CHAPTER 1

# Introduction

**1.1. Overview.** Higher order asymptotics deals with two sorts of closely related things. First, there are questions of approximation. One is concerned with expansions or inequalities for a distribution function, an asymptotic variance, the posterior density and integrated Bayes risk. Second, there are inferential issues. These involve, among other things, the application of the first set of ideas of the study of higher order efficiency, admissibility and minimaxity, conditional and adjusted likelihood and construction of noninformative priors by approximately matching posterior and frequentist probability. In the matter of expansions, it is as important to have usable, explicit formulas as a rigorous proof that the expansions are valid in the sense of truly approximating a target quantity up to the claimed degree of accuracy.

Classical asymptotics is based on the notion of asymptotic distribution, often derived from the central limit theorem, and usually the approximations are correct up to $O(n^{-1/2})$, where $n$ is the sample size. Higher order asymptotics provides refinements based on asymptotic expansions of the distribution or density function of an estimate and the posterior density function of the parameter. Posterior expansions are refinements of the Bernstein–von Mises theorem on asymptotic normality of the posterior, whereas the other expansions are rooted in the Edgeworth theory, which is itself a refinement of the central limit theorem.

When higher order asymptotics is correct up to $o(n^{-1/2})$, it is second order asymptotics. When further terms are picked up, so that the asymptotics is correct up to $o(n^{-1})$, it is third order asymptotics. In his pioneering papers, C. R. Rao coined the term second order efficiency for a concept that would now be called third order efficiency. The new terminology is essentially owing to Pfanzagl and Takeuchi.

The stress in the subsequent chapters is on basic concepts and main results, with enough technical details to make applications to specific examples fairly easy. For the main results, we provide a proof, or, where a proof is too long or technical, a sketch of the argument and a reference to where details are available. We do not strive for maximum generality.

As background, we assume little more than basic results on various modes of convergence, for which Chapter 1 of Serfling (1980) will suffice, and the basic theory of estimation in large samples under standard regularity conditions, as contained in, say, Chapter 5 of Rao (1973). Occasionally we need also some knowledge of basic facts about exponential families; see Lehmann [(1986), pages 56–66, 142–143].

Welcome aboard.

**1.2. First order efficiency.** To motivate the definition of higher order efficiency, we need to discuss briefly asymptotic efficiency or efficiency or first order efficiency. They all mean the same.

We consider $n$ r.v.'s $X_1, X_2, \ldots, X_n$. Although neither independence nor identical distribution is essential, we will assume, for simplicity, the $X$'s are i.i.d. The common p.d.f. or p.f. is $p(x|\theta)$, $\theta \in \Theta \subset R^p$, where $R^p$ is the $p$-dimensional Euclidean space and $\Theta$ is an open $p$-dimensional subset. Most of the time we will work with $p = 1$ and write $R$ for $R^1$. So, unless otherwise stated, $p = 1$.

The theory of asymptotic efficiency has been developed in a very general setting by Le Cam, the essential ingredient being the locally asymptotically normal (LAN) condition

$$(1) \qquad \log\left\{\frac{p(x_1, x_2, \ldots, x_n(\theta + \delta/\sqrt{n}))}{p(x_1, x_2, \ldots, x_n|\theta)}\right\} = \delta W_n - \frac{1}{2}\delta^2 I(\theta) + o_p,$$

where $\delta$ is real and $W_n$ is asymptotically normal, with mean zero and variance $I(\theta)$. Our development of the subject will be under the much stronger classical regularity conditions [see Rao (1973), page 364] and is in the spirit of Rao (1963). This leads to simplicity, and is convenient as an introduction to higher order efficiency.

