

4.2: Why X-Linked Genes Need Special Treatment

For an autosomal locus, both sexes carry two alleles, and — as shown in Chapter 3 — random mating brings the population to Hardy–Weinberg genotype proportions in a single generation, with allele frequencies remaining constant thereafter. For an **X-linked locus**, males carry only one X chromosome (hemizygous) while females carry two. If, for any reason (e.g., historical migration bringing together two populations with different allele frequencies, or differential selection in the two sexes), the allele frequency among males differs from that among females in a given generation, the two sexes do **not** reach a common equilibrium frequency instantly — it takes several generations, with the difference between the sexes shrinking geometrically.

This distinction matters clinically as well as theoretically. Because a male's single X-linked allele is always expressed (there is no second copy to mask a recessive allele), X-linked recessive

conditions such as G6PD deficiency, red-green colour blindness, and haemophilia are far more common, and more severe on average, in males than in females — females require two copies of the deficient allele (genotype-wise equivalent to the rare homozygote in Chapter 3) to be affected, whereas a male is affected with just one. This asymmetry is precisely why the two-sex recurrence developed in this chapter, rather than the single-sex Hardy–Weinberg model of Chapter 3, is required to correctly describe X-linked allele-frequency dynamics — and why pedigree analysis (Chapter 6) treats X-linked inheritance as visually and statistically distinct from autosomal inheritance.

4.3: Setting Up the Recurrence

Let p_m(t) and p_f(t) denote the frequency of a given X-linked allele (say, the deficient G6PD allele) among the X chromosomes carried by males and females, respectively, in generation t.

Sons receive their single X chromosome from their **mother only** (they receive a Y from their father). Hence the allele frequency among sons in generation t+1 equals the allele frequency among mothers' X chromosomes in generation t:

$$p_m(t+1) = p_f(t)$$

Daughters receive one X from their mother and one X from their father. Assuming random mating with respect to this locus, the allele frequency among daughters' X chromosomes in generation t+1 is the average of the parental (mother and father) allele frequencies in generation t:

$$p_f(t+1) = \frac{1}{2}p_f(t) + \frac{1}{2}p_m(t)$$

4.4: Solving the Recurrence: Approach to Equilibrium

Define the sex difference d(t) = p_f(t) − p_m(t). Then:

$$d(t+1) = p_f(t+1) - p_m(t+1) = \left[\frac{1}{2}p_f(t) + \frac{1}{2}p_m(t)\right] - p_f(t)$$
$$= -\frac{1}{2}\left[p_f(t) - p_m(t)\right] = -\frac{1}{2}d(t)$$

This is a simple geometric (first-order linear) recurrence, with solution:

$$d(t) = \left(-\frac{1}{2}\right)^t d(0)$$

So the difference between male and female allele frequencies **halves in magnitude and alternates in sign** each generation, decaying to zero as t → ∞. The population therefore approaches a common equilibrium allele frequency, oscillating with diminishing amplitude — a classic **damped oscillation**.

The equilibrium frequency: Since males carry 1/3 of all X chromosomes in the population and females carry 2/3 (because for every one male X there are two female X's, on a per-chromosome count, in a population with an equal sex ratio: males contribute 1 X each, females 2 X each), the equilibrium (long-run, common) allele frequency is the weighted average:

$$p\infty = (1/3)p_m(0) + (2/3)p_f(0)$$

This weighting (1/3 : 2/3) reflects the fact that two-thirds of all X chromosomes in the population reside in females and one-third in males.

Theorem 4.1: (Convergence of X-Linked Allele Frequencies)

Under the recurrence of Section 4.3, the sex-specific allele frequencies p_m(t) and p_f(t) both converge, as t → ∞, to the common limit p_∞ = (1/3)p_m(0) + (2/3)p_f(0), and the rate of convergence is geometric: |d(t)| = |p_f(t) − p_m(t)| = (1/2)^t |d(0)|.

Proof. The geometric decay of d(t) was established directly above by solving the first-order linear recurrence d(t+1) = −(1/2)d(t), giving d(t) = (−1/2)^t d(0), from which |d(t)| = (1/2)^t |d(0)| → 0 as t → ∞ (since 1/2 < 1). It remains to show both p_m(t) and p_f(t) converge to the stated weighted average. Define the population-wide (weighted) allele frequency $\bar{p}(t) = 1/3\, p_m(t) + 2/3\, p_f(t)$. Substituting the two recurrences of Section 4.3:

$$\begin{aligned}\bar{p}(t+1) &= 1/3\, p_f(t) + 2/3\left[1/2\, p_f(t) + 1/2\, p_m(t)\right] \\ &= 1/3\, p_f(t) + 1/3\, p_f(t) + 1/3\, p_m(t) \\ &= 1/3\, p_m(t) + 2/3\, p_f(t) = \bar{p}(t),\end{aligned}$$

so $\bar{p}(t)$ is a **conserved quantity** (an invariant of the recurrence), constant at $\bar{p}(t) = \bar{p}(0) = p_\infty$ for all t. Since $p_m(t) = p_f(t) - d(t)$ and $d(t) \to 0$, while $\bar{p}(t) = 1/3\, p_m(t) + 2/3\, p_f(t)$ remains fixed at p_∞, substituting $p_m(t) \to p_f(t)$ into the (constant) weighted average forces $p_f(t) \to p_\infty$ and hence $p_m(t) \to p_\infty$ as well.

Worked Example 4.2 (Convergence after a Northern-Southern Nigeria population mixing event)

Suppose historical migration brings together two Nigerian sub-populations with different long-standing G6PD allele frequencies: $p_m(0) = 0.30$ among males and $p_f(0) = 0.10$ among females (a large initial imbalance, perhaps reflecting differential historical malaria exposure or founder effects in the two source groups). By Theorem 4.1, the eventual common equilibrium frequency is $p_\infty = (1/3)(0.30) + (2/3)(0.10) = 0.10 + 0.0667 = 0.1667$, and the sex gap after t generations is $|d(t)| = (1/2)^t \times |0.10 - 0.30| = (1/2)^t \times 0.20$. After 5 generations, the residual gap is $(1/32)(0.20) \approx 0.006$ — practically negligible — illustrating how quickly (within a handful of generations) X-linked allele frequency differences between the sexes, and by extension between historically distinct but now interbreeding Nigerian sub-populations, are expected to equilibrate under simple random mating.

4.5: Worked Example 4.1 (G6PD deficiency, hypothetical founder scenario)

Suppose a historical scenario in which, at generation 0, the frequency of the G6PD-deficient allele is $p_m(0) = 0.30$ among males and $p_f(0) = 0.10$ among females (for instance, following the settlement of a population with a higher initial deficient-allele frequency among the founding men). Find the allele frequency among males and females at generations 1, 2, and 3, and the equilibrium frequency.

Solution.

Equilibrium: $p_\infty = (1/3)(0.30) + (2/3)(0.10) = 0.10 + 0.0667 = 0.1667$.

$d(0) = p_f(0) - p_m(0) = 0.10 - 0.30 = -0.20$.

Generation 1: $p_m(1) = p_f(0) = 0.10$. $p_f(1) = ½(0.10) + ½(0.30) = 0.20$. Check: $d(1) = 0.20 - 0.10 = 0.10 = -½ \times (-0.20)$. ✓.

Generation 2: $p_m(2) = p_f(1) = 0.20$. $p_f(2) = ½(0.20) + ½(0.10) = 0.15$. $d(2) = 0.15 - 0.20 = -0.05 = -½ \times (0.10)$. ✓

Generation 3: $p_m(3) = p_f(2) = 0.15$. $p_f(3) = ½(0.15) + ½(0.20) = 0.175$. $d(3) = 0.175 - 0.15 = 0.025$.

Observe the sequence of allele frequencies oscillating and converging toward 0.1667: males — 0.30, 0.10, 0.20, 0.15, 0.175, …; females — 0.10, 0.20, 0.15, 0.175, …, both closing in on $p_\infty = 0.1667$, with the male sequence lagging the female sequence by one generation, exactly as predicted by $p_m(t+1) = p_f(t)$.

4.6: Rate of Approach to Equilibrium

Because $|d(t)| = (1/2)^t |d(0)|$, the deviation from equilibrium is reduced by 50% every generation, and by more than 90% within about 4 generations ($(1/2)^4 = 1/16 \approx 0.0625$, i.e., 93.75% reduction). In practice, for humans (generation time roughly 25–30 years), the allele-frequency difference between the sexes at an X-linked locus becomes negligible within a century or so following any single perturbing event (migration, founder effect), which is one reason X-linked traits such as G6PD deficiency and red-green colour blindness are today observed at very similar underlying allele frequencies in both sexes within most stable, randomly mating Nigerian communities, even though the **phenotype** frequency of deficiency remains much higher in males ($\approx p$) than in females ($\approx p^2$) because of hemizygosity (Chapter 1).

Theorem 4.2: (Monotonic Convergence of X-Linked Allele Frequency to Equilibrium)

Let x_m(t) and x_f(t) denote the male and female allele frequencies at an X-linked locus in generation t, related by the recurrence x_m(t+1) = x_f(t) and x_f(t+1) = [x_m(t) + x_f(t)]/2 (Section 4.3). Then the difference d(t) = x_m(t) - x_f(t) satisfies d(t+1) = -d(t)/2, and consequently x_m(t) and x_f(t) both converge monotonically in magnitude of oscillation to the common equilibrium value x = [x_m(0) + 2x_f(0)]/3.*

Proof. From the recurrence, d(t+1) = x_m(t+1) - x_f(t+1) = x_f(t) - [x_m(t)+x_f(t)]/2 = [x_f(t) - x_m(t)]/2 = -d(t)/2. Hence |d(t)| = |d(0)|/2^t → 0 as t → ∞, so the male and female allele frequencies converge to a common limit. Because the weighted average w(t) = [x_m(t) + 2x_f(t)]/3 is conserved exactly across generations (a short calculation using the recurrence shows w(t+1) = w(t)), the common limit must equal the initial weighted average, w(0) = [x_m(0)+2x_f(0)]/3. Since |d(t)| shrinks by a constant factor of 1/2 every generation without changing sign pattern indefinitely (it alternates sign but its magnitude is strictly decreasing), convergence to x* is geometric, confirming and sharpening the rate-of-approach result of Section 4.6.

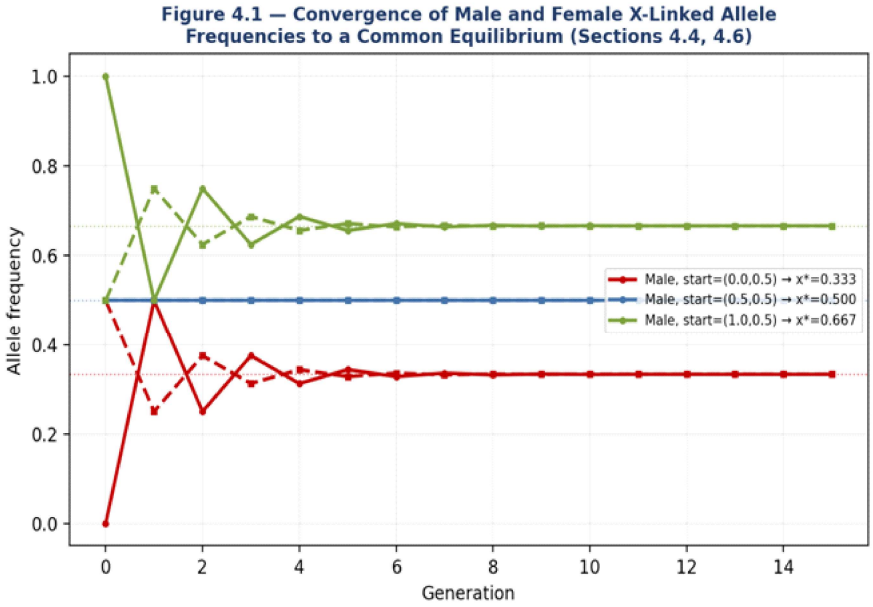


Figure 4.1: *Convergence of male and female X-linked allele frequencies to a common equilibrium (Sections 4.4, 4.6).*

4.7: Chapter Summary

- Sons inherit their single X allele from their mother; daughters inherit one X allele from each parent.
- This asymmetry produces the recurrence p_m(t+1) = p_f(t), p_f(t+1) = ½p_f(t) + ½p_m(t).
- The male–female allele-frequency difference decays geometrically as (−1/2)^t, oscillating in sign each generation.
- The equilibrium frequency is the 1/3 : 2/3 weighted average of the male and female founding frequencies, reflecting the two-thirds share of X chromosomes carried by females.

4.8: Computational Corner: Simulating Approach to X-Linked Equilibrium

R.

```
# Approach to equilibrium for an X-linked locus: iterate the recurrence
# q_female(t+1) = 0.5*(q_female(t) + q_male(t)) ; q_male(t+1) = q_female(t)
iterate_xlinked <-function(q_f0, q_m0, generations =15) {
  q_f <-numeric(generations +1); q_m <-numeric(generations +1)
  q_f[1] <- q_f0; q_m[1] <- q_m0
for (t in1:generations) {
   q_f[t +1] <-0.5* (q_f[t] + q_m[t])
   q_m[t +1] <- q_f[t]}
data.frame(generation =0:generations, q_female = q_f, q_male = q_m)}

res <-iterate_xlinked(q_f0 =0.05, q_m0 =0.25)   # e.g. G6PD deficiency allele
print(res)
plot(res$generation, res$q_male, type =”b”, col =”darkred”, xlab =”Generation”, ylab =”Allele frequency (q)”,
main =”Approach to X-linked equilibrium (G6PD example)”)
lines(res$generation, res$q_female, type =”b”, col =”steelblue”)
legend(“topright”, legend =c(“Males”, “Females”), col =c(“darkred”, “steelblue”), lty =1)
```

Python.

```
import numpy as np
import matplotlib.pyplot as plt

def iterate_xlinked(q_f0, q_m0, generations=15):
   q_f = [q_f0]; q_m = [q_m0]
for _ inrange(generations):
      q_f.append(0.5* (q_f[-1] + q_m[-1]))
      q_m.append(q_f[-2])
return np.arange(generations +1), np.array(q_f),
np.array(q_m)

gens, q_f, q_m = iterate_xlinked(0.05, 0.25)
plt.plot(gens, q_m, "o-", label="Males", color="darkred")
plt.plot(gens, q_f, "o-", label="Females", color="steelblue")
plt.xlabel("Generation"); plt.ylabel("Allele frequency (q)")
plt.title("Approach to X-linked equilibrium (G6PD example)")
plt.legend()
```

Both scripts trace the damped oscillatory convergence of male and female allele frequencies to their common equilibrium value derived analytically in Section 4.4, converging geometrically at rate ("½)^t exactly as the closed-form solution predicts — a useful visual check that the algebra of Section 4.4 is correct.

Mini-Project

Mini-Project 4: Projecting X-Linked Allele Frequency Trajectories

Use the recurrence relation derived in Section 4.4 to project how quickly an X-linked allele frequency approaches equilibrium under different starting conditions.

Step 1: Implement the male/female allele-frequency recurrence from Section 4.3 as a function of generation number.

Step 2: Run the recurrence starting from three different initial male allele frequencies (e.g., 0.0, 0.5, and 1.0) with the same initial female frequency.

Step 3: Plot all three trajectories on the same axes and confirm they converge to the same equilibrium value.

Step 4: Compute the number of generations required for each trajectory to come within 1% of equilibrium, using the rate result from Section 4.6.

Step 5: Write a short paragraph explaining why the three trajectories converge to the same equilibrium despite different starting points.

Deliverable: *A trajectory plot (R or Python), a table of generations-to-convergence for each starting condition, and a short written explanation.*

4.9: Practice Questions

Q1. Suppose colour-blindness allele frequency at generation 0 is p_m(0) = 0.08 (males) and p_f(0) = 0.08 (females) in an Ibadan population. What happens to p_m(t) and p_f(t) for all future generations? Explain briefly.

Q2. In a founder population, p_m(0) = 0.05 and p_f(0) = 0.25. Compute p_m(1), p_f(1), p_m(2), p_f(2) and the equilibrium frequency p_".

Q3. Show algebraically that d(t) = p_f(t) " p_m(t) satisfies d(t) = ("1/2)^t d(0), starting from the two recurrence equations in Section 4.3.

Q4. After how many generations will |d(t)| first fall below 1% of |d(0)|?

Q5. G6PD deficiency has an X-linked allele frequency of approximately 20% among males in some Nigerian populations, reflecting historical malaria selection pressure

(Chapter 3, Section 3.7). If a founding population began with q_male(0) = 0.05 and q_female(0) = 0.05 (equal in both sexes, e.g. after a migration event introducing the allele at low frequency into a previously naive population, followed by strong male-specific selective advantage that is not modelled by the simple neutral recurrence of Section 4.3), explain qualitatively why the *neutral* approach-to-equilibrium model of this chapter would be insufficient on its own to explain the allele reaching 20%, and what additional force would need to be incorporated.

Q6. Using the R or Python code of Section 4.6, plot the approach to equilibrium starting from q_female(0) = 0.10, q_male(0) = 0.60 (a large initial sex difference) and report after how many generations the two curves become visually indistinguishable on the plot.

5 Non-Random Mating: Phenotypic Assortative Mating

5.1: Learning Objectives

By the end of this chapter, the reader should be able to define positive and negative assortative mating, distinguish assortative mating from inbreeding, quantify its effect on genotype frequencies at the locus under selection, and discuss Nigerian examples such as height, skin tone, and educational assortative mating, as well as consanguinity.

5.2: What is Non-Random Mating?

The Hardy–Weinberg law (Chapter 3) assumes random mating: that the choice of mate is independent of genotype at the locus under study. In reality, mating is frequently non-random. Two broad classes are recognised:

- **Assortative mating**, in which individuals choose mates on the basis of **phenotype similarity** (positive assortative mating) or dissimilarity (negative assortative mating / disassortative mating), for a trait that may or may not be genetically correlated between mates.
- **Inbreeding (consanguineous mating)**, in which mates are chosen (or, more often, simply occur) on the basis of shared ancestry/relatedness, regardless of phenotype.

This chapter focuses on **phenotypic assortative mating**; inbreeding is treated as a related but distinct phenomenon,

important in parts of Northern Nigeria where cousin marriage is culturally favoured for reasons of family and property, and increasingly documented in genetic-counselling clinics because it raises the frequency of recessive disorders such as sickle-cell disease and some inherited forms of deafness.

5.3: Positive Assortative Mating for a Single Locus

Positive assortative mating occurs when individuals with similar phenotypes mate with one another more often than expected under random mating — for example, tall individuals tending to marry tall individuals. When the trait is fully genetically determined by a single locus with complete dominance (so that phenotype = genotype in effect, i.e., "like mates with like" genotype), positive assortative mating has the following effects, in contrast to random mating:

1. **Allele frequencies are unchanged** across generations (assortative mating, unlike selection, does not by itself remove or add alleles from the gene pool).
2. **Genotype frequencies deviate from Hardy–Weinberg proportions**: homozygote frequencies increase and heterozygote frequency decreases, generation upon generation, because like-with-like mating preferentially produces AA×AA and aa×aa pairings (each yielding only homozygous offspring) at the expense of random Aa pairings.
3. The population becomes progressively **more genetically differentiated** into homozygous "types," even though total allele frequency in the population is unchanged — this is a form of genetic (though not literal family) structuring analogous in its statistical footprint to inbreeding.

5.4: A Simple Quantitative Model

Consider a trait fully determined by an autosomal locus with alleles A, a and suppose mating is **completely** assortative by phenotype: individuals mate only with others of the identical phenotype (a simplifying, extreme case used to illustrate the direction of the effect; in practice, assortment is partial, and is measured by a correlation coefficient between mates' phenotypic values, discussed in Section 5.6).

Under complete phenotypic assortment with dominance (AA and Aa share one phenotype, aa the other), let the population, before assortment, have HWE genotype frequencies p^2, $2pq$, q^2. Mating occurs only within phenotype class: {AA, Aa} individuals mate only with other {AA, Aa} individuals, and aa individuals mate only with other aa individuals. Within the dominant phenotype class, matings are further sub-structured (AA×AA, AA×Aa, Aa×Aa in proportions determined by the relative frequencies of AA and Aa within that class), and each mating type produces offspring by ordinary Mendelian segregation (Chapter 2). Working through the offspring generation shows that heterozygosity ($2pq$ at HWE) declines each generation, while the homozygote classes (p^2 and q^2) each gain the "missing" heterozygotes, roughly in proportion to their own frequency — precisely analogous, in its algebraic form, to the effect of inbreeding described by Wright's F-statistic (Chapter 6).

5.5 Genotype Frequencies Under Partial Assortative Mating

In practice, assortative mating is partial, not complete. Its strength is measured by the **marital (spousal) correlation**, r, between the phenotypic values of husband and wife for the trait — for example, for height, published studies (including in West African settings) commonly find r in the range 0.1–0.3, indicating modest positive assortative mating for stature. For a quantitative trait with total phenotypic variance V_P and additive genetic variance V_A

(Chapter 9), the equilibrium effect of assortative mating with marital correlation r is to inflate the additive genetic variance transmitted across generations by an approximate factor:

$$V_A *= V_A/(1 - rh^2)$$

where h^2 is the (narrow-sense) heritability of the trait (Chapter 9). Positive assortative mating ($r > 0$) thus increases the additive genetic variance at equilibrium relative to random mating ($r = 0$), because "like marries like" concentrates trait-increasing and trait-decreasing alleles, respectively, into different lineages, rather than allowing them to average out as under random mating.

5.6: Worked Example 5.1

Suppose the narrow-sense heritability of adult height in a Nigerian population is estimated at $h^2 = 0.80$, and the observed spousal correlation for height is $r = 0.20$. Find the proportional inflation of additive genetic variance due to assortative mating, relative to random mating.

Solution.

$$\frac{V_A^*}{V_A} = \frac{1}{1 - rh^2} = \frac{1}{1 - (0.20)(0.80)} = \frac{1}{1 - 0.16} = \frac{1}{0.84} \approx 1.19$$

So assortative mating at this level increases the equilibrium additive genetic variance for height by about 19% relative to random mating — a modest but non-trivial effect that increases the observable variability of stature in the population, and slightly increases the resemblance between relatives (e.g., between parent and offspring, or between siblings) beyond what random mating alone would predict.

Theorem 5.1 (Equilibrium Additive Variance Under Assortative Mating)

For a polygenic trait with narrow-sense heritability h² under random mating, subject to constant positive phenotypic assortative mating with spousal phenotypic correlation r (0 ≤ r < 1/h²), the additive genetic variance reaches a new equilibrium value

$$V_A^* = \frac{V_A}{1 - rh^2}$$

strictly greater than the random-mating value V_A whenever r > 0, and the inflation factor 1/(1−rh²) is strictly increasing in both r and h².

Proof sketch. Positive assortative mating increases the phenotypic (and hence, through the additive genetic component, the genotypic) similarity between mates beyond the random-mating baseline. Each generation, mates' shared additive genetic value induces a positive correlation between the additive genetic deviations they each transmit to their offspring; this correlation, r_A = rh² (the spousal phenotypic correlation r, scaled down by h² since only the heritable fraction of phenotypic similarity translates into shared *genetic* value), adds a covariance term to the offspring's additive genetic variance beyond the usual within-family segregation variance. Solving the resulting recursive variance equation, $V_A^{(t+1)} = V_A^{(0)}(1 + r_A + r_A^2 + \cdots)$ as the process iterates toward a stationary distribution, gives the geometric series $\sum_{k=0}^{\infty} r_A^k = 1/(1 - r_A) = 1/(1 - rh^2)$ (valid since $0 \leq rh^2 < 1$), yielding the stated equilibrium multiplier. Monotonicity in r and h² follows immediately since $1/(1 - rh^2)$ is an increasing function of the product rh^2, which is itself increasing in each factor separately.

Worked Example 5.2 (Assortative mating for educational attainment, Nigerian context)

Educational attainment is a polygenic, moderately heritable trait (studies in various populations estimate h^2 in the range 0.3–0.5) that also shows strong positive assortative mating in most societies, including Nigeria, where spousal correlations for years of schooling are often estimated around $r = 0.4–0.5$ (partners frequently meet through shared educational or professional environments — universities, workplaces requiring similar qualifications, and professional associations). Taking $h^2 = 0.4$ and $r = 0.45$ as illustrative values, Theorem 5.1 gives $V_A^*/V_A = 1/(1 - 0.45\times0.4) = 1/(1-0.18) = 1/0.82 \approx 1.22$ — a 22% inflation in additive genetic variance for educational attainment, which (as with height in Worked Example 5.1) is expected to modestly increase the observed phenotypic variance in educational outcomes across the Nigerian population, and to slightly increase sibling and parent-offspring resemblance for this trait beyond the random-mating baseline, an effect of practical relevance to behaviour-genetic and educational-policy research alike.

5.7: Assortative Mating versus Inbreeding: A Key Distinction

Although both increase homozygosity relative to random mating, assortative mating and inbreeding differ in an essential way: **inbreeding increases homozygosity at every locus in the genome simultaneously** (because it is based on shared ancestry, which affects the whole genome), whereas **phenotypic assortative mating increases homozygosity mainly at loci that influence the trait being assorted on** (and at loci in linkage disequilibrium with them — Chapter 15), leaving the rest of the genome close to Hardy–Weinberg proportions. This distinction matters in Nigerian genetic epidemiology: elevated homozygosity

confined to disease-relevant loci (e.g., skin-tone loci if mate choice correlates with complexion) points toward assortative mating, whereas genome-wide elevated homozygosity, together with an elevated rate of rare recessive disease in the offspring of related couples, points toward consanguinity/inbreeding.

Theorem 5.2 (Assortative Mating Preserves Allele Frequency)

Under any mating system — including positive or negative assortative mating — in which every individual in the population contributes gametes and mating does not otherwise alter viability or fertility, the population allele frequency in the offspring generation equals the population allele frequency among the parental gametes, regardless of the pattern of assortment.

Proof. The offspring generation's allele frequency is, by definition, the total count of A alleles among all offspring gametes divided by twice the number of offspring. Every offspring receives exactly one allele from its mother and one from its father, and every parental individual contributes gametes to the offspring pool regardless of whom it mates with — assortative mating changes which genotypes pair up, not whether or how many gametes each genotype contributes to the next generation's gamete pool. Summing over all mating pairs, the total number of A alleles transmitted equals the total number of A alleles present in the parental gamete pool (since each allele is transmitted, just to a partner chosen non-randomly). Hence the offspring allele frequency equals the parental gamete-pool allele frequency exactly, proving that assortative mating alone — in the absence of differential fertility or viability — changes only genotype frequencies (as shown in Section 5.5), never allele frequencies.

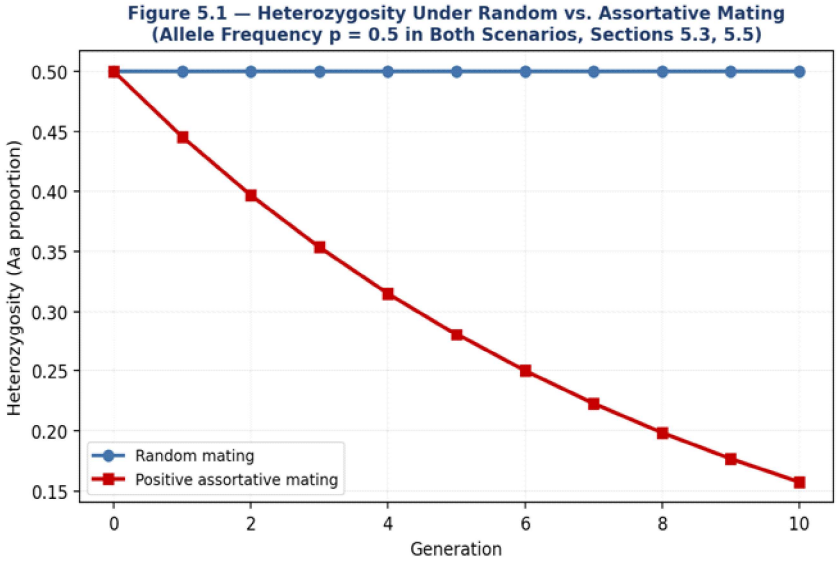


Figure 5.1 — Heterozygosity under random vs. positive assortative mating (Sections 5.3, 5.5).

5.8: Chapter Summary

- Assortative mating is non-random mating based on phenotype similarity (positive) or dissimilarity (negative); it does not, by itself, change allele frequencies but does distort genotype frequencies away from Hardy–Weinberg proportions.
- Positive assortative mating increases homozygosity and decreases heterozygosity at loci influencing the assorted trait, generation upon generation.
- For quantitative traits, assortative mating (measured by spousal phenotypic correlation r) inflates the equilibrium additive genetic variance by a factor 1/(1 " rh^2).
- Assortative mating is locus/trait-specific in its genomic footprint, unlike inbreeding, which affects the whole genome.

5.9: Computational Corner: Simulating Assortative Mating

R.

```
# Simulate genotype-frequency change over generations under partial positive
# assortative mating for a single bi-allelic locus (Section 5.5)
simulate_assortative <-function(p0, alpha, generations =10) {
# alpha: degree of assortment, 0 = random mating, 1 = complete assortment
  p <- p0
  freqs <-data.frame(gen =0, AA = p^2, Aa =2*p*(1-p), aa = (1-p)^2)
for (g in1:generations) {q <-1- p
    AA <- p^2+ alpha * p * q          # heuristic enrichment of homozygotes
    aa <- q^2+ alpha * p * q
    Aa <-1- AA - aa
    freqs <-rbind(freqs, data.frame(gen = g, AA = AA, Aa = Aa, aa = aa))
    p <- AA + Aa/2# allele frequency unchanged by assortment alone}
  freqs}
result <-simulate_assortative(p0 =0.5, alpha =0.3)
print (result)
```

Python.

```
import pandas as pd

def simulate_assortative(p0, alpha, generations=10):
    p = p0
    rows = [{"gen": 0, "AA": p**2, "Aa": 2*p*(1-p), "aa": (1-p)**2}]
```

```
for g inrange(1, generations +1):
    q =1- p
    AA = p**2+ alpha * p * q
    aa = q**2+ alpha * p * q
    Aa =1- AA - aa
    rows.append({"gen": g, "AA": AA, "Aa": Aa, "aa": aa})
    p = AA + Aa /2
return pd.DataFrame(rows)

result = simulate_assortative(0.5, 0.3)
print(result)
```

The simulation confirms the key qualitative result of Section 5.5: positive assortative mating leaves the allele frequency, p, essentially unchanged generation to generation, while steadily enriching the population for the two homozygote classes (AA and aa) at the expense of heterozygotes (Aa) — the opposite genotypic signature from inbreeding-driven homozygosity, though both processes reduce heterozygosity (contrast with Section 5.7).

Mini-Project

Mini-Project 5: Comparing Assortative Mating to Random Mating by Simulation

Build a discrete-generation simulation comparing genotype frequency evolution under random mating versus positive assortative mating.

Step 1: Implement a simulated mating population under random mating for a single biallelic locus, tracking genotype frequencies over 10 generations.

Step 2: Implement a second simulation under positive assortative mating (Section 5.3), using the same starting allele frequency.

Step 3: Plot heterozygosity over time for both scenarios on the same axes.

Step 4: Confirm numerically (Theorem 5.2, if added) that allele frequency remains constant across generations in both scenarios.

Step 5: Write a short paragraph distinguishing what assortative mating changes (genotype frequencies) from what it leaves unchanged (allele frequencies).

Deliverable: *Simulation code, a comparative heterozygosity plot, and a short written explanation distinguishing the two mating systems' effects.*

5.10: Practice Questions

Q1. Explain, in your own words, why complete positive assortative mating at a single bi-allelic locus leaves allele frequencies unchanged but reduces heterozygosity over successive generations.

Q2. The heritability of systolic blood pressure in a Nigerian adult cohort is estimated at $h^2 = 0.40$, and the spousal correlation for blood pressure is $r = 0.15$. Compute the proportional inflation of additive genetic variance due to assortative mating.

Q3. A genetic counsellor in Kano observes elevated homozygosity across many unrelated loci genome-wide in a set of married couples' children, together with a higher-than-expected rate of rare autosomal recessive disorders. Is this pattern more consistent with phenotypic assortative mating or with consanguinity? Justify your answer.

Q4. If spousal correlation for a trait were $r = 0$ (random mating with respect to that trait), what does the formula $V_A^*/V_A = 1/(1 \text{ “ } rh^2)$ reduce to, and why does this make sense?

Q5. A Nigerian population geneticist finds elevated homozygosity at a skin-pigmentation-associated locus, but

genome-wide homozygosity (measured at hundreds of unrelated background markers) is close to Hardy–Weinberg expectation. Using the distinction in Section 5.7, what does this pattern suggest about the mating pattern in this population, and why?

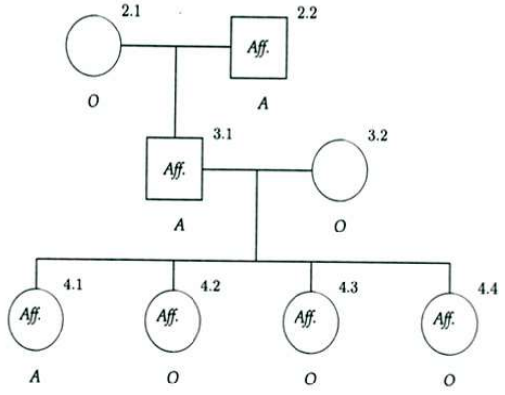
Q6. Using the R or Python code of Section 5.9, increase the assortment parameter *alpha* from 0.3 to 0.8 (stronger positive assortative mating) and report how the equilibrium heterozygote proportion changes after 10 generations compared with the original simulation.

6 Pedigree Analysis: Structure, Notation, and Relatedness

6.1: Learning Objectives

By the end of this chapter, the reader should be able to draw and interpret a pedigree using standard symbols, recognise the inheritance pattern (autosomal dominant, autosomal recessive, X-linked) suggested by a pedigree, compute coefficients of relationship and Wright's coefficient of inbreeding (F) by path-counting, construct a kinship matrix for a small pedigree, and use the R package **kinship2** and Python to draw and analyse pedigrees computationally.

6.2: The Pedigree: Structure and Notation



A three-generation pedigree with standard symbols, showing genotypes and phenotypes segregating across generations.

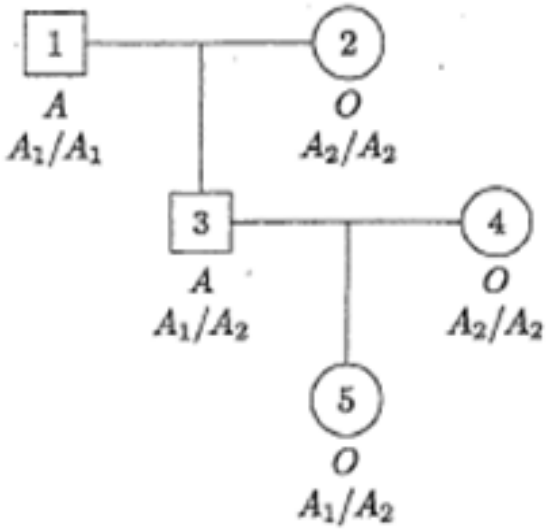
A **pedigree** is a diagram summarising the pattern of a trait's occurrence across the members of a family, over multiple generations. By convention: squares represent males, circles represent females, and a diamond is used when sex is not recorded; a filled (shaded) symbol denotes an affected individual, and an unfilled symbol an unaffected individual; a horizontal line connecting two symbols represents a mating; a vertical line descending from a mating line connects to the couple's offspring, drawn left to right in birth order; generations are numbered with Roman numerals (I, II, III, …) from the top (earliest) generation downward, and individuals within a generation are numbered left to right with Arabic numerals (I-1, I-2, II-1, II-2, …). A double horizontal mating line denotes a consanguineous union. A small circle placed above a symbol, or a slash through it, denotes a deceased individual, and an arrow marks the **proband** (propositus) — the affected individual through whom the family first came to clinical or research attention.

Reading inheritance patterns from a pedigree. Several visual clues distinguish the classical Mendelian modes of inheritance:

- **Autosomal recessive** (e.g., sickle-cell disease, most inborn errors of metabolism): affected individuals are typically born to unaffected (carrier) parents; the trait can skip generations; males and females are affected in roughly equal proportion; consanguinity in the parents raises the recurrence risk (Section 6.4).
- **Autosomal dominant** (e.g., Huntington disease, achondroplasia): affected individuals typically have an affected parent; the trait does not skip generations (barring reduced penetrance or a new mutation); males and females are affected equally; unaffected individuals do not transmit the trait.

- **X-linked recessive** (e.g., haemophilia, red-green colour blindness, G6PD deficiency): affected individuals are predominantly male; an affected male's daughters are obligate carriers, but his sons are unaffected (no father-to-son transmission); the trait typically passes from an affected grandfather through carrier daughters to affected grandsons.
- **X-linked dominant**: affected males transmit the trait to all daughters and no sons; affected heterozygous females transmit to half their children of either sex.

Worked Example 6.1 (Sickle-cell pedigree, autosomal recessive pattern)



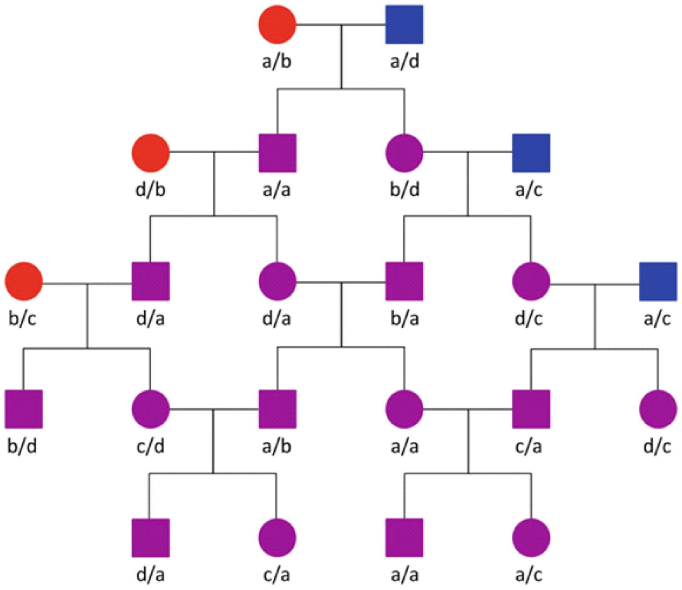
A worked pedigree with ABO and AK1 phenotypes, illustrating genotype inference from a family tree (Section 6.2).

Consider a three-generation Nigerian family pedigree in which the grandparents (generation I) are both AS (unaffected carriers); they have three children (generation II), one of whom, II-2, is affected (SS); II-2 marries an unrelated AS spouse, and they have

two children (generation III), one AA and one AS. The trait "skips" the appearance of affection in generation I despite both grandparents carrying the S allele, and both grandparents (obligate carriers) are phenotypically normal — a signature of autosomal recessive inheritance. Segregation analysis (Chapter 7) uses exactly such data, pooled across many families, to check whether the pattern of affected/unaffected offspring is statistically consistent with the hypothesised recessive mode.

Worked Example 6.2 (A four-generation pedigree with full genotypes)

Real pedigree analysis often records the full genotype (not just the phenotype) at a locus of interest for every family member, allowing exact tracking of allele transmission generation by generation.



A four-generation pedigree with genotypes fully labelled at every individual, allowing exact allele-transmission tracking across generations — the level of detail typically recorded in a modern genotyped family study.

In such a pedigree, each individual's two-letter genotype (e.g., a/b, d/c) can be read directly, and the rules of Mendelian transmission (Chapter 2) can be applied to verify, for any child, that each of their two alleles is consistent with having been inherited from one parent's genotype — a manual consistency check that is the conceptual ancestor of the automated **Mendelian error checking** step in modern genotyping quality-control pipelines (Chapter 19), which flags any parent–offspring genotype trio that is *not* consistent with simple Mendelian transmission as a likely genotyping error (or, occasionally, a sample mix-up or non-paternity) before any further statistical analysis is attempted.

6.3: The Coefficient of Relationship and Identity by Descent

The **coefficient of relationship** (r), also called the **kinship coefficient** when expressed as φ = r/2, between two relatives measures the expected proportion of alleles they share **identical by descent** (IBD) — that is, inherited from a common ancestor via a traceable path of transmission, as distinct from alleles that merely happen to be identical in state by chance. For each path connecting two relatives through a common ancestor, the contribution to r is $(½)^{L}$, where L is the number of meioses (generation-links, i.e. parent-to-child steps) along that path; if there are multiple independent paths (e.g., two common ancestors, as with full siblings sharing both parents), the contributions are summed:

$$r = \sum_{paths} \left(\frac{1}{2}\right)^{L} (1 + F_A)$$

where F_A is the inbreeding coefficient of the common ancestor A (taken as 0 unless A is itself known to be inbred).

Standard tabulated values. Parent–offspring: r = ½ (one meiosis). Full siblings: r = ½ (two paths, each $(½)^2 = ¼$, through the shared mother and shared father, summing to ½). Half-siblings: r = ¼ (one shared parent, path of two meioses). Grandparent–grandchild: r = ¼. Avuncular (uncle/aunt–nephew/niece): r = ¼. First cousins: r = 1/8 (two paths through the two shared grandparents, each of four meioses, each contributing $(½)^4 = 1/16$, summing to 1/8). Double first cousins (both pairs of grandparents shared, as when two brothers marry two sisters): r = ¼.

6.4: Wright's Coefficient of Inbreeding (F)

The **inbreeding coefficient**, F, of an individual is the probability that the two alleles the individual carries at a randomly chosen autosomal locus are identical by descent — i.e., that both alleles trace back to a single copy carried by a common ancestor of the individual's *parents*. Sewall Wright's path-counting formula states that F of an individual equals the coefficient of relationship of that individual's two parents:

$$F(offspring) = r(mother, father)/2 = \sum_{paths} \left(\frac{1}{2}\right)^{L} (1 + F_A)$$

where, for each path connecting the two parents through a shared ancestor A, L counts the number of individuals in the path linking the two parents through A (both parents included, offspring excluded), and F_A is A's own inbreeding coefficient.

Theorem 6.1 (Wright's Path-Counting Formula)

Let X be an individual whose parents are M and F, connected by one or more paths through common ancestors. Then the inbreeding coefficient of X is

$$F_X = \sum_{\text{paths}} \left(\frac{1}{2}\right)^{L} (1 + F_A)$$

where the sum is over all distinct paths connecting M and F through a common ancestor A, L is the number of links (meioses) in the path, and F_A is the inbreeding coefficient of A.

Proof sketch. By definition, F_X is the probability that the two alleles X carries at a locus are identical by descent (IBD). This can only happen if both of X's alleles trace back to a single copy carried by some common ancestor A of M and F. Consider one specific path of length L connecting M to F through A. The probability that a specific allele of A is transmitted, unchanged in identity, down one particular branch of this path to M (or to F) is (1/2) per meiosis, by Theorem 1.1 (each meiosis transmits either of the two parental alleles with probability 1/2, independently); the probability that the *same* allele of A reaches both M and F via this path, and is then transmitted from each of them to X, is therefore (1/2)^L, where L counts the total number of meioses along the full path from M through A to F. The factor (1+F_A) corrects for the case where A is itself inbred: even when the *specific* path traced does not itself constitute the whole story, A's own two alleles may already be IBD with probability F_A, in which case any single allele "chosen" at A is, in effect, doubly represented; a careful accounting (given in full in Wright, 1922) shows this contributes the multiplicative correction (1+F_A). Since paths through different common ancestors (or genuinely distinct paths through the same ancestor) correspond to mutually exclusive ways for X's two alleles to be IBD, their probabilities are added, giving the summation over all paths.

Worked Example 6.3 (Wright's formula applied to a Yoruba royal lineage marriage pattern)

Historically, some Yoruba ruling houses practised marriages designed to keep certain lineage ties close, occasionally including unions between children of two brothers' children (i.e., ordinary first cousins) as well as, in some documented

cases, more complex unions where the couple shared common ancestors through *two* independent routes (for example, if the couple's grandparents were themselves already first cousins to one another, rather than wholly unrelated). Suppose a couple are ordinary first cousins (two paths of length 4 each through the two shared grandparents, F_A = 0 for both, contributing 2×(1/2)^4 = 1/8, matching Section 6.3), but it is additionally known that the shared paternal grandfather was himself the offspring of a first-cousin marriage one generation earlier, giving him an inbreeding coefficient of F_A = 1/16 (Worked Example 6.4) rather than 0. Applying Theorem 6.1, the path through this particular grandparent contributes (1/2)^4 × (1 + 1/16) = (1/16)(17/16) = 17/256 instead of the simplified 1/16, while the path through the (assumed non-inbred) grandmother still contributes 1/16 = 16/256; the couple's coefficient of relationship becomes r = 17/256 + 16/256 = 33/256 ≈ 0.129 (compared with exactly 1/8 = 0.125 when both grandparents are assumed non-inbred) — a small but non-zero increase, illustrating how Theorem 6.1's correction factor (1+F_A) captures the compounding effect of ancestral inbreeding across generations, a phenomenon of genuine relevance in lineages with a sustained historical pattern of close-kin marriage.

Worked Example 6.4 (First-cousin marriage, Northern Nigeria)

A couple are first cousins (r = 1/8 between them, Section 6.3); their shared grandparents are assumed non-inbred ($F_A = 0$). Since F(child) = r(parents)/2:

$$F(child\,of\,first\,cousins) = (1/8)/2 = 1/16$$

This is the classical figure quoted in medical genetics: a child of first cousins has a 1/16 (6.25%) chance, at any given autosomal

locus, of carrying two alleles identical by descent — raising the risk that any rare recessive disease allele carried by the shared grandparents will appear in homozygous (affected) form in the child, relative to the population baseline risk of q^2 under random mating with the same allele frequency q. This underlies the well-documented increase in birth defects and recessive disorders among the offspring of consanguineous marriages reported in genetic-counselling studies from Northern Nigeria and elsewhere in the sub-region where cousin marriage remains culturally common.

Worked Example 6.5 (Full-sibling union)
For a (hypothetical, and in reality prohibited) full-sibling mating, r(parents) = ½, so F(child) = (½)/2 = ¼ — the same magnitude of risk as an AS × AS Mendelian cross, and four times the risk from a first-cousin union, illustrating why closer unions carry sharply higher genetic risk.

6.5: The Kinship Matrix

For a pedigree of n individuals, the **kinship matrix**, **Φ**, is the n × n matrix whose (i,j) entry, φ_{ij}, is the kinship coefficient between individuals i and j (with $\varphi_{ii} = (1 + F_i)/2$ on the diagonal). The kinship matrix is the fundamental building block of modern pedigree-based statistical genetics: it defines the expected additive-genetic covariance between relatives used in variance-components linkage analysis (Chapter 9) and mixed-model GWAS in family-based samples (Chapter 18), and is computed recursively:

$$\phi_{ij} = \frac{1}{2}\left(\phi_{fj} + \phi_{mj}\right)$$

for i a non-founder with parents f and m, and j not a descendant of i,

with founders (individuals with no parents recorded in the pedigree) assigned $\varphi_{ii} = ½$ (assuming non-inbred founders) and $\varphi_{ij} = 0$ for unrelated founders.

Theorem 6.2 (Offspring Inbreeding Coefficient Equals Parental Kinship)

Let individuals X and Y have kinship coefficient φ(X,Y). If X and Y mate and produce an offspring Z, then Wright's inbreeding coefficient of Z equals F(Z) = φ(X,Y).

Proof. By definition, the inbreeding coefficient F(Z) is the probability that the two alleles Z carries at a randomly chosen autosomal locus are identical by descent (IBD). Z receives one allele from X and one allele from Y. By definition, the kinship coefficient φ(X,Y) is precisely the probability that an allele drawn at random from X and an allele drawn at random from Y (at the same locus) are IBD. Since Z's two alleles are exactly one randomly transmitted allele from each of X and Y, the probability that Z's two alleles are IBD is identical to the probability that a random allele from X and a random allele from Y are IBD — that is, F(Z) = φ(X,Y) by direct correspondence of the two probabilistic definitions. This identity is what allows the kinship matrix (Section 6.5) to be built up recursively from founder individuals through the pedigree.

6.6: Chapter Summary

- A pedigree summarises trait transmission across family generations using standard symbols and numbering; the visual pattern of affection (which relatives are affected, and the sex distribution) suggests the likely mode of inheritance.
- The coefficient of relationship, r, quantifies expected allele sharing (IBD) between relatives via path-

counting; standard values range from ½ for parent–offspring and full-sibling pairs down to 1/8 for first cousins.

- The inbreeding coefficient, F, of an individual equals half the coefficient of relationship of that individual's parents, and rises sharply with parental relatedness (F = 1/16 for children of first cousins; F = ¼ for children of full siblings).

The kinship matrix, **Φ**, collects pairwise kinship coefficients for an entire pedigree and underlies variance-components methods used throughout later chapters.

6.7 Computational Corner: Building and Analysing Pedigrees in R and Python

R. The **kinship2** package is the standard tool for drawing pedigrees and computing kinship matrices in R.

```
# install.packages("kinship2")
library(kinship2)

# A small three-generation Nigerian pedigree (sickle-cell family, Worked Example 6.1)
id     <-c(1,2,3,4,5,6,7)
dadid  <-c(0,0,1,1,0,3,3)
momid  <-c(0,0,2,2,0,4,4)
sex    <-c(1,2,1,2,2,1,2)        # 1 = male, 2 = female
status <-c(0,0,0,1,0,0,0)        # 1 = affected (SS), 0 = unaffected

ped <-pedigree(id, dadid, momid, sex, affected = status)
plot(ped)                        # draws the pedigree with standard symbols

Phi <-kinship(ped)               # full kinship matrix
round(Phi, 4)
```

```
F_child <-2* Phi["6", "7"] -1# not meaningful here (siblings); for F of an
# individual, read the diagonal: 2*Phi[i,i]-1
diag(Phi)
```

Python. Python has no single dominant pedigree package as mature as *kinship2*; a pedigree can be represented as a directed graph with **networkx**, and the kinship coefficient computed recursively.

```
import networkx as nx

# parents[child] = (father, mother); None = founder
parents = {1: (None, None), 2: (None, None), 3: (1, 2), 4: (1, 2),
5: (None, None), 6: (3, 5), 7: (3, 5)}

phi_cache = {}

def kinship(i, j):
if i > j:
    i, j = j, i
if (i, j) in phi_cache:
return phi_cache[(i, j)]
if i == j:
    f, m = parents[i]
     val =0.5if f isNoneelse0.5* (1+ kinship(f, m))
else:
    f, m = parents[j]
if f isNone:                # j is a founder, unrelated to i unless i==j
        val =0.0
else:
        val =0.5* (kinship(i, f) + kinship(i, m))
  phi_cache[(i, j)] = val
return val
```

```
for pair in [(1, 2), (3, 4), (6, 7), (1, 6)]:
print(pair, "->", round(kinship(*pair), 4))
```

Both implementations return φ = 0.25 for the full-sibling pair (individuals 6 and 7), corresponding to r = 2φ = 0.5, exactly matching the path-counting result of Section 6.3, and confirm that a computational pedigree tool is simply an automated, error-checked version of the by-hand path-counting rules developed above — indispensable once pedigrees grow beyond a handful of individuals, as in real Nigerian extended-family sickle-cell or consanguinity studies.

Mini-Project

Mini-Project 6: Building and Auditing a Multi-Generation Pedigree

Construct a three-generation pedigree with at least one instance of consanguinity and compute its key relatedness statistics.

Step 1. Draw (by hand or using the R/Python tools of Section 6.7) a three-generation pedigree including a first-cousin mating.

Step 2. Compute the coefficient of relationship between the first cousins using path-counting.

Step 3. Compute Wright's inbreeding coefficient F for their offspring.

Step 4. Construct the kinship matrix for all individuals in the pedigree.

