2 Point Estimation and the Method of Maximum Likelihood

2.1 Point estimation

Given a vector of observations $\mathbf{X}$ from a family $\mathcal{F}$ of probability distributions, we shall often wish to estimate the underlying value of the vector of parameters, $\theta = (\theta_1, \theta_2, \dots, \theta_q)$, where $\theta \in \Theta$.

Definition
A *(point) estimator* $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ is a function of the observations, mapping the range, $\mathcal{X}$, of $\mathbf{X}$ into the parameter space Θ.

Note that an estimator is a special case of a statistic. Once a particular value $\mathbf{x}$ of $\mathbf{X}$ has been observed, the corresponding value of the estimator, $\hat{\theta}(\mathbf{x})$, is known as an *estimate.* Sometimes, an estimator is simply referred to as an estimate.

In simple cases, there may be obvious natural estimators to use. For example, if $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)$ is a random sample of size n from a $N(\mu, \sigma^2)$ distribution, with σ^2 known, then the natural estimator $\hat{\mu}$ of μ is given by

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

However, we might wish to be reassured that this is in some sense the best estimator to use.

In other situations, it may not be at all obvious how to estimate the unknown parameters. In general, there are a number of different methods available for finding suitable estimators or for finding best estimators according to some optimality criterion.

2.2 The method of moments

This is an old and simple method for quickly finding an estimator. It can produce estimators with good properties, but it can also go badly wrong. The method is often regarded as providing a first attempt at finding a suitable estimator. It may sometimes be used to find a starting point for an iterative procedure that finds a better estimate.

Let $\mathbf{X} \equiv (X_1, X_2, \dots, X_n)$ be a random sample of size n from some family of distributions depending upon the vector of q parameters, $\theta = (\theta_1, \theta_2, \dots, \theta_q)$. Assume that the first q moments $\mu_j,\ j = 1, \dots, q$ of the distributions exist,

$$\mu_j(\theta_1, \theta_2, \dots, \theta_q) = E[X_i^j; \theta_1, \theta_2, \dots, \theta_q] \qquad j = 1, \dots, q.$$

Let $m_j,\ j = 1, \dots, q$ be the corresponding sample moments,

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j \qquad j = 1, \dots, q.$$

According to the method of moments, we match the sample and population/distribution moments. We solve the following set of q simultaneous equations for the q unknown parameter values,

$$\begin{aligned} m_1 &= \mu_1(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q) \\ m_2 &= \mu_2(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q) \\ &\vdots \\ m_q &= \mu_q(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q), \end{aligned} \tag{1}$$

to obtain the estimator $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q)$.

- Note: the existence and uniqueness of a solution $\hat{\theta} \in \Theta$ of the set of Equations (1) is not guaranteed.

The first moment is just the mean, so that the first of the Equations (1) equates the sample and population means and may be written as

$$\bar{X} = \mu(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q). \tag{2}$$

The population variance σ^2 satisfies

$$\sigma^2 = \mu_2 - \mu_1^2.$$

Correspondingly, for the sample values,

$$\hat{\sigma}^2 = m_2 - m_1^2,$$

where

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

Subtracting the square of the first of the Equations (1) from the second of the Equations (1), we obtain

$$\hat{\sigma}^2 = \sigma^2(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q), \tag{3}$$

which we may use as an alternative to the second of the Equations (1).

- In Equation (3) we equate the sample variance (defined with denominator n) and the population variance.

Example 1

For a random sample $X_1, X_2, \ldots, X_n$ of size n from a $N(\mu, \sigma^2)$ distribution with the two unknown parameters μ, σ^2, we find directly from Equations (2) and (3) that the method of moments gives the estimators

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

and

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

Example 2

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, i.e., with p.d.f.

$$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \qquad x > 0.$$

This distribution has mean given by

$$\mu = \frac{\alpha}{\beta}$$

and variance given by

$$\sigma^2 = \frac{\alpha}{\beta^2}.$$

Using the method of moments, the estimators of the parameters α and β are obtained from the Equations (2) and (3), which become

$$\bar{X} = \frac{\hat{\alpha}}{\hat{\beta}} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{\hat{\alpha}}{\hat{\beta}^2},$$

respectively. Hence

$$\hat{\beta} = \frac{\bar{X}}{\hat{\sigma}^2} \qquad \text{and} \qquad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}.$$

2.3 The method of maximum likelihood

Definition

For each $\mathbf{x} \in \mathcal{X}$, let $\hat{\theta}(\mathbf{x})$ be such that, with $\mathbf{x}$ held fixed, the likelihood function $\{L(\theta; \mathbf{x}) : \theta \in \Theta\}$ attains its maximum value as a function of θ at $\hat{\theta}(\mathbf{x})$. The estimator $\hat{\theta}(\mathbf{X})$ is then said to be a *maximum likelihood estimator* (MLE) of θ.

