5 Properties of Point Estimators

5.1 Fisher Information; Minimum Variance and Efficient Estimators

We are returning now to point estimators and discuss some of their important properties.
Let us recall our framework: we study a population characteristic X, with pdf $f(x;\theta)$, mean $E(X) = \mu$ and variance $V(X) = \sigma^2$. The *target parameter* θ is to be estimated based a random sample of size n, i.e. *sample variables* $X_1, \ldots, X_n$, which are independent and identically distributed (iid), having the same pdf as X.
A *point estimator* for (the estimation of) the target parameter θ is a sample function (statistic)

$$\overline{\theta} = \overline{\theta}(X_1, X_2, \ldots, X_n).$$

One of the first properties that we want in a point estimator is that it is *unbiased*,

$$E(\overline{\theta}) \;\; = \;\; \theta. \tag{5.1}$$

The sample mean, the sample moments of order k and the sample variance are examples of unbiased estimators for the corresponding population characteristic.

Another desirable trait for a point estimator is that its values do not vary too much from the value of the target parameter, i.e. that it has a *low variance*. This "low" variance property can be measured in several ways.

Definition 5.1. *An estimator $\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n)$ is called an **absolutely correct** estimator for θ, if it satisfies the conditions*

$$\begin{array}{ll} \text{(i)} & E(\overline{\theta}) = \theta, \\ \text{(ii)} & \lim\limits_{n\to\infty} V(\overline{\theta}) = 0. \end{array} \tag{5.2}$$

This is saying that the variance of the estimator decreases as the sample size increases, so the estimation gets better.

Remark 5.2. The sample mean $\overline{X}$ is an absolutely correct estimator for the theoretical mean $\mu = E(X)$, since (see Proposition 1.4 in Lecture 9)

$$V\left(\overline{X}\right) = \frac{\sigma^2}{n} \overset{n\to\infty}{\longrightarrow} 0.$$

Definition 5.3. *An unbiased estimator $\overline{\theta} = \overline{\theta}(X_1, \ldots, X_n)$ for θ is called a **minimum-variance unbiased estimator (MVUE)**, if it has lower variance than any other unbiased estimator for θ,*

$$V(\overline{\theta}) \ \leq \ V(\hat{\theta}), \ \forall \hat{\theta} \text{ with } E(\hat{\theta}) = \theta. \tag{5.3}$$

Remark 5.4. It can be shown that if an unbiased estimator exists for a parameter, then a MVUE also exists and it is unique. However, they are not easy to produce! In what follows, we present a way of obtaining MVUE's via efficient estimators.

Definition 5.5. *The **likelihood function** of a sample $X_1, \ldots, X_n$ is the joint probability function of the sample (seen as a vector), i.e. the sample function*

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta) = f(X_1; \theta) f(X_2; \theta) \ldots f(X_n; \theta), \tag{5.4}$$

with value $L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$, representing the joint probability distribution (in the discrete case) or the joint density (in the continuous case) of the random vector $(X_1, \ldots, X_n)$.

Definition 5.6. *For a sample of size n, the **Fisher (quantity of) information** relative to θ, is the quantity*

$$I_n(\theta) \ = \ E\left[\left(\frac{\partial \ln L(X_1, \ldots, X_n; \theta)}{\partial \theta}\right)^2\right], \tag{5.5}$$

if the likelihood function L is differentiable with respect to θ.

Remark 5.7. The Fisher information is a way of measuring the amount of information that a random sample $X_1, \ldots, X_n$ carries about an unknown parameter θ, upon which the likelihood function depends. Formally, it is the expected value of the *observed information* (or the variance of the *score*).