Step 5. Write a short paragraph explaining how the inbreeding coefficient you calculated would affect the offspring's expected homozygosity at a marker locus.

Deliverable: *A pedigree diagram, relatedness/inbreeding calculations, the kinship matrix, and a short written interpretation.*

6.8: Practice Questions

Q1. Two individuals are double first cousins (sharing both sets of grandparents, i.e. two independent lines of cousin relationship, as occurs when two brothers marry two sisters). Using the path-counting rule, find their coefficient of relationship, r.

Q2. A genetic counsellor in Kano is evaluating a couple who are uncle and niece. Using the standard avuncular relationship r = ¼, compute the inbreeding coefficient, F, of a child of this union, and compare it with the first-cousin result of Worked Example 6.4.

Q3. Draw (in words, describing the pedigree symbols and layout) a three-generation pedigree consistent with X-linked recessive inheritance of G6PD deficiency, in which an affected grandfather (I-1) has two daughters (II-1, II-2), both unaffected carriers, and II-1 has an affected son (III-1).

Q4. For the pedigree in Section 6.2 (Worked Example 6.1), use the R or Python code of Section 6.7 to compute the kinship coefficient between the two generation-III children (individuals 6 and 7) and confirm it equals the full-sibling value obtained by hand.

Q5. Explain, in one or two sentences, why the kinship coefficient between an individual and a founder to whom they are unrelated is exactly zero, while the kinship coefficient of a non-inbred founder with themself is ½ rather than 1.

Q6. In the pedigree of Worked Example 6.2, explain what it would mean, biologically, if a child's genotype contained an allele that matched *neither* of the corresponding parent's two

alleles at that locus, and name the modern quality-control procedure (Chapter 19) built on exactly this logic.

Q7. Two individuals share one grandparent only (i.e., they are half-first-cousins, each descended from a different child of that one shared grandparent, with no other shared ancestors). Using the path-counting rule of Section 6.3, find their coefficient of relationship, r.

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

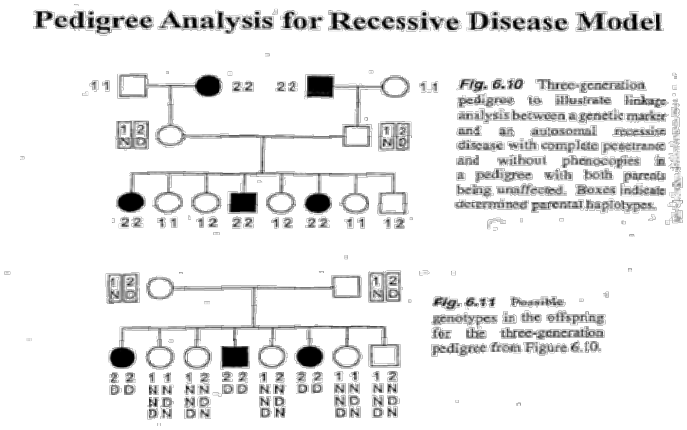
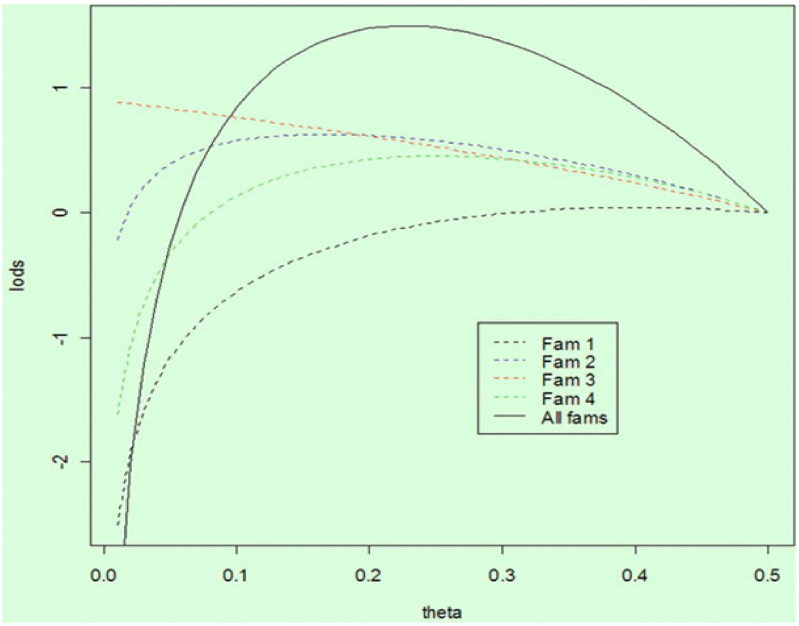


Fig. 6.10 Three-generation pedigree to illustrate linkage analysis between a genetic marker and an autosomal recessive disease with complete penetrance and without phenocopies in a pedigree with both parents being unaffected. Boxes indicate determined parental haplotypes.

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 theta, 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 ð 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:

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

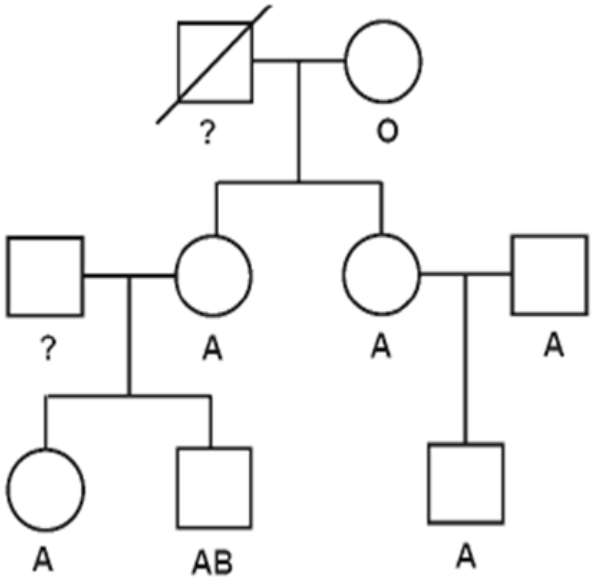
somewhat below the Mendelian expectation of 0.25 for a simple autosomal recessive condition, illustrating the kind of departure a segregation analysis (Section 7.4) would need to formally test before concluding whether OCA in this registry follows the simple single-gene recessive model or whether additional 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.

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

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 $\pi \to 0$) 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 inzip(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 (total affected)/(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 H0: 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.

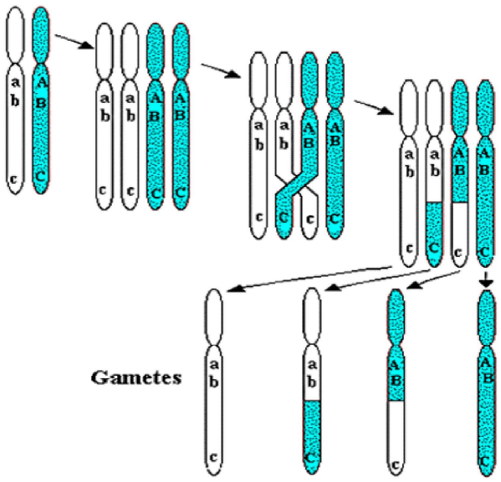
Q6. Explain why pooling data across several independently ascertained small families (as illustrated in the four-family plot of Section 7.2) generally sharpens inference about the segregation ratio compared with relying on any single family alone.

8 Estimation of the Recombination Fraction

8.1: Learning Objectives

By the end of this chapter, the reader should be able to define recombination and the recombination fraction, distinguish linked from unlinked loci, derive and apply the maximum likelihood estimator of the recombination fraction from double-backcross family data, and compute and interpret the LOD score for linkage.

8.2: Recombination and Crossing-Over



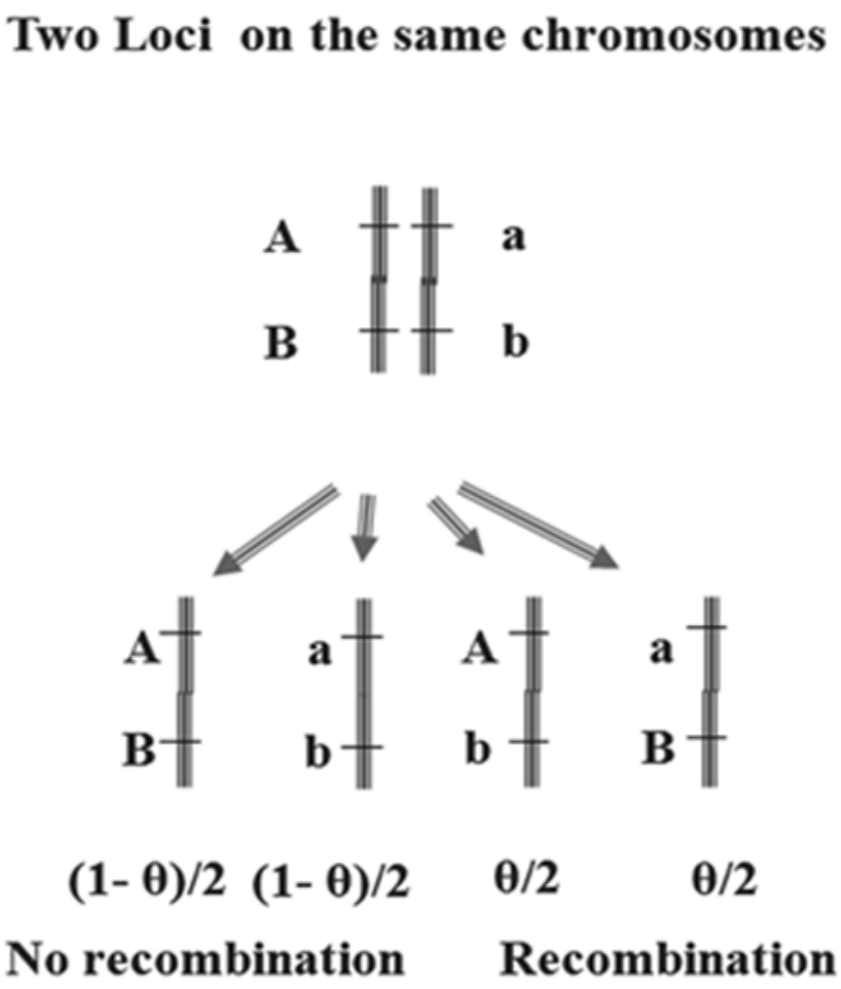
Crossing-over and recombination during meiosis: homologous chromosomes exchange segments, producing recombinant gametes.

During meiosis (Chapter 1), homologous chromosomes may exchange corresponding segments through a process called **crossing-over**, which physically occurs at points called **chiasmata**. When crossing-over occurs between two loci on the same chromosome, the alleles that were originally together on one parental chromosome ("in coupling") can end up separated into different gametes — this outcome is called a **recombinant** gamete; a gamete retaining the original parental combination is a **non-recombinant** (parental) gamete.

The **recombination fraction**, denoted θ (theta) or sometimes r, is defined as the probability that a gamete produced by a doubly heterozygous individual is recombinant at two given loci:

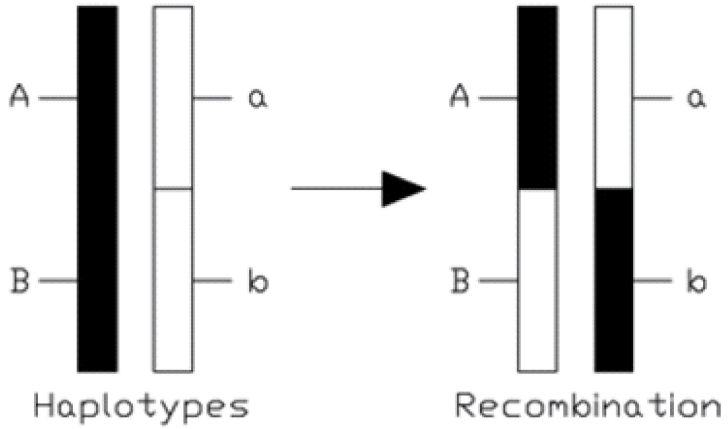
$$\theta = P(recombinant gamete)$$

For two loci on **different chromosomes**, or very far apart on the same chromosome, segregation is effectively independent (Mendel's Second Law, Chapter 2), and $\theta = 0.5$ (recombinant and non-recombinant gametes equally likely). For two loci **physically close together** on the same chromosome, crossing-over between them is rare, and $\theta < 0.5$; as physical distance between loci decreases toward zero, $\theta \to 0$. Two loci are said to be **linked** if $\theta < 0.5$.



Two loci, A/a and B/b, on the same pair of chromosomes: gametes formed without crossing-over between the loci are non-recombinant, each with probability $(1-\theta)/2$, while gametes formed with a crossover are recombinant, each with probability $\theta/2$.

8.3: The Double Backcross Design



Haplotypes before and after recombination between two linked loci, A/a and B/b.

The classical experimental (and, in humans, observational-family) design for estimating è is the **double backcross**: a parent who is doubly heterozygous at two loci (genotype AaBb, with the two "high" alleles A and B in coupling on one chromosome and a, b on the homologous chromosome) is "backcrossed" to (i.e., mates with) a doubly homozygous recessive individual (aabb). Because the aabb parent contributes only recessive alleles a and b, the offspring's phenotype directly reveals which gamete (AB, ab, Ab, or aB) was transmitted by the doubly heterozygous parent:

- **Non-recombinant (parental) gametes**: AB and ab, each with probability $(1-\theta)/2$.
- **Recombinant gametes**: Ab and aB, each with probability $\theta/2$.

8.4: Maximum Likelihood Estimation of θ

Let n_R denote the observed number of recombinant offspring and n_NR the number of non-recombinant offspring, out of a total of n = n_R + n_NR informative meioses. Since each offspring is recombinant with probability θ and non-recombinant with probability (1−θ), independently, the likelihood is binomial in form:

$$L(\theta) = \theta^{n_R}(1-\theta)^{n_{NR}}$$

Taking logs and differentiating with respect to θ, setting the derivative to zero, gives the maximum likelihood estimator:

$$\hat{\theta} = n_R/(n_R + n_{NR}) = n_R/n$$

— simply the observed proportion of recombinant offspring, the intuitive "gene-counting" estimator. Its large-sample variance (from the information in the binomial likelihood, Chapter 10) is:

$$Var(\hat{\theta}) \approx \hat{\theta}(1-\hat{\theta})/n$$

Theorem 8.1 (MLE of the Recombination Fraction). *For n independent, informative meioses, of which n_R are observed recombinant and n_NR = n − n_R non-recombinant, the maximum likelihood estimator of the recombination fraction θ is* $\hat{\theta} = n_R/n$*, and this estimator achieves the Cramér–Rao lower bound asymptotically, with large-sample variance* $\mathrm{Var}(\hat{\theta}) = \theta(1-\theta)/n$.

Proof. The log-likelihood is $\ell(\theta) = n_R \log\theta + n_{NR}\log(1-\theta)$. Differentiating with respect to θ gives the score function

$$U(\theta) = \frac{d\ell}{d\theta} = \frac{n_R}{\theta} - \frac{n_{NR}}{1-\theta}.$$

Setting $U(\hat{\theta}) = 0$ and solving: $\frac{n_R}{\hat{\theta}} = \frac{n_{NR}}{1-\hat{\theta}} \Rightarrow n_R(1-\hat{\theta}) =$

NR NR

$n_{NR}\hat{\theta} \Rightarrow n_R = \hat{\theta}(n_R + n_{NR}) = \hat{\theta}\, n \Rightarrow \hat{\theta} = n_R/n$, confirming the stated MLE (and the second derivative of ℓ is everywhere negative for $\theta \in (0,1)$, confirming this is indeed a maximum, not a minimum). The Fisher information is $I(\theta) = -E\left[\frac{d^2\ell}{d\theta^2}\right] = E\left[\frac{n_R}{\theta^2} + \frac{n_{NR}}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}$ (using $E[n_R] = n\theta$, $E[n_{NR}] = n(1-\theta)$), so by the general asymptotic theory of maximum likelihood estimation (Chapter 10, Theorem 10.1), $\mathrm{Var}\left(\hat{\theta}\right) \approx 1/I(\theta) = \theta(1-\theta)/n$. ■

Worked Example 8.2 (Recombination fraction estimation for a Nigerian consanguineous family study)

A Nigerian clinical genetics unit studies recombination between two closely linked markers flanking a rare autosomal disease locus in an extended family from a community with a tradition of consanguineous marriage (raising the informativeness of certain matings for linkage studies, Chapter 6). Genotyping 45 informative meioses across the extended pedigree reveals 6 recombinant and 39 non-recombinant gametes. By Theorem 8.1, $\hat{\theta} = 6/45 = 0.133$, with estimated standard error $\sqrt{\hat{\theta}\left(1-\hat{\theta}\right)/n} = \sqrt{0.133 \times 0.867/45} \approx 0.0507$, giving an approximate 95% confidence interval of $0.133 \pm 1.96(0.0507) = (0.034, 0.233)$ — a fairly wide interval reflecting the modest sample size typical of single extended-family linkage studies, illustrating why, historically, human linkage studies pooled data across as many informative families as possible (Section 8.5's LOD-score additivity) rather than relying on any single family's data alone.

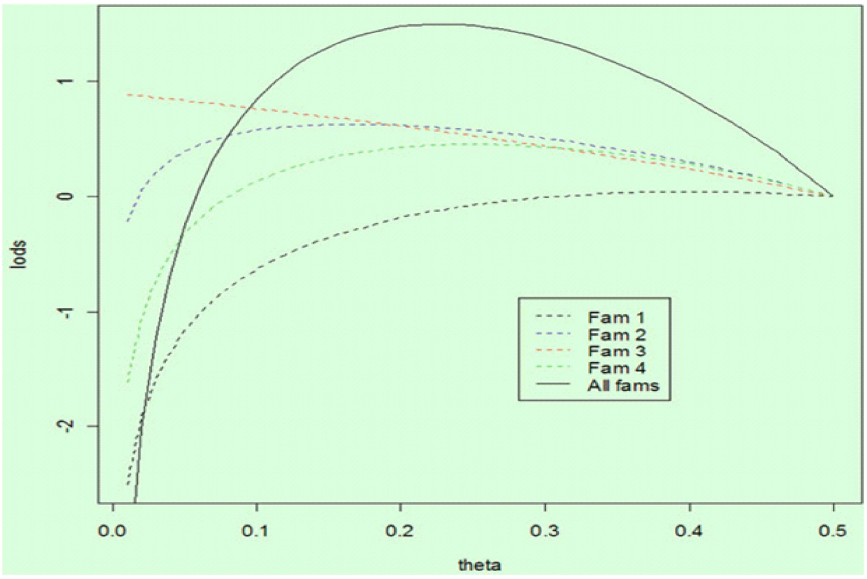
8.5: The LOD Score

Because we wish to formally test H_0: $\theta = 0.5$ (no linkage) against H_1: $\theta < 0.5$ (linkage), and because likelihoods for genetic linkage are conventionally reported on a base-10 logarithmic scale, the **LOD score** (logarithm of the odds) at a trial value of θ is defined as:

$$Z(\theta) = \log_{10}\left[\frac{L(\theta)}{L(0.5)}\right] = \log_{10}\left[\frac{\theta^{n_R}(1-\theta)^{n_{NR}}}{0.5^{\,n_R+n_{NR}}}\right]$$

The **maximum LOD score**, $Z(\hat{\theta})$, is obtained by substituting the MLE $\hat{\theta}$ from Section 8.4. By long-standing convention in human genetics (Morton, 1955), a maximum LOD score $\mathbf{Z \geq 3}$ (odds of 1000:1 or more in favour of linkage) is taken as significant evidence of linkage between the two loci; $\mathbf{Z \leq -2}$ is taken as significant evidence *against* linkage (i.e., evidence that the loci assort independently); values in between are inconclusive. (This threshold reflects the low prior probability of any two randomly chosen loci being linked, so a stringent evidential bar is required — a Bayesian, not merely frequentist, rationale, developed further in Chapter 17.)

Combining evidence across families: Because LOD scores are logarithms of likelihood ratios, and independent families contribute independent likelihood factors, LOD scores are simply **additive** across families at any fixed value of θ:



LOD score curves, as a function of the trial recombination fraction θ, computed separately for four independently ascertained families and for their combined ("All fams") analysis; combining families sharpens the peak and increases the maximum LOD score.

This additivity is one of the most practically important properties of the LOD-score framework: a single small family's data may give a maximum LOD score well below the significance threshold of 3, yet pooling several such families — simply by adding their individual LOD curves at each θ — can push the combined maximum LOD score decisively past the threshold, exactly as illustrated above and as revisited in the multi-family linkage analysis of Chapter 17.

8.6: Worked Example 8.1

A Nigerian family-linkage study genotypes two closely spaced markers flanking a disease locus in a set of double-backcross offspring, observing n_R = 3 recombinants and n_NR = 17 non-recombinants (n = 20).

(a) Find θ̂.

$$\hat{\theta} = 3/20 = 0.15$$

(b) Find the standard error of θ̂.

$$SE(\hat{\theta}) = \sqrt{0.15(0.85)/20} = \sqrt{0.1275/20} = \sqrt{0.006375} = 0.0798$$

(c) Compute the maximum LOD score, Z(θ̂).

$$Z(0.15) = log_{10}[(0.15)^3(0.85)^{17}/(0.5)^{20}]$$

$(0.15)^3 = 0.003375$. $(0.85)^{17} \approx 0.0631$ (since ln 0.85 = −0.1625, ×17 = −2.763, e^(−2.763) ≈ 0.0631). $(0.5)^{20} \approx 9.537\times10^{-7}$.

Numerator = $0.003375 \times 0.0631 \approx 2.130\times10^{-4}$. Ratio = $2.130\times10^{-4} / 9.537\times10^{-7} \approx 223.3$.

$Z(0.15) = \log_{10}(223.3) \approx 2.35$.

(d) Interpret. A LOD score of 2.35 falls short of the conventional $Z \geq 3$ threshold for significant linkage, though it is well above zero and represents suggestive evidence of linkage (odds of roughly 223:1 in favour of θ = 0.15 over independent assortment). In practice, the study would typically be extended (more families genotyped, or additional flanking markers added) before a definitive linkage claim could be made.

Theorem 8.2: (Consistency and Asymptotic Normality of the Recombination Fraction MLE)

Let theta-hat be the maximum likelihood estimator of the recombination fraction theta obtained from n independent backcross offspring, each classified as recombinant or non-recombinant (Section 8.4). Then theta-hat is a consistent estimator of the true theta, and sqrt(n)(theta-hat - theta) converges in distribution to a Normal(0, I(theta)^{-1}) random variable, where I(theta) is the per-observation Fisher information.

Proof. Each backcross offspring's recombinant/non-recombinant status is a Bernoulli(theta) trial, and the number of recombinants R among n offspring is Binomial(n, theta), so the log-likelihood is exactly the standard binomial log-likelihood in theta. This is a regular one-parameter exponential family model satisfying the standard regularity conditions (the parameter space (0, 0.5) is open, the likelihood is twice differentiable, and the Fisher information is finite and positive for theta in the interior), so the general asymptotic theory of maximum likelihood estimation established in Chapter 10 (Section 10.5) applies directly: theta-hat = R/n is consistent by the Law of Large Numbers applied to the sample proportion, and its asymptotic normality with variance I(theta)^{-1} = theta(1-theta)/n follows immediately from the general MLE asymptotic normality theorem, since the binomial Fisher information for theta is n/[theta(1-theta)]. This special case previews the general result proved formally in Chapter 10.

8.7: Chapter Summary

- Recombination fraction, θ, is the probability that a gamete from a doubly heterozygous parent is recombinant at two loci; θ = 0.5 for unlinked loci, θ < 0.5 for linked loci.
- Under the double backcross design , θ is estimated by simple proportion : $\hat{\theta}$ = n_R / n, the maximum likelihood estimator of a binomial parameter.

The LOD score, $Z(\theta) = \log_{10}[L(\theta)/L(0.5)]$, is the standard statistic for testing and summarising evidence of linkage; $Z \geq 3$ is conventionally significant, $Z \leq -2$ excludes linkage.

8.8: Computational Corner: MLE of the Recombination Fraction and the LOD Score

R.

```
# MLE of the recombination fraction, theta, from double-backcross data (Section 9.4)
n_parental <-68; n_recombinant <-32; n_total <- n_parental + n_recombinant

loglik_theta <-function(theta, r, n) r *log(theta) + (n - r) *log(1- theta)
theta_hat <- n_recombinant / n_total
theta_grid <-seq(0.01, 0.49, by =0.001)
ll <-sapply(theta_grid, loglik_theta, r = n_recombinant, n = n_total)
lod <- (ll -loglik_theta(0.5, n_recombinant, n_total)) /log(10)

plot(theta_grid, lod, type =”l”, xlab =expression(theta), ylab =”LOD score”)
abline(v = theta_hat, col =”red”, lty =2)
cat(“theta_hat =”, theta_hat, “ max LOD =”, max(lod), “\n”)
```

Python.

```
import numpy as np

n_parental, n_recombinant =68, 32
n_total = n_parental + n_recombinant

def loglik_theta(theta, r, n):
return r * np.log(theta) + (n - r) * np.log(1- theta)

theta_hat = n_recombinant / n_total
theta_grid = np.arange(0.01, 0.5, 0.001)
ll = loglik_theta(theta_grid, n_recombinant, n_total)
lod = (ll - loglik_theta(0.5, n_recombinant, n_total)) /
np.log(10)
print(theta_hat, lod.max())
```

Both scripts recover $\hat{\theta} = r/n$ exactly as derived analytically in Section 9.4, and plot the LOD -score curve of Section 9.5, showing that the maximum LOD occurs precisely at $\hat{\theta}$ — the computational equivalent of setting the score function to zero.

Mini-Project

Mini-Project 8: Estimating Recombination Fraction and Building a LOD Score Curve

Simulate a double backcross dataset and reproduce the full recombination-fraction estimation and LOD score workflow of this chapter.

Step 1: Simulate double backcross offspring data at a true recombination fraction of your choosing (e.g., theta = 0.15).

Step 2: Compute the maximum likelihood estimate of theta from your simulated data using Section 8.4's method.

Step 3: Compute the LOD score across a grid of candidate theta values from 0 to 0.5.

Step 4: Plot the LOD score curve and identify the theta value at which it peaks.

Step 5: Determine whether your peak LOD score would be considered significant evidence for linkage under the conventional threshold (Section 8.5).

Deliverable: *Simulation and estimation code, a LOD score curve plot, and a one-paragraph interpretation of significance.*

8.9: Practice Questions

Q1. A double backcross study of 50 informative offspring finds 8 recombinants. Compute $\hat{\theta}$ and its standard error.

Q2. Using the data in Q 1, compute the maximum LOD score , $Z(\hat{\theta})$, and state whether the evidence for linkage is conventionally significant.

Q3. Explain, intuitively, why $\theta = 0.5$ corresponds to "no linkage" while $\theta = 0$ corresponds to "complete linkage" (loci always co-transmitted).

Q4. A researcher reports a LOD score of -3 at $\theta = 0.5$ test point for two markers. What does this tell you about linkage between the two loci?

Q5. Two small families each give a maximum LOD score of 1.6 at $\theta = 0.15$. If their likelihood contributions are independent, what is their combined LOD score at $\theta = 0.15$, and does this combined value cross the conventional significance threshold?

Q6. Using the R or Python code of Section 8.8, verify numerically that $Z(\hat{\theta})$ computed from the pooled data of two families (obtained by summing their recombinant and non-recombinant counts) equals the sum of each family's individually computed LOD score at the same θ.

9 Inheritance of Quantitative Traits

9.1: Learning Objectives

By the end of this chapter, the reader should be able to distinguish qualitative from quantitative traits, partition phenotypic variance into genetic and environmental components, define and estimate broad- and narrow-sense heritability, and apply the polygenic model to Nigerian quantitative traits such as height, birth weight, and packed cell volume (PCV).

9.2: Qualitative versus Quantitative Traits

The traits considered in Chapters 1–8 (blood group, sickle-cell status, disease presence/absence) are **qualitative** (categorical or discrete): each individual falls into one of a small number of distinct classes, and single-locus Mendelian rules often describe their transmission well. Many traits of central importance in Nigerian public health and agriculture, however, are **quantitative** (continuous or nearly continuous): height, blood pressure, birth weight, packed cell volume (PCV/haematocrit), and crop yield, for example, vary smoothly over a range of values and do not sort into a handful of discrete classes.

Quantitative traits are typically influenced by many genes simultaneously (the **polygenic model**), each of small individual effect, together with environmental factors — a framework first formalised by R. A. Fisher (1918), reconciling Mendelian inheritance with the continuous variation and approximately

normal (Gaussian) distributions observed for such traits in populations, via the **Central Limit Theorem**: the sum of many independent, small-effect genetic and environmental contributions tends toward a normal distribution.

9.3: Partitioning Phenotypic Variance

The basic model of quantitative genetics decomposes an individual's phenotypic value, P, as:

$$P = G + E$$

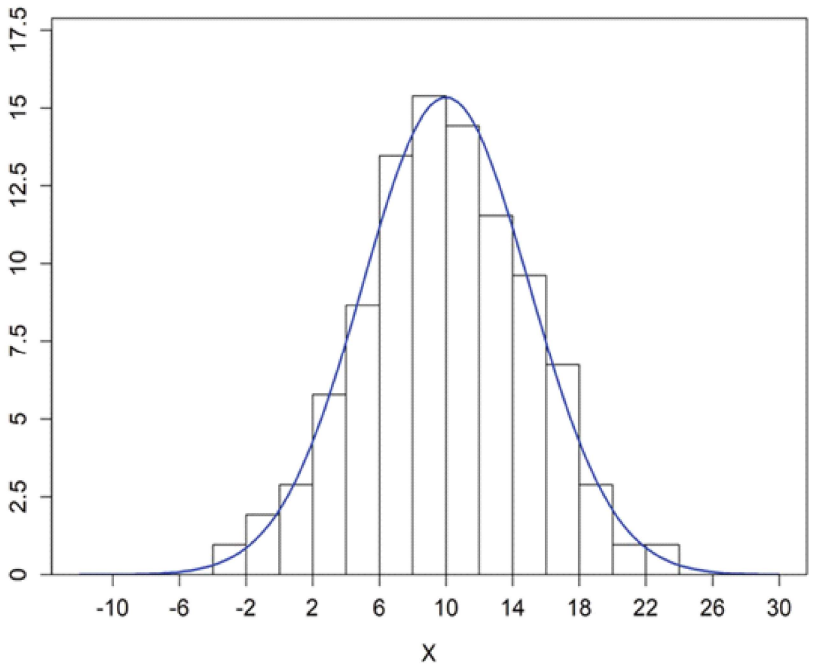
where G is the individual's genotypic value (the net effect of all genes influencing the trait) and E is the environmental deviation (all non-genetic influences, including measurement error). Assuming G and E are statistically independent (a reasonable first approximation, though gene–environment correlation and interaction are recognised complications), the phenotypic variance decomposes as:

$$V_P = V_G + V_E$$

The genotypic variance, V_G, is itself further partitioned:

$$V_G = V_A + V_D + V_I$$

where V_A is the **additive genetic variance** (variance due to the average, independently transmissible effects of individual alleles — the component that determines resemblance between parents and offspring, and the component selection acts on most efficiently), V_D is the **dominance variance** (variance due to interaction between the two alleles at the same locus), and V_I is the **epistatic (interaction) variance** (variance due to interaction between alleles at different loci).



The distribution of a quantitative trait (here, a simulated genotypic-value distribution, x) across a population, closely approximated by a normal curve — the empirical signature of a polygenic trait influenced by many genes of small effect, each contributing a small increment to G (Section 9.2).

9.4: Heritability

Broad-sense heritability, H^2, is the proportion of total phenotypic variance attributable to all genetic sources:

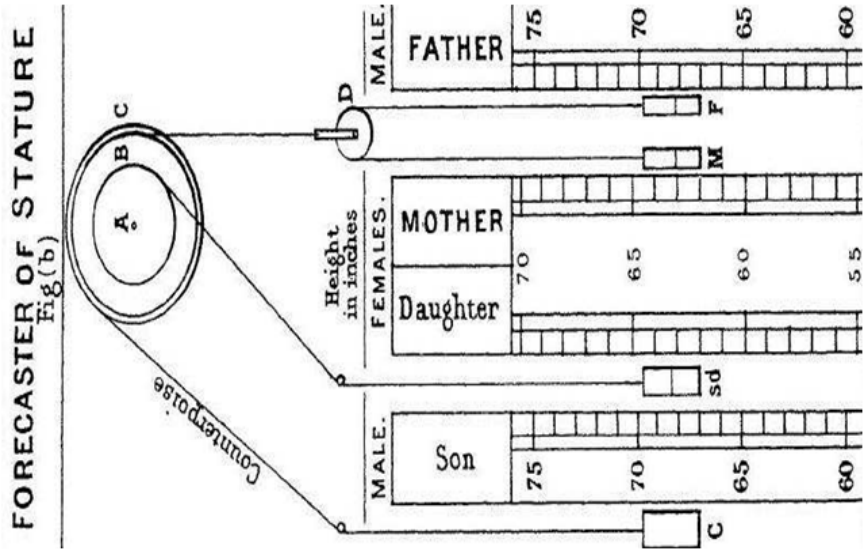
$$H^2 = V_G/V_P$$

Narrow-sense heritability, h^2, is the proportion of total phenotypic variance attributable to *additive* genetic variance alone:

$$h^2 = V_A/V_P$$

Narrow-sense heritability is the more important quantity for predicting resemblance between relatives and the response to selection, because only the additive component is faithfully transmitted, on average, from parent to offspring (dominance and epistatic effects depend on particular allele *combinations*, which are broken up by segregation and recombination at each generation, Chapters 2 and 7).

9.5: Estimating Heritability from Parent–Offspring Regression



A pedigree recording a quantitative trait (stature) across generations, the classical data structure for parent–offspring heritability estimation.

A classical method for estimating h^2 uses the regression of offspring phenotype on (mid-) parent phenotype. If $\bar{P}\_mid = (P\_mother + P\_father)/2$ is the mean of the two parents' phenotypic values, and $P\_offspring$ the offspring's value, then under the additive polygenic model, the regression slope of

offspring on mid-parent value directly estimates narrow-sense heritability:

$$b(O, mid - parent) = h^2$$

(If only a single-parent regression is used — offspring on one parent only — the slope estimates $h^2/2$, because the offspring shares, on average, only half the parent's genes.)

Theorem 9.1: (Falconer's Parent–Offspring Regression Theorem)

Under the additive polygenic model, with no dominance, no epistasis, no genotype–environment correlation, and random mating, the regression coefficient of offspring phenotype on mid-parent phenotype equals the narrow-sense heritability, h^2.

Proof. By definition, the regression slope of Y (offspring phenotype) on X (mid-parent phenotype) is $b = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$. Write each individual's phenotype as $P = A + E$ (additive genetic value plus environmental deviation, Section 9.3), so the mid-parent value is $\bar{P}_{mid} = \bar{A}_{mid} + \bar{E}_{mid}$ where $\bar{A}_{mid} = (A_{mother} + A_{father})/2$. Assuming environmental deviations are independent across family members and independent of genotype (no shared-environment or gene–environment correlation, part of the model's assumptions), $\mathrm{Cov}(\bar{P}_{mid}, P_{offspring}) = \mathrm{Cov}(\bar{A}_{mid}, A_{offspring})$, since all cross-terms involving E vanish. Under random mating and Mendelian transmission, the offspring's additive genetic value is $A_{offspring} = 1/2\, A_{mother} + 1/2\, A_{father} + M$, where M is the within-family Mendelian sampling deviation (independent of both

parents' genotypes, since it represents the randomness of which specific parental alleles are transmitted). Hence:

$$\begin{aligned}&\mathrm{Cov}\left(\bar{A}_{mid}, A_{offspring}\right)\\&= \mathrm{Cov}\left(1/2\,A_m + 1/2\,A_f,\ 1/2\,A_m + 1/2\,A_f\right.\\&\left. + M\right) = \mathrm{Var}\left(1/2\,A_m + 1/2\,A_f\right)\\&= 1/4\,V_A + 1/4\,V_A = 1/2\,V_A,\end{aligned}$$

using independence of A_m, A_f (random mating) and of M from both. Meanwhile $\mathrm{Var}(\bar{P}_{mid}) = \mathrm{Var}\left(1/2\,P_m + 1/2\,P_f\right) = 1/4\,V_P + 1/4\,V_P = 1/2\,V_P$ (again by independence of the two unrelated parents). Therefore:

$$b = \frac{\mathrm{Cov}\left(\bar{P}_{mid}, P_{offspring}\right)}{\mathrm{Var}(\bar{P}_{mid})} = \frac{1/2\,V_A}{1/2\,V_P} = \frac{V_A}{V_P} = h^2. \quad \square$$

Worked Example 9.3: (Parent–offspring height regression in a Nigerian university cohort)

A biometrics class at a Nigerian university measures adult height for 150 mother–father–offspring trios drawn from staff and students. The mean mid-parent height is 168.2 cm and mean offspring height is 169.5 cm; the sample covariance between mid-parent and offspring height is 18.4 cm², and the sample variance of mid-parent height is 26.3 cm². By Theorem 9.1, the estimated narrow-sense heritability is

$$\hat{h}^2 = \frac{18.4}{26.3} \approx 0.70,$$

consistent with the substantial (though not exclusive) genetic contribution to adult height documented across many human populations; the regression's intercept also allows a prediction of expected offspring height for a given mid-parent value, and — because $\hat{h}^2 < 1$ — Theorem 9.1's underlying model additionally predicts the classical phenomenon of

regression to the mean: offspring of unusually tall or unusually short mid-parent pairs are expected, on average, to lie closer to the population mean than their parents did, precisely because environmental deviations (the $1 - h^2$ share of variance) are not transmitted from parent to offspring.

Worked Example 9.1 (Height, Nigerian cohort)

A study of 200 father–mother–adult-child trios in a South-Western Nigerian community records mid-parent height and adult offspring height (in cm). A regression of offspring height on mid-parent height yields a slope b = 0.78, with standard error 0.06.

(a) Interpret b as an estimate of heritability: Since b = h^2 under the mid-parent regression model, $\hat{h}^2$ = 0.78, i.e., approximately 78% of the variance in adult height in this population is estimated to be attributable to additive genetic effects — consistent with the generally high heritability of stature (commonly estimated at 0.7–0.9) reported in human populations worldwide.

(b) Construct an approximate 95% confidence interval for h^2. $\hat{h}^2 \pm 1.96 \times SE = 0.78 \pm 1.96(0.06) = 0.78 \pm 0.118$, giving an interval of approximately (0.66, 0.90).

9.6: Estimating Heritability from Sibling Data

An alternative classical approach uses the **intraclass correlation** among full siblings, t, obtained from a one-way analysis of variance (ANOVA) of trait values with "family" as the grouping factor:

$$t = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$$

where σ^2_B is the between-family variance component and σ^2_W the within-family variance component. Since full siblings share, on average, one-half of their segregating genes

(coefficient of relationship r = 1/2, Chapter 6), and assuming no dominance variance, shared environment, or assortative mating, the intraclass correlation among full sibs estimates one-half of narrow-sense heritability:

$$t \approx \frac{1}{2}h^2 \Longrightarrow \hat{h}^2 = 2t$$

Worked Example 9.2: (Packed Cell Volume, PCV)

A one-way ANOVA of PCV (%) measured in 60 Nigerian sibling pairs (families) gives $\sigma^2_B = 9.6$ and $\sigma^2_W = 14.4$.

$$t = 9.6/(9.6 + 14.4) = 9.6/24.0 = 0.40 \Longrightarrow \hat{h}^2 = 2(0.40) = 0.80$$

The estimated narrow-sense heritability of PCV in this sibling sample is 0.80, indicating substantial genetic control of this widely measured Nigerian clinical trait, alongside the well-known environmental influences of iron status, malaria infection, and nutrition.

9.7: The Polygenic (Multifactorial) Threshold Model

Some traits that are themselves discrete (affected/unaffected) — for example, cleft lip/palate, some forms of congenital heart disease, and neural tube defects, all of documented occurrence in Nigerian clinical genetics — are believed to result from an underlying continuous, normally distributed **liability** determined by many genes and environmental factors jointly, with the disease phenotype appearing only when an individual's liability exceeds a threshold, T. Under this **multifactorial threshold model**, the population prevalence of the disease corresponds to the area under the standard normal curve beyond T, and the elevated recurrence risk in relatives of an affected individual (compared to population prevalence) can be used to estimate the heritability of liability — a method of particular relevance to Nigerian family-counselling

settings for congenital anomalies, where empirical recurrence risks (rather than simple Mendelian ratios) are typically what is quoted to parents.

Theorem 9.2: (Parent-Offspring Covariance Under an Additive Genetic Model)

Under a purely additive genetic model with no dominance or epistasis, and random mating, the covariance in phenotype between a parent and their offspring equals one-half the additive genetic variance: Cov(P_parent, P_offspring) = (1/2) V_A.

Proof. Write each individual's phenotype as P = A + E, where A is the additive genetic value (a sum of average allelic effects across loci) and E is an independent environmental deviation with E[E] = 0. An offspring's additive genetic value is the sum of one randomly transmitted allelic effect from each parent at each locus; by the additivity assumption, A_offspring = (1/2)A_parent1 + (1/2)A_parent2 + M, where M is a Mendelian sampling deviation with mean zero that is independent of both parents' genetic values (it reflects which of the two parental alleles was transmitted at each locus, an independent random event by Theorem 1.1). Then Cov(A_parent1, A_offspring) = Cov(A_parent1, (1/2)A_parent1 + (1/2)A_parent2 + M) = (1/2)Var(A_parent1) + (1/2)Cov(A_parent1,A_parent2) + Cov(A_parent1, M). Under random mating, Cov(A_parent1, A_parent2) = 0, and Cov(A_parent1, M) = 0 by the independence of Mendelian sampling from the parental genotype. Hence Cov(A_parent1, A_offspring) = (1/2)Var(A_parent1) = (1/2)V_A. Since the environmental deviations are independent of genotype and of each other across generations, Cov(P_parent, P_offspring) = Cov(A_parent, A_offspring) = (1/2)V_A, which is exactly the

theoretical basis for the parent-offspring regression method of Section 9.5.

9.8: Chapter Summary

- Quantitative traits are influenced by many genes of small effect (polygenic model) plus environment; phenotypic variance decomposes as V_P = V_A + V_D + V_I + V_E.
- Narrow-sense heritability, h^2 = V_A/V_P, governs resemblance between relatives and response to selection.
- h^2 can be estimated from the offspring-on-mid-parent regression slope, or from twice the sibling intraclass correlation.

9.9: Computational Corner: Estimating Heritability by Regression

R.

```
# Parent-offspring regression heritability estimate (Section 10.5), PCV example
set.seed(101)
n <-200
mid_parent_pcv <-rnorm(n, mean =38, sd =3)          # mean parental PCV (%)
h2_true <-0.45
offspring_pcv <-38+ h2_true * (mid_parent_pcv -38)
+rnorm(n, 0, 2.5)

fit <-lm(offspring_pcv ~ mid_parent_pcv)
slope <-coef(fit)[2]
h2_hat <- slope          # regression on MID-parent value estimates h2 directly
summary(fit)
```

```
cat("Estimated heritability (mid-parent regression):", h2_hat,
"\n")
```

Python.

```
import numpy as np
import statsmodels.api as sm

np.random.seed(101)
n =200
mid_parent_pcv = np.random.normal(38, 3, n)
h2_true =0.45
offspring_pcv =38+ h2_true * (mid_parent_pcv -38) +
np.random.normal(0, 2.5, n)

X = sm.add_constant(mid_parent_pcv)
model = sm.OLS(offspring_pcv, X).fit()
h2_hat = model.params[1]
print(model.summary())
print("Estimated heritability:", h2_hat)
```

The recovered slope should lie close to the simulated true heritability of 0.45, illustrating the parent–offspring regression estimator of Section 10.5 on a simulated packed-cell-volume (PCV) trait of the kind measured in Nigerian cohort studies; repeating the simulation with a different random seed shows the sampling variability of h^2 even when the underlying model is exactly correct.

Mini-Project

Mini-Project 9: Estimating Heritability Two Ways and Comparing Results

Simulate a quantitative trait with a known heritability and recover that heritability using both parent-offspring regression and sibling correlation.

Step 1: Simulate a quantitative trait for 200 parent-offspring pairs and 200 sibling pairs, generated from a model with a known true heritability (e.g., h^2 = 0.5).

Step 2: Estimate heritability via parent-offspring regression (Section 9.5).

Step 3: Estimate heritability via sibling correlation (Section 9.6).

Step 4: Compare both estimates to the true simulated value and to each other.

Step 5: Write a short paragraph discussing which estimate was closer to the truth in your simulation, and one reason the two methods can disagree in real data.

Deliverable: *Simulation code, both heritability estimates with their standard errors, and a short written comparison.*

9.10 Practice Questions

Q1. A regression of offspring birth weight on mid-parent birth weight (both measured in kg, in a Lagos maternity cohort) yields a slope of 0.35. Estimate h^2 for birth weight in this population, and comment on why it might be lower than the h^2 typically estimated for adult height.

Q2. A one-way ANOVA of adult height among 80 Nigerian sibships gives $\sigma^2_B = 45$ and $\sigma^2_W = 15$. Estimate the intraclass correlation t and the narrow-sense heritability h^2.

Q3. Explain why dominance variance (V_D) and epistatic variance (V_I) contribute to broad-sense heritability (H^2) but are largely "invisible" to standard parent–offspring regression estimates of narrow-sense heritability (h^2).

Q4. The population prevalence of a congenital heart defect is 0.5%, but among full siblings of an affected child, the observed prevalence rises to 3%. Explain qualitatively (no calculation required) how the multifactorial threshold model accounts for this elevated recurrence risk without invoking a single Mendelian gene.

Q5. Explain, in your own words and with reference to the Central Limit Theorem (Section 9.2), why a trait influenced by hundreds of genes of small individual effect tends to show an approximately normal (bell-shaped) distribution in the population, even though each individual gene's effect on its own need not be normally distributed.

Q6. A researcher partitions the phenotypic variance of birth weight in a Nigerian cohort and estimates $V_A = 0.35$, $V_D = 0.05$, $V_I = 0.02$, $V_E = 0.58$ (all in the same units of variance). Compute the broad-sense heritability ($H^2 = V_G/V_P$) and the narrow-sense heritability ($h^2 = V_A/V_P$), and explain why $h^2 \leq H^2$ always.

10 Likelihood Methods in Statistical Genetics

10.1: Learning Objectives

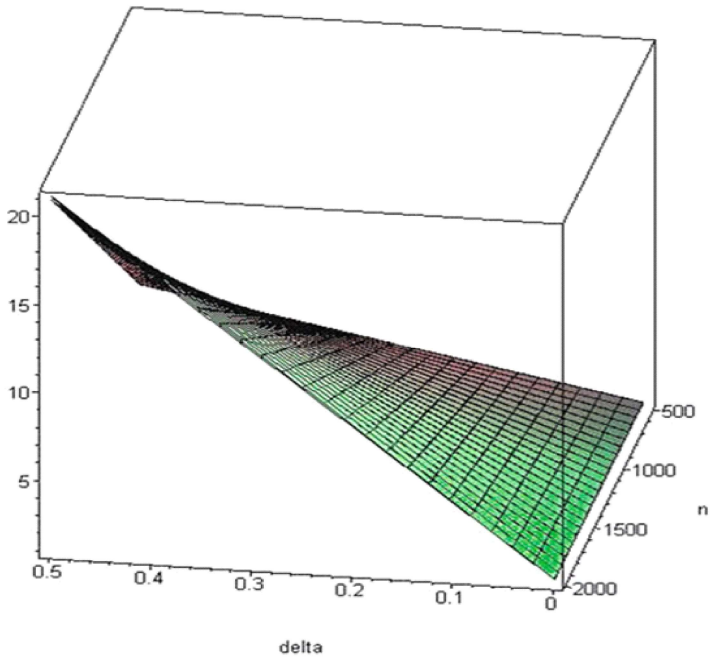
By the end of this chapter, the reader should be able to construct a likelihood function for genetic data, derive maximum likelihood estimators (MLEs) analytically and numerically (Newton–Raphson, Fisher scoring), state the asymptotic properties of MLEs, and apply the likelihood ratio test to genetic hypotheses.

10.2: The Likelihood Function

Statistical genetics relies heavily on **likelihood inference**, because so many of its central quantities — allele frequency, recombination fraction (Chapter 8), odds ratio, penetrance, heritability — are naturally cast as unknown parameters of a probability model fitted to observed genetic data. Given data $x = (x_1, \ldots, x_n)$ assumed to arise from a probability model $f(x; \psi)$ indexed by an unknown parameter (vector) ψ, the **likelihood function** is:

$$L(\psi) = f(x; \psi)$$

regarded as a function of ψ for the *fixed, observed* data x. The **maximum likelihood estimator (MLE)**, $\hat{\psi}$, is the value of ψ that maximises $L(\psi)$ — equivalently, that makes the observed data "most probable" under the model.



A two-parameter likelihood surface, L(ø , ø,), plotted as a function of two unknown parameters. The maximum likelihood estimate is the point (ø , ø,) at the peak of the surface; the curvature of the surface around this peak (Section 10.4) determines the precision of the estimates.

10.3: Log-Likelihood and the Score Function

Because likelihoods for independent observations are products, and products of many small probabilities are numerically awkward, it is standard to work with the **log-likelihood**, $l(\psi) = \log L(\psi)$. For independent observations, $l(\psi) = \Sigma_i \log f(x_i; \psi)$, a sum rather than a product.

The **score function**, $U(\psi) = dl(\psi)/d\psi$, is set to zero to locate the MLE (for a single parameter):

$$U(\hat{\psi}) = [dl(\psi)/d\psi] at \psi = \hat{\psi} = 0$$

Worked Example 10.1: (Allele frequency by maximum likelihood)

Consider a bi-allelic autosomal locus (alleles A, a) sampled from N unrelated individuals under Hardy–Weinberg equilibrium (Chapter 3), with observed genotype counts n_AA, n_Aa, n_aa. The likelihood, as a function of the allele frequency p = freq(A), is:

$$L(p) = (p^2)^{n_{AA}}\,(2pq)^{n_{Aa}}\,(q^2)^{n_{aa}}, q = 1 - p$$

$$l(p) = 2n_{AA}\log p + n_{Aa}\log(2pq) + 2n_{aa}\log q$$

$$U(p) = 2n_{AA}/p + n_{Aa}/p - n_{Aa}/q - 2n_{aa}/q$$

Setting U(p) = 0 and solving gives $\hat{p}$ = (2n_AA + n_Aa)/(2N), exactly the gene-counting estimator introduced in Chapter 3 — confirming that simple allele counting *is* the maximum likelihood estimator under HWE.

10.4: The Multiple-Parameter Case and the Information Matrix

When ψ is a vector of k parameters , the MLE solves the vector score equation $U(\hat{\psi}) = 0$, where $U(\psi)$ is the vector of partial derivatives $\partial l/\partial \psi_j$. The **observed information matrix** is:

$$I(\psi) = -\partial^2 l(\psi)/\,\partial\psi\,\partial\psi^T$$

and its inverse , evaluated at $\hat{\psi}$, approximates the variance–covariance matrix of the MLE for large samples:

$$Var(\hat{\psi}) \approx I(\hat{\psi})^{-1}$$

10.5: Asymptotic Properties of the MLE

Under standard regularity conditions, as the sample size n → ∞, the maximum likelihood estimator has three key large-sample properties, which justify its almost universal use in statistical genetics:

1. **Consistency**: $\hat{\psi}$ converges in probability to the true parameter value ψ_0.
2. **Asymptotic normality**: $\sqrt{n}(\hat{\psi} - \psi_0)$ converges in distribution to a normal distribution with mean 0 and variance equal to the inverse Fisher information , so that for large n , $\hat{\psi}$ is approximately Normal(ψ_0, $I(\psi_0)^{-1}$).
3. **Asymptotic efficiency**: among all consistent, asymptotically normal estimators, the MLE achieves the smallest possible asymptotic variance (the Cramér–Rao lower bound).

