Chapter 14
Quadratic Mean Differentiable Families

14.1 Introduction

As mentioned at the beginning of Chapter 11, the finite-sample theory of optimality for hypothesis testing is applied only to rather special parametric families, primarily exponential families and group families. On the other hand, asymptotic optimality will apply more generally to parametric families satisfying smoothness conditions. In particular, we shall assume a certain type of differentiability condition, called *quadratic mean differentiability*. Such families will be considered in Section 14.2. In Section 14.3, the notion of *contiguity* will be developed, primarily as a technique for calculating the limiting distribution or power of a test statistic under an alternative sequence, especially when the limiting distribution under the null hypothesis is easy to obtain. In Section 14.4, these techniques will then be applied to classes of tests based on the likelihood function, namely the Wald, Rao, and likelihood ratio tests. The asymptotic optimality of these tests will be established in Chapter 15.

14.2 Quadratic Mean Differentiability (q.m.d.)

Consider a parametric model $\{P_\theta, \theta \in \Omega\}$, where, throughout this section, Ω is assumed to be an open subset of $\mathbb{R}^k$. The probability measures P_θ are defined on some measurable space $(\mathcal{X}, \mathcal{C})$. Assume each P_θ is absolutely continuous with respect to a σ-finite measure μ, and set $p_\theta(x) = dP_\theta(x)/d\mu(x)$. In this section, smooth parametric models will be considered. To motivate the smoothness condition given in Definition 14.2.1 below, consider the case of n i.i.d. random variables $X_1, \ldots, X_n$ and the problem of testing a simple null hypothesis $\theta = \theta_0$ against a simple alternative θ_1 (possibly depending on n). The most powerful test rejects when the loglikelihood ratio statistic

$$\log[L_n(\theta_1)/L_n(\theta_0)]$$

E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer Texts in Statistics, https://doi.org/10.1007/978-3-030-70578-7_14

is sufficiently large, where

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i) \tag{14.1}$$

denotes the likelihood function. We would like to obtain certain expansions of the loglikelihood ratio, and the smoothness condition we impose will ensure the existence of such an expansion.

Example 14.2.1 (**Normal Location Model**) Suppose P_θ is $N(\theta, \sigma^2)$, where σ^2 is known. It is easily checked that

$$\log[L_n(\theta_1)/L_n(\theta_0)] = \frac{n}{\sigma^2}[(\theta_1 - \theta_0)\bar{X}_n - \frac{1}{2}(\theta_1^2 - \theta_0^2)] \,, \tag{14.2}$$

where $\bar{X}_n = \sum_{i=1}^{n} X_i/n$. By the Weak Law of Large Numbers, under θ_0,

$$(\theta_1 - \theta_0)\bar{X}_n - \frac{1}{2}(\theta_1^2 - \theta_0^2) \xrightarrow{P} (\theta_1 - \theta_0)\theta_0 - \frac{1}{2}(\theta_1^2 - \theta_0^2) = -\frac{1}{2}(\theta_1 - \theta_0)^2 \,,$$

and so $\log[L_n(\theta_1)/L_n(\theta_0)] \xrightarrow{P} -\infty$. Therefore, $\log[L_n(\theta_1)/L_n(\theta_0)]$ is asymptotically unbounded in probability under θ_0. As in Example 11.3.2, a more useful result is obtained if θ_1 in (14.2) is replaced by $\theta_0 + hn^{-1/2}$. We then find

$$\log[L_n(\theta_0 + hn^{-1/2})/L_n(\theta_0)] = \frac{hn^{1/2}(\bar{X}_n - \theta_0)}{\sigma^2} - \frac{h^2}{2\sigma^2} = hZ_n - \frac{h^2}{2\sigma^2} \,, \tag{14.3}$$

where $Z_n = n^{1/2}(\bar{X}_n - \theta_0)/\sigma^2$ is $N(0, 1/\sigma^2)$. Notice that the expansion (14.3) is a linear function of Z_n and a simple quadratic function of h, with the coefficient of h^2 nonrandom. Furthermore, $\log[L_n(\theta_0 + hn^{-1/2})/L_n(\theta_0)]$ is distributed as $N(-h^2/2\sigma^2, h^2/\sigma^2)$ under θ_0 for every n. (The relationship that the mean is the negative of half the variance will play a key role in the next section.) ■

The following more general family permits an asymptotic version of (14.3).

Example 14.2.2 (**One-parameter Exponential Family**) Let $X_1, \ldots, X_n$ be i.i.d. having density

$$p_\theta(x) = \exp[\theta T(x) - A(\theta)]$$

with respect to a σ-finite measure μ. Assume θ_0 lies in the interior of the natural parameter space. Then,

$$\log[L_n(\theta_0 + hn^{-1/2})/L_n(\theta_0)] = hn^{-1/2} \sum_{i=1}^{n} T(X_i) - n[A(\theta_0 + hn^{-1/2}) - A(\theta_0)] \,.$$

Recall (Problem 2.16) that $E_{\theta_0}[T(X_i)] = A'(\theta_0)$ and $Var_{\theta_0}[T(X_i)] = A''(\theta_0)$. By a Taylor expansion,

$$n[A(\theta_0 + hn^{-1/2}) - A(\theta_0)] = hn^{1/2}A'(\theta_0) + \frac{1}{2}h^2A''(\theta_0) + o(1)$$

as $n \to \infty$, so that

$$\log[L_n(\theta_0 + hn^{-1/2})/L_n(\theta_0)] = hZ_n - \frac{1}{2}h^2A''(\theta_0) + o(1)\ , \tag{14.4}$$

where, under θ_0,

$$Z_n = n^{-1/2}\sum_{i=1}^{n}\{T(X_i) - E_{\theta_0}[T(X_i)]\} \xrightarrow{d} N(0,\, A''(\theta_0))\ .$$

Thus, the loglikelihood ratio (14.4) behaves asymptotically like the loglikelihood ratio (14.3) from a normal location model. As we will see, such approximations allow one to deduce asymptotic optimality properties for the exponential model (or any model whose likelihood ratios satisfy an appropriate generalization of (14.4)) from optimality properties of the simple normal location model. ■