Proposition 5.8. *If the range of X does not depend on θ and the likelihood function L is twice differentiable with respect to θ, then*

$$I_n(\theta) \ = \ -E\left[\frac{\partial^2 \ln L(X_1, \ldots, X_n; \theta)}{\partial \theta^2}\right]. \tag{5.6}$$

Corollary 5.9. *If the range of X does not depend on θ, then*

$$I_n(\theta) \;=\; nI_1(\theta). \tag{5.7}$$

Proof. By (5.4), we have

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \ln f(X_i;\theta), \\ \frac{\partial^2 \ln L}{\partial \theta^2} &= \sum_{i=1}^{n} \frac{\partial^2 \ln f(X_i;\theta)}{\partial \theta^2}\,. \end{aligned}$$

By Proposition 5.8,

$$\begin{aligned} I_n(\theta) &= -\sum_{i=1}^{n} E\left[\frac{\partial^2 \ln f(X_i;\theta)}{\partial \theta^2}\right] \\ &= \sum_{i=1}^{n} I_1(\theta) \;=\; nI_1(\theta). \end{aligned}$$

□

Recall that we seek unbiased estimators with *small* variance. A MVUE has the *lowest* variance that an unbiased estimator can possibly have. The next result tells us exactly how low that can be, under certain conditions.

Theorem 5.10 (Cramér-Rao Inequality)**.** *Let X be a characteristic whose pdf $f(x;\theta)$ is differentiable with respect to θ and let $\overline{\theta} = \overline{\theta}(X_1,\ldots,X_n)$ be an absolutely correct estimator for θ. Then*

$$V(\overline{\theta}) \;\geq\; \frac{1}{I_n(\theta)}. \tag{5.8}$$

Definition 5.11. *Let $\overline{\theta} = \overline{\theta}(X_1,\ldots,X_n)$ be an absolutely correct estimator for θ. The **efficiency** of $\overline{\theta}$ is the quantity*

$$e(\overline{\theta}) \;=\; \frac{I_n^{-1}(\theta)}{V(\overline{\theta})} = \frac{1}{I_n(\theta)V(\overline{\theta})}. \tag{5.9}$$

*The estimator $\overline{\theta}$ is said to be **efficient** for θ, if $e(\overline{\theta}) = 1$.*

Remark 5.12.

1. So, by Theorem 5.10, the efficiency $e(\overline{\theta})$ is the minimum possible variance for an unbiased

estimator $\overline{\theta}$ divided by its actual variance. Its value is always $e(\overline{\theta}) \leq 1$. An efficient estimator has the maximum possible efficiency.

2. An efficient estimator may not exist, but if it does, it is also the MVUE. This is because an efficient estimator maintains equality on the Cramér-Rao inequality for all parameter values, which means it attains the minimum variance for all parameters. So, this is one way to obtain MVUE's. The MVUE, even if it exists, is not necessarily efficient.

Example 5.13. Let X be a characteristic with pdf

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

for $x > 0$ and 0, otherwise, where $\theta > 0$ is unknown. For a random sample $X_1, \ldots, X_n$, consider the estimator $\overline{\theta} = \frac{1}{2}\overline{X}$. Show that it is absolutely correct and find its efficiency.

Solution. First, let us see that $f(x;\theta)$ is indeed a density function (and recall some properties along the way).

$$\int_{\mathbb{R}} f(x)\, dx = \frac{1}{\theta^2} \int_0^\infty x e^{-\frac{x}{\theta}}\, dx,$$

which, with the change of variables $u = \dfrac{x}{\theta}$, is equal to

$$\begin{aligned} &= \frac{1}{\theta^2} \int_0^\infty (\theta u) e^{-u} (\theta du) \\ &= \int_0^\infty u e^{-u}\, du \\ &= \Gamma(2) \;=\; 1, \end{aligned}$$

where $\Gamma(a) = \displaystyle\int_0^\infty x^{a-1} e^{-x}\, dx$ is Euler's Gamma function (see Seminar 1). Recall that $\Gamma(n+1) = n!$. With the same change of variables, we compute

$$E(X) \;=\; \int_{\mathbb{R}} x f(x)\, dx \;=\; \frac{1}{\theta^2} \int_0^\infty x^2 e^{-\frac{x}{\theta}}\, dx \;=\; \theta \int_0^\infty u^2 e^{-u}\, du \;=\; \theta\, \Gamma(3) \;=\; 2\theta,$$