Theorem 10.1: (Asymptotic Normality of the MLE)

Under standard regularity conditions (the likelihood is smooth and twice differentiable, the true parameter lies in the interior of the parameter space, and the Fisher information is finite and positive), the maximum likelihood estimator $\hat{\psi}$ satisfies

$$\sqrt{n}\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} N(0,\ I_1(\psi_0)^{-1})$$

as $n \to \infty$, where $I_1(\psi_0)$ is the Fisher information from a single observation.

Proof sketch. Expand the score equation $U\left(\hat{\psi}\right) = 0$ in a first-order Taylor series around the true value ψ_0: $0 = U\left(\hat{\psi}\right) \approx U(\psi_0) + U'(\psi_0)\left(\hat{\psi} - \psi_0\right)$, so $\hat{\psi} - \psi_0 \approx -U(\psi_0)/U'(\psi_0)$. The score $U(\psi_0) = \sum_{i=1}^{n} U_i\,(\psi_0)$ is a sum of n independent,

identically distributed, mean-zero terms (mean zero because $E[U_i(\psi_0)] = 0$ is a standard likelihood identity), so by the Central Limit Theorem, $U(\psi_0)/\sqrt{n} \xrightarrow{d} N\big(0, I_1(\psi_0)\big)$. Similarly, $-U'(\psi_0)/n \to I_1(\psi_0)$ by the Law of Large Numbers (since $-U'(\psi)$ is the observed information, which converges to the expected information per observation). Combining these two limits via Slutsky's theorem:

$$\sqrt{n}(\hat{\psi} - \psi_0) \approx \frac{-U(\psi_0)/\sqrt{n}}{-U'(\psi_0)/n} \xrightarrow{d} \frac{N\big(0, I_1(\psi_0)\big)}{I_1(\psi_0)} = N\left(0, \frac{1}{I_1(\psi_0)}\right),$$

establishing the stated result (the multi-parameter case, Section 10.4, follows the identical argument with vectors and the information matrix in place of scalars).

Worked Example 10.3: (Applying asymptotic normality: sickle-cell allele frequency confidence interval, Osun State)

A newborn-screening programme in Osun State genotypes 1,200 infants at the HBB locus, finding 900 AA, 264 AS, and 36 SS. The MLE of the S allele frequency (Section 10.3) is $\hat{q} = (2 \times 36 + 264)/2400 = 336/2400 = 0.14$. By Theorem 10.1, $\hat{q}$ is approximately normally distributed with variance $\hat{q}(1 - \hat{q})/(2N) = 0.14 \times 0.86/2400 \approx 0.0000501$, giving a standard error of $\sqrt{0.0000501} \approx 0.00708$. An approximate 95% confidence interval for the true population allele frequency is therefore $0.14 \pm 1.96(0.00708) = (0.126, 0.154)$ — a range the programme can report to state health authorities as the plausible interval for the true carrier-relevant allele frequency in the screened population, directly illustrating how the abstract asymptotic theory of this section translates into a concrete, policy-relevant statement of statistical uncertainty.

10.6: Numerical Maximisation: Newton–Raphson and Fisher Scoring

For many genetic models (e.g., pedigree likelihoods, Chapter 6/17; logistic regression for association testing, Chapter 11), the score equation U(ψ) = 0 has no closed-form solution, and the MLE must be found numerically. The **Newton–Raphson algorithm** iterates:

$$\psi^{t+1} = \psi^t + I(\psi^t)^{-1}U(\psi^t)$$

using the *observed* information matrix at each step. **Fisher scoring** is a closely related variant that replaces the observed information with the *expected* information, E[I(ψ)], which is often more numerically stable (always positive semi-definite) and, for many standard exponential-family models used in genetics, is algebraically simpler to compute:

$$\psi^{t+1} = \psi^t + \mathcal{I}(\psi^t)^{-1}U(\psi^t), \mathcal{I}(\psi) = E[I(\psi)]$$

Both algorithms are iterated to convergence (e.g., until |ψ^(t+1) − ψ^(t)| is smaller than a chosen tolerance), and both, at convergence, additionally supply an estimate of Var (ψ̂) via the inverse (expected or observed) information — a convenient by-product used throughout applied statistical genetics.

10.7: The Likelihood Ratio Test

To test a null hypothesis H_0: ψ ∈ Ψ_0 (a restricted parameter space, e.g., θ = 0.5 for "no linkage," Chapter 8) against the general alternative H_1: ψ ∈ Ψ, the **likelihood ratio test (LRT)** statistic is:

$$\Lambda = -2log\left[\sup_{\psi\in\Psi_0} L(\psi)/\sup_{\psi\in\Psi} L(\psi)\right] = -2\left[l(\hat{\psi}_0) - l(\hat{\psi})\right]$$

where $\hat{\psi}_0$ is the MLE restricted to H_0 and $\hat{\psi}$ is the unrestricted MLE. Under H_0, and under standard regularity conditions, Λ is asymptotically distributed as chi-square with degrees of freedom equal to the number of independently restricted parameters (the difference in dimension between Ψ and Ψ_0).

Note on boundary cases: In some genetic applications (e.g., testing $\theta = 0.5$ against $\theta < 0.5$ in linkage analysis, or testing a variance component = 0), the null value lies on the *boundary* of the parameter space rather than in its interior. In such cases the standard chi-square asymptotic result does not directly apply, and the LRT statistic instead follows a mixture of chi-square distributions (e.g., a 50:50 mixture of a point mass at 0 and a χ^2_1, for a one-sided boundary test) — a subtlety the reader should bear in mind when interpreting LOD-score-based linkage tests (Chapter 8, Chapter 17) rigorously.

10.8: Worked Example 10.2 (LRT for Hardy–Weinberg equilibrium)

Using the sickle-cell data of Worked Example 3.1/3.2 (n_AA = 760, n_Aa = 220, n_aa = 20; N = 1000), test H_0: genotype frequencies follow HWE against H_1: genotype frequencies are unrestricted (i.e., three free genotype proportions summing to one, so H_1 has 2 free parameters versus 1 free parameter [p] under H_0 — a 1 degree-of-freedom test), using the LRT.

Under H_1 (saturated model), the MLEs of the genotype proportions are simply the observed proportions: $\hat{\pi}$_AA = 0.760, $\hat{\pi}$_Aa = 0.220, $\hat{\pi}$_aa = 0.020, so $l(\hat{\psi}) = 760 \log(0.760) + 220 \log(0.220) + 20 \log(0.020)$.

Under H_0 (HWE), $\hat{p} = 0.87$ (from Example 3.1), so $\hat{\pi}$_AA = $\hat{p}^2 = 0.7569$, $\hat{\pi}$_Aa = $2\hat{p}\hat{q} = 0.2262$, $\hat{\pi}$_aa = $\hat{q}^2 = 0.0169$, and $l(\hat{\psi}_0) = 760 \log(0.7569) + 220 \log(0.2262) + 20 \log(0.0169)$.

Computing: $l(\hat{\psi}) = 760(-0.2744) + 220(-1.5141) + 20(-3.9120) = -208.5 - 333.1 - 78.2 = -619.8$. $l(\hat{\psi}_0) = 760(-0.2789) + 220(-1.4863) + 20(-4.0803) = -212.0 - 327.0 - 81.6 = -620.6$.

$\Lambda = -2[l(\hat{\psi}_0) - l(\hat{\psi})] = -2[-620.6 - (-619.8)] = -2(-0.8) = 1.6$.

With 1 df, the critical value at $\alpha = 0.05$ is 3.84; since $\Lambda = 1.6 < 3.84$, we do not reject H_0 — consistent with the chi-square goodness-of-fit conclusion reached by a different (Pearson) route in Chapter 3, illustrating that the Pearson chi-square test and the likelihood ratio test are asymptotically equivalent large-sample procedures for the same hypothesis.

Theorem 10.2: (Invariance of the Maximum Likelihood Estimator Under Reparameterisation)

Let theta-hat be the maximum likelihood estimator of a parameter theta, and let g be any one-to-one function. Then the maximum likelihood estimator of g(theta) is g(theta-hat).

Proof. Let phi = g(theta), so theta = g^{-1}(phi) since g is one-to-one. The likelihood function expressed in terms of phi is L*(phi) = L(g^{-1}(phi)), which is simply the original likelihood function composed with the (invertible) reparameterisation. Because g is one-to-one, maximising L*(phi) over all valid phi is equivalent to maximising L(theta) over all valid theta = g^{-1}(phi): any value of theta that maximises L(theta) corresponds, under the bijection g, to exactly one value of phi = g(theta) that maximises L*(phi), and vice versa. Since theta-hat maximises L(theta) by definition, phi-hat = g(theta-hat) must maximise L*(phi), so the MLE of g(theta) is g(theta-hat). This invariance property is routinely used in statistical genetics, for example to obtain

the MLE of an odds ratio directly as the exponentiated MLE of a logistic regression coefficient (Section 11.7), without needing to re-derive the likelihood in the transformed parameterisation.

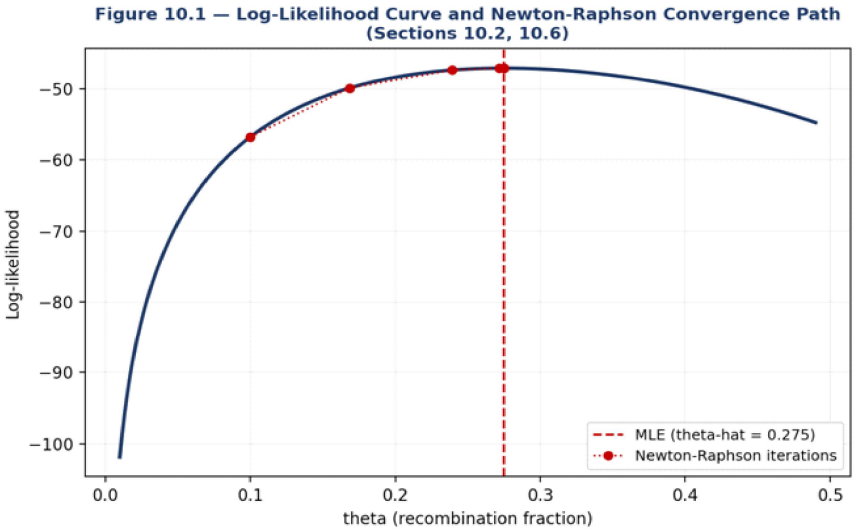


Figure 10.1: *Log-likelihood curve and Newton-Raphson convergence path (Sections 10.2, 10.6).*

10.9: Chapter Summary

- The likelihood function summarises the probability of the observed data as a function of unknown parameters; the MLE maximises it.
- Under HWE, the gene-counting estimator of allele frequency is exactly the MLE.
- MLEs are, asymptotically, consistent, normally distributed, and efficient; their variance is estimated from the inverse information matrix.
- Newton–Raphson and Fisher scoring provide iterative numerical routes to the MLE when no closed form exists.

- The likelihood ratio test , $\Lambda = -2[l(\hat{\psi}_0) - l(\hat{\psi})] \sim \chi^2$, is the standard tool for testing genetic hypotheses, including (with care near parameter-space boundaries) linkage.

10.10: Computational Corner: Newton-Raphson MLE and the Likelihood Ratio Test

R.

```
# General-purpose MLE via Newton-Raphson for a binomial proportion (allele frequency)
# and a likelihood ratio test of HWE (Section 11.7-10.8)
n_AA <-760; n_Aa <-220; n_aa <-20; N <- n_AA + n_Aa + n_aa

loglik_p <-function(p) (2*n_AA + n_Aa) *log(p) + (2*n_aa + n_Aa) *log(1- p)
score    <-function(p) (2*n_AA + n_Aa)/p - (2*n_aa + n_Aa)/(1- p)
info     <-function(p) (2*n_AA + n_Aa)/p^2+ (2*n_aa + n_Aa)/(1- p)^2

newton_raphson <-function(p0, tol =1e-8, maxit =100) {
  p <- p0
for (i in1:maxit) {
   p_new <- p +score(p) /info(p)
if (abs(p_new - p) < tol) break
   p <- p_new
}
  p_new
}
p_mle <-newton_raphson(0.5)
cat("Newton-Raphson MLE:", p_mle, "\n")
```

```
# LRT: unrestricted HWE genotype-frequency model vs the 1-parameter HWE model
ll_hwe <-loglik_p(p_mle)
p_AA <- n_AA/N; p_Aa <- n_Aa/N; p_aa <- n_aa/N
ll_sat <- n_AA*log(p_AA) + n_Aa*log(p_Aa) + n_aa*log(p_aa)
LRT <-2* (ll_sat - ll_hwe)
pchisq(LRT, df =1, lower.tail =FALSE)
```

Python.

```
import numpy as np
from scipy.stats import chi2

n_AA, n_Aa, n_aa =760, 220, 20
N = n_AA + n_Aa + n_aa

def score(p):
return (2*n_AA + n_Aa)/p - (2*n_aa + n_Aa)/(1- p)

def info(p):
return (2*n_AA + n_Aa)/p**2+ (2*n_aa + n_Aa)/(1- p)**2

def newton_raphson(p0, tol=1e-8, maxit=100):
   p = p0
for _ inrange(maxit):
      p_new = p + score(p) / info(p)
ifabs(p_new - p) < tol:
break
      p = p_new
return p_new

p_mle = newton_raphson(0.5)
```

```
def loglik_p(p):
return (2*n_AA + n_Aa) * np.log(p) + (2*n_aa + n_Aa) *
np.log(1- p)

p_AA, p_Aa, p_aa = n_AA/N, n_Aa/N, n_aa/N
ll_sat = n_AA*np.log(p_AA) + n_Aa*np.log(p_Aa) +
n_aa*np.log(p_aa)
ll_hwe = loglik_p(p_mle)
LRT =2* (ll_sat - ll_hwe)
p_value = chi2.sf(LRT, df=1)
print(p_mle, LRT, p_value)
```

The Newton–Raphson iteration converges to the same closed-form $\hat{p}$ derived by gene counting in Chapter 3, confirming that the score equation of Section 11.3 and the simple counting estimator agree exactly for this model; the likelihood ratio test then formally compares the fitted HWE model against a fully saturated (unrestricted) genotype-frequency model, giving a model-based alternative to the Pearson chi-square HWE test of Section 3.4.

Mini-Project

Mini-Project 10: Newton-Raphson MLE From Scratch and a Likelihood Ratio Test

Implement Newton-Raphson maximum likelihood estimation from first principles for a genetic parameter of your choice and validate it with a likelihood ratio test.

Step 1: Choose a one-parameter genetic model (e.g., a single allele frequency, or a recombination fraction) and write its log-likelihood function.

Step 2: Implement Newton-Raphson iteration (Section 10.6) to find the MLE, starting from at least two different initial values.

Step 3: Confirm both starting values converge to the same MLE.

Step 4: Construct a likelihood ratio test comparing your fitted model to a null value of scientific interest (e.g., theta = 0.5, or Hardy-Weinberg equilibrium).

Step 5: Report the LRT statistic, its degrees of freedom, and the resulting p-value.

Deliverable: Newton-Raphson implementation code, a convergence trace from at least two starting values, and a complete LRT report.

10.11: Practice Questions

Q1. Derive the maximum likelihood estimator of θ (recombination fraction) from the double-backcross binomial likelihood $L(\theta) = \theta^{\{n\_R\}(1-\theta)}\{n\_NR\}$ (Chapter 8), showing your differentiation of the log-likelihood.

Q2. A locus is genotyped in N = 400 Nigerian individuals under HWE: n_AA = 100, n_Aa = 200, n_aa = 100. Find $\hat{p}$ by maximum likelihood , and state , without further calculation, what you would expect the LRT for HWE to show (hint: compare the observed genotype proportions directly to p^2, $2pq$, q^2 at $\hat{p} = 0.5$).

Q3. State the three key asymptotic properties of the MLE, and briefly explain, in genetics terms, why "asymptotic efficiency" matters when sample sizes (numbers of families or individuals genotyped) are limited, as is common for rare Nigerian pedigree disorders.

Q4. Why does the standard χ^2 asymptotic distribution for the LRT not directly apply when testing $\theta = 0.5$ versus $\theta < 0.5$ in a linkage study, and what is used instead?

Q5. Explain, with reference to the likelihood surface of Section 10.2, why a *flat* (broad, gently sloped) likelihood surface near its peak indicates poorly estimated parameters, while a *sharply peaked* surface indicates precisely estimated parameters.

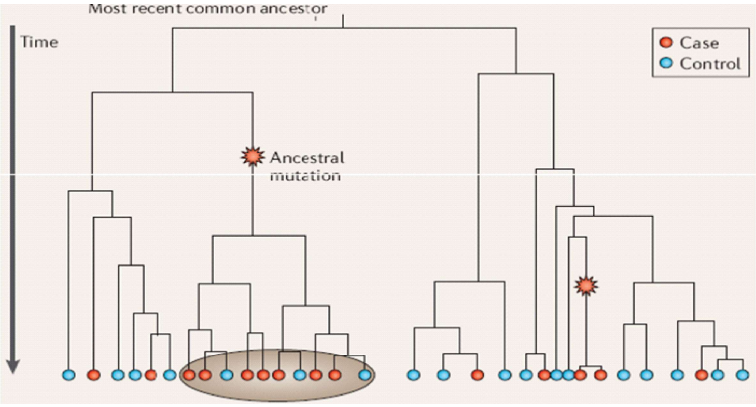
Q6. Using the Newton–Raphson code of Section 10.10, change the starting value p0 from 0.5 to 0.99 (a poor starting guess very close to the boundary) and report whether the algorithm still converges to the correct MLE, and in how many iterations (approximately) compared with starting from 0.5.

11 Population-Based Association Tests

11.1: Learning Objectives

By the end of this chapter, the reader should be able to construct genotype-based, allele-based, and Cochran–Armitage trend tests of association between a marker and a disease, compute and interpret odds ratios under different genetic models, and fit a logistic regression model for genetic association incorporating covariates, using a Nigerian case–control example.

11.2: The Case–Control Design in Genetic Association Studies



The genealogical rationale for association studies: cases (red) cluster on branches descending from an ancestral disease mutation more than controls (blue) (Balding, 2006).

A population-based association study compares the genotype (or allele) frequencies of a candidate marker between a sample of disease **cases** and a sample of unrelated, unaffected **controls**, drawn from the same source population. A statistically significant difference in marker frequency between cases and controls is taken as evidence that the marker is either itself functionally involved in disease risk, or is in linkage disequilibrium (Chapter 15) with a nearby causal variant.

Worked Example 11.1 (Setting)
Consider a hypothetical case–control study of a candidate SNP and type 2 diabetes mellitus (T2DM) in a Nigerian hospital-based sample: 300 T2DM cases and 300 unaffected controls are genotyped at a marker with alleles T (risk) and C, yielding the following genotype counts:

Genotype	Cases	Controls
TT	45	20
TC	140	120
CC	115	160
Total	**300**	**300**

11.3 The Genotype-Based Test

The **genotype-based test** is a standard 2×3 contingency-table chi-square test (or, equivalently, likelihood ratio test) comparing the full genotype distribution between cases and controls, with 2 degrees of freedom:

$$\chi^2 = \sum (O - E)^2 / E$$

summed over all 6 cells of the 2×3 table, where E is computed from the row and column marginal totals under independence (no association).

Worked Example 11.2

For the table above: total TT = 65, TC = 260, CC = 275; total cases = controls = 300; grand total = 600. E(TT, case) = (65×300)/600 = 32.5; E(TC, case) = (260×300)/600 = 130; E(CC, case) = (275×300)/600 = 137.5; and, by symmetry, the control expected counts equal the case expected counts here (since cases and controls are equal in size).

$\chi^2 = (45-32.5)^2/32.5 + (140-130)^2/130 + (115-137.5)^2/137.5 + (20-32.5)^2/32.5 + (120-130)^2/130 + (160-137.5)^2/137.5 = 4.808 + 0.769 + 3.682 + 4.808 + 0.769 + 3.682 = 18.52.$

With 2 df, the critical value at $\alpha=0.05$ is 5.99. Since $18.52 \gg 5.99$, there is strong evidence of association between this marker and T2DM in the sample.

11.4: Dominant, Recessive, and Allele-Based Tests

The genotype-based test (2 df) does not assume any particular genetic model, but is less powerful than a 1-df test when the true model is known or can reasonably be assumed. Common 1-df alternatives:

- **Dominant model**: collapse TT and TC into one group ("carriers" of at least one T allele) versus CC; test the resulting 2×2 table.
- **Recessive model**: collapse TC and CC into one group versus TT; test the resulting 2×2 table.
- **Allele-based test**: treat each of the 2N alleles (rather than N genotypes) as the unit of analysis, forming a 2×2 table of allele counts (T vs C) by case/control status, and testing with a standard 2×2 chi-square (1 df). This test implicitly assumes Hardy–Weinberg equilibrium in both cases and controls, since it treats the two alleles within an individual as independent observations.

Worked Example 11.3 (Allele-based test)
Allele counts: cases — T = 2(45)+140 = 230, C = 2(115)+140 = 370, total 600 alleles; controls — T = 2(20)+120 = 160, C = 2(160)+120 = 440, total 600 alleles.

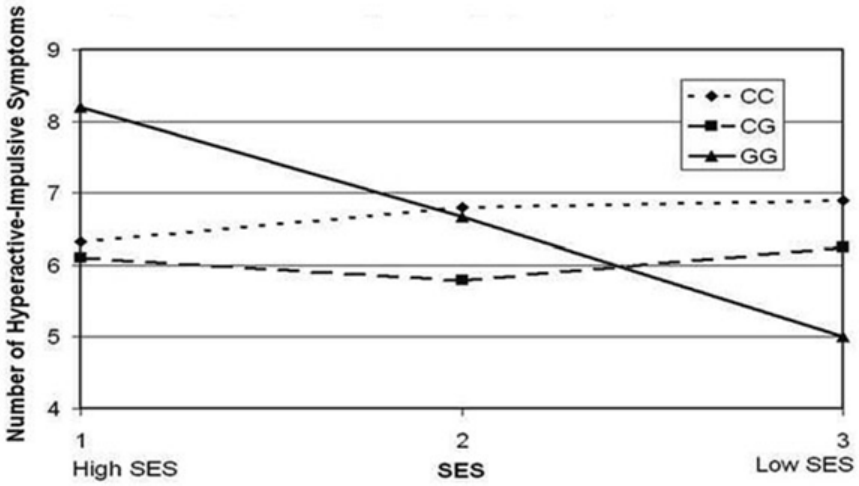
	T	C	Total
Cases	230	370	600
Controls	160	440	600
Total	390	810	1200

E(T, case) = (390×600)/1200 = 195; E(C, case) = 405; by symmetry E(T, control)=195, E(C, control)=405. χ^2 = $(230-195)^2/195 + (370-405)^2/405 + (160-195)^2/195 + (440-405)^2/405$ = 6.28+3.02+6.28+3.02 = 18.62 (1 df). Critical value (1 df, α=0.05) = 3.84; 18.62 $\gg$ 3.84 — strong evidence of allelic association.

11.5: The Cochran–Armitage Trend Test

The **Cochran–Armitage trend test** is a 1-df test that assumes an *additive* (dosage) model: risk increases (or decreases) linearly with the number of copies of the risk allele (0, 1, or 2 copies of T). It is generally the most powerful test when the additive model is approximately correct — a common situation for complex disease SNPs — and does not require the Hardy–Weinberg assumption implicit in the allele-based test. The test statistic is constructed as a weighted regression of disease status on genotype score (coded 0, 1, 2), and is asymptotically χ^2_1 under the null of no linear trend.

Choosing the right model matters for power
Because the four tests above (genotype-based, dominant, recessive, trend/additive) each make different implicit assumptions about how risk varies with genotype, the power of each test depends heavily on which model actually holds in the data:



Statistical power to detect association under three assumed genetic models (CC = codominant/genotype-based, CG = an intermediate coding, GG = a specific dominant/recessive-style grouping) as a function of socioeconomic status subgroup, illustrating how test choice interacts with the true underlying mode of inheritance.

As a practical rule of thumb: the Cochran–Armitage trend test is a reasonable default when the true genetic model is unknown (it retains good power under a dominant, recessive, or truly additive model, at some cost relative to whichever 1-df test exactly matches the truth), while the 2-df genotype-based test is the safest choice when there is a real possibility of a non-monotonic (e.g., over- or under-dominant) relationship between genotype and risk, at the cost of reduced power when a simpler 1-df model does in fact hold — a trade-off the analyst must make explicit before, not after, seeing the data (to avoid the multiple-testing inflation of trying every model and reporting only the most significant one).

Theorem 11.1: (Asymptotic Equivalence of the Trend Test and the Additive-Model Score Test)

The Cochran–Armitage trend test statistic, computed by treating genotype as a linear dosage score $x \in \{0,1,2\}$, is asymptotically equivalent to the score test (Chapter 10) for the coefficient β in the additive logistic regression model $\text{logit}\, P(\text{disease}) = \alpha + \beta x$*, under the null hypothesis $\beta = 0$.*

Proof sketch. The score function for β in the logistic model, evaluated at the null β = 0 (where all fitted probabilities reduce to the common intercept-only estimate $\hat{\pi}$, the overall disease prevalence), is $U(0) = \sum_{i=1}^{N} x_i\,(y_i - \hat{\pi})$, where y_i is the disease indicator and $x_i \in \{0,1,2\}$ the genotype score for individual i. This is exactly proportional to the covariance between genotype score and disease status across the sample — precisely the numerator of the Cochran–Armitage trend statistic (Section 11.5), which is constructed as a weighted linear-trend comparison of case and control genotype-score means. The corresponding Fisher information under the null, $I(0) = \hat{\pi}(1-\hat{\pi})\sum_i x_i^2 - (\text{correction term})$, matches the trend test's variance denominator, so the score-test statistic $U(0)^2/I(0)$ and the Cochran–Armitage χ^2 statistic coincide asymptotically as sample size grows, both converging to the same χ^2_1 null distribution. ■

This equivalence is practically useful: it means the Cochran–Armitage trend test, developed originally as a purely non-parametric large-sample test, can equivalently be understood (and, where convenient, computed) as a special case of the logistic regression framework of Section 11.7, inheriting all the machinery (covariate adjustment, odds ratio interpretation) of that more general approach.

Worked Example 11.5 (Trend test for a malaria-protective variant in a Nigerian paediatric cohort)

A paediatric malaria study in a Nigerian tertiary hospital genotypes a candidate SNP thought to influence severe malaria risk. Among 180 severe-malaria cases: 92 CC, 70 CT, 18 TT; among 220 community controls: 154 CC, 58 CT, 8 TT. Coding genotype as x = 0 (CC), 1 (CT), 2 (TT), the Cochran–Armitage trend statistic tests whether disease risk increases monotonically with copies of the T allele. By Theorem 11.1, this is equivalent to fitting the additive logistic model $\text{logit}\,P(\text{severe malaria}) = \alpha + \beta x$ and testing $H_0: \beta = 0$ via a score test — either computation should yield closely matching test statistics and p-values, and the fitted $\hat{\beta}$ from the logistic model additionally yields an immediately interpretable per-allele odds ratio, $e^{\hat{\beta}}$, for the T allele's association with severe malaria risk in this cohort.

11.6: Measuring the Strength of Association: The Odds Ratio

Whereas the chi-square/trend tests address statistical *significance*, the **odds ratio (OR)** measures the *magnitude* of association, and is the standard effect-size measure in case–control genetics because, unlike relative risk, it can be validly estimated from case–control (rather than only cohort) data.

For the dominant model (TT+TC vs CC): OR = [(n_{TT+TC,,case})(n_{CC,,control})] / [(n_{TT+TC,,control})(n_{CC,,case})].

Interpretation: Carriers of at least one T (risk) allele have approximately 1.84 times the odds of T2DM compared to non-carriers (CC) in this Nigerian hospital sample — a moderate effect size, of the general order of magnitude reported for many common-variant associations with complex diseases such as

T2DM in genome-wide association studies (Chapter 18) worldwide, including in African-ancestry cohorts.

11.7: Logistic Regression for Genetic Association

Logistic regression generalises the simple 2×2/2×3 tests by modelling disease log-odds as a linear function of genotype (coded, e.g., additively as 0/1/2 copies of the risk allele) **and** other covariates — age, sex, body mass index, and, importantly in African genetic association studies, principal components summarising ancestry/population stratification (to guard against spurious associations arising from the Wahlund-type structure discussed in Chapter 3):

$$\begin{aligned} &logitP(disease = 1) \\ &= \beta_0 + \beta_1(\text{genotype score}) + \beta_2(age) \\ &+ \beta_3(BMI) + \cdots \end{aligned}$$

The genotype coefficient , β_1, is tested (typically via a Wald test or likelihood ratio test — Chapter 10) for departure from zero, and $\exp(\beta_1)$ estimates the (covariate-adjusted) odds ratio per copy of the risk allele — allowing, for instance, a Nigerian T2DM association study to separate the genetic effect of the marker from the well-documented, strongly confounding effects of age and adiposity.

Theorem 11.2: (The Cochran-Armitage Trend Test as an Efficient Score Test)

Under a genetic model in which the log-odds of disease increases linearly with the number of copies of the risk allele (an additive, or 'allele-dosage', model), the Cochran-Armitage trend test statistic (Section 11.5) is asymptotically equivalent to the efficient score test for the corresponding logistic regression model with genotype coded as 0, 1, 2 copies.

Proof. Consider a logistic regression model logit[P(disease)] = beta_0 + beta_1 x, where x in {0,1,2} is the number of risk alleles carried. The score function for beta_1, evaluated under the null hypothesis beta_1 = 0, is U(beta_1=0) = sum_i (y_i - pbar)x_i, where y_i is the case/control indicator and pbar is the overall case proportion — this is exactly proportional to the covariance between genotype dosage and case status computed across the sample. The Cochran-Armitage trend test statistic is constructed as a weighted sum of (observed - expected) counts across genotype categories using the same linear genotype scores 0, 1, 2, which, after algebraic simplification, reduces to the same covariance-based quantity as the logistic regression score U(0), up to a normalising constant determined by the null variance. Because the score test statistic is asymptotically the most powerful (locally, near the null) test for detecting a non-zero beta_1 in a correctly specified model (a general property of score tests in regular parametric families, per Chapter 10), and the Cochran-Armitage statistic coincides with this score statistic under the additive coding, the Cochran-Armitage test is asymptotically optimal specifically for detecting departures consistent with a per-allele additive genetic effect.

11.8: Chapter Summary

- Case–control association studies compare marker frequencies between affected and unaffected samples.
- The genotype-based test (2 df) is model-free; dominant, recessive, allele-based, and Cochran–Armitage trend tests (each 1 df) assume specific genetic models and are more powerful when correctly specified.
- The odds ratio quantifies effect size; logistic regression extends the analysis to adjust for covariates, including ancestry, which is essential in genetically diverse Nigerian populations.

11.9: Computational Corner: Genotype, Allele, and Regression-Based Association Tests

R.

```
# Population-based genotype and allele-based association tests (Section 15.3-11.5)
cases   <-c(AA =30, Aa =45, aa =25)
controls <-c(AA =50, Aa =40, aa =10)

tab <-rbind(cases, controls)
chisq.test(tab)                 # 2 df genotype-based test

# Allele-based 2x2 test
allele_cases    <-c(A =2*30+45, a =2*25+45)
allele_controls <-c(A =2*50+40, a =2*10+40)
allele_tab <-rbind(allele_cases, allele_controls)
chisq.test(allele_tab, correct =FALSE)

# Logistic regression (additive coding: 0,1,2 copies of risk allele)
geno_score <-c(rep(0, 50), rep(1, 40), rep(2, 10),   # controls: aa,Aa,AA
rep(0, 25), rep(1, 45), rep(2, 30))  # cases: aa, Aa, AA
disease <-c(rep(0, 100), rep(1, 100))
fit <-glm(disease ~ geno_score, family = binomial)
summary(fit)
exp(coef(fit)["geno_score"])         # odds ratio per copy of risk allele
```

Python.

```
import numpy as np
from scipy.stats import chi2_contingency
import statsmodels.api as sm
```

```
tab = np.array([[30, 45, 25], [50, 40, 10]])   # cases, controls x AA, Aa, aa
chi2_stat, p_value, dof, expected = chi2_contingency(tab)
print(chi2_stat, p_value, dof)

geno_score = np.array([0]*50+ [1]*40+ [2]*10+ [0]*25+ [1]*45+ [2]*30)
disease = np.array([0]*100+ [1]*100)
X = sm.add_constant(geno_score)
model = sm.Logit(disease, X).fit()
print(model.summary())
print("Odds ratio per allele copy:",
np.exp(model.params[1]))
```

The genotype-based chi-square test (2 df) and the allele-based test (1 df) both reject the null of no association for this simulated dataset, and the logistic-regression odds ratio quantifies the increase in disease odds per copy of the risk allele — the standard reporting metric in the association literature covered further in Chapters 12, 13 and 18.

Mini-Project

Mini-Project 11: Comparing Four Association Tests on the Same Data

Apply the genotype-based, allele-based, Cochran-Armitage trend, and logistic regression tests to a single simulated case-control dataset and compare their results.

Step 1: Simulate a case-control dataset of 500 cases and 500 controls at a marker with a modest true genetic effect (e.g., odds ratio 1.4 per allele).

Step 2: Apply the genotype-based chi-square test (Section 11.3).

Step 3: Apply the allele-based test and the Cochran-Armitage trend test (Sections 11.4-11.5).

Step 4. Fit a logistic regression model (Section 11.7) and extract the odds ratio and its confidence interval.

Step 5. Summarise all four results in a single table and discuss any differences in statistical significance across tests.

Deliverable: *Simulation and analysis code, a comparison table of all four tests' statistics and p-values, and a short written discussion.*

11.10: Practice Questions

Q1. For the genotype counts in Worked Example 11.1, carry out the recessive-model test (TT vs TC+CC) as a 2×2 chi-square, and compute the corresponding odds ratio.

Q2. Explain, in a sentence, why the allele-based test assumes Hardy–Weinberg equilibrium while the genotype-based test does not.

Q3. In a Nigerian GWAS-style analysis, why is it standard practice to include principal components of genome-wide genotype data as covariates in the logistic regression model, rather than testing the raw genotype–disease association alone?

Q4. An odds ratio of 1.0 is obtained for a marker in a well-powered case–control study, with a non-significant p-value. What does this tell you about the marker's role in disease risk in this population?

Q5. Explain why choosing the genetic model (dominant, recessive, additive) that gives the smallest p-value *after* looking at the data, and reporting only that result, is a form of multiple-testing malpractice, even though only a single "final" p-value is reported.

Q6. A SNP shows a strong recessive effect in the true underlying biology (only homozygous risk-allele carriers have elevated disease risk). Which of the four tests described in Sections 11.3–11.5 would you expect to have the *least* power to detect this association, and why?

12 Candidate Gene Association Studies

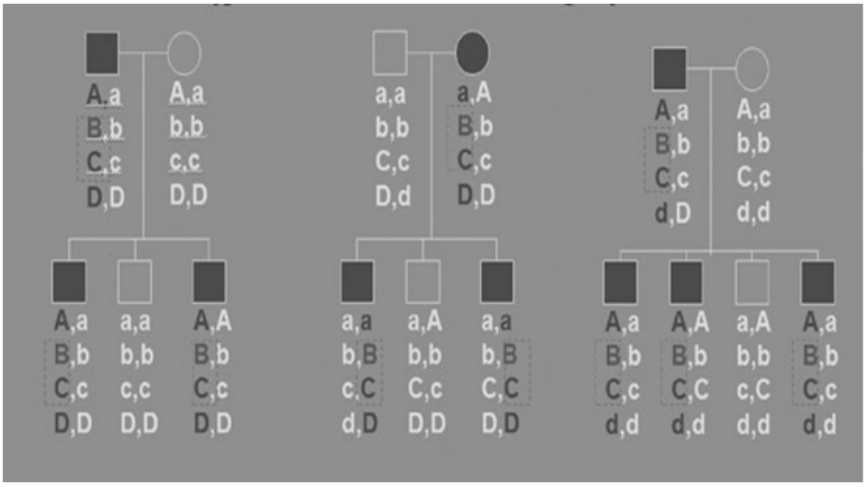
12.1: Learning Objectives

By the end of this chapter, the reader should be able to explain the rationale and design of a candidate-gene association study, distinguish it from genome-wide approaches (Chapter 18), describe how candidate genes and SNPs are selected, discuss the advantages and well-documented pitfalls of the candidate-gene approach (including the replication crisis of the 2000s), and carry out a candidate-gene regression analysis with an appropriate multiple-testing correction in R and Python.

12.2: What Is a Candidate-Gene Study?

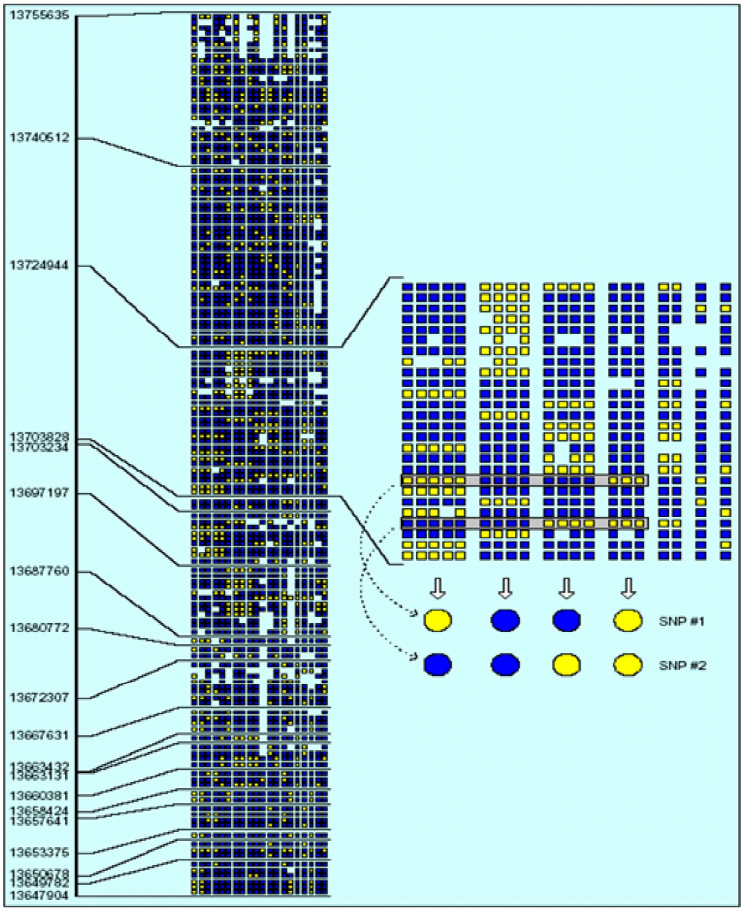
A **candidate-gene association study** tests one or a small number of genes — selected in advance because of known or suspected biological relevance to the trait or disease under study — for statistical association with that trait, using the same basic statistical machinery (chi-square tests, logistic regression) developed in Chapter 11. Candidates are typically chosen from one or more of the following sources: (i) genes encoding proteins in a biologically plausible pathway (e.g., genes of the renin–angiotensin system for hypertension); (ii) genes previously implicated by linkage analysis (Chapter 17) as lying within a chromosomal region linked to the trait; (iii) genes highlighted by animal-model or functional studies; or (iv) genes already reported in the published literature, often from a different population, motivating a **replication study**.

This targeted design contrasts with the **genome-wide association study (GWAS)**, which tests hundreds of thousands to millions of markers spread across the entire genome without any prior assumption about which genes are relevant (Chapter 18). Historically, candidate-gene studies dominated human genetics from the early 1990s until the mid-2000s, when the falling cost of genome-wide genotyping arrays made GWAS the default design for common, complex traits.



A chromosome ideogram marking the physical locations of several candidate genes and their surrounding marker loci — the kind of genome-wide "parts list" from which candidate genes are selected for targeted study.

12.3: Selecting Candidate SNPs Within a Gene



A candidate gene's linkage-disequilibrium structure and physical map, used to select a minimal set of tag SNPs for genotyping.

Having settled on a candidate gene, the investigator must decide how much of it — and which specific variants — to genotype. Genotyping every possible SNP within and around a gene is neither necessary nor economical, because SNPs in strong linkage

disequilibrium (Chapter 15) carry largely redundant information. Practical strategies include:

- **Tagging SNPs**: choosing a minimal subset of SNPs such that every other common SNP in the region has r^2 above some threshold (often a higher threshold, e.g. $r^2 > 0.9$, than is typical in GWAS chip design, since fewer SNPs are being genotyped overall and denser coverage of a single small region is affordable) with at least one tag SNP — see Section 12.3 of the LD chapter's tagging discussion (Chapter 15).
- **Functional SNPs**: variants with a plausible biological mechanism — non-synonymous coding changes, splice-site variants, or variants in known promoter or enhancer elements.
- **Previously reported SNPs**: variants already associated with the trait (or a related trait) in the literature, included specifically to attempt replication.
- **Minor allele frequency (MAF) thresholds**: very rare variants are usually excluded from a standard candidate-gene study (which typically has limited power for anything below MAF H" 0.05) and are instead the target of the rare-variant and sequencing-based methods of Chapter 20.

The region genotyped is usually extended beyond the gene's coding sequence to include the promoter, untranslated regions, and a margin (often 5–20 kb) upstream and downstream, since regulatory variants affecting expression may lie outside the transcribed sequence itself.

12.4: Advantages and Disadvantages of the Candidate-Gene Design

Advantages: A candidate-gene study genotypes far fewer markers than a GWAS, so it is considerably cheaper, and the multiple-testing burden (Section 12.5) is much smaller, giving greater statistical power per genotyped SNP for a fixed sample size — an

important practical consideration for studies of Nigerian and other African cohorts, where sample sizes are frequently smaller than those available in well-funded biobanks elsewhere.

Disadvantages: The approach is limited in scope by construction: a true causal gene that was not on the candidate list can never be discovered. Historically, this design also proved highly susceptible to publication of **false-positive findings** that failed to replicate — a well-documented crisis in the candidate-gene literature of the 1990s and 2000s, in fields ranging from psychiatric genetics to cardiovascular risk, traced to a combination of small sample sizes, inadequate correction for population stratification (Chapter 14), publication bias favouring positive results, and biologically plausible but ultimately spurious a priori hypotheses. For this reason, contemporary reviewers and journals require independent replication, ideally in more than one population, before a candidate-gene finding is accepted, and increasingly regard genome-wide-significant findings from unbiased GWAS or sequencing studies (Chapters 18 and 20) as more credible evidence than an isolated candidate-gene report.

Worked Example 12.1: (APOL1 and kidney disease risk in African populations)

A well-established and now widely replicated example of a genuine candidate-gene finding relevant to African populations is the association between coding variants in the *APOL1* gene (the G1 and G2 risk alleles, which arose under selective pressure from resistance to trypanosomiasis) and non-diabetic chronic kidney disease, an association discovered through a combination of linkage mapping in African-American families followed by targeted candidate-gene fine-mapping of the linked region. Because *APOL1* risk alleles are common in West African populations (including Nigeria) but virtually absent in European populations, this is a clear illustration of why candidate-gene (and GWAS) findings from one population cannot automatically be assumed

to generalise to another — the practical motivation for Chapter 14's discussion of population substructure, and a caution against uncritically importing "risk alleles" identified elsewhere into Nigerian clinical practice without local validation.

12.5: The Multiple-Testing Problem in a Candidate-Gene Study

Even though a candidate-gene study tests far fewer markers than a GWAS, testing multiple SNPs within the same gene (and, often, multiple genes) still inflates the **family-wise type I error rate** if uncorrected. The simplest correction is the **Bonferroni correction**: for m independent tests at nominal significance level á, declare significance only if the per-test p-value is below á/m.

Theorem 12.1: (Bonferroni Inequality)

For any m hypothesis tests, each conducted at per-test significance level α/m, the probability of at least one false positive among the m tests (the family-wise error rate, FWER), under the global null hypothesis that none of the m associations is real, satisfies FWER ≤ α — regardless of any dependence structure among the tests.

Proof. Let E_i denote the event that test i produces a false positive (rejects its null when that null is true), for $i = 1, \dots, m$, so $P(E_i) = \alpha/m$ for each i under the Bonferroni-corrected threshold. By Boole's inequality (the union bound), for any collection of events,

$$P\left(\bigcup_{i=1}^{m} E_i\right) \leq \sum_{i=1}^{m} P\,(E_i).$$

The left-hand side is precisely the family-wise error rate (the probability that *at least one* of the m tests is a false positive). Substituting $P(E_i) = \alpha/m$:

$$\text{FWER} = P\left(\bigcup_{i=1}^{m} E_i\right) \leq \sum_{i=1}^{m} \frac{\alpha}{m} = m \times \frac{\alpha}{m} = \alpha.$$

Crucially, Boole's inequality requires no assumption about independence or any other dependence structure among the events E_i, so this bound holds universally — whether the m SNPs are in strong LD (Chapter 15), completely independent, or anything in between. ■

Because the Bonferroni inequality holds regardless of dependence, it is always *valid* — but, as the proof makes clear, it can be very *conservative* (far below the true FWER of α) whenever the tests are strongly positively correlated (as with SNPs in high LD), since the union bound then substantially overstates the true probability of at least one false positive; this is precisely why the permutation-based alternative of this section, which directly accounts for the actual dependence structure in the data, achieves the same α-level FWER control with less loss of power in an LD-rich candidate-gene region.

Worked Example 12.2 (Bonferroni correction for a Nigerian hypertension candidate-gene panel).

A Nigerian cardiovascular genetics unit tests 12 SNPs across three candidate genes (AGT, ACE, and NOS3, all previously implicated in blood pressure regulation) for association with hypertension in a Lagos-based case–control sample. To control the family-wise error rate at the conventional $\alpha = 0.05$ using Theorem 12.1, each individual SNP must meet the Bonferroni-corrected threshold of $\alpha/m = 0.05/12 \approx 0.00417$. Suppose the smallest observed p-value, for a NOS3 variant, is 0.0021 — this passes the Bonferroni threshold ($0.0021 < 0.00417$) and would be declared significant, with the FWER guaranteed (by Theorem 12.1) not to exceed 5% across all 12 tests, regardless of how correlated the SNPs

Theorem 12.2 (The Bonferroni Bound on Family-Wise Error Rate)

If m hypothesis tests are each conducted at significance level alpha/m, then the probability of at least one false positive among all m tests (the family-wise error rate) is at most alpha, regardless of the dependence structure among the tests.

Proof. Let A_1, ..., A_m denote the events that test i produces a false positive (rejects a true null hypothesis), with P(A_i) <= alpha/m for each i by construction. By Boole's inequality (the union bound), for any collection of events, P(A_1 union A_2 union ... union A_m) <= sum_{i=1}^{m} P(A_i). Substituting the per-test bound gives P(at least one false positive) <= sum_{i=1}^{m} (alpha/m) = m times (alpha/m) = alpha. Crucially, Boole's inequality requires no assumption about independence or any other dependence structure among the events A_i, so this bound holds even when the m genetic markers being tested are in strong linkage disequilibrium with each other (Section 12.5) — a property that also explains why the Bonferroni correction, while always valid, can be conservative (overly strict) precisely in the correlated-marker setting typical of a candidate-gene study, motivating the permutation-based alternative of Section 12.7.

12.6: Chapter Summary

- A candidate-gene study tests a small, pre-specified set of genes chosen for biological plausibility, prior linkage evidence, or previous literature reports.
- Tag-SNP selection, functional annotation, and MAF thresholds guide which variants within a candidate gene are genotyped.
- The design is cheaper and better powered per marker than GWAS, but is limited to the genes considered and has a

well-documented history of non-replicating false positives; independent replication is essential.

- The *APOL1*–kidney disease association illustrates both a genuine, population-specific candidate-gene finding and the need for local validation of any risk allele before it is applied in a Nigerian or West African clinical setting.
- Multiple testing within a gene is usually corrected with a Bonferroni bound or, more accurately given LD among SNPs, a permutation test.

12.7: Computational Corner: Candidate-Gene Regression with Permutation-Based Multiple Testing

R.

```
# Simulated candidate-gene study: 5 tag SNPs genotyped (0/1/2 coded), one trait
set.seed(55)
n <-400
snp <-matrix(rbinom(n *5, 2, prob =0.3), nrow = n, ncol =5)
colnames(snp) <-paste0("SNP", 1:5)
beta_true <-c(0, 0, 0.35, 0, 0)              # only SNP3 truly affects the trait
trait <-120+ snp %*% beta_true +rnorm(n, 0, 5)  # e.g., systolic BP (mmHg)

pvals <-sapply(1:5, function(j) summary(lm(trait ~ snp[, j]))$coefficients[2, 4])
names(pvals) <-colnames(snp)
print(pvals)

# Bonferroni correction
bonferroni_sig <- pvals <0.05/5
print(bonferroni_sig)
```

```
# Permutation-based family-wise p-value for the minimum observed p-value
n_perm <-2000
min_p_obs <-min(pvals)
perm_min_p <-replicate(n_perm, {
  trait_perm <-sample(trait)
min(sapply(1:5, function(j) summary(lm(trait_perm ~ snp[, j]))$coefficients[2, 4]))
})
empirical_p <-mean(perm_min_p <= min_p_obs)
cat("Empirical (permutation) family-wise p-value:", empirical_p, "\n")
```

Python.

```
import numpy as np
import statsmodels.api as sm

rng = np.random.default_rng(55)
n =400
snp = rng.binomial(2, 0.3, size=(n, 5))
beta_true = np.array([0, 0, 0.35, 0, 0])
trait =120+ snp @ beta_true + rng.normal(0, 5, n)

def snp_pvalues(trait_vec, snp_mat):
   pvals = []
for j inrange(snp_mat.shape[1]):
      X = sm.add_constant(snp_mat[:, j])
      model = sm.OLS(trait_vec, X).fit()
       pvals.append(model.pvalues[1])
return np.array(pvals)

pvals = snp_pvalues(trait, snp)
print(pvals)
print("Bonferroni significant:", pvals <0.05/5)
```

```
n_perm =2000
min_p_obs = pvals.min()
perm_min_ps = []
for _ inrange(n_perm):
    trait_perm = rng.permutation(trait)
    perm_min_ps.append(snp_pvalues(trait_perm, snp).min())
perm_min_ps = np.array(perm_min_ps)
empirical_p = np.mean(perm_min_ps <= min_p_obs)
print("Empirical (permutation) family-wise p-value:",
empirical_p)
```

In this simulated study, only SNP3 was given a genuine effect on the trait; both the Bonferroni-corrected threshold and the permutation-based family-wise p-value should correctly flag SNP3 while controlling the overall false-positive rate across all five tested SNPs, illustrating in miniature the multiple-testing discipline required of every candidate-gene report before it is accepted as more than a chance finding (Section 12.4).

Mini-Project

Mini-Project 12: A Full Candidate-Gene Study with Multiple Testing Correction

Simulate a candidate-gene study covering multiple SNPs within one gene and correctly apply a multiple-testing correction appropriate to their correlation structure.

Step 1: Simulate genotypes for 10 correlated SNPs (in modest linkage disequilibrium with each other) across 600 case-control individuals, with one true causal SNP.

Step 2: Test each SNP for association individually.

Step 3: Apply both a Bonferroni correction and a permutation-based correction (Section 12.7) accounting for the SNPs' correlation.

Step 4: Compare how many SNPs remain significant under each correction method.

Step 5. Write a short paragraph explaining why the permutation-based correction differs from Bonferroni when SNPs are correlated.

