# ST2132 Parameter Estimation II

Maximum likelihood

Semester 1 2022/2023

# Introduction

- Recall the estimation problem:

  Assuming data $x_1, \ldots, x_n$ are realisations of IID RVs $X_1, \ldots, X_n$ with density/mass $f(x|\theta)$. The parameter $\theta$ lies in a parameter space $\Theta \subset \mathbb{R}^p$.

  The task is to estimate $\theta$, estimate the SE and bias, and perhaps construct a CI for $\theta$.

- Parameter Estimation I introduces the MOM estimator. Bootstrap is needed to approximate the SE, as before. Monte Carlo approximate is also necessary if there is no formula for the SE.

- Now we learn another general method: maximum likelihood.

# Poisson likelihood

- Let $x_1, \ldots, x_n$ be realisations of IID Poisson($\lambda$) RV's $X_1, \ldots, X_n$.

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots$$

The joint probability of data is

$$f(x_1|\lambda) \times \cdots \times f(x_n|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{x_1! \cdots x_n!}$$

- The likelihood is the joint probability, but only depends on $\lambda$:

$$L(\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{x_1! \cdots x_n!}$$

What is the domain of the likelihood?

# Normal likelihood

- Let $x_1, \ldots, x_n$ be realisations of IID $\mathrm{N}(\mu, \sigma^2)$ RV's $X_1, \ldots, X_n$.

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The joint density at the data is

$$f(x_1|\mu, \sigma) \cdots f(x_n|\mu, \sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

- What is the domain of the likelihood function $L(\mu, \sigma)$?

# Likelihood function

Let $x_1, \ldots, x_n$ be realisations of IID RV's $X_1, \ldots, X_n$ with density/mass function $f(x|\theta)$, $\theta \in \Theta \subset \mathbb{R}^p$.

- The likelihood function is a product of density/mass terms

$$\begin{aligned} L(\theta) &= f(x_1|\theta) f(x_2|\theta) \cdots f(x_n|\theta) \\ &= \prod_{i=1}^{n} f(x_i|\theta) \end{aligned}$$

- $L$ maps $\Theta$ to $\mathbb{R}_+$. The dependence on data is suppressed in the notation.

# Loglikelihood function

The loglikelihood function is

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

i.e., a sum of log density/mass terms.

Poisson:

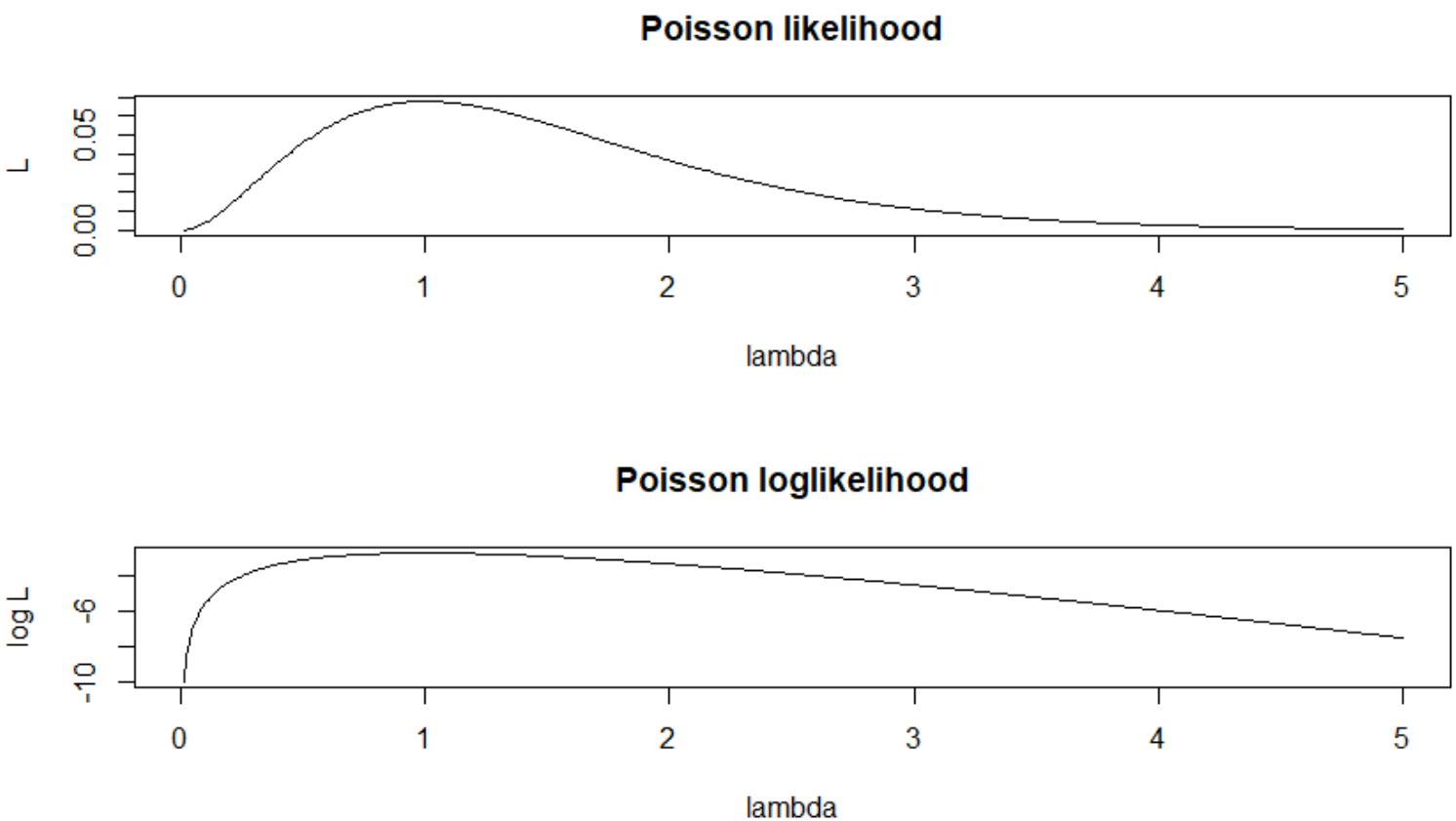
$$\ell(\lambda) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

