Mathematical Statistics II

Maximum Likelihood Estimation

Jesse Wheeler

Introduction

Overview

- The next approach we will discuss is Maximum Likelihood Estimation (MLE).
- As we will see, the MLE has several desirable properties, and as a result is often favored over approaches like the method of moments.
- The material for this section largely comes from Chapter 8.5 of Rice (2007), and various sections in Pawitan (2001).

Likelihood: an introduciton

What is likelihood?

- The term “likelihood” is often used colloquially to mean something analogous to probability. E.g., “What is the likelihood that it rains tomorrow?”
- When we use this term in statistics / mathematics, we mean something specific that isn’t the same thing as probability.
- The use of the term “likelihood” was first made by R. A. Fisher, who was the architect and primary proponent of “likelihood-based-inference”.
- We will start with the treatment of likelihood in the text “In all Likelihood” (Pawitan, 2001), which is a fantastic resource on the subject. (This will lead to some review...)

What is likelihood? II

Coin Flips

We will revisit this example, as it is a great starting point to connect with existing understanding.

Consider flipping a coin $N = 10$ times.

What is likelihood? III

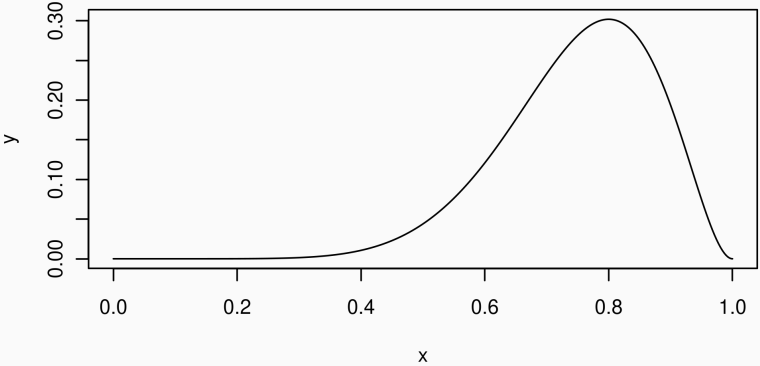
- For our specific coin-flipping example with $N = 10$, $X = 8$, the likelihood function is

$$L(\theta) = P_\theta(X = 8).$$

- This is plotted in the following way:

```r
x <- seq(1e-8, 1-1e-8, length.out = 1000)
y <- dbinom(8, 10, x)
plot(x = x, y = y, type = 'l')
```

What is likelihood? IV



What is likelihood? V

- From the figure, we see that p is unlikely to be less than 0.5,or greater than 0.95.
- Given the data alone (no prior), we should prefer a value somewhere in the middle of these values.
- We still have some uncertainty about the value of p, but the likelihood gives us a numerical way to compare values of θ. Stochastic uncertainty as a result of sampling is captured in the likelihood function $L(\theta)$.

What is likelihood? VI

- The likelihood is not a probability. Though it came from a probability, the likelihood function (a function of θ) does not satisfy the requirements to be a probability. In our previous example, we have:

$$\int_0^1 L(\theta)\,d\theta = 1/11 \neq 1.$$

- For discrete probability, the likelihood was continuous. Discrete likelihoods are possible, arising when we want to select from a list $\{\theta_1, \theta_2, \ldots\}$.

What is likelihood? VII

- The idea behind maximum likelihood estimation (MLE) is simple: our estimate is the value of θ that maximizes the likelihood function $L(\theta)$.
- The MLE is considered a *frequentist* approach. Why? It quantifies a maximum belief about a parameter, which is more Bayesian in nature than Frequentist.
- As we'll see later, the MLE has nice theoretical Frequentist *properties*, and as a result can be justified via the frequentist paradigm.
- Still, it has close connection to Bayesian estimation and interpretation. In fact, we'll discuss connections between the MLE and Bayesian statistics later.

What is likelihood? VIII

- Often, maximizing the likelihood directly is challenging, so we maximize the log-likelihood instead.
- Other times, the likelihood has to be maximized numerically.

MLE of coin toss problem

Suppose we have $N = 10$ total tosses, and n total heads. Find the MLE of p, the probability of heads.