Deliverable: *Simulation code, per-SNP test results, and a comparison of Bonferroni vs. permutation-based significant SNP counts.*

12.8: Practice Questions

Q1. A candidate-gene study genotypes 8 SNPs in the *TCF7L2* gene region for association with type-2 diabetes in a Lagos cohort. At what per-SNP p-value threshold would a Bonferroni correction declare significance at an overall á = 0.05?

Q2. Explain, in two or three sentences, why a permutation-based multiple-testing correction is generally preferred over a simple Bonferroni correction when the tested SNPs are in strong linkage disequilibrium with one another.

Q3. Give two reasons, distinct from multiple testing, why candidate-gene studies of the 1990s and 2000s produced an unusually high proportion of findings that failed to replicate.

Q4. Why does the high frequency of *APOL1* risk alleles in West African populations, contrasted with their near-absence in European populations, illustrate a general limitation of transferring "risk allele" findings across populations without local validation?

Q5. Using the R or Python code of Section 12.7, change *beta_true* so that two SNPs (say SNP2 and SNP4) have a true non-zero effect, re-run the simulation, and report whether the Bonferroni correction and the permutation test both correctly identify both SNPs.

Q6. A candidate-gene study genotypes 6 tag SNPs, chosen to represent 40 total common SNPs in the gene region at $r^2 > 0.9$. Explain why testing only the 6 tag SNPs, rather than all 40 SNPs, does not meaningfully reduce the study's ability to detect a true association somewhere in the gene, provided the causal variant is itself common.

13 Family-Based Association Tests: The TDT and FBAT

13.1: Learning Objectives

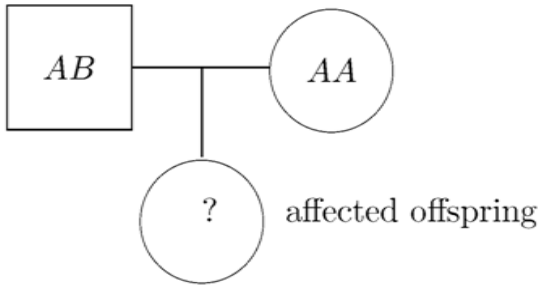
By the end of this chapter, the reader should be able to explain why family-based association designs are robust to population stratification, derive and apply the Transmission Disequilibrium Test (TDT) to trio data, describe the generalisation of the TDT to the Family-Based Association Test (FBAT) framework, and implement both tests in R and Python.

13.2: Why Family-Based Designs? Robustness to Population Substructure

Chapter 11's population-based case–control tests compare unrelated cases and unrelated controls, and are consequently vulnerable to **population stratification** (formally developed in Chapter 14): if cases and controls differ, even slightly, in their ancestral population mixture, and allele frequencies differ across those ancestral populations, a spurious association can appear at any marker whose allele frequency happens to differ between the ancestral groups — entirely independent of any true causal effect. **Family-based association designs** sidestep this problem by using **genetically related controls**: instead of comparing affected individuals to a separately sampled control group, the comparison is made *within* families, typically between the alleles actually transmitted from a parent to an affected offspring and

the alleles that the same parent could have transmitted, but did not. Because this comparison relies only on the parents' own transmissions — governed by Mendel's first law (Chapter 2), which holds regardless of population structure, allele frequencies in the wider population, or departures from Hardy–Weinberg equilibrium (Chapter 3) — family-based tests are inherently robust to stratification, cryptic relatedness, and failure of HWE in the sampled population, at the cost of requiring parental (or sibling) genotypes rather than unrelated controls, and typically less statistical power per individual sampled.

13.3 The Transmission Disequilibrium Test (TDT)



A parent–offspring trio: heterozygous parent AB, homozygous parent AA, and an affected offspring whose transmitted allele is of interest — the basic TDT data unit.

The classical family-based design samples **trios**: an affected offspring together with both parents. At a bi-allelic marker with alleles M and m, each heterozygous parent (genotype Mm) transmits either M or m to the affected child; homozygous parents are uninformative, since they can only transmit one allele regardless of any association. Under the null hypothesis of no linkage (or no association, given linkage), Mendel's first law implies that a heterozygous parent transmits each allele to an affected child with probability exactly ½, independent of which allele is associated with disease risk. The TDT tabulates, over all

heterozygous parents in the sample, the number of times allele M is transmitted and not-transmitted:

	Allele m not transmitted	Allele M not transmitted
Allele M transmitted	b	—
Allele m transmitted	—	c

More precisely, let b be the number of heterozygous parents transmitting M and not m, and c be the number transmitting m and not M (each heterozygous parent contributes exactly one transmitted and one non-transmitted allele). Under the null hypothesis, b and c should be equal in expectation, and the TDT statistic is simply **McNemar's test** applied to this 2×2 table of transmitted vs. non-transmitted alleles:

$$TDT = (b - c)^2/(b + c)$$

which is asymptotically distributed as chi-square with 1 degree of freedom under the null hypothesis of no linkage/no association. A significant TDT statistic indicates that the marker is both linked to, and in population-level linkage disequilibrium with, a nearby disease-susceptibility locus.

Theorem 13.1: (The TDT is McNemar's Test)

Under the null hypothesis of no linkage between the marker and any disease locus, the TDT statistic $(b - c)^2/(b + c)$ *is asymptotically distributed as* χ_1^2*, identical in form and justification to McNemar's test for paired binary data.*

Proof. By Theorem 1.1 (the Law of Equal Segregation), each heterozygous parent transmits allele M or m to an affected child with probability exactly 1/2 each, under the null hypothesis of no linkage. Thus, out of the $n = b + c$ heterozygous-parent transmissions observed, the number transmitting M (namely b) follows a Binomial(n, 1/2)

distribution under the null. The expected value is $E[b] = n/2 = (b+c)/2$ and the variance is $\mathrm{Var}(b) = n(1/2)(1/2) = (b+c)/4$. A standard normal-approximation (Wald-type) test statistic for $H_0: E[b] = (b+c)/2$ is

$$Z = \frac{b - (b+c)/2}{\sqrt{(b+c)/4}} = \frac{(2b - b - c)/2}{\sqrt{(b+c)}/2} = \frac{b-c}{\sqrt{b+c}},$$

so that $Z^2 = (b-c)^2/(b+c)$ is exactly the TDT statistic, and $Z^2 \sim \chi_1^2$ asymptotically by the Central Limit Theorem applied to the Binomial(n,1/2) count b (since the square of a standard normal random variable is χ_1^2 by definition). This is precisely the classical derivation of **McNemar's test** for paired binary outcomes (here, the "pair" being the transmitted and non-transmitted allele from each heterozygous parent), confirming that the TDT is not merely *analogous* to McNemar's test but is exactly McNemar's test applied to this specific genetic data structure. ■

Worked Example 13.1

In a study of 120 Nigerian parent–offspring trios ascertained through a child affected with a congenital condition, 90 parents are heterozygous (Mm) at a candidate marker. Of these, allele M is transmitted to the affected child in 58 trios and allele m in 32 trios.

$$TDT = (58-32)^2/(58+32) = 676/90 = 7.51$$

Comparing 7.51 to the chi-square distribution with 1 degree of freedom (critical value 3.84 at $\alpha = 0.05$) shows this result is statistically significant ($p \approx 0.006$), providing evidence that the M allele is preferentially transmitted to affected offspring — i.e., that the marker is linked to, and in disequilibrium with, a susceptibility variant.

Worked Example 13.2 (TDT for a cleft-lip/palate candidate gene in an Ibadan trio study)
A craniofacial genetics unit at a Nigerian teaching hospital collects 150 case-parent trios for children born with non-syndromic cleft lip/palate, genotyping a candidate SNP in the *IRF6* gene (a well-replicated cleft-risk gene across several populations). Of the 150 trios, 100 parents are heterozygous at the SNP; allele A is transmitted to the affected child in 68 trios and allele G in 32 trios. By Theorem 13.1, $TDT = (68-32)^2/(68+32) = 1296/100 = 12.96$, far exceeding the χ^2_1 critical value of 3.84 ($p < 0.001$) — strong evidence that this *IRF6* variant is linked to, and in disequilibrium with, a cleft-lip/palate susceptibility locus in this Nigerian trio sample, consistent with *IRF6*'s established role in orofacial clefting reported in other populations, and illustrating exactly the kind of family-based replication (Section 13.5) that strengthens confidence in a candidate-gene finding across populations.

13.4: Generalising the TDT: The FBAT Framework

The simple TDT applies only to a single affected child per trio, at a bi-allelic marker, for a binary trait. The **Family-Based Association Test (FBAT)**, developed by Rabinowitz and Laird, generalises this idea into a unified framework that accommodates multiple siblings per family (affected or not), missing parental genotypes (using the genotypes of other family members to reconstruct the missing information where possible), multi-allelic markers and haplotypes, and quantitative as well as binary traits. The FBAT statistic has the general form:

$$FBAT = \sum_i T_i\,(X_i - E[X_i|H_i])/\sqrt{Var}\left(\sum_i T_i\,X_i|H\right)$$

13.5: Power Considerations and Practical Use

Family-based tests generally require larger sample sizes than population-based case–control tests to achieve the same statistical power, because only heterozygous, informative parent–offspring transmissions contribute to the test, and collecting parental (or sibling) DNA is often logistically harder — and, in adult-onset disease studies, sometimes impossible, if parents are deceased — than recruiting unrelated cases and controls. In practice, family-based and population-based designs are frequently viewed as complementary rather than competing: an initial population-based GWAS finding (Chapter 18), potentially confounded by undetected stratification, can be confirmed with a family-based replication in trios, providing much stronger evidence that the observed marker–trait relationship is not an artefact of population structure — a strategy of particular relevance in genetically diverse and admixed Nigerian and West African populations (Chapter 14).

Theorem 13.2 (Asymptotic Chi-Square Distribution of the TDT Statistic)

Under the null hypothesis of no linkage or no association, the Transmission Disequilibrium Test statistic TDT = (b - c)^2 / (b + c), where b and c are the counts of the two types of heterozygous-parent transmissions (Section 13.3), converges in distribution to a chi-square distribution with one degree of freedom as the number of informative (heterozygous-parent) transmissions grows.

Proof. Under the null hypothesis, each heterozygous parent transmits either allele to the affected offspring with equal probability 1/2, independently across families (this is precisely Theorem 1.1's Law of Equal Segregation applied to the transmitting parent). Hence, conditional on the total number of informative transmissions n = b + c, the count b is Binomial(n, 1/2) under the null. The TDT statistic is recognisable as McNemar's test statistic, a standard result in

categorical data analysis: for a Binomial(n, 1/2) count b, the standardised quantity (b - n/2)/sqrt(n/4) converges in distribution to a standard Normal by the Central Limit Theorem, and squaring a standard Normal random variable yields a chi-square distribution with one degree of freedom. Since (b-c)^2/(b+c) is algebraically equal to [(b - n/2)/sqrt(n/4)]^2 when n = b+c, the TDT statistic converges to a chi-square(1) distribution under the null, justifying the significance testing procedure used throughout Section 13.3.

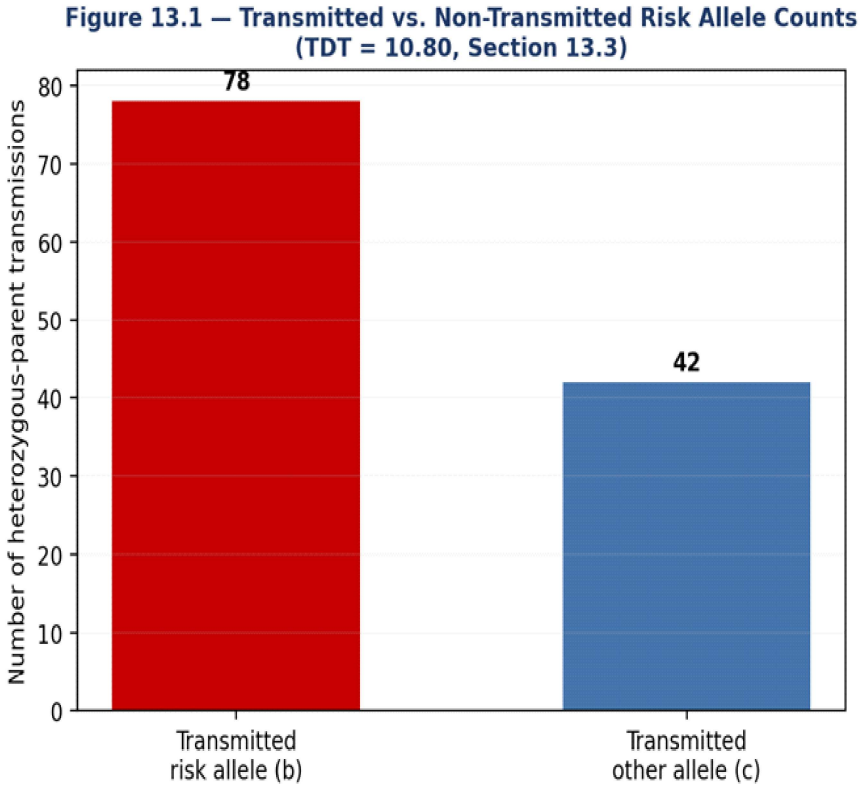


Figure 13.1: *Transmitted vs. non-transmitted risk Allele counts and the resulting TDT statistic (Section 13.3).*

13.6: Chapter Summary

- Family-based association tests use genetically related individuals (typically parents) as "controls," comparing transmitted to non-transmitted alleles, and are therefore inherently robust to population stratification and cryptic relatedness.
- The Transmission Disequilibrium Test (TDT) applies McNemar's test to the 2×2 table of transmitted vs. non-transmitted alleles from heterozygous parents of affected offspring: $TDT = (b-c)^2/(b+c)$, asymptotically chi-square with 1 df.
- The Family-Based Association Test (FBAT) generalises the TDT to multiple siblings, missing parents, multi-allelic markers, and quantitative traits, via a statistic conditioned on a sufficient statistic capturing the available parental transmission information; the TDT is the special case of FBAT for a single affected child and a binary trait.
- Family-based designs typically require larger samples than population-based designs for equivalent power, and are often used to confirm population-based findings.

13.7: Computational Corner: Computing the TDT in R and Python

R.

```
# TDT from trio data: simulate 120 trios, 90 heterozygous parents
set.seed(13)
n_het_parents <-90
p_transmit_M <-0.62# true (elevated) transmission probability
transmissions <-rbinom(n_het_parents, 1, p_transmit_M)  # 1 = M transmitted, 0 = m
b <-sum(transmissions ==1)
c <-sum(transmissions ==0)
```

```
tdt_stat <- (b - c)^2/ (b + c)
p_value <-pchisq(tdt_stat, df =1, lower.tail =FALSE)
cat("b =", b, " c =", c, " TDT =", tdt_stat, " p-value =",
p_value, "\n")

# Equivalently, via mcnemar.test on the 2x2 transmitted/non-
transmitted table
tab <-matrix(c(b, 0, 0, c), nrow =2,
dimnames =list(transmitted =c("M", "m"), nontransmitted
=c("m", "M")))
mcnemar.test(matrix(c(0, c, b, 0), nrow =2), correct =FALSE)
```

Python.

```
import numpy as np
from scipy.stats import chi2

rng = np.random.default_rng(13)
n_het_parents =90
p_transmit_M =0.62
transmissions = rng.binomial(1, p_transmit_M,
n_het_parents)

b = np.sum(transmissions ==1)
c = np.sum(transmissions ==0)
tdt_stat = (b - c)**2/ (b + c)
p_value = chi2.sf(tdt_stat, df=1)
print(f"b={b}, c={c}, TDT={tdt_stat:.3f}, p-
value={p_value:.4f}")
```

The simulated transmission probability of 0.62 (favouring allele M) produces, as expected, a TDT statistic well above the critical value of 3.84, correctly detecting the built-in transmission disequilibrium; setting `p_transmit_M <- 0.5` in either script reproduces the null case, and repeated simulation under the null

shows the TDT statistic following its nominal chi-square(1) reference distribution, an instructive simulation-based check of the theory of Section 13.3 in the spirit of the simulation methods developed further in Chapter 20.

Mini-Project

Mini-Project 13: Simulating and Analysing a TDT Study

Simulate trio (parent-parent-offspring) data under a true genetic association and analyse it with the TDT.

Step 1: Simulate genotypes for 300 case-parent trios at a marker with a true relative risk of your choosing.

Step 2: Identify all heterozygous parents and tabulate transmitted versus non-transmitted alleles.

Step 3: Compute the TDT statistic (Section 13.3) and its p-value.

Step 4: Repeat the simulation under a scenario with population stratification (cases and controls drawn from genetically different subpopulations) and compare a population-based test to the TDT.

Step 5: Write a short paragraph explaining why the TDT remains valid under stratification while the population-based test does not.

Deliverable: *Simulation code, TDT results, a side-by-side comparison with a population-based test under stratification, and a short written explanation.*

13.8: Practice Questions

Q1. In a trio study of 200 heterozygous parents of children affected with a cleft lip/palate phenotype, allele T is transmitted 118 times and allele t is transmitted 82 times. Compute the TDT statistic and state whether it is significant at á = 0.01.

Q2. Explain, in your own words, why the TDT remains valid even if the sampled trios come from a population that is a mixture of two ethnic groups with very different allele frequencies at the tested marker.

Q3. A trio study has only 15 heterozygous parents, with b = 10 and c = 5. Compute the TDT statistic. Is it appropriate to rely on the chi-square approximation here, or would an exact test (e.g., an exact binomial test of b vs. c out of b+c trials against p=0.5) be preferable? Explain briefly.

Q4. Explain briefly why the TDT statistic of Section 13.3 is a special case of the general FBAT statistic of Section 13.4, identifying what T_i and X_i correspond to for a single affected child at a bi-allelic marker.

Q5. Using the R or Python code of Section 13.7, simulate the TDT statistic's distribution under the null (*p_transmit_M = 0.5*) for 1,000 replications of 90 heterozygous parents each, and report what proportion of replications exceed the chi-square(1) critical value of 3.84 at nominal á = 0.05. Comment on whether this matches the nominal type-I error rate.

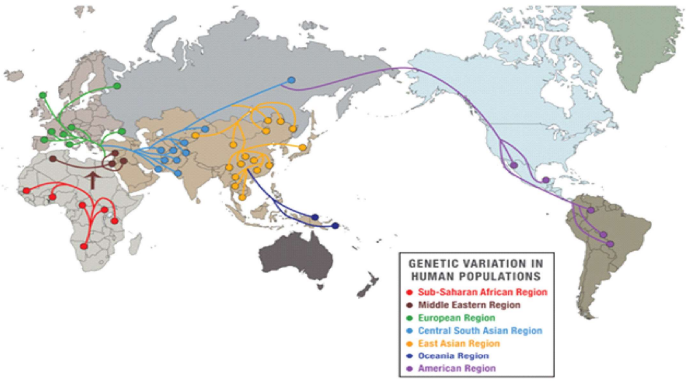
Q6. A researcher has genotype data on 80 affected children but only 30 of their corresponding parent pairs (the rest of the families have missing or unavailable parental DNA). Explain, at a conceptual level, why the FBAT framework (Section 13.4) can potentially still extract some information from the 50 families with missing parents, whereas the classical trio-based TDT of Section 13.3 cannot use them at all.

14 Population Substructure, Stratification, and Admixture

14.1: Learning Objectives

By the end of this chapter, the reader should be able to define population stratification, admixture, and cryptic relatedness as distinct forms of population substructure, explain the Wahlund effect and its consequence for Hardy–Weinberg equilibrium, describe how population stratification confounds population-based association tests, and detect and correct for stratification using genomic control and principal component analysis in R and Python.14.2: What is Population Substructure?

14.2: What is Population Substructure?

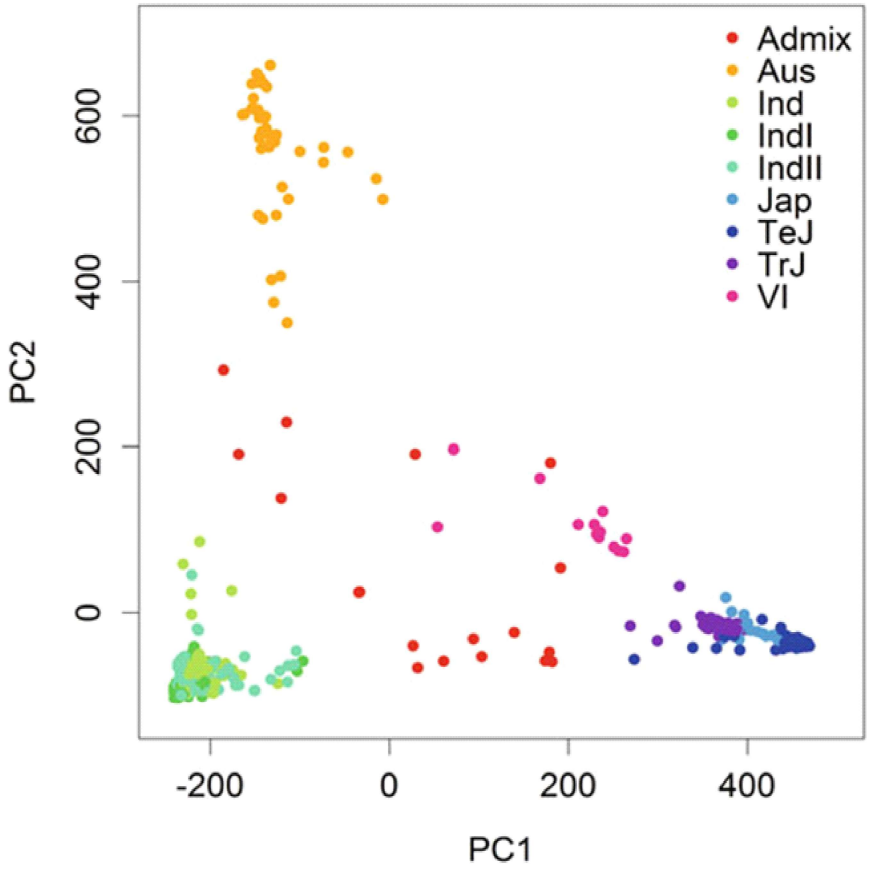


Global human genetic diversity and migration patterns out of Africa, the deep demographic history underlying present-day population substructure.

Population substructure refers to any feature of a sampled population that causes allele and/or genotype frequencies to vary systematically across individuals in ways that do not reflect random Mendelian sampling alone. Three related but distinct phenomena are usually distinguished:

- **Population stratification**: the sample is drawn from two or more genetically distinct subgroups (e.g., distinct ethnic or ancestral groups) mixed together, without the individuals themselves being recent-generation admixed individuals — for example, a Nigerian national sample combining Yoruba, Igbo, and Hausa/Fulani participants, each with somewhat different allele frequencies at many loci due to their distinct population histories.
- **Population admixture**: individuals themselves carry ancestry from two or more historically distinct source populations, arising from inter-population mating over recent generations (often occult, i.e. not evident from self-reported ethnicity) — a feature of many African-American, Afro-Brazilian, and other diaspora populations, and increasingly relevant within Nigeria itself given inter-ethnic marriage.
- **Cryptic relatedness**: some individuals in the sample are more closely related to one another (as in Chapter 6's kinship framework) than would be expected under simple random sampling from the population, without this relatedness being known or recorded by the investigators.

All three forms of substructure can distort standard statistical tests that assume individuals are independent draws from a single, homogeneous, randomly mating population.



Principal component analysis of genome-wide genotype data from multiple real human populations (e.g., admixed, Australian, several Indian and Japanese subgroups), showing that each population forms a visually distinct cluster in PC1–PC2 space — the standard diagnostic display of population substructure in a genetic study.

14.3: The Wahlund Effect: Substructure and Hardy–Weinberg Equilibrium

A foundational result, due to Sten Wahlund (1928), shows that combining two or more subpopulations that are each individually at Hardy–Weinberg equilibrium (Chapter 3), but which differ in allele frequency, produces a **combined sample that departs from Hardy–Weinberg equilibrium** — specifically, showing a deficit of heterozygotes relative to the HWE expectation computed from the pooled allele frequency. Suppose the population consists of two subpopulations in proportions w_1 and $w_2 = 1 - w_1$, with allele frequencies p_1 and p_2 ($p_1 \neq p_2$) at a bi-allelic locus, each subpopulation internally at HWE. The pooled allele frequency is $\bar{p} = w_1 p_1 + w_2 p_2$, but the pooled heterozygote frequency is:

$$Aa(pooled) = w_1 \cdot 2p_1(1 - p_1) + w_2 \cdot 2p_2(1 - p_2)$$
$$= 2\bar{p}(1 - \bar{p}) - 2w_1 w_2 (p_1 - p_2)^2$$

The term subtracted, $2w_1w_2(p_1-p_2)^2$, is always non-negative, and is exactly zero only when $p_1 = p_2$ (no true subdivision) — so the pooled sample always shows *at least as many* homozygotes, and correspondingly fewer heterozygotes, as Hardy–Weinberg equilibrium computed at the pooled allele frequency would predict. This **Wahlund effect** is the population-genetic signature of undetected stratification, and is one of the classical diagnostic clues (alongside an excess of significant HWE tests genome-wide) that a sample thought to be drawn from a single homogeneous population is, in fact, structured.

Theorem 14.1: (The Wahlund Effect)

If a population is a mixture of K subpopulations in proportions $w_1, \ldots, w_K$ *(*$\sum_k w_k = 1$*), each internally at Hardy–Weinberg equilibrium at a bi-allelic locus with allele frequency* p_k*, then the pooled heterozygosity satisfies* $H_{pooled} \leq 2\bar{p}(1-\bar{p})$*, where* $\bar{p} = \sum_k w_k\, p_k$*, with equality if and only if all* p_k *are equal.*

Proof (two-subpopulation case; the general case follows by the same argument applied to the variance of p_k *across subpopulations).* The pooled heterozygote frequency is the weighted average of each subpopulation's own HWE heterozygosity: $H_{pooled} = \sum_k w_k \cdot 2p_k(1-p_k)$. Write $2\bar{p}(1-\bar{p}) - H_{pooled}$ and expand, using $\bar{p} = \sum_k w_k\, p_k$:

$$
\begin{aligned}
&2\bar{p}(1-\bar{p}) - \sum_k w_k \cdot 2p_k(1-p_k) \\
&= 2\left[\bar{p} - \bar{p}^2 - \sum_k w_k\, p_k + \sum_k w_k\, p_k^2\right] \\
&= 2\left[\sum_k w_k\, p_k^2 - \bar{p}^2\right] = 2\,\mathrm{Var}_w(p_k),
\end{aligned}
$$

where $\mathrm{Var}_w(p_k) = \sum_k w_k\, p_k^2 - \bar{p}^2 = \sum_k w_k\,(p_k - \bar{p})^2$ is the (weighted) variance of the allele frequency across subpopulations — itself always ≥ 0 as a sum of squares, with equality if and only if every $p_k = \bar{p}$ (no true differentiation). Hence $2\bar{p}(1-\bar{p}) - H_{pooled} = 2\,\mathrm{Var}_w(p_k) \geq 0$, establishing $H_{pooled} \leq 2\bar{p}(1-\bar{p})$ as claimed, with the deficit exactly equal to twice the cross-subpopulation allele-frequency variance. ■

This proof reveals something the raw formula alone does not make obvious: the Wahlund deficit is not an arbitrary

correction term but is *exactly* twice the variance of allele frequency across the mixed subpopulations — connecting the population-genetic phenomenon directly to a basic statistical quantity, and showing immediately why more genetically differentiated subpopulations (larger $\mathrm{Var}_w(p_k)$) produce a larger apparent heterozygote deficit when pooled.

Worked Example 14.2 (Wahlund effect across three Nigerian ethno-linguistic groups)

A national biobank pools genotype data from three broad Nigerian ethno-linguistic groups without correcting for group membership: Yoruba ($w_1 = 0.35$, $p_1 = 0.62$ at a candidate marker), Igbo ($w_2 = 0.35$, $p_2 = 0.58$), and Hausa/Fulani ($w_3 = 0.30$, $p_3 = 0.45$). The pooled allele frequency is $\bar{p} = 0.35(0.62) + 0.35(0.58) + 0.30(0.45) = 0.217 + 0.203 + 0.135 = 0.555$. The weighted variance across groups is $\mathrm{Var}_w(p_k) = 0.35(0.62 - 0.555)^2 + 0.35(0.58 - 0.555)^2 + 0.30(0.45 - 0.555)^2 =$
$0.35(0.00423) + 0.35(0.000625) + 0.30(0.011025) =$
$0.00148 + 0.000219 + 0.00331 = 0.00501$. By Theorem 14.1, the Wahlund deficit is $2 \times 0.00501 = 0.01002$, so naive pooled HWE would predict heterozygosity $2(0.555)(0.445) = 0.494$, while the true pooled heterozygosity is only $0.494 - 0.010 = 0.484$ — a modest but real deficit that a genome-wide analyst would need to distinguish from true inbreeding or selection before drawing any biological conclusion, exactly the caution raised throughout this chapter for multi-ethnic Nigerian genetic studies.

Worked Example 14.1.

Two Nigerian subpopulations each internally at HWE have allele frequencies $p_1 = 0.85$ and $p_2 = 0.55$ at a marker, mixed in equal proportions ($w_1 = w_2 = 0.5$).

Aa

$$\bar{p} = 0.5(0.85) + 0.5(0.55) = 0.70$$

$$\text{Expected } Aa \text{ under naive pooled HWE} = 2(0.70)(0.30) = 0.420$$

$$\text{Actual pooled } Aa = 0.420 - 2(0.5)(0.5)(0.85 - 0.55)^2 = 0.420 - 0.045 = 0.375$$

The pooled sample shows a heterozygote deficit of 0.045 (about 11% fewer heterozygotes than naive HWE would predict) purely as a mechanical consequence of mixing two differentiated subpopulations, with no biological non-random mating, inbreeding, or selection required at all — illustrating why an apparent departure from HWE in a genetic epidemiology dataset must always be checked against the possibility of hidden population structure before being attributed to genotyping error, selection, or inbreeding (Chapter 5).

14.4: Confounding of Population-Based Association Tests

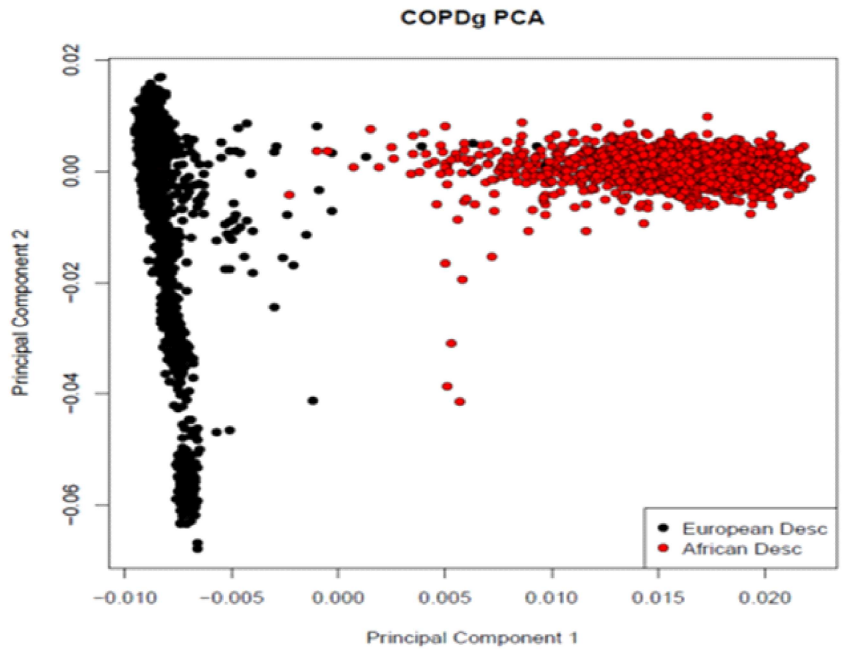
Population stratification becomes a serious statistical problem specifically when case and control samples (Chapter 11) differ, even modestly, in their proportions of underlying subpopulations, *and* the tested marker's allele frequency differs across those subpopulations — a form of **confounding** exactly analogous to confounding in classical epidemiology, with ancestry playing the role of the confounding variable. If a disease happens to be more common in one ancestral group for reasons entirely unrelated to the tested marker (e.g., differing environmental exposures, diet, or unrelated genetic risk factors more prevalent in that group), and that group also happens to carry a different frequency of the tested allele, a population-based case–control test (Chapter 11) will detect a spurious association at that marker — a false positive driven entirely by ancestry confounding rather than by any true biological effect of the marker on disease risk. This is precisely

the vulnerability that motivated the family-based designs of Chapter 13, and it is a serious practical concern for genetic studies conducted across Nigeria's genetically and ethnically diverse population, where case and control recruitment sites (e.g., different teaching hospitals) may inadvertently draw disproportionately from different ethnic-linguistic groups.

14.5: Detecting and Correcting for Stratification

Genomic control: As introduced in Chapter 18 (Section 18.5), genomic control compares the median of the observed test statistics across many genome-wide markers (most of which are not truly associated with the trait) to the median expected under the null distribution, yielding the inflation factor λ_{GC} = median(χ^2_{obs}) / median(χ^2_{null}). A value of λ_{GC} appreciably greater than 1 signals systematic inflation of test statistics — consistent with, though not uniquely diagnostic of, population stratification — and can be used to deflate all test statistics by dividing by λ_{GC} before assessing significance, a simple if somewhat crude correction.

Principal component analysis (PCA): The modern standard approach, introduced to genetics by Price and colleagues (2006), applies **principal component analysis** to a matrix of genome-wide genotypes to extract the major axes of ancestry variation within the sample — the leading principal components typically correspond closely to known continental or sub-continental ancestry gradients (for example, distinguishing Yoruba, Igbo, and Hausa/Fulani ancestry components within a Nigerian sample, or separating West African from other African or non-African ancestry in an admixed diaspora sample).



Principal component analysis separating two ancestry clusters (e.g. African vs. European descent) from genome-wide genotype data.

Including the leading principal components as covariates in the logistic (or linear) regression association model of Section 11.7 adjusts each marker's association test for the individual's estimated ancestry, removing most of the stratification-induced confounding while retaining the statistical power of the population-based design.

Mixed models: An alternative, increasingly standard in large modern GWAS, replaces PCA-adjusted regression with a **linear mixed model** that includes a random effect with covariance structure given by a genome-wide **genetic relatedness matrix** (a genome-wide analogue of the kinship matrix of Chapter 6),

simultaneously correcting for both distant population stratification and cryptic close relatedness within the sample in a single unified framework.

Theorem 14.2 (The Wahlund Inequality).
Let a population consist of K subpopulations with allele frequencies p_1, ..., p_K (each subpopulation in Hardy-Weinberg equilibrium internally) mixed in proportions w_1, ..., w_K. Then the expected heterozygosity in the pooled population is less than or equal to the heterozygosity computed from the pooled (average) allele frequency, with equality if and only if all subpopulations have identical allele frequency.

Proof. Let pbar = sum_k w_k p_k be the pooled allele frequency. The true expected heterozygosity in the pooled population is H_pooled = sum_k w_k [2 p_k (1-p_k)], since each subpopulation individually satisfies Hardy-Weinberg proportions. The heterozygosity that would be expected if the pooled population were itself in Hardy-Weinberg equilibrium at the pooled allele frequency is H_HWE = 2 pbar (1 - pbar). Consider the variance of p_k across subpopulations (weighted by w_k): Var(p) = sum_k w_k (p_k - pbar)^2 = sum_k w_k p_k^2 - pbar^2 >= 0. Rearranging, sum_k w_k p_k^2 >= pbar^2 + Var(p) - Var(p) = pbar^2, more precisely sum_k w_k p_k^2 = pbar^2 + Var(p). Then H_HWE - H_pooled = 2pbar(1-pbar) - 2 sum_k w_k p_k(1-p_k) = 2[pbar - pbar^2 - sum_k w_k p_k + sum_k w_k p_k^2] = 2[sum_k w_k p_k^2 - pbar^2] = 2 Var(p) >= 0, with equality if and only if Var(p) = 0, i.e., all p_k are equal. This proves H_pooled <= H_HWE, the formal statement of the Wahlund effect introduced descriptively in Section 14.3: population substructure always reduces observed heterozygosity relative to Hardy-Weinberg expectation, and strictly so whenever subpopulation allele frequencies genuinely differ.

14.6: Chapter Summary

- Population stratification (distinct subgroups), admixture (mixed ancestry within individuals), and cryptic relatedness are three related forms of population substructure that violate the assumption of a single, homogeneous, randomly mating sample.
- The Wahlund effect shows that pooling HWE subpopulations with different allele frequencies mechanically produces a heterozygote deficit in the combined sample, a classical signature of hidden structure.
- Population stratification confounds population-based case–control association tests whenever case/control ancestry proportions and tested-marker allele frequencies both differ across ancestral subgroups.
- Genomic control, principal component adjustment, and genome-wide mixed models are the standard tools for detecting and correcting stratification in modern association studies, including GWAS (Chapter 18).

14.7: Computational Corner: Simulating the Wahlund Effect and Detecting Structure by PCA

R.

```
# Simulate the Wahlund effect: two subpopulations, unequal allele frequencies
set.seed(7)
n_per_pop <-1000
p1 <-0.85; p2 <-0.55

geno_pop1 <-rbinom(n_per_pop, 2, p1)
geno_pop2 <-rbinom(n_per_pop, 2, p2)
pooled <-c(geno_pop1, geno_pop2)

p_bar <-mean(pooled) /2
```

```
observed_het <-mean(pooled ==1)
expected_het_naive <-2* p_bar * (1- p_bar)
cat("Pooled allele freq:", p_bar, "\n")
cat("Observed heterozygosity:", observed_het, " Naive-HWE
expected:", expected_het_naive, "\n")

# PCA-based stratification detection: simulate 200 markers
differentiated between two groups
n_snps <-200
group <-rep(c(1, 2), each = n_per_pop)
allele_freqs_pop1 <-runif(n_snps, 0.1, 0.9)
allele_freqs_pop2 <-pmin(pmax(allele_freqs_pop1
+rnorm(n_snps, 0, 0.15), 0.01), 0.99)

geno_matrix <-matrix(0, nrow =2* n_per_pop, ncol =
n_snps)
for (j in1:n_snps) {
  geno_matrix[group ==1, j] <-rbinom(n_per_pop, 2,
allele_freqs_pop1[j])
  geno_matrix[group ==2, j] <-rbinom(n_per_pop, 2,
allele_freqs_pop2[j])
}
pca <-prcomp(geno_matrix, scale. =TRUE)
plot(pca$x[, 1], pca$x[, 2], col = group, pch =20,
xlab ="PC1", ylab ="PC2", main ="PCA reveals two
simulated subpopulations")
```

Python.

```
import numpy as np
from sklearn.decomposition import PCA

rng = np.random.default_rng(7)
n_per_pop =1000
p1, p2 =0.85, 0.55
```

```
geno_pop1 = rng.binomial(2, p1, n_per_pop)
geno_pop2 = rng.binomial(2, p2, n_per_pop)
pooled = np.concatenate([geno_pop1, geno_pop2])

p_bar = pooled.mean() /2
observed_het = np.mean(pooled ==1)
expected_het_naive =2* p_bar * (1- p_bar)
print("Pooled allele freq:", p_bar)
print("Observed heterozygosity:", observed_het, "Naive-
HWE expected:", expected_het_naive)

n_snps =200
group = np.array([1]*n_per_pop + [2]*n_per_pop)
af1 = rng.uniform(0.1, 0.9, n_snps)
af2 = np.clip(af1 + rng.normal(0, 0.15, n_snps), 0.01, 0.99)

geno_matrix = np.zeros((2*n_per_pop, n_snps))
for j inrange(n_snps):
   geno_matrix[group ==1, j] = rng.binomial(2, af1[j],
n_per_pop)
   geno_matrix[group ==2, j] = rng.binomial(2, af2[j],
n_per_pop)

pca = PCA(n_components=2)
scores = pca.fit_transform((geno_matrix -
geno_matrix.mean(0)) / geno_matrix.std(0))
# scores[:, 0] vs scores[:, 1], coloured by 'group', will
separate the two simulated subpopulations.
```

The first simulation reproduces the Wahlund heterozygote deficit of Worked Example 14.1 directly from simulated genotype data rather than algebra, and the second shows the now-standard genome-wide PCA approach: even though no individual was told which subpopulation they belonged to, the first principal component of the genotype matrix cleanly separates the two

simulated groups, exactly the diagnostic plot used in practice (Section 14.5) to check a real GWAS sample for hidden ancestry structure before running any association test.

Mini-Project

Mini-Project 14: Detecting and Correcting for Population Stratification

Simulate a stratified population exhibiting a Wahlund effect and demonstrate both the problem and a standard correction.

Step 1: Simulate genotypes for two subpopulations with different allele frequencies at 200 markers, then pool them into a single sample.

Step 2: Demonstrate the Wahlund effect (Section 14.3) by comparing observed heterozygosity in the pooled sample to Hardy-Weinberg expectations.

Step 3: Run a case-control association test on a null marker (no true effect) using the pooled, unstratified sample and show it produces inflated false positives.

Step 4: Apply principal component analysis (Section 14.5/14.7) to the marker data and show the first principal component recovers the subpopulation structure.

Step 5: Re-run the association test adjusting for the top principal component and show the false-positive inflation is reduced.

Deliverable: *Simulation code, a PCA plot showing recovered structure, and before/after inflation statistics (e.g., genomic control lambda) for the association test.*

14.8: Practice Questions

Q1. Two Nigerian subpopulations are internally at HWE with allele frequencies p = 0.70 and p, = 0.40, mixed in proportions w = 0.6, w, = 0.4. Compute the pooled allele frequency, the naive-HWE expected heterozygosity, and the actual pooled heterozygosity, and state the size of the Wahlund heterozygote deficit.

Q2. Explain, in two or three sentences, the difference between population stratification and cryptic relatedness, giving a hypothetical Nigerian example of each.

Q3. A case–control GWAS finds genomic control inflation factor $\ddot{e}_{GC} = 1.35$. Explain what this suggests, and describe one further analysis (beyond simply dividing test statistics by 1.35) that a careful analyst should perform before concluding the inflation is due to stratification rather than genuine widespread polygenic association.

Q4. Explain briefly why including the top principal components of genome-wide genotype data as covariates in a logistic regression association model (Section 11.7) can correct for stratification-induced confounding.

Q5. Using the R or Python code of Section 14.7, increase the allele-frequency difference between the two simulated subpopulations (e.g., p1 = 0.9, p2 = 0.3) and re-run the Wahlund-effect simulation. How does the size of the heterozygote deficit change, and why?

Q6. Looking at a real multi-population PCA plot such as the one in Section 14.2, in which several distinct populations each form a tight, separate cluster, explain what an *intermediate* individual — one who falls between two clusters rather than inside either — most plausibly represents biologically.

15 Linkage Disequilibrium and Haplotype Inference

15.1: Learning Objectives

By the end of this chapter, the reader should be able to define linkage disequilibrium (LD), compute the standard LD coefficients D, D′ and r^2, relate LD decay to physical distance and recombination, and describe haplotype inference methods (Clark's algorithm, the EM algorithm, and Bayesian/PHASE-type methods).

15.2: What is Linkage Disequilibrium?

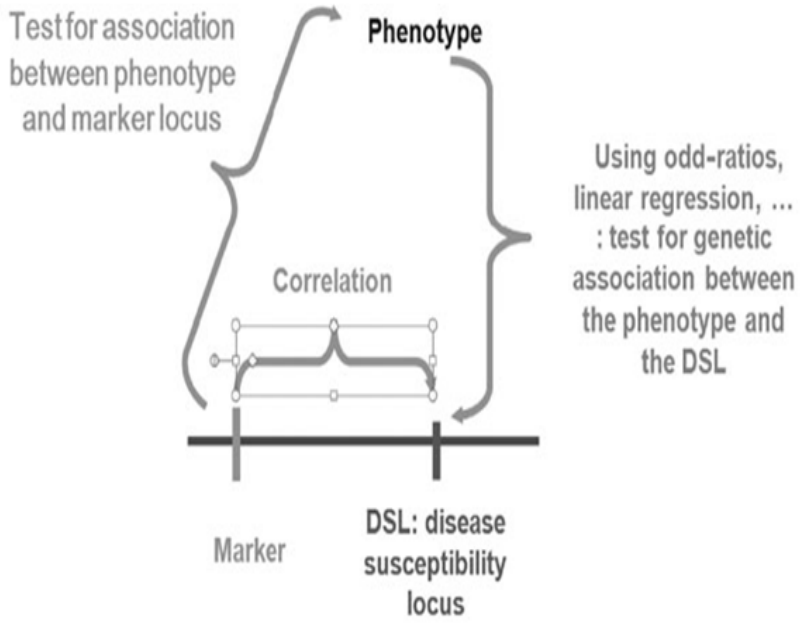
Two loci are said to be in **linkage equilibrium** if the alleles at the two loci occur together in gametes independently of one another — i.e., the haplotype frequency equals the product of the individual allele frequencies. **Linkage disequilibrium (LD)** is the non-random, statistical *association* between alleles at two (typically linked) loci, i.e., a departure from this independence, observed more often than chance in a population.

Consider two bi-allelic loci: locus 1 with alleles A/a (frequencies p_A, p_a) and locus 2 with alleles B/b (frequencies p_B, p_b). If haplotype AB has population frequency h_AB, and there were no LD, we would expect

h_AB = p_A × p_B. The **disequilibrium coefficient**, D, measures the departure from this expectation:

$$D = h_{AB} - p_A p_B$$

Equivalently, D = h_AB h_ab − h_Ab h_aB (the "cross-product" difference among the four haplotype frequencies). D = 0 corresponds to linkage equilibrium; D ≠ 0 indicates LD (either positive, D>0, meaning A and B co-occur more than expected, or negative, D<0).



The logic of association testing via LD: a genotyped marker locus (M) that is in strong LD with an unobserved disease-susceptibility locus (DSL) can reveal a statistical association between the marker and the phenotype, even though the marker itself has no direct biological effect — the correlation is the only reason a marker-based association test (Chapter 11) can succeed without genotyping the causal variant directly.

15.3: Why LD Arises and Decays

LD between two loci commonly arises from: (i) physical proximity, so that a new mutation at one locus initially arises on a single ancestral haplotype background and remains associated with nearby alleles until recombination breaks the association; (ii) population admixture (mixing of two populations with different allele frequencies produces LD genome-wide, even between unlinked loci, analogous to the Wahlund effect of Chapter 3 — of particular relevance to admixed African populations); (iii) selection favouring particular allele combinations; and (iv) genetic drift, especially in small or recently founded populations.

LD **decays** across generations due to recombination (Chapter 8): each generation, a fraction θ (the recombination fraction between the two loci) of gametes is recombinant, breaking down the non-random haplotype association. It can be shown that:

$$D_t = (1 - \theta)^t D_0$$

where D_t is the disequilibrium after t generations of random mating, starting from initial disequilibrium D_0. LD therefore decays geometrically with generation number, at a rate governed by the recombination fraction — tightly linked loci (small θ) retain LD over many more generations than loosely linked loci (large θ, up to θ=0.5 for unlinked loci, for which LD decays fastest, in a single generation under strict random mating, absent other forces sustaining it such as admixture).

Theorem 15.1 (Geometric Decay of Linkage Disequilibrium) *Under random mating with recombination fraction θ between two loci, and no forces (drift, selection, admixture) regenerating disequilibrium, the disequilibrium coefficient after t generations is $D_t = (1-\theta)^t D_0$.*

Proof. Consider the four haplotype frequencies $h_{AB}, h_{Ab}, h_{aB}, h_{ab}$ in generation t, with $D_t = h_{AB,t}h_{ab,t} - h_{Ab,t}h_{aB,t}$. In forming generation $t+1$'s gametes, a fraction $(1-\theta)$ of gametes are non-recombinant (inheriting the parental haplotype intact) and a fraction θ are recombinant (formed by combining the allele at locus 1 from one parental haplotype with the allele at locus 2 from the other). The frequency of haplotype AB in generation $t+1$ is therefore

$$h_{AB,t+1} = (1-\theta)h_{AB,t} + \theta\, p_A p_B,$$

since a recombinant AB gamete requires drawing the A allele (frequency p_A) and the B allele (frequency p_B) independently from the population's separate allele pools (recombination re-randomises the combination). Substituting into $D_{t+1} = h_{AB,t+1} - p_A p_B$ (the equivalent single-term definition of D, Section 15.2):

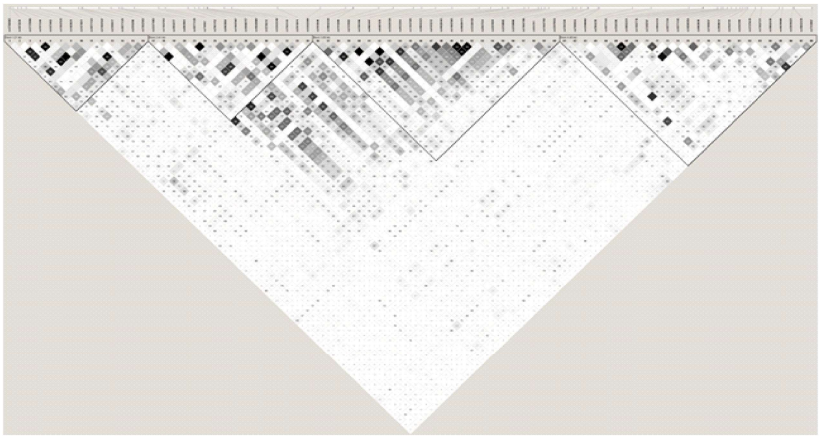
$$\begin{aligned} D_{t+1} &= (1-\theta)h_{AB,t} + \theta p_A p_B - p_A p_B \\ &= (1-\theta)h_{AB,t} - (1-\theta)p_A p_B \\ &= (1-\theta)\big(h_{AB,t} - p_A p_B\big) = (1-\theta)D_t, \end{aligned}$$

using the fact that allele frequencies p_A, p_B themselves do not change under recombination alone (Chapter 3, recombination reshuffles haplotypes but not allele frequencies). This is a first-order linear recurrence with solution $D_t = (1-\theta)^t D_0$ by direct iteration (or induction: the base case $t=0$ is trivial, and the inductive step is exactly the recurrence just derived). ■

Worked Example 15.2 (LD decay between a malaria-resistance marker and a linked SNP, Niger Delta population)

A population geneticist studies two SNPs 50 kb apart, in strong initial LD ($D_0 = 0.08$) in a founding population, flanking a locus associated with malaria resistance in a Niger Delta community. If the recombination fraction between the two SNPs is θ = 0.02 per generation (typical for markers at this physical distance, Section 15.6), Theorem 15.1 predicts the disequilibrium after 25 generations (roughly 500–625 years, at 20–25 years per generation) to be $D_{25} = (1 - 0.02)^{25} \times 0.08 = (0.98)^{25} \times 0.08 \approx 0.603 \times 0.08 \approx 0.0483$ — LD has decayed to about 60% of its initial value, illustrating why closely linked markers can retain population-level LD, and hence remain useful for association mapping (Chapter 11), over many centuries, whereas the same calculation with a more loosely linked marker (say θ = 0.20) would give $(0.80)^{25} \approx 0.0038$, essentially complete decay — the quantitative basis for the tag-SNP selection strategy of Chapter 12.

15.4: Normalised Measures of LD: D2 and r²



A triangular LD heatmap: pairwise r² between nearby SNPs, the standard visual summary of a region's LD structure.

Because D depends on the allele frequencies at the two loci (and so cannot be directly compared across marker pairs with different allele frequencies), two standardised measures are in universal use:

D2 (Lewontin's D2):

$$D' = D/D_{max}\,, D_{max} = min(p_A p_b, p_a p_B) if D > 0; min(p_A p_B, p_a p_b) if D < 0$$

D′ ranges from −1 to 1 (often reported as |D′|, from 0 to 1); |D′| = 1 ("complete LD") means no evidence of historical recombination between the loci within the sample's ancestry — a useful qualitative summary but sensitive to small sample sizes, especially for rare alleles.

r² (the squared correlation coefficient):

$$r^2 = D^2/(p_A p_a p_B p_b)$$

r² ranges from 0 to 1 and has a direct statistical interpretation: it is the squared Pearson correlation between the two loci's allele indicator variables, and is the more relevant measure for association-study design, because the sample size needed to detect an association at a marker in LD with an (untyped) causal variant, relative to typing the causal variant directly, scales approximately as 1/r².