Normal:

$$\ell(\mu, \sigma) = -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

# Poisson example: $x_1 = 0, x_2 = 2$

Write down $L$ and $\ell$.



# Maximising Poisson likelihood

What can you say about the value of $\lambda$ which maximises $L$ (based on data $x_1, \ldots, x_n$), and that which maximises $\ell = \log L$?

Find the maximiser of $L$.

# Maximising normal likelihood

Find the value of $(\mu, \sigma)$ that maximises the normal likelihood.

# Maximum likelihood estimation

- The maximiser of $L$ is the **maximum likelihood estimate** of $\theta$. It is a function of the data $x_1, \ldots, x_n$, hence is a realisation of the **maximum likelihood estimator** $\hat{\theta}$.

- Write down the ML estimate and estimator of
(a) $\lambda$ for Poisson

(b) $(\mu, \sigma)$ for normal

# Deriving ML estimate

Often, $\ell(\theta)$ is easier to maximise than $L(\theta)$. $\ell(\theta)$ can be quickly obtained by summing log terms.

- Logarithm of the Poisson($\lambda$) density:

- Loglikelihood:

- Obtain the ML estimate of $\lambda$.

# Refined definitions of $L$ and $\ell$

In statistical applications, the maximisers of $L$ and $\ell$ are of primary interest; the maximum values of the functions are seldom used.

Refinement: constant factors in $L(\theta)$ may be left out. Likewise, additive constants in $\ell(\theta)$ may be left out.

Poisson: instead of

$$L(\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{x_1! \cdots x_n!}$$

we may define

$$L(\lambda) = \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}$$

Instead of $\ell(\lambda) = \sum_{i=1}^n x_i \log \lambda - n\lambda - \sum_{i=1}^n \log(x_i!)$, we may define

$$\ell(\lambda) = \sum_{i=1}^n x_i \log \lambda - n\lambda$$

# Gamma distribution

- Let $x_1, \ldots, x_n$ be realisations of IID Gamma$(\alpha, \lambda)$ RV's $X_1, \ldots, X_n$.
- Logarithm of density:
- Loglikelihood:

