

Probability and Statistics

Fall 2000

LECT. 11: METHOD OF MOMENT AND MAXIMUM LIKELIHOOD ESTIMATION,
CB 7.1,7.2.1, 7.2.2

From now on we consider a different approach to estimation. The context is the following: We have a random sample of size N from a distribution with pdf or pmf $f(x; \theta^*)$. We know that $\theta^* \in \Theta$, but we do not know the actual value of θ^* . In the Bayesian approach we looked for the conditional distribution of the unknown parameters given the data. Now, we look at the conditional distribution of the data given the parameter. This will have the advantage that we do not have to come up with a prior distribution. We look at functions of the observations, statistics, and look for statistics that are good point estimators. By good we mean good according to repeated sampling criteria: we look at the behavior (distribution) of the estimator when we repeatedly get new random samples of the same size. We would like the distribution of the estimator to be concentrated around θ^* .

Note: We use different notation here. We use θ^* here to denote the true value of the parameter that generated the data, and use θ to denote any element of the parameter space. This distinction will be important as we will evaluate the density function sometimes at the true value of the parameter, and sometimes at arbitrary values.

The first approach to systematically find estimators is the method of moments. Consider a set of independent and identically distributed random variables with a pdf/pmf $f(x; \theta)$. Suppose θ is a scalar. The mean of this distribution is

$$E[X] = g(\theta^*),$$

where

$$g(\theta) = \int x \cdot f(x; \theta) dx.$$

The function $g(\cdot)$ is clearly a known function given knowledge of $f(\cdot)$. Now suppose that

we calculate the average of our random sample, $\bar{x} = \sum x_i/N$. With N reasonably large this average should be close to $g(\theta^*)$. Therefore the value of θ that solves

$$g(\theta) = \bar{x},$$

would appear to be a sensible estimator for θ^* .

Let us look at some examples. Suppose that X_i has a Bernoulli distribution with probability p^* . Then the expectation of X is p^* , so the method of moment estimator solves

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

This is an unbiased estimator: its expectation is equal to the unknown parameter.

Suppose that X has an exponential distribution with arrival rate λ . The expectation of X is $1/\lambda$, so the method of moments estimator solves:

$$g(\hat{\lambda}) = 1/\hat{\lambda} = \bar{X},$$

of

$$\hat{\lambda} = 1/\bar{X}.$$

This estimator is not unbiased: the expectation of X is $1/\lambda$, so the expectation of $1/X$ is greater than λ by Jensen's inequality.

Result 1 (JENSEN'S INEQUALITY)

Let $g(x)$ be a convex function. Then $E[g(X)] \geq g(E[X])$, with equality only in the case of a linear function.

Proof: This is for the continuous case, with $g(x)$ twice continuously differentiable, so that $g''(x) \geq 0$. We can expand $g(x)$ around $E[X]$:

$$g(x) = g(E[X]) + g'(E[X]) \cdot (x - E[X]) + g''(\tilde{x}) \cdot (x - E[X])^2,$$

for some \tilde{x} in between x and $E[X]$. Because of the convexity, the last term is non-positive, so

$$g(x) \leq g(E[X]) + g'(E[X]) \cdot (x - E[X]).$$

Then

$$\begin{aligned} E[g(X)] &\leq E[g(E[X]) + g'(E[X]) \cdot (x - E[X])] \\ &= g(E[X]) + g'(E[X]) \cdot E[(x - E[X])] = g(E[X]), \end{aligned}$$

which completes the proof. \square

Now suppose we have two parameters, that is, θ is a two dimensional vector. In that we calculate the first two moments:

$$g_1(\theta) = E[X] = \int x \cdot f_X(x; \theta) dx,$$

$$g_2(\theta) = E[X^2] = \int x^2 \cdot f_X(x; \theta) dx,$$

and equate them with the corresponding sample moments:

$$g_1(\hat{\theta}) = \bar{x},$$

$$g_2(\hat{\theta}) = \overline{x^2}.$$

For example, suppose that X has a normal distribution with mean μ and variance σ^2 . The $g(\cdot)$ functions are:

$$g_1(\mu, \sigma^2) = \int x f_X(x; \mu, \sigma^2) dx = \mu,$$

$$g_2(\mu, \sigma^2) = \int x^2 f_X(x; \mu, \sigma^2) dx = \mu^2 + \sigma^2.$$

Hence the method of moments estimators are

$$\hat{\mu} = \bar{x},$$

$$\hat{\sigma}^2 = \overline{x^2} - (\bar{x})^2.$$

Again these appear reasonable and intuitive estimators. In other cases the estimators are somewhat less obvious and less attractive. Consider the case where the distribution of X is multinomial with parameters k and p both unknown (a relatively unusual problem in practice). The first two moments are:

$$g_1(k, p) = \sum x f_X(x; k, p) = k \cdot p,$$

$$g_2(k, p) = \sum x^2 f_X(x; k, p) = k \cdot p \cdot (1 - p) + k^2 \cdot p^2.$$

Thus we solve

$$\bar{x} = \hat{k} \cdot \hat{p},$$

$$\overline{x^2} = \hat{k} \cdot \hat{p} \cdot (1 - \hat{p}) + \hat{k}^2 \cdot \hat{p}^2.$$

