

Math 4261: Mathematical Statistics I

Lecture 11: Maximum likelihood estimation
(References: Rice 8.5; DS 7.5; Wasserman 9.3)

Fall 2026

Outline

Recap

Maximum likelihood estimators

Examples with i.i.d. data

Example with dependent data

Outline

Recap

Maximum likelihood estimators

Examples with i.i.d. data

Example with dependent data

Recap: Parameter estimation

Parameter estimation

Suppose data $X_1, \dots, X_n$ came from joint distribution with PMF/PDF parametrized by $\theta \in \Omega \subseteq \mathbb{R}^k$:

$$f(x_1, \dots, x_n | \theta)$$

- ▶ Goal: estimate θ . Ω is the parameter space.
- ▶ If data i.i.d. from $f(x|\theta)$, joint distribution is $\prod_{i=1}^n f(x_i|\theta)$.

An **estimator** $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ is a function of the data.

1. $\hat{\theta}$ is a random variable.
2. **Unbiased**: $\mathbb{E}_{\theta}[\hat{\theta}] = \theta$ for every $\theta \in \Omega$.
3. **Consistent**: $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \rightarrow \infty$.

Outline

Recap

Maximum likelihood estimators

Examples with i.i.d. data

Example with dependent data

Definitions

Suppose $X_1, \dots, X_n$ have joint distribution $f(x_1, \dots, x_n | \theta)$.

The **likelihood function** is

$$\text{lik}(\theta) = f(X_1, \dots, X_n | \theta)$$

Often more convenient to work with the **log-likelihood**

$$\ell_n(\theta) = \log \text{lik}(\theta) = \log f(X_1, \dots, X_n | \theta).$$

When data are **i.i.d.** $f(x | \theta)$

$$\text{lik}(\theta) = \prod_{i=1}^n f(X_i | \theta), \quad \text{and} \quad \ell_n(\theta) = \sum_{i=1}^n \log f(X_i | \theta)$$

MLE procedure

The **maximum likelihood estimator** is

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Omega} \ell_n(\theta)$$

- ▶ MLE $\hat{\theta}$ is the value of θ that **maximizes the joint PMF/PDF at the observed data**.
- ▶ Equivalent to maximize $\operatorname{lik}(\theta)$. But log-likelihood $\ell_n(\theta)$ typically more **convenient**.
- ▶ Equivalently, for i.i.d. data, the MLE **minimizes**

$$\frac{1}{n} \sum_{i=1}^n -\log f(X_i \mid \theta)$$

This is an instance of **empirical risk minimization**: minimizing the average of the loss function $-\log f(x \mid \theta)$ over the data.

- ▶ The MLE $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ is a random variable. Having observed data $X_1 = x_1, \dots, X_n = x_n$, we derive a **concrete estimate** $\hat{\theta}(x_1, \dots, x_n)$.

