1 Maximum Likelihood Estimators (MLEs)

A typical process for computing an MLE based on a random sample is as follows:

(a) Write out the joint likelihood $L(\theta)$. You'll probably want to simplify as much as possible in this stage. You can use proportionality α_θ to simplify even further, if you want.

(b) Take the log, obtaining the log-likelihood $l(\theta)$.

(c) Take the derivative with respect to θ, to get $l'(\theta)$.

(d) Set the derivative $l'(\theta)$ equal to 0, solving for the critical value(s) $\hat{\theta}$ of the parameter.

(e) Verify that $\hat{\theta}$ globally maximizes.

- If there is one critical value, verify that it is a maximizer through either the first derivative test (l' is positive for values less than $\hat{\theta}$ and negative for values greater than $\hat{\theta}$) or second derivative test ($l''(\hat{\theta}) < 0$)
- If there is more than one critical value, then check all the values.

Note that this process does not always work, depending on some technicalities that can arise sometimes.

1.1 Some Technicalities

The above process relies on differentiability. If the likelihood/log-likelihood is not differentiable, then we are unable to use our favorite tool of calculus to optimize. You would then have to examine the likelihood analytically, plot it, etc.

Another issue that can arise is that the MLE may exist on the *boundary* of the parameter space. This typically isn't a problem, since many of our commonly used probability distributions have parameter spaces that are open sets (e.g., $\mu \in \mathbb{R}$ for the normal distribution or $\alpha \in (0, \infty)$ for the Gamma distribution). That's why most people don't even bring up endpoints and just jump straight to calculating the joint likelihood (which is what I do too, when I see that it's not an issue for a particular distribution). It often arises though when there are *constraints* on the values the parameter can take. There are a few examples below that highlight this.

Also it's possible that the MLE does not exist. Even if the MLE does exist, it's possible that we can't solve for it analytically, in which case we would have to use numerical optimization methods on the computer.

MLE for Bernoulli. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$, where $0 < p < 1$. Find the MLE for p.

Solution. The density for a Bernoulli RV is $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$, for $p \in (0, 1)$. So the parameter space is (0, 1), an open set, and we don't have to worry about endpoints, so we can just follow our typical process for getting the MLE. The joint likelihood then is

$$\begin{aligned} L(p) = f(X_1, \ldots, X_n; p) &= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} \\ &= p^{\sum_{i=1}^n X_i}(1-p)^{\sum_{i=1}^n (1-X_i)} \\ &= p^{\sum X_i}(1-p)^{n-\sum X_i}. \end{aligned}$$

The log-likelihood is

$$l(p) = log(L(p)) = (\sum X_i)\log(p) + (n - \sum X_i)\log(1-p).$$

Differentiating with respect to p, we have

$$l'(p) = \frac{\sum X_i}{p} + \frac{\sum X_i - n}{1-p}.$$

Setting it equal to 0 and solving for p, we have

$$\begin{aligned} \frac{\sum X_i}{p} + \frac{\sum X_i - n}{1-p} = 0 &\iff \\ \frac{(1-p)\sum X_i + p(\sum X_i - n)}{p(1-p)} = 0 &\iff \\ (1-p)\sum X_i + p(\sum X_i - n) = 0 &\iff \\ \sum X_i - p\sum X_i + p\sum X_i - np = 0 &\iff \\ \sum X_i - np = 0 &\iff \\ p = \frac{\sum X_i}{n} = \bar{X}_n. \end{aligned}$$

Thus, $\hat{p} = \bar{X}_n$ is a global optimizer.

Let us verify that $\hat{p}$ is a global maximizer with the second derivative test:

$$l''(p) = \frac{-\sum X_i}{p^2} + \frac{n - \sum X_i}{(1-p)^2}.$$

Plugging in $\hat{p}$, we have

$$\begin{aligned} l''(\hat{p}) &= \frac{-\sum X_i}{\bar{X}^2} + \frac{n - \sum X_i}{(1-\bar{X})^2} \\ &= \frac{-n\bar{X}}{\bar{X}^2} + \frac{n - n\bar{X}}{(1-\bar{X})^2} \\ &= \frac{-n}{\bar{X}} + \frac{n(1-\bar{X})}{(1-\bar{X})^2} < 0, \end{aligned}$$

as long as $\bar{X}$ is not exactly 0 or 1 (which would only happen if all of our $X_i's$ were 0 or all of our $X_i's$ were 1s).

Thus, $\hat{p} = \bar{X}$ is indeed the MLE for p.

MLE for Constrained Bernoulli. Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$ but where we have the constraint that p cannot exceed $1/2$. Find the MLE for p.