# ML estimation

- The ML estimates of $(\alpha, \lambda)$ satisfy

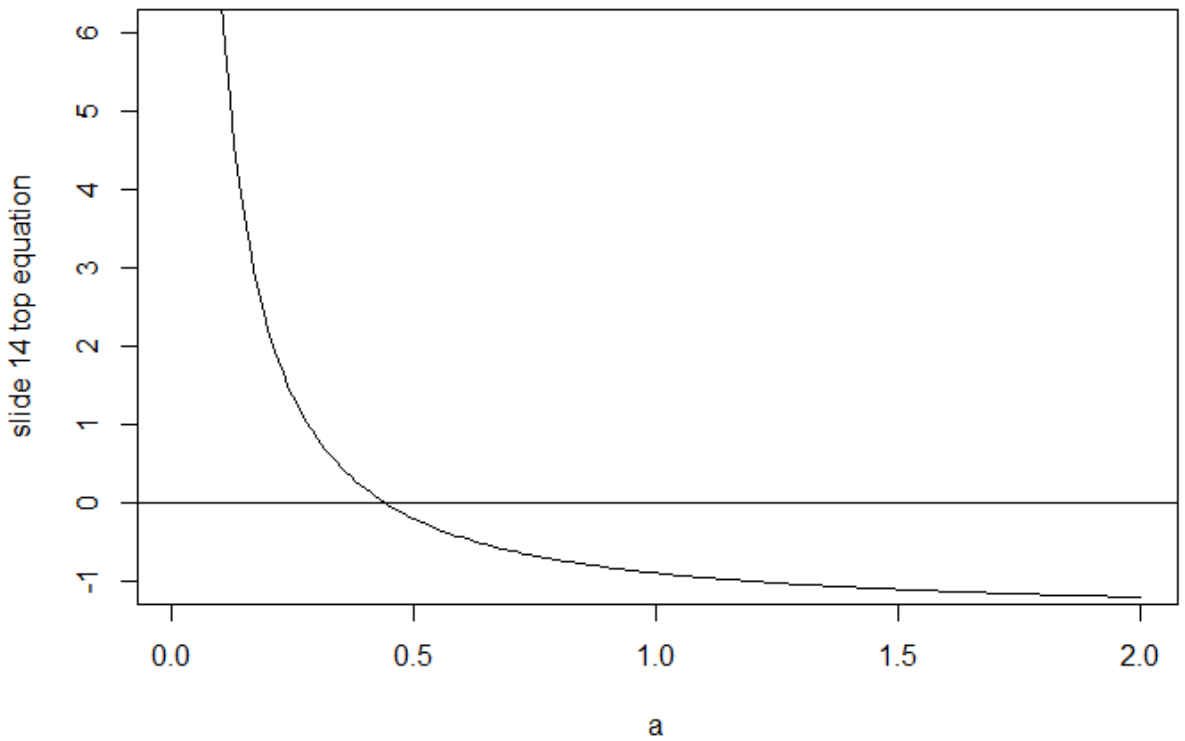
$$\log\left(\frac{\alpha}{\bar{x}}\right) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \bar{y} = 0, \qquad \lambda = \frac{\alpha}{\bar{x}}$$

  where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$. A numerical method is needed to estimate $\alpha$.

- The ML estimators $(\hat{\alpha}, \hat{\lambda})$ satisfy

# Revisiting rainfall

- For the 227 amounts of rainfall, the mean is $\bar{x} = 0.2244$ and the mean of log is $\bar{y} = 2.9642$.




R function uniroot(): ML estimate of $\alpha$ is about 0.44, so ML estimate of $\lambda$ is $0.44/0.2244 \approx 1.96$.

# What next?

- How large is the random error in the ML estimates?
- How large is the systematic error in the ML estimates?
- How to answer these questions?

# SE and bias: bootstrap

- Bootstrap approximation:

  SE in 0.44 is $\mathrm{SD}(\hat{\alpha}) \approx \mathrm{SD}(\hat{\alpha}^*)$.

  Bias in 0.44 is $\mathrm{E}(\hat{\alpha}) - \alpha \approx \mathrm{E}(\hat{\alpha}^*) - \alpha^*$.

  SE in 1.96 is $\mathrm{SD}(\hat{\lambda}) \approx \mathrm{SD}(\hat{\lambda}^*)$.

  Bias in 1.96 is $\mathrm{E}(\hat{\lambda}) - \lambda \approx \mathrm{E}(\hat{\lambda}^*) - \lambda^*$.

- $\hat{\alpha}^*$ and $\hat{\lambda}^*$ are based on IID RV's $X_1^*, \ldots, X_{227}^*$, with what distribution?

# SE and bias: Monte Carlo

- Monte Carlo approximation:

  $\mathrm{SD}(\hat{\alpha}^*) \approx 0.03.$

  $\mathrm{E}(\hat{\alpha}^*) - \alpha^* \approx 0.00.$

  $\mathrm{SD}(\hat{\lambda}^*) \approx 0.26.$

  $\mathrm{E}(\hat{\lambda}^*) - \lambda^* \approx 0.04.$

- Conclusion:

  $\alpha$ is around $0.44 \pm 0.03$. The bias is negligible.