Continuous models

- The interpretation of the likelihood function as the “the probability of the observed data x^*, considered as a function of θ” makes perfect sense in the discrete model case.
- For continuous models, the technical issue arises that the probability of any point value x is zero.
- We resolve the problem similar to what was done in Math 4450 and the John Rice text: approximate the probability by discretizing into small, discrete intervals:

$$x^* \in (x^* - \epsilon/2, x^* + \epsilon/2),$$

Continuous models II

thus, the probability of observing something ϵ-close to the data is:

$$L(\theta) = P_\theta\big(X \in (x^* - \epsilon/2, x^* + \epsilon/2)\big)$$
$$= \int_{x^*-\epsilon/2}^{x^*+\epsilon/2} f(x;\theta)\, d\theta \approx \epsilon f(x^*;\theta).$$

- Then, since the likelihood is only meaningful up to a constant (we will discuss likelihood ratios later), then this has the same behavior as $L(\theta) = f(x^*;\theta)$.
- There are more advanced approaches to this problem, but this simple argument justifies the use of the pdf of a continuous random variable as the likelihood $L(\theta)$.

Continuous models III

- **Going forward:** we once again will generalize a model $f(x;\theta)$ to mean either the pmf or pdf of a random variable. I will often say "density" as a blanket term, even if this corresponds to a pmf, not a density.
- Further, when we "integrate" a density, this means either:

$$\int f(x;\theta)\,dx, \quad \text{If continuous}$$

or

$$\sum_x f(x;\theta), \quad \text{If discrete.}$$

Joint Probabilities

Likelihood with multiple observations

- Often the data we observe is multi-dimensional, rather than summarized as a single observation.
- In this case, the likelihood θ is still determined via the joint model:

$$L(\theta) = f(x^*; \theta) = f_{X_{1:N}}(x_1^*, x_2^*, \ldots, x_N^*; \theta).$$

- We are mostly focused in this class in the case were the observations are independent, meaning the likelihood factors:

$$L(\theta) = \prod_{i=1}^{N} f_{X_i}(x_i^*; \theta).$$

Likelihood with multiple observations II

- We often further simplify this by assuming the data are identically distributed:

$$L(\theta) = \prod_{i=1}^{N} f_{X_1}(x_i^*; \theta).$$

- As we've seen, it's generally easier to maximize the log-likelihood. In the IID case:

$$\ell(\theta) = \log \prod_{i=1}^{N} f_{X_1}(x_i^*; \theta) = \sum_{i=1}^{n} \log f_{X_1}(x_i^*; \theta).$$

Examples

Examples of finding the MLE

Traffic data: Poisson Model

Returning to a motivating example, suppose we model traffic accidents in a given week as $X_1, X_2, \ldots, X_N$, where the data are iid Poisson(λ). Obtain the MLE for λ.

Examples of finding the MLE II

Two parameter model: Gaussian model

Suppose we model observations $X_1, \ldots, X_N$ as IID $N(\mu, \sigma^2)$ random variables. Find the MLE of $\theta = (\mu, \sigma^2)$.

Plotting Normal Likelihood

- The likelihood function (not just to point estimate) will be used to measure uncertainty.
- For models with a single parameter, we often plot the likelihood curve.
- With more than one parameter, however, we have a likelihood surface.
- For the iid Normal(μ, σ^2) model, code for plotting this surface is available with course source-code.

Plotting Normal Likelihood II

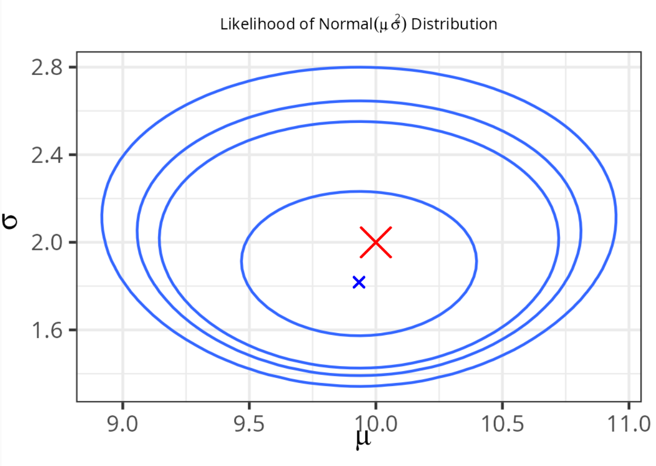


Figure 1: Likelihood surface of data generated from normal distribution.

