7.5 Maximum Likelihood Estimators

The Maximum Likelihood Estimation provides a method for choosing estimators of parameters that avoids using prior distributions or loss functions. Instead, MLE selects as an estimate the value that maximizes the likelihood function. This is one of the most widely used estimation methods in statistics (and is relatively consistent with our intuition).

Ex 1: Suppose a possibly biased coin with probability p of heads is flipped 5 times, and shows heads on exactly 1 of those flips. Is it possible this coin is unbaised? (Yes, this outcome would occur about 15% of the time). But for what value of p is this an unsurprising result? (If $p = .2$, this would occur about 40% of the time) Would you have **good** reason to believe that $p = .99$?

Informally, the method of maximum likelihood looks at all possible models, and selects the one for which the data is most consistent.

However, this is not the same as saying that the given value of θ was most likely to have produced the data! To do so, we would need a prior distribution for θ.

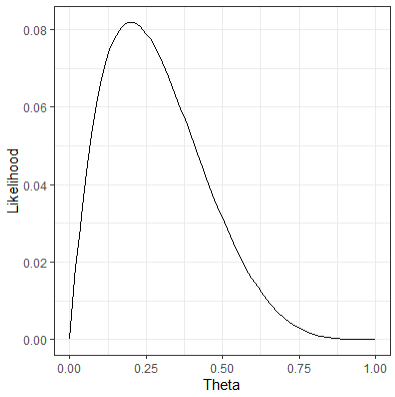
Def: For each possible observed vector $\mathbf{x}$, let $\delta(\mathbf{x}) \in \Omega$ denote the value of $\theta \in \Omega$ for which the likelihood function $f(\mathbf{x}|\theta)$ is maximum. The **maximum likelihood estimator** $\hat{\theta}$ is defined as $\hat{\theta} = \delta(\mathbf{X})$.

Ex 2: Again, suppose a coin with probability θ of landing heads is flipped 5 times, and exactly 1 heads is obtained. Find the likelihood function and the MLE.

Solution. Let X denote the number of heads obtained in 5 flips, and note that $X|\theta \sim \text{Bin}(5, \theta)$, and so

$$f(x|\theta) = \binom{5}{x}\theta^x(1-\theta)^{5-x}$$

With $x = 1$, $f(1|\theta) \propto \theta(1-\theta)^4$, which has a graph



Differentiating,

$$\frac{\partial}{\partial\theta}f(1|\theta) = (1-\theta)^4 - 4\theta(1-\theta)^3 = (1-\theta)^3(1-5\theta)$$

which has 0's at $\theta = \frac{1}{5}, 1$. Using 1st derivative test, $f(1|\theta)$ is max at $\theta = \frac{1}{5}$. ■

In many cases, it is more convenient to maximize the **log-likelihood function** $\log f(\mathbf{x}|\theta)$ (why?) and note that any max of $\ln f$ is also a max of f (also why?)

Thm: Suppose $X_1, X_2, \ldots, X_n$ form an iid sample from a Normal distribution, with unknown parameters μ and σ^2. The maximum likelihood estimator for $\theta = (\mu, \sigma^2)$ is

$$\hat{\theta} = \left(\bar{X}_n, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right)$$

Proof. By assumption, $X_1, \ldots, X_n$ are conditionally iid $N(\mu, \sigma^2)$, with

$$\begin{aligned} f(x_i|\mu, \sigma^2) =& \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\} \\ f(\mathbf{x}|\mu, \sigma^2) =& \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\ \log f(\mathbf{x}|\mu, \sigma^2) =& -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

We first seek to maximize over μ, with σ^2 fixed:

$$\frac{\partial}{\partial \mu} \log f(\mathbf{x}|\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = -n\frac{\mu}{\sigma^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i = \frac{n}{\sigma^2}(\bar{x} - \mu)$$

which has a zero when $\mu = \bar{x}$. Note that the estimator for μ doesn't depend on σ^2.