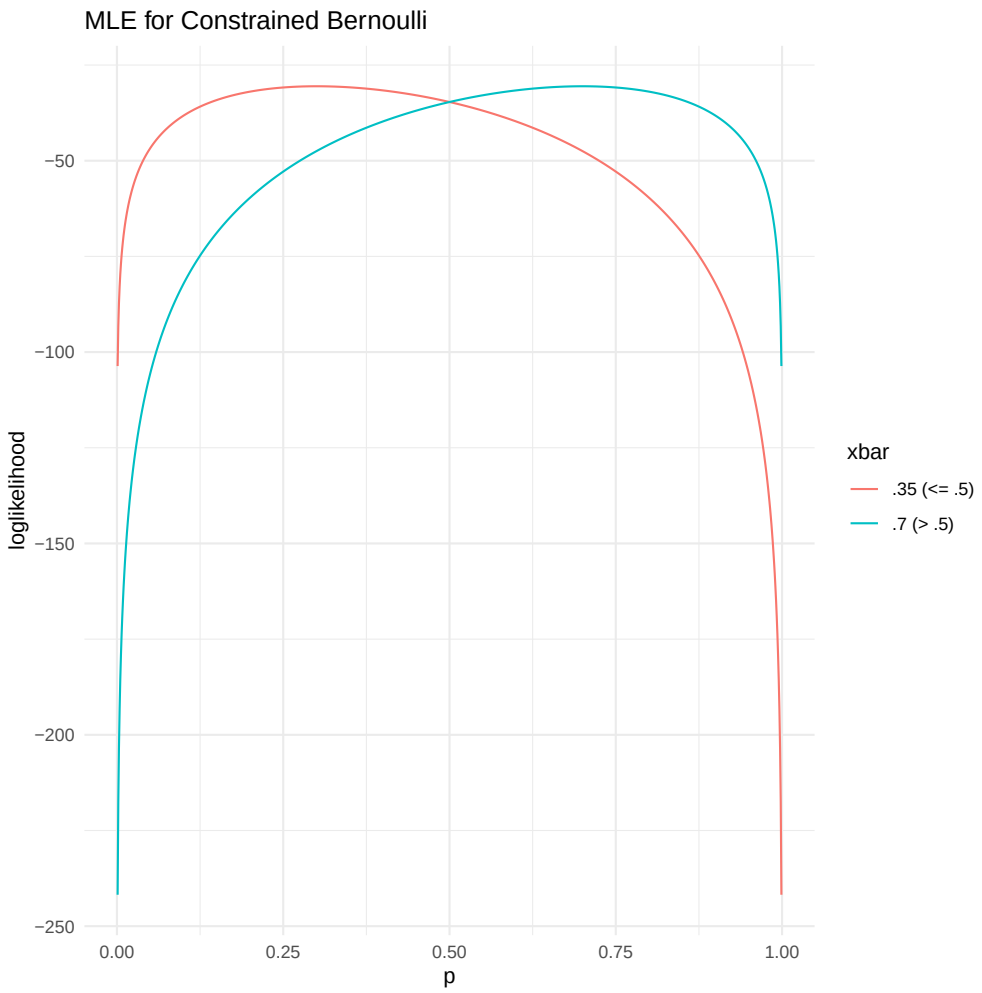
Solution. The density is

$$f(x_i; p) = p^{x_i}(1-p)^{1-x_i},$$

where now p is only allowed to be in $(0, 1/2]$. Since we now have a half-open interval, we do need to worry about the endpoints.

Because the likelihood function for Bernoullis is concave, it follows that the MLE is $\min\{\bar{X}, 1/2\}$.

For example, suppose $n = 50$. Let's look at the log-likelihood if $\bar{x} > .5$ vs. if $\bar{x} \leq .5$. The below plot compares the log-likelihoods for $\bar{x} = .7$ and $\bar{x} = .35$, over $p \in (0, 1)$. We can see that over $(0, 1)$, $\bar{X}$ achieves the maximum. But since p is restricted to be less than or equal to .5, only the left half of the plot is relevant here. On this constrained domain of $(0, 1/2]$, we can see that when $\bar{x} = .7$ (blue line), the maximum value for the likelihood is attained for .5. When $\bar{x} = .35$, the maximum value of the likelihood is attained for .35.



We discussed another approach in section for computing the MLE when there are constraints on the parameter. In this approach, from the beginning, write the likelihood with the constraints encoded into indicator functions. But keep in mind that the presence of indicators makes a function non-differentiable, so we have to consider the different cases and maximize in that way. One of the exercises shows an example of this.

1.2 Properties of the MLE

The MLE has many desirable properties, including that

- it exists and is consistent with probability 1
- it is asymptotically normal
- it is asymptotically efficient.

