STAT 414 / 415

Probability Theory and Mathematical Statistics

Maximum Likelihood Estimation

Printer-friendly version (../../print/book/export/html/191/)

Statement of the Problem

Suppose we have a random sample $X_1, X_2,..., X_n$ whose assumed probability distribution depends on some unknown parameter θ. Our primary goal here will be to find a point estimator $u(X_1, X_2,..., X_n)$, such that $u(x_1, x_2,..., x_n)$ is a "good" point estimate of θ, where $x_1, x_2,..., x_n$ are the observed values of the random sample. For example, if we plan to take a random sample $X_1, X_2,..., X_n$ for which the X_i are assumed to be normally distributed with mean μ and variance σ^2, then our goal will be to find a good estimate of μ, say, using the data $x_1, x_2,..., x_n$ that we obtained from our specific random sample.

The Basic Idea

It seems reasonable that a good estimate of the unknown parameter θ would be the value of θ that **maximizes** the probability, errrr... that is, the **likelihood**... of getting the data we observed. (So, do you see from where the name "maximum likelihood" comes?) So, that is, in a nutshell, the idea behind the method of maximum likelihood estimation. But how would we implement the method in practice? Well, suppose we have a random sample $X_1, X_2,..., X_n$ for which the probability density (or mass) function of each X_i is $f(x_i; \theta)$. Then, the joint probability mass (or density) function of $X_1, X_2,..., X_n$, which we'll (not so arbitrarily) call $\boldsymbol{L(\theta)}$ is:

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

The first equality is of course just the definition of the joint probability mass function. The second equality comes from that fact that we have a random sample, which implies by definition that the X_i are independent. And, the last equality just uses the shorthand mathematical notation of a product of indexed terms. Now, in light of the basic idea of maximum likelihood estimation, one reasonable way to proceed is to treat the "**likelihood function**" $L(\theta)$ as a function of θ, and find the value of θ that maximizes it.

Is this still sounding like too much abstract gibberish? Let's take a look at an example to see if we can make it a bit more concrete.

Example

Suppose we have a random sample $X_1, X_2,..., X_n$ where:

- X_i = 0 if a randomly selected student does not own a sports car, and
- X_i = 1 if a randomly selected student does own a sports car.

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

Solution. If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

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

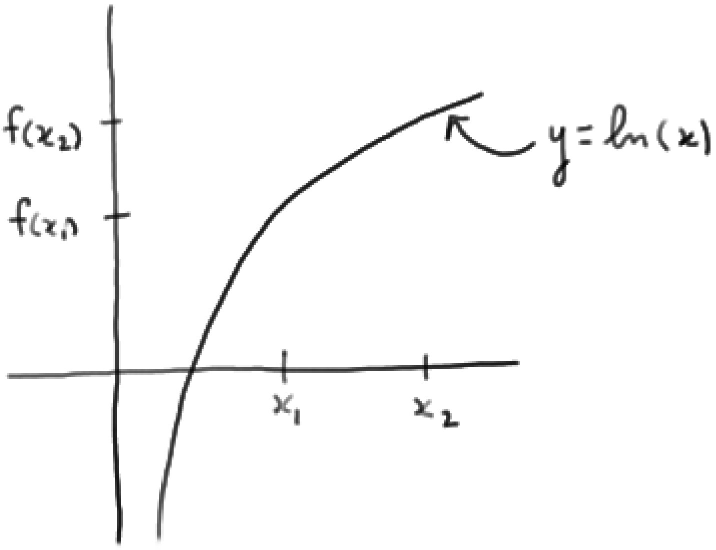
for x_i = 0 or 1 and 0 < p < 1. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p)=\prod_{i=1}^{n} f(x_i;p)=p^{x_1}(1-p)^{1-x_1}\times p^{x_2}(1-p)^{1-x_2}\times\cdots\times p^{x_n}(1-p)^{1-x_n}$$

for 0 < p < 1. Simplifying, by summing up the exponents, we get :

$$L(p)=p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:



That is, if $x_1 < x_2$, then $f(x_1) < f(x_2)$. That means that the value of p that maximizes the natural logarithm of the likelihood function $ln(L(p))$ is also the value of p that maximizes the likelihood function $L(p)$. So, the "trick" is to take the derivative of $ln(L(p))$ (with respect to p) rather than taking the derivative of $L(p)$. Again, doing so often makes the differentiation much easier. (By the way, throughout the remainder of this course, I will use either $ln(L(p))$ or $log(L(p))$ to denote the natural logarithm of the likelihood function.)

In this case, the natural logarithm of the likelihood function is:

$$\log L(p)=(\sum x_i)\log(p)+(n-\sum x_i)\log(1-p)$$

Now, taking the derivative of the log likelihood, and setting to 0, we get:

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} \overset{SET}{\equiv} 0$$

Now, multiplying through by $p(1-p)$, we get:

$$(\sum x_i)(1-p) - (n - \sum x_i)p = 0$$

Upon distributing, we see that two of the resulting terms cancel each other out:

$$\sum x_i - p\cancel{\sum x_i} - np + p\cancel{\sum x_i} = 0$$

leaving us with:

$$\sum x_i - np = 0$$

Now, all we have to do is solve for p. In doing so, you'll want to make sure that you always put a hat ("^") on the parameter, in this case p, to indicate it is an estimate:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or, alternatively, an estimator:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Oh, and we should technically verify that we indeed did obtain a maximum. We can do that by verifying that the second derivative of the log likelihood with respect to p is negative. It is, but you might want to do the work to convince yourself!

Now, with that example behind us, let us take a look at formal definitions of the terms (1) likelihood function, (2) maximum likelihood estimators, and (3) maximum likelihood estimates.