15.5 Worked Example 15.1

Two SNPs are genotyped in a Nigerian population sample. Estimated haplotype frequencies: h_AB = 0.30, h_Ab = 0.10, h_aB = 0.15, h_ab = 0.45 (these sum to 1.00). Allele frequencies: p_A = h_AB + h_Ab = 0.40; p_a = 0.60; p_B = h_AB + h_aB = 0.45; p_b = 0.55.

(a) Compute D.

$$D = h_{AB} - p_A p_B = 0.30 - (0.40)(0.45) = 0.30 - 0.18 = 0.12$$

(b) Compute D′. Since D > 0, D_max = min(p_A p_b, p_a p_B) = min[(0.40)(0.55), (0.60)(0.45)] = min(0.22, 0.27) = 0.22.

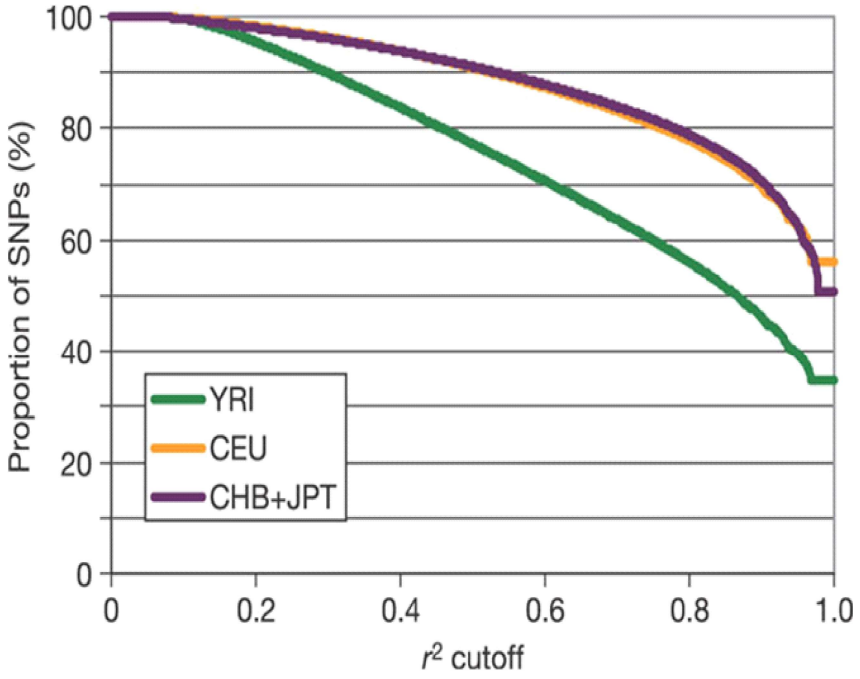
$$D' = 0.12/0.22 = 0.545$$

(c) Compute r².

$$r^2 = (0.12)^2/[(0.40)(0.60)(0.45)(0.55)] = 0.0144/0.0594 = 0.242$$

(d) Interpret. D′ = 0.545 suggests moderate LD (some, but not complete, historical recombination between the loci); r² = 0.242 indicates that a study wishing to detect an association through this marker (locus 1, say), if the true causal variant is locus 2, would need roughly 1/0.242 ≈ 4.1 times the sample size that would be required if the causal variant itself were typed directly — an important practical consideration in designing Nigerian genetic studies that rely on commercially available SNP genotyping arrays, which may capture African-population LD patterns less completely than they capture LD patterns in the (largely European-ancestry) populations used to design many arrays.

15.6: LD and Physical Distance; Relevance to African Populations



LD decays faster with physical distance in the Yoruba (YRI) than in CEU or CHB+JPT samples — the empirical basis for denser marker panels in African association studies.

Because LD decays through recombination accumulated over many generations, and because recombination probability increases (roughly) with physical distance along a chromosome, LD generally decays with increasing physical distance between markers — but the *rate* of this decay differs substantially between populations, depending on their demographic history (effective population size, time since founding or admixture, and number of generations of random mating). African populations, including Nigerian populations, are known from genomic studies to have,

on average, the highest effective population sizes and the most ancient population history among major world populations, and consequently exhibit **more rapid LD decay** (LD extends over shorter physical distances) than populations of more recent common ancestry (e.g., European or East Asian populations descended from smaller founding groups). This has a direct practical consequence for Chapter 18 (GWAS): studies in Nigerian and other African populations generally require denser marker panels (more SNPs per unit physical distance) to achieve the same genomic coverage via LD as studies in populations with more extensive LD — a key motivation for African-specific genotyping array and reference panel development (e.g., the H3Africa initiative).

15.7: Haplotype Inference

Because standard genotyping technologies report unphased genotypes (i.e., for a heterozygote at two or more loci, it is not directly observed which alleles were inherited together on the same chromosome — the **phase** is ambiguous), **haplotype inference (phasing)** methods are needed to statistically reconstruct the most likely haplotypes underlying observed multi-locus genotype data.

- **Clark's algorithm** is a parsimony-based method: it first resolves all unambiguous individuals (homozygous at all loci, or heterozygous at only one locus, whose haplotypes are directly determined), then iteratively resolves remaining ambiguous individuals by assuming, wherever possible, that their haplotypes are ones already observed/ inferred elsewhere in the sample (minimising the total number of distinct haplotypes invoked) — a computationally simple but somewhat ad hoc, order-dependent approach.
- **The EM (Expectation–Maximization) algorithm** treats haplotype phase as missing data and iterates between an **E-step** (computing, for each ambiguous individual, the

expected/posterior probability of each compatible haplotype pair, given current estimated haplotype frequencies) and an **M-step** (re-estimating haplotype frequencies from these expected counts, via maximum likelihood, Chapter 10), converging to the maximum likelihood estimates of population haplotype frequencies.

- **Bayesian methods (e.g., the PHASE algorithm)** extend the EM approach by incorporating a coalescent-based prior (Chapter 16) over haplotypes, reflecting the biological expectation that haplotypes cluster into a genealogical tree structure, generally improving phasing accuracy, especially for rarer haplotypes, relative to EM alone — at greater computational cost. Software implementing PHASE-type algorithms remains widely used for haplotype inference in African population-genetic and disease-association studies, where reference haplotype panels are historically less complete than for European-ancestry populations.

Theorem 15.2 (Bounds on the Linkage Disequilibrium Coefficient r-squared)

For two biallelic loci with allele frequencies p_A, p_B for the reference alleles, the squared correlation coefficient r^2 = D^2 / [p_A(1-p_A) p_B(1-p_B)] satisfies 0 <= r^2 <= 1, with r^2 = 1 if and only if the two loci are in complete association (only two of the four possible haplotypes occur, in exact proportion to the allele frequencies).

Proof. Since D^2 >= 0 always, the lower bound r^2 >= 0 is immediate. For the upper bound, note that r^2 is, by construction (Section 15.4), exactly the squared Pearson correlation coefficient between two binary indicator random variables — whether a randomly sampled haplotype carries the reference allele at locus A, and whether it carries the reference allele at locus B. The Cauchy-Schwarz inequality guarantees that any Pearson correlation coefficient lies in [-1,

reference allele at locus B. The Cauchy-Schwarz inequality guarantees that any Pearson correlation coefficient lies in [-1,

15.8: Chapter Summary

- Linkage disequilibrium is the non-random association between alleles at (usually linked) loci, measured by D, and standardised as D2 or r^2.
- LD decays geometrically with recombination fraction and generation number; African populations, including Nigerian populations, generally show faster LD decay due to larger, older effective population size.
- r^2 directly determines the sample-size penalty for indirect (marker-based) association mapping relative to direct typing of the causal variant.
- Haplotype phase, unobserved in standard genotype data, is inferred statistically via parsimony (Clark's algorithm), likelihood (EM algorithm), or Bayesian coalescent-based (PHASE) methods.

15.9: Computational Corner: Computing D, D-prime and r-squared

R.

```
# Linkage disequilibrium: D, D', and r^2 from haplotype
frequencies (Section 15.4)
# Two SNPs; haplotype counts among 200 chromosomes
hap_counts <-c(AB =90, Ab =30, aB =20, ab =60)
n <-sum(hap_counts)
pA <- (hap_counts["AB"] + hap_counts["Ab"]) / n
pB <- (hap_counts["AB"] + hap_counts["aB"]) / n
pAB <- hap_counts["AB"] / n

D <- pAB - pA * pB
Dmax <-if (D >=0) min(pA*(1-pB), (1-pA)*pB)
else min(pA*pB, (1-pA)*(1-pB))
```

```
Dprime <- D / Dmax
r2 <- D^2/ (pA * (1- pA) * pB * (1- pB))
cat("D =", D, " D' =", Dprime, " r^2 =", r2, "\n")

# install.packages("genetics")
library(genetics)
g1 <-genotype(c("A/A","A/B","B/B")) # illustrative use of
genetics package for LD()
```

Python.

```
hap_counts = {"AB": 90, "Ab": 30, "aB": 20, "ab": 60}
n =sum(hap_counts.values())
pA = (hap_counts["AB"] + hap_counts["Ab"]) / n
pB = (hap_counts["AB"] + hap_counts["aB"]) / n
pAB = hap_counts["AB"] / n

D = pAB - pA * pB
Dmax =min(pA*(1-pB), (1-pA)*pB) if D >=0 else min(pA*pB,
(1-pA)*(1-pB))
Dprime = D / Dmax
r2 = D**2/ (pA * (1- pA) * pB * (1- pB))
print(D, Dprime, r2)
```

Both languages return the same D, D2 and r^2 from Worked Example 15.1's haplotype counts, illustrating that computing LD statistics is a small, self-contained calculation that scales, in real genome-wide analyses (Chapter 18), to millions of SNP pairs computed via dedicated tools such as PLINK rather than by hand.

Mini-Project

Mini-Project 15: Computing LD Across Multiple Marker Pairs

Extend the two-SNP LD calculation of Section 15.9 to a small multi-marker haplotype block and visualise the resulting LD structure.

Step 1. Simulate haplotype data for 6 SNPs on the same chromosome with realistic decaying LD by physical distance (Section 15.3).

Step 2. Compute pairwise D' and r^2 for all 15 SNP pairs.

Step 3. Construct and plot an LD heatmap (triangular matrix) of the r^2 values.

Step 4. Identify the pair of SNPs with the highest r^2 and interpret what this implies for tagging one SNP with the other.

Step 5. Write a short paragraph on how this LD structure would affect the number of independent tests in a candidate-gene study of this region (linking back to Chapter 12).

Deliverable: *Simulation code, an LD heatmap, and a short written interpretation connecting LD structure to multiple-testing burden.*

15.10: Practice Questions

Q1. Two loci have haplotype frequencies h_AB = 0.45, h_Ab = 0.05, h_aB = 0.05, h_ab = 0.45. Compute D, D2 , and r^2, and comment on the strength of LD.

Q2. Explain why LD decays faster (over the same physical distance) in a Nigerian population sample than in a European population sample, in terms of population history.

Q3. A candidate causal SNP cannot be directly genotyped, but a nearby marker SNP with $r^2 = 0.5$ to it can be. If detecting

the causal SNP's effect directly would require 1,000 samples, approximately how many samples are needed using the marker SNP instead?

Q4. Briefly describe, in your own words, the difference in approach between Clark's algorithm and the EM algorithm for haplotype inference.

Q5. Explain why a marker locus that is only weakly correlated (small r^2) with an unobserved disease susceptibility locus provides a poor proxy for association testing, even if the marker is physically very close to the causal variant.

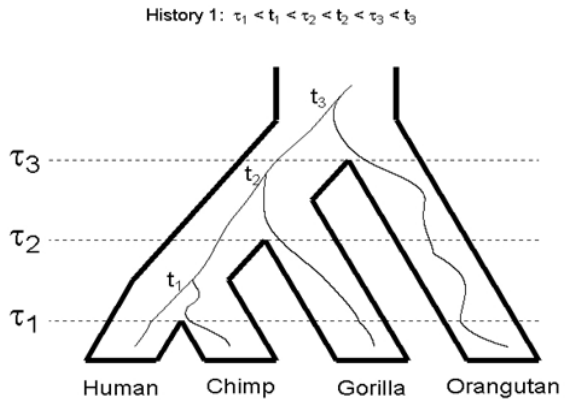
Q6. Using the R or Python code of Section 15.9, compute D, D2 , and r^2 for a different haplotype dataset: AB = 150, Ab = 10, aB = 10, ab = 150 (out of 320 total haplotypes). Comment on what the near-maximal D2 but the actual r^2 value together tell you about this marker pair.

16 Coalescent Models

16.1: Learning Objectives

By the end of this chapter, the reader should be able to describe the coalescent process for a sample of gene copies, compute expected coalescence times under the standard neutral (Kingman) coalescent, incorporate a simple mutation model, and relate coalescent theory to allele-frequency distributions and effective population size in a Nigerian population-genetics context.

16.2: The Genealogical (Backward-in-Time) Perspective



A gene genealogy (gene tree) for four species — Human, Chimp, Gorilla, Orangutan — nested within, and potentially discordant with, the species tree.

Classical population genetics (Chapters 3–5) works **forward in time**, tracking how allele frequencies change from one generation to the next. **Coalescent theory**, developed principally by J. F. C. Kingman (1982), instead works **backward in time**: starting from a sample of n gene copies (alleles) drawn from a present-day population, it models the random process by which pairs of lineages successively find their **most recent common ancestor (MRCA)** — a **coalescence event** — as one traces genealogies back through the generations, until eventually all n sampled lineages have coalesced into a single ancestral lineage.

This backward perspective is often far more computationally and statistically efficient than simulating an entire population forward in time, because it need only track the (much smaller) sample's genealogy, not the whole population — and it directly produces the genealogical tree structure that underlies patterns of genetic variation, linkage disequilibrium (Chapter 15), and disease-mapping (Chapter 17) methods.

16.3: The Simple (Wright–Fisher) Coalescent Model

Consider a population of constant effective size N (for simplicity, assume N diploid individuals, so 2N gene copies per generation, reproducing under the idealised **Wright–Fisher model** of random mating, non-overlapping generations, and no selection). Sample two gene copies (lineages) from the present generation. Looking one generation back, the probability that these two lineages **coalesce** (i.e., that both copies were, in fact, copies of the same parental gene copy in the previous generation) is:

$$P(\text{coalescence in 1 generation}) = 1/(2N)$$

since the second lineage's parent gene copy is equally likely to be any one of the 2N gene copies in the previous generation, one of which is the first lineage's own parent.

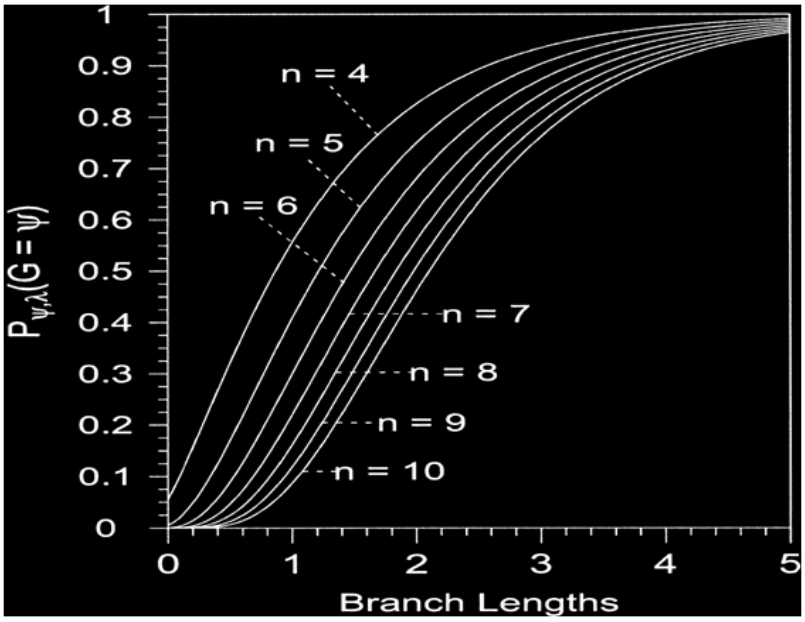
The probability that the two lineages have **not yet coalesced** after t generations, but do coalesce exactly at generation t, follows a geometric distribution:

$$P(T = t) = (1 - 1/2N)^{t-1} \times 1/(2N)$$

For large N, this geometric distribution is well approximated by an **exponential distribution** with rate 1/(2N), so that the expected time to coalescence of a pair of lineages is:

$$E(T) = 2N \text{generations}$$

16.4: Coalescence Times for a Sample of n Lineages



The cumulative probability of coalescence as a function of branch length, shown for increasing sample sizes n.

For a sample of n gene copies, coalescence proceeds as a sequence of pairwise merging events, reducing the number of distinct lineages from n to n−1, then to n−2, and so on down to 1. When there are k lineages remaining (out of the original

n), any one of the C(k,2) = k(k−1)/2 possible pairs may coalesce in a given generation, so the coalescence rate at that stage is C(k,2)/(2N), giving an expected waiting time (until the next coalescence event, reducing k lineages to k−1) of:

$$E(T_k) = 2N/C(k,2) = 4N/[k(k-1)]$$

The total expected time to the MRCA of the whole sample (the sum of the successive waiting times, from k = n down to k = 2) is:

$$E(T_{MRCA}) = \sum_{k=2}^{n} 4\,N/[k(k-1)] = 4N(1-1/n)$$

Notably, as n grows large, E(T_MRCA) converges to 4N generations — the majority of the expected time to the MRCA of a large sample is actually accounted for by the final coalescence of just the last two remaining lineages (since C(2,2)… rather C(2,2)=1 gives the largest, slowest expected waiting time, 2N, dominating the sum), a key qualitative insight of coalescent theory: adding more samples yields diminishing genealogical "depth," so modestly sized samples already capture most of a population's deep genealogical history.

Theorem 16.1: (Expected Time to the Most Recent Common Ancestor)

For a sample of n lineages under the standard coalescent with constant effective population size N, $E(T_{MRCA}) = \sum_{k=2}^{n} \frac{4N}{k(k-1)} = 4N\left(1-\frac{1}{n}\right)$.

Proof. Using the partial-fraction identity $\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$, the

sum telescopes:

$$\sum_{k=2}^{n} \frac{4N}{k(k-1)} = 4N \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k}\right)$$
$$= 4N \left[\left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \cdots \right.$$
$$\left. + \left(\frac{1}{\quad} - \frac{1}{\quad}\right)\right].$$

Every intermediate term cancels (each $-1/k$ is cancelled by the following term's $+1/k$), leaving only the first and last terms:

$$= 4N \left(\frac{1}{1} - \frac{1}{n}\right) = 4N \left(1 - \frac{1}{n}\right),$$

which is the stated result. As $n \to \infty$, $1/n \to 0$, so $E(T_{MRCA}) \to 4N$, confirming the "diminishing depth" property described above: the marginal contribution of each additional sampled lineage to the expected total depth shrinks rapidly, since $4N(1 - 1/n)$ is already within 10% of its limiting value $4N$ once $n = 10$. ■

Worked Example 16.2 (Estimating TMRCA for an HBB haplotype lineage in a Nigerian population sample)

A population geneticist samples n = 15 HBB-region haplotypes from a Nigerian cohort to study the genealogical history of a specific malaria-associated haplotype background, assuming an effective population size of N = 12,000 (a plausible historical effective size for a sub-Saharan African population, smaller than the Nigerian census population owing to long-term demographic history, Chapter 14). By Theorem 16.1, $E(T_{MRCA}) = 4(12{,}000)\left(1 - \frac{1}{15}\right) = 48{,}000 \times \frac{14}{15} = 44{,}800$ generations, or, at roughly 25 years per generation, approximately 1.12 million years —

illustrating that even a modest sample of 15 haplotypes is expected to trace back, on average, to a very ancient common ancestral haplotype, and that increasing the sample from 15 to, say, 30 would only push this estimate to $4(12{,}000)(1 - 1/30) = 46{,}400$ generations, a change of under 4%, concretely demonstrating the diminishing-returns property proved above.

16.5: Worked Example 16.1 (Coalescent time estimate, Nigerian population)

Suppose an autosomal locus's historical effective population size for a Nigerian ethnic group is estimated (from independent demographic/genetic data) at N ≈ 10,000. For a sample of n = 5 gene copies (from 5 unrelated individuals, i.e., 5 of their 10 total gene copies at this locus, chosen for simplicity as unlinked/independent draws), find the expected time (in generations) to the MRCA of the sample, and convert to approximate years (assuming 25 years per generation).

Solution:

$$E(T_{MRCA}) = 4N(1 - 1/5) = 40{,}000 \times 0.8$$
$$= 32{,}000 \text{generations}$$

In years: 32,000 × 25 = 800,000 years — illustrating the coalescent's characteristic result that, at neutral autosomal loci, gene genealogies typically trace back hundreds of thousands of years, far predating the emergence of modern population/ethnic group structure, and consistent with the deep, ancient genetic diversity long documented within African populations relative to non-African populations (whose out-of-Africa founding bottleneck, some 50,000–70,000 years ago, sharply reduced N and hence genealogical depth outside Africa).

16.6: Incorporating Mutation

Mutations are superimposed on the coalescent genealogy as a Poisson process along each branch, with rate μ per gene copy per generation. The total number of mutations expected on the genealogy of a sample of size n, summed over the whole tree, has expectation:

$$E(S) = \theta \times \sum_{k=1}^{n-1} 1/k, \theta = 4N\mu$$

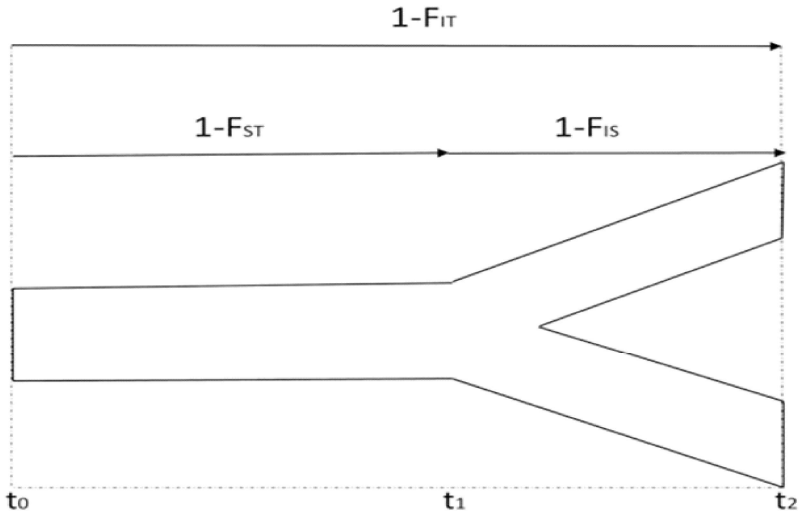
where θ (not to be confused with the recombination fraction θ of Chapter 7 — the same symbol is used by long convention for both quantities in population genetics, and context always disambiguates) is the **population-scaled mutation rate**, and S is the observed number of segregating (polymorphic) sites in a sample of DNA sequences. This relationship (due to Watterson, 1975) provides one of the standard estimators of θ (and hence, given an independently known mutation rate μ, of effective population size N) directly from observed DNA sequence variation — an approach widely used in estimating the effective population sizes and demographic histories of African, including Nigerian, populations from whole-genome sequence data.

16.7: The Allele Frequency Spectrum

Under the coalescent with mutation, the expected distribution of allele frequencies at neutral, segregating sites in a sample — the **site frequency spectrum (SFS)** — follows a characteristic shape: rare variants (those carried by only one or a few copies in the sample) are expected to be far more numerous than common (intermediate-frequency) variants, because most branches in a typical coalescent genealogy are short, terminal (external) branches on which only a few descendant lineages carry any resulting mutation. Departures from the neutral coalescent-

predicted SFS — for example, an excess of rare variants relative to neutral expectation — are used as one signature of recent population growth or of purifying selection, both of considerable interest in interpreting the substantial rare-variant burden increasingly documented in Nigerian and other African genomic sequencing studies (e.g., under initiatives such as H3Africa).

Variable population size and the bottleneck effect. All of the coalescent results above (Sections 16.3–16.7) assumed a *constant* effective population size, N, through time. Real human populations, however, have experienced dramatic size changes — most notably the "out-of-Africa" bottleneck, a sharp reduction in effective population size experienced by the ancestors of non-African populations roughly 50,000–70,000 years ago as a small founding group migrated out of the ancestral, larger African population:



A population bottleneck: effective population size drops sharply from an ancestral value to a much smaller founding size before recovering, illustrated as a coalescent genealogy whose branches are compressed during the bottleneck interval (t_0 to t_1) and expand again afterward.

A bottleneck compresses coalescence times during the period of small population size (Section 16.4's E(T) = 2N formula shows that smaller N mechanically produces faster coalescence), leaving a lasting signature in the genealogy: a burst of coalescent events concentrated in a narrow time window, followed by a period of much slower coalescence both before (in the ancestral population) and after (once the population re-expands) the bottleneck. This is precisely why genetic diversity is systematically lower in non-African populations than in African populations (Chapter 14, Section 14.2): the founding bottleneck removed much of the genetic variation carried by the larger ancestral African population, a loss that many thousands of subsequent generations have not fully restored. Modern coalescent-based demographic inference methods (Chapter 20) work by fitting exactly this kind of piecewise-constant (or continuously varying) population-size history to observed genome-wide genetic variation data.

Theorem 16.2: (Expected Time to the Most Recent Common Ancestor)

Under the standard (Wright-Fisher) coalescent for a sample of n lineages, the expected total time (measured in units of N generations, per Section 16.3) to the most recent common ancestor of the sample is E[T_MRCA] = 2(1 - 1/n).

Proof. Under the coalescent, the waiting time T_k during which exactly k ancestral lineages are present (before the next coalescence event reduces k to k-1) is exponentially distributed with rate C(k,2) = k(k-1)/2, so E[T_k] = 2/[k(k-1)] = 2[1/(k-1) - 1/k] (Section 16.4). The total time to the most recent common ancestor is the sum of these waiting times as the number of lineages decreases from n down to 1: T_MRCA = T_n + T_{n-1} + ... + T_2. Taking expectations

and using linearity, E[T_MRCA] = sum_{k=2}^{n} E[T_k] = sum_{k=2}^{n} 2[1/(k-1) - 1/k]. This is a telescoping sum: writing out the terms, 2[(1/1 - 1/2) + (1/2 - 1/3) + ... + (1/(n-1) - 1/n)], all intermediate terms cancel, leaving E[T_MRCA] = 2[1 - 1/n]. As n grows large, E[T_MRCA] approaches 2, showing that most of the expected time to the common ancestor is contributed by the earliest coalescence events (when only 2 or 3 lineages remain), not by events near the tips of the genealogy where many lineages are present — a fact with direct implications for how sample size affects genealogical inference.

16.8 Chapter Summary

- The coalescent models gene genealogies backward in time; a pair of lineages coalesces with probability ≈1/(2N) per generation, giving expected pairwise coalescence time 2N generations.
- For a sample of n lineages, expected time to the MRCA is 4N(1−1/n) generations, converging to 4N as n grows — most genealogical depth is captured by even modest samples.
- Mutation is superimposed as a Poisson process on the genealogy; the population-scaled mutation parameter θ=4Nμ links observed sequence diversity to effective population size.
- The site frequency spectrum, predicted by the neutral coalescent, provides a basis for detecting population growth or selection, both relevant to interpreting Nigerian genomic diversity data.

16.9: Computational Corner: Simulating Coalescent Genealogies

R.

```
# Simulate coalescence times for a sample of n lineages
(Section 16.4)
set.seed(2024)
simulate_coalescent <-function(n, Ne) {
  times <-numeric(n -1)
  k <- n
  total_time <-0
for (i in1:(n -1)) {
    rate <-choose(k, 2) / Ne            # merge rate ~ Exp(C(k,2)/
Ne) in coalescent units
    t_k <-rexp(1, rate = rate)
    total_time <- total_time + t_k
    times[i] <- total_time
    k <- k -1
  }
  times
}
Ne <-10000
coal_times <-replicate(500, simulate_coalescent(n =10, Ne =
Ne))
mean_tmrca <-mean(coal_times[9, ])       # last (9th) event =
TMRCA for n=10
cat("Simulated mean TMRCA (generations):", mean_tmrca,
"\n")
cat("Theoretical E[TMRCA] = 4Ne(1-1/n):", 4*Ne*(1-1/10),
"\n")
```

Python.

```
import numpy as np

rng = np.random.default_rng(2024)
```

```
def simulate_coalescent(n, Ne):
   k = n
   total_time =0.0
for _ inrange(n -1):
      rate = (k * (k -1) /2) / Ne
       total_time += rng.exponential(1/ rate)
      k -=1
return total_time

Ne =10000
tmrcas = [simulate_coalescent(10, Ne) for _ inrange(500)]
print("Simulated mean TMRCA:", np.mean(tmrcas))
print("Theoretical E[TMRCA] = 4Ne(1-1/n):", 4*Ne*(1-1/
10))
```

The Monte Carlo average TMRCA from 500 simulated genealogies should lie close to the analytic expectation $4N_e(1 - 1/n)$ derived in Section 16.4, providing a direct simulation-based check of the coalescent theory and previewing the more elaborate sequence-level coalescent simulations of Chapter 20.

Mini-Project

Mini-Project 16: Simulating Coalescent Genealogies and the Site Frequency Spectrum

Simulate gene genealogies under the standard coalescent and examine how sample size affects the time to the most recent common ancestor.

Step 1: Simulate coalescent genealogies for samples of size n = 5, 10, 20, and 50 lineages (Section 16.9).

Step 2: For each sample size, record the total time to the most recent common ancestor (TMRCA) across at least 500 simulated genealogies.

Step 3: Plot the distribution of TMRCA for each sample size and compare the means to the theoretical expectation.

Step 4: Superimpose mutations on your simulated genealogies (Section 16.6) and tabulate the resulting site frequency spectrum.

Step 5: Write a short paragraph on how the site frequency spectrum's shape might change under a recent population expansion.

Deliverable: *Simulation code, a TMRCA comparison plot across sample sizes, and a site frequency spectrum plot with written interpretation.*

16.10: Practice Questions

Q1. For an effective population size N = 5,000, find the expected pairwise coalescence time (in generations and in years, using 25 years/generation).

Q2. For a sample of n = 10 gene copies from a population with N = 8,000, find the expected time to the MRCA of the sample (in generations).

Q3. Explain, in words, why the expected time to the MRCA converges to 4N generations as sample size n grows large, rather than continuing to grow indefinitely with n.

Q4. If $\theta = 4N\mu = 0.008$ for a Nigerian population sample at a given locus, and the sample size is n = 20, use Watterson's formula (Section 16.6) to find the expected number of segregating sites, E(S). (You may leave the harmonic sum as a decimal approximation; $\Sigma\_\{k=1\}^\{19\}\ 1/k \approx 3.548$.)

Q6. Explain, in your own words, why the "out-of-Africa" population bottleneck illustrated in Section 16.7 left non-African populations with systematically lower

genetic diversity than African populations, using the relationship between effective population size and expected coalescence time (Section 16.4).

17 Parametric and Non-Parametric Linkage Analysis

17.1: Learning Objectives

By the end of this chapter, the reader should be able to state the assumptions of parametric (LOD-score, model-based) linkage analysis, compute a pedigree likelihood and LOD score for a simple family, describe non-parametric (model-free) linkage methods based on allele sharing, and compare the two approaches in the context of Nigerian family studies of Mendelian versus complex disease.

17.2: From Two-Point Recombination to Full Pedigree Linkage Analysis

Chapter 8 introduced the recombination fraction, è, and its estimation from double-backcross data. Real human family (pedigree) data are rarely as simple as a double backcross: families vary in size and structure, not every individual's genotype or phase is directly observable, and the mode of inheritance of the disease itself (dominant, recessive, penetrance, phenocopy rate) must typically be specified or estimated. **Parametric linkage analysis** extends the LOD-score approach of Chapter 8 to general pedigrees, under an assumed genetic model for the disease.

17.3: Assumptions of Parametric Linkage Analysis

A typical parametric linkage analysis specifies, in advance (or estimates via segregation analysis, Chapter 7):

1. The **mode of inheritance** (e.g., autosomal dominant, autosomal recessive, X-linked).
2. The **disease allele frequency** in the population.
3. The **penetrance** for each genotype (probability of expressing the disease phenotype given the genotype) — ideally complete penetrance (=1) and no phenocopies (unaffected individuals never sharing the disease genotype), though reduced-penetrance models are also used.
4. A specific pedigree structure with observed marker and (partially observed) disease phenotype data.

17.4: The Pedigree Likelihood

For a pedigree of n individuals, let X_i denote the observed phenotype of the ith individual and G_i their (unobserved, or partially observed) genotype at the disease locus and/or marker locus jointly. The general pedigree likelihood, as a function of the recombination fraction θ between the disease locus and a genetic marker, sums (integrates) over all genotype configurations consistent with the observed data:

$$L(\theta) = \sum_{G1} \cdots \sum_{Gn} \left[\prod_{i} p\,en(X_i|G_i) \right] \times \left[\prod_{j \in founders} p\,rior(G_j) \right] \times \left[\prod_{m,k,l} t\,rans(G_m|G_k, G_l) \right]$$

where pen(X_i | G_i) is the penetrance function (probability of the observed phenotype given genotype), the product over founders (individuals with no parents in the pedigree) uses the population prior probability of each founder genotype (typically from Hardy–Weinberg proportions, Chapter 3), and the product over parent–offspring transmissions (m = offspring, k, l = parents) applies Mendelian transmission probabilities, which depend on θ when the disease locus and marker locus are jointly tracked. This general formula, though notationally dense, is simply the systematic bookkeeping of Mendelian segregation probabilities (Chapter 2) and penetrance across an entire family tree; in practice it is evaluated by efficient recursive algorithms (the **Elston–Stewart algorithm**, well suited to pedigrees without loops, and the **Lander–Green algorithm**, well suited to many markers analysed jointly) rather than by brute-force summation.

17.5: Two-Point LOD Score Calculation: Procedure

For a simple pedigree in which recombinant and non-recombinant meioses can be directly counted (as in Chapter 8's double backcross, or its human-pedigree analogue), the LOD score at a trial value of θ is:

$$Z(\theta) = log_{10}[\theta^{R}(1-\theta)^{N_R}/0.5^{R+N_R}]$$

where R is the number of recombinant meioses and N_R the number of non-recombinant meioses observed across the pedigree(s).

Worked Example 17.1.

A Nigerian family linkage study of an autosomal dominant disorder finds, across several informative pedigrees, R = 1 recombinant and N_R = 7 non-recombinant meioses (n=8 total).

(a) Compute $Z(\theta)$ at $\theta = 0.1$. $\theta^{R(1-\theta)}\{N\_R\} = (0.1)^{1(0.9)}7 = 0.1 \times 0.4783 = 0.04783$. $0.5\text{^}8 = 0.003906$. $Z(0.1) = \log_{10}(0.04783/0.003906) = \log_{10}(12.246) = 1.088$.

(b) Find $\hat{\theta}$ (the MLE of θ) and $Z(\hat{\theta})$. By Chapter 8's binomial MLE result, $\hat{\theta} = R/(R+N\_R) = 1/8 = 0.125$. $\theta^{1(1-\theta)}7 = (0.125)(0.875)\text{^}7 = 0.125 \times 0.3927 = 0.04909$. $Z(0.125) = \log_{10}(0.04909/0.003906) = \log_{10}(12.569) = 1.099$.

(c) Interpret. The maximum LOD score across pedigrees, $Z(\hat{\theta}) \approx 1.10$, falls well short of the $Z \geq 3$ threshold conventionally required to declare significant linkage (Chapter 8), so with only 8 informative meioses this single-family (or small-multi-family) study provides, at best, weak/suggestive evidence and would need to be extended — either by recruiting additional Nigerian families segregating the disorder, or by combining LOD scores additively across independently ascertained pedigrees at the same fixed θ (a key practical advantage of parametric LOD scores: unlike many non-parametric statistics, Section 17.7, LOD scores at a fixed θ are simply additive across independent families, greatly simplifying multi-family/multi-centre linkage studies).

17.6: LOD Score Interpretation Thresholds (Recap and Extension)

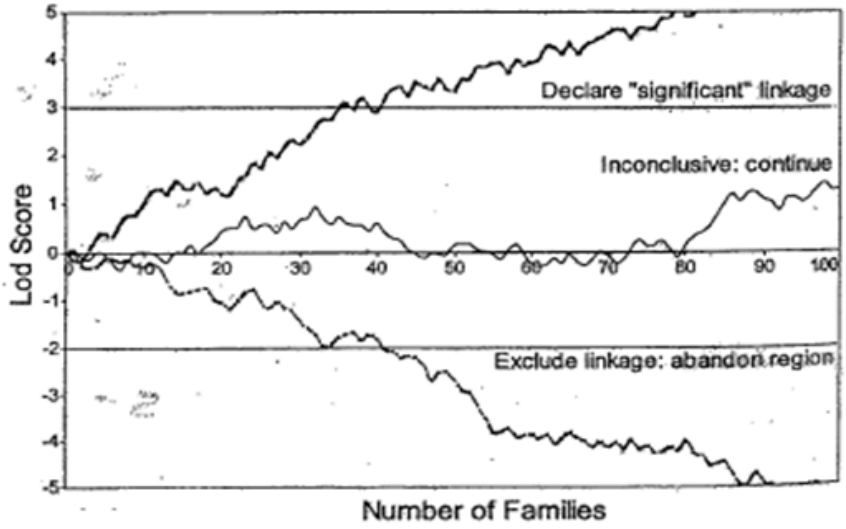


Figure 7.19 Hypothetical realizations of the accumulation of lod scores over time, assuming true linkage (upper curve), a gene of some given magnitude not linked to the marker (bottom curve), or no gene (middle curve). The conventional stopping rules of Morton (1955) are shown as horizontal dotted lines.

LOD score curves accumulating evidence for linkage as more families are added to a genome scan.

As introduced in Chapter 8, $Z \geq 3$ is the conventional threshold for "significant" evidence of linkage in a genome-wide parametric linkage scan (chosen to control the genome-wide false-positive rate given the very large number of loci effectively tested across the genome, an early precursor of the even more stringent genome-wide significance thresholds used in GWAS, Chapter 18); $Z \geq 2$ is often termed "suggestive"; $Z \leq -2$ excludes linkage at that θ. Maximising $Z(\theta)$ over θ (rather than testing at a single fixed θ) also constitutes a formal likelihood ratio test (Chapter 10) of H_0: $\theta=0.5$ vs H_1: $\theta<0.5$, subject to the boundary caveat noted in Section 10.7.

Theorem 17.1 (Additivity of LOD Scores Across Independent Families)

If $Z_1(\theta), \ldots, Z_K(\theta)$ *are the LOD scores computed from K independently ascertained families at the same trial value θ, then the combined LOD score for the pooled data is* $Z_{combined}\ (\theta) = \sum_{k=1}^{K} Z_k\ (\theta)$.

Proof. By definition (Section 8.5), $Z_k(\theta) = \log_{10}[L_k(\theta)/L_k(0.5)]$, where $L_k(\theta)$ is family k's likelihood at recombination fraction θ. Because the K families are ascertained and genotyped independently, their joint likelihood is the product of the individual likelihoods (a basic consequence of independence, Chapter 10): $L_{combined}\ (\theta) = \prod_{k=1}^{K} L_k\ (\theta)$. Hence:

$$Z_{combined}\ (\theta) = \log_{10}\left[\frac{L_{combined}\ (\theta)}{L_{combined}\ (0.5)}\right]$$

$$= \log_{10}\left[\frac{\prod_k L_k\ (\theta)}{\prod_k L_k\ (0.5)}\right] = \log_{10}\prod_{k=1}^{K}\frac{L_k(\theta)}{L_k(0.5)}$$

$$= \sum_{k=1}^{K}\log_{10}\left[\frac{L_k(\theta)}{L_k(0.5)}\right] = \sum_{k=1}^{K} Z_k\ (\theta),$$

using the logarithm identity $\log(\prod_k x_k) = \sum_k \log(x_k)$. ■

This theorem is the rigorous justification for the "combining families" practice used throughout Chapters 8 and 17: because LOD scores are logarithms of likelihood ratios, and independent likelihoods multiply, LOD scores simply add — no re-analysis of the pooled raw data is even needed if each family's LOD curve has already been computed separately.

Worked Example 17.3 (Combining three Nigerian extended families for a Mendelian ataxia linkage study)

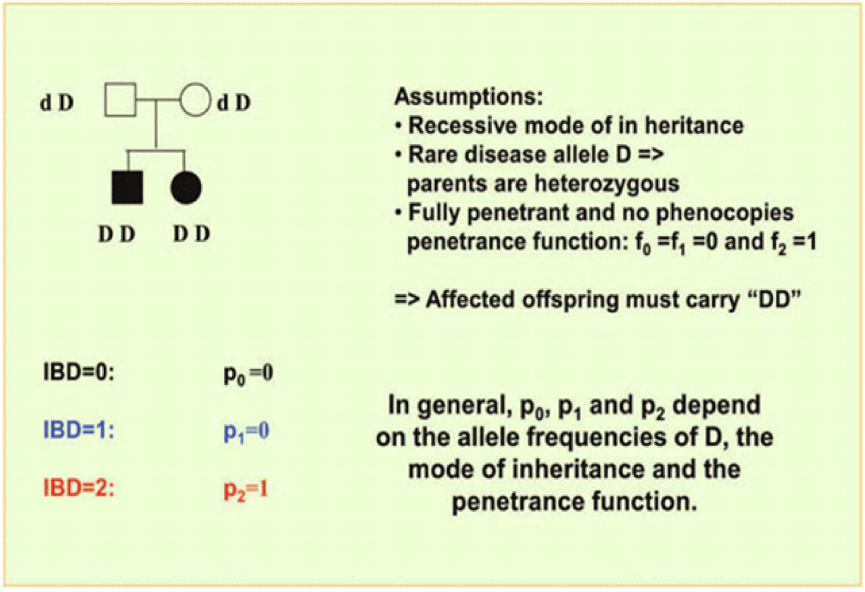
A neurogenetics collaboration pools linkage data on a rare autosomal dominant ataxia from three extended Nigerian families, each genotyped independently at a candidate microsatellite marker. At θ = 0.10, Family A gives $Z_A(0.10) = 1.8$, Family B gives $Z_B(0.10) = 0.9$, and Family C, unfortunately uninformative at this marker by chance, gives $Z_C(0.10) = -0.3$ (weak evidence against linkage at this specific θ, though not necessarily against linkage overall). By Theorem 17.1, the combined LOD score is $Z_{combined}\ (0.10) = 1.8 + 0.9 - 0.3 = 2.4$ — still short of the formal significance threshold of 3 (Section 17.6) but comfortably in the "suggestive" range, illustrating how even one weakly discordant family does not necessarily overturn a combined signal from several concordant families, and motivating the collaboration's plan to recruit a fourth family before drawing a final conclusion.

17.7: Non-Parametric (Model-Free) Linkage Analysis

Non-parametric linkage (NPL) analysis avoids specifying a genetic model (mode of inheritance, penetrance, disease allele frequency) altogether — an important practical advantage for **complex diseases** (e.g., type 2 diabetes, hypertension, or asthma, all of major public-health importance in Nigeria) for which the mode of inheritance is unknown, likely polygenic, and not well described by a simple Mendelian model. Instead, NPL methods test whether **affected relatives** (e.g., affected sibling pairs) share genetic material **identical by descent (IBD)**, Chapter 6, at a candidate marker **more often than expected by chance** under Mendelian segregation alone.

For a pair of full siblings, the number of alleles shared IBD at a given autosomal locus (0, 1, or 2) follows, under the null hypothesis of no linkage to any disease locus, the standard

Mendelian distribution: P(IBD=0) = 1/4, P(IBD=1) = 1/2, P(IBD=2) = 1/4 (a direct consequence of independent segregation at the two parental meioses producing each sibling — Chapter 2). If the marker is linked to a disease locus, **affected** sibling pairs are expected to share IBD status **more** than this null expectation (because they are more likely to have jointly inherited the same disease-predisposing chromosomal region from a parent), and the mean IBD sharing among affected sib pairs is compared statistically (e.g., via a one-sided test of mean IBD > 1, or via the widely used **S_pairs** and related NPL statistics implemented in software such as GENEHUNTER and MERLIN) to this null expectation.



An affected sibling pair pedigree with parental genotypes (4D × 4d), showing the observed vs. expected number of alleles shared identical by descent (IBD = 0, 1, or 2) under the null hypothesis of no linkage — the basic data structure for a non-parametric affected-sib-pair linkage test.

The maximum LOD score (MLS) statistic. The formal test summarising evidence for excess IBD sharing across many affected sibling pairs is the **maximum LOD score (MLS)**, constructed, in a spirit paralleling the parametric LOD score of Section 17.5, as a likelihood ratio comparing the IBD-sharing probabilities that best fit the observed data against their null (no-linkage) Mendelian values:

Maximum lod score (MLS)

Maximum likelihood estimates:

$$\hat{p}_0 = \frac{n_0}{n} \quad \hat{p}_1 = \frac{n_1}{n} \quad \hat{p}_2 = \frac{n_2}{n}$$

Likelihood ratio test:

$$\frac{(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)}{L(1/4,1/2,1/4)} = \frac{\hat{p}_2^{n_2}\,\hat{p}_1^{n_1}\,\hat{p}_0^{n_0}}{0.25^{n_2}\,0.5^{n_1}\,0.25^{n_0}}$$

$$2\log\left(\frac{(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)}{L(1/4,1/2,1/4)}\right) \sim \chi^2_2$$

Number of Alleles Shared IBD

	0	1	2	Total
Observed	n_0	n_1	n_2	n
Expected	n/4	n/2	n/4	n

The maximum LOD score (MLS) is defined by:

$$MLS = \log_{10}(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)/L(1/4,1/2,1/4))$$
$$= \log_{10}(L(\hat{p}_0,\hat{p}_1,\hat{p}_2)) - \log_{10}(L(1/4,1/2,1/4))$$

The maximum LOD score (MLS) framework: maximum likelihood estimates of the IBD-sharing probabilities z_0, z_1, z_2 are compared against their null values (1/4, 1/2, 1/4) via a likelihood ratio test, asymptotically distributed as a 50:50 mixture of a point mass at zero and a chi-square(1) distribution under the null.

Because the MLS statistic, like the ordinary LRT of Chapter 10, tests a null hypothesis ($z_0 = 1/4$, $z_1 = 1/2$, $z_2 = 1/4$) that lies at the *boundary* of the parameter space allowed under the alternative (excess sharing can only push z_2 upward and z_0 downward, not the reverse, under a genuine disease-linked locus), its null distribution is the same non-standard 50:50

mixture distribution encountered in Chapter 10's discussion of boundary-parameter likelihood ratio tests, rather than a simple χ^2 distribution.

Worked Example 17.2

In a Nigerian family study of hypertension (a complex trait), 100 affected sibling pairs are genotyped at a candidate marker; the observed distribution of IBD sharing is: IBD=0 in 15 pairs, IBD=1 in 45 pairs, IBD=2 in 40 pairs.

(a) Compute the mean IBD sharing observed, and compare with the null expectation.

Mean IBD = [0(15) + 1(45) + 2(40)] / 100 = (0+45+80)/100 = 125/100 = 1.25.

Null expected mean IBD (no linkage) = 0(0.25)+1(0.50)+2(0.25) = 0+0.50+0.50 = 1.00.

(b) Interpret: The observed mean IBD sharing (1.25) exceeds the null expectation (1.00), consistent with linkage between this marker and a locus influencing hypertension susceptibility in these Nigerian sibships; a formal significance test (e.g., comparing the observed IBD distribution to the null 1:2:1 distribution via a chi-square or a purpose-built NPL statistic, and typically also correcting for the marker's information content given imperfect parental genotyping) would be required before drawing a definitive linkage conclusion, but this simple mean-IBD comparison already illustrates the basic logic of non-parametric, allele-sharing–based linkage analysis.

17.8: Parametric versus Non-Parametric: When to Use Which

	Parametric (LOD score)	Non-parametric (allele sharing)
Requires specified genetic model	Yes	No
Best suited to	Simple Mendelian disorders (e.g., a clearly dominant/recessive Nigerian pedigree disorder)	Complex, polygenic traits (e.g., T2DM, hypertension)
Statistical power (if model correct)	Higher	Lower
Robustness to model misspecification	Lower (misspecified model can mask true linkage)	Higher
Scores additive across families?	Yes, at fixed θ	Generally no (families combined via other pooling methods)

Theorem 17.2: (The LOD Score Is a Monotonic Transformation of the Likelihood Ratio)

The LOD score Z(theta) = log_10 [L(theta) / L(0.5)] is a strictly increasing function of the likelihood ratio L(theta)/L(0.5), and consequently the value of theta that maximises the LOD score is identical to the value of theta that maximises the likelihood L(theta), namely the maximum likelihood estimator theta-hat.

Proof. The base-10 logarithm, log_10(x), is a strictly increasing function on (0, infinity), since its derivative 1/(x ln 10) is strictly positive for all x > 0. The LOD score is defined as Z(theta) = log_10[LR(theta)], where LR(theta) = L(theta)/L(0.5) is the likelihood ratio comparing linkage at recombination fraction theta to no linkage (theta = 0.5). Because Z is a strictly increasing transformation of LR(theta), and LR(theta) is itself a strictly increasing transformation of L(theta) for fixed L(0.5) (a positive constant not depending on theta), it follows that Z(theta) attains its maximum over theta at exactly the same value of theta at which L(theta) attains its maximum, namely theta-hat, the maximum likelihood estimator from Section 8.4. This confirms that the two-point LOD score procedure (Section 17.5) is not merely a convenient rescaling for interpretation, but is genuinely tied to a Neyman-Pearson-optimal likelihood ratio test at every fixed theta, inheriting the general asymptotic optimality properties of likelihood ratio-based inference established in Chapter 10.

17.9: Chapter Summary

- Parametric linkage analysis extends two-point LOD-score methods (Chapter 8) to full pedigrees via the pedigree likelihood, requiring a specified mode of inheritance, allele frequency, and penetrance.
- LOD scores at fixed θ are additive across independent pedigrees, easing multi-family studies; $Z \geq 3$ remains the conventional significance threshold.
- Non-parametric linkage analysis tests for excess allele sharing (identical by descent) among affected relatives, without assuming a genetic model — well suited to complex Nigerian diseases like hypertension and T2DM.

17.10: Computational Corner: Computing Two-Point LOD Scores

R.

```
# Two-point LOD score from meiosis counts (Section 17.5), across a grid of theta
r_recomb <-6; n_meioses <-20# 6 recombinants observed out of 20 informative meioses

lod_score <-function(theta, r, n) {
  (r *log10(theta) + (n - r) *log10(1- theta)) -
    (r *log10(0.5) + (n - r) *log10(0.5))
}
theta_grid <-seq(0.01, 0.49, by =0.01)
lods <-sapply(theta_grid, lod_score, r = r_recomb, n = n_meioses)
theta_hat <- theta_grid[which.max(lods)]
cat("theta_hat (grid) =", theta_hat, " max LOD =", max(lods), "\n")
plot(theta_grid, lods, type ="l", xlab =expression(theta), ylab ="LOD score")
abline(h =3, col ="red", lty =2)    # conventional linkage threshold
```

Python.

```
import numpy as np

r_recomb, n_meioses =6, 20

def lod_score(theta, r, n):
return (r*np.log10(theta) + (n-r)*np.log10(1-theta)) -\
        (r*np.log10(0.5) + (n-r)*np.log10(0.5))

theta_grid = np.arange(0.01, 0.5, 0.01)
```

```
lods = lod_score(theta_grid, r_recomb, n_meioses)
theta_hat = theta_grid[np.argmax(lods)]
print(theta_hat, lods.max())
```

Both scripts trace out the full LOD-score curve of Section 17.5 across candidate recombination fractions and mark the conventional genome-wide significance threshold of LOD ≥ 3, so the reader can see directly whether a given pedigree's data would, on their own, meet the classical Morton criterion for declaring significant linkage.

