CHAPTER 3

# Third Order Efficiency for Curved Exponentials

**3.1. The main result.** Consider a curved exponential as introduced in the previous chapter, that is, except for a factor involving $x$ only,

$$p(x|\theta) = c(\theta)\exp\left\{\sum_{i=1}^{k} \beta_i(\theta) f_i(x)\right\}.$$

For the time being, $\Theta \subset R$. Extensions to higher dimensions will be considered briefly later in the chapter. Recall also that $Z_{ij} = f_i(x_j)$, $Z_j = (Z_{1j}, \dots, Z_{kj})$, $\bar{Z} = n\ ^{1}\sum_1^n Z_j$, $\mu_i(\theta) = E(Z_{ij}|\theta)$, $\mu = (\mu_1, \dots, \mu_k)_\theta$ and $[\sigma_{ii'}(\theta)]$ is the dispension matrix of $Z_j$. Moreover, $\beta(\theta)$ lies in the interior of the natural parameter space of the multiparameter exponential, which is, apart from a factor of $x$,

$$p(x|\beta) = d(\beta)\exp\left\{\sum \beta_i f_i(x)\right\}.$$

So $p(x|\beta)$ is (real) analytic in $\beta$, and, hence, by our assumption of thrice continuous differentiability of $\beta(\theta)$, it follows that $p(x|\theta)$ is thrice continuously differentiable in $\theta$. Also by assumption, $1, f_1, f_2, \dots, f_k$ are linearly independent, so that, among other things, $[\sigma_{ii'}]$ is positive definite.

Fisher consistent estimates $T_n$ are of the form

$$T_n = H(\bar{Z}),$$

where $H(\mu(\theta)) = \theta$ and $H$ is thrice continuously differentiable in a neighborhood of $\mu(\theta)$ for all $\theta$. In Section 2.5 the mle was exhibited as a Fisher consistent estimate.

We will use frequently the calculations of Ghosh and Subramanyam (1974). So it is convenient to occasionally use the notations there, namely, $p$ or $p^n$ for $\bar{Z}$ for $\pi$ for $\mu$. Of the three interpretations for expansions considered in Section 2.7, the third is the most convenient, and so we also introduce a notation for expectation with truncation, namely,

$$E^U(Y|\theta_0) = E(I_U Y|\theta_0),$$

where $Y$ is a random variable depending on $p$ only, $U$ is a neighborhood of $\pi(\theta_0)$ which also contains neighborhoods of $\pi(\theta)$ for all $|\theta - \theta_0| < \delta$ and $I_U$ is the indicator of $U$. By Chernoff's inequality (1952),

$$E^U(Y|\theta) = E(Y^t|\theta) + O(\rho^n),$$

where $|\theta - \theta_0| < \delta$, $Y^t$ is $Y$ truncated in any arbitrary way so that $|Y|$ is bounded on the complement of $U$ and $0 < \rho < 1$. If $Y_n$ is a sequence of random variables such that $E^U(Y_n^2|\theta_0)$ is $o(a_n^2)$ or $O(a_n^2)$, we will write $Y_n$ is $o(a_n)$ or $O(a_n)$ accordingly, If $Y_n$ is $o(a_n)$ or $O(a_n)$, $Z_n$ is $O(b_n)$, then by Cauchy–Schwarz, $Y_n Z_n$ is $o(a_n b_n)$ or $O(a_n b_n)$. With some abuse of notation, we will write $E$ for $E^U$, since in this chapter [except in (3.1) to (3.4)] $E$ will stand for $E^U$. The set $U$ depends on the function $H$. However, given a finite collection $H_1, H_2, \dots, H_m$, we can find a single $U$ to accommodate all the $T$'s. Since we will only be comparing two estimates at a time, some $T$ and $\hat{\theta}$, this is all that we need.

Recall the formulation of third order efficiency in Chapter 1. Among all Fisher consistent estimates $T_n = H(p)$ with a fixed asymptotic bias $b(\theta)/n$,

$$E\{(T_n - \theta)|\theta\} = b(\theta)/n + o(n^{-1}), \tag{3.1a}$$

$$E\{(T_n - \theta - b(\theta)/n)^2|\theta\} = 1/nI(\theta) + \psi(H,\theta)/n^2 + o(n^{-2}) \tag{3.1b}$$

for $|\theta - \theta_0| < \delta$, minimize $\psi(H,\theta)$. A FC estimate $T_n$ for which minimization holds for all $\theta_0 \in \Theta$ is TOE.

Alternatively, as in Rao (1963), given an efficient FC estimate $T_n$, make it unbiased up to $o(n^{-1})$,

$$T_n^* = T_n - b(T_n)/n, \tag{3.1c}$$

and among such estimates minimize the coefficient of $n^{-2}$ in the expansion of variance

$$E\{(T_n^* - \theta)^2|\theta\} = \frac{1}{nI(\theta)} + \frac{\psi^*(H,\theta)}{n^2} + o(n^{-2}). \tag{3.1d}$$

In Ghosh and Subramanyam (1974), $\psi^*(H,\theta)$ is denoted as $\psi(\{T_n^*\},\theta)$.

Note that $b(\cdot)$ is a real analytic function of $\beta(\theta)$ and hence thrice differentiable in $\theta$. Hence the remarks on expansions associated with $T_n'$ in Section 2.7 apply to $T_n^*$ and other such perturbations of $T_n$. We note for later use

$$E(T_n^*|\theta) = \theta + o(n^{-1}), \tag{3.1e}$$

$$\begin{aligned} &E\{(T_n^* - \theta)^2|\theta\} \\ &\quad = E\left[\left\{(T_n - \theta - b(\theta)/n) - \frac{b'(\theta)}{n}(T_n - \theta)\right\}^2 \middle| \theta\right] + o(n^{-2}) \\ &\quad = \frac{1}{nI(\theta)} + \frac{\psi(H,\theta)}{n^2} - \frac{2b'(\theta)}{n^2 I(\theta)} + o(n^{-2}), \end{aligned} \tag{3.1f}$$

so that

$$\psi^*(H,\theta) = \psi(H,\theta) - \frac{2b'(\theta)}{I^2(\theta)}. \tag{3.1g}$$

