

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 $p = \text{freq}(A)$ and $q = \text{freq}(a)$, 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:

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

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

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

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

$$\begin{aligned} p_1 &= P(AA) + 1/2 P(Aa) = p_0^2 + 1/2 (2p_0 q_0) = p_0^2 + p_0 q_0 \\ &= p_0(p_0 + q_0) = p_0 \times 1 = p_0. \end{aligned}$$

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 , $2p_0 q_0$, 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 $\hat{p} = (2 \times 592 + 184) / 1600 = 0.86$, $\hat{q} = 0.14$. 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.

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

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: $E_{AA} = 1000(0.87)^2 = 756.9$; $E_{Aa} = 2(1000)(0.87)(0.13) = 226.2$; $E_{aa} = 1000(0.13)^2 = 16.9$.

$$\begin{aligned}\chi^2 &= (760 - 756.9)^2/756.9 + (220 - 226.2)^2/226.2 + (20 - 16.9)^2/16.9 \\ &= 0.013 + 0.170 + 0.569 = 0.752\end{aligned}$$

Critical value ($\chi^2_{0.05, 1 \text{ df}} = 3.84$). 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 $A_1, \dots, A_k$ with frequencies $p_1, \dots, p_k$ ($\sum p_i = 1$), random mating gives equilibrium genotype frequencies:

$$\begin{aligned}P(A_i A_i) &= p_i^2 (\text{homozygote}), P(A_i A_j) \\ &= 2p_i p_j (i \neq j) (\text{heterozygote})\end{aligned}$$

obtained by expanding $(p_1 + p_2 + \dots + p_k)^2 = 1$. 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 $P(O)$: $r^2 = 0.620 \Rightarrow r = \sqrt{0.620} = 0.787$. From $P(A) + P(O) = (p+r)^2 \Rightarrow p + r = \sqrt{(0.620+0.210)} = \sqrt{0.830} = 0.911 \Rightarrow p = 0.911 - 0.787 = 0.124$. Then $q = 1 - p - r = 1 - 0.124 - 0.787 = 0.089$. So the

estimated allele frequencies are $I^A \approx 0.124$, $I^B \approx 0.089$, $i \approx 0.787$ — 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 $\hat{q} = (2 \times 0 + 5) / (2 \times 500) = 0.005$, 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 $p \times p = p^2$; 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()

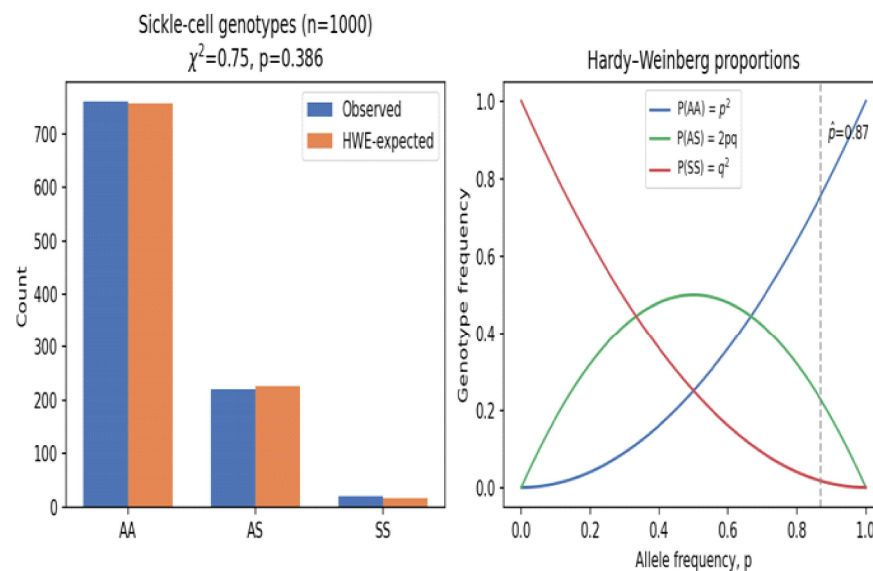
p_grid = np.linspace(0, 1, 200)
axes[1].plot(p_grid, p_grid**2, label="P(AA) =  $p^2$ ",
color="#4C72B0")
axes[1].plot(p_grid, 2*p_grid*(1-p_grid), label="P(AS) =
2pq", color="#55A868")

```

```

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
HWEExact(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 in range(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 $\alpha = 0.05$.

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.