Asymptotic normality is very helpful for computing confidence interals, so for now, let's review that a bit more. If $\hat{\theta}_n$ is the MLE for θ, then

$$\hat{\theta}_n \implies N(\theta, 1/I_n(\theta)),$$

where $I_n(\theta)$ is the Fisher Information, defined as

$$I_n(\theta) = \text{Var}(l'(\theta)).$$

Under some assumptions, it turns out that

$$I_n(\theta) = \mathbb{E}\left((l'(\theta))^2\right) = -\mathbb{E}(l''(\theta)).$$

In the case of iid data, it also turns out that

$$I_n(\theta) = nI(\theta),$$

where $I(\theta)$ is the Fisher Information computed for just ONE observation (rather than the joint sample). That is,

$$I(\theta) = \text{Var}(l_i'(\theta)).$$

In this case, you can write that

$$\hat{\theta}_n \implies N(\theta, 1/nI(\theta)).$$

Often it's easier to use this formula.

Caution: be very careful when using $1/I_n(\theta)$ vs. $1/nI(\theta)$ for the asymptotic variance of the MLE! In the case of iid data, they are the same, but just be sure that you know when to multiply by n or not, depending on if you calculated the Fisher Information for one observation or for all the observations. Otherwise, you will end up over or under counting the variance.

Also, the Fisher Information often ends up depending on θ...which is what we were trying to estimate in the first place! In that case, just plug in $\hat{\theta}_n$ for θ to get the plug-in estimator.

Problem 1. Estimation for Pareto Distribution

The Pareto distribution has been used in economics as a distribution with a slowly decaying tail. A simplified version of the density can be written as

$$f(x;\theta) = \theta\beta^{\theta}x^{-\theta-1}$$

where $\theta > 1$ is a parameter, $\beta > 0$ is a known constant, and $x \geq \beta$. Suppose we have an iid sample $X_1, X_2, \ldots, X_n$ from this distribution.

(a) Find the MLE of θ.

(b) Find the asymptotic variance of the MLE.

(c) Find an approximate 90% confidence interval for θ.

Solution

Part (a)

The likelihood can be written as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(X_i;\theta) \text{ (since } X_1, \ldots, X_n \text{ are iid)} \\ &= \prod_{i=1}^{n} \theta\beta^{\theta}X_i^{-\theta-1} \\ &= \theta^n\beta^{n\theta}\left(\prod_{i=1}^{n} X_i\right)^{-\theta-1}. \end{aligned}$$

The log-likelihood is

$$\begin{aligned} l(\theta) &= n\log(\theta) + n\theta\log(\beta) - (\theta+1)\log\left(\prod_{i=1}^{n} X_i\right) \\ &= n\log(\theta) + n\theta\log(\beta) - (\theta+1)\sum_{i=1}^{n}\log(X_i). \end{aligned}$$

Taking the derivative with respect to θ, we have

$$l'(\theta) = \frac{n}{\theta} + n\log(\beta) - \sum_{i=1}^{n}\log(X_i).$$

Setting equal to 0, we have

$$\frac{n + n\theta \log(\beta) - \theta \sum_{i=1}^{n} \log(X_i)}{\theta} = 0 \implies$$
$$\theta \left(n \log(\beta) - \sum_{i=1}^{n} \log(X_i) \right) = -n \implies$$
$$\theta = \frac{-n}{n \log(\beta) - \sum_{i=1}^{n} \log(X_i)}.$$

It remains to verify that this is indeed a maximum, rather than a minimum. Taking the second derivative of the log-likelihood, we have

$$l''(\theta) = -\frac{n}{\theta^2} < 0,$$

which implies that, indeed, the MLE is

$$\hat{\theta}_{ML} = \frac{n}{\sum_{i=1}^{n} \log(X_i) - n \log(\beta)}.$$

Part (b)

The approximate variance of the MLE is $\frac{1}{nI(\theta)}$, where $I(\theta)$ is the Fisher information. The regularity assumptions are met for this distribution, so I will use the formula $I(\theta) = -\mathbb{E}(l_i''(\theta))$. The likelihood for one observation is

$$L_i(\theta) = \theta \beta^{\theta} X_i^{-\theta-1},$$

so the log-likelihood is

$$l_i(\theta) = \log(\theta) + \theta \log(\beta) - (\theta + 1) \log(X_i).$$

Taking the derivative with respect to θ, we have

$$l_i'(\theta) = \frac{1}{\theta} + \log(\beta) - \log(X_i).$$

The second derivative is

$$l_i''(\theta) = -\frac{1}{\theta^2},$$

so the Fisher Information is

$$I(\theta) = -\mathbb{E}[l_i''(\theta)] = \frac{1}{\theta^2}.$$

Thus, the approximate variance of the MLE is $\frac{\theta^2}{n}$.

