

31. Maximum Likelihood Estimation and Method of Moments

Definition of MLE's

Easy Examples

Trickier Examples

Invariance Property of MLE's

Method of Moments

Definition of MLE's

Definition: Consider an i.i.d. random sample X_1, \dots, X_n , where each X_i has p.d.f./p.m.f. $f(x)$. Further, suppose that θ is some unknown parameter from X_i . The **likelihood function** is $L(\theta) \equiv \prod_{i=1}^n f(x_i)$.

Definition: The **maximum likelihood estimator** (MLE) of θ is the value of θ that maximizes $L(\theta)$. The MLE is a function of the X_i 's and is therefore a RV.

Easy Examples

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$. Find the MLE for λ .

$$L(\lambda) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right).$$

Now maximize $L(\lambda)$ with respect to λ .

Could take the derivative and plow through all of the horrible algebra. Too tedious. Need a trick....

Useful Trick: Since the natural log function is one-to-one, it's easy to see that the λ that maximizes $L(\lambda)$ also maximizes $\ln(L(\lambda))$!

$$\ln(L(\lambda)) = \ln\left(\lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right)\right) = n\ln(\lambda) - \lambda \sum_{i=1}^n x_i$$

This makes our job less horrible.

$$\frac{\partial}{\partial \lambda} \ln(L(\lambda)) = \frac{\partial}{\partial \lambda} \left(n\ln(\lambda) - \lambda \sum_{i=1}^n x_i \right) = \frac{n}{\lambda} - \sum_{i=1}^n x_i \equiv 0.$$

This implies that the MLE is $\hat{\lambda} = 1/\bar{X}$.

Remarks: (1) $\hat{\lambda} = 1/\bar{X}$ makes sense since $E[X] = 1/\lambda$.

(2) At the end, we put a little $\hat{\text{hat}}$ over λ to indicate that this is the MLE.

(3) At the end, we make all of the little x_i 's into big X_i 's to indicate that this is a RV.

(4) Just to be careful, you probably ought to perform a second-derivative test, but I won't blame you if you don't.

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(p)$. Find the MLE for p .

Useful trick for this problem: Since

$$X_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1 - p \end{cases},$$

we can write the p.m.f. as

$$f(x) = p^x(1 - p)^{1-x}, \quad x = 0, 1.$$

Thus,

$$\begin{aligned}L(p) &= \prod_{i=1}^n f(x_i) \\&= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \\&= p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}\end{aligned}$$

\Rightarrow

$$\ln(L(p)) = \sum_{i=1}^n x_i \ln(p) + (n - \sum_{i=1}^n x_i) \ln(1-p)$$

\Rightarrow

$$\frac{\partial}{\partial p} \ln(L(p)) = \frac{\sum_i x_i}{p} - \frac{n - \sum_i x_i}{1 - p} \equiv 0.$$

 \Rightarrow

$$(1 - p) \left(\sum_{i=1}^n x_i \right) - p \left(n - \sum_{i=1}^n x_i \right) = 0$$

 \Rightarrow

$$\hat{p} = \bar{X}.$$

This makes sense since $E[X] = p$.

Trickier Examples

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$. Find the *simultaneous* MLE's for μ and σ^2 .

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

This \Rightarrow

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

\Rightarrow (by the chain rule)

$$\frac{\partial}{\partial \mu} \ln(L(\mu, \sigma^2)) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) \equiv 0$$

\Rightarrow

$$\hat{\mu} = \bar{X}.$$

Now do the same thing for σ^2 ...

Similarly, take the partial w/rt σ^2 (*not* σ),

$$\frac{\partial}{\partial \sigma^2} \ln(L(\mu, \sigma^2)) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \hat{\mu})^2 \equiv 0.$$

\Rightarrow

$$-n\sigma^2 + \sum_{i=1}^n (x_i - \bar{x})^2 = 0.$$

After a bit more algebra, we get

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

Recap:

$$\hat{\mu} = \bar{X}, \quad \widehat{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}.$$

Remark: Notice how close $\widehat{\sigma^2}$ is to the (unbiased) sample variance

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{n}{n-1} \widehat{\sigma^2}$$

$\widehat{\sigma^2}$ is a little bit biased, but it has slightly less variance than S^2 . Anyway, as n gets big, S^2 and $\widehat{\sigma^2}$ become the same.

Example: The p.d.f. of the Gamma distrn w/parameters r and λ is

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x > 0.$$

Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Gam}(r, \lambda)$. Find the MLE's for r and λ .

$$L(r, \lambda) = \prod_{i=1}^n f(x_i) = \frac{\lambda^{nr}}{[\Gamma(r)]^n} \left(\prod_{i=1}^n x_i \right)^{r-1} e^{-\lambda \sum_i x_i}.$$

This \Rightarrow

$$\ln(L) = rn \ln(\lambda) - n \ln(\Gamma(r)) + (r-1) \ln\left(\prod_i x_i\right) - \lambda \sum_i x_i$$

\Rightarrow

$$\frac{\partial}{\partial \lambda} \ln(L) = \frac{rn}{\lambda} - \sum_{i=1}^n x_i \equiv 0,$$

so that $\hat{\lambda} = \hat{r} / \bar{X}$.