We would like to obtain an approximate result like (14.4) for more general families. Classical smoothness conditions usually assume that, for fixed x, the function $p_\theta(x)$ is differentiable in θ at θ_0; that is, for some function $\dot{p}_\theta(x)$,

$$p_{\theta_0+h}(x) - p_{\theta_0}(x) - \langle \dot{p}_{\theta_0}(x), h\rangle = o(|h|)$$

as $|h| \to 0$. In addition, higher order differentiability is typically assumed with further assumptions on the remainder terms. In order to avoid such strong assumptions, it turns out to be useful to work with square roots of densities. For fixed x, differentiability of $p_\theta^{1/2}(x)$ at $\theta = \theta_0$ requires the existence of a function $\eta(x, \theta_0)$ such that

$$R(x, \theta_0, h) \equiv p_{\theta_0+h}^{1/2}(x) - p_{\theta_0}^{1/2}(x) - \langle \eta(x, \theta_0), h\rangle = o(|h|)\ .$$

To obtain a weaker, more generally applicable condition, we will not require $R^2(x, \theta_0, h) = o(|h|^2)$ for every x, but we will impose the condition that $R^2(X, \theta_0, h)$ averaged with respect to μ is $o(|h|^2)$. Let $L^2(\mu)$ denote the space of functions g such that $\int g^2(x)\, d\mu(x) < \infty$. The convenience of working with square roots of densities is due in large part to the fact that $p_\theta^{1/2}(\cdot) \in L^2(\mu)$, a fact first exploited by Le Cam; see Pollard (1997) for an explanation. The desired smoothness condition is now given by the following definition.

Definition 14.2.1 The family $\{P_\theta, \theta \in \Omega\}$ is *quadratic mean differentiable* (abbreviated q.m.d.) at θ_0 if there exists a vector of real-valued functions $\eta(\cdot, \theta_0) = (\eta_1(\cdot, \theta_0), \ldots, \eta_k(\cdot, \theta_0))^\top$ such that

$$\int_{\mathcal{X}} \left[\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - < \eta(x, \theta_0), h >\right]^2 d\mu(x) = o(|h|^2) \tag{14.5}$$

as $|h| \to 0$.[1]

The vector-valued function $\eta(\cdot, \theta_0)$ will be called the quadratic mean derivative of P_θ at θ_0. Clearly, $\eta(x, \theta_0)$ is not unique since it can be changed on a set of x values having μ-measure zero. If q.m.d. holds at all θ_0, then we say the family is q.m.d.

The following are useful facts about q.m.d. families.

Lemma 14.2.1 *Assume $\{P_\theta, \theta \in \Omega\}$ is q.m.d. at θ_0. Let $h \in \mathbb{R}^k$.*

(i) Under P_{θ_0}, $\langle \frac{\eta(X,\theta_0)}{p_{\theta_0}^{1/2}(X)}, h \rangle$ is a random variable with mean 0; i.e., satisfying

$$\int p_{\theta_0}^{1/2}(x) \langle \eta(x, \theta_0), h \rangle d\mu(x) = 0 \, .$$

(ii) The components of $\eta(\cdot, \theta_0)$ are in $L^2(\mu)$; that is, for $i = 1, \ldots, k$,

$$\int \eta_i^2(x, \theta_0) \, d\mu(x) < \infty \, .$$

PROOF. In the definition of q.m.d., replace h by $hn^{-1/2}$ to deduce that

$$\int \left\{ n^{1/2} \left[p_{\theta_0 + hn^{-1/2}}^{1/2}(x) - p_{\theta_0}^{1/2}(x) \right] - \langle \eta(x, \theta_0), h \rangle \right\}^2 d\mu(x) \to 0$$

as $n \to \infty$. But, if $\int (g_n - g)^2 \, d\mu \to 0$ and $\int g_n^2 d\mu < \infty$, then $\int g^2 d\mu < \infty$ (Problem 14.3). Hence, for any $h \in \mathbb{R}^k$, $\langle \eta(x, \theta_0), h \rangle \in L^2(\mu)$. Taking h equal to the vector of zeros except for a 1 in the ith component yields (*ii*). Also, if $\int (g_n - g)^2 d\mu \to 0$ and $\int p^2 d\mu < \infty$ then $\int p g_n d\mu \to \int p g \, d\mu$ (Problem 14.4). Taking $p = p_{\theta_0}^{1/2}$ and $g_n = n^{1/2} \left[p_{\theta_0 + hn^{-1/2}}^{1/2}(x) - p_{\theta_0}^{1/2}(x) \right]$ yields

$$\begin{aligned}
&\int p_{\theta_0}^{1/2}(x) \langle \eta(x, \theta_0), h \rangle d\mu(x) \\
&= \lim_{n \to \infty} n^{1/2} \int p_{\theta_0}^{1/2}(x) [p_{\theta_0 + hn^{-1/2}}^{1/2}(x) - p_{\theta_0}^{1/2}(x)] \, d\mu(x) \\
&= \lim_{n \to \infty} n^{1/2} \left[\int p_{\theta_0}^{1/2}(x) \, p_{\theta_0 + hn^{-1/2}}^{1/2}(x) \, d\mu(x) - 1 \right] \\
&= -\tfrac{1}{2} \lim n^{-1/2} \, n \int [p_{\theta_0}^{1/2}(x) - p_{\theta_0 + hn^{-1/2}}^{1/2}(x)]^2 d\mu(x) \, .
\end{aligned}$$

But,

[1] The definition of q.m.d. is a special case of Fréchet differentiability of the map $\theta \to p_\theta^{1/2}(\cdot)$ from Ω to $L^2(\mu)$.

$$n \int \left[p_{\theta_0}^{1/2}(x) - p_{\theta_0 + hn^{-1/2}}^{1/2}(x) \right]^2 d\mu(x)$$
$$\to \int |\langle \eta(x, \theta_0), h \rangle|^2 d\mu(x) < \infty , \tag{14.6}$$

and (i) follows. ■

Note that Lemma 14.2.1 (i) asserts that the finite-dimensional set of vectors $\{\langle \eta(\cdot, \theta_0), h \rangle,\ h \in \mathbb{R}^k\}$ in $L^2(\mu)$ is orthogonal to $p_{\theta_0}^{1/2}(\cdot)$.