In addition, there is Fisher's original formulation via loss of information in $T_n$, which is defined as the limit of

$$\begin{aligned}&\big(nI(\theta) - I_{T_n}(\theta)\big)\\ &\quad= E\left\{\left(\frac{d\log p(X_1,\dots,X_n|\theta)}{d\theta} - \frac{d\log p(T_n|\theta)}{d\theta}\right)^2 \middle|\, \theta\right\},\end{aligned} \tag{3.2}$$

where $p(T_n|\theta)$ is the marginal p.f. or p.d.f. of $T_n$ and

$$I_{T_n}(\theta) = E_\theta\left(\frac{d\log p(T_n|\theta)}{d\theta}\right)^2$$

is the Fisher information carried by $T_n$. The limiting difference is the limiting difference in the total information in the sample and the information in $T_n$. The relation (3.2) follows from

$$E\left(\frac{d\log p(X_1,\dots,X_n|\theta)}{d\theta}\middle|\, T_n = t,\theta\right) = p_{T_n}(t|\theta). \tag{3.3}$$

To prove (3.3), note that for all bounded (measurable) $U(T_n)$,

$$\begin{aligned}E\left(\frac{p(X_1,X_2,\dots,X_n|\theta')}{p(X_1,X_2,\dots,X_n|\theta)}U(T_n)\middle|\,\theta\right) &= E(U(T_n)|\theta')\\ &= \int\frac{p_{T_n}(t|\theta')}{p_{T_n}(t|\theta)}U(t)P(dt|\theta)\end{aligned} \tag{3.3a}$$

so that

$$E\left(\frac{p(X_1,X_2,\dots,X_n|\theta')}{p(X_1,X_2,\dots,X_n|\theta)}\middle|\, T_n = t,\theta\right) = \frac{p_{T_n}(t|\theta')}{p_{T_n}(t|\theta)}. \tag{3.3b}$$

Hence the left-hand side of (3.2) is, assuming all interchanges of operations below are justified,

$$\begin{aligned}&\lim_{h\to 0} E\left\{\left(\frac{p(X_1,X_2,\dots,X_n|\theta+h)}{p(X_1,X_2,\dots,X_n|\theta)} - 1\right)\frac{1}{h}\middle|\, T_n = t,\theta\right\}\\ &\quad= \lim_{n\to 0}\left(\frac{p_{T_n}(t|\theta+h)}{p_{T_n}(t|\theta)} - 1\right)\frac{1}{h}\\ &\quad= \frac{d\log p_{T_n}(t|\theta)}{d\theta}.\end{aligned} \tag{3.4}$$

Note that (3.3) shows that, with respect to squared error loss, $(d \log p_{T_n}(t|\theta))/d\theta$ is the best predictor of $(d\log p(X_1, X_2, \dots, X_n|\theta))/d\theta$ based on $T_n$. If $T_n$ is sufficient, the loss of information is zero, so loss of information is a measure of the sufficiency of $T_n$. A FC estimate which minimizes this limiting loss may be called TOE in Fisher's sense. It appears Fisher was sure that $\hat{\theta}$ is TOE in this sense. It seemed so obvious to him—not to us—that he did not even sketch a proof.

Since densities or p.f.'s are involved in this measure, one needs expansions for these, rather than expansions for probabilities or distribution functions. Even in the continuous case, expansions for densities of $T_n$ would be difficult to justify, and we would need much stronger assumptions than those made of Theorem 2.1. To avoid this, especially to have a theory for the (discrete) multinomial of Chapter 1, in which Rao's main interest lay, Rao (1961) proposed a modified formulation, which leads to the Fisher–Rao criterion

$$E_2 = \inf_{\lambda} \lim_{n\to\infty} \operatorname{Var}_\theta \left\{ \frac{d\log p(X_1, X_2, \dots, X_n|\theta)}{d\theta} - nI(\theta)(T_n - \theta) - \lambda n (T_n - \theta)^2 \right\} \tag{3.5}$$

in which the best predictor of Fisher is being replaced by a quadratic in $T_n$. If $T_n$ were exactly normal with mean $\theta$ and variance $(nI)^{-1}$, $I$ being free of $\theta$, then the linear term in $(T_n - \theta)$ would be exactly equal to the best predictor. So the quadratic in $(T_n - \theta)$ is a natural refinement compensating for approximate normality and, presumably, for the dependence of $I$ on $\theta$. The significance of the linear term can also be seen from the following fact. A FC (continuously differentiable) $T_n$ is efficient if and only if

$$\frac{1}{\sqrt{n}} \frac{d \log p(X_1, X_2, \dots, X_n|\theta)}{d\theta} - \sqrt{n}\, I(\theta)(T_n - \theta) \to_p 0. \tag{3.6}$$

Rao (1961, 1973) often takes this as the definition of efficiency, that is, of FOE. A proof of this appears in the course of proving Theorem 3.1; see Step 3 in the proof.

A FC $T_n$ is TOE in the Fisher–Rao sense if it minimizes $E_2$. The first statement below merely repeats (3.1b). We shall prove only parts (iii)(a) and (iv).

THEOREM 3.1.

(i) $$E\{(T_n^* - \theta_0)^2|\theta_0\} = \frac{1}{nI(\theta_0)} + \frac{\psi^*(H, \theta_0)}{n^2} + o(n^{-2}),$$

*where $T_n^*$ is defined in* (3.1c) *and is unbiased up to* $o(n^{-1})$.

(ii) $E_2$ *for* $T_n$ *equals*

$$\psi^*(H, \theta_0)I^2(\theta_0) - 2(J(\theta_0)/2 + \mu_{11}(\theta_0))^2/I^2(\theta_0),$$

*where*

$$J(\theta) = E\left(\frac{d^3 \log p(X_1|\theta)}{d\theta^3}\middle|\theta\right),$$

$$\mu_{ij}(\theta) = E\left\{\left(\frac{d \log p(X_1|\theta)}{d\theta}\right)^i \cdot \left(\frac{d^2 \log p(X_1|\theta)}{d\theta^2}\right)^j \middle|\theta\right\}.$$

(iii)(a) *$\psi^*$ is minimized for all $\theta_0$ when the FC, FOE $T_n$ is the mle $\hat{\theta}$.*