$$E(X^2) \;=\; \int_{\mathbb{R}} x^2 f(x)\, dx \;=\; \frac{1}{\theta^2} \int_0^\infty x^3 e^{-\frac{x}{\theta}}\, dx \;=\; \theta^2 \int_0^\infty u^3 e^{-u}\, du \;=\; \theta^2\, \Gamma(4) \;=\; 6\theta^2,$$

$$V(X) \;=\; E(X^2) - (E(X))^2 \;=\; 6\theta^2 - 4\theta^2 \;=\; 2\theta^2.$$

Then for $\overline{\theta}$ we have

$$E(\overline{\theta}) \;=\; \frac{1}{2}E(\overline{X}) \;=\; \frac{1}{2}E(X) \;=\; \theta,$$

which means $\overline{\theta}$ is unbiased and

$$V(\overline{\theta}) \;=\; \frac{1}{4}V(\overline{X}) \;=\; \frac{1}{4}\frac{V(X)}{n} \;=\; \frac{\theta^2}{2n} \to 0, \text{ as } n \to \infty,$$

so $\overline{\theta}$ is absolutely correct.

To compute the Fisher information, since the range of X does not depend on θ, we use formulas (5.6)-(5.7). We have

$$\begin{aligned} L(X_1;\theta) &= \frac{1}{\theta^2}X_1 e^{-\frac{1}{\theta}X_1}, \\ \ln L &= -2\ln\theta + \ln X_1 - \frac{1}{\theta}X_1, \\ \frac{\partial \ln L}{\partial \theta} &= -\frac{2}{\theta} + \frac{1}{\theta^2}X_1, \\ \frac{\partial^2 \ln L}{\partial \theta^2} &= \frac{2}{\theta^2} - \frac{2}{\theta^3}X_1. \end{aligned}$$

Then

$$I_1(\theta) \;=\; -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right) \;=\; -\frac{2}{\theta^2} + \frac{2}{\theta^3}E(X_1) \;=\; -\frac{2}{\theta^2} + \frac{4}{\theta^2} \;=\; \frac{2}{\theta^2}.$$

Thus

$$I_n(\theta) \;=\; \frac{2n}{\theta^2} \text{ and } e(\overline{\theta}) \;=\; \frac{1}{\frac{2n}{\theta^2}\cdot\frac{\theta^2}{2n}} \;=\; 1,$$

so $\overline{\theta} = \frac{1}{2}\overline{X}$ is an efficient estimator and, by Remark 5.12, also the MVUE for θ.

■

5.2 Methods of Point Estimation

So far, we have discussed desirable properties of point estimators, how to distinguish "good" from "bad" or "better" estimators, based on how reliable they are in approximating the value of a population parameter. But *how* to *actually* find an estimator, an approximating value $\overline{\theta} = \overline{\theta}(X_1, X_2, \ldots, X_n)$ for a target parameter θ, based on sample variables $X_1, X_2, \ldots, X_n$? Sometimes, such a value may be "guessed" from past experience or from observing many samples over

time. But statisticians wanted more rigorous, more mathematical ways of producing a point estimator, which can then be analyzed from the various points of view discussed in the previous section. This question will be addressed in this section.

We present two of the most popular methods of finding point estimators: the *method of moments* and the *method of maximum likelihood*. We will also discuss advantages and disadvantages of each method.

Method of Moments

This is one of the oldest and easiest methods for obtaining point estimators, first formalized by K. Pearson in the late 1800's.

Let us recall, for a population characteristic X, we define the *moments of order* k as

$$\nu_k \;=\; E\left(X^k\right) \;=\; \begin{cases} \displaystyle\sum_{i\in I} x_i^k p_i, & \text{if } X \text{ is discrete with pdf } X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I} \\ \\ \displaystyle\int_{\mathbb{R}} x^k f(x)\,dx, & \text{if } X \text{ is continuous with pdf } f:\mathbb{R}\to\mathbb{R}. \end{cases} \tag{5.10}$$

For a sample drawn from the distribution of X, i.e. sample variables $X_1, \ldots, X_n$ (iid), the *sample moments of order* k are defined by

$$\overline{\nu}_k \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^k. \tag{5.11}$$

Also, let us recall (from Proposition 1.8 in Lecture 9) that

$$E(\overline{\nu}_k) \;=\; \nu_k, \tag{5.12}$$

so $\overline{\nu}_k$ is an *unbiased* estimator for ν_k and

$$V(\overline{\nu}_k) \;=\; \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right) \;\to\; 0, \text{ as } n\to\infty. \tag{5.13}$$

By (5.12)-(5.13), the sample moment of order k is an *absolutely correct* estimator for the population moment of the same order.