Specifically, we make the following assumptions:

ASSUMPTION $A_1$. $p(x|\theta)$ is thrice continuously differentiable as a function of $\theta$, interchanges of differentiation with respect to $\theta$ and integration with respect to $x$ are valid, and

$$\left|\frac{d^3 \log p}{d\theta^3}\right|_{\theta'} < M(x), \quad E(M(X)|\theta) < K$$

for all $\theta'$ in a neighborhood of $\theta$. Moreover, the Fisher information (per observation)

$$I(\theta) \underset{\text{def}}{=} E\left\{\left(\frac{d \log p}{d\theta}\right)^2 \middle| \theta\right\} = -E\left\{\frac{d^2 \log p}{d\theta^2} \middle| \theta\right\}$$

is finite and positive for all $\theta$. In addition, as in Rao (1963), we require the further assumption:

ASSUMPTION $A_2$. Each estimate $T_n$ that we consider is asymptotically normal (A.N.) with mean $\theta$ and variance $\nu(\theta)/n$ [A.N. $(\theta, \nu(\theta)/n)$], uniformly on compact $\theta$-sets.

The function $\nu(\theta)$ in $A_2$ depends on the estimate $T_n$. Uniformity means the following: Given any compact set $C$ and $\varepsilon > 0$,

$$\lim_{n \to \infty} \sup_{\theta \in C} \sup_{y} \left| P\left(\sqrt{n}\,(T_n - \theta) < y\right) - \Phi(y|0, \nu(\theta)) \right| < \varepsilon,$$

where $\Phi(\cdot|\mu, \sigma^2)$ is the normal distribution function with mean $\mu$ and variance $\sigma^2$.

Sometimes, again following Rao (1963), instead of Assumption $A_2$, we will need Assumption $A_3$:

ASSUMPTION $A_3$. $T_n$ is Fisher consistent.

If $X_i$'s are multinomial, assuming $(k+1)$ distinct values $a_1, a_2, \ldots, a_{k+1}$ with probability $\pi_1(\theta), \pi_2(\theta), \ldots, \pi_{k+1}(\theta)$, then $T_n$ is said to be Fisher consistent if

$$T_n = H(p_1, \ldots, p_k),$$

where $p_i$ is the proportion of $X_1, \ldots, X_n$ that equals $a_i$ and $H$ is a real valued function such that

$$H(\pi_1(\theta), \ldots, \pi_k(\theta)) = \theta \qquad \forall\, \theta,$$

that is, $T_n$ depends on the observations only through the sufficient statistic $(p_1, \ldots, p_k)$, this dependence is free of $n$ and $T_n$ evaluated at the true proportions gives the true value of $\theta$. The concept is owing to Fisher and the explicit definition is owing to Rao. We shall define later a curved exponential family and Fisher consistent estimates in that setting, following Ghosh and Subramanyam (1974) and Efron (1975). Assumption $A_3$ will be taken to imply we are in the multinomial or, more generally, the curved exponential setting.

For a Fisher consistent $T_n$ to be consistent in the usual sense, that is, for $T_n \to_p \theta$, one needs $H$ to be continuous. If one also has a continuously differentiable $H$, then, by the so-called delta method, that is, Taylor expansion, $T_n$ is asymptotically normal.

The following facts are owing to Rao; see Ghosh (1985) for detailed review and references.

THEOREM 1.1. *Assume Assumptions* $A_1$, *and* $A_2$ *or* $A_3$, *with* $H$ *continuously differentiable. Then*

(i) $\nu(\theta) \geq 1/I(\theta)$.

(ii) (a) $\lim_{n\to\infty} P\{-y < \sqrt{n}(T_n - \theta) < y|\theta\} \leq \int_{-y}^{y} \Phi(dy|0, 1/I(\theta))$,

(b) $\liminf_{n\to\infty} E\{(l(\sqrt{n}(T_n - \theta)))\} \geq \int_{-\infty}^{\infty} l(y)\Phi(dy|0, 1/I(\theta))$, *where* $l$ *is a nonnegative loss function and* $l(0) = 0$, $l(y) = l(-y)$ *and* $l(y) \leq l(z)$ *if* $0 \leq y < z$.