Now, differentiating the log likelihood function with respect to σ^2:

$$\frac{\partial}{\partial \sigma^2} \log f(\mathbf{x}|\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

which has a zero when

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

Since both $\frac{\partial}{\partial \mu} \log f$ and $\frac{\partial}{\partial \sigma^2} \log f$ must be simultaneously zero in order to maximize $\log f$, then we can substitute the solution $\mu = \bar{x}$. This gives a critical point of

$$\hat{\theta} = \left(\bar{x}, \frac{1}{n} \sum (x_i - \bar{x})^2 \right)$$

To classify this critical point, we compute the Hessian matrix:

$$H(\theta, \sigma^2) = \begin{pmatrix} \frac{\partial^2}{\partial \mu^2} \log f & \frac{\partial^2}{\partial (\sigma^2) \partial \mu} \log f \\ \frac{\partial^2}{\partial \mu \partial (\sigma^2)} \log f & \frac{\partial^2}{\partial (\sigma^2)^2} \log f \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n}{(\sigma^2)^2}(\bar{x} - \mu) \\ -\frac{n}{(\sigma^2)^2}(\bar{x} - \mu) & \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix}$$

Evaluating this at $\theta = \bar{x}$ and $\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \hat{\sigma}^2$ gives the matrix:

$$H(\bar{x}, \hat{\sigma}^2) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n}{(\hat{\sigma}^2)^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix}$$

which has determinant

$$\det H(\bar{x}, \hat{\sigma}^2) = \frac{n^2}{2(\hat{\sigma}^2)^3} > 0$$

Since

$$\frac{\partial^2}{\partial \mu^2} = -\frac{n}{\hat{\sigma}^2} < 0$$

then the solution

$$\hat{\theta} = \left(\bar{x}, \frac{1}{n}\sum (x_i - \bar{x})^2\right)$$

is a local maximum of the log likelihood function. □

The maximum likelihood estimator need not be unique:

Ex 3: Suppose $X_1, X_2, \ldots, X_n \sim f(x|\theta)$ where

$$f(x|\theta) = \frac{1}{2}e^{-|x-\theta|}$$

Then the likelihood and log-likelihood functions are

$$f(\mathbf{x}|\theta) = \frac{1}{2^n}e^{-\sum |x_i - \theta|} \qquad \log f(\mathbf{x}|\theta) = -n\log 2 - \sum |x_i - \theta|$$

The log-likelihood function is maximized when the expression $\sum |x_i - \theta|$ is minimized. If n is odd, this occurs exactly when θ is the unique median of the x_i. But if n even, this occurs when θ is ANY median of the x_i.

7.6 Properties of Maximum Likelihood Estimators

Maximum Likelihood Estimators have an important **invariance property**:

Thm: If $\hat{\theta}$ is the MLE of θ and g is a one-to-one function, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Proof. Let $\Gamma = g(\Omega)$. Since g is one-to-one, it has an inverse h on Γ. The likelihood function for $\psi = g(\theta)$ is conditional distribution of $\mathbf{x}$ given ψ. But we know the conditional distribution of $\mathbf{x}$ given θ: $f(\mathbf{x}|\theta)$, and since $\theta = h(\psi)$, the likelihood function for ψ is

$$f(\mathbf{x}|h(\psi))$$

This function is maximized when $\theta = \hat{\theta} = h(g(\hat{\theta}))$, and so is maximized when $g(\theta) = g(\hat{\theta})$. □

Ex 4: Suppose $X_1, \ldots, X_n$ are conditionally iid Pois(θ). Find the MLE for $p = P(X_i = 0)$.

Solution. Note that $p = P(X_i = 0) = e^{-\theta}$, so by the invariance principal, it suffices to find the MLE for θ. Note that the likelihood and log likelihood functions for θ are

$$f(\mathbf{x}|\theta) \propto e^{-n\theta}\theta^{x_1+\cdots+x_n} \qquad \log f(\mathbf{x}|\theta) = c - n\theta + (x_1 + \cdots + x_n)\log\theta$$