It turns out that, when q.m.d. holds, the integrals of products of the components of $\eta(\cdot, \theta)$ play a vital role in the theory of asymptotic efficiency. Such values (multiplied by 4 for convenience) are gathered into a matrix, which we call the *Fisher Information matrix*. The use of the term *information* is justified by Problem 14.5.

Definition 14.2.2 For a q.m.d. family with derivative $\eta(\cdot, \theta)$, define the *Fisher Information matrix* to be the matrix $I(\theta)$ with (i, j) entry

$$I_{i,j}(\theta) = 4 \int \eta_i(x, \theta) \eta_j(x, \theta)\, d\mu(x) .$$

The existence of $I(\theta)$ follows from Lemma 14.2.1 (ii) and the Cauchy–Schwarz inequality. Furthermore, $I(\theta)$ does not depend on the choice of dominating measure μ (Problem 14.8).

Lemma 14.2.2 *For any* $h \in \mathbb{R}^k$,

$$\int |\langle h, \eta(x, \theta_0) \rangle|^2\, d\mu(x) = \tfrac{1}{4} \langle h, I(\theta_0) h \rangle .$$

PROOF. Of course

$$\langle h, \eta(x, \theta_0) \rangle = \Sigma h_i \eta_i(x, \theta_0) .$$

Square it and integrate. ■

Next, we would like to determine simple sufficient conditions for q.m.d. to hold. Assuming that the pointwise derivative of $p_\theta(x)$ with respect to θ exists, one would expect that the quadratic mean derivative $\eta(\cdot, \theta_0)$ is given by

$$\eta_i(\cdot, \theta) = \frac{\partial}{\partial \theta_i} p_\theta^{1/2}(x) = \tfrac{1}{2} \frac{\frac{\partial}{\partial \theta_i} p_\theta(x)}{p_\theta^{1/2}(x)} . \tag{14.7}$$

In fact, Hájek (1972) gave sufficient conditions where this is the case, and the following result for the case $k = 1$ is based on his argument.

Theorem 14.2.1 *Suppose Ω is an open subset of $\mathbb{R}$ and fix $\theta_0 \in \Omega$. Assume $p_\theta^{1/2}(x)$ is an absolutely continuous function of θ in some neighborhood of θ_0, for μ-almost all x.*[2] *Also, assume for μ-almost all x, the derivative $p'_\theta(x)$ of $p_\theta(x)$ with respect to θ exists at $\theta = \theta_0$. Define*

$$\eta(x, \theta) = \frac{p'_\theta(x)}{2p_\theta^{1/2}(x)} \tag{14.8}$$

if $p_\theta(x) > 0$ and $p'_\theta(x)$ exists and define $\eta(x, \theta) = 0$ otherwise. Also, define

$$I(\theta) = 4 \int \eta^2(x, \theta)\mu(x) \, ,$$

and assume that $I(\theta)$ is finite and continuous in θ at θ_0. Then, $\{P_\theta\}$ is q.m.d. at θ_0 with quadratic mean derivative $\eta(\cdot, \theta_0)$ and so $I(\theta)$ is the Fisher Information.

PROOF. If $p_\theta(x) > 0$ and $p'_\theta(x)$ exists, then from standard calculus it follows that

$$\frac{d}{d\theta} p_\theta^{1/2}(x) = \eta(x, \theta) \, .$$

Also, if $p_\theta(x) = 0$ and $p'_\theta(x)$ exists, then $p'_\theta(x) = 0$ (since $p_\theta(\cdot)$ is nonnegative). Now, if x is such that $p_\theta^{1/2}(x)$ is absolutely continuous in $[\theta_0, \theta_0 + \delta]$, then

$$\left\{ \frac{1}{\delta}[p_{\theta_0+\delta}^{1/2}(x) - p_{\theta_0}^{1/2}(x)] \right\}^2 = \frac{1}{\delta^2} \left[\int_0^\delta \eta(x, \theta_0 + \lambda) d\lambda \right]^2 \leq \frac{1}{\delta} \int_0^\delta \eta^2(x, \theta_0 + \lambda) d\lambda \, .$$

Integrating over all x with respect to μ yields

$$\int \left\{ \frac{1}{\delta}[p_{\theta_0+\delta}^{1/2}(x) - p_{\theta_0}^{1/2}(x)] \right\}^2 d\mu(x) \leq \frac{1}{4\delta} \int_0^\delta I(\theta_0 + \lambda) d\lambda \, .$$

By continuity of $I(\theta)$ at θ_0, the right-hand side tends to

$$\frac{1}{4} I(\theta_0) = \int \eta^2(x, \theta_0) d\mu(x)$$

as $\delta \to 0$. But, for μ-almost all x,

$$\frac{1}{\delta}[p_{\theta_0+\delta}^{1/2}(x) - p_{\theta_0}^{1/2}(x)] \to \eta(x, \theta_0) \, .$$

[2] A real-valued function g defined on an interval $[a, b]$ is absolutely continuous if $g(\theta) = g(a) + \int_a^\theta h(x)dx$ for some integrable function h and all $\theta \in [a, b]$; Problem 2 on p. 182 of Dudley (1989) clarifies the relationship between this notion of absolute continuity of a function and the general notion of a measure being absolute continuous with respect to another measure, as defined in Section 2.2.

The result now follows by Vitali's Theorem (Corollary 2.2.1). ■

Corollary 14.2.1 *Suppose μ is Lebesgue measure on $\mathbb{R}$ and that $p_\theta(x) = f(x-\theta)$ is a location model, where $f^{1/2}(\cdot)$ is absolutely continuous. Let*

$$\eta(x,\theta) = \frac{-f'(x-\theta)}{2f^{1/2}(x-\theta)}$$

if $f(x-\theta) > 0$ and $f'(x-\theta)$ exists; otherwise, define $\eta(x,\theta) = 0$. Also, let

$$I = 4\int_{-\infty}^{\infty} \eta^2(x,0)dx\ ,$$

and assume $I < \infty$. Then, the family is q.m.d. at θ_0 with quadratic mean derivative $\eta(x,\theta_0)$ and constant Fisher Information I.

The assumption that $f^{1/2}$ is absolutely continuous can be replaced by the assumption that f is absolutely continuous; see Hájek (1972), Lemma A.1. For other conditions, see Le Cam and Yang (2000), Section 7.3.