Outline

Recap

Maximum likelihood estimators

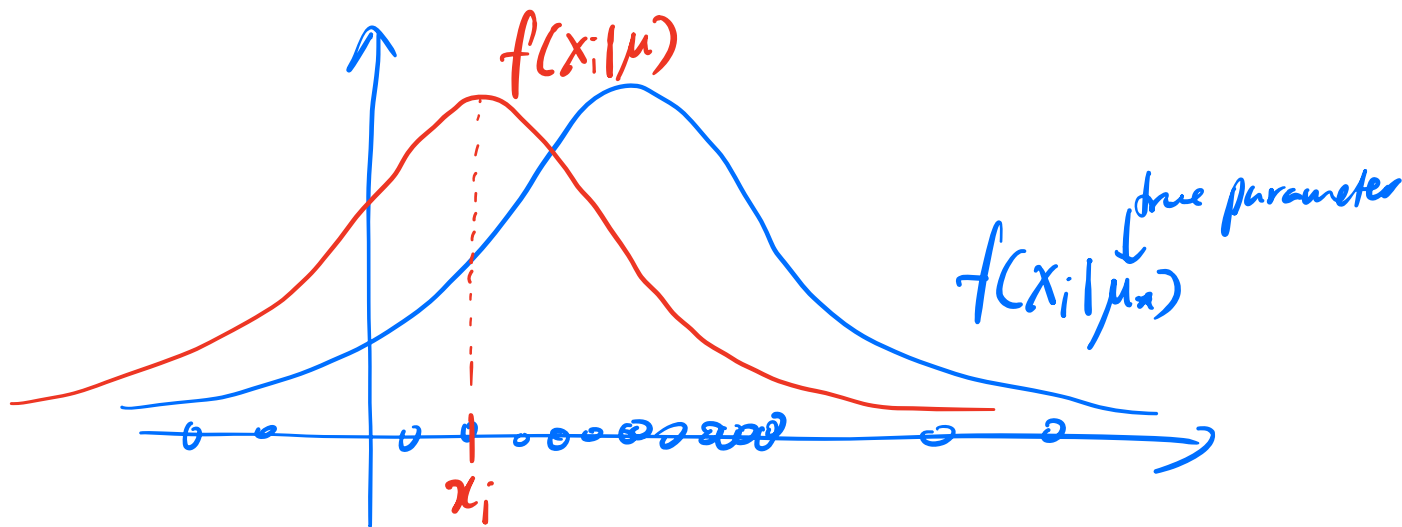
Examples with i.i.d. data

Example with dependent data

Example: Gaussian location model

Suppose $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where $\sigma > 0$ is **known**. Find the MLE.

$$\begin{aligned} \text{Likelihood: } \text{Lik}(\mu) &= \prod_{i=1}^n f(x_i | \mu) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right\} \end{aligned}$$



Example: Gaussian location model

$$\text{Lik}(\mu) = \underbrace{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n}_{\text{pos. factor}} \underbrace{\exp}_{\text{increasing func.}} \left\{ \underbrace{-\frac{1}{2\sigma^2}}_{\text{negative factor}} \sum_{i=1}^n (X_i - \mu)^2 \right\}$$

Hence

$$\underset{\mu}{\operatorname{argmax}} \text{Lik}(\mu) = \underset{\mu}{\operatorname{argmin}} \underbrace{\sum_{i=1}^n (X_i - \mu)^2}_{g(\mu)}$$

$$g'(\mu) = -2 \sum_{i=1}^n (X_i - \mu) \stackrel{\text{set}}{=} 0$$

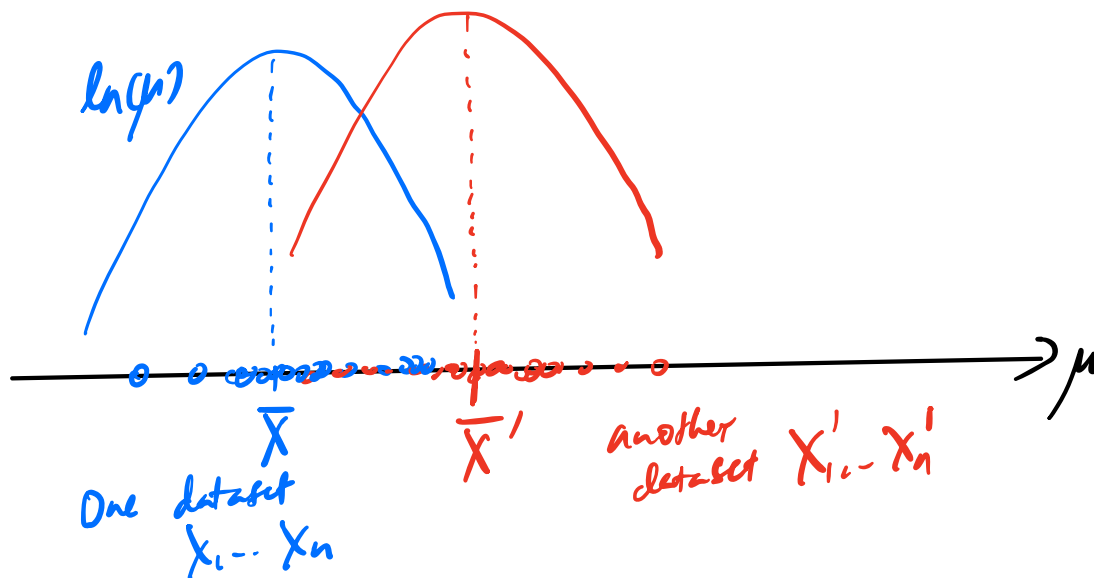
$$\Rightarrow \text{MLE} \quad \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \leftarrow \begin{array}{|l} \text{This doesn't} \\ \text{depend on} \\ \sigma \end{array}$$

[same conclusion if take $\log$ and $\underset{\mu}{\operatorname{argmax}} \ln(\mu)$]

Example: Gaussian location model

$$\ell_n(\mu) = C_{n,\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

Different data lead to a different log-likelihood surface



Example: Bernoulli

Suppose $X_1, \dots, X_n$ are i.i.d. **Bernoulli(p)**, $p \in [0, 1]$. Find the MLE $\hat{p}$.