(b) *The minimum value of $\psi^*$ is*

$$\psi^*(\theta_0) = \{I(\theta_0)\mu_{02}(\theta_0) - \mu_{11}^2(\theta_0)\}/I^4(\theta_0)$$
$$+ 2\{J(\theta_0)/2 + \mu_{11}(\theta_0)\}^2/I^4(\theta_0)$$

*and the minimum value of $E_2$ is $E_2(\theta_0) = \{I(\theta_0)\mu_{02}(\theta_0) - \mu_{11}(\theta_0)^2\}/I^2(\theta_0)$.*

(iv) *Given any FC, FOE $T_n$, one can find a function $c(\theta)$, such that*

$$E\left\{\left(\hat{\theta} + c(\hat{\theta})/n - \theta_0\right)^2|\theta_0\right\} \le E\{(T_n - \theta)^2|\theta_0\} + o(n^{-2})$$

*for all $\theta_0$.*

The function $c(\cdot)$ is found by matching the bias of $T_n$, that is, if

(3.7) $$E\left(\hat{\theta}_n|\theta\right) = \theta + b_0(\theta)/n + o(n^{-1}),$$

then

(3.7a) $$c(\cdot) = b(\cdot) - b_0(\cdot).$$

Note that $c(\cdot)$ is thrice continuously differentiable since so are $b(\cdot)$ and $b_0(\cdot)$. From (3.9) of Step 1 of the proof of Theorem 3.1, it will be seen that

(3.7b) $$b_0(\theta_0) = \{J(\theta_0)/2 + \mu_{11}(\theta_0)\}/I^2(\theta_0).$$

Note its appearance in (iii)(b) of Theorem 3.1. In the following we will write

(3.7c) $$\hat{\theta}_n^* = \hat{\theta}_n - b_0\left(\hat{\theta}_n\right)/n.$$

In the proof of Theorem 3.1 we will need the following simple facts. Let $Y_{1j}, Y_{2j}, Y_{3j}, Y_{4j}$ be i.i.d. real r.v.'s with zero expectations and finite moments of order 4. Let $\overline{Y}_i = (\sum_{j=1}^n Y_{ij})n^{-1}$. Then, up to $o(n^{-2})$,

(3.7d) $$\operatorname{Cov}\left(\overline{Y}_1^2, \overline{Y}_2\,\overline{Y}_3\right) = 2\operatorname{Cov}(Y_{11}, Y_{21})\operatorname{Cov}(Y_{11}, Y_{13})/n^2,$$

(3.7e) $$\operatorname{Cov}\left(\overline{Y}_1\,\overline{Y}_2, \overline{Y}_3\,\overline{Y}_4\right) = \{\operatorname{Cov}(Y_{11}, Y_{31})\operatorname{Cov}(Y_{21}, Y_{41})$$
$$+\operatorname{Cov}(Y_{11}, Y_{41})\operatorname{Cov}(Y_{21}, Y_{31})\}/n^2.$$

Proof of Theorem 3.1. Statement (i) follows from easy direct calculation by the delta method, which is easily justified since $E$ stands for $EI_U$, and (ii) follows from fairly direct but involved calculations, making use of the proof of part (iii)(a) below. We refer the reader to Ghosh and Subramanyam [(1974), pages 344 and 345].

(iv) follows easily from (iii)(a). To see this, note, by (3.7a),

$$\begin{aligned}\hat{\theta} + c(\hat{\theta})/n &= \hat{\theta}^* + b(\hat{\theta})/n \\ &= \hat{\theta}^* + b(\theta_0)/n + b'(\theta_0)\big(\hat{\theta} - \theta_0\big)/n + \text{smaller terms.}\end{aligned}$$

By the delta method,

$$\begin{aligned}&E\left\{\left(\hat{\theta} + \frac{c(\hat{\theta})}{n} - \theta_0\right)^2 \middle| \theta_0\right\} \\ &\quad = E\left\{\left(\hat{\theta}^* - \theta_0\right)^2 | \theta_0\right\} + \frac{b^2(\theta_0)}{n^2} + 2\frac{b'(\theta_0)}{n}E\left(\hat{\theta} - \theta_0\right)^2 + o(n^{-2}) \\ &\quad = E\left\{\left(\hat{\theta}^* - \theta_0\right)^2 | \theta_0\right\} + \frac{b_2(\theta_0)}{n^2} + \frac{2b'(\theta_0)}{n^2 I(\theta_0)} + o(n^{-2}).\end{aligned}$$

Similarly,

$$\begin{aligned}E\left\{(T_n - \theta_0)^2 | \theta_0\right\} = E\left\{(T_n^* - \theta_0)^2 | \theta_0\right\} &+ b^2(\theta_0)/n^2 \\ &+ 2b'(\theta_0)/n^2 I(\theta_0) + o(n^{-2}).\end{aligned}$$

Hence (iv) follows from parts (i) and (iii)(a).

(iii)(a) The idea behind the proof is to show

$$T_n^* - \theta_0 = \hat{\theta}_n^* - \theta_0 + R_n,$$

where $R_n$ is orthogonal to $(\hat{\theta}_n^* - \theta_0)$ in the sense $E\{R_n(\hat{\theta}_n^* - \theta_0)|\theta_0\} = o(n^{-2})$.

STEP 1. Let $I = I(\theta_0)$,

$$(3.8a) \qquad Z_n = n^{-1} \frac{d \log p(X_1, X_2, \ldots, X_n|(\theta))}{d\theta}\bigg|_{\theta_0} = \sum \beta'^{(i)}(p_i - \pi_i(\theta_0)),$$

$$(3.8b) \qquad W_n = n^{-1} \frac{d^2 \log p}{d\theta^2}\bigg|_{\theta_0} + I = \sum \beta''^{(i)}(p_i - \pi_i(\theta_0)),$$

$$(3.8c) \qquad S_n' = I^{-2}(\theta_0)(Z_n W_n) + (2J^3)^{-1} J Z_n^2,$$

where $J = J(\theta_0)$ is defined in (iii)(b) of the theorem.

Essentially by Taylor expansion, as in Section 2.6,

$$(3.9a) \qquad \hat{\theta}_n - \theta_0 = Z_n/I + S_n' + \hat{R}_n,$$

where the remainder $\hat{R}_n = O(n^{-3/2})$ as defined earlier in this chapter, that is, $E(|\hat{R}_n|^2|\theta_0) = O(n^3)$.