Differentiating

$$\frac{\partial}{\partial\theta}\log f = -n + \frac{x_1 + \cdots + x_n}{\theta}$$

which has a zero at $\theta = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}$. Hence, the MLE for p is

$$\hat{p} = e^{-\bar{x}}$$

■

We would like to extend this result to arbitrary functions g. But one problem is that if $g(\theta)$ is not one-to-one, and the statistical model is parameterized by θ, then the likelihood function for $g(\theta)$ isn't well-defined. To rectify, we introduce a more general notion of likelihood function:

Def: Let $g(\theta)$ be an arbitrary function of the parameter, and let $\Gamma = g(\Omega)$. For each $\gamma \in \Gamma$, let $T_\gamma = g^{-1}(\gamma) = \{\theta : g(\theta) = \gamma\}$. Define the **induced log-likelihood function** $L^*(\gamma)$ by

$$L^*(\gamma) = \max_{\theta \in T_\gamma} \log f(\mathbf{x}|\theta)$$

Define the MLE of $g(\theta)$ to be $\hat{\gamma}$ where

$$\hat{\gamma} = \arg\max_{\gamma \in \Gamma} L^*(\gamma)$$

Thm: Let $\hat{\theta}$ be an MLE of θ and let $g(\theta)$ be a function of θ. Then an MLE of $g(\theta)$ is $g(\hat{\theta})$

Ex 5: Suppose $X_1, \ldots, X_n$ are a random sample from Bern(θ), a distribution with mean θ and variance $\nu = \theta(1-\theta)$. If $\hat{\theta}$ is the MLE for θ, then $\hat{\theta}(1-\hat{\theta})$ is an MLE for ν.

While under reasonable assumptions, the maximum likelihood estimator is consistent and has the invariance property, it can be biased:

Def: An estimator $\delta(\mathbf{X})$ of $g(\theta)$ is **unbiased** if

$$E[\delta(\mathbf{X})] = g(\theta)$$

Ex 6: The MLE for variance in a Normal distribution is biased.

Solution. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Recall that the MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$

We note that

$$\begin{aligned}
n\hat{\sigma}^2 &= \sum (X_i - \bar{X})^2 \\
&= \sum \Big((X_i - \mu) - (\bar{X} - \mu) \Big)^2 \\
&= \sum (X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2 \\
&= \Big\{ \sum (X_i - \mu)^2 \Big\} - \Big\{ 2 \sum (X_i - \mu)(\bar{X} - \mu) \Big\} + \Big\{ \sum (\bar{X} - \mu)^2 \Big\} \\
&= \Big\{ \sum (X_i - \mu)^2 \Big\} - \Big\{ 2(\bar{X} - \mu) \sum (X_i - \mu) \Big\} + \big\{ n(\bar{X} - \mu)^2 \big\} \\
&= \Big\{ \sum (X_i - \mu)^2 \Big\} - \big\{ 2(\bar{X} - \mu) n(\bar{X} - \mu) \big\} + \big\{ n(\bar{X} - \mu)^2 \big\} \\
&= \Big\{ \sum (X_i - \mu)^2 \Big\} - \big\{ n(\bar{X} - \mu)^2 \big\}
\end{aligned}$$

Therefore,

$$\begin{aligned}
E[\hat{\sigma}^2] &= \frac{1}{n} E\left[\Big\{ \sum (X_i - \mu)^2 \Big\} - \big\{ n(\bar{X} - \mu)^2 \big\} \right] \\
&= \frac{1}{n} \left\{ \sum E\left[(X_i - \mu)^2 \right] - nE[(\bar{X} - \mu)^2] \right\} \\
&= \frac{1}{n} \left\{ n\sigma^2 - n\frac{\sigma^2}{n} \right\} \\
&= \frac{n-1}{n} \sigma^2
\end{aligned}$$

The fix?

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

But is s^2 an MLE for σ^2? No.

Although $\hat{\sigma}^2$ is a biased estimator, it is still consistent:

$$\begin{aligned}\hat{\sigma}^2 =& \frac{1}{n}\left[\sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] \\ =& \frac{\sum (X_i - \mu)^2}{n} - \frac{\sum n(\bar{X} - \mu)^2}{n}\end{aligned}$$

The left expression converges in probability to σ^2 and the right expression converges in probability to 0, by the weak law of large numbers. ■