The trouble is, we need to find $\hat{r} \dots$

Similar to the above work, we get

$$\frac{\partial}{\partial r} \ln(L) = n \ln(\lambda) - \frac{n}{\Gamma(r)} \frac{d}{dr} \Gamma(r) + \ln\left(\prod_i x_i\right) \equiv 0.$$

Note that $\Psi(r) \equiv \Gamma'(r)/\Gamma(r)$ is the *digamma* function.

At this point, substitute in $\hat{\lambda} = \hat{r}/\bar{X}$, and use a *computer* to search for the value of r that solves

$$n \ln(r/\bar{X}) - n\Psi(r) + \ln\left(\prod_i x_i\right) \equiv 0.$$

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Unif}(0, \theta)$. Find the MLE for θ .

First of all, the p.d.f. is $f(x) = 1/\theta$, $0 < x < \theta$, and you need to beware of the funny limits.

In any case,

$$L(\theta) = \prod_{i=1}^n f(x_i) = \begin{cases} 1/\theta^n & \text{if } 0 \leq x_i \leq \theta, \forall i \\ 0 & \text{otherwise} \end{cases}$$

In order to have $L(\theta) > 0$, we must have $0 \leq x_i \leq \theta$, $\forall i$. In other words, we must have $\theta \geq \max_i x_i$.

Subject to this constraint, $L(\theta) = 1/\theta^n$ is maximized at the smallest possible θ value, namely, $\hat{\theta} = \max_i X_i$.

This makes sense in light of the similar (unbiased) estimator, $Y_2 = \frac{n+1}{n} \max_i X_i$, from the previous module.

Remark: We used very little calculus in this example!

Invariance Property of MLE's

Theorem (Invariance Property): If $\hat{\theta}$ is the MLE of some parameter θ and $h(\cdot)$ is a one-to-one function, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

Remark: We noted before that such a property does *not* hold for unbiasedness. For instance, although $E[S^2] = \sigma^2$, it is usually the case that $E[\sqrt{S^2}] \neq \sigma$.

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$.

We saw that the MLE for σ^2 is $\widehat{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$.

If we consider the one-to-one function $h(y) = +\sqrt{y}$, then the invariance property says that the MLE of σ is

$$\widehat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}.$$

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Exp}(\lambda)$.

We saw that the MLE for λ is $\hat{\lambda} = 1/\bar{X}$.

Meanwhile, we define the **survival function** as

$$\bar{F}(x) = \Pr(X > x) = 1 - F(x) = e^{-\lambda x}.$$

Then the invariance property says that the MLE of $\bar{F}(x)$ is

$$\widehat{\bar{F}}(x) = e^{-\hat{\lambda}x} = e^{-x/\bar{X}}.$$

The Method of Moments

Recall: The k th **moment** of a RV X is

$$E[X^k] = \begin{cases} \sum_x x^k f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathfrak{R}} x^k f(x) dx & \text{if } X \text{ is cts} \end{cases}$$

Definition: Suppose X_1, \dots, X_n are i.i.d. from p.d.f. / p.m.f. $f(x)$. Then the **method of moments** (MOM) estimator for $E[X^k]$ is $\sum_{i=1}^n X_i^k / n$.

Examples:

The MOM estimator for $\mu = E[X_i]$ is $\bar{X} = \sum_{i=1}^n X_i/n$.

The MOM estimator for $E[X_i^2]$ is $\sum_{i=1}^n X_i^2/n$.

The MOM estimator for $\text{Var}(X_i) = E[X_i^2] - (E[X_i])^2$ is

$$\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n} = \frac{n-1}{n} S^2.$$

(Of course, it's also OK to use S^2 .)

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Pois}(\lambda)$.

Since $\lambda = E[X_i]$, a MOM estimator for λ is \bar{X} .

But also note that $\lambda = \text{Var}(X_i)$, so another MOM estimator for λ is $\frac{n-1}{n}S^2$ (or plain old S^2).

Usually use the easier-looking estimator if you have a choice.

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$.

MOM estimators for μ and σ^2 are \bar{X} and $\frac{n-1}{n}S^2$ (or S^2), respectively.

For this example, these estimators are the same as the MLE's.

Let's finish up with a less-trivial example...

Example: Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Beta}(a, b)$.

The p.d.f. is

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \quad 0 < x < 1.$$

It turns out (after lots of alg) that

$$E[X] = \frac{a}{a+b}, \quad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Let's estimate a and b via MOM.

We have

$$\mathbb{E}[X] = \frac{a}{a+b} \Rightarrow a = \frac{b\mathbb{E}[X]}{1-\mathbb{E}[X]} \approx \frac{b\bar{X}}{1-\bar{X}}, \quad (*)$$

so

$$\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)} = \frac{\mathbb{E}[X]b}{(a+b)(a+b+1)}.$$

Plug into the above \bar{X} for $\mathbb{E}[X]$, S^2 for $\text{Var}(X)$, and $\frac{b\bar{X}}{1-\bar{X}}$ for a .

We can now solve for b (tho it'll take lots of alg).

After some work, we get

$$b \approx \frac{(1 - \bar{X})^2 \bar{X}}{S^2} - 1 + \bar{X}.$$

To finish up, you can plug back into (*) to get the MOM estimator for a .

Example (Hayter): Suppose we take a bunch of observations from a Beta distrn and it turns out that $\bar{X} = 0.3007$ and $S^2 = 0.01966$. Then the MOM estimators for a and b are 2.92 and 6.78, respectively.