Let

$$S_n = S_n' - E(S_n'|\theta_0)$$

and note that

$$E(S_n'|\theta_0) = b_0(\theta_0)/n.$$

It follows that

$$\hat{\theta}_n - \theta_0 = Z_n/I + S_n + b_0(\theta_0)/n + \hat{R}_n. \tag{3.9b}$$

STEP 2. On the set $U$ introduced earlier in this chapter, the FC $T_n$ estimate has a Taylor expansion around $H(\pi(\theta)) = \theta$,

$$\begin{aligned} T_n = H(p) &= \theta + \sum l_i(p_i - \pi_i(\theta)) \\ &\quad + \sum l_{ij}(p_i - \pi_i(\theta))(p_j - \pi_j(\theta)) \\ &\quad + o(n^{-3/2}) \\ &\quad + \sum l_{ijk}(p_i - \pi_i(\theta))(p_j - \pi_j(\theta))(p_k - \pi_k(\theta)), \end{aligned} \tag{3.10}$$

where $l_i = l_i(\pi(\theta))$, and so on. We record for later use

$$\sum l_i(\pi(\theta))\pi_i'(\theta) = \frac{dH}{d\theta} = \frac{d\theta}{d\theta} = 1. \tag{3.10a}$$

Let the sum of the first three terms on the right side of (3.10) be denoted by $H_2(p)$, and let $H_3(p)$ stand for the sum when the fourth term is included. The terms on the right side of (3.10) will be denoted as the constant, linear, quadratic, and cubic terms. Write $E$ for $E\{|\theta\}$ and note

$$E(T_n) = E(H_2) + o(n^{-1}) \tag{3.11}$$

and

$$E(H_2) = \theta + b(\theta)/n. \tag{3.12}$$

It follows, at least formally, on differentiating under the expectation sign, that

$$\begin{aligned} E(H_2 Z_n) &= E(T_n Z_n) + o(n^{-2}) \\ &= \frac{1}{n}\frac{d}{d\theta}E(T_n) + o(n^{-2}) \\ &= \frac{1}{n}\frac{d}{d\theta}E(H_2) + o(n^{-2}) \\ &= \frac{1}{n} + \frac{b'(\theta)}{n^2} + o(n^{-2}), \end{aligned} \tag{3.12a}$$

that is,

$$E(H_2 Z_n) = \frac{1}{n} + \frac{b'(\theta)}{n^2} + o(n^{-2}). \tag{3.13}$$

The second and third line in (3.12a) are hard to justify, so we prove (3.13) directly. Clearly

$$E(H_2 - H_3)Z_n = o(n^{-2}). \tag{3.13a}$$

So to prove (3.13), it suffices to show

$$E(H_3 Z_n) = \frac{1}{n} + \frac{b'(\theta)}{n^2} + o(n^{-2}). \tag{3.13b}$$

Toward this end, note, using (3.10a),

$$E\{\text{linear term in } H_3\}Z_n = \frac{1}{n}\sum l_i\pi_i'(\theta) = \frac{1}{n}. \tag{3.14a}$$

Also

$$E(\theta Z_n) = 0. \tag{3.14b}$$

Moreover,

$$\begin{aligned} &E\{(\text{quadratic term of } H_3)Z_n\} \\ &\quad = \frac{1}{2n}\frac{d}{d\theta}E\Big\{\sum_{i,j} l_{ij}(p_i - \pi_i(\theta))(p_j - \pi_j(\theta))\Big\} \\ &\qquad + \frac{1}{2n}E\Big\{\sum_{i,j} l_{ij}\frac{d}{d\theta}(p_i - \pi_i(\theta))(p_j - \pi_j(\theta))\Big\} \\ &\qquad - \frac{1}{2n}E\Big\{\sum_{i,j,k} l_{ijk}\pi_k'(\theta)(p_i - \pi_i(\theta))(p_j - \pi_j(\theta))\Big\}. \end{aligned} \tag{3.14c}$$

To prove this, apply Leibnitz's rule to evaluate

$$\frac{d}{d\theta}\{l_{ij}(p_i - \pi_i(\theta))(p_j - \pi_j(\theta))p(X_1, X_2, \dots, X_n|\theta)\}$$

and observe that interchange of differentiation and expectation is justified for polynomials in $p$, since $\beta(\theta)$ is in the interior of the natural parameter space. The first term on the right side of (3.14c) is $b'(\theta)/n^2$, and the second term on the right side of (3.14c) is

$$\frac{2}{2n}\sum_{i,j} l_{ij}\pi_i'(\theta)(p_j - \pi_j(\theta)) = 0.$$

So,

$$\begin{aligned} &E\{(\text{quadratic term in } H_3)Z_n\} \\ &\quad = \frac{b'(\theta)}{n^2} - \frac{1}{2n}E\Big\{\sum_{i,j,k} l_{ijk}\pi_k'(p_i - \pi_i)(p_j - \pi_j)\Big\}. \end{aligned} \tag{3.14d}$$

Finally, in a similar way,

$$\begin{aligned} &E\{(\text{cubic term in } H_3)Z_n\} \\ &\quad = \frac{1}{n}\frac{d}{d\theta}E(\text{cubic term}) \\ &\qquad - \frac{1}{6n}E\Big\{\sum_{i,j,k}\Big(\frac{d}{d\theta}l_{ijk}\Big)(p_i - \pi_i)(p_j - \pi_j)(p_k - \pi_k)\Big\} \\ &\qquad + \frac{3}{6n}E\Big\{\sum_{i,j,k} l_{ijk}\pi_k'(p_i - \pi_i)(p_j - \pi_j)\Big\}. \end{aligned} \tag{3.14e}$$

The first two terms on the right side of (3.14e) are $o(n^{-2})$, while the last term here cancels the last term in (3.14d). Hence, adding up the right sides of (3.14a), (3.14b), (3.14d) and (3.14e), we get (3.13b). Hence, via (3.13a), (3.13) is proved.