Example 14.2.3 (**Cauchy Location Model**) The previous corollary applies to the Cauchy location model, where $p_\theta(x) = f(x-\theta)$ and $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$, and $I(\theta) = 1/2$ (Problem 14.9). ■

Example 14.2.4 (**Double Exponential Location Model**) Consider the location model $p_\theta(x) = f(x-\theta)$ where $f(x) = \frac{1}{2}\exp(-|x|)$. Although $f(\cdot)$ is not differentiable at 0, the corollary shows the family is q.m.d. Also, $I(\theta) = 1$ (Problem 14.9). ■

Example 14.2.5 Consider the location model $p_\theta(x) = f(x-\theta)$, where

$$f(x) = C(\beta)\exp\{-|x|^\beta\},$$

where β is a fixed positive constant and $C(\beta)$ is a normalizing constant. By the previous corollary, this family is q.m.d. if $\beta > \frac{1}{2}$. In fact, one can check that

$$\int_{-\infty}^{\infty} \frac{[f'(x)]^2}{f(x)}dx < \infty$$

if and only if $\beta > \frac{1}{2}$ (Problem 14.10). This suggests that q.m.d. fails if $\beta \le \frac{1}{2}$, which is the case; see Rao (1968) or Le Cam and Yang (2000), pp. 188–190. ■

In the k-dimensional case, sufficient conditions for a family to be q.m.d. in terms of "ordinary" differentiation can be obtained by an argument similar to the proof of Theorem 14.2.1. As an example, we state the following (Problem 14.11, or Bickel et al. (1993), Proposition 2.1).

Theorem 14.2.2 *Suppose Ω is an open subset of $\mathbb{R}^k$, and P_θ has density $p_\theta(\cdot)$ with respect to a measure μ. Assume $p_\theta(x)$ is continuously differentiable in θ for μ-almost all x, with gradient vector $\dot{p}_\theta(x)$ (of dimension $1 \times k$). Let*

$$\eta(x, \theta) = \frac{\dot{p}_\theta(x)}{2p_\theta^{1/2}(x)} \tag{14.9}$$

if $p_\theta(x) > 0$ and $\dot{p}_\theta(x)$ exists, and set $\eta(x, \theta) = 0$ otherwise. Assume the Fisher Information matrix $I(\theta)$ exists and is continuous in θ. Then, the family is q.m.d. with derivative $\eta(x, \theta)$.

Example 14.2.6 (**Exponential Families in Natural Form**) Suppose

$$\frac{dP_\theta}{d\mu}(x) = p_\theta(x) = C(\theta) \exp[\langle \theta, T(x) \rangle],$$

where

$$\Omega = \text{int}\{\theta \in \mathbb{R}^k : \int \exp[\langle \theta, T(x) \rangle] \, d\mu(x) < \infty\}$$

and $T(x) = (T_1(x), \ldots, T_k(x))^\top$ is a Borel vector-valued function on the space $\mathcal{X}$ where μ is defined. This family is q.m.d. ■

Example 14.2.7 (**Three-Parameter Lognormal Family**) Suppose P_θ is the distribution of $\gamma + \exp(X)$, where $X \sim N(\mu, \sigma^2)$. Here, $\theta = (\gamma, \mu, \sigma)$, where γ and μ may take on any real-value and σ any positive value. Note the support of the distribution varies with θ. Theorem 14.2.2 yields that this family is q.m.d., even though the likelihood function is unbounded. ■

Example 14.2.8 (**Uniform Family**) Suppose P_θ is the uniform distribution on $[0, \theta]$. This family is *not* q.m.d., which can be seen by the fact that the convergence (14.6) fails for any choice of η. Indeed, for $h > 0$,

$$n \int [p_{\theta_0}^{1/2}(x) - p_{\theta_0 + hn^{-1/2}}^{1/2}(x)]^2 dx \geq n \int_{\theta_0}^{\theta_0 + hn^{-1/2}} \frac{1}{\theta_0 + hn^{-1/2}} dx \to \infty .$$

In fact, it is quite typical that families whose support depends on unknown parameters will not be q.m.d., though Example 14.2.7 is an exception. ■

We are now in a position to obtain an asymptotic expansion of the loglikelihood ratio whose asymptotic form corresponds to that of the normal location model in Example 14.2.1. First, define the *score function* (or *score vector*) $\tilde{\eta}(x, \theta)$ by

$$\tilde{\eta}(x, \theta) = \frac{2\eta(x, \theta)}{p_\theta^{1/2}(x)} \tag{14.10}$$

if $p_\theta(x) > 0$ and $\tilde{\eta}(x, \theta) = 0$ otherwise. Under the conditions of Theorem 14.2.2, $\tilde{\eta}(x, \theta)$ can often be computed as the gradient vector of $\log p_\theta(x)$. Also, define the normalized score vector Z_n by

$$Z_n = Z_{n,\theta_0} = n^{-1/2} \sum_{i=1}^{n} \tilde{\eta}(X_i, \theta_0) \, . \tag{14.11}$$

The following theorem, due to Le Cam, is the main result of this section.

Theorem 14.2.3 *Suppose $\{P_\theta, \theta \in \Omega\}$ is q.m.d. at θ_0 with derivative $\eta(\cdot, \theta_0)$ and Ω is an open subset of $\mathbb{R}^k$. Suppose $I(\theta_0)$ is nonsingular. Fix θ_0 and consider the likelihood ratio $L_{n,h}$ defined by*

$$L_{n,h} = \frac{L_n(\theta_0 + hn^{-1/2})}{L_n(\theta_0)} = \prod_{i=1}^{n} \frac{p_{\theta_0 + hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} \, , \tag{14.12}$$

where the likelihood function $L_n(\cdot)$ is defined in (14.1).