$P[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid p] = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$

Handwritten notes:

- $x_i \in \{0, 1\}$ (blue arrow pointing to x_i)
- "parametrized by p " (green arrow pointing to p)
- $p^{x_i} (1-p)^{1-x_i} = \begin{cases} p & \text{if } x_i = 1 \\ 1-p & \text{if } x_i = 0 \end{cases}$ (purple text and circle around the product term)

Likelihood: $lik(p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$

log likelihood $\ln(p) = \log lik(p) = \sum_{i=1}^n x_i \log p + (1-x_i) \log(1-p)$

$$\ln'(p) = \sum_{i=1}^n \left[\frac{x_i}{p} - \frac{1-x_i}{1-p} \right] \stackrel{\text{set}}{=} 0$$

Solving MLE $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X}$

Example: Poisson counts

Suppose $X_1, \dots, X_n$ are i.i.d. $\text{Poisson}(\lambda)$. Find the MLE $\hat{\lambda}$.

$$P[X_i = x_i | \lambda] = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

Going straight to log-likelihood:

$$\begin{aligned} \ln(\lambda) &= \sum_{i=1}^n \log \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \sum_{i=1}^n -\lambda + x_i \log \lambda - \log x_i! \\ &= (\log \lambda) \cdot \sum_{i=1}^n x_i - n\lambda + \text{Const} = \sum_{i=1}^n \log x_i! \end{aligned}$$

does not depend on λ

$$\ln'(\lambda) = \frac{\sum x_i}{\lambda} - n \stackrel{\text{Set}}{=} 0$$

$$\text{Deduce MLE } \hat{\lambda} = \frac{\sum x_i}{n} = \bar{X}$$

Example: Gaussian location and scale model

Suppose $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, where both parameters μ and σ^2 are to be estimated. For nonconstant observations, the MLEs are

$$\hat{\mu} = \overline{X}$$
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2$$

- ▶ Derivation hint: the MLE $\hat{\mu}$ does not depend on σ , so it is the same over all σ .
- ▶ MoM and MLE coincide in this case.

Example: Gamma

Suppose $X_1, \dots, X_n$ are i.i.d. **Gamma**(α, λ), $\alpha, \lambda > 0$, which has density

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad 0 < x < \infty.$$

Find the MLEs $\hat{\alpha}$ and $\hat{\lambda}$.

$$\begin{aligned} \text{log-likelihood: } \ln(\alpha, \lambda) &= \sum_{i=1}^n \ln \frac{\lambda^\alpha x_i^{\alpha-1} e^{-\lambda x_i}}{\Gamma(\alpha)} \\ &= \sum_{i=1}^n \alpha \ln \lambda + (\alpha-1) \ln x_i - \lambda x_i - \ln \Gamma(\alpha) \end{aligned}$$

Partial derivatives

$$\textcircled{1} \quad \frac{\partial \ln}{\partial \alpha} = n \ln \lambda + \sum_{i=1}^n \ln x_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \stackrel{\text{set}}{=} 0$$

$$\textcircled{2} \quad \frac{\partial \ln}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^n x_i \stackrel{\text{set}}{=} 0 \iff \boxed{\hat{\lambda} = \frac{2}{\bar{X}}}$$

$$\text{Sub } \hat{\lambda} \text{ into } \textcircled{1}: \boxed{n \ln \hat{\alpha} - n \ln \bar{X} + \sum_{i=1}^n \ln x_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0}$$

Example: Gamma

MLEs for i.i.d. **Gamma**(α, λ): $(\hat{\alpha}, \hat{\lambda}) = (\hat{\alpha}, \hat{\alpha}/\bar{X})$, where $\hat{\alpha}$ satisfies

$$n \log \hat{\alpha} - n \log \bar{X} + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0$$

- ▶ Differs from MoM estimates, which were in closed form.
- ▶ An iterative root-finding procedure can be used. E.g. secant method and method of bisection (not tested in this class). Use MoM estimators as initializers.