Part (ii)(a) is an immediate consequence of the first inequality, often called Fisher's inequality, and then (ii)(b) follows from (ii)(a) by an easy argument involving integration by parts.

An alternative route, due to Le Cam and Hájek, is to consider all estimates, not necessarily satisfying Assumption $A_2$ or $A_3$, but using the local minimax criterion

$$\lim_{\delta \downarrow 0} \liminf_{n\to\infty} \sup_{|\theta' - \theta| < \delta} E\{l(\sqrt{n}(T_n - \theta'))|\theta'\},$$

which is always greater than or equal to the right-hand side of (ii)(b), under the same assumptions on $l$. For the proof of this, one needs only the LAN condition; see Le Cam and Yang (1990) or Ibragimov and Has'minskii [(1981), Chapter 2].

DEFINITION 1.1. $T_n$ is efficient or first order efficient (FOE) if $T_n$ is A.N. $(\theta, 1/nI(\theta))$ and satisfies $A_2$ or $A_3$.

Note that without Assumption $A_2$ or $A_3$, Fisher's inequality $\nu(\theta) \geq 1/I(\theta)$ need not hold. In fact, as first pointed out by Hodges in the early 1950s, one takes an estimate $T_n$ that is A.N. $(\theta, 1/nI(\theta))$, for example, the maximum likelihood estimate (mle), and shrinks it suitably toward a fixed point $\theta_0$ so that for the new "superefficient" estimate, $\nu(\theta_0) < (I(\theta_0))^{-1}$ and $\nu(\theta) = (I(\theta))^{-1}$ for $\theta \neq \theta_0$. Such shrinkage estimates, including the famous James–Stein estimates, which improve on the sample mean for a $p$-variate normal with $p > 2$, are excluded by Assumption $A_2$ or $A_3$.

**1.3. Third order efficiency.** Many classical estimates, including the mle, are FOE. To distinguish between them, we introduce the notion of estimates that are third order efficient (TOE), which is the modern name for the concept for which Rao had coined the term second order efficient. We will clarify later why second order efficiency fails to distinguish between FOE estimates.

Consider two FOE estimates $T_{in}$, $i = 0, 1$. Typically the asymptotic mean and variance of $T_{in}$ will have expansions of the form

$$\theta + \frac{b_i(\theta)}{n} + o(n^{-1})$$

and

$$\frac{1}{nI(\theta)} + \frac{A_i(\theta)}{n^2} + o(n^{-2}),$$

respectively. A precise interpretation of these expansions, as well as a method for calculating $b_i$ and $A_i$, will be given in the next chapter.

It makes sense to restrict higher order comparison of variance to FOE estimates which have the same bias up to $o(n^{-1})$, that is, assume $b_0(\theta) = b_1(\theta) = b(\theta)$. Informally, $T_{on}$ is TOE among FOE estimates with same asymptotic bias up to $o(n^{-1})$, if

$$A_0(\theta) \leq A_1(\theta) \qquad \forall\, \theta$$

for all competing FOE estimates $T_{1n}$.

One natural choice is $b(\theta) \equiv 0$. In this case, one is confining attention to FOE estimates which are asymptotically unbiased up to $o(n^{-1})$. There are other natural choices. For example, one starts with a particular FOE $T_n$, say, and takes the bias term for $T_n$ as $b(\theta)$. The TOE estimate in this class will be an improvement on $T_n$. Finding a TOE estimate is rather like using a Rao–Blackwell theorem when a complete sufficient statistic exists—one gets either a minimum variance unbiased estimate or improves upon a given estimate.

No single estimate is TOE among all FOE estimates. Rather, there is a sort of complete class of TOE's from which you can always get one to beat a given FOE.

The above informal definition of TOE and the subsequent remarks need the following qualification. Usually, in theorems on TOE, there will be additional regularity conditions on the FOE's being considered, which involve strengthening of Assumption $A_2$ or $A_3$.