Mathematical Statistics II

Maximum Likelihood Estimation

Jesse Wheeler

Contents

1 Introduction **1**

2 Likelihood: an introduciton **1**

3 Joint Probabilities **5**

1 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).

2 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...)

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.

- In a probability class, we might assume the probability of heads $\theta = p$ is some number, say $p = 0.4$.

- If X is the number of heads, then $P_\theta(X = 8) = 0.011$ is something we can calculate exactly. In any interpretation of probability, this means that if we repeat the experiment 10,000 time, we expect around 110 times, the experiment will have 8 heads.
- Now what if, in a single experiment, we observe $X = 8$? **Based on this information alone**, what can we say about the parameter p? Information we have about p is *incomplete*(we've already done a pure frequentist interpretation of this, as well as a Bayesian interpretation. The MoM estimate is also easy to compute).
- However, we can conclude that p **cannot be zero or one**. This fact is available *deductively*, since if p were zero or one, then $X = 0$ or $X = 10$.
- Similarly, we are fairly certain that p is not close to zero, since observing $X = 10$ would be extremely unlikely in this case. For instance, the probability of observing $X = 8$ if $p = 0.10$ is only 3.6×10^{-7}, meaning we would only expect about 1 out of 3 million replicates to give $X = 8$ (or similar for observing $X \in \{8, 9, 10\}$).
- In contrast, $\theta = 0.6$ or $\theta = 0.7$ are *likely*, since $P_\theta(X = 8)$ would be 0.12, 0.23, respectively.
- Thus, we have found a *deductive* way of comparing different values of p, based on the observed data: Compare the probability of observed data under different values of p.
- Concretely, we take the probability model $P_\theta(X = 8)$, and treat it as a function of θ.
- Note this is opposite of what we did last semester, where we treat the pmf for a fixed θ as a function of x (the possible output).
- This is how we define the likelihood function (for this model, and more generally, for any pmf):

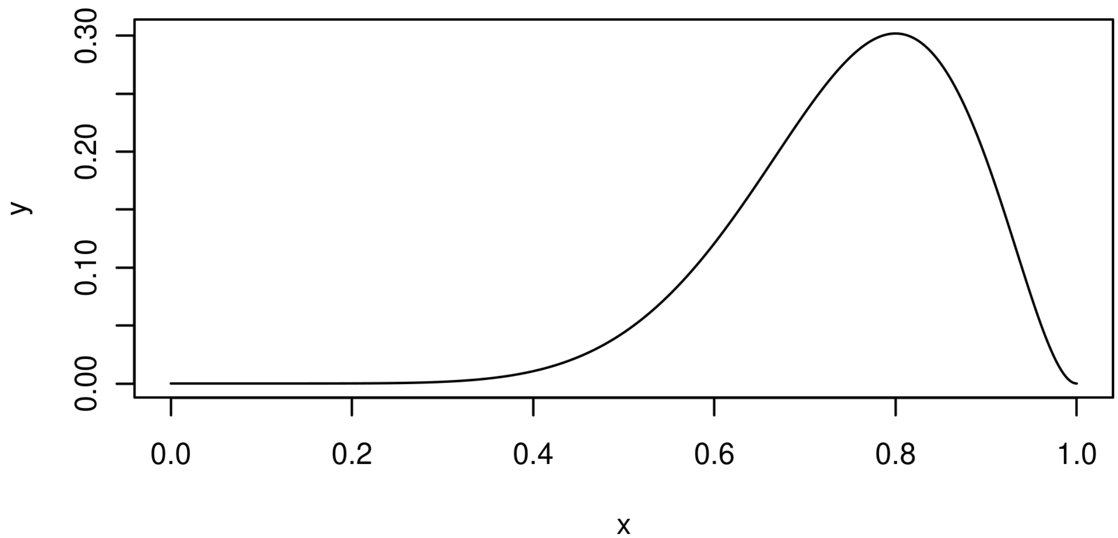
$$L(\theta) = P_\theta(X = 8) = f(x^*; \theta).$$

- 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:

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



- 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 $L(\theta)$.
- *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 $\{\theta_1, \theta_2, \ldots\}$.
- 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.
- 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.