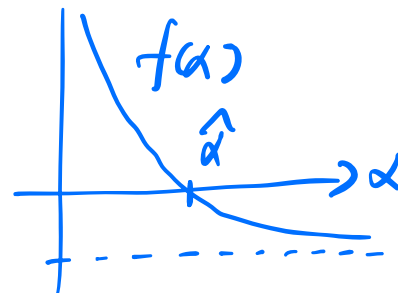
Example: Gamma

MLEs for i.i.d. $\text{Gamma}(\alpha, \lambda)$: $(\hat{\alpha}, \hat{\lambda}) = (\hat{\alpha}, \hat{\alpha}/\bar{X})$, where $\hat{\alpha}$ satisfies $f(\hat{\alpha}) = 0$ where

$$f(\alpha) = \log \alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \bar{X} + \frac{1}{n} \sum_{i=1}^n \log X_i.$$

One may check that

- (i) f is decreasing over $\alpha \in (0, \infty)$
- (ii) $\lim_{\alpha \rightarrow 0} f(\alpha) = \infty$
- (iii) $\lim_{\alpha \rightarrow \infty} f(\alpha) = -\log \bar{X} + \frac{1}{n} \sum_{i=1}^n \log X_i$



Jensen's inequality $\frac{1}{n} \sum_{i=1}^n \log X_i < \log \sum_{i=1}^n X_i = \log \bar{X}$

So $-\log \bar{X} + \frac{1}{n} \sum_{i=1}^n \log X_i < 0$ for all large α

Hence $f(\alpha) = 0$ has unique solution $\hat{\alpha}$.

Example: Gamma

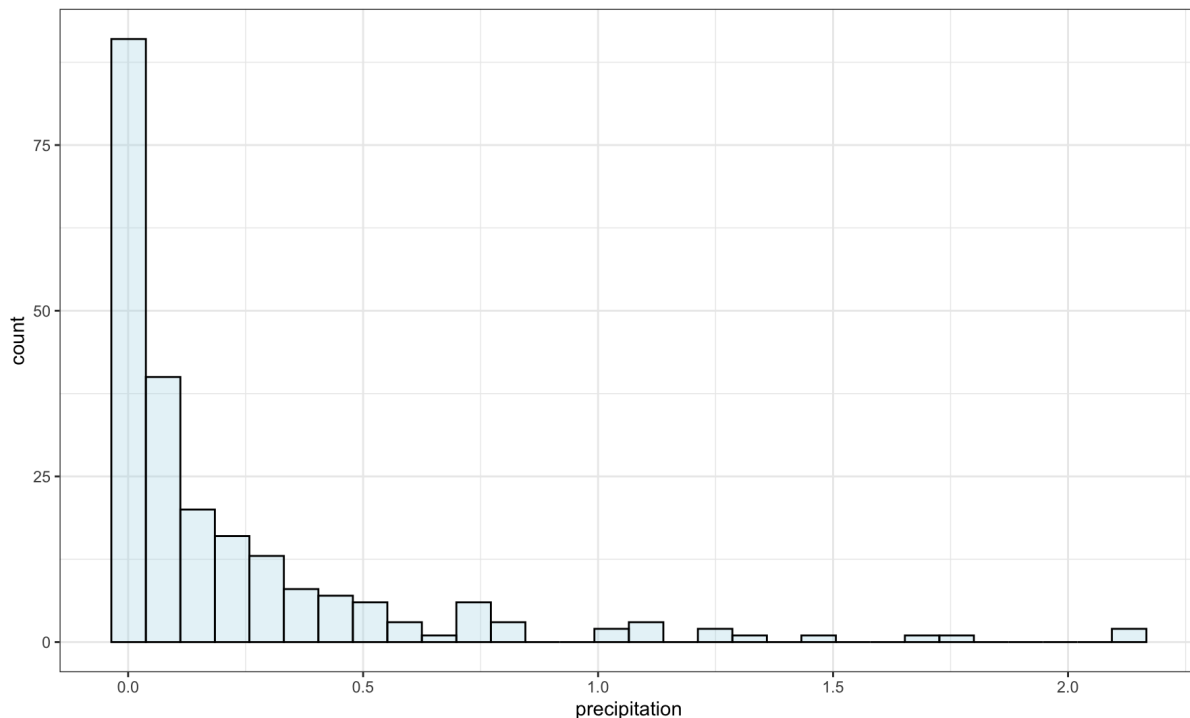
In Lecture 9 (and Midterm 1), we showed that for i.i.d. $\text{Gamma}(\alpha, \lambda)$ data, the **MoM estimators** are in *closed form*

$$\hat{\alpha}_{\text{MoM}} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}, \quad \hat{\lambda}_{\text{MoM}} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

Here $\hat{\mu}_1 = \overline{X}$ and $\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$.