Numeric Optimization

Numeric Optimization

- In the previous examples, the MLE was available *analytically*.
- In many cases, however, there is no closed-form solution for the MLE, and it must be computed numerically.
- The next example demonstrates this, and then we will discuss optimization strategies.

Numeric Optimization II

Example: Gamma likelihood

Suppose we want to model data $X_1, X_2, \ldots, X_n$ as iid Gamma(α, λ), which has the density function:

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

Find the MLE of $\theta = (\alpha, \lambda)$.

Numeric Optimization III

- The previous example leads us to consider numeric techniques for optimization and root finding.
- Note that this class is not an optimization course, so we'll only cover some of the most basic ideas.
- More modern and efficient numeric optimization techniques are readily available in R (or any other statistical software).
- For this class, we'll introduce some basic ideas like the Newton-Raphsom approach for root finding and optimization, as well as some other basic methods.

Newton-Raphsom root-finding algorithm

- Idea: start at a point θ_0, and approximate find the tangent line of $f(\theta)$ at the point θ_0:

$$y - f(\theta_0) = f'(\theta_0)\big(\theta - \theta_0\big)$$

- Then, find the root of the tangent line by setting $y = 0$, and solving for θ:

$$\theta = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}.$$

Newton-Raphsom root-finding algorithm II

- This root of the tangent line will be closer than our original guess θ_0, so we set:

  $$\theta_1 = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)},$$

  and repeat:

  $$\theta_{n+1} = \theta_n - \frac{f(\theta_n)}{f'(\theta_n)}.$$

- We stop based on some convergence criteria, often something like $|\theta_{n+1} - \theta_n| < \epsilon$, for a small choice of ϵ.
- (In class, check out wikipedia or draw a picture).

Newton-Raphsom root-finding algorithm III

- We need now a starting point θ_0.
- Really we can pick anything, but it's best if we are close to the MLE.
- For our current problem (Gamma distribution), we could use the MoM estimator:

$$\hat{\alpha}_{\mathsf{MoM}} = \theta_0 = \frac{\left(\bar{x}_n^*\right)^2}{\frac{1}{n}\sum_{i=1}^n \left(x_i^* - \bar{x}_n^*\right)^2}.$$

Newton-Raphsom root-finding algorithm IV

```
NR_root <- function(theta0, fn, deriv, tol = 1e-8, maxiter = 100
  iter <- 0
  theta_old <- theta0
  theta_new <- theta_old + 10 * tol

  while(abs(theta_old - theta_new) > tol && iter < maxiter) {
    iter <- iter + 1
    theta_old <- theta_new
    theta_new <- theta_old - fn(theta_old) / deriv(theta_new)
  }
  cat("iters: ", iter, "\n")
  theta_new
}
```

Newton-Raphsom root-finding algorithm V

```r
alpha_fn <- function(alpha, data) {
  n <- length(data)
  n * log(alpha) - n * log(mean(data)) + sum(log(data)) - n * di
}

alpha_deriv <- function(alpha, data) {
  n <- length(data)
  (n/alpha) - n * psigamma(alpha, 1)
}

set.seed(123)
data <- rgamma(n = 23, 1, 2)
```

Newton-Raphsom root-finding algorithm VI

```
# Not the exact MoM estimate, but close enough:
alpha_mom <- (mean(data)^2) / sd(data)

NR_root(
  theta0 = 0.86,
  fn = function(x) alpha_fn(x, data = data),
  deriv = function(x) alpha_deriv(x, data = data),
  tol = 1e-10
)

iters:  5
[1] 0.9728019
```

References and Acknowledgements

Pawitan Y (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA.

- Compiled on January 22, 2026 using R version 4.5.2.
-

References and Acknowledgements II

- We acknowledge students and instructors for previous versions of this course / slides.