- Given any observed value $\mathbf{x}$, the value of a MLE, $\hat{\theta}(\mathbf{x})$, can be interpreted as the most likely value of θ to have given rise to the observed $\mathbf{x}$ value.

- To find a MLE, it is often more convenient to maximize the log likelihood function, $\ln L(\theta; \mathbf{x})$, which is equivalent to maximizing the likelihood function.

- It should be noted that a MLE may not exist — there may be an $\mathbf{x} \in \mathcal{X}$ such that there is no θ that maximises the likelihood function $\{L(\theta; \mathbf{x}) : \theta \in \Theta\}$.

- The MLE, if it does exist, may not be unique — given $\mathbf{x} \in \mathcal{X}$, there may be more than one value of θ that maximises the likelihood function $\{L(\theta; \mathbf{x}) : \theta \in \Theta\}$.

- There will be simple cases in which the MLE may be found analytically, possibly just using the standard methods of calculus. In other cases, especially if the dimension q of the vector of parameters is large, the maximisation will have to be carried out using numerical methods within computer packages.

- The maximum of the likelihood function may be in the interior of Θ, but there will be cases in which it lies on the boundary of Θ. In the former case, routine analytical or numerical methods are generally available, but the possibility that the latter case may occur has to be borne in mind.

Assuming that the likelihood function $L(\theta; \mathbf{x})$ is a continuously differentiable function of θ, given $\mathbf{x} \in \mathcal{X}$, an interior stationary point of $\ln L(\theta; \mathbf{x})$ or of $L(\theta; \mathbf{x})$ is given by a solution of the *likelihood equations*,

$$\frac{\partial \ln L(\theta; \mathbf{x})}{\partial \theta_j} = 0, \qquad j = 1, \ldots, q. \tag{4}$$

A solution of Equations (4) may or may not be unique, and may or may not give us a MLE. In many of the standard cases, solution of the Equations (4) does give us the MLE, but we cannot take this for granted.

We may wish to check that a solution of Equation (4) gives at least a local maximum of the likelihood function. If $L(\theta; \mathbf{x})$ is twice continuously differentiable, the criterion to check is that the Hessian matrix, the matrix of second order partial derivatives,

$$\frac{\partial^2 \ln L(\theta; \mathbf{x})}{\partial \theta_j \partial \theta_k} \qquad j, k = 1, \ldots, q.$$

is negative definite at the solution point.

However, even if this condition can be shown to be satisfied, it does not prove that the solution point is a global maximum. If possible, it is better to use alternative, direct methods of proving that a global maximum has been found.

Example 1 (continued)

For a random sample $X_1, X_2, \ldots, X_n$ of size n from a $N(\mu, \sigma^2)$ distribution, we found in Chapter 1 that the log likelihood function is given by

$$\ln L(\mu, \sigma^2; \mathbf{x}) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(s_x^2 + n(\bar{x} - \mu)^2), \tag{5}$$

where $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$. We could find the MLE by solving the likelihood equations, but the following argument provides a clear-cut demonstration that we really have found the MLE.

For any given $\sigma^2 > 0$, the expression (5) is maximized with respect to μ by taking the value $\hat{\mu} = \bar{x}$ of μ. Using this fact, we reduce the problem of finding a MLE to the univariate problem of maximizing

$$\ln L(\hat{\mu}, \sigma^2; \mathbf{x}) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{s_x^2}{2\sigma^2} \tag{6}$$

with respect to $\sigma^2 > 0$. Differentiating the expression (6) w.r.t σ^2, we find

$$\begin{aligned} \frac{\partial \ln L(\hat{\mu}, \sigma^2; \mathbf{x})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{s_x^2}{2\sigma^4} \\ &= \frac{n}{2\sigma^4}\left(\frac{s_x^2}{n} - \sigma^2\right). \end{aligned}$$

Hence

$$\frac{\partial \ln L(\hat{\mu}, \sigma^2; \mathbf{x})}{\partial \sigma^2} \quad \begin{cases} > 0 & \text{for } \sigma^2 < s_x^2/n \\ = 0 & \text{for } \sigma^2 = s_x^2/n \\ < 0 & \text{for } \sigma^2 > s_x^2/n \end{cases}$$

Thus $\ln L(\hat{\mu}, \sigma^2; \mathbf{x})$ is a unimodal function of σ^2 which attains its maximum at the value $\hat{\sigma^2} = s_x^2/n$ of σ^2. It follows that the MLE of (μ, σ^2) is given by

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

which is exactly the same estimator as that given by the method of moments.