Example: Rainfall in Illinois by Gamma modeling

In a study of the natural variability of rainfall, the **rainfall of summer storms** was measured by a network of rain gauges in southern Illinois for the years 1960–1964 (Changnon and Huff, Le Cam and Neyman 1967).

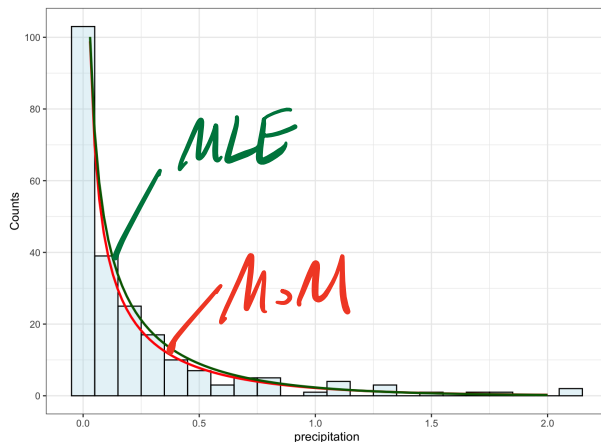


Example: Rainfall in Illinois

MLE estimates are

$$\hat{\alpha}_{\text{MLE}} = 0.441, \quad \hat{\lambda}_{\text{MLE}} = 1.96$$

Differs from MoM estimates ($\hat{\alpha}_{\text{MoM}} = 0.375$ and $\hat{\lambda}_{\text{MoM}} = 1.674$).



*Bootstrap estimate*¹ (size 1000) shows that the estimated standard errors are $s_{\hat{\alpha}} = 0.03$ and $s_{\hat{\lambda}} = 0.26$ which are better than MoM's 0.06 and 0.34. MLE is expected to be more precise.

¹We haven't covered this yet.

Outline

Recap

Maximum likelihood estimators

Examples with i.i.d. data

Example with dependent data

Example: Multinomial

Here is a situation where the data are *not i.i.d.*

Suppose we run n trials, where each trial can have m different outcomes. The j^{th} outcome has probability p_j of occurring. Let Y_j count the number of trials with j^{th} outcome. Then

$$(Y_1, \dots, Y_m) \sim \text{Mult}(n, \{p_1, \dots, p_m\}).$$

- ▶ **Multinomial distribution** with n (independent) trials and m possible outcomes (categories).
- ▶ $\{p_1, \dots, p_m\}$ is PMF associated with each trial: $\sum_{j=1}^m p_j = 1$
- ▶ $m = 2$ case: $Y_1 \sim \text{Bin}(n, p_1)$ and $Y_2 = n - Y_1$
- ▶ **Marginally**, each Y_j is $\text{Bin}(n, p_j)$
- ▶ $Y_1, \dots, Y_m$ **not** independent.
(in Binomial, # heads is not independent of # tails).

Example: Multinomial

$n=6, M=3 : A \text{ or } B \text{ or } C$ 

$$\mathbb{P}[Y_A=3, Y_B=1, Y_C=2] = \frac{6!}{3!1!2!} p_A^3 p_B^1 p_C^2$$

The **PMF** of $\text{Mult}(n, \{p_1, \dots, p_m\})$, i.e. the joint distribution of $Y_1, \dots, Y_m$ is

$$\begin{aligned} f(y_1, \dots, y_m | p_1, \dots, p_m) &= \mathbb{P}[Y_1 = y_1, \dots, Y_m = y_m] \\ &= \frac{n!}{y_1! \cdots y_m!} \prod_{j=1}^m p_j^{y_j}, \end{aligned}$$

for nonnegative integers $y_1, \dots, y_m$ satisfying $\sum_{j=1}^m y_j = n$, and zero otherwise.