It follows, using (3.13) and expanding $b(H_2)$,

$$\begin{aligned} E(T_n^* Z_n | \theta_0) &= E\{(H_2 - b(H_2)/n) Z_n | \theta_0\} + o(n^{-2}) \\ &= \frac{1}{n} + \frac{b'(\theta_0)}{n^2} - \frac{b'(\theta_0)}{n} E(H_2 Z_n | \theta_0) + o(n^{-2}), \end{aligned} \tag{3.15}$$

$$E(H_2 Z_n | \theta_0) = \frac{1}{n} + o(n^{-1}). \tag{3.16}$$

Hence, using (3.16) in (3.15), we get

$$E(T_n^* Z_n | \theta_0) = \frac{1}{n} + o(n^{-2}). \tag{3.17}$$

STEP 3. We now use the fact that $T_n$ is FOE. The asymptotic variance of $\sqrt{n}(T_n - \theta_0)$ equals

$$\mathrm{Var}\left(\sqrt{n} \sum l_i (p_i - \pi_i(\theta_0)) | \theta_0\right),$$

which, by Cauchy–Schwarz, (3.15) and (3.16),

$$\geq \frac{1}{I(\theta_0)} \tag{3.18}$$

with equality if and only if

$$\sum l_i (p_i - \pi_i(\theta_0)) = Z_n / I. \tag{3.19}$$

Hence, by (3.8a),

$$l_i = \beta'^{(i)}(\theta_0) / I(\theta_0). \tag{3.20}$$

Since (3.20) is true for all $\theta_0$, we may differentiate it once more, getting

$$\sum l_{ij} \pi_j'(\theta_0) = \frac{d}{d\theta} \left( \frac{\beta'^{(i)}(\theta)}{I(\theta)} \right) \bigg|_{\theta_0}. \tag{3.21}$$

By (3.10) and (3.14),

$$T_n^* - \theta_0 = T_n - \theta_0 - b(\theta)/n + O(n^{-3/2}) = Z_n / I + O(n^{-1}), \tag{3.22}$$

$$\hat{\theta}_n^* - \theta_0 = \hat{\theta} - \theta_0 - b_0(\theta)/n + O(n^{-3/2}) = Z_n / I + O(n^{-1}) \tag{3.23}$$

so that

$$(T_n^* - \theta_0) - (\hat{\theta}_n^* - \theta_0) = O(n^{-1}). \tag{3.24}$$

Also, by (3.9b) and (3.7b),

$$\hat{\theta}_n^* - \theta_0 = (Z_n / I)(1 - b_0'(\theta_0)/n) + S_n + O(n^{-3/2}). \tag{3.25}$$

STEP 4. Let $A_n = T_n - \theta_0 - Z_n/I$. By (3.20),

$$A_n = \tfrac{1}{2}\sum l_{ij}(p_i - \pi_i(\theta_0))(p_j - \pi_j(\theta_0)) + O(n^{-3/2}). \tag{3.26}$$

We claim

$$E(A_n S_n|\theta_0) = \text{a constant (not depending on } T_n) + o(n^{-2}). \tag{3.27}$$

In view of the definition of $S_n$, it is enough to prove the above with $S_n$ replaced by $Z_n^2 - I/n$ or $Z_nW_n - E(Z_nW_n)$, and in view of (3.22) and the fact that $S_n = O(n^{-1})$, it is enough to replace $A_n$ by $\Sigma l_{ij}(p_i - \pi_i(\theta_0))(p_j - \pi_j(\theta_0))$. In the following, $E$ stands for $E\{|\theta_0\}$:

$$\begin{aligned}
&E\left\{\left(\sum_{i,j} l_{ij}(p_i - \pi_i)(p_j - \pi_j)\right)(Z_n^2 - I/n)\right\} \\
&\quad= \sum_{i,j} l_{ij}\operatorname{Cov}\{Z_n^2, (p_i - \pi_i)(p_j - \pi_j)\} \\
&\quad= \sum_{i,j} l_{ij}\operatorname{Cov}\{Z_n, p_i\}\cdot\operatorname{Cov}\{Z_n, p_j\} + o(n^{-2}) \quad [\text{by (3.7c)}] \\
&\quad= \sum_i \operatorname{Cov}\{Z_n, p_i\}\sum_i l_{ij}\pi_j' + o(n^{-2}) \quad [\text{by (3.15)}] \\
&\quad= \sum_i \operatorname{Cov}\{Z_n, p_i\}\frac{d}{d\theta}\left(\frac{\beta'^{(i)}}{I}\right)\Bigg|_{\theta_0} + o(n^{-2}) \quad [\text{by (3.21)}].
\end{aligned} \tag{3.28}$$

This does not involve $H$ or its derivatives. The proof with $Z_nW_n - E(Z_nW_n)$ replacing $Z_n^2 - I/n$ is quite similar; it makes use of (3.7d), (3.15) and (3.21). Details are omitted. [The interested reader may consult Lemma 4 of Ghosh and Subramanyam (1974).]

STEP 5.

$$\begin{aligned}
&E\left\{(T_n^* - \theta_0) - (\hat\theta_n^* - \theta_0)\right\}(\hat\theta_n^* - \theta_0) \\
&\quad= E\left\{(T_n^* - \theta_0) - (\hat\theta_n^* - \theta_0)\right\}\left\{\frac{Z_n}{I}\left(1 - \frac{b_0'(\theta_0)}{n}\right) + S_n\right\} + o(n^{-2}) \\
&\qquad\qquad [\text{by (3.24) and (3.25)}] \\
&\quad= E\left\{(T_n^* - \theta_0) - (\hat\theta_n^* - \theta_0)\right\}\{S_n\} + o(n^{-2}) \\
&\qquad\qquad \left(\text{by Step 2, applied to } T_n^* \text{ and } \hat\theta_n^*\right) \\
&\quad= E\left\{(T_n - \theta_0) - (\hat\theta_n - \theta_0), S_n|\theta_0\right\} + o(n^{-2}) \\
&\qquad\qquad [\text{by (3.22), (3.23) and } E(S_n) = 0] \\
&\quad= o(n^{-2}) \quad \left(\text{by Step 4 applied to } T_n \text{ and } \hat\theta_n\right).
\end{aligned}$$

This completes the proof of (iii)(a). The proof (iii)(b) follows from straightforward direct calculation; see Ghosh and Sumbramanyam [(1974), page 344]. □

The fact that $\hat{\theta}$ minimizes $E_2$ among FC estimates is the Fisher–Rao theorem. The fact that the third order unbiased estimate $\hat{\theta}_n^*$ based on $\hat{\theta}_n$ minimizes the variance [up to $o(n^{-2})$] among third order unbiased estimates $T_n^*$ based on FC, FOE $T_n$, is Rao's theorem. These results were proved for a multinomial by Rao, (1961, 1962, 1963). The present proof is taken from Ghosh and Subramanyam (1974), except that whenever we use their Lemma 3, we provide complete justification and streamline the proof of (iii)(a) in Theorem 3.1. Even if specialized to the multinomial, the present argument is different from Rao's.