(i) Then, as $n \to \infty$,

$$\log(L_{n,h}) - \left[\langle h, Z_n \rangle - \frac{1}{2} \langle h, I(\theta_0) h \rangle \right] = o_{P^n_{\theta_0}}(1). \tag{14.13}$$

(ii) Under $P^n_{\theta_0}$, $Z_n \xrightarrow{d} N(0, I(\theta_0))$ and so

$$\log(L_{n,h}) \xrightarrow{d} N\left(-\tfrac{1}{2} \langle h, I(\theta_0) h \rangle, \langle h, I(\theta_0) h \rangle \right). \tag{14.14}$$

PROOF. Consider the triangular array $Y_{n,1}, \ldots, Y_{n,n}$, where

$$Y_{n,i} = \frac{p^{1/2}_{\theta_0 + hn^{-1/2}}(X_i)}{p^{1/2}_{\theta_0}(X_i)} - 1.$$

Note that $E_{\theta_0}(Y^2_{n,i}) \le 2 < \infty$ and

$$\log(L_{n,h}) = 2 \sum_{i=1}^{n} \log(1 + Y_{n,i}) \, . \tag{14.15}$$

But,

$$\log(1 + y) = y - \tfrac{1}{2} y^2 + y^2 r(y) \, ,$$

where $r(y) \to 0$ as $y \to 0$, so that

$$\log(L_{n,h}) = 2\sum_{i=1}^{n} Y_{n,i} - \sum_{i=1}^{n} Y_{n,i}^2 + 2\sum_{i=1}^{n} Y_{n,i}^2\, r(Y_{n,i})\ .$$

The idea of expanding the likelihood ratio in terms of variables involving square roots of densities is known as Le Cam's square root trick; see Le Cam (1969). The proof of (*i*) will follow from the following four convergence results:

$$\sum_{i=1}^{n} E_{\theta_0}(Y_{n,i}) \to -\tfrac{1}{8}\langle h, I(\theta_0)h\rangle \tag{14.16}$$

$$\sum_{i=1}^{n}[Y_{n,i} - E_{\theta_0}(Y_{n,i})] - \frac{1}{2}\langle h, Z_n\rangle \overset{P_{\theta_0}^n}{\to} 0 \tag{14.17}$$

$$\sum_{i=1}^{n} Y_{n,i}^2 \overset{P_{\theta_0}^n}{\to} \frac{1}{4}\langle h, I(\theta_0)h\rangle \tag{14.18}$$

$$\sum Y_{n,i}^2\, r(Y_{n,i}) \overset{P_{\theta_0}^n}{\to} 0\ . \tag{14.19}$$

Once these four convergences have been established, part (*ii*) of the theorem follows by the Central Limit Theorem and the facts that

$$E_{\theta_0}[\langle\tilde{\eta}(X_1,\theta_0), h\rangle] = 0 \qquad \text{by Lemma 14.2.1}(i)$$

and

$$\mathrm{Var}_{\theta_0}[\langle\tilde{\eta}(X_1,\theta_0), h\rangle] = \langle h, I(\theta_0)h\rangle \qquad \text{by Lemma 14.2.2.}$$

(a) To show (14.16),

$$\begin{aligned}\sum_{i=1}^{n} E_{\theta_0}(Y_{n,i}) &= n\int\left[\frac{p_{\theta_0+hn^{-1/2}}^{1/2}(x)}{p_{\theta_0}^{1/2}(x)} - 1\right] p_{\theta_0}(x)\,d\mu(x)\\ &= -\tfrac{n}{2}\int\left[p_{\theta_0+hn^{-1/2}}^{1/2}(x) - p_{\theta_0}^{1/2}(x)\right]^2 d\mu(x)\\ &\to -\tfrac{1}{2}\int|\langle\eta(x,\theta_0), h\rangle|^2 d\mu(x)\end{aligned}$$

by (14.6). This last expression is equal to $-\frac{1}{8}\langle h, I(\theta_0)h\rangle$ by Lemma 14.2.2, and (14.16) follows.
(b) To show (14.17), write

$$Y_{n,i} = \frac{1}{2} n^{-1/2} \langle h, \tilde{\eta}(X_i, \theta_0) \rangle + n^{-1/2} \frac{R_n(X_i)}{p_{\theta_0}^{1/2}(X_i)} , \tag{14.20}$$

where $\int R_n^2(x)\, d\mu(x) \to 0$ (by q.m.d.). Hence,

$$\sum_{i=1}^{n} [Y_{n,i} - E_{\theta_0}(Y_{n,i})] = \frac{1}{2} \langle h, Z_n \rangle + h n^{-1/2} \sum_{i=1}^{n} \left[\frac{R_n(X_i)}{p_{\theta_0}^{1/2}(X_i)} - E_{\theta_0} \left(\frac{R_n(X_i)}{p_{\theta_0}^{1/2}(X_i)} \right) \right].$$

The last term, under $P_{\theta_0}^n$, has mean 0 and variance bounded by

$$h^2 E_{\theta_0} \left[\frac{R_n^2(X_i)}{p_{\theta_0}(X_i)} \right] = h^2 \int R_n^2(x)\, d\mu(x) \to 0 .$$

So, (14.17) follows.
(c) To prove (14.18), by the Weak Law of Large Numbers, under θ_0,

$$\tfrac{1}{n} \sum_{i=1}^{n} [\langle h, \tilde{\eta}(X_i, \theta_0) \rangle]^2 \xrightarrow{P} E_{\theta_0} \{ [\langle h, \tilde{\eta}(X_1, \theta_0) \rangle]^2 \} = \langle h, I(\theta_0) h \rangle . \tag{14.21}$$

Now using Equation (14.20), we get

$$\sum_{i=1}^{n} Y_{n,i}^2 = \tfrac{1}{4n} \sum_{i=1}^{n} [\langle h, \tilde{\eta}(X_i, \theta_0) \rangle]^2 + \tfrac{1}{n} \sum_{i=1}^{n} \frac{R_n^2(X_i)}{p_{\theta_0}(X_i)}$$

$$+ \tfrac{1}{n} \sum_{i=1}^{n} [\langle h, \tilde{\eta}(X_i, \theta_0) \rangle] \sum_{j=1}^{n} \frac{R_n(X_j)}{p_{\theta_0}^{1/2}(X_j)} . \tag{14.22}$$