- First, describing what the likelihood function is. We already did this for this model, but a quick review, reinforcing common notation.
- We'll use a binomial(N, p) model; the parameter of interest is $\theta = p$.
- Let X be a random variable, denoting the number of heads. Thus, pmf of the model is:

$$P(X = x) = f(x; \theta) = \binom{N}{x} p^x (1-p)^{N-x}.$$

- We assume we observe a specific dataset x^*. In this case, we observe $x^* = n$. Fixing the model at the observed data, we can describe the probability of the observed data as a function of $\theta = p$:

$$L(\theta) = f(x^*; \theta) = \binom{N}{n} p^n (1-p)^{N-n},$$

  and the MLE is:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta),$$

  meaning that our estimate $\hat{\theta}$ is the *argument* that *maximizes* the likelihood function $L(\theta)$.

- As often is the case, this will be easiest if we take the natural logarithm, giving the log-likelihood. Note this is a monotonic transformation, so the argument that maximizes the log-likelihood is the same as the argument that maximizes the likelihood.
- Typically, we denote the log-likelihood using $\ell(\theta)$. In latex, this symbol is given by `\ell`.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) = \underset{\theta}{\operatorname{argmax}} \log\big(L(\theta)\big) = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta).$$

- For our specific model, the log-likelihood is:

$$\ell(\theta) = \log \binom{N}{n} p^n (1-p)^{N-n} = \log\left(\binom{N}{n}\right) + n\log(p) + (N-n)\log(1-p).$$

- Now we see another recurring theme when finding the MLE: we end up with terms that don't impact the maximization. That is, the binomal term out front does not change the maximum value of p, so it can be ignored.
- To maximize, we take the derivative, and set equal to zero:

$$\frac{\partial \ell}{\partial \theta} = \frac{n}{p} - \frac{N-n}{1-p} = 0,$$

  or

$$\frac{n}{p} = \frac{N-n}{1-p} \implies n - np = Np - np.$$

- Solving for p, we get the MLE:

$$p = \frac{n}{N} \implies \hat{p} = \frac{n}{N} = \underset{\theta}{\operatorname{argmax}}\, L(\theta).$$

- A few notes:

- The MLE matches the frequentist-derived estimate. In most models, a first-principle estimate using frequentist interpretation of probability is not available, but the MLE will still match what the estimate *would be* if such an estimate were available.
- Setting a derivative equal to zero and solving only gives a critical point. Why did I assume it was a maximum? First, we plotted like likelihood function, and it is clear any critical point will be a maximum. More importantly, many of the classic probability models are part of what is known as the *exponential family*. In this family, all critical points of $L(\theta)$ will correspond to maximums. Next, theory we will establish later shows that as the number of observations increases, the likelihood function will be concave down around the MLE, suggesting that for even complex models, critical points in the region of high-likelihood will always be maximums.

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:

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

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

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

- 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 $L(\theta)$.
- **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

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

3 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:

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

- 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:

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

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.

- If X_i are iid poisson, then the pmf of a single observation is:

$$f(x;\, \theta) = P_\theta(X_i = x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

- In this case, the parameter set of interest, θ is a single value: $\theta = \lambda$.
- Due to the IID assumption, the likelihood of each observation is given by:

$$L(\theta) = f_{X_{1:N}}(x_{1:N}^*;\, \theta) = \prod_{i=1}^{N} \frac{\lambda^{x_i^*} e^{-\lambda}}{x_i^*!}.$$