  $\lambda$ is around $1.96 \pm 0.26$. A less biased estimate is $1.96 - 0.04 = 1.92$.

# Monte Carlo

Let $X_1^*, \ldots, X_{227}^*$ be IID Gamma$(\alpha^*, \lambda^*)$ random variables.
$\alpha^* = 0.44, \lambda^* = 1.96$.

(1) Generate realisations $x_1^*, \ldots, x_{277}^*$.

(2) Calculate a realisation each from $\hat{\alpha}^*$ and $\hat{\lambda}^*$.

(3) Repeat (1) and (2) to get many realisations.

(4) Estimate the expectations and SD's of $\hat{\alpha}^*$ and $\hat{\lambda}^*$.

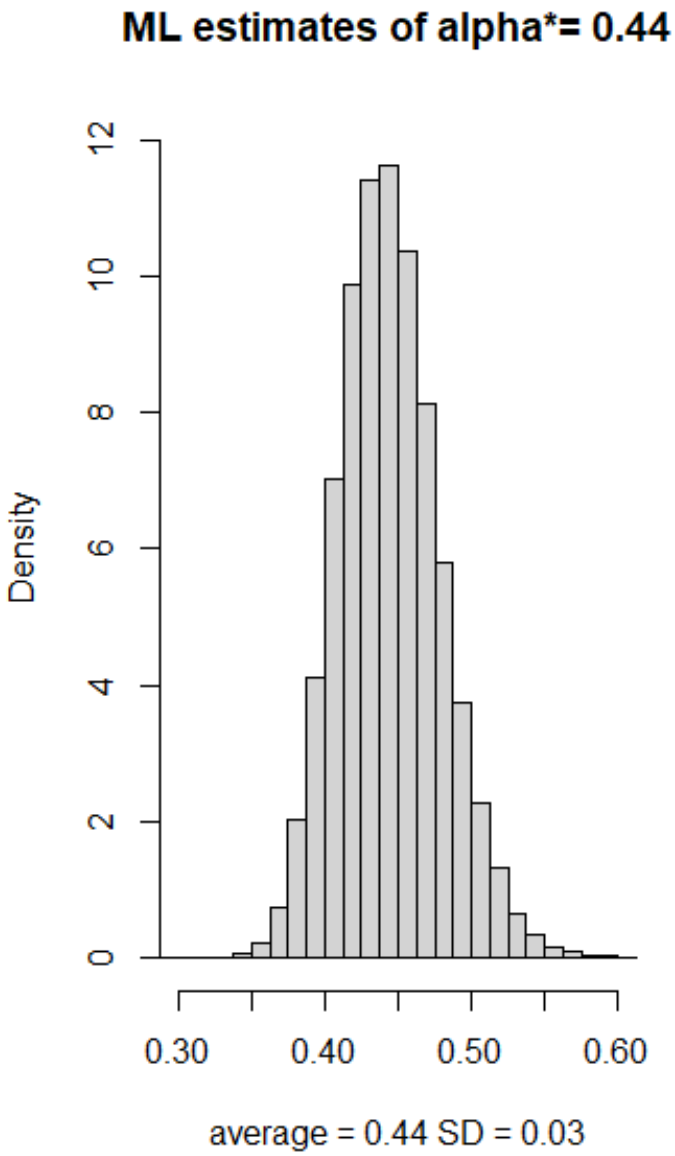
# Gamma data: MLE vs MOM

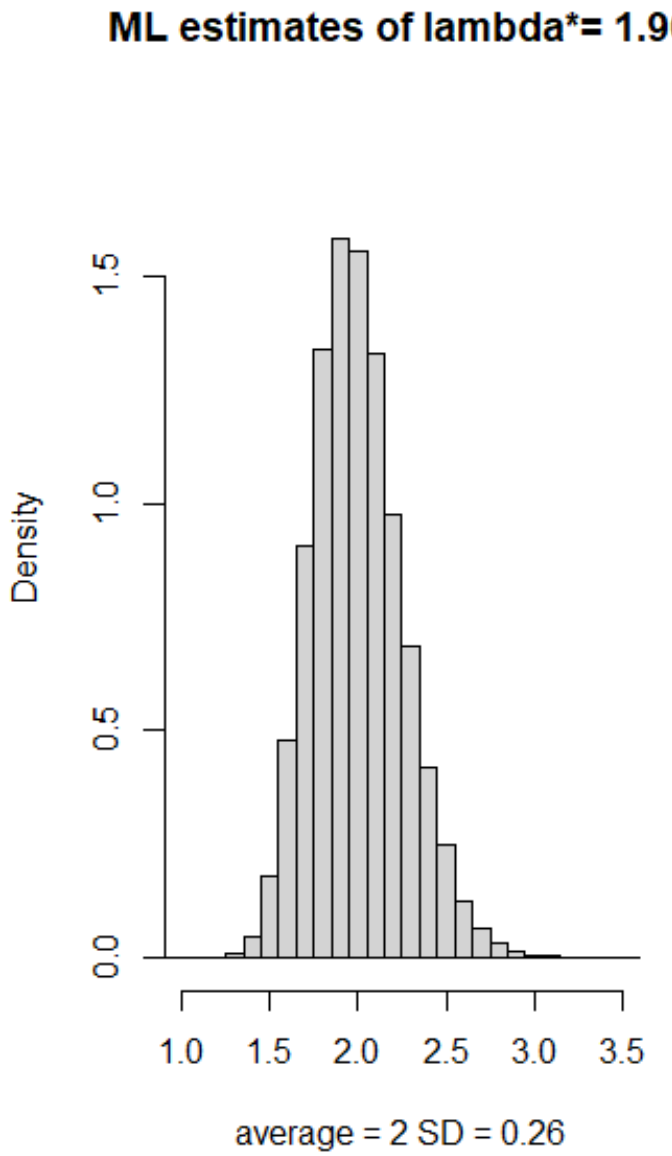
- Like MOM, the bias and SE of the ML estimates are approximated using bootstrap, and Monte Carlo. Unlike MOM, the ML estimates cannot be written down as formulae.
- ML beats MOM on both SE and bias.

| | *ML* | | | *MOM* | | |
|---|---|---|---|---|---|---|
| *Parameter* | *Est.* | *SE* | *Bias* | *Est.* | *SE* | *Bias* |
| $\alpha$ | 0.44 | 0.03 | 0.00 | 0.38 | 0.06 | 0.02 |
| $\lambda$ | 1.96 | 0.26 | 0.04 | 1.67 | 0.35 | 0.10 |

- ML estimates are functions of $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$ (slide 14). MOM never uses $\bar{y}$.

# Distribution of gamma MLE



# Multinomial data

Let $(x_1, \ldots, x_r)$ be a realisation of
$(X_1, \ldots, X_r) \sim \text{Multinomial}(n, (p_1, \ldots, p_r))$.

Find the ML estimate and estimator of $p_1, \ldots, p_r$.

$$L(p_1, \ldots, p_r) \quad =$$

$$\ell(p_1, \ldots, p_r) \quad =$$