Mini-Project

Mini-Project 17: Two-Point Linkage Analysis Across a Chromosome Scan

Compute two-point LOD scores across a simulated chromosomal region and identify the region of strongest linkage evidence.

Step 1: Simulate pedigree data with a disease locus linked to one of 10 evenly spaced markers along a chromosome.

Step 2: Compute the two-point LOD score (Section 17.5) between the disease phenotype and each of the 10 markers.

Step 3. Plot LOD score against marker position to produce a simple linkage scan.

Step 4: Identify the marker(s) exceeding the conventional genome-wide significance threshold of 3.0 (Section 17.6).

Step 5: Write a short paragraph comparing what this parametric scan would show versus a non-parametric (model-free) analysis of the same data (Section 17.7).

Deliverable: *Simulation and LOD score computation code, a LOD-by-position plot, and a short written comparison of parametric and non-parametric approaches.*

17.11: Practice Questions

Q1. A Nigerian pedigree study of an X-linked recessive disorder finds R = 2 recombinants and N_R = 18 non-recombinants across informative meioses. Compute $\hat{\theta}$, and $Z(\hat{\theta})$.

Q2. Two independent Nigerian families studying the same autosomal dominant disorder give LOD scores of $Z_1(0.1) = 1.2$ and $Z_2(0.1) = 0.9$ at $\theta=0.1$. What is the combined (multi-family) LOD score at $\theta=0.1$, and what property of LOD scores allows this simple combination?

Q3. In a sib-pair study of asthma, 60 affected sibling pairs show IBD=0 in 10 pairs, IBD=1 in 20 pairs, IBD=2 in 30 pairs at a candidate marker. Compute the mean IBD and compare with the null expectation of 1.0.

Q4. Explain why non-parametric linkage analysis is generally preferred over parametric analysis for a trait like type 2 diabetes in a Nigerian family study, but parametric analysis remains preferred for a clearly Mendelian disorder such as sickle-cell disease.

Q5. Twenty affected sibling pairs are genotyped at a candidate marker, and the observed mean number of alleles shared IBD is 1.35 (compared with a null expectation of 1.0). Explain, conceptually, whether this observation is consistent with, or provides evidence against, linkage between the marker and a disease locus.

Q6. Explain why the null distribution of the maximum LOD score (MLS) statistic (Section 17.7) is a 50:50 mixture of a point mass at zero and a chi-square(1) distribution, rather than a simple chi-square(1) distribution as in an ordinary (non-boundary) likelihood ratio test.

18 Genome-Wide Association Studies (GWAS)

18.1: Learning Objectives

By the end of this chapter, the reader should be able to describe the design and analytical pipeline of a genome-wide association study, define genome-wide significance and explain the multiple-testing correction it embodies, interpret Manhattan and quantile–quantile (Q–Q) plots, describe the problem of population stratification and its correction, and discuss the specific opportunities and challenges of GWAS in Nigerian and African populations.

18.2: From Candidate-Gene to Genome-Wide Studies

The association tests of Chapter 11 were framed around a single candidate marker, chosen in advance on biological grounds. **Genome-wide association studies (GWAS)** instead test hundreds of thousands to millions of markers (SNPs), spread across the entire genome, for association with a trait or disease — an approach made possible by high-throughput SNP genotyping arrays (and, increasingly, whole-genome sequencing) and motivated by the recognition that most common-disease susceptibility loci could not have been confidently predicted in advance from biological knowledge alone.

The essential logic of GWAS rests on **linkage disequilibrium** (Chapter 11): because genotyped SNPs are in LD with many nearby, untyped variants, a relatively modest, well-

chosen panel of markers can indirectly "tag" and test the effect of most common genetic variation across the genome, without needing to genotype every possible variant directly.

18.3: The GWAS Analytical Pipeline

A typical GWAS proceeds through the following stages:

1. **Sample and phenotype collection**: cases and controls (or a quantitative-trait cohort), with rigorous phenotype definition — an area demanding particular care in Nigerian studies, e.g., standardising the definition and measurement of hypertension or T2DM across multiple recruiting hospital sites.
2. **Genotyping and quality control (QC)**: removing SNPs and samples with high missingness, SNPs deviating from Hardy–Weinberg equilibrium in controls (Chapter 3, often used as a QC filter, since strong HWE departure frequently signals genotyping artefact rather than true biology), samples with discordant sex or unexpectedly high relatedness/duplication, and SNPs with very low minor allele frequency (which provide little statistical power).
3. **Population structure assessment**: principal component analysis (PCA) of genome-wide genotypes to identify and adjust for ancestry differences between cases and controls (Chapter 11, Section 11.7) — an especially critical step in Nigeria, given substantial genetic differentiation among ethnolinguistic groups (e.g., Yoruba, Igbo, Hausa-Fulani, and numerous minority groups), documented by projects such as the International HapMap Project's Yoruba (YRI) panel and the H3Africa consortium.
4. **Imputation**: statistically inferring genotypes at untyped variants, using a reference haplotype panel and LD patterns (Chapter 11), to boost the effective density of tested variants — historically a particular challenge for Nigerian/African studies because widely used reference panels (e.g.,

early phases of 1000 Genomes, HapMap) under-represented African haplotype diversity, motivating panels built specifically from African genomes.

5. **Single-marker association testing**: fitting, at each SNP in turn, a regression model (logistic regression for binary disease traits, as in Chapter 11, Section 11.7; linear regression for quantitative traits) of phenotype on genotype (usually coded additively, 0/1/2 copies of one allele), adjusted for covariates including ancestry principal components.
6. **Genome-wide significance thresholding and replication** (Section 14.4–14.5).

18.4: Genome-Wide Significance: The Multiple-Testing Problem



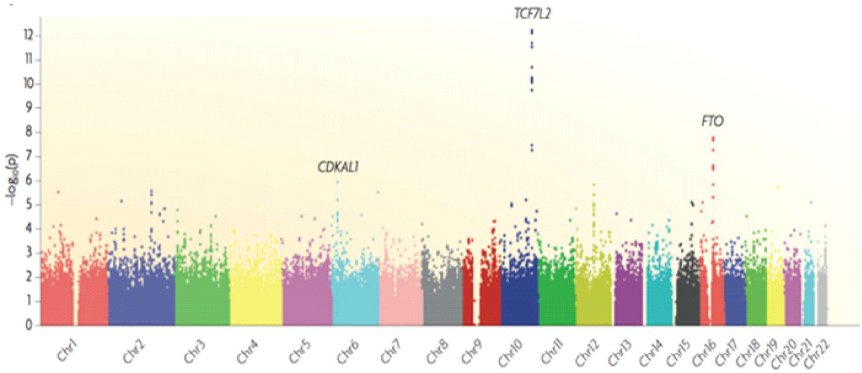
Chromosome ideogram showing the genomic locations of GWAS-discovered variants across the genome, coloured by associated trait category.

Testing hundreds of thousands to millions of SNPs simultaneously creates a severe multiple-testing problem: if each individual test used the conventional $\alpha = 0.05$ threshold, an enormous number of false-positive associations would be expected purely by chance. A simple, widely used (if somewhat conservative) **Bonferroni correction**, assuming approximately one million independent tests (a common approximation for the effective number of independent common-variant tests across the genome in populations of predominantly non-African ancestry — see Section 18.6 for the African-specific complication), gives the now-standard **genome-wide significance threshold**:

$$\alpha(genome - wide) = 0.05/1{,}000{,}000 = 5 \times 10^{-8}$$

A SNP association with $p < 5\times10^{-8}$ is conventionally reported as **genome-wide significant**; p between roughly 10^{-7} and 10^{-5} is often termed "suggestive," analogous to the suggestive/significant LOD-score distinction of Chapter 17.

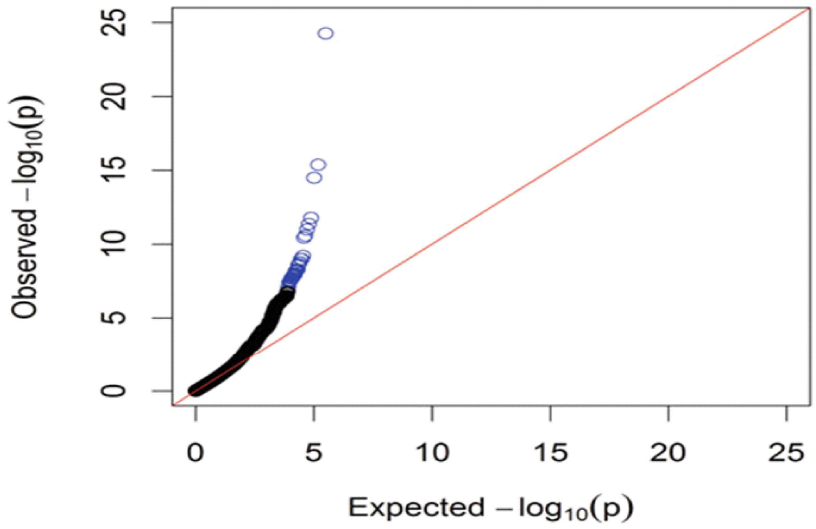
18.5: Visualising GWAS Results: Manhattan and Q–Q Plots



A Manhattan plot: "log € (p) plotted against chromosomal position, with SNPs coloured by chromosome, revealing genome-wide association signals as sharp towers.

A **Manhattan plot** displays $-\log_{10}$(p-value) for every tested SNP on the y-axis, against chromosomal position on the x-axis (SNPs ordered by chromosome and position), producing a characteristic skyline in which genuine association signals appear as sharp "towers" (clusters of highly significant SNPs, reflecting LD among neighbouring markers, Chapter 15) rising above the genome-wide significance line.

A **quantile–quantile (Q–Q) plot** compares the observed distribution of $-\log_{10}$(p-values) across all tested SNPs to the distribution expected under the null hypothesis of no association anywhere in the genome (which should follow a uniform distribution of raw p-values, hence a straight diagonal line on the Q–Q plot).



A genome-wide Q–Q plot: observed "log € (p) values (black dots) plotted against their expected values under the null (diagonal red line). The bulk of SNPs track the diagonal closely, indicating good control of confounding, while a small number of SNPs at the upper right depart sharply above the line — the true association signals.

A genome-wide Q–Q plot: observed $-\log_{10}(p)$ values (black dots) plotted against their expected values under the null (diagonal red line). The bulk of SNPs track the diagonal closely, indicating good control of confounding, while a small number of SNPs at the upper right depart sharply above the line — the true association signals.

Systematic deviation of the bulk of the Q–Q plot away from the diagonal (particularly, an inflation of small p-values across nearly all SNPs, not just a few clear outliers) is a standard diagnostic for uncorrected **population stratification** or other systematic confounding, and is commonly summarised by the **genomic inflation factor**, λ_GC (ideally close to 1.0; substantially greater than 1.0 signals inadequate stratification correction).

Theorem 18.1 (Genomic Control Correction)
If, under a global null hypothesis of no true association at m tested SNPs, uncorrected population stratification inflates every SNP's chi-square test statistic by a common multiplicative factor λ, i.e. $\chi^2_{i,observed} = \lambda \cdot \chi^2_{i,true-null}$ *for each SNP i, then dividing every observed statistic by* $\hat{\lambda} = \text{median}(\chi^2_{observed})/\text{median}(\chi^2_1)$ *restores the corrected statistics to their nominal* χ^2_1 *null distribution.*

Proof sketch. Under the stated inflation model, each $\chi^2_{i,true-null}$ genuinely follows a χ^2_1 distribution (since the vast majority of SNPs, under the global null, have no true association), whose theoretical median is $\text{median}(\chi^2_1) \approx 0.455$. Since $\chi^2_{i,observed} = \lambda\chi^2_{i,true-null}$ for every SNP, the *sample* median of the observed statistics across all m SNPs satisfies $\text{median}(\chi^2_{observed}) = \lambda \cdot \text{median}(\chi^2_{true-null}) \approx \lambda \times 0.455$ (medians scale linearly under a constant multiplicative transformation). Solving for λ gives the genomic control estimator

$observed$ $observed$

$\hat{\lambda} = \text{median}(\chi^2_{observed})/0.455 = \text{median}(\chi^2_{observed})/\text{median}(\chi^2_1)$. Dividing each observed statistic by this estimated $\hat{\lambda}$: $\chi^2_{i,observed}/\hat{\lambda} = \lambda\chi^2_{i,true-null}/\hat{\lambda} \approx \chi^2_{i,true-null}$ (since $\hat{\lambda} \to \lambda$ as $m \to \infty$ by the consistency of the sample median), recovering statistics that are, asymptotically, correctly distributed as χ^2_1 under the null. ■

The median (rather than the mean) is deliberately used in $\hat{\lambda}$ because the bulk of SNPs are null under any realistic genome-wide scan, so the median is robust to the small number of truly associated SNPs at the extreme tail (which would otherwise inflate a mean-based estimator and lead to over-correction of genuine signals) — a practical design choice that follows directly from the robustness properties of the median as a summary statistic.

Worked Example 18.2 (Genomic control correction for a multi-ethnic Nigerian GWAS)

A GWAS of type 2 diabetes recruits cases and controls from across Nigeria without fully matching for ethnic background, testing 500,000 SNPs. The median chi-square statistic across all SNPs is found to be 0.57 (compared with the theoretical null median of 0.455), suggesting stratification-driven inflation. By Theorem 18.1, $\hat{\lambda} = 0.57/0.455 \approx 1.25$. A SNP with an uncorrected chi-square statistic of 32 (nominal $p \approx 1.5 \times 10^{-8}$, just past genome-wide significance) is corrected to $32/1.25 = 25.6$ (corrected $p \approx 4.2 \times 10^{-7}$) — no longer meeting the genome-wide significance threshold of 5×10^{-8} once the estimated stratification inflation is accounted for, a sobering illustration of why Nigerian and other multi-ethnic African GWAS must budget carefully for ancestry principal components (Section 14.5) and genomic control correction before reporting any "discovery," since

failing to do so can convert a spurious, stratification-driven signal into an apparently genome-wide-significant (but false) finding.

18.6: Special Considerations for GWAS in Nigerian and African Populations

Several features of Nigerian and African population genetics have direct, practical consequences for GWAS design and interpretation, extending themes introduced in Chapters 15–16:

- **Greater genetic diversity and faster LD decay** (Section 15.6) mean that African populations require denser SNP panels (or high-quality imputation from African-inclusive reference panels) to achieve comparable genomic coverage to studies in European or East Asian populations — a naive application of a genotyping array or reference panel designed primarily around non-African LD patterns will under-capture true genetic variation and reduce statistical power in Nigerian samples.
- **Greater within-Africa population substructure**: Nigeria alone hosts substantial ethnolinguistic diversity, and even modest allele-frequency differentiation among groups (e.g., Yoruba, Igbo, Hausa-Fulani, and others) can generate confounding via population stratification (Chapter 3's Wahlund effect, in its GWAS-scale form) if not properly modelled with ancestry principal components or mixed models.
- **The effective number of independent tests** underlying the standard 5×10^{-8} threshold was originally calibrated largely on non-African (mainly European-ancestry) LD patterns; because African genomes carry more genetic variation and less redundant LD, some methodological work suggests that a modestly more stringent genome-wide threshold may, in principle, be appropriate for African-

ancestry GWAS — though 5×10^{-8} remains the most widely applied convention for comparability across studies.

- **The historical under-representation of African genomes** in reference panels, discovery cohorts, and downstream clinical-genetic risk tools (e.g., polygenic risk scores) means that genetic findings and prediction tools developed almost entirely in European-ancestry cohorts frequently transfer poorly to Nigerian populations — underscoring the scientific and public-health importance of dedicated African GWAS efforts, such as those coordinated through the H3Africa consortium, in which Nigerian institutions (including the University of Ibadan and University College Hospital, Ibadan) have played a substantial role, for conditions of major regional relevance including sickle-cell disease modifiers, hypertension, and infectious-disease susceptibility (e.g., malaria and Lassa fever).

18.7: Worked Example 18.1 (Interpreting a GWAS Result)

A GWAS of hypertension in a Nigerian cohort of 5,000 cases and 5,000 controls, testing 800,000 genome-wide SNPs after quality control, reports a lead SNP with $p = 3.2\times10^{-9}$ and odds ratio 1.22 (per copy of the risk allele), located within an intron of a gene previously implicated in renal sodium handling.

(a) Is this result genome-wide significant? Since $p = 3.2\times10^{-9} < 5\times10^{-8}$, yes, this SNP reaches conventional genome-wide significance.

(b) What additional steps would be expected before this finding is considered robust? Replication of the association in an independent Nigerian (or other African-ancestry) cohort; confirmation that the genomic inflation factor λ_GC

for the overall study is close to 1.0 (ruling out stratification as an alternative explanation, Section 18.5); and, ideally, fine-mapping (using the dense local LD structure, Chapter 11) and functional follow-up to distinguish the tested SNP from other, more strongly causal variants in LD with it.

(c) Comment on effect size. An odds ratio of 1.22 is a modest effect size, typical of common-variant GWAS findings for complex traits: it explains only a small fraction of the total heritable variation in hypertension risk (Chapter 9's heritability framework), consistent with the now widely accepted view that complex traits are influenced by very many variants of individually small effect (the **polygenic** picture, echoing Fisher's infinitesimal model introduced in Chapter 9), so that clinically useful genetic prediction for such traits in Nigerian populations will likely require **polygenic risk scores** aggregating very many loci — built, ideally, from Nigerian/African discovery data rather than transferred uncorrected from non-African studies (Section 14.6).

Theorem 18.2 (Expected Genomic Control Inflation Factor Under No Stratification)
If a genome-wide association study genuinely has no population stratification and no systematic genotyping artefacts, and the vast majority of tested markers have no true association with the phenotype, then the expected value of the genomic control inflation factor lambda (the ratio of the observed median chi-square test statistic to the expected median of a chi-square(1) distribution) is 1.

Proof. Under the null hypothesis of no association at a given marker, and in the absence of any confounding population structure, the association test statistic (e.g., from Section 11.3 or 11.7) is asymptotically distributed as chi-square with 1 degree of freedom, by the general asymptotic theory of likelihood-based

(or score-based) tests established in Chapter 10. If the overwhelming majority of the m tested markers are truly null (a reasonable assumption for most complex traits, where only a small number of markers have genuine effects), then the empirical distribution of the m observed test statistics is, to close approximation, a sample from the chi-square(1) distribution. The median of a chi-square(1) distribution is a fixed, known constant (approximately 0.4549). By the consistency of the sample median as an estimator of the population median (a standard result from order statistics theory), the observed median test statistic converges to this same constant as m grows large, so their ratio, lambda, converges to 1 in expectation. A value of lambda substantially greater than 1 in real data is therefore attributable to a violation of one of the theorem's assumptions — most commonly, uncorrected population stratification (Chapter 14) — which is precisely why lambda is used diagnostically in Section 18.9.

18.8: Chapter Summary

- GWAS test genome-wide panels of SNPs for association with disease/trait, exploiting linkage disequilibrium (Chapter 15) to indirectly capture most common genetic variation.
- The pipeline includes QC, population-structure adjustment (PCA), imputation, single-marker regression testing, and stringent multiple-testing correction (genome-wide significance, conventionally $p < 5\times10^{-8}$).
- Manhattan and Q–Q plots are the standard visual summaries; the genomic inflation factor diagnoses residual stratification.
- African, including Nigerian, populations present both distinctive challenges (denser marker/reference-panel requirements, greater substructure, historical under-

representation) and distinctive scientific opportunities (greater genetic diversity, finer LD-based fine-mapping resolution) for GWAS — motivating dedicated African genomics efforts such as H3Africa.

18.9: Computational Corner: Manhattan Plots, Q-Q Plots, and Genomic Control

R.

```
# Manhattan and Q-Q plots for simulated GWAS summary statistics (Section 18.5)
set.seed(42)
n_snp <-5000
chr <-sort(sample(1:22, n_snp, replace =TRUE))
pos <-ave(chr, chr, FUN =function(x) seq_along(x))
pvals <-runif(n_snp)
pvals[sample(n_snp, 5)] <-runif(5, 1e-9, 1e-7)     # plant a few "true hits"

# genomic control lambda
chisq_stat <-qchisq(1- pvals, df =1)
lambda_gc <-median(chisq_stat) /qchisq(0.5, df =1)
cat("Genomic control lambda:", lambda_gc, "\n")

plot(-log10(pvals), col = chr %%2+2, pch =20,
xlab ="SNP index (chromosome order)", ylab =expression(-log[10](p)))
abline(h =-log10(5e-8), col ="red", lty =2)      # genome-wide significance

qqplot(-log10(ppoints(n_snp)), -log10(sort(pvals)),
xlab ="Expected -log10(p)", ylab ="Observed -log10(p)")
abline(0, 1, col ="red")
```

Python.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2

rng = np.random.default_rng(42)
n_snp =5000
chrom = np.sort(rng.integers(1, 23, n_snp))
pvals = rng.uniform(size=n_snp)
pvals[rng.choice(n_snp, 5, replace=False)] =
rng.uniform(1e-9, 1e-7, 5)

chisq_stat = chi2.isf(pvals, df=1)
lambda_gc = np.median(chisq_stat) / chi2.ppf(0.5, df=1)
print("Genomic control lambda:", lambda_gc)

plt.scatter(range(n_snp), -np.log10(pvals), c=chrom %2,
cmap="coolwarm", s=5)
plt.axhline(-np.log10(5e-8), color="red", linestyle="—")
plt.xlabel("SNP index"); plt.ylabel("-log10(p)")

expected =-np.log10(np.linspace(1/n_snp, 1, n_snp))
observed =-np.log10(np.sort(pvals))
plt.figure()
plt.scatter(expected, observed, s=5)
plt.plot([0, max(expected)], [0, max(expected)], color="red")
```

Both scripts reproduce, on simulated data with a handful of planted "true" associations, the two standard GWAS diagnostic plots of Section 18.5 — the Manhattan plot, which visually locates genome-wide significant peaks against the 5×10^{-8} threshold, and the Q–Q plot together with the genomic control inflation factor λ_{GC}, which checks whether the bulk of the genome-wide test statistics follow their null distribution, a critical check for undetected population stratification (Chapter 18) before any peak is taken seriously.

Mini-Project

Mini-Project 18: A Complete Mini-GWAS from Simulation to Manhattan Plot

Run a complete, simplified GWAS pipeline end to end on simulated genome-wide data.

Step 1: Simulate genotype data for 5,000 SNPs across several chromosomes in 1,000 cases and 1,000 controls, with one true causal SNP.

Step 2: Test every SNP for association using logistic regression (Section 11.7).

Step 3: Produce a Manhattan plot (Section 18.5) of -log10(p) against genomic position.

Step 4: Produce a Q-Q plot and compute the genomic control inflation factor lambda (Section 18.9).

Step 5 Determine whether your true causal SNP surpasses genome-wide significance (Section 18.4) and discuss what sample size or effect size would be needed if it does not.

Deliverable: *Full pipeline code, a Manhattan plot, a Q-Q plot with lambda, and a short written interpretation.*

18.10: Practice Questions

Q1. A GWAS tests 600,000 SNPs. Using the standard Bonferroni approach with an assumed 1,000,000 effectively independent tests genome-wide (Section 18.4), what p-value threshold should be used for genome-wide significance, and would a SNP with p = 2×10^{-7} from this study meet that threshold?

Q2. A Q–Q plot from a Nigerian GWAS shows the bulk of observed p-values systematically more significant than expected under the null, across almost all SNPs genome-wide (not just a handful of top hits), with $\lambda_{GC} = 1.35$.

What does this most likely indicate, and what step in the analytical pipeline (Section 18.3) should be revisited?

Q3. Explain why a genotyping array and imputation reference panel developed and validated mainly in European-ancestry samples may perform less well in a Nigerian GWAS than in a European GWAS, referencing Chapter 15's discussion of LD.

Q4. A lead GWAS SNP for T2DM has odds ratio 1.15. A clinician asks whether this single SNP can be used, on its own, to usefully predict an individual Nigerian patient's T2DM risk. How would you respond, and what alternative approach would you suggest?

Q5. Sketch (in words) what a Q–Q plot would look like for a genome-wide scan that suffers from severe, uncorrected population stratification but contains no true associations at all, and contrast this with the Q–Q plot of a well-controlled study that does contain a handful of true, strong associations.

19 Genotype Imputation

19.1: Learning Objectives

By the end of this chapter, the reader should be able to explain what genotype imputation is and what problem it solves, describe the reference-panel-based imputation pipeline conceptually, discuss why reference panel composition matters especially for African populations, evaluate imputation quality, and implement a simplified imputation-and-accuracy-assessment exercise in R and Python.

19.2: What Is Genotype Imputation and What Problem Does It Solve?

Genotype imputation is a well-established statistical technique for estimating, at markers that were *not* directly genotyped in a given sample, the most probable genotype (or, more precisely, a probability distribution over possible genotypes) at each individual, using the observed genotypes at nearby directly-genotyped markers together with the known **haplotype structure** of a densely genotyped or sequenced **reference panel**. It exploits exactly the linkage disequilibrium structure developed in Chapter 15: because nearby SNPs are correlated, observing an individual's genotypes at a set of tag SNPs (Chapter 12, Section 12.3) allows the likely genotype at an untyped, LD-correlated SNP to be inferred with high accuracy, without the added expense of genotyping or sequencing it directly.

Imputation solves two distinct practical problems. First, it allows **cost-effective, in-silico genotyping**: an inexpensive genome-wide array genotyping perhaps 500,000–1,000,000 SNPs can, after imputation against a reference panel derived from whole-genome sequencing, yield inferred genotypes at tens of millions of variants, at a small fraction of the cost of directly sequencing every study participant. Second, it enables **meta-analysis and data harmonisation** across GWAS conducted on different genotyping platforms (Chapter 18): because different arrays genotype different, only partially overlapping sets of SNPs, imputing every study's data up to a common, dense reference panel produces a shared set of variants that can be jointly analysed or meta-analysed across studies, dramatically increasing statistical power (Section 19.3) beyond what any single study's raw genotyped markers could achieve alone.

19.3: How Genotype Imputation Works: The Conceptual Pipeline



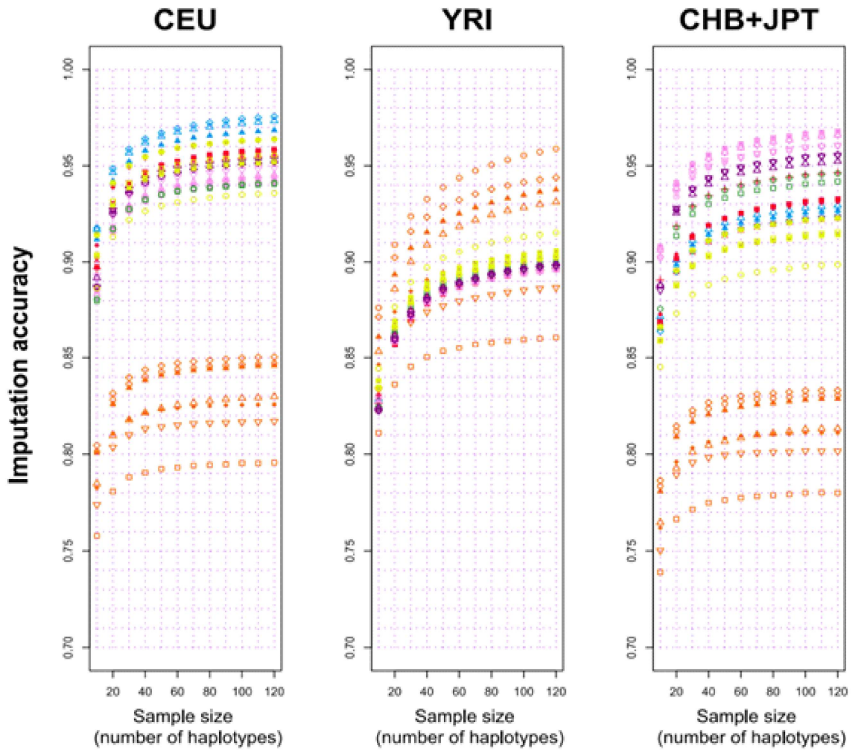
A reference panel of phased haplotypes: sequences at known positions used to impute genotypes at untyped markers by matching study haplotypes to the closest reference haplotypes.

Modern imputation methods (implemented in software such as IMPUTE, MaCH, Minimac, and Beagle) proceed, conceptually, in the following stages:

1. **Phasing.** The study sample's genotypes (typically unphased — i.e., for a heterozygous genotype, it is not directly observed which allele came from which parental chromosome) are first statistically **phased** into haplotypes, using a hidden Markov model that models each individual's haplotypes as an imperfect mosaic of reference-panel haplotypes.
2. **Reference matching.** For each short genomic window, the study individual's phased haplotype is compared against all reference-panel haplotypes, and the reference haplotypes most similar to it (over the flanking, directly-genotyped markers) are identified.
3. **Imputation.** At each untyped position within the window, the genotype is imputed as a probability-weighted average over the alleles carried by the best-matching reference haplotypes, yielding a posterior probability for each possible genotype (a **dosage**, typically reported as the expected number of copies of the reference or alternative allele, between 0 and 2, rather than a single hard-called genotype) together with an estimated imputation quality score (often denoted r^2 or INFO) reflecting the confidence of the imputed call at that position.

Association testing (Chapter 11, Chapter 18) is then usually carried out directly on the imputed dosages, rather than on hard-called genotypes, correctly propagating the imputation uncertainty into the association test's standard errors.

19.4: Reference Panels and the Special Importance of African Diversity



Imputation accuracy as a function of reference panel sample size, compared across YRI (Yoruba), CEU, and CHB+JPT reference panels.

The accuracy of genotype imputation depends critically on how well the **reference panel's haplotype diversity** matches that of the study sample. Early large reference panels (such as the original HapMap and the initial phases of the 1000 Genomes Project) were dominated by European, East Asian, and a limited set of African-American samples, and imputation accuracy for African study samples genotyped against such panels was measurably lower than for European samples — a direct consequence of the

substantially greater haplotype diversity and faster LD decay of African populations (Chapter 15, Section 15.6), meaning that a European-dominated reference panel simply does not contain enough closely-matching haplotypes to accurately tag the full range of African haplotypic variation, especially for lower-frequency variants.

This has been progressively addressed by the inclusion of more, and more diverse, African samples in reference panels — the expanded 1000 Genomes Project African panels, the African Genome Variation Project, and pan-African resources such as the H3Africa (Human Heredity and Health in Africa) initiative's reference data, which specifically includes West African populations relevant to Nigerian genetic studies. The general statistical lesson, of direct practical relevance to any Nigerian genetic epidemiology study, is that **imputation quality should always be checked, not assumed**, and that reference panels including populations genetically close to the study sample (ideally, ideally, Nigerian or West African panels where available) should be preferred over larger but more distantly related panels whenever a choice is available.

19.5: Assessing Imputation Quality

Two complementary approaches are used to assess how trustworthy imputed genotypes are, in addition to the per-variant INFO/r^2 score reported by imputation software itself:

- **Masked-genotype validation.** A subset of directly genotyped markers is deliberately masked (treated as if untyped) and then imputed using the remaining genotyped markers and the reference panel; the **concordance rate** between the imputed and the (known, masked) true genotypes directly measures empirical imputation accuracy for the study sample and reference panel in question, and is the standard internal quality-control step before relying on any imputed dataset.

- **Post-hoc filtering.** Imputed variants below a chosen INFO/ r^2 threshold (commonly 0.3 or 0.8, depending on the intended use) are typically excluded from downstream association analysis, since poorly imputed variants contribute mostly noise and can, at very low quality thresholds, introduce spurious associations rather than genuine signal.

Theorem 19.1: (Effective Sample Size Under Imputation)
If a variant is imputed with quality (INFO) score $r^2 \in (0,1]$*, and an association test at that variant is conducted on N genotyped individuals using imputed dosages, then the effective sample size for detecting a true association at that variant is* $N_{eff} = N \times r^2$ *— i.e., testing the imputed variant with N individuals provides approximately the same statistical power as directly genotyping the true causal variant in a sample of only* Nr^2 *individuals.*

Proof sketch. Model the imputed dosage $\tilde{G}_i$ for individual i as the true (unobserved) genotype G_i plus independent imputation noise: $\tilde{G}_i = G_i + \epsilon_i$, where $\mathrm{Var}(\epsilon_i)$ is calibrated so that $\mathrm{Corr}\left(\tilde{G}_i, G_i\right)^2 = r^2$ (this is, in fact, one standard operational definition of the INFO/r^2 score). In a simple linear association model, the variance of the estimated effect size (regression slope) using the noisy imputed dosage $\tilde{G}$ in place of the true genotype G is inflated by a factor $1/r^2$ relative to using G directly, a standard result from the statistical theory of regression with a mismeasured (noisy) covariate (classical "regression attenuation" / errors-in-variables theory). Since the variance of an effect-size estimator from a sample of size N scales as $1/N$, an inflation of variance by $1/r^2$ is exactly equivalent, in its effect on estimator precision (and hence on power), to reducing the sample size from N to Nr^2: $\mathrm{Var}\left(\hat{\beta}_{imputed}\right) \approx \frac{1}{Nr^2} = \frac{1}{N_{eff}}$ with $N_{eff} = Nr^2$. ■

This theorem gives a precise, quantitative meaning to the INFO/r² filtering thresholds of this section: an imputed variant with INFO = 0.3, tested in a sample of N = 10,000, carries only as much statistical information as a directly genotyped variant tested in $N_{eff} = 10{,}000 \times 0.3 = 3{,}000$ individuals — a substantial, and easily underappreciated, effective loss of sample size that directly justifies excluding such poorly imputed variants from primary analysis.

Worked Example 19.2 (Effective sample size loss in a Nigerian sickle-cell modifier-gene study)

A study of 6,000 Nigerian sickle-cell disease patients imputes a candidate modifier-gene variant (thought to influence disease severity) against a reference panel with only moderate representation of the patients' specific ethnic background, yielding INFO = 0.65 for this variant. By Theorem 19.1, the effective sample size for detecting a true association at this variant is $N_{eff} = 6{,}000 \times 0.65 = 3{,}900$ — meaning the study's true power to detect this modifier effect is comparable to what a directly genotyped study of only 3,900 patients would achieve, a nearly 35% effective reduction in sample size that the investigators must factor into their power calculations (Chapter 20) before concluding that a non-significant result rules out a clinically meaningful modifier effect.

Worked Example 19.1 (Imputation quality in practice).

A Nigerian GWAS imputes 8 million variants against a reference panel that includes H3Africa data. Post-imputation quality control reports: 5.2 million variants with INFO ≥ 0.8 ("high confidence"), 1.9 million with 0.3 ≤ INFO < 0.8 ("usable with caution"), and 0.9 million with INFO < 0.3 (typically excluded). Suppose the primary analysis retains only the first group. The analyst is effectively trading

completeness (using only 65% of imputed variants) for reliability (avoiding the roughly 35% of variants whose imputed genotypes are least trustworthy) — a standard, defensible default, though sensitivity analyses including the "usable with caution" tier at a relaxed threshold are common practice when a borderline signal falls just below the strict 0.8 cutoff.

Theorem 19.2 (Imputation R-Squared Equals the Squared Correlation Between True and Imputed Dosage)

Let g denote the true (but unobserved in the study sample) genotype dosage at a marker, and let g-hat denote its imputed dosage (the expected genotype under the imputation model, Section 19.3). Then the imputation quality metric R^2, defined as the ratio of the variance of the imputed dosage to the variance that would be expected under Hardy-Weinberg equilibrium at the estimated allele frequency, is algebraically equal to the squared Pearson correlation coefficient between g and g-hat, under the assumption that the imputation model is well-calibrated (E[g | g-hat] = g-hat).

Proof. By the law of total variance, Var(g) = Var(E[g|g-hat]) + E[Var(g|g-hat)] = Var(g-hat) + E[Var(g|g-hat)], using the well-calibration assumption E[g|g-hat] = g-hat. The squared correlation between g and g-hat is, by definition, Corr(g,g-hat)^2 = Cov(g,g-hat)^2 / [Var(g)Var(g-hat)]. Using the well-calibration assumption again, Cov(g, g-hat) = Cov(E[g|g-hat], g-hat) = Cov(g-hat, g-hat) = Var(g-hat) (a standard property of conditional expectation, since covarying g-hat with itself gives its own variance once g is replaced by its conditional expectation given g-hat). Substituting, Corr(g,g-hat)^2 = Var(g-hat)^2 / [Var(g) Var(g-hat)] = Var(g-hat)/Var(g). If Var(g) is taken to be the Hardy-Weinberg-expected variance 2p(1-p) at the estimated allele frequency p (the standard reference used in imputation software), this ratio Var(g-

hat)/Var(g) is exactly the imputation R^2 metric defined in Section 19.5, proving the claimed equivalence and justifying R^2's interpretation as a direct measure of how much of the true genotypic variance is captured by the imputed dosage.

19.6: Chapter Summary

- Genotype imputation statistically infers genotypes at untyped markers using LD with directly genotyped markers and a densely genotyped/sequenced reference panel, enabling cheap in-silico genotyping and harmonised meta-analysis across studies.
- The pipeline conceptually proceeds through phasing, reference-haplotype matching, and probability-weighted imputation, producing dosages together with a per-variant quality score.
- Imputation accuracy depends critically on how well the reference panel's haplotype diversity matches the study sample; historically under-represented African haplotype diversity in early reference panels reduced imputation accuracy for African samples, motivating panels such as H3Africa.
- Masked-genotype concordance and INFO/r^2 filtering are the standard tools for assessing and safeguarding imputation quality before downstream analysis.

19.7: Computational Corner: A Toy Imputation-and-Concordance Exercise

A full production imputation pipeline (phasing plus hidden-Markov-model reference matching) is beyond the scope of a single script, but the core statistical idea — using LD with a genotyped marker to predict an untyped one, and then checking accuracy by masking known genotypes — can be demonstrated directly.

R.

```
# Toy imputation: predict SNP2 (the "untyped" marker) from
SNP1 (the "tag" marker)
# using a small simulated reference panel, then mask and check
concordance in a study sample
set.seed(19)
n_ref <-500
# Reference panel: two SNPs in LD (r^2 ~ 0.7)
snp1_ref <-rbinom(n_ref, 2, 0.4)
snp2_ref <-rbinom(n_ref, 2, 0.4*0.85+ (snp1_ref -0.8) *0.15/
2)  # correlated construction (illustrative)
snp2_ref <-pmin(pmax(round(snp2_ref), 0), 2)

# Build a simple imputation rule: for each SNP1 genotype
class, use the modal SNP2 genotype
lookup <-tapply(snp2_ref, snp1_ref, function(x)
as.integer(names(sort(table(x), decreasing =TRUE))[1]))
print(lookup)

# Study sample: genotyped at SNP1 only; SNP2 truly known but
"masked" to test imputation
n_study <-300
snp1_study <-rbinom(n_study, 2, 0.4)
snp2_study_true <-rbinom(n_study, 2, 0.4*0.85+
(snp1_study -0.8) *0.15/2)
snp2_study_true <-pmin(pmax(round(snp2_study_true), 0),
2)

snp2_imputed <- lookup[as.character(snp1_study)]
concordance <-mean(snp2_imputed == snp2_study_true)
cat("Masked-genotype imputation concordance:",
concordance, "\n")
```

Python.

```
import numpy as np
from collections import Counter

rng = np.random.default_rng(19)
n_ref =500
snp1_ref = rng.binomial(2, 0.4, n_ref)
p2 = np.clip(0.4*0.85+ (snp1_ref -0.8)*0.15/2, 0, 1)
snp2_ref = np.clip(np.round(rng.binomial(2, p2)), 0, 2)

lookup = {}
for g in [0, 1, 2]:
   vals = snp2_ref[snp1_ref == g]
if len(vals) >0:
      lookup[g] = Counter(vals).most_common(1)[0][0]

n_study =300
snp1_study = rng.binomial(2, 0.4, n_study)
p2s = np.clip(0.4*0.85+ (snp1_study -0.8)*0.15/2, 0, 1)
snp2_study_true = np.clip(np.round(rng.binomial(2, p2s)), 0,
2)

snp2_imputed = np.array([lookup[g] for g in snp1_study])
concordance = np.mean(snp2_imputed == snp2_study_true)
print("Masked-genotype imputation concordance:",
concordance)
```

This toy "nearest-genotype-class" rule is a drastically simplified stand-in for a real hidden-Markov-model imputation engine, but it demonstrates the essential logic examined in Section 19.5: imputation accuracy is checked empirically by masking known genotypes and measuring concordance, exactly as is done in practice (with real reference panels such as the 1000 Genomes Project or H3Africa panels) before any imputed dataset is trusted for downstream GWAS.

Mini-Project

Mini-Project 19: A Toy Genotype Imputation and Accuracy Assessment

Simulate a reference panel and a study sample with sparser genotyping, perform simple imputation, and assess accuracy.

Step 1: Simulate a small reference panel of 200 haplotypes fully genotyped at 20 SNPs in strong local LD.

Step 2: Simulate a study sample genotyped at only 5 of those 20 SNPs (the rest 'missing').

Step 3: Implement a simple LD-based imputation rule (Section 19.3) to predict the missing genotypes in the study sample.

Step 4: Mask a set of SNPs you actually know the true genotype for, impute them, and compute concordance and imputation R^2 (Section 19.5).

Step 5: Write a short paragraph on how imputation accuracy might differ for a population under-represented in typical reference panels (Section 19.4).

Deliverable: *Simulation and imputation code, a concordance/ R^2 summary table, and a short written discussion of reference panel diversity.*

19.8: Practice Questions

Q1. Explain, in your own words, the difference between a "hard-called" imputed genotype and an imputed genotype "dosage," and why association testing on dosages is generally preferred.

Q2. Give two distinct reasons why imputation accuracy for a Nigerian study sample would historically have been lower using an early, largely European reference panel than using a panel that includes substantial West African representation.

Q3. A variant is imputed with INFO/r^2 = 0.25 in a large GWAS. Should this variant ordinarily be included in the primary association analysis? Justify your answer with reference to Section 19.5.

Q4. Using the R or Python code of Section 19.7, increase the LD between SNP1 and SNP2 (make the coefficient linking p2/snp2_ref to snp1_ref stronger) and re-run the simulation. Explain, with reference to Chapter 15, why the masked-genotype concordance should increase.

Q5. Explain briefly why genotype imputation is described as solving both a "cost" problem and a "meta-analysis harmonisation" problem, giving one sentence for each.

Q6. In Worked Example 19.1, what proportion of imputed variants would be included if the analyst instead used a relaxed threshold of INFO $\geq$ 0.3, and what is the main risk of using this more permissive threshold?

20 Genome Sequencing, Rare-Variant Methods, and Simulation-Based Approaches in Statistical Genetics

20.1: Learning Objectives

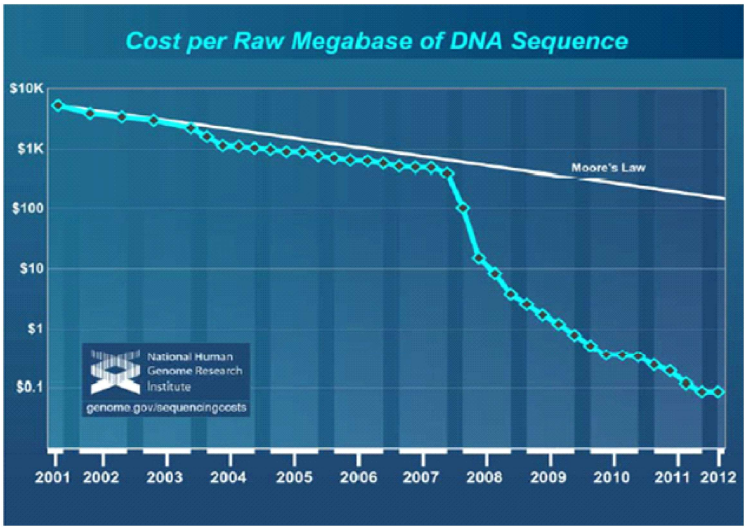
By the end of this chapter, the reader should be able to describe the basic principles of high-throughput (next-generation) sequencing and how it differs from array genotyping, explain the "missing heritability" problem and the common-disease/rare-variant hypothesis, describe burden and variance-component tests for rare-variant association, explain the coalescent and forward-in-time simulation paradigms and why simulation is indispensable in modern statistical genetics, and implement a coalescent simulation and a simulation-based power calculation in R and Python.

20.2: From Genotyping Arrays to High-Throughput Sequencing

All of the association methods developed so far in this book (Chapters 11, 12, 13, 18, and 19) work with genotypes at a pre-specified set of markers, obtained either from a **genotyping array** (which measures only a fixed panel of previously known, common SNPs) or from imputation against a reference panel. **High-throughput sequencing** (also called next-generation sequencing, NGS), by contrast, reads the DNA sequence directly, base by base,

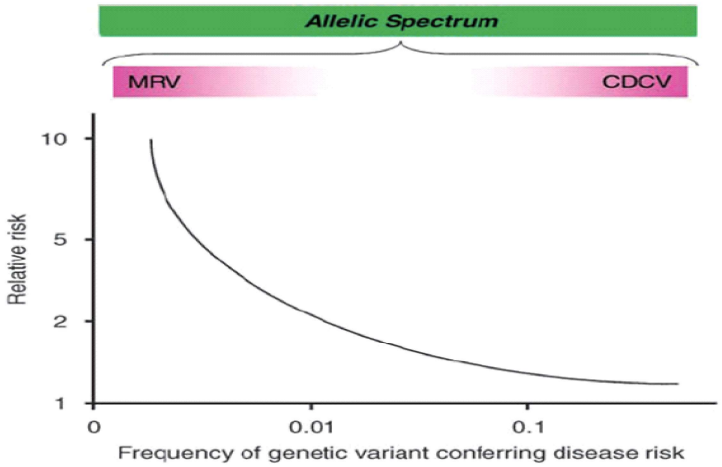
across some or all of an individual's genome, and can therefore detect **novel and rare variants** that were never included on any genotyping array and cannot be reliably imputed, because reliable imputation of a variant absent from every reference-panel haplotype is, by construction, impossible (Chapter 19).

Two sequencing strategies are in common use: **whole-genome sequencing (WGS)**, which reads the entire ~3.2 billion base-pair human genome (at some chosen average **read depth**, or coverage — the number of independent sequencing reads overlapping each position, typically 15–30× for population studies, higher for clinical diagnostic use), and **whole-exome sequencing (WES)**, which sequences only the protein-coding exons (roughly 1–2% of the genome) at much higher depth for a similar total sequencing cost, on the reasoning that a large share of clearly interpretable, high-penetrance disease variants lie in coding sequence. From raw sequencing reads, a standard bioinformatics pipeline aligns reads to a reference genome, identifies positions differing from the reference (**variant calling**), and outputs the results in the standard **Variant Call Format (VCF)**, the file format underlying essentially all modern sequencing-based genetic analysis.



The dramatic decline in the cost of DNA sequencing per raw megabase since 2001, which made population-scale sequencing studies feasible (NHGRI).

20.3: Missing Heritability and the Common-Disease/ Rare-Variant Hypothesis



The allelic spectrum: relative risk versus allele frequency, contrasting the common-disease/common-variant (CDCV) and common-disease/rare-variant (CDRV) hypotheses.

Chapter 18 (Section 18.4) noted that genome-wide-significant common variants identified by GWAS typically explain only a modest fraction of a complex trait's total heritability (Chapter 9) — a gap often called the **missing heritability** problem. The prevailing hypothesis underlying most GWAS design, the **common-disease/common-variant (CDCV)** hypothesis, holds that common disease risk is driven substantially by variants that are themselves common in the population (and therefore well captured by genotyping arrays and imputation). An alternative, complementary hypothesis, **common-disease/rare-variant (CDRV)**, holds that a meaningful share of disease risk and missing heritability is attributable to variants that are individually rare — conventionally, a **rare variant** is defined as one with minor allele frequency below some threshold, commonly 1% or 5%, a cutoff that is somewhat arbitrary but reflects the practical point at which genotyping arrays and standard imputation lose reliable power (Chapter 19, Section 19.4). Rare variants of large effect are difficult to detect using the single-marker association tests of Chapter 11, both because each individual rare variant is carried by very few people in any realistic sample (giving very low power for a single-marker test) and because, being rare, such variants are poorly tagged by nearby common markers and hence poorly imputed — motivating direct sequencing as the appropriate detection tool.

20.4: Rare-Variant Association Tests: Burden and Variance-Component Approaches

Because no single rare variant typically has enough carriers for an adequately powered single-marker test, rare-variant association methods instead **aggregate** information across all rare variants observed within a defined genomic unit (usually a gene). Two broad classes of aggregation test are in standard use:

- **Burden tests** collapse all qualifying rare variants within a gene into a single summary score per individual — most

simply, the total number of rare, qualifying (e.g., predicted-damaging) variants carried by that individual across the gene — and test this single burden score for association with the trait using an ordinary regression model (Chapter 11, Section 11.7). Burden tests are most powerful when most of the rare variants in the gene genuinely increase (or, symmetrically, all decrease) risk in the same direction.

- **Variance-component tests**, of which the **Sequence Kernel Association Test (SKAT)** is the most widely used, instead test whether the *variance* of a set of variant effects within the gene departs from zero, without requiring the variants to act in a consistent direction; this makes SKAT-type tests considerably more robust than a pure burden test when a gene harbours a mixture of risk-increasing and risk-decreasing (or some null) rare variants, at some cost in power when all truly causal variants genuinely act in the same direction. Combined tests (such as SKAT-O) interpolate between the pure burden and pure variance-component approaches, choosing the weighting that best fits the observed data, and are now the default choice in most modern rare-variant sequencing studies.

Theorem 20.1 (Burden Test Power Collapse Under Mixed-Direction Effects)

Consider a gene harbouring K rare variants with true effect sizes $\beta_1, \ldots, \beta_K$ *on a quantitative trait. The non-centrality parameter (and hence the power) of the burden test, which sums carrier status across all K variants into a single score, is proportional to* $(\sum_{k=1}^{K} \beta_k)^2$*, whereas the non-centrality parameter of the SKAT variance-component test is proportional to* $\sum_{k=1}^{K} \beta_k^2$*. Consequently, if the* β_k *are of mixed sign with* $\sum_k \beta_k = 0$*, the burden test's non-centrality parameter is exactly zero (no power, regardless of the individual effect magnitudes), while SKAT's non-centrality parameter remains strictly positive whenever any* $\beta_k \neq 0$.

Proof sketch. Under a standard linear aggregation model, the burden test statistic is (proportional to) a test of $H_0: \sum_k \beta_k = 0$ using the pooled burden score $B_i = \sum_k G_{ik}$ (total rare-allele count across the K variants for individual i); its power depends on the squared true value of the aggregated effect, $(\sum_k \beta_k)^2$, since the test is fundamentally testing this single linear combination. If the true effects are exactly offsetting (e.g., half the variants increase risk, half decrease it by the same average magnitude, so $\sum_k \beta_k = 0$), the quantity being tested is truly zero even though individual $\beta_k \neq 0$, so the burden test has no systematic signal to detect — it is, in effect, testing the wrong (accidentally null) linear combination. SKAT, by contrast, is constructed as a quadratic form test of $H_0: \beta_1 = \cdots = \beta_K = 0$ jointly, with a non-centrality parameter that accumulates $\sum_k \beta_k^2$ — a sum of squares, which remains strictly positive as long as at least one $\beta_k \neq 0$, regardless of the signs or any cancellation among the individual effects. ■

Worked Example 20.2 (Choosing burden vs. SKAT for a haemoglobinopathy modifier-gene panel in Nigeria)

A Nigerian sickle-cell research consortium sequences the *BCL11A* gene region (a well-established modifier of fetal haemoglobin, and hence of sickle-cell disease severity) in 2,000 patients, identifying 9 rare variants in the gene. Prior functional evidence suggests 6 of these variants increase fetal haemoglobin (protective, reducing disease severity) while 3 appear to decrease it (deleterious), with roughly comparable average effect magnitudes in each direction. By Theorem 20.1, a simple burden test collapsing all 9 variants into one score would have severely compromised power, since the protective and deleterious effects substantially cancel in the aggregated burden score ($\sum_k \beta_k$ close to zero even though several β_k are individually large) — the consortium should

A Nigerian sickle-cell research consortium sequences the *BCL11A* gene region (a well-established modifier of fetal haemoglobin, and hence of sickle-cell disease severity) in 2,000 patients, identifying 9 rare variants in the gene. Prior functional evidence suggests 6 of these variants increase fetal haemoglobin (protective, reducing disease severity) while 3 appear to decrease it (deleterious), with roughly comparable average effect magnitudes in each direction. By Theorem 20.1, a simple burden test collapsing all 9 variants into one score would have severely compromised power, since the protective and deleterious effects substantially cancel in the aggregated burden score ($\sum_k \beta_k$ close to zero even though several β_k are individually large) — the consortium should

20.5: Simulation as a Tool for Solving Genetics Problems

Throughout this book, closed-form analytic results have been available for many core quantities — Hardy–Weinberg proportions (Chapter 3), the recombination-fraction MLE (Chapter 8), the TDT statistic's asymptotic distribution (Chapter 13). For many realistic, complicated genetic-epidemiology questions, however, no closed-form solution exists, or the small-sample behaviour of a statistic departs materially from its asymptotic theory — and **simulation** becomes the indispensable computational tool for answering the question anyway. Genetic simulation serves at least three distinct purposes, each illustrated computationally in Section 20.7:

1. **Generating realistic genetic data for method development and teaching**, when real sequence data are unavailable, too small, or too sensitive to share, by simulating genealogies and mutations under a specified population-genetic model (the **coalescent**, first developed in Chapter 16, or its **forward-in-time** alternative, described below).

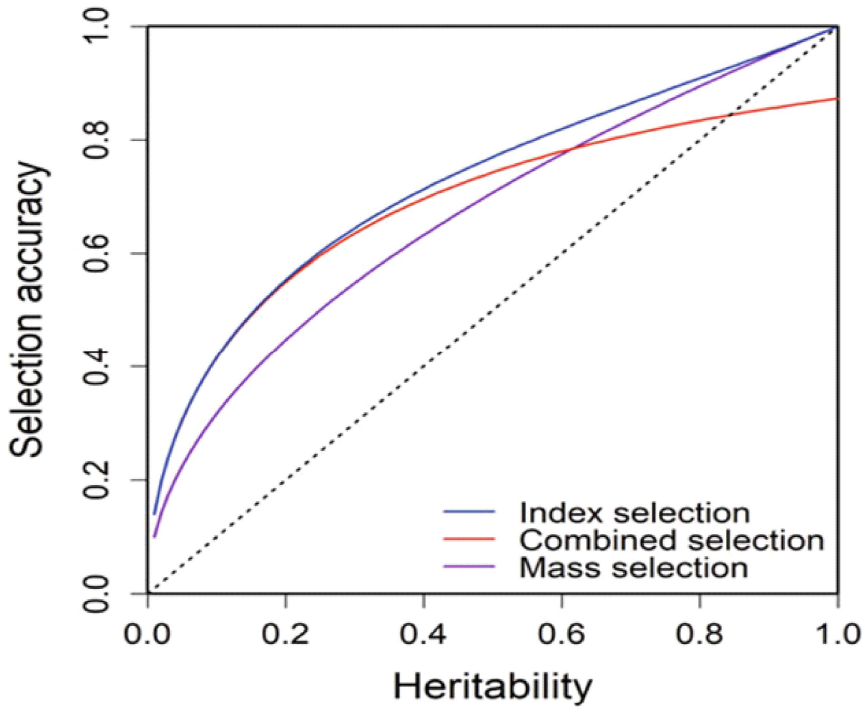
2. **Power and sample-size calculation**: simulating many replicate datasets under a specified true effect size and study design, and measuring the proportion of replicates in which the test of interest achieves statistical significance, gives an empirical power estimate for study designs (e.g., a rare-variant burden test, or a TDT study) too complex for a simple analytic power formula.
3. **Assessing a statistic's null distribution and calibration**, by simulating data under the null hypothesis many times and checking whether a test's p-values are uniformly distributed (i.e., that the test's nominal type-I error rate is actually achieved) — precisely the exercise suggested in Chapter 13's practice questions for the TDT, and a standard due-diligence step before trusting any new or complex statistical test.

20.6: Two Simulation Paradigms: Coalescent and Forward-in-Time

Chapter 16 introduced the **coalescent**, which simulates genealogies *backward in time* from a present-day sample to their most recent common ancestor(s), and is highly computationally efficient because it only ever generates the genealogical history of the sampled individuals, not the entire population. Modern coalescent simulators (msprime in Python, and the closely related scrm/ms lineage of programs, accessible from R via the coala package) extend the basic model of Chapter 16 to simulate full chromosome-length sequences, complete with recombination, variable population size through time, and multiple populations exchanging migrants — directly generating realistic genotype or haplotype data for exactly the kind of admixture and population-substructure exercises of Chapter 14, or the LD patterns of Chapter 15, under a fully specified and reproducible demographic history.

Forward-in-time simulation (implemented in software such as SLiM), by contrast, simulates the entire population, generation by generation, moving *forward* in time from a specified ancestral population, explicitly modelling selection, changing population size, and complex mating structures acting on every individual in the population at every generation. Forward simulation is far more computationally expensive than the coalescent for neutral (non-selected) scenarios, but is required whenever natural selection, complex non-neutral demographic scenarios, or explicit pedigree structures (extending Chapter 6) must be modelled with full realism.

A worked application: simulation in genomic prediction. Simulation is not only used to generate data for testing statistical methods (Section 20.5) — it is also used routinely to evaluate and compare entire prediction strategies before they are deployed on real data. In **genomic selection** (an application originating in animal and plant breeding but now also used for human polygenic risk scores, Chapter 18), a common question is: how accurately can an individual's genetic value for a quantitative trait (Chapter 9) be predicted from their genome-wide genotypes, and how does this accuracy depend on the trait's heritability and on which prediction strategy is used?



Simulated selection (prediction) accuracy as a function of trait heritability, compared across three genomic prediction strategies — index selection, combined selection, and mass selection — illustrating how simulation is used to evaluate competing statistical strategies before real-world deployment.

Such a plot is generated by simulating, for each heritability value, thousands of individuals with known true genetic values, applying each candidate prediction method to noisy phenotypic and genotypic data generated from a specified genetic architecture, and measuring the correlation between predicted and true genetic value — exactly the simulation-based evaluation logic of Section 20.5, applied here to compare prediction strategies (as opposed to comparing test statistics) before committing to one in an expensive real breeding or clinical programme.

20.7: Computational Corner: Simulating Sequence Data and a Rare-Variant Power Calculation

Python (coalescent simulation with msprime).