Similar theorems have been obtained by Efron (1975) and also by Akahira and Takeuchi and Pfanzagl and his co-workers, at about the same time. Their results are described in Akahira and Takeuchi (1981) and Pfanzagl (1979).

The fact that $\hat{\theta}$ minimizes $E_2$ is interesting as well as amenable to geometrical analysis. However, the properties (iii) and (iv) seem the most convenient if one wants to apply these results to get better estimates: (iii) is relevant if unbiasedness is a concern and (iv) is relevant in the absence of that. It is worth mentioning that the $o(n^{-2})$ term in (i) is uniformly so on compact $\theta_0$-sets. This is true of (iv) also.

If one compares a FC, FOE estimate $T_n$ directly with $\hat{\theta}_n$ [i.e., without either of the adjustments in (iii) or (iv)], then, in general, neither is better than the other in the third order for all $\theta_0$.

**3.2. Third order efficient approximate solution of likelihood equation.** Since the mle is often hard to calculate, one may wish to have approximate solutions of the likelihood equations which will continue to be TOE. Let $T_n$ be a $\sqrt{n}$-consistent estimate of $\theta$, that is, $\sqrt{n}(T_n - \theta) = O_p(1)$. The likelihood equation may be written as

$$0 = \frac{d \log L_n}{d\theta}\bigg|_{\hat{\theta}} = \frac{d \log L_n}{d\theta}\bigg|_{T_n} + \left(\hat{\theta} - T_n\right)\frac{d^2 \log L_n}{d\theta^2}\bigg|_{T_n}$$

$$+ \frac{\left(\hat{\theta} - T_n\right)^2}{2}\frac{d^3 \log L_n}{d\theta^3}\bigg|_{T_n} + \frac{\left(\hat{\theta} - T_n\right)^3}{6}\frac{d^4 \log L_N}{d\theta^3}\bigg|_{T_n} + O_p\left(\frac{1}{n}\right).$$

Then the successive approximations to $\sqrt{n}(\hat{\theta}_n - \theta)$ are

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \sqrt{n}(T_n - \theta) + U_{in} + R_{in}, \qquad i = 1,2,3,$$

where $R_{in} = O_p(n^{-i/2})$,

$$U_{1n} = \frac{\left((1/\sqrt{n})(d \log L_n/d\theta)|_{T_n}\right)}{\left(-(1/n)(d^2 \log L_n/d\theta^2)|_{T_n}\right)},$$

$$U_{2n} = \left\{\frac{1}{\sqrt{n}}\frac{d \log L_n}{d\theta}\bigg|_{T_n} + \frac{U_{1n}^2}{2}\frac{1}{n^{3/2}}\frac{d^3 \log L_n}{d\theta^3}\bigg|_{T_n}\right\}\left(-\frac{1}{n}\frac{d^2 \log L_n}{d\theta^2}\bigg|_{T_n}\right)^{-1},$$

$$U_{3n} = \left\{ \frac{1}{\sqrt{n}} \frac{d \log L_n}{d\theta}\bigg|_{T_n} + \frac{U_{2n}^2}{2} \frac{1}{n^{3/2}} \frac{d^2 \log L_n}{d\theta^3}\bigg|_{T_n} + \frac{U_{2n}^3}{6} \frac{1}{n^2} \frac{d^4 \log L_n}{d\theta^4}\bigg|_{T_n} \right\}$$
$$\times \left\{ -\frac{1}{n} \frac{d^2 \log L_n}{d\theta^2}\bigg|_{T_n} \right\}^{-1}.$$

Our approximate solution is $\tilde{\theta} = T_n + U_{3n}$.

If $T_n$ is FC with a thrice continuously differentiable $H$, in addition to being $\sqrt{n}$ consistent, we can Taylor expand $\tilde{\theta}$ in $U$ and show, with our present interpretation of $E = EI_U$, that

$$E\{(\tilde{\theta} - \theta)|\theta\} = E\{(\hat{\theta} - \theta)|\theta\} + o(n^{-1}),$$
$$E\{(\tilde{\theta} - \theta)^2|\theta\} = E\{(\tilde{\theta} - \theta)^2|\theta\} + o(n^{-2}),$$

so that $\tilde{\theta}$ will inherit third order optimal properties of $\hat{\theta}$.

Alternatively, in addition to $\sqrt{n}$-consistency, one may require $T_n$ to have a moderate deviation property, that is, there exists a sufficiently large $c$ such that $P\{|T_n - \theta| > c\sqrt{\log n} / \sqrt{n}\} = o(n^{-1})$ uniformly on compact sets of $\theta$. In this case one can show $R_{3n}$ satisfies the conditions of Lemma 2.1 with $s = 4$. Hence if $\sqrt{n}(\hat{\theta} - \theta)$ has a valid Edgeworth expansion for $s = 4$, then so will $\sqrt{n}(\tilde{\theta} - \theta)$. This means the first interpretation of expansions of moments in Section 2.7 is applicable. We may also use the interpretation based on the result of Götze and Hipp (1978).