# Maximising $\ell$

- Since $p_1 + \cdots + p_r = 1$, differentiating $\ell$ does not work.
- Instead, define the Lagrangian function

  $$\mathcal{L}(p_1, \ldots, p_r, \lambda) = x_1 \log p_1 + \cdots + x_r \log p_r + \lambda(p_1 + \cdots + p_r - 1)$$

  Now find the maximum of $\mathcal{L}$, treating $p_1, \ldots, p_r, \lambda$ as if they are unconstrained.
- $$\frac{\partial \mathcal{L}}{\partial p_i} = \qquad\qquad , \qquad i = 1, \ldots, r$$

  $$\frac{\partial \mathcal{L}}{\partial \lambda} =$$

# Estimates and estimators

- For $i = 1, \ldots, r$, the ML estimate of $p_i$ is _____.
- The ML estimator of $p_i$ is _____.
- The SE of the estimate is _____, estimated by _____.
- Do the above look familiar?

# Genetics

- Chromosomes often come in pairs, one from each parent. The main ingredient of a chromosome is the DNA molecule, which may be viewed as a sequence consisting of bases: A, C, G, T.
- DNA sequences very long. A locus is a subsequence on a chromosome. Alleles are different versions of bases at a locus. An unordered pair of alleles is called a genotype.
- The *ABO* locus on chromosome 9 has three alleles: $a$, $b$ and $o$. Blood types are determined as follows:

| *Genotype* | $aa, ao$ | $bb, bo$ | $ab$ | $oo$ |
|---|---|---|---|---|
| *Blood type* | *A* | *B* | *AB* | *O* |

# Gregor Mendel

- Mendel discovered the laws of inheritance:
  (1) The maternal allele is randomly chosen from her two alleles; similarly for the paternal allele.
  (2) The two choices are independent.