By (14.21), the first term converges in probability under θ_0 to $\frac{1}{4} \langle h, I(\theta_0) h \rangle$. The second term is nonnegative and has expectation under θ_0 equal to

$$\int R_n^2(x) \mu(dx) \to 0 ;$$

hence, the second term goes to 0 in probability under $P_{\theta_0}^n$ by Markov's inequality. The last term goes to 0 in probability under $P_{\theta_0}^n$ by the Cauchy–Schwarz inequality and the convergences of the first two terms. Thus, (14.18) follows. By taking expectations in (14.22), a similar argument shows

$$n E_{\theta_0}(Y_{n,i}^2) = \frac{1}{4} \langle h, I(\theta_0) h \rangle + o(1) \tag{14.23}$$

as $n \to \infty$, which also implies $E_{\theta_0}(Y_{n,i}) \to 0$.
(d) Finally, to prove (14.19), note that

$$\left|\sum_{i=1}^{n} Y_{n,i}^2\, r(Y_{n,i})\right| \le \max_{1\le i\le n} |r(Y_{n,i})| \sum_{i=1}^{n} Y_{n,i}^2 .$$

So, it suffices to show $\max_i |r(Y_{n,i})| \to 0$ in probability under θ_0, which follows if we can show

$$\max_{1\le i\le n} |Y_{n,i}| \overset{P_{\theta_0}^n}{\to} 0 . \tag{14.24}$$

But, $\sum_{i=1}^{n}[Y_{n,i} - E_{\theta_0}(Y_{n,i})]$ is asymptotically normal by (14.17) and the Central Limit Theorem. Hence, Corollary 11.2.2 is applicable with $s_n^2 = O(1)$, which yields the Lindeberg Condition

$$nE_{\theta_0}[|Y_{n,i} - E_{\theta_0}(Y_{n,i})|^2 I\{|Y_{n,i} - E_{\theta_0}(Y_{n,i})| \ge \epsilon\}] \to 0 \tag{14.25}$$

for any $\epsilon > 0$. But then,

$$P_{\theta_0}\{\max_{1\le i\le n} |Y_{n,i} - E_{\theta_0}(Y_{n,i})| > \epsilon\} \le nP_{\theta_0}\{|Y_{n,i} - E_{\theta_0}(Y_{n,i})|^2 > \epsilon^2\} ,$$

which can be bounded by the expression on the left side of (14.25) divided by ϵ^2, and so $\max_{1\le i\le n} |Y_{n,i} - E_{\theta_0}(Y_{n,i})| \to 0$ in probability under θ_0. The result (14.24) follows, since $E_{\theta_0}(Y_{n,i}) \to 0$. ■

Remark 14.2.1 Since the theorem concerns the local behavior of the likelihood ratio near θ_0, it is not entirely necessary to assume Ω is open. However, it is important to assume θ_0 is an interior point; see Problem 14.14.

Remark 14.2.2 The theorem holds if h is replaced by h_n on the left side of each part of the theorem where $h_n \to h$. Under further assumptions, it is plausible that the left side of (14.13) tends to 0 in probability uniformly in h as long as h varies in a compact set; that is, for any $c > 0$, the supremum over h such that $|h| \le c$ of the absolute value of the left side of (14.13) tends to 0 in probability under θ_0; see Problem 15.12.

14.3 Contiguity

Contiguity is an asymptotic form of a probability measure Q being absolutely continuous with respect to another probability measure P. In order to motivate the concept, suppose P and Q are two probability measures on some measurable space $(\mathcal{X}, \mathcal{F})$. Assume that Q is absolutely continuous with respect to P. This means that $E \in \mathcal{F}$ and $P(E) = 0$ implies $Q(E) = 0$.

Suppose $T = T(X)$ is a random vector from $\mathcal{X}$ to $\mathbb{R}^k$, such as an estimator, test statistic, or test function. How can one compute the distribution of T under Q if you know how to compute probabilities or expectations under P? Specifically, suppose it is required to compute $E_Q[f(T)]$, where f is some measurable function from $\mathbb{R}^k$ to $\mathbb{R}$. Let p and q denote the densities of P and Q with respect to a common measure μ. Then, assuming Q is absolutely continuous with respect to P,

$$E_Q[f(T(X))] = \int_{\mathcal{X}} f(T(x))dQ(x) \tag{14.26}$$

$$= \int_{\mathcal{X}} f(T(x))\frac{q(x)}{p(x)}p(x)d\mu(x) = E_P[f(T(X))L(X)]\,, \tag{14.27}$$

where $L(X)$ is the usual likelihood ratio statistic:

$$L(X) = \frac{q(X)}{p(X)}\ . \tag{14.28}$$

Hence, the distribution of $T(X)$ under Q can be computed if the joint distribution of $(T(X), L(X))$ under P is known. Let $F^{T,L}$ denote the joint distribution of $(T(X), L(X))$ under P. Then, by taking f to be the indicator function $f(T(X)) = I_B[T(X)]$ defined to be equal to one if $T(X)$ falls in B and equal to zero otherwise, we obtain:

$$Q\{T(X) \in B\} = \int_{\mathcal{X}} I(T(x) \in B)L(x)p(x)\mu(dx) \tag{14.29}$$

$$= E_P[I(T(X) \in B)L(X)] = \int_{B\times\mathbb{R}} rdF^{T,L}(t,r)\ . \tag{14.30}$$

Thus, under absolute continuity of Q with respect to P, the problem of finding the distribution of $T(X)$ under Q can in principle be obtained from the joint distribution of $T(X)$ and $L(X)$ under P.

More generally, if $f = f(t,r)$ is a function from $\mathbb{R}^k \times \mathbb{R}$ to $\mathbb{R}$,

$$E_Q[f(T(X), L(X))] = \int_{\mathbb{R}^k\times\mathbb{R}} f(t,r)rdF^{T,L}(t,r) \tag{14.31}$$

(Problem 14.18).

Contiguity is an asymptotic version of absolute continuity that permits an analogous asymptotic statement. Consider sequences of pairs of probabilities $\{P_n, Q_n\}$, where P_n and Q_n are probabilities on some measurable space $(\mathcal{X}_n, \mathcal{F}_n)$. Let T_n be some random vector from $\mathcal{X}_n$ to $\mathbb{R}^k$. Suppose the asymptotic distribution of T_n under P_n is easily obtained, but the behavior of T_n under Q_n is also required. For example, if T_n represents a test function for testing P_n versus Q_n, the power of T_n is the

expectation of T_n under Q_n. Contiguity provides a means of performing the required calculation. An example may help fix ideas.