That is precisely the idea of this method. Since the theoretical (population) moments in (5.10) contain the target parameters that are to be estimated, while the sample moments in (5.11) are all

known, computable from the sample data, simply set the two to be equal and solve the resulting system. To estimate k parameters, equate the first k population and sample moments:

$$\begin{cases} \nu_1 &= \overline{\nu}_1 \\ \dots &\dots &\dots \\ \nu_k &= \overline{\nu}_k \end{cases} \tag{5.14}$$

The left-hand sides of these equations depend on the distribution parameters. The right-hand sides can be computed from data. The **method of moments estimator** is the solution of this $k \times k$ system of equations.

Example 5.14. Let us consider the characteristic X from Example 5.13, with pdf

$$f(x;\theta) = \frac{1}{\theta^2} x e^{-\frac{x}{\theta}},\ x > 0,$$

where $\theta > 0$ is unknown. Based on a random sample $X_1, \dots, X_n$, find the method of moments estimator $\overline{\theta}$ for θ. For the sample data $\{2.3, 3.7, 1.44, 2.16\}$, find the numerical estimation of θ.

Solution. There is only one unknown parameter, θ, so we will have only one equation in system (5.14),

$$\begin{aligned} \nu_1 &= \overline{\nu}_1, \text{ i.e.} \\ E(X) &= \overline{X}. \end{aligned}$$

In our work in Example 5.13, we computed

$$E(X) = \int_{\mathbb{R}} x f(x)\, dx = 2\theta.$$

So, we solve the equation

$$2\theta = \overline{X},$$

to find the method of moments estimator

$$\overline{\theta} = \frac{1}{2}\overline{X}.$$

Notice that it is an unbiased estimator, since

$$E\left(\overline{\theta}\right) \;=\; E\left(\frac{1}{2}\overline{X}\right) \;=\; \frac{1}{2}E\left(\overline{X}\right) \;=\; \frac{1}{2}E(X) \;=\; \frac{1}{2}\cdot 2\theta \;=\; \theta.$$

For the sample data $x_1 = 2.3, x_2 = 3.7, x_3 = 1.44$ and $x_4 = 2.16$, we have

$$\overline{x} \;=\; \frac{x_1 + x_2 + x_3 + x_4}{4} \;=\; \frac{9.6}{4} \;=\; 2.4,$$

so the numerical value of our estimator is

$$\overline{\theta} \;=\; 1.2.$$

■

Example 5.15. Let us use the method of moments to estimate *both* parameters of the Normal $N(\mu, \sigma)$ distribution.

Solution. Now we have a characteristic X with pdf

$$f(x;\mu,\sigma) \;\;=\;\; \frac{1}{\sigma\sqrt{2\pi}}e^{-\dfrac{(x-\mu)^2}{2\sigma^2}},\; x \in \mathbb{R},$$

with $\mu \in \mathbb{R}$ and $\sigma > 0$, both unknown.

To estimate two parameters, we need two equations in system (5.14),

$$\begin{cases} \nu_1 \;=\; \overline{\nu}_1 \\ \nu_2 \;=\; \overline{\nu}_2 \end{cases}$$

In the first equation, we have

$$\begin{aligned} \nu_1 \;&=\; E(X) \;=\; \mu \text{ and} \\ \overline{\nu}_1 \;&=\; \overline{X}, \end{aligned}$$

since for a Normal $N(\mu, \sigma)$ variable the first parameter is its expectation. We also know that the variance of a $N(\mu, \sigma)$ variable is equal to σ^2. But recall the computational formula for the variance (in general)

$$V(X) \;=\; E\left(X^2\right) - \left(E(X)\right)^2 \;=\; \nu_2 - \nu_1^2.$$

From here, we get

$$\nu_2 \;=\; V(X) + \nu_1^2 \;=\; \sigma^2 + \mu^2,$$

in this case.