- Suppose the father has genotype *oo*, the mother has genotype *ao*. What is the probability that their child has blood type *A*?

# Genotypic vs allelic proportions

- Suppose in a population there are $k$ alleles at a locus: $a_1, \ldots, a_k$. How many possible genotypes are there?

- Let $k = 2$. Calculate the genotype and allele proportions in these populations:
  (1) $(a_1, a_1)$, $(a_1, a_2)$,$(a_1, a_2)$, $(a_2, a_2)$

  (2) $(a_1, a_1)$, $(a_1, a_1)$,$(a_2, a_2)$, $(a_2, a_2)$

- Given the genotype proportions, can you calculate the allele proportions?

- Given the allele proportions, can you calculate the genotype proportions?

# Hardy-Weinberg equilibrium

- A population is in HWE at a locus if the genotype proportions are

$$f(a_i a_j) = \begin{cases} p_i^2 & i = j \\ 2p_i p_j & i \neq j \end{cases}$$

where $p_i$ is the proportion of allele $a_i$.

- Suppose that in a large population, the allelic proportions among males are the same as those among females, and mating is random. Then the genotype proportions of the next generation are given by the above. These genotype proportions persist through subsequent generations from random mating.

G. H. Hardy (1908): "To The Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists...."

# HWE: 2 alleles

- Suppose that in a large population, a locus in HWE has two alleles $A$ and $a$, and the proportion of $a$ is $\theta$. Hence the genotype proportions are

$$AA : (1-\theta)^2, \qquad Aa : 2\theta(1-\theta), \qquad aa : \theta^2$$

- What is the distribution of the number of $a$ alleles in a randomly chosen individual from the population?

- How about the total number of $a$ alleles in an SRS of size $n$?

# 1937 Hong Kong study

- The sample frequencies were

$$AA : 342, \quad Aa : 500, \quad aa : 187$$

- Assume that data came from an SRS, and that HWE held. Then the frequencies are approximately realisations from

$$(X_1, X_2, X_3) \sim \text{Multinomial}(1029, ((1-\theta)^2, 2\theta(1-\theta), \theta^2)))$$

where $\theta$ is the population proportion of $a$.

# ML estimation

$$L(\theta) \quad =$$

$$\ell(\theta) \quad =$$

ML estimate of $\theta$ is

$$\frac{x_2 + 2x_3}{2n} = \frac{500 + 2 \times 187}{2 \times 1029} \approx 0.42$$

ML estimator is

$$\hat{\theta} = \frac{X_2 + 2X_3}{2n}$$

# SE estimation

- $X_2 + 2X_3$ is the number of $a$ alleles: Binomial$(2n, \theta)$.

$$\text{var}(\hat{\theta}) = \frac{\theta(1-\theta)}{2n}$$

SE in 0.42 is, by the bootstrap, approximately

$$\sqrt{\frac{0.42 \times 0.58}{2 \times 1029}} \approx 0.01$$

- Monte Carlo is unnecessary.
- How might MOM be applied here?

# Conclusion (1)

Estimation problem: Suppose $x_1, \ldots, x_n$ are realisations of IID RVs $X_1, \ldots, X_n$ with density $f(x|\theta)$, where the parameter $\theta$ lies $\Theta \subset \mathbb{R}$. Estimate $\theta$.

- MOM and ML use the same ideas as sampling survey: estimate, estimator, SE, bootstrap.
- Simplest case:

| Parameter | Estimator | SE | Monte Carlo | Bias |
|---|---|---|---|---|
| $\mu/p$ | $\bar{X}$ | $\sigma/\sqrt{n}$ | No | No |

- Gamma data:

| Method | Parameter | Estimator | SE | Monte Carlo | Bias |
|---|---|---|---|---|---|
| MOM | $\alpha$ | $\bar{X}^2/\hat{\sigma}^2$ | ? | Yes | Yes |
| ML | $\alpha$ | ? | ? | Yes | Yes |

# Conclusion (2)

- ML makes more sense in some problems, like the HWE, though it is numerically more demanding, like for Gamma data.
- Why use ML estimators? They are asymptotically unbiased and normally distributed in practically all models in applications. Hence large-sample CIs are feasible.
- Just because you write complex code to do Monte Carlo approximation of SE or bias of some sophisticated estimate, does not mean the randomness assumption is solid. If it is doubtful, then your estimate, SE, CI are dubious.