**3.3. Example 2.3 (Berkson's bioassay example) revisited.** Suppose, for simplicity, $\beta$ is known and the parameter of interest if $\alpha$. The likelihood equation is

$$0 = \sum_{i=1}^{k} n(p_i - \pi_i(\alpha)).$$

Let $\hat{\alpha}_n$ be the mle. Let $L_i = \log\{\pi_i/(1 - \pi_i)\} = \alpha + \beta d_i$ and $l_i = \log\{p_i/(1 - p_i)\}$. Minimizing $\Sigma p_i(1 - p_i)(l_i - L_i)^2$ with respect to $\alpha$, one gets Berkson's minimum logit chi-square estimate $T_n$. $T_n$ is the solution of

$$\Sigma p_i(1 - p_i)(l_i - L_i^*) = 0,$$

where $L_i^* = T_n + \beta d_i$. Clearly $T_n$ is easy to calculate explicitly, unlike the mle $\hat{\alpha}_n$. It is also FC and FOE. So on the grounds of easy calculation, one might prefer $T_n$ to $\hat{\alpha}_n$ if first order efficiency were the only other concern. If we invoke third order efficiency, we can do strictly better than $T_n$ by using a suitable perturbation of $\hat{\alpha}_n$; see Theorem 3.1, part (iv). If third order unbiasedness is also a concern, we can perturb $T_n$ to make it unbiased up to $o(n^{-1})$, and do the same with $\hat{\alpha}_n$. By Theorem 3.1, part (iii)(a), we know $\hat{\alpha}_n^*$ is third order better than $T_n^*$. Actually, in this example $\hat{\alpha}_n^*$ is strictly better in the third order sense. Some simulations reported in Subramanyam's 1980 thesis (submitted to the Indian Statistical Institute) bear this out.

This example is actually a linear exponential and so a complete sufficient statistic exists, namely, $\Sigma_1^k p_i$. $T_n^*$ is not a function of it, but $\alpha_n^*$ is. In fact $\hat{\alpha}_n^*$ may be thought of as an approximation to the Rao–Blackwellized estimate $E(T_n^* | \Sigma p_i)$, which is superior to $T_n^*$.

Since $T_n$ is FC and FOE, we may use it to approximate $\hat{\alpha}_n$ or $\hat{\alpha}_n^*$; see Section 3.2.

**3.4. Multiparameter extensions.** To fix ideas, suppose $\theta = (\theta_1, \theta_2)$, $T_n = H(\bar{Z})$ is FC, FOE (in the sense that each component of $T_n$ is FC, FOE) and $\hat{\theta} = H_0(\bar{Z})$ is mle. If one considers the expansions of the dispersion matrix of $T_n^* = T_n - b(T_n)/n$ and $\hat{\theta}_n^* = \hat{\theta}_n - b_0(\theta)/n$, then the difference of the coefficient matrices of $n\ ^{2}$ is positive semidefinite. This is a generalization of Rao's theorem.

To generalize the Fisher–Rao theorem, consider the analogous of $E_2$, namely, $\inf_\lambda E_{12}(\lambda, H, \theta_0)$ and $\inf_\gamma E_{22}(\gamma, H, \theta_0)$, where

$$L_n = p(X_1, X_2, \dots, X_n | \theta),$$

$$E_{12}(\lambda, H, \theta_0) = \lim_{n \to \infty} \operatorname{Var}_{\theta_0} \left[ \frac{\partial \log L_N}{\partial \theta_1} \bigg|_{\theta_0} - \sum n(T_{in} - \theta_{i0}) I_{1i}(\theta_0) \right. \tag{3.29}$$
$$\left. - \sum n \lambda_{ij} (T_{in} - \theta_{j0})(T_{jn} - \theta_{j0}) \right],$$

$$E_{22}(\gamma, H, \theta_0) = \lim_{n \to \infty} \operatorname{Var}_{\theta_0} \left[ \frac{\partial \log L_N}{\partial \theta_2} \bigg|_{\theta_0} - \sum_i n(T_{in} - \theta_{i0}) I_{2i}(\theta_0) \right. \tag{3.30}$$
$$\left. - \sum n \gamma_{ij} (T_{in} - \theta_{i0})(T_{jn} - \theta_{i0}) \right].$$

Let the random variables within [ ] above be denoted as $Y_1(n, \lambda, H, \theta_0)$ and $Y_2(n, \gamma, H, \theta_0)$. Then one can show that for each $\lambda, \gamma$, the limiting dispersion matrix of

$$\{Y_1(n, \lambda, H, \theta_0) - Y_1(n, \lambda, H_0, \theta_0)\}$$

and

$$\{Y_2(n, \gamma, H, \theta_0) - Y_2(n, \gamma, H_0, \theta_0)\}$$

is positive semidefinite. (This implies that $E_{12}$, $E_{22}$ are minimized by $\hat{\theta}$.) This is the generalization of the Fisher-Rao theorem.

The proof in both cases is exactly identical.

The extensions are taken from Ghosh and Subramanyam (1974).

**3.5. Example 2.4 (Behrens–Fisher) revisited.** The likelihood equations reduce to

$$\hat{\sigma}_1^2 = s_1^2 + (\bar{U} - \hat{\mu})^2, \tag{3.31}$$

$$\sigma_2^2 = s_2^2 + (\bar{U} - \hat{\mu})^2, \tag{3.32}$$

$$0 = A(\hat{\mu}), \tag{3.33}$$

where

$$A(\mu) = (\overline{U} - \mu)\left(s_1^2 + (\overline{U} - \mu)^2\right) + (\overline{V} - \mu)\left(s_2^2 + (\overline{V} - \mu)^2\right),$$

and $s_1^2 = n^{-1}\Sigma_1^n(U_j - \overline{U})^2$, $s_2^2 = n^{-1}\Sigma_1^n(V_j - \overline{V})^2$, $\overline{U} = n^{-1}\Sigma_1^n U_j$ and $\overline{V} = n^{-1}\Sigma V_j$.

Equation (3.33) means $\hat{\mu}$ is a weighted mean of $\overline{U}$ and $\overline{V}$ with the weights being proportional to the inverse of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. Note that (3.33) is a cubic equation, $A(\mu)$ changes sign only when $\mu$ is between $\overline{U}$ and $\overline{V}$, $A(\overline{U})$ and $A(\overline{V})$ have opposite signs and $A'(\mu) < 0$ for $\mu$ between $\overline{U}$ and $\overline{V}$. It follows that (3.33) has a unique real root, and the root lies between $\overline{U}$ and $\overline{V}$. This root must be consistent as both $\overline{U}$ and $\overline{V}$ are, so this is $\hat{\mu}$.

It is also clear that if we define $T_n$ as the weighted mean of $\overline{U}$ and $\overline{V}$, with the weights proportional to the inverse of $s_1^2$ and $s_2^2$, then $T_n$ is FC and FOE. So $T_n$ may be used to approximate $\hat{\mu}$.