So system (5.14) becomes

$$\begin{cases} \mu & = & \overline{X} \\ \mu^2 + \sigma^2 & = & \overline{\nu}_2 \end{cases},$$

a system of two equations in two unknowns, with solution

$$\begin{cases} \overline{\mu} & = & \overline{X} \\ \overline{\sigma} & = & \sqrt{\overline{\nu}_2 - \overline{X}^2} \;=\; \sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 - \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i\right)^2} \end{cases}.$$

■

Remark 5.16. Method of moments estimates are typically easy to compute. However, on rare occasions, when k equations are not enough to estimate k parameters, higher moments (i.e. more equations) can be considered.

Method of Maximum Likelihood

Maximum-likelihood estimation was first recommended, analyzed and then vastly popularized by R. A. Fisher in the 1920's, although it had been used earlier by Gauss and Laplace. For a fixed random sample from an underlying probability distribution, the maximum likelihood method picks the values of the population parameters that make the data "more likely" than any other values of the parameters would make them.

Let us illustrate it, first, with a simple example, to understand the underlying ideas.

Example 5.17. Suppose there are 5 balls in a box, black or white, the number of each being unknown. Suppose further, that we randomly select 3 of them, without replacement, and we get all three white. What would be a good estimate, $\overline{w}$, for the number of white balls in the box, w?

Solution.
Obviously, $w \in \{3, 4, 5\}$.

If the true value was $w = 3$, then the probability of randomly selecting 3 white balls without replacement, would be (by the Hypergeometric model)

$$p_1 = \frac{C_3^3 C_2^0}{C_5^3} = \frac{1}{10}.$$

If the true value was $w = 4$, then the probability of us randomly selecting 3 white balls without replacement, would be

$$p_2 = \frac{C_4^3 C_1^0}{C_5^3} = \frac{4}{10}.$$

And, finally, if the true value was $w = 5$, then the probability of randomly selecting 3 white balls without replacement, would be

$$p_3 = \frac{C_5^3 C_0^0}{C_5^3} = 1.$$

So, it would seem reasonable to choose $\overline{w} = 5$ as our estimate for w, since this would *maximize* the probability of obtaining our observed sample.

■

This, in essence, describes the *method of maximum likelihood* estimation. Now let us write it formally.

Recall that the probability of obtaining an observed sample is measured by the *likelihood function* of a sample:

$$L(X_1, \ldots, X_n; \Theta) = \prod_{i=1}^{n} f(X_i; \Theta),$$

where now *all* unknown target parameters are contained in a vector $\Theta = (\theta_1, \ldots, \theta_l)$.

This method chooses the values of an estimator $\overline{\Theta} = (\overline{\theta}_1, \ldots, \overline{\theta}_l) = \overline{\Theta}(X_1, \ldots, X_n)$ that maximize the function $L(X_1, \ldots, X_n; \Theta)$. So, if L is twice differentiable with respect to each $\theta_1, \ldots, \theta_l$, we find the solutions of the maximum-likelihood system

$$\frac{\partial L(X_1, \ldots, X_n; \theta_1, \ldots, \theta_l)}{\partial \theta_j} = 0, \quad j = \overline{1, l}, \tag{5.15}$$

or, equivalently, but easier to compute, the maximum-likelihood equations

$$\frac{\partial \ln L(X_1, \ldots, X_n; \theta_1, \ldots, \theta_l)}{\partial \theta_j} = 0, \quad j = \overline{1, l}. \tag{5.16}$$

If the system (5.16) has a solution, then it is unique and it is called the **maximum likelihood (MLE)** estimator.

Example 5.18. Consider again the situation in Example 5.14, so a characteristic with pdf

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

for $x > 0$, with $\theta > 0$ is unknown. Based on a random sample $X_1, \ldots, X_n$, let us now find the MLE $\hat{\theta}$ for θ.