The following two theorems, which we state without proof, show two general properties of MLEs, which give support to the use of MLEs. The first theorem formally demonstrates that the method of maximum likelihood is consistent with the sufficiency principle, something that is not generally true, for example, for the method of moments.

Theorem 1 *Let $\mathbf{X}$ be a vector of observations from a family $\mathcal{F}$ of distributions and suppose that a MLE $\hat{\theta}$ exists and is unique, i.e, for each $\mathbf{x} \in \mathcal{X}$ there exists a unique $\hat{\theta}(\mathbf{x})$ that maximizes $\ln L(\theta; \mathbf{x})$. For any sufficient statistic T, the estimator $\hat{\theta}$ is a function of T.*

Rather than estimating the vector of parameters θ, we may wish to estimate some function $\tau(\theta)$ of the parameters. Another useful result about MLEs is the following.

Theorem 2 (The Invariance Property of MLEs) *If $\hat{\theta}$ is a MLE of θ then, for any function $\tau(\theta)$ of θ, $\tau(\hat{\theta})$ is a MLE of $\tau(\theta)$.*

To give a simple illustration of the use of Theorem 2, for a random sample of size n from a $N(\mu, \sigma^2)$ distribution, as we have seen, $\sum (X_i - \bar{X})^2/n$ is the MLE of σ^2. It follows that $\sqrt{\sum (X_i - \bar{X})^2/n}$ is the MLE of σ.

The invariance property of Theorem 2 will not in general hold for other types of estimator which we introduce later.

The MLE appears to be a sensible estimator to use, but how good is it? How does it compare with other estimators? We shall discuss later other properties of MLEs, which further justify their widespread use.

2.4 The mean squared error

Given observations from a family $\mathcal{F}$ of probability distributions, how do we evaluate and compare the various estimators of θ that are available to us? In the present discussion, we shall restrict attention to the estimation of a single parameter θ or, more generally, to a real-valued function $\tau(\theta)$ of the vector of parameters.

Definition

The *mean squared error* (MSE) of an estimator $\hat{\tau}$ of $\tau(\theta)$ is the function $\mathrm{MSE}(\hat{\tau}; \theta)$ of θ defined by

$$\mathrm{MSE}(\hat{\tau}; \theta) = E\left[(\hat{\tau}(\mathbf{X}) - \tau(\theta))^2; \theta \right] \qquad \theta \in \Theta.$$

- We are assuming the existence of finite second moments.
- For any given $\theta \in \Theta$, the MSE is the expectation of the square of the difference between the estimator $\hat{\tau}$ and $\tau(\theta)$. Broadly speaking, the smaller the MSE of an estimator the better.

- As is generally the case in statistical theory, the square of the difference is quite a mathematically tractable measure of discrepancy.

Definition

The *bias* of an estimator $\hat{\tau}$ of $\tau(\theta)$ is the function $b(\theta)$ defined by

$$b(\theta) = E\left[\hat{\tau};\theta\right] - \tau(\theta) \qquad\qquad \theta \in \Theta.$$

Definition

An estimator $\hat{\tau}$ for which $b(\theta) = 0$ for all $\theta \in \Theta$ is said to be an *unbiased* estimator of $\tau(\theta)$.

Equivalently, $\hat{\tau}$ is an unbiased estimator of $\tau(\theta)$ if and only if

$$E\left[\hat{\tau};\theta\right] = \tau(\theta) \qquad\qquad \theta \in \Theta.$$

In a frequentist approach to statistical inference, for any given $\theta \in \Theta$, imagine carrying out a sequence of independent repetitions of the statistical experiment that generates observations from a family $\mathcal{F}$ of distributions. In the long run, the average value of an unbiased estimator $\hat{\tau}$ is equal to the true value of the function $\tau(\theta)$ that it is estimating.