Example 14.3.1 (**The Wilcoxon Signed-Rank Statistic**) Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with common density $f(\cdot)$. Assume that $f(\cdot)$ is symmetric about θ. The problem is to test the null hypothesis that $\theta = 0$ against the alternative hypothesis that $\theta > 0$. Consider the Wilcoxon signed-rank statistic defined by:

$$W_n = W_n(X_1, \ldots, X_n) = n^{-3/2} \sum_{i=1}^{n} R_{i,n}^{+} \text{sign}(X_i) \,, \tag{14.32}$$

where $\text{sign}(X_i)$ is 1 if $X_i \geq 0$ and is -1 otherwise, and $R_{i,n}^{+}$ is the rank of $|X_i|$ among $|X_1|, \ldots, |X_n|$. Note that $W_n = 2n^{-3/2}[V_n - n(n+1)]$, where V_n was previously studied in Example 12.3.6. Under the null hypothesis, the behavior of W_n is fairly easy to obtain. (Alternatively, one can use Example 12.3.6.) If $\theta = 0$, the variables $\text{sign}(X_i)$ are i.i.d., each 1 or -1 with probability 1/2, and are independent of the variables $R_{i,n}^{+}$. Hence, $E_{\theta=0}(W_n) = 0$. Define $\tilde{I}_k$ to be 1 if the kth largest $|X_i|$ corresponds to a positive observation and -1 otherwise. Then, we have

$$Var_{\theta=0}(W_n) = n^{-3} Var(\sum_{k=1}^{n} k\tilde{I}_k) \tag{14.33}$$

$$= n^{-3} \sum_{k=1}^{n} k^2 = n^{-3} \frac{n(n+1)(2n+1)}{6} \to \frac{1}{3} \tag{14.34}$$

as $n \to \infty$. Not surprisingly, $W_n \xrightarrow{d} N(0, \frac{1}{3})$. To see why, note that (Problem 14.19)

$$W_n - n^{-1/2} \sum_{i=1}^{n} U_i \text{sign}(X_i) = o_P(1) \,, \tag{14.35}$$

where $U_i = G(|X_i|)$ and G is the c.d.f. of $|X_i|$. But, under the null hypothesis, U_i and $\text{sign}(X_i)$ are independent. Moreover, the random variables $U_i \text{sign}(X_i)$ are i.i.d., and so the Central Limit Theorem is applicable. Thus, W_n is asymptotically normal with mean 0 and variance 1/3, and this is true whenever the underlying distribution has a symmetric density about 0. Indeed, the exact distribution of W_n is the same for all distributions symmetric about 0. Hence, the test that rejects the null hypothesis if W_n exceeds $3^{-1/2} z_{1-\alpha}$ has limiting level $1 - \alpha$. Of course, for finite n, critical values for W_n can be obtained exactly. Suppose now that we want to approximate the power of this test. The above argument does not generalize to even close alternatives since it heavily uses the fact that the variables are symmetric about zero. Contiguity provides a fairly simple means of attacking this problem, and we will reconsider this example later. ■

We now return to the general setup.

Definition 14.3.1 Let P_n and Q_n be probability distributions on $(\mathcal{X}_n, \mathcal{F}_n)$. The sequence $\{Q_n\}$ is *contiguous* to the sequence $\{P_n\}$ if $P_n(E_n) \to 0$ implies $Q_n(E_n) \to 0$ for every sequence $\{E_n\}$ with $E_n \in \mathcal{F}_n$.

The following equivalent definition is sometimes useful. The sequence $\{Q_n\}$ is contiguous to $\{P_n\}$ if for every sequence of real-valued random variables T_n such that $T_n \to 0$ in P_n-probability we also have $T_n \to 0$ in Q_n-probability.

If $\{Q_n\}$ is contiguous to $\{P_n\}$ and $\{P_n\}$ is contiguous to $\{Q_n\}$, then we say the sequences $\{P_n\}$ and $\{Q_n\}$ are *mutually contiguous*, or just contiguous.

Example 14.3.2 Suppose P_n is the standard normal distribution $N(0, 1)$ and Q_n is $N(\xi_n, 1)$. Unless ξ_n is bounded, P_n and Q_n cannot be contiguous. Indeed, suppose $\xi_n \to \infty$ and consider $E_n = \{x : |x - \xi_n| < 1\}$. Then, $Q_n(E_n) \approx 0.68$ for all n, but $P_n(E_n) \to 0$. Note that, regardless of the values of ξ_n, P_n and Q_n are mutually absolutely continuous for every n. ■

Example 14.3.3 Suppose P_n is the joint distribution of n i.i.d. observations $X_1, \ldots, X_n$ from $N(0, 1)$ and Q_n is the joint distribution of n i.i.d. observations from $N(\xi_n, 1)$. Unless $\xi_n \to 0$, P_n and Q_n cannot be contiguous. For example, suppose $\xi_n > \epsilon > 0$ for all large n. Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and consider $E_n = \{\bar{X}_n > \epsilon/2\}$. By the law of large numbers, $P_n(E_n) \to 0$ but $Q_n(E_n) \to 1$. As will be seen shortly, in order for P_n and Q_n to be contiguous, it will be necessary and sufficient for $\xi_n \to 0$ in such a way so that $n^{1/2}\xi_n$ remains bounded. ■

We now would like a useful means of determining whether or not Q_n is contiguous to P_n. Suppose P_n and Q_n have densities p_n and q_n with respect to μ_n. For $x \in \mathcal{X}_n$, define the *likelihood ratio* of Q_n with respect to P_n by

$$L_n(x) = \begin{cases} \frac{q_n(x)}{p_n(x)} & \text{if} p_n(x) > 0 \\ \infty & \text{if } p_n(x) = 0 < q_n(x) \\ 1 & \text{if } p_n(x) = q_n(x) = 0. \end{cases} \tag{14.36}$$

Under P_n or Q_n, the event $\{p_n = q_n = 0\}$ has probability 0, so it really doesn't matter how L_n is defined in this case (as long as it is measurable). Note that L_n is regarded as an extended random variable, which means it is allowed to take on the value ∞, at least under Q_n. Of course, under P_n, L_n is finite with probability one.

Observe that

$$E_{P_n}(L_n) = \int_{\mathcal{X}_n} L_n(x) p_n(x) \mu_n(dx) = \int_{\{x:\ p_n(x)>0\}} q_n(x) \mu_n(dx)$$

$$= Q_n\{x :\ p_n(x) > 0\} = 1 - Q_n\{x :\ p_n(x) = 0\} \le 1\,, \tag{14.37}$$

with equality if and only if Q_n is absolutely continuous with respect to P_n.

