

10

Likelihood Methods in Statistical Genetics

10.1: Learning Objectives

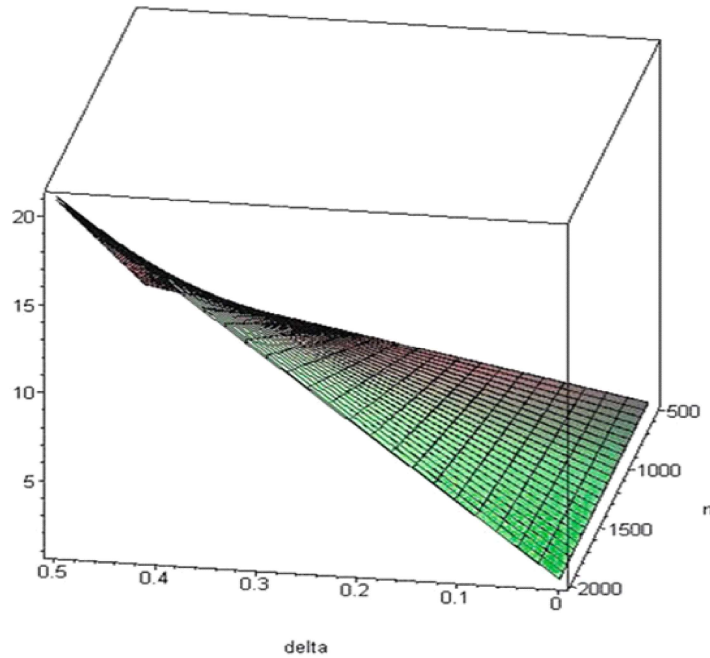
By the end of this chapter, the reader should be able to construct a likelihood function for genetic data, derive maximum likelihood estimators (MLEs) analytically and numerically (Newton–Raphson, Fisher scoring), state the asymptotic properties of MLEs, and apply the likelihood ratio test to genetic 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 $\mathbf{x} = (x_1, \dots, x_n)$ assumed to arise from a probability model $f(\mathbf{x}; \boldsymbol{\psi})$ indexed by an unknown parameter (vector) $\boldsymbol{\psi}$, the **likelihood function** is:

$$L(\boldsymbol{\psi}) = f(\mathbf{x}; \boldsymbol{\psi})$$

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



A two-parameter likelihood surface, $L(\theta, \phi)$, plotted as a function of two unknown parameters. The maximum likelihood estimate is the point $(\hat{\theta}, \hat{\phi})$ 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, $l(\psi) = \sum_i \log f(x_i; \psi)$, 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):

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

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

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

10.5: Asymptotic Properties of the MLE

Under standard regularity conditions, as the sample size $n \rightarrow \infty$, 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 $I(\psi_0)^{-1}$, so that for large n , $\hat{\psi}$ is approximately $\text{Normal}(\psi_0, I(\psi_0)^{-1})$.
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

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

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

Proof sketch. Expand the score equation $U(\hat{\psi}) = 0$ in a first-order Taylor series around the true value ψ_0 : $0 = U(\hat{\psi}) \approx U(\psi_0) + U'(\psi_0)(\hat{\psi} - \psi_0)$, 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, $U(\psi_0)/\sqrt{n} \xrightarrow{d} N(0, I_1(\psi_0))$. Similarly, $-U'(\psi_0)/n \rightarrow I_1(\psi_0)$ 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:

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

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 $U(\psi) = 0$ 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, $E[I(\psi)]$, 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:

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

Both algorithms are iterated to convergence (e.g., until $|\psi^{(t+1)} - \psi^{(t)}|$ is smaller than a chosen tolerance), and both, at convergence, additionally supply an estimate of $\text{Var}(\hat{\psi})$ 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 $H_0: \psi \in \Psi_0$ (a restricted parameter space, e.g., $\theta = 0.5$ for “no linkage,” Chapter 8) against the general alternative $H_1: \psi \in \Psi$, the **likelihood ratio test (LRT)** statistic is:

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

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

Proof. Let $\phi = g(\theta)$, so $\theta = g^{-1}(\phi)$ since g is one-to-one. The likelihood function expressed in terms of ϕ 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 ϕ is equivalent to maximising $L(\theta)$ over all valid $\theta = g^{-1}(\phi)$: any value of θ 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 $\hat{\theta}$ maximises $L(\theta)$ by definition, $\hat{\phi} = g(\hat{\theta})$ must maximise $L^*(\phi)$, so the MLE of $g(\theta)$ is $g(\hat{\theta})$. 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.

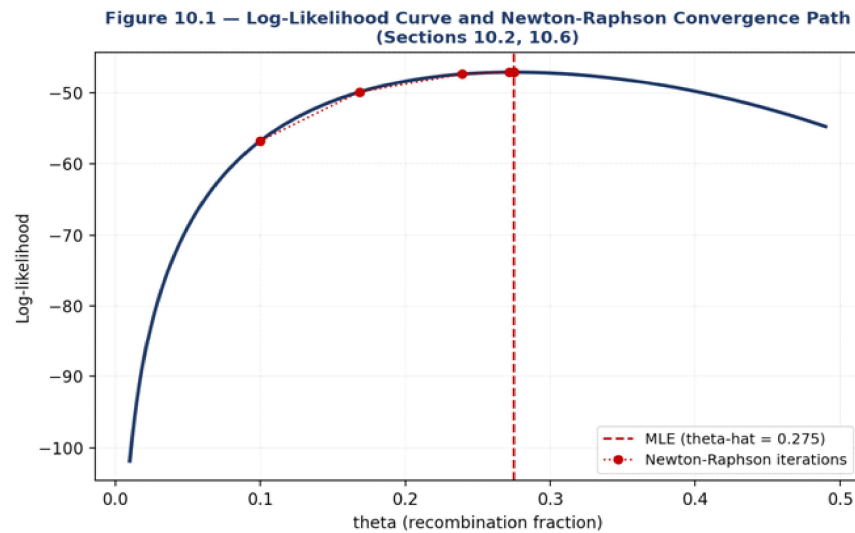


Figure 10.1: *Log-likelihood curve and Newton-Raphson convergence path (Sections 10.2, 10.6).*

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 in 1: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 _ in range(maxit):
```

```
    p_new = p + score(p) / info(p)
```

```
if abs(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 p_0 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.