Solution. The likelihood function is given by

$$\begin{aligned} L(X_1, \ldots, X_n; \theta) &= \prod_{i=1}^{n} \left(\frac{1}{\theta^2} X_i e^{-\frac{X_i}{\theta}} \right) \\ &= \left(\prod_{i=1}^{n} X_i \right) \frac{1}{\theta^{2n}} e^{-\frac{1}{\theta} \sum_{i=1}^{n} X_i} \\ &= K \frac{1}{\theta^{2n}} e^{-\frac{n\overline{X}}{\theta}}, \end{aligned}$$

where $K = \prod_{i=1}^{n} X_i$ is a constant with respect to θ.
Take the logarithm, to make computations easier and differentiate it with respect to θ (the only unknown).

$$\begin{aligned} \ln L &= \ln K - 2n \ln \theta - \frac{n\overline{X}}{\theta} \\ \frac{\partial \ln L}{\partial \theta} &= -\frac{2n}{\theta} + \frac{n\overline{X}}{\theta^2}. \end{aligned}$$

Then system (5.16) becomes

$$-\frac{2n}{\theta} + \frac{n\overline{X}}{\theta^2} = 0,$$

whose solution is the MLE

$$\hat{\theta} = \frac{1}{2}\overline{X},$$

the same as the method of moments estimator $\hat{\theta}$.

■

Example 5.19. Let us also find the MLE's for the parameters of the Normal $N(\mu, \sigma)$ distribution

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \; x \in \mathbb{R},$$

with $\mu \in \mathbb{R}$ and $\sigma > 0$, both unknown.

Solution. We find the likelihood function and its logarithm:

$$\begin{aligned} L(X_1,\ldots,X_n;\mu,\sigma) &= \prod_{i=1}^{n}\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}, \\ \ln L(\mu,\sigma) &= -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}X_i^2 - 2\mu\sum_{i=1}^{n}X_i + n\mu^2\right) \\ &= -n\ln\sigma - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}X_i^2 - 2n\mu\overline{X} + n\mu^2\right). \end{aligned}$$

The maximum likelihood system will consist of two equations

$$\begin{cases} \dfrac{\partial \ln L(\mu,\sigma)}{\partial\mu} &= 0 \\ \dfrac{\partial \ln L(\mu,\sigma)}{\partial\sigma} &= 0, \end{cases}$$

i.e.

$$\begin{cases} -\dfrac{1}{2\sigma^2}\left(-2n\overline{X}+2n\mu\right) &= 0 \\ -n\dfrac{1}{\sigma} + \dfrac{1}{\sigma^3}\left(\displaystyle\sum_{i=1}^{n}X_i^2 - 2n\mu\overline{X} + n\mu^2\right) &= 0, \end{cases}$$

From the first equation, we get

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

Substituting that into the second equation, we find

$$\begin{aligned} n &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}X_i^2 - 2n\hat{\mu}\overline{X} + n\hat{\mu}^2\right), \\ \sigma^2 &= \frac{1}{n}\left(\sum_{i=1}^{n}X_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) \\ &= \overline{\nu}_2 - \overline{X}^2 \\ \hat{\sigma} &= \sqrt{\overline{\nu}_2 - \overline{X}^2} \end{aligned}$$

So, again, the MLE's coincide with the method of moments estimators.

■

Remark 5.20. In both our examples, the two methods yielded the same point estimator. That is not always the case. If they differ, the natural question is: which one is better? In some respects, when estimating parameters of a known family of probability distributions, the method of moments is superseded by Fisher's method of maximum likelihood, because maximum likelihood estimators have higher probability of being close to the quantities to be estimated. However, in some cases, the likelihood equations (5.16) may be intractable without computers, whereas the method of moments estimators can be quickly and easily calculated by hand as seen above. Estimates by the method of moments may be used as the first approximation to the solutions of the likelihood equations (5.16), and successive improved approximations may then be found by some iterative approximation methods (like the Newton method). In this way, the method of moments and the method of maximum likelihood are symbiotic. In some cases, infrequent with large samples but not so infrequent with small samples, the estimates given by the method of moments are outside of the parameter space and it does not make sense to rely on them then. That problem never arises in the method of maximum likelihood.