Example: Multinomial

To find the MLE $\hat{p}_1, \dots, \hat{p}_m$, we write the log-likelihood

$$\ell_n(p_1, \dots, p_m) = \log n! - \sum_{j=1}^m \log Y_j! + \sum_{j=1}^m Y_j \log p_j$$

Maximize this over all $\{p_1, \dots, p_m\}$ that **sum to one**. Incorporate this constraint by introducing Lagrange multiplier λ

$$\mathcal{L}(p_1, \dots, p_m, \lambda) = \text{Const.}(n, Y_1, \dots, Y_m) + \sum_{j=1}^m Y_j \log p_j + \lambda \left(\sum_{j=1}^m p_j - 1 \right)$$

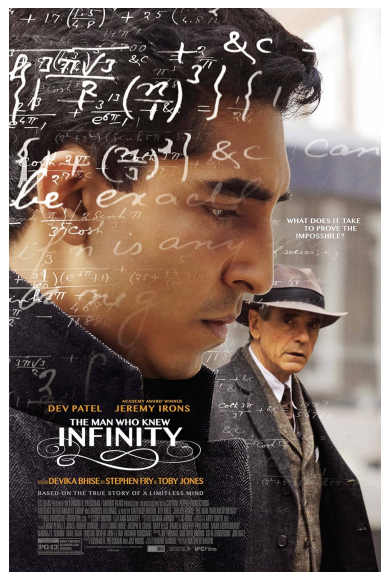
Setting all partials to zero and solving, we have the MLEs

$$\boxed{\hat{p}_j = \frac{Y_j}{n}}$$

Example: Hardy–Weinberg equilibrium

‘I have never done anything “useful”. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world. ’

G. H. Hardy (1940)



Example: Hardy–Weinberg equilibrium

In genetics, it is often assumed that the genotypes AA , Aa , and aa at a single locus satisfy *Hardy–Weinberg equilibrium*: they occur with probabilities $(1 - \theta)^2$, $2\theta(1 - \theta)$ and θ^2 , where $\theta \in [0, 1]$ represents the frequency of the recessive allele. Thus the counts of these three genotypes in n samples may be modeled as

$$(Y_{AA}, Y_{Aa}, Y_{aa}) \sim \text{Mult}\left(n, \underbrace{\{(1 - \theta)^2, 2\theta(1 - \theta), \theta^2\}}_{\text{sums to 1}}\right).$$

	A $1-\theta$	a θ
A	$(1-\theta)^2$	$\theta(1-\theta)$
a	$\theta(1-\theta)$	θ^2

Example

A dataset of bloodtypes showing erythrocyte antigens M and N from a population in Hong Kong (1937):

Blood type				
	M	MN	N	Total
Count	342	500	187	1029
	y_1	y_2	y_3	n

Model: the probability of a person having M , MN , and N genotype is $(1 - \theta)^2$, $2\theta(1 - \theta)$, and θ^2 respectively.

Find the MLE $\hat{\theta}$.

$$(y_1, y_2, y_3) \sim \text{Mult}(n, \{(1-\theta)^2, 2\theta(1-\theta), \theta^2\})$$

(1) Write down the **log-likelihood**.

$$\begin{aligned} \text{Likelihood } P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 | \theta] \\ = \frac{n!}{y_1! y_2! y_3!} (1-\theta)^{2y_1} \cdot (2\theta(1-\theta))^{y_2} \theta^{2y_3} \end{aligned}$$

log-likelihood :

$$\ln(\theta) = \underbrace{\text{const.}}_{\text{does not depend on } \theta} + 2y_1 \ln(1-\theta) + y_2 \ln[2\theta(1-\theta)] + 2y_3 \ln \theta$$

Example

(2) Differentiate log-likelihood and set to zero

$$\ln(\theta) = \text{const.} + (2Y_1 + Y_2) \ln(1-\theta) + (Y_2 + 2Y_3) \ln \theta$$

$$\ln'(\theta) = -\frac{2Y_1 + Y_2}{1-\theta} + \frac{Y_2 + 2Y_3}{\theta} \stackrel{\text{Set}}{=} 0$$

$$\Rightarrow \text{MLE } \hat{\theta} = \frac{2Y_3 + Y_2}{2n}$$

Example: Hardy–Weinberg equilibrium

The **MLE**

$$\hat{\theta} = \frac{2Y_3 + Y_2}{2Y_1 + 2Y_2 + 2Y_3} = \frac{2Y_3 + Y_2}{2n}$$

- ▶ Plugging in the actual data gives $\hat{\theta} = 0.4247$
- ▶ This gives expected cell counts

$$(M, MN, N) = (340.6, 502.8, 185.6)$$