Unbiasedness is often regarded as a desirable property of an estimator. However, as we shall see, there may be good reasons to prefer a biased to an unbiased estimator. There is also the problem that an unbiased estimator does not always exist.

Theorem 3

$$MSE(\hat{\tau};\theta) = \text{var}(\hat{\tau};\theta) + b(\theta)^2 \qquad\qquad \theta \in \Theta.$$

Proof. For any random variable Y with a finite second moment and for any constant α, it is an elementary result that

$$E[(Y-\alpha)^2] = \text{var}(Y) + (\alpha - \mu)^2,$$

where μ is the mean of Y. The result of the lemma for any $\theta \in \Theta$ follows from taking $Y = \hat{\tau}(\mathbf{X})$ and $\alpha = \tau(\theta)$.

We see from the result of Theorem 3 that the MSE of an estimator may be thought of as the sum of two components, its variance and the square of its bias. The MSE of an unbiased estimator is identical with its variance. It is possible, however, for there to be a biased estimator that has not only a smaller variance but also a smaller MSE than an unbiased estimator for all $\theta \in \Theta$.

Given observations from a family $\mathcal{F}$ of probability distributions and a function $\tau(\theta)$ of the parameters, we might consider finding an estimator of $\tau(\theta)$ with minimum MSE. However, the MSE is a function of θ, and it will not in general be possible to find an estimator $\hat{\tau}$ that simultaneously minimises the MSE for all $\theta \in \Theta$, unless we impose some restriction on the class of estimators to be considered.

If we restrict attention to the class of unbiased estimators then there is a simple and elegant theory for finding *minimum variance unbiased estimators.* This will be dealt with in the next chapter.

Example 1 (continued)

For a random sample $X_1, X_2, \ldots, X_n$ of size n from a $N(\mu, \sigma^2)$ distribution with the two unknown parameters μ, σ^2, consider the estimation of σ^2. Restrict attention to the class of estimators $\hat{\tau}$ of σ^2 of the form

$$\hat{\tau} = aS_x^2,$$

where a is an arbitrary constant and $S_x^2 = \sum(X_i - \bar{X})^2$.

Recall that S_x^2/σ^2 has the χ^2_{n-1} distribution. It follows that

$$E[S_x^2] = (n-1)\sigma^2$$

and

$$\mathrm{var}(S_x^2) = 2(n-1)\sigma^4.$$

Hence

$$E[\hat{\tau}] = a(n-1)\sigma^2$$

and

$$\mathrm{var}(\hat{\tau}) = 2a^2(n-1)\sigma^4.$$

Thus the bias b of $\hat{\tau}$ as an estimator of σ^2 is given by

$$b(\mu, \sigma^2) = [a(n-1) - 1]\sigma^2. \tag{7}$$

From Theorem 3, the MSE is given by

$$\begin{aligned} \mathrm{MSE}(\hat{\tau}; \mu, \sigma^2) &= 2a^2(n-1)\sigma^4 + [a(n-1)-1]^2\sigma^4 \\ &= [(n^2-1)a^2 - 2(n-1)a + 1]\sigma^4 \end{aligned} \tag{8}$$

From Equation (7) we see that $\hat{\tau}$ is unbiased if and only if $a = 1/(n-1)$. A commonly used estimator of σ^2 is the unbiased estimator,

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

The corresponding MSE is $2\sigma^4/(n-1)$.

The MLE,

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

corresponds to $a = 1/n$ and is biased. However, from Equation (8) we see that its MSE is $(2n-1)\sigma^4/n^2$, less than the MSE of S^2.

Minimizing the expression of Equation (8) w.r.t. to a, we find that the MSE is minimized for all values of μ and σ^2 by taking $a = 1/(n+1)$. Thus, in the class of estimators that we are considering, the one with minimum MSE is

$$\frac{1}{n+1}\sum(X_i - \bar{X})^2.$$

The corresponding MSE is $2\sigma^4/(n+1)$.

A drawback to the use of the property of unbiasedness is that if $\hat{\theta}$ is an unbiased estimator of θ and $\tau(\theta)$ is some function of θ then it is not in general the case that $\tau(\hat{\theta})$ is an unbiased estimator of $\tau(\theta)$. Thus unbiased estimators do not have the invariance property of MLEs stated in Theorem 2. For instance, in Example 1, $S^2 \equiv \sum(X_i - \bar{X})^2/(n-1)$ is an unbiased estimator of σ^2, but S is not an unbiased estimator of σ.