Part (c)

By the asymptotic normality of the MLE, a 90% confidence interval for θ would be

$$\hat{\theta}_{ML} \pm z_{.05} \frac{1}{nI(\theta)}.$$

But the Fisher information depended on θ, so let us instead use the plug-in estimator, obtaining

$$\hat{\theta}_{ML} \pm z_{.05} \frac{\hat{\theta}^2_{ML}}{n}.$$

Problem 2. MLE for the Uniform Distribution

Suppose we have the continuous random variables $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta]$ with unknown parameter θ. (Recall that if $X \sim \text{Uniform}(a, b]$, then the likelihood function is $f(x; a, b) = \frac{1}{b-a}$ for any $x \in (a, b]$.) What is the MLE for θ?

Solution

This problem is a bit tricky, since we have a constraint on θ, in that the support of the distribution is involving θ.

With this in mind, let us rewrite the likelihood for a Uniform random variable as $f(x; a, b) = \frac{1}{b-a} 1(a < x \leq b)$, where $1(\cdot)$ is the indicator function.

Now we are ready to write the likelihood for our sample:

$$\begin{aligned} L(\theta) &= f(X_1, \ldots, X_n; \theta) \\ &= \prod_{i=1}^{n} f(X_i; \theta) \text{ (since the observations are iid)} \\ &= \prod_{i=1}^{n} \frac{1}{\theta} 1(0 < X_i \leq \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} 1(X_i > 0)(X_i \leq \theta) \\ &= \frac{1}{\theta^n} 1(\min_{i=1,\ldots,n} X_i > 0) 1(\max_{i=1,\ldots,n} X_i \leq \theta). \end{aligned}$$

Since the indicator $1(\min_{i=1,\ldots,n} X_i > 0)$ does not depend on θ, we can write

$$L(\theta) \propto_\theta \frac{1}{\theta^n} 1(\max_{i=1,\ldots,n} X_i \leq \theta).$$

Now there is only one indicator left, but the presence of the indicator depending on θ makes the entire function non-differentiable.

We can split this function into cases:

$$l(\theta) \propto_\theta \begin{cases} \frac{1}{\theta^n} & \text{if } \max_{i=1,\ldots,n} X_i \leq \theta \\ 0 & \text{if } \max_{i=1,\ldots,n} X_i > \theta \end{cases}.$$

Let's take a closer look at this. First, observe that $\frac{1}{\theta^n} > 0$ for all $\theta > 0$, which means that the likelihood will always be greatest when we are in the first case. So we ought to choose θ such that we are always in the first case, and we can ensure that by choosing $\theta \geq \max X_i$. Second, observe that $\frac{1}{\theta^n}$ decreases as θ increases, which means that the likelihood will be greatest for the smallest possible value of θ (subject to the constraint that $\theta \geq \max(X_i)$...but that's $\max X_i$ itself!

Thus, the MLE is $\hat{\theta} = \max_{i=1,\ldots,n} X_i$.

Problem 3. Optional: Intuition for Estimation for Uniform RVs

Is there anything that seems weird about the estimator that we got for the MLE of the Uniform distribution? Can you think of another estimator for θ?

Solution

Some context: Estimating θ for the Uniform distribution was actually done during World War II and was known as the "German Tank Problem". The Allied forces were trying to estimate the total number of tanks that Nazi Germany had, based on broken parts of tanks with consecutive serial numbers that were found on battlefields. This can be posed as estimating θ, the upper bound of the support for iid Uniform random variables.

With this context in mind, the MLE for θ would simply be the largest serial number observed from the sample of broken tank parts. So if the largest serial number was 208, we would just guess there are exactly 208 tanks...but intuitively, this does seem somewhat odd. I would guess the true number of tanks is probably 208 + some extra bit.

Indeed, the MLE is a biased estimator of θ, in that it underestimates θ. An unbiased estimator of θ turns out to be

$$\tilde{\theta} = \max(X_i)(1 + n^{-1}) - 1$$

Actually, this estimator also has the minimum variance among all other unbiased estimators! So it's a pretty good estimator.

It can also be written as $\tilde{\theta} = \max(X_i) + \frac{\max(X_i) - n}{n}$, which is the MLE + the average gap between consecutive observations in the sample. In other words, it is the MLE + some extra bit.

But ultimately, the MLE is doing exactly what it intends to. It gives us the most likely value of θ *based on the observed data*. This goes to show that we often want to consider a variety of properties when determining the best estimator to use for a particular situation.