- We want to maximize this function with respect to λ; this is easier done after taking the log:

$$\begin{aligned}
\ell(\theta) &= \log f_{X_{1:N}}(x_{1:N}^*;\, \theta) \\
&= \sum_{i=1}^{N} \log f_{X_1}(x_i^*;\, \theta) \\
&= \sum_{i=1}^{N} \left(x_i^* \log \lambda - \lambda - \log(x_i^*!) \right) \\
&= \log \lambda \sum_{i=1}^{N} x_i^* - N\lambda - \sum_{i=1}^{N} \log(x_i^*!) \\
&= \log \lambda \sum_{i=1}^{N} x_i^* - N\lambda - c(x^*).
\end{aligned}$$

- In the last step, we replaced the sum of log-factorials with some constant function $c(x^*)$. As a function of θ, the maximum of $\ell(\theta)$ does not change with this constant, so it can be ignored for the purposes of maximization.
- Now to find the maximum, we can take the derivative and set equal to zero:

$$\ell'(\theta) = \frac{1}{\lambda}\sum_{i=1}^{N} x_i^* - N = 0,$$

  Thus,

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} x_i^* = \bar{x}_N^*.$$

- We formally need to check if this is indeed a maximum and not just a critical point. The second derivative with respect to λ is always negative (since x_i^* is non-negative), implying that it is indeed a maximum.
- We'll see later that as $N \to \infty$, then the likelihood function is *always* concave-down around the maximum, under fairly week conditions.
- We could also plot the likelihood / log-likelihood, which is informative on it's own right.
- Finally, note that this is the same estimator that was obtained via the MoM approach. This sometimes happens.

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 $\theta = (\mu, \sigma^2)$.

- Though we have two parameters, the approach for finding the MLE is similar: find the likelihood function of the entire dataset, take the logarithm, and the compute (partial) derivatives and set equal to zero.
- In the final step, we might end up with a system of equations rather than just a single equation.
- There's also an interesting feature of this example that is applicable elsewhere: should we find the MLE of σ, or the MLE of σ^2? Fortunately, you get the same result either way (we'll show this later).
- In this case, either approach is straightforward. As practice, you may want to consider finding the MLE of σ^2. Here, we will treat σ as the parameter of interest, though sometimes it's easier to differentiate with respect to something like: $\theta_2 = \sigma^2$.
- Because of the IID assumption joint-likelihood function is:

$$L(\mu, \sigma) = f(x^*;\, \mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\big((x_i^* - \mu)/\sigma\big)^2\Big),$$

- The log-likelihood is therefore:

$$\begin{aligned}\ell(\mu, \sigma) &= \sum_{i=1}^{N}\Big(\log(1) - \log(\sigma\sqrt{2\pi}) - \frac{1}{2}\big((x_i^* - \mu)/\sigma\big)^2\Big) \\ &= -n\log(\sigma) - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i^* - \mu)^2.\end{aligned}$$

- The partial derivatives are given by:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i^* - \mu) = 0$$
$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{N}(x_i^* - \mu)^2.$$

- Because $\sigma \geq 0$ (and zero if and only if the data are the same value), the first equation will only be zero if $\sum_{i=1}^{N}(x_i^* - \mu) = 0$, or if $\mu = \bar{x}_N^*$, which is independent of σ. Thus, the MLE is:

$$\hat{\mu} = \bar{x}_N^*.$$

- For the second equation, the partial derivative is zero if and only if

$$\frac{n}{\sigma} = \sigma^{-3}\sum_{i=1}^{N}(x_i^* - \mu)^2,$$

  or if

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{N}(x_i^* - \mu)^2.$$

  Since we are looking for the MLE of $\theta = (\theta, \sigma)$, we can now replace μ with it's maximizer $\hat{\mu} = \bar{x}_N^*$, giving us:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i^* - \hat{\mu})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i^* - \bar{x}_N^*)^2}.$$

- Like the Poisson example, the MLE for the iid normal model has the same estimator as the MoM procedure.

- The sampling distribution for θ is therefore:

$$\hat{\mu} \sim N(\mu, \sigma^2/N), \quad N\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-1},$$

  and the two distributions are independent.

Acknowledgments

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

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

References

Pawitan Y (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press. 1, 2

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA. 1