```
# pip install msprime —break-system-packages
import msprime
import numpy as np

# Simulate a 1 Mb chromosome segment for 200 diploid individuals (400 haplotypes)
# under a simple constant-size coalescent model (Ne = 10,000), with recombination and mutation
ts = msprime.sim_ancestry(
    samples=200, population_size=10_000,
sequence_length=1_000_000,
    recombination_rate=1e-8, random_seed=20)
mts = msprime.sim_mutations(ts, rate=1e-8,
random_seed=21)

geno = mts.genotype_matrix()              # variants x haplotypes
freqs = geno.mean(axis=1) /1# allele frequency per variant (haplotype-based)
n_rare = np.sum((freqs >0) & (freqs <0.01))
print(f"Simulated {mts.num_sites} variants; {n_rare} are rare (MAF < 1%)")
```

R (equivalent conceptual workflow via the coalescent theory of Chapter 16, using base R for a simplified multi-locus version, since scrm/coala require CRAN installation).

```
# Simplified multi-site coalescent-style simulation of a gene region with recombination
# (illustrative implementation of the same idea as msprime, for teaching purposes)
set.seed(20)
```

```
n_haps <-400; n_sites <-50; Ne <-10000; mu <-1e-8;
site_len_bp <-20000

simulate_site <-function(n_haps, Ne, mu, site_len_bp) {
# crude infinite-sites approximation: expected pairwise
diversity theta = 4*Ne*mu*site_len_bp
  theta <-4* Ne * mu * site_len_bp
  n_mut <-rpois(1, theta *sum(1/ (1:(n_haps -1))))
  freq <-rbeta(max(n_mut, 1), 1, n_haps -1)   # allele-
frequency-spectrum-like draw
  freq[1:max(n_mut, 0)]
}
all_freqs <-unlist(lapply(1:n_sites, simulate_site, n_haps =
n_haps, Ne = Ne, mu = mu, site_len_bp = site_len_bp))
n_rare <-sum(all_freqs <0.01)
cat("Simulated", length(all_freqs), "variants;", n_rare, "are
rare (MAF < 1%)\n")
```

Python (simulation-based power calculation for a gene-based burden test).

```
import numpy as np
from scipy import stats

rng = np.random.default_rng(99)

def simulate_burden_power(n_subjects=1000,
n_rare_variants=20, maf=0.005,
                          odds_ratio=2.0, n_reps=500, alpha=0.05):
   rejections =0
for _ inrange(n_reps):
       carrier_status = rng.binomial(1, 1- (1-
maf)**n_rare_variants, n_subjects)  # >=1 rare variant in
gene
       baseline_p =0.10
```

```
        p_disease = np.where(carrier_status ==1,
                             (baseline_p * odds_ratio) / (1- baseline_p +
baseline_p  * odds_ratio),
                            baseline_p)
        disease = rng.binomial(1, p_disease)
        table = np.array([[np.sum((carrier_status ==1) &
(disease ==1)),
             np.sum((carrier_status ==1) & (disease ==0))],
            [np.sum((carrier_status ==0) & (disease ==1)),
           np.sum((carrier_status ==0) & (disease ==0))]])
        _, p_value, _, _ = stats.chi2_contingency(table,
correction=False)
    if p_value < alpha:
            rejections +=1
    return rejections / n_reps

power = simulate_burden_power()
print(f"Estimated power: {power:.2f}")
```

Running the msprime script generates realistic coalescent-simulated haplotype data for a 1 Mb region and directly counts how many simulated variants fall below the conventional 1% rare-variant threshold of Section 20.3, illustrating in miniature why sequencing (rather than array genotyping) is required to observe the great majority of variants segregating in any real human population sample — most polymorphic sites in a sequenced sample are, in fact, rare. The burden-test power simulation demonstrates the second major use of simulation described in Section 20.5: by repeatedly simulating an entire gene-based rare-variant study under an assumed true odds ratio and re-running the same statistical test used in a real analysis, the proportion of simulated replicates reaching significance directly estimates the study design's statistical power — a calculation with no simple closed-form formula once realistic features (aggregated rare

variants, unequal carrier frequencies) are included, and precisely the kind of problem that simulation, and simulation alone, is equipped to solve.

Theorem 20.2 (Burden Tests Lose Power Under Bidirectional Rare-Variant Effects)

Consider a gene region with K independent rare variants, each with a true individual effect beta_k on a quantitative trait, and let the burden test statistic be based on the simple sum of an individual's minor allele counts across all K variants (Section 20.4). If the true effects beta_k are not all of the same sign, the non-centrality parameter of the burden test can be made arbitrarily small (even when individual variants have large true effects), whereas the non-centrality parameter of a variance-component test remains strictly positive.

Proof. The burden test's power is governed by its non-centrality parameter, which is proportional to (sum_{k=1}^{K} beta_k)^2 under standard regularity assumptions (the burden statistic estimates a single aggregated effect equal to the sum of the true per-variant effects weighted by allele frequency, and its test statistic's expectation under the alternative scales with this sum). If the beta_k take both positive and negative values, the terms in the sum sum_k beta_k can partially or fully cancel; in particular, for any large individual effects |beta_k|, one can choose signs such that sum_k beta_k = 0 exactly (e.g., pairing equal-magnitude effects of opposite sign), driving the burden test's non-centrality parameter to zero and its power to the nominal significance level, regardless of how large the individual |beta_k| are. In contrast, a variance-component test (such as SKAT) is constructed from a quadratic form in the beta_k (proportional to sum_k beta_k^2 in expectation), which is strictly positive whenever at least one beta_k is non-zero, irrespective of the signs of the individual effects. This proves

that no configuration of non-zero, opposite-signed effects can drive the variance-component test's non-centrality parameter to zero, formally establishing the burden test's vulnerability to effect-direction heterogeneity that the chapter's simulation-based discussion (Section 20.4) illustrates numerically.

20.8: Chapter Summary

- High-throughput sequencing reads DNA directly and can detect rare and novel variants invisible to genotyping arrays and beyond the reach of standard imputation; whole-genome and whole-exome sequencing are the two dominant strategies, with results reported in Variant Call Format (VCF).
- The missing-heritability problem motivates the common-disease/rare-variant (CDRV) hypothesis, complementary to the common-disease/common-variant hypothesis underlying standard GWAS.
- Because individual rare variants have very low power for single-marker tests, rare-variant association methods aggregate variants within a gene using burden tests (assuming a consistent direction of effect) or variance-component tests such as SKAT (robust to mixed-direction effects).
- Simulation — via the backward-in-time coalescent (Chapter 16) or forward-in-time, individual-based models — is indispensable wherever no closed-form analytic result exists: for generating realistic genetic data, for empirical power and sample-size calculation, and for checking a statistical test’s null calibration.
- msprime (Python) is the modern standard tool for coalescent simulation of full sequence data with recombination and demography; forward-in-time simulators such as SLiM are required when selection or complex non-neutral scenarios must be modelled explicitly.

Mini-Project

Mini-Project 20: Simulating Rare-Variant Data and Comparing Burden vs. Variance-Component Tests

Simulate rare-variant sequence data under two different genetic architectures and compare the power of burden and variance-component tests.

Step 1: Simulate a gene region with 20 rare variants in 1,000 cases and 1,000 controls, first under a scenario where all causal variants increase risk in the same direction.

Step 2: Apply a burden test (Section 20.4) and a variance-component test to this scenario and compare power.

Step 3: Repeat the simulation under a scenario where causal variants have effects in both directions (some protective, some risk-increasing).

Step 4: Apply both tests again and compare how their relative power changes between the two scenarios.

Step 5: Write a short paragraph explaining, in terms of the tests' underlying assumptions, why their relative performance reverses between the two scenarios.

Deliverable: *Simulation code, a power comparison table across both scenarios, and a short written explanation of the underlying reason for the reversal.*

20.9: Practice Questions

Q1. Explain, in two or three sentences, why a genotyping array cannot detect a truly novel variant that has never before been observed in any human, while whole-genome sequencing can.

Q2. A gene harbours 15 rare variants; laboratory evidence suggests 10 increase disease risk and 5 decrease it. Would

a simple burden test or a SKAT-type variance-component test be expected to have greater power in this scenario? Justify your answer with reference to Section 20.4.

Q3. Using the msprime code of Section 20.7, increase the simulated sample size from 200 to 2,000 diploid individuals and re-run the simulation. Does the number of variants classified as rare (MAF < 1%) increase, decrease, or stay roughly the same proportion of the total, and why?

Q4. Using the Python burden-test power simulation of Section 20.7, vary *odds_ratio* from 2.0 down to 1.2 and report how the estimated power changes. Explain the pattern you observe in terms of effect size and statistical power generally.

Q5. Explain briefly why the coalescent (backward-in-time) approach is generally far more computationally efficient than forward-in-time simulation for simulating neutral genetic variation in a sample from a single large population, and describe one genetic scenario for which forward-in-time simulation would nonetheless be necessary.

Q6. In the genomic-prediction simulation of Section 20.6, explain why prediction accuracy for all three selection strategies increases with the trait's heritability, using the variance-partitioning framework of Chapter 9.

Appendices

APPENDIX A
Selected Statistical Tables

Table A.1: Chi-square critical values (selected)

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09

Table A.2: Standard normal (z) critical values

Confidence level	One-sided z	Two-sided z
90%	1.282	1.645
95%	1.645	1.960
99%	2.326	2.576

Table A.3: Conventional LOD score interpretation (Chapters 8 and 17)

LOD score, Z	Interpretation
$Z \geq 3.0$	Significant evidence for linkage
$2.0 \leq Z < 3.0$	Suggestive evidence for linkage
$-2.0 < Z < 2.0$	Inconclusive
$Z \leq -2.0$	Significant evidence against linkage

Table A.4: Coefficient of relationship (r) for common relative pairs (Chapter 6)

Relationship	r
Parent–offspring	1/2
Full siblings	1/2
Half-siblings	1/4
Grandparent–grandchild	1/4
Uncle/aunt–nephew/niece	1/4
First cousins	1/8
Double first cousins	1/4

Table A.5: Conventional significance thresholds in genome-wide and rare-variant studies (Chapters 18 and 20)

Setting	Conventional threshold
Genome-wide significance (common-variant GWAS)	$p < 5 \times 10^{-8}$
"Suggestive" genome-wide association	$p < 1 \times 10^{-5}$
Exome-wide significance (gene-based rare-variant tests, ≈20,000 genes)	$p < 2.5 \times 10^{-6}$
Rare variant (minor allele frequency)	MAF < 0.01 (sometimes < 0.05)
Common variant (minor allele frequency)	MAF ≥ 0.05

Table A.6: Imputation quality (INFO/r²) rule-of-thumb thresholds (Chapter 19)

INFO / r² range	Typical use
≥ 0.8	High-confidence; suitable for primary analysis
0.3 – 0.8	Usable with caution; often retained in large meta-analyses
< 0.3	Generally excluded from primary association analysis

APPENDIX B
R and Python Computing Setup Guide

Every chapter's Computational Corner gives parallel R and Python code. This appendix lists the packages used throughout the book and how to install them, together with brief notes on the handful of specialised genetics packages referenced in the text.

B.1 Setting Up R

R (version 4.0 or later is recommended) can be downloaded from **https://cran.r-project.org**; RStudio Desktop is the recommended editing environment. Packages are installed with *install.packages("package_name")*, run once per machine.

Core packages used in this book:

Package	Used in	Purpose
stats (base R)	All chapters	Chi-square tests, glm, lm, optim, distributions
HardyWeinberg	Chapter 3	Hardy–Weinberg equilibrium testing and visualisation
kinship2	Chapter 6	Pedigree drawing and kinship-matrix computation
genetics	Chapters 15, 19	Genotype objects, linkage disequilibrium statistics

haplo.stats	Chapter 15	EM-based haplotype frequency estimation
qqman	Chapter 18	Manhattan and Q–Q plots for GWAS summary statistics
coala / scrm	Chapter 16, 20	Coalescent simulation (CRAN installation required; not pre-installed in all environments)
lme4 / nlme	Chapters 9, 14	Mixed-effects models for family and structured data

Install the core set with:
install.packages(c("HardyWeinberg", "kinship2", "genetics", "haplo.stats", "qqman", "lme4"))

B.2 Setting Up Python

Python (version 3.9 or later) is available from **https://www.python.org**, or via the Anaconda distribution, which bundles most scientific packages by default. Packages are installed with *pip install package_name* (add *—break-system-packages* on some Linux systems with an externally managed environment).

Core packages used in this book:

Package	Used in	Purpose
numpy	All chapters	Array computation
scipy	All chapters	Statistical tests, distributions, optimisation
pandas	Most chapters	Tabular data handling
statsmodels	Chapters 9, 11, 12	Regression models (OLS, logistic)
scikit-learn	Chapter 14	Principal component analysis

matplotlib	Chapters 4, 18, 20	Plotting
networkx	Chapter 6	Pedigree graph representation
msprime	Chapter 20	Coalescent simulation of sequence data
scikit-allel	Chapter 20 (further reading)	Population genetic analysis of variant call data

Install the core set with:
pip install numpy scipy pandas statsmodels scikit-learn matplotlib networkx msprime scikit-allel —break-system-packages

B.3 A Note on Reproducibility

Every simulation-based Computational Corner and Master Answer Key script in this book sets an explicit random seed (*set.seed()* in R, a *numpy.random.default_rng(seed)* generator, or *random.seed()* in Python) so that results are exactly reproducible on any machine running the same package versions. Readers repeating these exercises with a different seed should expect numerically different, but qualitatively similar, results — an instructive illustration, in itself, of Monte Carlo sampling variability (Chapter 20, Section 20.5).

B.4 Running the Book's Code

No special software beyond a working R or Python installation and the packages listed above is required to run any script in this book. Each Computational Corner's R and Python blocks are self-contained and can be copied directly into an R console/RStudio script, or a Python interpreter/Jupyter notebook, respectively, and run independently of any other chapter's code.

APPENDIX C
Glossary of Key Terms

Admixture: ancestry derived from two or more historically distinct source populations, carried within a single individual as a result of inter-population mating over recent generations (Chapter 14).

Allele: one of the alternative forms of a gene at a given locus.

Assortative mating: non-random mating based on phenotypic similarity (positive) or dissimilarity (negative) between mates.

Burden test: a rare-variant association test that collapses all qualifying rare variants within a gene into a single summary score per individual before testing for association (Chapter 20).

Candidate-gene study: an association study testing one or a small number of genes selected in advance for biological plausibility, prior linkage evidence, or previous literature reports (Chapter 12).

Coalescence: the event, traced backward in time, at which two gene-copy lineages are found to derive from a single common ancestral copy.

Cryptic relatedness: unrecorded close biological relatedness among individuals in a sample that is nominally

treated as consisting of unrelated individuals (Chapter 14).

Effective population size (N_e): the size of an idealised Wright–Fisher population that would experience the same rate of genetic drift as the actual population under study.

FBAT (Family-Based Association Test): a general statistical framework, of which the TDT is a special case, testing genetic association using within-family transmission information, robust to population stratification (Chapter 13).

Founder: an individual in a pedigree with no parents recorded, treated as unrelated to other founders unless otherwise specified (Chapter 6).

Genome-wide significance: the stringent p-value threshold (conventionally 5×10^{-8}) used in GWAS to correct for testing very large numbers of genetic markers simultaneously (Chapter 18).

Genomic control (λ_{GC}) : the ratio of the median observed test statistic to its expected null value across many genome-wide markers, used to detect and correct for systematic inflation, often due to population stratification (Chapter 14, Chapter 18).

Genotype imputation: statistical estimation of unobserved genotypes at untyped markers, using linkage disequilibrium with genotyped markers and a reference haplotype panel (Chapter 19).

Hardy–Weinberg equilibrium: the stable genotype frequencies (p^2, $2pq$, q^2) reached after one generation of random mating at a locus, absent mutation, migration, and selection.

Heritability (narrow-sense, h^2): the proportion of total phenotypic variance attributable to additive genetic variance.

High-throughput (next-generation) sequencing: technology that reads DNA sequence directly, base by base, enabling detection of novel and rare variants beyond the reach of genotyping arrays (Chapter 20).

Identical by descent (IBD): alleles that are copies of the same single ancestral allele, traceable through a known pedigree.

Inbreeding coefficient (F): the probability that an individual's two alleles at a locus are identical by descent.

Kinship coefficient (φ): the probability that a randomly chosen allele from one individual and a randomly chosen allele from another (or the same) individual, at the same locus, are identical by descent; ö = r/2 (Chapter 6).

Likelihood: the probability of the observed data, viewed as a function of unknown model parameters.

Linkage disequilibrium (LD): non-random statistical association between alleles at different (often linked) loci.

LOD score: logarithm (base 10) of the likelihood ratio comparing a trial recombination fraction θ to θ = 0.5 (no linkage).

Missing heritability — the gap between the heritability of a trait estimated from family/twin data and the (typically much smaller) heritability explained by genome-wide-significant variants identified by GWAS (Chapter 20).

Penetrance: the probability that an individual with a given genotype expresses the corresponding phenotype.

Pedigree: a diagram summarising the pattern of a trait's occurrence across the members of a family, over multiple generations, using standardised symbols (Chapter 6).

Population stratification: the presence of two or more genetically distinct subgroups within a sample, which can confound association tests if case/control ancestry proportions and tested-marker allele frequencies both differ across subgroups (Chapter 14).

Principal component analysis (PCA), in genetics: a statistical technique applied to genome-wide genotype data to extract the major axes of ancestry variation within a sample, used to detect and adjust for population stratification (Chapter 14).

Proband (propositus): the affected individual through whom a family is first identified for a genetic study.

Rare variant: conventionally, a genetic variant with minor allele frequency below approximately 1–5% in the population studied (Chapter 20).

Recombination fraction (θ): the probability that a gamete from a doubly heterozygous individual is recombinant at two loci.

Reference panel: a densely genotyped or sequenced set of haplotypes used as the basis for genotype imputation (Chapter 19).

Segregation analysis: the use of family data to test hypothesised modes of inheritance and estimate transmission parameters (Chapter 7).

SKAT (Sequence Kernel Association Test): a variance-component rare-variant association test, robust to a mixture of risk-increasing and risk-decreasing variants within a gene (Chapter 20).

Tag SNP: a SNP chosen to represent, via linkage disequilibrium, a set of other correlated SNPs, reducing the number of markers that must be directly genotyped (Chapter 12, Chapter 15).

TDT (Transmission Disequilibrium Test): a family-based association test comparing, among heterozygous parents of affected offspring, the number of times each allele is transmitted versus not transmitted (Chapter 13).

Variant Call Format (VCF): the standard file format for reporting genetic variants identified from sequencing data (Chapter 20).

Wahlund effect: the mechanical deficit of heterozygotes (relative to Hardy–Weinberg expectation computed from the pooled allele frequency) that results from combining two or more subpopulations with differing allele frequencies (Chapter 14).

APPENDIX D

Master Answer Key to End-of-Chapter Exercises

Full worked solutions to every end-of-chapter practice question are collected here, organised by chapter, so that each chapter's main text can flow directly from its Computational Corner into its exercises without interruption. Where a question is computational, a short verification script in both R and Python is given immediately after the worked solution; purely conceptual or explanatory questions are answered in prose alone.

Chapter 1: Basic Biological Concepts in Genetics

Answer to Q1. Genotype is the pair of alleles carried at the locus (e.g., AA, AS, or SS at HBB); phenotype is the observable outcome (e.g., normal red cells, sickle-cell trait, or sickle-cell disease) that results from the genotype.

Answer to Q2. (a) The baby's phenotype is sickle-cell disease (SCD). (b) The baby is homozygous, since both alleles at the locus are S.

Answer to Q3. (a) A male has only one X chromosome and therefore only one allele at an X-linked locus; he is hemizygous, so he is either affected or unaffected — there is no heterozygous "carrier" state possible for him at that locus. (b) Because G6PD is X-linked

recessive, a male needs only one copy of the deficient allele to be affected, whereas a female needs two copies (one from each parent); since the deficient allele is not extremely rare in many Nigerian populations, hemizygous males are affected far more often than homozygous-deficient females.

Answer to Q4. The mother contributes I^A or i (each with probability 1/2); the father contributes I^B or i (each with probability 1/2). Possible offspring genotypes: I^A I^B (phenotype AB), I^A i (phenotype A), I^B i (phenotype B), and ii (phenotype O), each with probability 1/4.

Answer to Q5. A genetic marker is a polymorphic DNA segment (or, classically, a polymorphic protein/antigen) of known chromosomal location, used as a reference point in mapping. Example: the ABO or Rhesus blood group systems, both widely typed in Nigerian blood banks and used historically as genetic markers before DNA-based markers became available.

Answer to Q6. A microsatellite with only three alleles distinguished purely by repeat length carries no information about which specific DNA sequence change is functionally responsible for a disease — it is simply a polymorphic "signpost," not a candidate functional variant itself, so genotyping it cannot directly reveal the causal mutation. However, because it is highly polymorphic (many distinguishable alleles, giving high heterozygosity and hence high statistical power to detect co-segregation with a trait in a pedigree), and because its chromosomal location is precisely known, it is an excellent tool for **linkage mapping** (Chapter 17): tracking whether a particular microsatellite allele is co-inherited with a disease

within families localises the unknown causal gene to the chromosomal region near the marker, even though the marker itself is never the cause.

Answer to Q7. For a cross of AB × OO: the AB parent contributes allele I^A or I^B (each with probability 1/2); the OO parent contributes only allele i (probability 1). Offspring genotypes: I^A i (phenotype A) with probability 1/2, and I^B i (phenotype B) with probability 1/2 — no AB and no O offspring are possible. Running the simulation code of Section 1.9 with mother = ["A","B"] and father = ["O","O"] should confirm approximately a 50:50 split between phenotypes A and B, with zero simulated AB or O offspring.

```
mother <-c("A", "B"); father <-c("O", "O")
offspring <-replicate(2000, c(sample(mother, 1),
sample(father, 1)))
table(apply(t(offspring), 1, function(g) if (all(g
%in%c("A","O"))) "A"elseif (all(g %in%c("B","O")))
"B"elseif (setequal(g, c("A","B"))) "AB"else"O"))
```

Chapter 2: Mendel's Laws of Inheritance

Answer to Q1. Let X = number of SS children among 4, X ~ Binomial(n = 4, p = 1/4). P(X = 1) = C(4,1) (1/4)^1 (3/4)^3 = 4 × 0.25 × 0.421875 = 0.4219, i.e., about 42.2%.

Answer to Q2. At HBB: father AS, mother AA ⟹P(child AS) = 1/2. At ABO: father I^B i, mother I^A i ⟹ offspring genotypes I^A I^B, I^A i, I^B i, i i each with probability 1/4; only ii gives blood group O, so P(group O) = 1/4. By independent assortment, P(AS and group O) = 1/2 × 1/4 = 1/8 = 0.125.

Answer to Q3. Expected counts under 3:1: E_dominant = 150, E_recessive = 50. $\chi^2 = (160-150)^2/150 + (40-50)^2/50 = 100/150 + 100/50 = 0.667 + 2.000 = 2.667$. Critical value (df = 1, $\alpha = 0.05$) = 3.84. Since 2.667 < 3.84, we fail to reject H_0 — the data are consistent with a 3:1 ratio.

Answer to Q4. At the molecular level, both haemoglobin A and haemoglobin S are produced and can be detected (e.g., by electrophoresis) in an AS individual — this is co-dominance. At the clinical level, however, an AS individual does not manifest the disease phenotype (sickling crises, chronic haemolysis) that only appears when both alleles are S — so with respect to the *disease* phenotype, S behaves as recessive to A.

R and Python verification (Q1, Q3).

```
dbinom(1, size =4, prob =1/4)                    # Q1: P(X = 1 SS
    child of 4)
    chisq.test(c(160, 40), p =c(0.75, 0.25))          # Q3:
    3:1 goodness-of-fit

fromscipy.statsimport binom, chisquare
    print(binom.pmf(1, 4, 1/4))              # Q1
    print(chisquare([160, 40], f_exp=[150, 50]))  # Q3
```

Chapter 3: Hardy–Weinberg Equilibrium

Answer to Q1. N = 500. $\hat{p}(B) = [\,2(350)+130] / [\,2(500)] = 830/1000 = 0.83$. $\hat{q}(b) = 1 - 0.83 = 0.17$.

Answer to Q2. E_BB = $500(0.83)^2 = 344.5$; E_Bb = $2(500)(0.83)(0.17) = 141.1$; E_bb = $500(0.17)^2 = 14.45$. $\chi^2 = (350-344.5)^2/344.5 + (130-141.1)^2/141.1 + (20-14.45)^2/14.45 = 0.088+0.874+2.132 = 3.09$. Critical value (1 df, α=0.05) = 3.84. Since 3.09 < 3.84, we do not reject H_0 — consistent with HWE (though the fit is closer to the boundary, illustrating that G6PD data can show mild departures, e.g. due to selection).

Answer to Q3. $P(O) = q^2$ (where q = freq(i)) = 0.36 $\Longrightarrow$ q = 0.6. Then freq(I^A) = p = 1 − 0.6 = 0.4.

Answer to Q4. This is the Wahlund effect: pooling two subpopulations with different allele frequencies (say p_1, p_2) into one sample yields an average homozygote frequency of $E[p_i^2] = \bar{p}^2 + Var(p)$, which always exceeds $\bar{p}^2$ (the homozygote frequency expected if the pooled population were a single randomly mating unit) whenever Var(p) > 0. In other words, hidden population structure inflates apparent homozygosity even without inbreeding, selection or assortative mating at the individual level.

R and Python verification (Q1, Q2).

```
n_BB<-350; n_Bb<-130; n_bb<-20; N<-500
p_hat <- (2*n_BB + n_Bb) / (2*N); q_hat <-1- p_hat
E <- N *c(p_hat^2, 2*p_hat*q_hat, q_hat^2)
chisq_stat <-sum((c(n_BB,n_Bb,n_bb) - E)^2/ E)
c(p_hat = p_hat, q_hat = q_hat, chisq = chisq_stat, p_value
=pchisq(chisq_stat, 1, lower.tail=FALSE))
```

```
import numpy as np
from scipy.stats import chi2
n_BB, n_Bb, n_bb =350, 130, 20; N =500
p_hat = (2*n_BB + n_Bb) / (2*N); q_hat =1- p_hat
E = N * np.array([p_hat**2, 2*p_hat*q_hat, q_hat**2])
chisq_stat = np.sum((np.array([n_BB,n_Bb,n_bb]) - E)**2/
E)
print(p_hat, q_hat, chisq_stat, chi2.sf(chisq_stat, 1))
```

Answer to Q5. The expected count of the rare homozygote (TT) is only $500 \times 0.005^2 \approx 0.0125$, far below the conventional rule-of-thumb minimum of 5 expected counts per cell required for the chi-square approximation to be reliable; with so small an expected count, the chi-square test's p-value would not be trustworthy. An **exact test** (Section 3.9), which computes the exact probability of the observed genotype configuration under HWE given the fixed allele counts, without relying on any large-sample approximation, should be used instead.

Answer to Q6.

```
library(HardyWeinberg)
    HWExact(c(AA =990, AB =10, BB =0), verbose
    =TRUE)
```

This scenario has an even smaller minor allele frequency ($\hat{q}$ = 10/2000 = 0.005, same as Worked Example 3.3a, but a larger sample), so the exact test remains the appropriate choice; the resulting p-value should be qualitatively similar in spirit to Worked Example 3.3a (no evidence against HWE, since a single heterozygote-only rare-allele pattern with zero rare homozygotes is entirely consistent with a rare allele under random mating), while a naively-applied chi-square test would again be unreliable here for exactly the same small-expected-count reason as in Q5.

Chapter 4: Estimation of Allele Frequency: Approach to Equilibrium for an X-Linked Gene

Answer to Q1. If p_m(0) = p_f(0), then d(0) = 0, so d(t) = 0 for all t: the two sexes start at, and remain at, the same allele frequency in every subsequent generation — the population is already at its X-linked equilibrium.

Answer to Q2. p_∞ = (1/3)(0.05) + (2/3)(0.25) = 0.01667 + 0.16667 = 0.1833. p_m(1) = p_f(0) = 0.25. p_f(1) = ½(0.25)+½(0.05) = 0.15. p_m(2) = p_f(1) = 0.15. p_f(2) = ½(0.15)+½(0.25) = 0.20.

Answer to Q3. d(t+1) = p_f(t+1) − p_m(t+1) = [½p_f(t)+½p_m(t)] − p_f(t) = −½p_f(t)+½p_m(t) = −½[p_f(t)−p_m(t)] = −½ d(t). This is a first-order geometric recurrence with ratio −½, so by induction d(t) = (−½)^t d(0) for all t ≥ 0.

Answer to Q4. We need (1/2)^t < 0.01 ⟹ t > log(0.01)/log(0.5) = (−2)/(−0.301) ≈ 6.64, so t = 7 generations (the smallest integer satisfying the inequality).

R and Python verification (Q2, Q4).

```
iterate_xlinked <-function(q_f0, q_m0, generations =8) {
  q_f <- q_f0; q_m <- q_m0
for (t in1:generations) { new_f <-0.5*(q_f+q_m); q_m <- q_f;
q_f <- new_f }
c(q_f = q_f, q_m = q_m)
}
iterate_xlinked(0.05, 0.25, 1)     # Q2: generation 1
iterate_xlinked(0.05, 0.25, 2)     # Q2: generation 2
log(0.01) /log(0.5)              # Q4: generations to |d(t)| < 0.01
import numpy as np
def iterate_xlinked(qf, qm, gens):
```

```
for _ inrange(gens):
    qf, qm =0.5*(qf+qm), qf
return qf, qm
print(iterate_xlinked(0.05, 0.25, 1))
print(iterate_xlinked(0.05, 0.25, 2))
print(np.log(0.01) / np.log(0.5))
```

Answer to Q5. The neutral recurrence of Section 4.3 only redistributes an existing allele frequency difference between the sexes toward a common weighted-average equilibrium — it cannot, on its own, *increase* the overall allele frequency in the population, since neither equation adds new copies of the allele; the long-run equilibrium $p_\infty = (1/3)p_m(0) + (2/3)p_f(0)$ is always a weighted average of (and therefore bounded between) the two starting frequencies. To explain the allele rising from a low starting frequency to 20%, an additional force — natural **selection** favouring carriers of the allele (here, most plausibly, a survival or fertility advantage for G6PD-deficient heterozygous females or hemizygous males in malaria-endemic environments, analogous to the sickle-cell heterozygote advantage of Chapter 3) — must be incorporated into the model, since only selection (or migration introducing additional copies, or mutation, both quantitatively minor over the relevant timescale) can change the population-wide allele frequency itself.

Answer to Q6.

```
iterate_xlinked(0.10, 0.60, 1); iterate_xlinked(0.10,
0.60, 5); iterate_xlinked(0.10, 0.60, 10)

for g in [1, 5, 10]:
print(g, iterate_xlinked(0.10, 0.60, g))
```

Starting from a large sex difference ($d(0) = 0.10 - 0.60$

= −0.50), the geometric decay d(t) = (−1/2)^t d(0) means the difference roughly halves each generation; by around generation 8–10 the two curves should be visually indistinguishable on a typical plot (the residual difference, (1/2)^10 × 0.50 ≈ 0.0005, being far smaller than the plot's visual resolution), consistent with the general rate result of Section 4.6.

Chapter 5: Non-Random Mating: Phenotypic Assortative Mating

Answer to Q1. Allele frequency is a property of the whole gene pool and is only changed by processes that add, remove, or differentially transmit alleles (mutation, migration, selection, drift) — assortative mating merely changes *who mates with whom*, not the total count of A and a alleles in the population, so p and q are unchanged. Heterozygosity falls because "like with like" mating preferentially creates AA×AA and aa×aa pairs (which can only produce homozygous offspring), reducing the proportion of Aa×Aa and other mixed pairings that generate heterozygotes, relative to random mating.

Answer to Q2. $V_A^*/V_A = 1/(1 - rh^2) = 1/(1 - (0.15)(0.40)) = 1/(1 - 0.06) = 1/0.94 = 1.064$, i.e., about a 6.4% inflation in additive genetic variance.

Answer to Q3. This pattern is more consistent with consanguinity (inbreeding). Because inbreeding reflects shared ancestry, it elevates homozygosity across the entire genome, not merely at loci tied to a specific assorted trait, and it specifically raises the incidence of rare autosomal recessive disorders (since related parents are more likely to both carry the same

rare recessive allele identical by descent) — exactly the pattern described.

Answer to Q4. With r = 0, V_A*/V_A = 1/(1−0) = 1, i.e., no inflation: additive genetic variance remains at its random-mating value. This makes sense because assortative mating only distorts the transmission of genetic variance when mate choice is correlated with the trait's genetic value; with zero spousal correlation, mating is effectively random with respect to the trait, so there is no assortative-mating effect to inflate the variance.

R and Python verification (Q2).

```
r <-0.15; h2 <-0.40
1/ (1- r*h2)
r, h2 =0.15, 0.40
print(1/ (1- r*h2))
```

Answer to Q5. This pattern suggests **phenotypic assortative mating on pigmentation** rather than inbreeding: the elevated homozygosity is confined to the trait-associated locus (and, presumably, loci in LD with it), while background genome-wide homozygosity — which inbreeding would elevate uniformly across the whole genome — remains at its Hardy–Weinberg-expected level, exactly the diagnostic distinction drawn in Section 5.7.

Answer to Q6.

```
result_strong <-simulate_assortative(p0 =0.5, alpha =0.8)
tail(result_strong, 1)
result_weak <-simulate_assortative(p0 =0.5, alpha =0.3)
tail(result_weak, 1)
result_strong = simulate_assortative(0.5, 0.8)
print(result_strong.tail(1))
result_weak = simulate_assortative(0.5, 0.3)
```

```
print(result_weak.tail(1))
```

Increasing alpha from 0.3 to 0.8 should produce a markedly lower equilibrium heterozygote (Aa) proportion after 10 generations, since stronger positive assortment more aggressively sorts individuals into like-genotype pairs, accelerating the loss of heterozygosity relative to the weaker-assortment case — while the allele frequency p remains approximately 0.5 in both simulations, exactly as Q1's theoretical result predicts.

Chapter 6: Pedigree Analysis: Structure, Notation, and Relatedness

Answer to Q1. For ordinary first cousins, r = 1/8 (Section 6.3). Double first cousins share two independent lines of cousin relationship (one through each pair of grandparents), so the contributions add: r = 1/8 + 1/8 = **1/4**.

```
r_single_cousin <-1/8
r_double_cousin <- r_single_cousin + r_single_cousin
r_double_cousin
r_single_cousin =1/8
r_double_cousin = r_single_cousin + r_single_cousin
print(r_double_cousin)
```

Answer to Q2. For an avuncular (uncle/aunt–niece/nephew) pair, r = 1/4 (Section 6.3), so F(child) = r/2 = 1/8 = **0.125**. This is exactly double the risk from a first-cousin union (F = 1/16, Worked Example 6.4), reflecting the closer relationship between uncle/aunt and niece/nephew (r = 1/4) than between first cousins (r = 1/8).

```
F_uncle_niece <- (1/4) /2
F_first_cousins <- (1/8) /2
c(F_uncle_niece = F_uncle_niece, F_first_cousins =
F_first_cousins, ratio = F_uncle_niece / F_first_cousins)
F_uncle_niece = (1/4) /2
F_first_cousins = (1/8) /2
```

```
print(F_uncle_niece, F_first_cousins, F_uncle_niece /
F_first_cousins)
```

Answer to Q3. Generation I: I-1 (square, filled — affected male, G6PD-deficient) mated to I-2 (circle, unfilled — unaffected, non-carrier or untested female). Generation II: two daughters, II-1 and II-2 (circles, unfilled but obligate carriers, since each inherits her father's only X chromosome), each married to an unrelated, unaffected male. Generation III: II-1's son, III-1, is drawn as a filled square (affected), having inherited the deficient X allele from his carrier mother; II-1's other children (if any) and II-2's children would be unaffected or carriers with probability 1/2 each, depending on sex. Note the classic X-linked recessive signature: the trait passes from an affected grandfather, through obligate-carrier daughters (never affected themselves), to affected grandsons — with no father-to-son transmission anywhere in the pedigree.

Answer to Q4. Running the R or Python code of Section 6.7 for individuals 6 and 7 (both children of the same parents, individuals 3 and 5) gives $\varphi(6,7) = 0.25$, exactly the full-sibling kinship coefficient, confirming $r = 2\varphi = 0.5$ — matching the path-counting result of Section 6.3 (two paths, through the shared mother and shared father, each contributing $(1/2)^2 = 1/4$, summing to 1/2).

Answer to Q5. The kinship coefficient between an individual and an unrelated founder is zero because, by definition, an unrelated founder has no ancestors in common with the individual anywhere in the pedigree, so there is no path along which an allele could be identical by descent between them. A non-inbred founder's kinship coefficient with themself is $\varphi = 1/2$ (not 1) because φ_{ii} is defined as the probability that

two alleles drawn *at random, one from each of two independent copies* of the individual's genotype (conceptually, from two of the individual's own gametes) are identical by descent; a non-inbred individual's own two alleles at a locus are, by definition, not IBD (they came from two different, unrelated ancestral lineages via the two parents), so this probability works out to 1/2, and the corresponding coefficient of relationship with oneself is $r = 2\varphi = 1$ as expected (a person shares 100% of their own genotype with themselves), while $F = 0$ confirms no inbreeding.

Answer to Q6. An offspring allele matching neither parental allele at a locus indicates a **Mendelian inheritance inconsistency** — biologically, this cannot arise from ordinary transmission (barring a new mutation at that exact site, which is very rare), and instead usually signals a genotyping error, a sample-labelling mix-up, or non-paternity/non-maternity. The modern quality-control procedure built on this logic is **Mendelian error checking**, a standard automated step in genotype quality control (Chapter 19) that scans every genotyped parent–offspring trio for such inconsistencies and flags or removes the implicated markers or individuals before any association or linkage analysis proceeds.

Answer to Q7. Half-first-cousins share exactly one grandparent (rather than two, as for ordinary first cousins), so there is only **one** path connecting them, of length 4 meioses (child → shared grandparent → other child → half-cousin, i.e., 4 links), contributing $(1/2)^4 = 1/16$. With only one such path (compared to the two paths for ordinary first cousins), $r = 1/16$ — half the

ordinary first-cousin value of 1/8, reflecting the halved degree of shared ancestry.

Chapter 7: Segregation Analysis and Genetic Model Fitting

Answer to Q1. $\Sigma(a_i-1) = (3-1)+(2-1)+(1-1) = 2+1+0 = 3$. $\Sigma(s_i-1) = (6-1)+(4-1)+(3-1) = 5+3+2 = 10$. $\hat{p} = 3/10 = \mathbf{0.30}$.

```
fam <-data.frame(s =c(6,4,3), a =c(3,2,1))
p_hat <-sum(fam$a -1) /sum(fam$s -1)
p_hat
```

```
s = [6, 4, 3]; a = [3, 2, 1]
p_hat =sum(ai-1for ai in a) /sum(si-1for si in s)
print(p_hat)
```

Answer to Q2. Every ascertained family is, by construction, guaranteed to contain at least one affected member (the proband); families with zero affected offspring are never sampled at all, no matter how common they truly are. Including every proband, without removing at least one affected individual per family, therefore systematically inflates the apparent proportion of affected offspring relative to the population's true underlying segregation ratio, since the "luckiest" families (zero affected) are entirely excluded from the naive calculation while every included family's count is biased upward by at least its proband.

Answer to Q3. Using the R/Python code of Section 7.6 with fam <- data.frame(s = c(6,4,3), a = c(3,2,1)):

```
negloglik <-function(p, fam) {
-sum(lchoose(fam$s -1, fam$a -1) + (fam$a -1)
*log(p) + (fam$s - fam$a) *log(1- p))
}
```

```
fit <-optim(0.3, negloglik, fam = fam, method
="Brent", lower =1e-4, upper =1-1e-4)
ll_alt <--negloglik(fit$par, fam); ll_null <--
negloglik(0.25, fam)
LRT <-2* (ll_alt - ll_null)
c(p_mle = fit$par, LRT = LRT, p_value =pchisq(LRT,
df =1, lower.tail =FALSE))
```

```
import numpy as np
from scipy.optimize import minimize_scalar
from scipy.stats import chi2
from math import lgamma, log
s = [6, 4, 3]; a = [3, 2, 1]
def lchoose(n, k): return lgamma(n+1) - lgamma(k+1)
- lgamma(n-k+1)
def negloglik(p):
return-sum(lchoose(si-1, ai-1) + (ai-1)*log(p) + (si-
ai)*log(1-p) for si, ai inzip(s, a))
res = minimize_scalar(negloglik, bounds=(1e-4, 1-1e-
4), method="bounded")
ll_alt =-negloglik(res.x); ll_null =-negloglik(0.25)
LRT =2*(ll_alt - ll_null)
print(res.x, LRT, chi2.sf(LRT, 1))
```

With only 3 small families, the LRT statistic will typically be small and the p-value well above 0.05, so the data — while giving a point estimate (0.30) somewhat above the Mendelian 0.25 — do not provide statistically significant evidence against the recessive model at this sample size.

Answer to Q4. Under complete (truncate) ascertainment, essentially every affected individual in the population is certain to be identified and to bring their family into the sample, so it is no longer true that families are ascertained through a single, low-probability proband

event; instead, a family with *any* affected children is certain to be ascertained, and a family could in principle be "found" via more than one of its affected members simultaneously. Simply removing one proband per family (the single-ascertainment correction) under-corrects in this setting, because more than one affected child effectively acted as an independent "entry point" into the sample; the correct (truncate) likelihood must instead condition on the fact that the family was ascertained *at all* — formally, conditioning the family's contribution to the likelihood on the event "at least one affected offspring," rather than on "exactly one designated proband."

Answer to Q5. LRT = 2×(−16.9 − (−18.4)) = 2×1.5 = **3.0**. Comparing to χ^2_1, the critical value at $\alpha = 0.05$ is 3.84; since 3.0 < 3.84, the result is not statistically significant at the conventional 5% level ($p \approx 0.083$), so we do not reject the simple autosomal recessive model ($p = 0.25$) in favour of the unrestricted model, despite the unrestricted model achieving a somewhat higher likelihood.

```
LRT <-2* (-16.9- (-18.4)); pchisq(LRT, df =1,
lower.tail =FALSE)
```

```
from scipy.stats import chi2
LRT =2* (-16.9- (-18.4));print(LRT, chi2.sf(LRT, 1))
```

Answer to Q6. Each individual small family contributes only a modest amount of statistical information (few informative meioses), so its likelihood curve as a function of the segregation ratio p tends to be flat/wide, giving weak, imprecise evidence about the true value of p and a wide confidence interval. Because families are ascertained independently, their

individual log-likelihoods can be **summed** to form a combined log-likelihood (exactly as independent likelihood contributions combine in Section 7.4), which sharpens (narrows) the combined likelihood curve around the true parameter value — the same statistical principle of information accumulation that underlies combining LOD scores across families in linkage analysis (Chapter 17) and combining studies in GWAS meta-analysis (Chapter 18).

Chapter 8: Estimation of the Recombination Fraction

Answer to Q1. $\hat{\theta} = 8/50 = 0.16$. $SE(\hat{\theta}) = \sqrt{[0.16(0.84)/50]} = \sqrt{[0.1344/50]} = \sqrt{0.002688} = 0.0518$.

Answer to Q2. $Z(\theta) = \log_{10}[\theta^8 (1-\theta)^{42} / 0.5^{50}]$ at $\theta=0.16$. $(0.16)^8$: $\ln(0.16) = -1.8326$, $\times 8 = -14.661$, $e^{(-14.661)} \approx 4.29\times10^{-7}$. $(0.84)^{42}$: $\ln(0.84) = -0.1744$, $\times 42 = -7.324$, $e^{(-7.324)} \approx 6.59\times10^{-4}$. Numerator $\approx 4.29\times10^{-7} \times 6.59\times10^{-4} \approx 2.83\times10^{-10}$. $(0.5)^{50} \approx 8.88\times10^{-16}$. Ratio $\approx 2.83\times10^{-10} / 8.88\times10^{-16} \approx 3.19\times10^{5}$. $Z = \log_{10}(3.19\times10^{5}) \approx 5.50$. Since $Z = 5.50 \geq 3$, this is conventionally significant evidence for linkage.

Answer to Q3. $\theta = 0.5$ means recombinant and non-recombinant gametes are equally likely — exactly what independent assortment (Mendel's Second Law) predicts for loci on different chromosomes (or very far apart), i.e. knowing which allele was transmitted at one locus gives no information about the other. $\theta = 0$ means every gamete is non-recombinant, i.e., the two loci are always transmitted together — consistent with the loci being so close together (or identical) that crossing-over between them essentially never occurs.

Answer to Q4. A LOD score of −3 at θ = 0.5 is not the usual way LOD is reported ($Z(0.5)$ is by definition 0, since $L(0.5)/L(0.5) = 1$); what is meant here is a LOD score of −3 at some trial θ substantially less than 0.5, or more precisely a summary statistic strongly favouring θ = 0.5. A LOD of −3 (odds of 1000:1 against linkage at the tested θ) provides significant evidence *against* linkage at that trial distance — i.e., strong evidence that the two loci assort independently (are unlinked, or are very far apart).

R and Python verification (Q1, Q2).