Let

$$\hat{W}_1 = \hat{\sigma}_2^2/\left(\hat{\sigma}_1^2 + \hat{\sigma}_2^2\right), \qquad W_1 = s_1^2/\left(s_1^2 + s_2^2\right).$$

Then

$$(3.34) \quad n(\hat{\mu} - T_n) = \sqrt{n}\left(\hat{W}_1 - W_1\right)\left\{\sqrt{n}\left(\overline{U} - \mu\right) - \sqrt{n}\left(\overline{V} - \mu\right)\right\} \to_P 0.$$

Also, using the fact that $s_1^2, s_2^2$ are independent of $\overline{U}, \overline{V}$,

$$(3.35) \qquad E(T_n|\theta) = \mu.$$

Using (3.34) and (3.35), it is easy to show

$$E(\hat{\mu}|\theta) = \mu + o(n^{-1}),$$

that is, $\hat{\mu}$ is unbiased up to third order. Since

$$n^{3/2}\left[\left(\overline{U} - \hat{\mu}\right)^2 - \left(\overline{U} - T_n\right)^2\right] \to_P 0,$$

$$n^{3/2}\left[\left(\overline{V} - \hat{\mu}\right)^2 - \left(\overline{V} - T_n\right)^2\right] \to_P 0,$$

let

$$\tilde{\sigma}_1^2 = s_1^2 + \left(\overline{U} - T_n\right)^2, \qquad \tilde{\sigma}_2^2 = s_2^2 + \left(\overline{V} - T_n^2\right)$$

and $\tilde{\mu}$ be the weighted mean of $\overline{U}, \overline{V}$ with weights inversely proportional to $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$. Then $(\tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \tilde{\mu})$ has the third order properties of $\hat{\theta}$, and, in particular, $\tilde{\mu}$ is third order efficient among all third order unbiased estimates.

Easy calculation shows

$$E\left(\tilde{\sigma}_1^2|\theta\right) = \sigma_1^2 + \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}\right) + o(n^{-1}),$$

and similarly for $\tilde{\sigma}_2^2$.

**3.6. Hodges–Lehmann deficiency.** Comparing the coefficient of $n^{-2}$ in the expansion of variance or mean square can be interpreted in terms of "equivalent" sample sizes in the manner of Hodges and Lehmann (1970).

Suppose $T_{1n}$ and $T_{2n}$ are two estimates and

$$E\{(T_{in} - \theta)^2|\theta\} = \frac{a_1}{n} + \frac{a_{i2}}{n^2} + o(n^{-2}).$$

Suppose we define two sample sizes $n_1$ and $n_2$ to be equivalent if $T_{1n_1}$ and $T_{2_{n2}}$ have (nearly) equal accuracy in the sense

$$\frac{a_1}{n_1} + \frac{a_{12}}{n_1^2} = \frac{a_1}{n_2} + \frac{a_{22}}{n_2^2} + o(n_2^{-2}).$$

Then, as shown in Hodges and Lehmann (1970),

$$\lim_{n_2 \to \infty} \frac{n_1}{n_2} = 1, \qquad \lim_{n_2 \to \infty} (n_2 - n_1) = d.$$

Then $d$ must satisfy

$$\frac{a_1}{n_1} - \frac{a_1}{n_1 + d} = \frac{a_{22}}{(n_1 + d)^2} - \frac{a_{12}}{n_1^2} + o(n_1^{-2}),$$

which, on simplification, leads to

$$a_1 \cdot d = \frac{(a_{22} - a_{12})n_1^2 + o(n_1^2)}{n_1^2 + o(n_1^2)},$$

so that

$$d = \frac{a_{22} - a_{12}}{a_1},$$

a formula due to Hodges and Lehmann (1970). As noted in Ghosh and Subramanyam [(1974), page 347], Theorem 3.1 then shows $T_n^*$ is deficient with respect to $\hat{\theta}^*$ in the sense of requiring $(\psi^*(H, \theta_0) - \psi^*(\theta_0))/I(\theta_0)$ additional observations for the same accuracy.

**3.7. Asymptotic sufficiency.** The following is a slight correction of a result stated without proof in Ghosh and Subramanyam (1974). Under suitable regularity conditions, there is a compact neighborhood $\Theta_0$ of $\theta_0$ and densities $q_{\theta, n}$, $\theta \in \Theta_0$, such that

$$\text{(A)} \qquad \hat{\theta}, \left.\frac{d^i \log p(X_1, X_2, \dots, X_n|\theta)}{d\theta^i}\right|_{\hat{\theta}} \equiv L_i, \qquad i = 2, 3,$$

are sufficient for $q_{\theta, n}$ and

$$\text{(B)} \quad \sup_{\theta \in \Theta_0} \int \cdots \int |p(x_1, x_2, \dots, x_n|\theta) - q_{\theta, n}(x_1, x_2, \dots, x_n|dx_1, \dots, dx_n)| \\ = o(n^{-1}).$$

The proof is by Taylor expanding to get an approximation and then normalizing to get a density.

In a similar sense, $\hat{\theta}$ and $L_2$ are asymptotically sufficient up to $o(n^{-1/2})$. Asymptotic sufficiency of $\hat{\theta}$ and $L_2$ in a different sense appears in Chapter 4.

Since the $q_{\theta,n}$'s are curved exponentials, in a sense Theorem 3.1 and the above approximation theorem can be used to generalize third order results to general densities. We follow a different route in Chapter 6.

**3.8. Third order efficiency and Bhattacharya bounds.** Since first order efficiency has a small sample analogue in the Cramér–Rao inequality, it is natural to ask if the third order results have an analogue in the Bhattacharya bounds. That is not the case.

Ghosh and Subramanyam [(1974), page 350] note that if one regresses $T_n$ on $d \log p/d\theta$ and $(1/p)d^2p/d\theta^2$ as in Bhattacharya bounds, one gets

$$E\{(T_n^* - \theta_0)^2(\theta_0)\} \geq \frac{1}{nI(\theta_0)} + \frac{1}{n^2I^4(\theta_0)}\left(\frac{J(\theta_0)}{2} + \mu_{11}(\theta_0)\right) + o(n^{-2}),$$

which, in view of Theorem 3.1, is not sharp. In fact, it can be attained up to $o(n^{-2})$ for all $\theta$ if and only if the curved exponential is linear.

There is no small sample analogue of third order efficiency.