The solutions are:

$$\hat{p} = \bar{x} / \hat{k},$$

and

$$\hat{k} = \frac{\bar{x} \cdot \overline{x^2}}{\overline{x^2} + \bar{x} - \bar{x}^2}.$$

A problem is that these estimates can in fact be negative if the variance is larger than the mean, which is obviously not such a good estimate for a probability and a number of trials.

The second general approach to estimation we consider is maximum likelihood estimation. The likelihood function is the density function viewed as a function of the unknown parameters, rather than as a function of the random variable. Let X be a random variable with pdf/pmf

$$f_X(x; \theta^*),$$

where θ^* is the unknown true value of the parameter θ . The likelihood function is then:

$$\mathcal{L}(\theta) = f_X(x; \theta).$$

If we have more than one random variable, say X_1, X_2, \dots, X_N , the likelihood function is based on the joint probability density/mass function:

$$\mathcal{L}(\theta) = f_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \theta).$$

If the random variables are independent and identically distributed, the likelihood function obviously simplifies:

$$\mathcal{L}(\theta) = f_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \theta) = \prod_{i=1}^N f_X(x_i; \theta).$$

Often we prefer to work with the logarithm of the likelihood function, the log likelihood function, e.g., in the case of N independent and identically distributed random variables:

$$L(\theta) = \ln \mathcal{L}(\theta) = \ln f_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_N; \theta) = \sum_{i=1}^N \ln f_X(x_i; \theta).$$

The maximum likelihood estimator or mle for θ^* is the value of θ that maximizes the likelihood function, or equivalently, the log likelihood function (because the logarithm is a one to one, strictly monotone, transformation the maximand of one is the maximand of the other.)

Let us first look at a simple example. Suppose X has a normal distribution with unknown mean μ and known variance 1. The probability density function is

$$f_X(x; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu)^2\right).$$

The likelihood function is

$$\mathcal{L}(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu)^2\right).$$

The log likelihood function is

$$L(\mu) = \ln \mathcal{L}(\mu) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2}(x - \mu)^2.$$

The value of μ that maximizes the log likelihood function is $\hat{\mu}_{mle} = x$.

Now suppose we have N random variables independent and all normally distributed with mean μ and variance 1. In that case the likelihood function is

$$\mathcal{L}(\mu) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right),$$

and the log likelihood function is

$$L(\mu) = -N \frac{1}{2} \ln(2\pi) - \sum_{i=1}^N \frac{1}{2}(x_i - \mu)^2.$$

The maximum likelihood estimator is $\hat{\mu}_{mle} = \bar{x}$, the sample average.

In both these cases the resulting estimator appears very reasonable, similar to the method of moments estimators for this case. In general, however, it is not obvious why maximizing the likelihood function is a sensible strategy. Here are two motivations.

First, compare the mle to the posterior distributions considered before. If we use a flat prior distribution, that is, a prior distribution that does not change with the value of the parameter θ , the maximum likelihood estimator corresponds to the posterior mode. If the flat prior based posterior distribution is approximately normal, and the log likelihood function therefore approximately quadratic, it also corresponds approximately to the posterior mean.

A second motivation is more direct and relies on large sample approximations. First we need to recall Jensen's inequality: For a convex function $g(\cdot)$, and any random variable X ,

the expected value of $g(X)$ is greater than or equal to the function evaluated at the expected value of X :

$$E[g(X)] \geq g(E[X]).$$

Now consider the random variable Y defined as the ratio of the density function at some arbitrary value of θ to the density function at θ^* , both evaluated at the random variable X :

$$Y = f_X(X; \theta) / f_X(X; \theta^*).$$

Take $g(\cdot)$ to be minus the logarithmic function: $g(a) = -\ln(a)$, so $g'(a) = -1/a$, and $g''(a) = 1/a^2 > 0$ and $g(\cdot)$ is convex, Then, by Jensen's inequality

$$E[-\ln Y] \geq -\ln E[Y],$$

implying

$$E\left[-\ln\left(\frac{f_X(X; \theta)}{f_X(X; \theta^*)}\right)\right] \geq -\ln\left(E\left[\frac{f_X(X; \theta)}{f_X(X; \theta^*)}\right]\right),$$

where the expectation is over the distribution of X , that is the density $f_X(x; \theta^*)$. The expectation on the right therefore simplifies:

$$E\left[\frac{f_X(X; \theta)}{f_X(X; \theta^*)}\right] = \int \frac{f_X(x; \theta)}{f_X(x; \theta^*)} \cdot f_X(x; \theta^*) dx = \int f_X(x; \theta) dx = 1,$$

for all values of θ , so that after taking the log, the righthand side is equal to zero, and thus

$$E\left[-\ln\left(\frac{f_X(X; \theta)}{f_X(X; \theta^*)}\right)\right] \geq 0$$

implying

$$E[\ln f_X(X; \theta)] \leq E[\ln f_X(X; \theta^*)].$$

This implies that the expected value of the log likelihood is maximized at the true value of θ , and therefore there is some hope that the actual log likelihood function is maximized at a value close to θ^* , and therefore at a value that is a good estimate of θ^* .

Note that we always take the expectation over the true distribution, $f(x; \theta^*)$, even if we are taking expectations of functions of θ , evaluated at all possible values of θ , and in particular evaluated at values other than the true value θ^* .

Let us look at an example that illustrates this argument. Let X_1, X_2, \dots, X_N be Bernoulli trials with unknown probability p^* . The joint density is

$$f_{X_1, X_2, \dots, X_N}(x_1, \dots, x_N; p) = p^{\sum x_i} (1-p)^{N-\sum x_i}.$$

The log likelihood function is

$$L(p) = \sum x_i \ln p + (N - \sum x_i) \cdot \ln(1-p).$$

The value that maximizes the log likelihood function is the sample average $\hat{p}_{mle} = \bar{x}$. Now consider the expected value of the log likelihood function:

$$\begin{aligned} E[L(p)] &= E \sum X_i \ln p + (N - \sum X_i) \cdot \ln(1-p), \\ &= N \cdot p^* \cdot \ln(p) + N \cdot (1-p^*) \cdot \ln(1-p). \end{aligned}$$

This expected log likelihood function is maximized at $p = p^*$, which is the true value. So, while the actual log likelihood function is not necessarily maximized at p^* , its expectation is. If we multiply the log likelihood function by $1/N$ it is a sample average and as such subject to law of large numbers. (multiplying the log likelihood function by a constant does not affect the location of the maximum.) Hence its expectation may be reasonably close to the average, and that is one motivation for considering maximum likelihood estimators.

Let us consider two more examples. Suppose that X_1, X_2, \dots, X_N are iid with uniform distribution on the interval zero to θ^* . The likelihood function is

$$\mathcal{L}(\theta) = (1/\theta^N) \cdot 1_{\{\max\{x_i\} \leq \theta\}}.$$

The likelihood function is strictly decreasing for values where it differs from zero. The maximum likelihood estimator is equal to $\max_i\{x_i\}$.

Suppose X_1, \dots, X_N are normal with unknown mean μ and unknown variance σ^2 . The likelihood function is

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right),$$

with logarithm

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

To maximize this over μ and σ^2 jointly we try to solve the first order conditions:

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

$$\frac{\partial L}{\partial \sigma^2}(\mu, \sigma^2) = \sum_{i=1}^N -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x_i - \mu)^2.$$

Setting both derivatives equal to zero leads to

$$\hat{\sigma}_{mle}^2 = \sum_{i=1}^N (x_i - \bar{x})^2 / N,$$

and, as before

$$\hat{\mu}_{mle} = \bar{x}.$$

In principle we also have to check that the matrix of second derivatives is negative definite to ensure that we have located a maximum and not a minimum or saddle point. In this case the matrix of second derivatives is indeed negative definite and this is the global maximum.

Finally, an important property of maximum likelihood estimators is their invariance: If $\hat{\theta}_{mle}$ is the maximum likelihood estimator for θ^* , then $\hat{\pi}_{mle} = g(\hat{\theta}_{mle})$ is the maximum likelihood estimator for $\pi^* = g(\theta^*)$.

For example of this consider the exponential distribution. Let X_1, X_2, \dots, X_N be iid with an exponential distribution with arrival rate λ . The log likelihood function is:

$$L(\lambda) = N \cdot \ln \lambda - \sum x_i \cdot \lambda.$$

The mle is $\hat{\lambda}_{mle} = 1/\bar{x}$. Now suppose we are interested in the mle for the mean of this distribution, and therefor parametrize the distribution in terms of $\mu = 1/\lambda$. we can go back to the likelihood function:

$$L(\mu) = -N/\ln \mu - \sum x_i/\mu,$$

and we end up with the maximum likelihood estimator.

$$\hat{\mu}_{mle} = \bar{x}.$$

The invariance property implies the same result: $\hat{\mu}_{mle} = 1/\hat{\lambda}_{mle} = (1/(1/\bar{x})) = \bar{x}$.