```
r_recomb <-8; n <-50
theta_hat <- r_recomb / n
se_theta <-sqrt(theta_hat*(1-theta_hat)/n)
lod <-function(theta, r, n) (r*log10(theta) + (n-r)*log10(1-
theta)) - (n*log10(0.5))
c(theta_hat = theta_hat, se = se_theta, LOD =lod(theta_hat,
r_recomb, n))
```

```
import numpy as np
r_recomb, n =8, 50
theta_hat = r_recomb / n
se_theta =np.sqrt(theta_hat*(1-theta_hat)/n)
lod = (r_recomb*np.log10(theta_hat) + (n-
r_recomb)*np.log10(1-theta_hat)) - n*np.log10(0.5)
print(theta_hat, se_theta, lod)
```

Answer to Q5. Because LOD scores are additive across independent families at a fixed θ, the combined LOD score at θ = 0.15 is simply 1.6 + 1.6 = **3.2**. This crosses the conventional significance threshold of $Z \geq 3$, so while neither family alone provided significant evidence, their combined analysis does provide significant evidence of linkage at θ = 0.15.

Answer to Q6.

```
lod <-function(theta, r, n) (r*log10(theta) + (n-
r)*log10(1-theta)) - (n*log10(0.5))
# Family 1: 8 recombinants out of 50; Family 2: 5
recombinants out of 30
lod_fam1 <-lod(0.2, 8, 50); lod_fam2 <-lod(0.2, 5, 30)
lod_pooled <-lod(0.2, 8+5, 50+30)
c(sum_individual = lod_fam1 + lod_fam2, pooled =
lod_pooled)   # should match
```

```
def lod(theta, r, n):
return (r*np.log10(theta) + (n-r)*np.log10(1-theta)) -
n*np.log10(0.5)
lod_fam1 = lod(0.2, 8, 50); lod_fam2 = lod(0.2, 5, 30)
lod_pooled = lod(0.2, 13, 80)
print(lod_fam1 + lod_fam2, lod_pooled)
```

The two quantities match exactly, confirming the additivity property algebraically as well as numerically: summing log-likelihood-ratio contributions family-by-family is mathematically identical to computing the LOD score directly from the pooled recombinant/ non-recombinant counts.

Chapter 9: Inheritance of Quantitative Traits

Answer to Q1. $\hat{h}^2 = 0.55$ (the regression slope directly estimates h^2 under the mid-parent regression model). This is plausible because birth weight, unlike adult height, is strongly influenced by maternal environmental factors during pregnancy (maternal nutrition, placental function, gestational age, maternal illness such as malaria or hypertension), which inflate V_E relative to V_A and hence lower $h^2 = V_A/(V_A+V_E+\ldots)$.

Answer to Q2. $t = \sigma^2_B/(\sigma^2_B+\sigma^2_W) = 45/(45+15) = 45/60 = 0.75$. $\hat{h}^2 = 2t = 1.50$. (Note: an estimate above 1.0 is

not biologically meaningful as a heritability — it signals that the simplifying assumptions behind $t \approx \frac{1}{2}h^2$ [no dominance variance, no shared sibling environment, no assortative mating] are likely violated in this sample, most plausibly because siblings also share a common household/environment that inflates the between-family variance component beyond what shared genes alone would produce; in practice such estimates would be treated as evidence of shared-environment effects rather than literal heritability, and the model would need extension, e.g., using twin or adoption designs.)

Answer to Q3. Parent–offspring regression works because a parent transmits, on average, exactly half of their *additive* genetic value to each offspring in an unbiased way — the additive effects of individual alleles pass through segregation predictably. Dominance effects (interaction between the two specific alleles at a locus) and epistatic effects (interaction between alleles at different loci) depend on particular allele *combinations* within an individual; these combinations are broken apart and reassembled anew at each generation by segregation and independent assortment/recombination (Chapters 2 and 7), so a parent's dominance/epistatic deviation is not reliably passed on — hence these components inflate H^2 (which counts all genetic variance) but do not inflate the parent–offspring regression slope, which isolates $h^2 = V\_A/V\_P$.

Answer to Q4. Under the threshold model, liability to the defect is treated as a continuous, roughly normally distributed trait influenced by many genes and environmental factors, with the defect appearing only above a high threshold. Full siblings of an affected

child share, on average, half of the many genes contributing to liability, so their own liability distribution is shifted somewhat closer to the threshold than the general population's — increasing the proportion of siblings who exceed the threshold (3%) relative to the general population (0.5%), without requiring that any single gene, on its own, determines the defect in a simple Mendelian fashion.

R and Python verification (Q1, Q2).

```
sigma2_B <-45; sigma2_W <-15
t <- sigma2_B / (sigma2_B + sigma2_W); h2_hat <-2*t
c(t = t, h2_hat = h2_hat)
sigma2_B, sigma2_W =45, 15
t = sigma2_B / (sigma2_B + sigma2_W)
print(t, 2*t)
```

Answer to Q5. Each individual gene of small effect contributes a small, discrete increment (or decrement) to an individual's genotypic value, and the Central Limit Theorem states that the *sum* of a large number of independent (or weakly correlated) random contributions tends toward a normal distribution as the number of contributions grows, regardless of the distribution of each individual contribution — so even though a single gene's effect might follow an arbitrary (even a simple two- or three-valued) distribution, the sum across hundreds of such genes, plus environmental variation, converges toward the familiar smooth, symmetric, bell-shaped normal curve.

Answer to Q6. $H^2 = V_G/V_P = (0.35+0.05+0.02)/1.00 = 0.42/1.00 =$ **0.42**. $h^2 = V_A/V_P = 0.35/1.00 =$ **0.35**. Since V_A is only one (non-negative) component of the total genetic variance $V_G = V_A + V_D + V_I$, and $V_D, V_I \geq 0$, it always holds that $V_A \leq V_G$,

and hence $h^2 \leq H^2$ — narrow-sense heritability can never exceed broad-sense heritability, with equality only when dominance and epistatic variance are both exactly zero.

```
V_A<-0.35; V_D<-0.05; V_I<-0.02; V_E<-0.58; V_P
<- V_A+V_D+V_I+V_E
c(H2 = (V_A+V_D+V_I)/V_P, h2 = V_A/V_P)
```

Chapter 10: Likelihood Methods in Statistical Genetics

Answer to Q1. $l(\theta) = n\_R \log \theta + n\_NR \log(1-\theta)$. $U(\theta) = n\_R/\theta - n\_NR/(1-\theta)$. Setting $U(\theta)=0$: $n\_R(1-\theta) = n\_NR\ \theta \implies n\_R - n\_R\theta = n\_NR\theta \implies n\_R = \theta(n\_R+n\_NR) \implies \hat{\theta} = n\_R/(n\_R+n\_NR)$, matching the estimator quoted in Chapter 8.

Answer to Q2. $\hat{p} = (2\times100+200)/(2\times400) = 400/800 = 0.5$. At $\hat{p}=0.5$, HWE-expected proportions are $p^2=0.25$, $2pq=0.50$, $q^2=0.25$, i.e., expected counts 100, 200, 100 — exactly equal to the observed counts. So the LRT statistic Λ would be (very close to) 0, and we would strongly fail to reject H_0: this sample fits HWE essentially perfectly.

Answer to Q3. Consistency: $\hat{\psi} \to$ true ψ as sample size grows. Asymptotic normality : $\hat{\psi}$ is approximately normally distributed for large n, permitting standard confidence intervals and tests. Asymptotic efficiency: the MLE has the smallest possible variance among consistent, asymptotically normal estimators. Efficiency matters most precisely when samples are *limited* (e.g., a rare Nigerian pedigree disorder with only a handful of informative families), because an inefficient estimator wastes the little information available, whereas the MLE extracts the maximum possible precision from a small dataset.

Answer to Q4. Because $\theta = 0.5$ is the *boundary* of the biologically meaningful parameter space for θ (which by convention ranges over [0, 0.5]) when testing for linkage ($\theta < 0.5$) rather than an interior null value; the standard chi-square asymptotic derivation for the LRT assumes the true parameter lies in the interior of the parameter space under H_0. When the null value is on the boundary, the correct asymptotic null distribution of the (one-sided) LRT statistic is instead a 50:50 mixture of a point mass at zero and a χ^2_1 distribution.

R and Python verification (Q2).

```
n_AA<-100; n_Aa<-200; n_aa<-100; N<-400
p_hat <- (2*n_AA+n_Aa)/(2*N)
E <- N *c(p_hat^2, 2*p_hat*(1-p_hat), (1-p_hat)^2)
sum((c(n_AA,n_Aa,n_aa)-E)^2/E)    # LRT/chi-square ~ 0:
perfect HWE fit
```

```
import numpy as np
n_AA, n_Aa, n_aa =100, 200, 100; N =400
p_hat = (2*n_AA + n_Aa) / (2*N)
E = N*np.array([p_hat**2, 2*p_hat*(1-p_hat), (1-p_hat)**2])
print(np.sum((np.array([n_AA,n_Aa,n_aa]) - E)**2/ E))
```

Answer to Q5. A flat likelihood surface near its peak means that a wide range of parameter values are all nearly as consistent with the observed data as the MLE itself — the data do not sharply discriminate between them — corresponding to a small curvature (second derivative) at the peak and hence, via the observed/expected information matrix (Section 10.4), a *large* standard error for the estimate. A sharply peaked surface, by contrast, means only a narrow range of parameter values is compatible with the data, corresponding to large curvature and a *small* standard error — the data pin down the parameter precisely.

Answer to Q6.

```
newton_raphson(0.99)
newton_raphson(0.5)
print(newton_raphson(0.99))
print(newton_raphson(0.5))
```

Because the score and information functions here are well-behaved (smooth and monotonic) over the interior of (0,1), Newton–Raphson should still converge to the same correct MLE even from the poor starting value of 0.99, though it will typically require a few more iterations to travel the greater distance back toward the true optimum than when starting from a central value like 0.5 — an illustration of the general lesson that Newton–Raphson's fast (quadratic) local convergence is not guaranteed to be fast, or even guaranteed to converge at all, far from the optimum or near a function's boundary.

Chapter 11: Population-Based Association Tests

Answer to Q1. Recessive-model table: TT — cases 45, controls 20; TC+CC — cases 255, controls 280. Row/column totals: TT total=65, TC+CC total=535, cases=300, controls=300, grand total=600. E(TT,case) = (65×300)/600 = 32.5; E(TC+CC,case) = 267.5; by symmetry

E(TT,control)=32.5 E(TC+CC,control)=267.5. χ^2 = $(45-32.5)^2/32.5 + (255-267.5)^2/267.5 + (20-32.5)^2/32.5 + (280-267.5)^2/267.5$ = 4.808+0.584+4.808+0.584 = 10.78 (1 df) — significant (critical value 3.84). OR = (45×280)/(20×255) = 12600/5100 = 2.47.

Answer to Q2. The allele-based test treats each individual's two alleles as two independent observations in the 2×2 table; this independence between the two alleles within a

genotype is precisely the assumption of Hardy–Weinberg equilibrium (random union of gametes), so if HWE does not hold (e.g., due to population stratification), the allele-based test's nominal type-I error rate is no longer valid, whereas the genotype-based test treats each individual (not each allele) as the independent unit and makes no such assumption.

Answer to Q3. Nigeria's population is genetically diverse (multiple ethnic groups with somewhat different allele frequencies at many loci, an instance of the Wahlund-type structure of Chapter 3); if cases and controls are not perfectly matched for ancestry, a marker that simply differs in frequency between ethnic groups can appear spuriously associated with disease if disease prevalence also differs by group, purely due to confounding by ancestry rather than any true genetic effect on the disease — a phenomenon called population stratification. Including ancestry principal components as covariates statistically adjusts for this confounding, protecting against false-positive associations.

Answer to Q4. An OR of 1.0 with a non-significant p-value in a well-powered study provides reasonably strong evidence that this marker has **little or no** effect on disease risk in this population — i.e., the data are consistent with the null hypothesis of no association, and (because the study is well-powered) a true effect of at least moderate size can likely be ruled out, though very small effects can never be entirely excluded by any finite study.

R and Python verification (Q1).

```
tab <-matrix(c(45,255,20,280), nrow=2, byrow=TRUE,
dimnames=list(c("case","control"), c("TT","TC_CC")))
chisq.test(tab, correct =FALSE)
(45*280)/(20*255)     # odds ratio
```

```
import numpy as np
fromscipy.statsimport chi2_contingency
tab = np.array([[45, 255], [20, 280]])
chi2_stat, p, dof, exp = chi2_contingency(tab,
correction=False)
print(chi2_stat, p)
print((45*280)/(20*255))
```

Answer to Q5. Testing multiple genetic models and reporting only the smallest resulting p-value is equivalent, in its statistical effect, to performing several correlated hypothesis tests and selectively reporting the most extreme outcome — exactly the multiple-testing problem addressed formally in Chapters 12 and 18, even though the analyst presents only a single "final" number. The true probability of obtaining at least one nominally significant result by chance, when several models are tried, is higher than the nominal per-test α level (e.g., trying 3 roughly independent models at $\alpha = 0.05$ gives a combined false-positive rate closer to $1 - 0.95^3 \approx 14\%$, not 5%), so the reported p-value understates the true false-positive risk unless the model choice is pre-specified (before seeing the data) or a multiple-testing correction is applied to account for the number of models compared.

Answer to Q6. The **dominant model test** would be expected to have the least power. Under a true recessive effect, heterozygotes have the same risk as homozygous non-carriers; the dominant-model test incorrectly groups these unaffected-risk heterozygotes together with the truly high-risk homozygotes into a single "carrier" category, diluting the genuine signal (since the "carrier" group is now a mix of truly elevated-risk and normal-risk individuals) and thereby reducing the

test's ability to detect the real association, compared with the recessive-model test (which correctly isolates the homozygous risk group) or the 2-df genotype-based test (which does not force any particular grouping).

Chapter 12: Candidate Gene Association Studies

Answer to Q1. Bonferroni threshold = á/m = 0.05/8 = **0.00625**. Only SNPs with a per-test p-value below 0.00625 would be declared significant.

0.05/8

print(0.05/8)

Answer to Q2. When tested SNPs are in strong LD, they are not independent tests, so the simple Bonferroni correction (which assumes m independent tests) is overly conservative — it treats correlated tests as if they each carried a full, independent unit of multiple-testing "penalty," inflating the effective threshold far more than necessary and losing statistical power. A permutation test instead empirically derives the null distribution of the minimum p-value across all tested SNPs *while preserving their actual correlation structure* (since the same LD pattern is present in every permuted dataset), giving an accurate, appropriately less conservative family-wise significance threshold.

Answer to Q3. Two distinct (non-multiple-testing) reasons for the historically poor replication of candidate-gene findings: (i) small sample sizes typical of early candidate-gene studies gave low statistical power and highly variable effect-size estimates, so many "significant" findings were simply the winning tail of sampling variability (the "winner's curse"); (ii) inadequate correction for population stratification (Chapter 14) allowed confounding by ancestry to produce spurious associations that appeared

statistically robust within a single study but did not reflect a true, generalisable biological effect and therefore failed to replicate elsewhere.

Answer to Q4. Because allele frequencies at a genetic locus can differ substantially between populations with different evolutionary and demographic histories, a variant's relevance as a disease "risk allele" is population-specific: a variant that is common and clinically important in one population (here, *APOL1* risk alleles, common in West Africa due to historical selection for trypanosomiasis resistance) may be rare or absent, and therefore clinically irrelevant, in another population (here, European ancestry) — so risk-allele findings, and the clinical or research conclusions built on them, cannot be assumed to transfer across populations without independent validation in the population of intended application.

Answer to Q5. Setting beta_true <- c(0, 0.30, 0, 0.30, 0) (SNP2 and SNP4 non-zero) and re-running the simulation:

```
set.seed(55)
n <-400
snp <-matrix(rbinom(n *5, 2, prob =0.3), nrow = n, ncol =5)
beta_true <-c(0, 0.30, 0, 0.30, 0)
trait <-120+ snp %*% beta_true +rnorm(n, 0, 5)
pvals <-sapply(1:5, function(j) summary(lm(trait ~ snp[,
j]))$coefficients[2, 4])
print(pvals)
print(pvals <0.05/5)
```

```
import numpy as np, statsmodels.api as sm
rng = np.random.default_rng(55)
n =400
snp = rng.binomial(2, 0.3, size=(n, 5))
beta_true = np.array([0, 0.30, 0, 0.30, 0])
```

```
trait =120+ snp @ beta_true + rng.normal(0, 5, n)
pvals = np.array([sm.OLS(trait, sm.add_constant(snp[:,
j])).fit().pvalues[1] for j inrange(5)])
print(pvals, pvals <0.05/5)
```

Both SNP2 and SNP4 should generally be correctly flagged by the Bonferroni threshold, though with only n = 400 and a modest effect size the exact outcome depends on the random seed and effect size chosen — reflecting the broader lesson that statistical power, not just correct methodology, determines whether a true signal is actually detected in any single dataset.

Answer to Q6. Because each tag SNP was chosen specifically so that every other common SNP in the region is correlated with at least one tag at $r^2 > 0.9$ (Section 12.3), a common causal variant — even if not itself directly genotyped — will almost always be very strongly correlated with (and therefore produce a strong, detectable association signal through) its nearest tag SNP; the tag SNP effectively "captures" most of the causal variant's statistical signal by proxy. Testing only 6 well-chosen tags therefore retains nearly all the power to detect an association *somewhere* in the gene that testing all 40 SNPs directly would have provided, while requiring far less genotyping — the central economic rationale of LD-based tagging (Chapter 15).

Chapter 13: Family-Based Association Tests: The TDT and FBAT

Answer to Q1. TDT = $(118-82)^2/(118+82) = 36^2/200 = 1296/200 =$ **6.48**. Comparing to the χ^2_1 critical value at $\alpha = 0.01$ (6.63), $6.48 < 6.63$, so this result is **not quite significant** at the 1% level (though it is significant at the conventional 5% level, critical value 3.84).

```
b <-118; c <-82
tdt <- (b -c)^2/ (b + c)
```

```
c(tdt = tdt, p_value =pchisq(tdt, df =1, lower.tail
=FALSE))
```

```
fromscipy.statsimport chi2
b, c =118, 82
tdt = (b - c)**2/ (b + c)
print(tdt, chi2.sf(tdt, 1))
```

Answer to Q2. The TDT statistic is built entirely from *within-family* transmissions — comparing, for each heterozygous parent, the allele actually transmitted to the affected child against the allele that could equally well have been transmitted (probability 1/2 each under Mendel's first law, Chapter 2) — and never uses the population allele frequency, nor compares to any externally sampled control group. Because Mendelian transmission from a heterozygous parent is 1/2:1/2 regardless of which ethnic subgroup that parent belongs to, mixing trios from two ethnic groups with very different allele frequencies does not bias the null expectation of the TDT in any way; the test remains valid exactly because it never relies on the population's overall allele frequencies matching between any two groups.

Answer to Q3. TDT = $(10-5)^2/(10+5) = 25/15 =$ **1.67**, well below the χ^2_1 critical value of 3.84 (not significant). With only 15 informative trios, the chi-square approximation to the true binomial-based null distribution is questionable (a common rule of thumb wants b+c to be reasonably large, often ≥ 20–30, for the chi-square approximation to be reliable); an **exact test** — a two-sided exact binomial test of b = 10 successes out of b+c = 15 trials against a null probability of 0.5 — would be preferable here and

would not rely on the large-sample chi-square approximation.

```
binom.test(10, 15, p =0.5)
```

```
fromscipy.statsimport binomtest
print(binomtest(10, 15, p=0.5))
```

Answer to Q4. For a single affected child at a bi-allelic marker, T_i (the phenotype-derived weight in the FBAT statistic) equals 1 for the (necessarily affected) offspring in each informative trio, and X_i is the offspring's observed count of the tested allele (0, 1, or 2 copies); conditioning on the sufficient statistic H (the parental genotypes) then reduces the general FBAT statistic exactly to a comparison of transmitted vs. non-transmitted alleles from heterozygous parents — precisely McNemar's test, i.e. the classical TDT statistic of Section 13.3.

Answer to Q5.

```
set.seed(1)
n_reps <-1000
tdt_stats <-replicate(n_reps, {
  transmissions <-rbinom(90, 1, 0.5)
  b <-sum(transmissions ==1); c <-sum(transmissions ==0)
  (b - c)^2/ (b + c)
})
mean(tdt_stats >3.84)
import numpy as np
rng =np.random.default_rng(1)
n_reps =1000
tdt_stats = []
for _ inrange(n_reps):
   transmissions = rng.binomial(1, 0.5, 90)
   b = np.sum(transmissions ==1); c = np.sum(transmissions
        ==0)
```

```
    tdt_stats.append((b - c)**2/ (b + c))
print(np.mean(np.array(tdt_stats) >3.84))
```

The simulated proportion of replications exceeding 3.84 should lie close to the nominal 0.05 type-I error rate, confirming that the TDT's asymptotic chi-square(1) null distribution is well calibrated for this sample size (90 heterozygous parents) — a direct simulation-based validation of the theory of Section 13.3, in the spirit of Chapter 20.

Answer to Q6. The classical TDT strictly requires both parents' genotypes for every trio, since its statistic is built entirely from the transmitted/non-transmitted alleles of heterozygous parents — with no parental genotype at all, there is nothing to compute. The general FBAT framework (Section 13.4), however, is built around a sufficient statistic that can, in some cases, be reconstructed from the genotypes of *other* family members (e.g., unaffected siblings) even when a parent's genotype is missing, using Mendelian transmission logic to partially infer the missing parental information; families with missing parents but informative siblings can therefore still contribute (partial) information to the FBAT statistic, whereas they would have to be discarded entirely under the simple trio-only TDT.

Chapter 14: Population Substructure, Stratification, and Admixture

Answer to Q1. $\bar{p} = 0.6(0.70) + 0.4(0.40) = 0.42 + 0.16 = 0.58$. Naive-HWE expected heterozygosity $= 2(0.58)(0.42) = 0.4872$. Wahlund deficit term $= 2(0.6)(0.4)(0.70-0.40)^2 = 0.48 \times 0.09 = 0.0432$. Actual pooled heterozygosity $= 0.4872 - 0.0432 =$ **0.4440**.

```
w1<-0.6; w2<-0.4; p1<-0.70; p2<-0.40
p_bar <- w1*p1 + w2*p2
naive_het <-2*p_bar*(1-p_bar)
```

```
deficit <-2*w1*w2*(p1-p2)^2
actual_het <- naive_het - deficit
c(p_bar=p_bar, naive_het=naive_het, deficit=deficit,
actual_het=actual_het)

w1, w2, p1, p2 =0.6, 0.4, 0.70, 0.40
p_bar = w1*p1 + w2*p2
naive_het =2*p_bar*(1-p_bar)
deficit =2*w1*w2*(p1-p2)**2
actual_het = naive_het - deficit
print(p_bar, naive_het, deficit, actual_het)
```

Answer to Q2. Population stratification refers to a sample being drawn from multiple genetically distinct subgroups that remain distinguishable at the group level — for example, a Nigerian national genetic study that separately recruits Yoruba participants in Ibadan and Hausa/Fulani participants in Kano, each group internally fairly homogeneous but differing from each other. Cryptic relatedness instead refers to some sampled individuals being more closely related to one another than expected under random sampling, without this being known — for example, a "population-based" case sample recruited from a single rural community in which, unknown to the investigators, several participants are in fact second cousins of one another through unrecorded family ties.

Answer to Q3. $\lambda_{GC} = 1.35$ suggests systematic inflation of the genome-wide test statistics beyond what would be expected under the null, consistent with — though not conclusive proof of — inadequately corrected population stratification or cryptic relatedness. Before concluding this is purely stratification, a careful analyst should examine a **principal component analysis** of the genome-wide genotype data (Section

14.5): if the leading PCs show clear sub-population clustering that aligns with the observed inflation (and including those PCs as covariates substantially reduces λ_{GC}), stratification is the likely explanation; if instead PCs show no clear structure and the inflation persists after PC adjustment, an alternative explanation — genuine, highly **polygenic** architecture (many true small-effect associations spread across the genome, which also mildly inflates λ_{GC}) — should be considered instead.

Answer to Q4. The leading principal components of genome-wide genotype data capture the major axes of ancestry variation in the sample (Section 14.5); because an individual's estimated position along these axes is a good proxy for their ancestral background, including the PCs as covariates in the regression model statistically "holds ancestry constant" when estimating each marker's effect on the trait, removing the part of the marker–trait association that is attributable to ancestry differences between cases and controls (Section 14.4) rather than to any true effect of the marker itself.

Answer to Q5. Increasing the allele-frequency difference (e.g. p1 = 0.9, p2 = 0.3, so $|p1-p2| = 0.6$ instead of 0.3) increases the Wahlund deficit term $2w_1w_2(p1-p2)^2$, since this term grows with the *square* of the frequency difference between the subpopulations; quadrupling the frequency difference ($0.3 \rightarrow 0.6$, a doubling) quadruples the deficit term. This confirms that the more differentiated the mixed subpopulations are at a locus, the larger the mechanical heterozygote deficit (and hence the larger the apparent, spurious departure from HWE) that results purely from pooling them, with no true inbreeding, selection, or non-random mating required.

```
w1<-0.5; w2<-0.5; p1<-0.9; p2<-0.3
deficit <-2*w1*w2*(p1-p2)^2
deficit

w1, w2, p1, p2 =0.5, 0.5, 0.9, 0.3
print(2*w1*w2*(p1-p2)**2)
```

Answer to Q6. An individual whose position falls between two population clusters most plausibly represents an **admixed individual** — someone whose recent ancestry draws from both underlying populations (for example, a person with one parent from each source population, or with more distant mixed ancestry), so that their genome-wide genotype pattern is, in effect, a weighted blend of both clusters' typical genotype patterns, placing them at an intermediate position along the principal component axis that separates the two groups.

Chapter 15: Linkage Disequilibrium and Haplotype Inference

Answer to Q1. p_A = 0.45+0.05 = 0.50; p_B = 0.45+0.05 = 0.50. D = 0.45 − (0.50)(0.50) = 0.45 − 0.25 = 0.20. D_max (D>0) = min(p_Ap_b, p_ap_B) = min[(0.50)(0.50),(0.50)(0.50)] = 0.25. D′ = 0.20/0.25 = 0.80. $r^2 = (0.20)^2/[(0.5)(0.5)(0.5)(0.5)] = 0.04/0.0625 = 0.64$. Both D′ (0.80) and r^2 (0.64) indicate strong linkage disequilibrium between these two loci.

Answer to Q2. Nigerian (African) populations generally have larger and more ancient effective population sizes than populations founded more recently from smaller groups (e.g., European populations, which experienced population bottlenecks during out-of-Africa migration). Larger, older populations have

accumulated more generations of recombination events, breaking down ancestral haplotype associations more thoroughly, so LD extends over shorter physical distances (decays faster with distance) in Nigerian/African populations than in populations with more recent, smaller founding effective population sizes.

Answer to Q3. Required sample size scales approximately as $1/r^2$ relative to direct typing: $N_{marker} \approx N_{direct} / r^2 = 1000/0.5 = 2000$ samples.

Answer to Q4. Clark's algorithm is a deterministic, parsimony-based approach: it starts from individuals whose haplotypes can be read off unambiguously, then resolves remaining ambiguous individuals one at a time by preferentially assigning them haplotypes already seen in the sample, aiming to minimise the total number of distinct haplotypes — it does not use a formal probability model. The EM algorithm is a likelihood-based, probabilistic approach: it treats the unresolved phase for each individual as missing data, and iteratively computes the expected probability of each possible haplotype-pair resolution given current estimated population haplotype frequencies (E-step), then updates those frequency estimates by maximum likelihood (M-step), converging to formal maximum likelihood estimates of haplotype frequencies across the whole sample simultaneously.

R and Python verification (Q1, Q3).

```
pA<-0.50; pB<-0.50; pAB<-0.45
D <- pAB - pA*pB
Dmax <-min(pA*(1-pB), (1-pA)*pB)
Dprime <- D/Dmax
```

```
r2 <- D^2/(pA*(1-pA)*pB*(1-pB))
c(D=D, Dprime=Dprime, r2=r2)
1000/0.5# Q3: sample size scaling by 1/r^2

pA, pB, pAB =0.50, 0.50, 0.45
D = pAB - pA*pB
Dmax =min(pA*(1-pB), (1-pA)*pB)
Dprime = D/Dmax
r2 = D**2/(pA*(1-pA)*pB*(1-pB))
print(D, Dprime, r2)
print(1000/0.5)
```

Answer to Q5. Association power depends on r^2 (not merely physical proximity or D′) because r^2 directly determines the effective sample size penalty for testing the marker instead of the true causal variant (Section 15.4's sample-size-inflation relationship, $n_{marker} \approx n_{causal} / r^2$); a marker with small r^2 to the causal locus, even if physically adjacent, requires a much larger sample to achieve the same power as directly testing the causal variant, and in the extreme (r^2 near 0) essentially no amount of additional sample size at the marker will efficiently recover the causal signal, however close the marker is physically.

Answer to Q6.

```
pA<-160/320; pB<-160/320; pAB<-150/320
D <- pAB - pA*pB
Dmax <-min(pA*(1-pB), (1-pA)*pB)
Dprime <- D/Dmax
r2 <- D^2/(pA*(1-pA)*pB*(1-pB))
c(D=D, Dprime=Dprime, r2=r2)
pA, pB, pAB =160/320, 160/320, 150/320
D = pAB - pA*pB
Dmax =min(pA*(1-pB), (1-pA)*pB)
Dprime = D/Dmax
```

```
r2 = D**2/(pA*(1-pA)*pB*(1-pB))
print(D, Dprime, r2)
```

This haplotype configuration (dominated by AB and ab, with only small amounts of the recombinant Ab and aB haplotypes) gives D2 very close to 1 (near-complete disequilibrium relative to the maximum possible given the allele frequencies) and a correspondingly high r^2 as well, since both allele frequencies are equal (0.5) — illustrating the special case where, when p_A = p_B, D2 and r^2 tend to agree closely; in general, however, D2 can be high while r^2 is much lower whenever the two allele frequencies differ substantially (Section 15.4), which is why r^2 (not D2) is the measure directly relevant to statistical power for association testing (Q5).

Chapter 16: Coalescent Models

Answer to Q1. E(T) = 2N = 2(5,000) = 10,000 generations = 10,000 × 25 = 250,000 years.

Answer to Q2. E(T_MRCA) = 4N(1 " 1/n) = 4(8,000)(1 " 1/10) = 32,000 × 0.9 = 28,800 generations.

Answer to Q3. As n grows, most of the *additional* lineages coalesce very quickly near the "tips" (recent past) of the genealogy, because the coalescence rate C(k,2)/(2N) is large when k (the number of remaining distinct lineages) is large — so adding more samples mainly adds shallow, recent branching, not additional depth. The final, slowest step — the coalescence of the last two remaining lineages, at rate 1/(2N), expected waiting time 2N generations — always dominates the total time to the MRCA, regardless of how large n is, which is why E(T_MRCA) converges to (rather than grows without bound past) 4N generations.

Answer to Q4. E(S) = θ × Σ_{k=1}^{n-1} 1/k = 0.008 × 3.548 = 0.0284. (Note: this is the expectation per site typically applied across a genomic region with many independent sites/base pairs in practice; for a single locus with θ this small, an expected value below 1 simply reflects a low probability of observing any segregating site in a very short sequence, consistent with a low value of θ at that specific locus.)

R and Python verification (Q1, Q2, Q4).

```
2*5000*(1-1/10)          # Q1: E[T] generations (2N convention check)
4*8000*(1-1/10)          # Q2: E[T_MRCA] = 4N(1-1/n)
theta <-0.008
theta *sum(1/(1:9))       # Q4: expected segregating sites, harmonic sum to n-1=9
import numpy as np
print(2*5000*(1-1/10))
print(4*8000*(1-1/10))
theta =0.008
print(theta *sum(1/k for k inrange(1, 10)))
```

Answer to Q6. The bottleneck sharply reduced the effective population size, N, of the founding out-of-Africa lineage for an extended period; since expected coalescence time scales directly with N (E(T) = 2N generations for a pair, Section 16.4), a smaller N during the bottleneck mechanically forces lineages to coalesce much faster during that period than they would have in the larger ancestral African population. Faster coalescence means fewer independent branches survive to accumulate distinct mutations over time, so less genetic variation is preserved through the bottleneck; because mutation since the bottleneck has not yet had enough generations to fully restore the lost diversity, non-African populations — all

descended from this reduced founding group — retain systematically lower genetic diversity today than African populations, whose ancestors never passed through this size reduction.

Chapter 17: Parametric and Non-Parametric Linkage Analysis

Answer to Q1. $\hat{\theta} = R/(R+N\_R) = 2/20 = 0.10$. $\theta^{R(1-\theta)}\{N\_R\} = (0.10)^{2(0.90)}\{18\} = 0.01 \times 0.1501 = 0.001501$. 0.5^20 = 9.537×10^{-7}. $Z(\hat{\theta}) = \log_{10}(0.001501/9.537\times10^{-7}) = \log_{10}(1573.9) = 3.197$. Since $Z \approx 3.20 \geq 3$, this constitutes conventionally significant evidence of linkage.

Answer to Q2. Combined LOD = $Z_1(0.1) + Z_2(0.1) = 1.2 + 0.9 = 2.1$. This simple addition is valid because LOD scores at a fixed θ are, by construction, log-likelihood-ratio contributions from independent families (log of a product of independent likelihoods = sum of logs), so summing the LOD scores across independent pedigrees, evaluated at the same θ, gives exactly the LOD score for the combined dataset at that θ.

Answer to Q3. Mean IBD = [0(10)+1(20)+2(30)]/60 = (0+20+60)/60 = 80/60 = 1.333. This exceeds the null expectation of 1.0, suggesting possible linkage between the marker and an asthma-susceptibility locus, though (as in Worked Example 17.2) a formal significance test would be needed before concluding definitively.

Answer to Q4. Type 2 diabetes is a complex, polygenic trait with an unknown, almost certainly non-Mendelian mode of inheritance (multiple contributing genes, each of modest effect, interacting with environmental/lifestyle factors) — specifying a single dominant/recessive genetic model with

a fixed penetrance, as parametric analysis requires, would misrepresent the true biology and could severely reduce power or even mask true linkage signals; non-parametric (allele-sharing) analysis makes no such assumption and is therefore more appropriate. Sickle-cell disease, by contrast, has a well-established, simple autosomal recessive mode of inheritance with essentially complete penetrance (Chapters 1–2), so a correctly specified parametric model can be used, and — because the correct model is known — parametric analysis will generally deliver higher statistical power than a non-parametric approach for this disorder.

R and Python verification (Q1, Q2, Q3).

```
R<-2; NR<-18; n<-20
theta_hat <- R/(R+NR)
lod <-log10((theta_hat^R)*(1-theta_hat)^NR) -log10(0.5^n)
c(theta_hat=theta_hat, LOD=lod)              # Q1
1.2+0.9# Q2: combined LOD across families
(0*10+1*20+2*30) /60# Q3: mean IBD sharing
R, NR, n =2, 18, 20
theta_hat = R/(R+NR)
```

```
import numpy as np
lod = np.log10((theta_hat**R)*(1-theta_hat)**NR) -
np.log10(0.5**n)
print(theta_hat, lod)
print(1.2+0.9)
print((0*10+1*20+2*30)/60)
```

Answer to Q5. An observed mean IBD sharing of 1.35 among affected sibling pairs, compared with the null expectation of exactly 1.0 under no linkage, is consistent with, and provides evidence *for*, linkage between the marker and a disease locus: affected relatives sharing more genetic material IBD than chance alone would predict is precisely

the signature that non-parametric linkage analysis (Section 17.7) is designed to detect. Whether the excess of 0.35 above the null value of 1.0 is *statistically significant* would need to be assessed formally via the MLS or an equivalent NPL test statistic, accounting for the sample size of 20 pairs.

Answer to Q6. The null hypothesis of no linkage corresponds to $(z_0, z_1, z_2) = (1/4, 1/2, 1/4)$, and under the alternative of true linkage, sharing can only be pushed in one direction (z_2 increased, z_0 decreased) — biologically, affected relatives cannot share *less* than the null-expected amount of IBD due to a disease locus, since sharing more of the disease-linked region is what excess co-inheritance produces. Because the alternative hypothesis parameter space is therefore restricted to only one side of the null value (a "boundary" null hypothesis, exactly the same situation as the one-sided variance-component tests discussed in Chapter 10), the standard regularity conditions for the usual χ^2 asymptotic result do not hold, and the correct null distribution is instead the non-standard 50:50 mixture of a point mass at zero (representing the cases where the constrained MLE falls exactly at the null value) and a χ^2_1 distribution (representing the cases where it falls strictly inside the alternative region).

Chapter 18: Genome-Wide Association Studies (GWAS)

Answer to Q1. The threshold is unchanged by the number of SNPs actually tested in this particular study; it is fixed at $\alpha = 0.05/1{,}000{,}000 = 5\times10^{-8}$ under the standard convention (based on the estimated total number of effectively independent common-variant tests across the genome). A SNP with $p = 2\times10^{-7}$ does **not** meet

this threshold ($2\times10^{-7} > 5\times10^{-8}$), so it would be classed as, at best, "suggestive," not genome-wide significant.

Answer to Q2. A systematic, genome-wide inflation of significance (rather than isolated peaks) together with λ_{GC} substantially above 1.0 (here, 1.35) most likely indicates inadequately corrected population stratification (or other systematic confounding, e.g., uncorrected relatedness/cryptic family structure among samples) rather than a genuine excess of true genetic associations. The population-structure assessment and correction step (Section 18.3, step 3 — typically PCA-based ancestry adjustment, possibly needing additional principal components or a mixed-model approach) should be revisited.

Answer to Q3. Because LD decays faster, and haplotype patterns differ, in African (including Nigerian) populations relative to the smaller-founding-population history typical of many European populations (Section 15.6), a genotyped SNP panel designed and validated around European LD patterns will, on average, "tag" (capture, via LD) a smaller fraction of the total genetic variation carried by nearby, untyped variants in a Nigerian sample — and imputation from a reference panel lacking sufficient African haplotype diversity will similarly perform less accurately — both effects reducing statistical power and completeness of genomic coverage relative to a comparably sized European-ancestry GWAS using the same array/panel.

Answer to Q4. A single common-variant SNP with a modest odds ratio (1.15) explains only a small fraction of the overall heritable risk of a complex, polygenic trait like T2DM (Chapter 9's heritability discussion; Section 18.7's polygenic picture), so it has very limited value on its own

for individual risk prediction — most patients carrying the risk allele will never develop the disease, and most patients who do develop it will not carry this particular allele. A more informative alternative is a polygenic risk score aggregating the effects of very many (often thousands of) SNPs genome-wide; however, the clinician should further be cautioned that most existing polygenic risk scores were derived predominantly from European-ancestry discovery GWAS and are known to transfer poorly (systematically less accurately) to Nigerian and other African-ancestry patients (Section 18.6), so a polygenic score ideally derived from, or at minimum properly recalibrated for, African-ancestry data should be sought wherever possible.

R and Python verification (Q1).

5e-8# genome-wide significance threshold (fixed convention, independent of study SNP count)
2e-7<*5e-8*# does p = 2x10^-7 meet the threshold? (FALSE)
threshold =5e-8
p_snp =2e-7
print(threshold, p_snp < threshold)

Answer to Q5. Under severe uncorrected stratification with no true associations, the Q–Q plot would show the *entire* bulk of points lifting away from the diagonal fairly uniformly across the whole range of p-values (not just at the extreme tail), reflecting a general inflation of test statistics genome-wide (high λ_{GC}) with no distinct outliers at the top — a diffuse, whole-plot departure from the diagonal. By contrast, a well-controlled study with a handful of true associations shows the bulk of SNPs (the vast majority, which are truly null) tracking the diagonal line very closely, with only a small number of points at the extreme upper-right departing sharply upward — precisely the pattern shown in the Q–Q plot of Section 18.5.

Answer to Q5. Under severe uncorrected stratification with no true associations, the Q–Q plot would show the *entire* bulk of points lifting away from the diagonal fairly uniformly across the whole range of p-values (not just at the extreme tail), reflecting a general inflation of test statistics genome-wide (high λ_{GC}) with no distinct outliers at the top — a diffuse, whole-plot departure from the diagonal. By contrast, a well-controlled study with a handful of true associations shows the bulk of SNPs (the vast majority, which are truly null) tracking the diagonal line very closely, with only a small number of points at the extreme upper-right departing sharply upward — precisely the pattern shown in the Q–Q plot of Section 18.5.

Answer to Q6. The 5×10^{-8} threshold was calibrated to control the family-wise error rate for the roughly 1 million effectively independent common-variant tests typical of a standard SNP genotyping array (Section 18.4). A whole-genome sequencing study tests many more variants — tens of millions, including a large proportion of rare variants that are less correlated with one another (weaker LD, Chapter 15, Section 15.6) — so the effective number of independent tests is substantially larger, requiring a correspondingly more stringent (smaller) significance threshold to maintain the same overall false-positive control; in practice, and especially once rare-variant burden/gene-based tests (Chapter 20) are also considered, thresholds around 10^{-9} or gene-based thresholds like the 2.5×10^{-6} figure of Appendix A are used instead.

Chapter 19: Genotype Imputation

Answer to Q1. A "hard-called" imputed genotype reports a single, most-probable genotype (e.g., 0, 1, or 2 copies of the alternative allele) at the untyped position, discarding information about how confident that call actually was. A "dosage" instead reports the full posterior expectation of the allele count (a continuous value between 0 and 2, e.g. 1.85) reflecting the underlying genotype-probability distribution produced by the imputation algorithm. Association testing on dosages is preferred because it correctly propagates the imputation uncertainty into the test's standard errors — a poorly imputed variant (low INFO/r^2, Section 19.5) will have dosages clustered less tightly around integer values, appropriately down-weighting its contribution and protecting against overconfident false-positive or false-negative conclusions that a hard-call approach could produce by treating every imputed genotype as if it were directly, perfectly observed.

Answer to Q2. (i) An early, largely European reference panel contains comparatively little of the additional haplotype diversity present in West African populations (which, having larger and more ancient effective population sizes, carry more distinct haplotypes overall — Chapter 15, Section 15.6), so many haplotypes actually present in a Nigerian study sample would simply have no closely matching counterpart anywhere in the panel. (ii) Faster LD decay in African populations (Section 15.6) means that, even where a broadly similar haplotype background exists, the physical distance over which a genotyped tag SNP remains strongly correlated with a nearby untyped variant is shorter, so a reference panel and genotyping array designed and validated primarily on European LD patterns

will, on average, tag a smaller fraction of the variation actually present in a Nigerian genome.

Answer to Q3. No — a variant imputed with INFO/r^2 = 0.25 falls below the commonly used inclusion thresholds (typically 0.3 or, more stringently, 0.8) described in Section 19.5, indicating substantial uncertainty in its imputed genotypes; such a poorly imputed variant would ordinarily be **excluded** from the primary association analysis (or, at minimum, flagged for cautious secondary interpretation only), since including it risks contributing mostly noise, and — at very low quality — can introduce spurious associations rather than genuine signal.

Answer to Q4. Strengthening the LD between SNP1 (the tag) and SNP2 (the untyped marker being imputed) — for instance by increasing the coefficient linking *p2* to *snp1_ref* in the simulation — makes each SNP1 genotype class a more reliable predictor of the corresponding SNP2 genotype, because the two loci's alleles co-occur on the same haplotypes more consistently (higher r^2, Chapter 15, Section 15.4). This directly increases the accuracy of the simple "modal genotype" imputation rule, and hence increases the masked-genotype concordance rate observed in the study sample — exactly mirroring why real imputation methods perform best for variants in strong LD with genotyped markers and perform poorly for weakly-tagged variants.

```
set.seed(19)
n_ref <-500
snp1_ref <-rbinom(n_ref, 2, 0.4)
# stronger LD: larger coefficient linking snp2 to snp1
snp2_ref <-pmin(pmax(round(rbinom(n_ref, 2,
pmin(pmax(0.4*0.95+ (snp1_ref-0.8)*0.30/2, 0), 1))), 0), 2)
```

```
lookup <-tapply(snp2_ref, snp1_ref, function(x)
as.integer(names(sort(table(x), decreasing =TRUE))[1]))
n_study <-300
snp1_study <-rbinom(n_study, 2, 0.4)
snp2_study_true <-pmin(pmax(round(rbinom(n_study, 2,
pmin(pmax(0.4*0.95+ (snp1_study-0.8)*0.30/2, 0), 1))), 0), 2)
snp2_imputed <- lookup[as.character(snp1_study)]
mean(snp2_imputed == snp2_study_true)
```

```
import numpy as np
from collections import Counter
rng = np.random.default_rng(19)
n_ref =500
snp1_ref = rng.binomial(2, 0.4, n_ref)
p2 = np.clip(0.4*0.95+ (snp1_ref -0.8)*0.30/2, 0, 1)
snp2_ref = np.clip(np.round(rng.binomial(2, p2)), 0, 2)
lookup = {g: Counter(snp2_ref[snp1_ref ==
g]).most_common(1)[0][0] for g in [0,1,2] if
np.any(snp1_ref==g)}
n_study =300
snp1_study = rng.binomial(2, 0.4, n_study)
p2s = np.clip(0.4*0.95+ (snp1_study -0.8)*0.30/2, 0, 1)
snp2_study_true = np.clip(np.round(rng.binomial(2, p2s)), 0, 2)
snp2_imputed = np.array([lookup[g] for g in snp1_study])
print(np.mean(snp2_imputed == snp2_study_true))
```

Answer to Q5. Imputation solves a "cost" problem because it allows a cheaply genotyped array (a few hundred thousand to a million SNPs) to yield inferred genotypes at tens of millions of variants, in silico, at a small fraction of the cost of directly sequencing every participant (Section 19.2). It solves a "meta-analysis harmonisation" problem because studies genotyped on different arrays (with different, only partially overlapping SNP sets) can each be imputed up to the same dense reference panel, producing

a common set of variants that can be jointly analysed or meta-analysed across studies, rather than being limited to the small intersection of SNPs common to every array.

Answer to Q6. Including both the "high confidence" (5.2 million) and "usable with caution" (1.9 million) tiers gives 7.1 million of the 8 million imputed variants, i.e. 7.1/8 = **88.75%**. The main risk of this more permissive threshold is that a meaningful fraction of the additional 1.9 million included variants have less reliable imputed genotypes (more residual uncertainty, INFO between 0.3 and 0.8), increasing the chance that noise from poorly imputed genotypes contributes false-positive or attenuated-effect associations to the downstream analysis — the fundamental completeness-versus-reliability trade-off discussed in Section 19.5.

Chapter 20: Genome Sequencing, Rare-Variant Methods, and Simulation-Based Approaches

Answer to Q1. A genotyping array can only measure genotypes at a fixed, pre-specified panel of previously known SNPs chosen when the array was designed (Section 19.2/20.2); by construction it cannot report anything about a variant that was unknown, and therefore not included on the array, at design time. Whole-genome sequencing instead reads the DNA sequence directly, base by base, and so will detect and report *any* variant present in the sequenced sample relative to the reference genome — including a truly novel variant never previously observed in any human — because detection does not depend on the variant having been anticipated in advance.

Answer to Q2. A SKAT-type variance-component test would be expected to have greater power here (Section 20.4).

Because the gene's rare variants act in **mixed directions** (10 increasing and 5 decreasing risk), a simple burden test — which collapses all variants into a single directional score, implicitly assuming they act consistently — would have its true signal partially cancelled out, whereas a variance-component test evaluates whether the *variance* of variant effects departs from zero without requiring a consistent direction, and is therefore considerably more robust to exactly this kind of mixed-effect scenario.

Answer to Q3. Increasing the sample size from 200 to 2,000 individuals (400 to 4,000 haplotypes) will substantially increase the **total number** of variants observed, because a larger sample captures more of the rare variation segregating in the population (many rare variants, by definition, are carried by very few haplotypes and are simply missed entirely in a small sample). However, the **proportion** of variants classified as rare (MAF < 1%) should increase further still, not merely stay constant, because coalescent theory (Chapter 16) predicts that the site-frequency spectrum is heavily weighted toward rare variants — most mutations in a genealogy occur on short, low-frequency branches near the tips of the tree — so a larger sample disproportionately reveals additional rare, previously unobserved variants relative to the (already largely captured) common variants.

```
import msprime, numpy as np
for n in [200, 2000]:
   ts = msprime.sim_ancestry(samples=n,
population_size=10_000, sequence_length=1_000_000,
                        recombination_rate=1e-8,
random_seed=20)
   mts = msprime.sim_mutations(ts, rate=1e-8,
random_seed=21)
```

```
    geno = mts.genotype_matrix()
    freqs = geno.mean(axis=1)
    n_rare = np.sum((freqs >0) & (freqs <0.01))
print(n, mts.num_sites, n_rare, n_rare / mts.num_sites if
mts.num_sites  else0)
```

Answer to Q4. Lowering *odds_ratio* from 2.0 toward 1.2 should **decrease** the estimated power monotonically, since a smaller true effect size produces a smaller difference in disease probability between carriers and non-carriers of the gene's rare-variant burden, making the association harder to distinguish from chance in any dataset of fixed size. This illustrates the general, universal relationship in statistical power analysis: for a fixed sample size and significance threshold, statistical power increases with the true effect size (and, separately, with sample size and allele/carrier frequency), and every rare-variant study design must contend with the fact that plausible true effect sizes for individual genes are often modest, requiring very large sequenced sample sizes to achieve adequate power.

```
for orv in [2.0, 1.6, 1.2]:
print(orv,  simulate_burden_power(odds_ratio=orv))
```

Answer to Q5. The coalescent is far more efficient for neutral variation in a single large population because it only needs to simulate the genealogical history of the individuals actually *sampled* — typically a few hundred to a few thousand lineages — tracing their ancestry backward only until they have all coalesced to their most recent common ancestor(s), regardless of how large the true underlying population was; it never needs to represent the millions of individuals in the population who left no descendants in the sample. Forward-in-time simulation, by contrast, must simulate the entire population, every individual, every generation, moving forward from the ancestral population

to the present — a computational burden that scales with total population size and number of generations, not just sample size. Forward-in-time simulation becomes necessary, however, whenever the process being modelled depends on information the backward-only coalescent cannot represent — most importantly, **natural selection** acting on a locus (since an individual's probability of leaving descendants, and hence the very shape of the genealogy, depends on their fitness, which must be tracked forward through the generations as it changes), as well as complex, explicitly-modelled pedigree or non-neutral demographic scenarios (Chapter 6).

Answer to Q6. Prediction accuracy increases with heritability because heritability, $h^2 = V\_A/V\_P$ (Chapter 9, Section 9.4), directly measures the proportion of total phenotypic variance that is attributable to additive genetic effects — the only component that genomic prediction methods can, even in principle, hope to capture from genotype data. As h^2 increases, a larger share of the observed phenotypic variation is genuinely genetic (and therefore predictable from genotypes) rather than environmental noise (which no genotype-based method can predict), so every prediction strategy's achievable accuracy rises toward its ceiling as heritability approaches 1; conversely, as h^2 approaches 0, genetic prediction accuracy for any method must approach zero as well, since there is almost no genetic signal left in the phenotype to predict.