Definition. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$. Suppose that $(\theta_1, \theta_2, \ldots, \theta_m)$ is restricted to a given parameter space Ω. Then:

(1) When regarded as a function of $\theta_1, \theta_2, \ldots, \theta_m$, the joint probability density (or mass) function of $X_1, X_2, \ldots, X_n$:

$$L(\theta_1, \theta_2, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$$

($(\theta_1, \theta_2, \ldots, \theta_m)$ in Ω) is called the **likelihood function**.

(2) If:

$$[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)]$$

is the *m*-tuple that maximizes the likelihood function, then:

$$\hat{\theta}_i = u_i(X_1, X_2, \ldots, X_n)$$

is the **maximum likelihood estimator** of θ_i, for *i* = 1, 2, ..., *m*.

(3) The corresponding observed values of the statistics in (2), namely:

$$[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)]$$

are called the **maximum likelihood estimates** of θ_i, for *i* = 1, 2, ..., *m*.

Example

Suppose the weights of randomly selected American female college students are normally distributed with unknown mean *μ* and standard deviation *σ*. A random sample of 10 American female college students yielded the following weights (in pounds):

```
115   122   130   127   149   160
152   138  149   180
```

Based on the definitions given above, identify the likelihood function and the maximum likelihood estimator of *μ*, the mean weight of all American female college students. Using the given sample, find a maximum likelihood estimate of *μ* as well.

Solution. The probability density function of X_i is:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

for −∞ < *x* < ∞. The parameter space is Ω = {(*μ*, *σ*): −∞ < *μ* < ∞ and 0 < *σ* < ∞}. Therefore, (you might want to convince yourself that) the likelihood function is:

$$L(\mu, \sigma) = \sigma^{-n}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

for −∞ < *μ* < ∞ and 0 < *σ* < ∞. It can be shown (we'll do so in the next example!), upon maximizing the likelihood function with respect to *μ*, that the maximum likelihood estimator of *μ* is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Based on the given sample, a maximum likelihood estimate of *μ* is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(115 + \cdots + 180) = 142.2$$

pounds. Note that the only difference between the formulas for the maximum likelihood estimator and the maximum likelihood estimate is that:

- the estimator is defined using capital letters (to denote that its value is random), and
- the estimate is defined using lowercase letters (to denote that its value is fixed and based on an obtained sample)

Okay, so now we have the formal definitions out of the way. The first example on this page involved a joint probability mass function that depends on only one parameter, namely *p*, the proportion of successes. Now, let's take a look at an example that involves a joint probability density function that depends on two parameters.

Example

Let X_1, X_2,..., X_n be a random sample from a normal distribution with unknown mean μ and variance σ^2. Find maximum likelihood estimators of mean μ and variance σ^2.

Solution. In finding the estimators, the first thing we'll do is write the probability density function as a function of $\theta_1 = \mu$ and $\theta_2 = \sigma^2$:

$$f(x_i;\theta_1,\theta_2)=\frac{1}{\sqrt{\theta_2}\sqrt{2\pi}}\text{exp}\left[-\frac{(x_i-\theta_1)^2}{2\theta_2}\right]$$

for $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. We do this so as not to cause confusion when taking the derivative of the likelihood with respect to σ^2. Now, that makes the likelihood function:

$$L(\theta_1,\theta_2)=\prod_{i=1}^{n}f(x_i;\theta_1,\theta_2)=\theta_2^{-n/2}(2\pi)^{-n/2}\text{exp}\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right]$$

and therefore the log of the likelihood function:

$$\text{log}L(\theta_1,\theta_2)=-\frac{n}{2}\text{log}\theta_2-\frac{n}{2}\text{log}(2\pi)-\frac{\sum(x_i-\theta_1)^2}{2\theta_2}$$

Now, upon taking the partial derivative of the log likelihood with respect to θ_1, and setting to 0, we see that a few things cancel each other out, leaving us with:

$$\frac{\partial \log L(\theta_1,\theta_2)}{\partial\theta_1}=\frac{\cancel{-2}\sum(x_i-\theta_1)(\cancel{-1})}{\cancel{2}\theta_2}\overset{\text{SET}}{=}0$$

Now, multiplying through by θ_2, and distributing the summation, we get:

$$\sum x_i - n\theta_1 = 0$$

Now, solving for θ_1, and putting on its hat, we have shown that the maximum likelihood estimate of θ_1 is:

$$\hat{\theta}_1=\hat{\mu}=\frac{\sum x_i}{n}=\bar{x}$$

Now for θ_2. Taking the partial derivative of the log likelihood with respect to θ_2, and setting to 0, we get:

$$\frac{\partial \log L(\theta_1,\theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{\Sigma(x_i-\theta_1)^2}{2\theta_2^2} \overset{SET}{\equiv} 0$$

Multiplying through by $2\theta_2^2$:

$$\frac{\partial \log L(\theta_1,\theta_2)}{\partial \theta_2} = \left[-\frac{n}{2\theta_2} + \frac{\Sigma(x_i-\theta_1)^2}{2\theta_2^2} \overset{SET}{\equiv} 0 \right] \times 2\theta_2^2$$

we get:

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

And, solving for θ_2, and putting on its hat, we have shown that the maximum likelihood estimate of θ_2 is:

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

(I'll again leave it to you to verify, in each case, that the second partial derivative of the log likelihood is negative, and therefore that we did indeed find maxima.) In summary, we have shown that the maximum likelihood estimators of μ and variance σ^2 for the normal model are:

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

respectively.

Note that the maximum likelihood estimator of σ^2 for the normal model is not the sample variance S^2. They are, in fact, competing estimators. So how do we know which estimator we should use for σ^2 ? Well, one way is to choose the estimator that is "unbiased." Let's go learn about unbiased estimators now.

‹ Definitions (../190/) up (../189/) Unbiased Estimation › (../192/)