Example 14.3.4 (**Contiguous but not absolutely continuous sequence**) Suppose P_n is uniformly distributed on $[0, 1]$ and Q_n is uniformly distributed on $[0, \theta_n]$, where $\theta_n > 1$. Then, Q_n is not absolutely continuous with respect to P_n. Note that the likelihood ratio L_n is equal to $1/\theta_n$ with probability one under P_n, and so

$$E_{P_n}(L_n) = \frac{1}{\theta_n} < 1 \, .$$

It will follow from Theorem 14.3.1 that Q_n is contiguous to P_n if $\theta_n \to 1$. ■

The notation $\mathcal{L}(T|P)$ refers to the distribution of a random variable (or possibly an extended random variable) $T = T(X)$ when X is governed by P. Let $G_n = \mathcal{L}(L_n|P_n)$, the distribution of the likelihood ratio under P_n. Note that G_n is a tight sequence, because by Markov's inequality,

$$P_n\{L_n > c\} \le \frac{E_{P_n}(L_n)}{c} \le \frac{1}{c} \, , \tag{14.38}$$

where the last inequality follows from (14.37).

The statement that $E_{P_n}(L_n) = 1$ implies that Q_n is absolutely continuous with respect to P_n, by (14.37). The following result, known as Le Cam's First Lemma, may be regarded as an asymptotic version of this statement.

Theorem 14.3.1 *Given P_n and Q_n, consider the likelihood ratio L_n defined in (14.36). Let G_n denote the distribution of L_n under P_n. Suppose G_n converges weakly to a distribution G. If G has mean 1, then Q_n is contiguous to P_n.*

PROOF. Suppose $P_n(E_n) = \alpha_n \to 0$. Let ϕ_n be a most powerful level α_n test of P_n versus Q_n. By the Neyman–Pearson Lemma, the test is of the form

$$\phi_n = \begin{cases} 1 & \text{if } L_n > k_n \\ 0 & \text{if } L_n < k_n, \end{cases} \tag{14.39}$$

for some k_n chosen so the test is level α_n. Since ϕ_n is at least as powerful as the test that has rejection region E_n,

$$Q_n\{E_n\} \le \int \phi_n d Q_n \, ,$$

so it suffices to show the right side tends to zero. Now, for any $y < \infty$,

$$\int \phi_n d Q_n = \int_{L_n \le y} \phi_n d Q_n + \int_{L_n > y} \phi_n d Q_n$$

$$\le y \int \phi_n d P_n + \int_{L_n > y} d Q_n \le y \int \phi_n d P_n + 1 - \int_{L_n \le y} d Q_n$$

$$= y\alpha_n + 1 - \int_{L_n \le y} L_n dP_n = y\alpha_n + 1 - \int_0^y x dG_n(x) \, .$$

Fix any $\epsilon > 0$ and take y to be a continuity point of G with

$$\int_0^y x dG(x) > 1 - \frac{\epsilon}{2} \, ,$$

which is possible since G has mean 1. But G_n converges weakly to G implies

$$\int_0^y x dG_n(x) \to \int_0^y x dG(x) \, , \tag{14.40}$$

by an argument like that in Example 11.4.4 (Problem 14.27). Thus, for sufficiently large n,

$$1 - \int_0^y x dG_n(x) < \frac{\epsilon}{2}$$

and $y\alpha_n < \epsilon/2$. It follows that, for sufficiently large n,

$$\int \phi_n dQ_n < \epsilon \, ,$$

as was to be proved. ■

The following result summarizes some equivalent characterizations of contiguity. The notation $\mathcal{L}(T|P)$ refers to the distribution (or law) of a random variable T under P.

Theorem 14.3.2 *The following are equivalent characterizations of* $\{Q_n\}$ *being contiguous to* $\{P_n\}$.

(i) For every sequence of real-valued random variables T_n *such that* $T_n \to 0$ *in* P_n*-probability, it also follows that* $T_n \to 0$ *in* Q_n*-probability.*
(ii) For every sequence T_n *such that* $\mathcal{L}(T_n|P_n)$ *is tight, it also follows that* $\mathcal{L}(T_n|Q_n)$ *is tight.*
(iii) If G *is any limit point* [3] *of* $\mathcal{L}(L_n|P_n)$*, then* G *has mean 1.*

PROOF. First, we show that (ii) implies (i). Suppose $T_n \to 0$ in P_n-probability; that is, $P_n\{|T_n| > \delta\} \to 0$ for every $\delta > 0$. Then, there exists $\epsilon_n \downarrow 0$ such that $P_n\{|T_n| > \epsilon_n\} \to 0$. So, $|T_n|/\epsilon_n$ is tight under $\{P_n\}$. By hypothesis, $|T_n|/\epsilon_n$ is also tight under $\{Q_n\}$. Assume the conclusion that $T_n \to 0$ in Q_n-probability fails; then, one could find $\epsilon > 0$ such that $Q_n\{|T_n| > \epsilon\} > \epsilon$ for infinitely many n. Then, of course, $Q_n\{|T_n| > \sqrt{\epsilon_n}\} > \epsilon$ for infinitely many n. Since $1/\sqrt{\epsilon_n} \uparrow \infty$, it follows that $|T_n|/\epsilon_n$ cannot be tight under $\{Q_n\}$, which is a contradiction.

[3] G is a limit point of a sequence G_n of distributions if G_{n_j} converges in distribution to G for some subsequence n_j.

Conversely, to show that (i) implies (ii), assume that $\mathcal{L}(T_n|P_n)$ is tight. Then, given $\epsilon > 0$, there exists k such that $P_n\{|T_n| > k\} < \epsilon/2$ for all n. If $\mathcal{L}(T_n|Q_n)$ is not tight, then for every j, $Q_n\{|T_n| > j\} > \epsilon$ for some n. That is, there exists a subsequence n_j such that $Q_{n_j}\{|T_{n_j}| > j\} > \epsilon$ for every j. As soon as $j > k$,

$$P_{n_j}\{|T_{n_j}| > j\} \leq P_{n_j}\{|T_{n_j}| > k\} < \frac{\epsilon}{2} ,$$

a contradiction.

To show (iii) implies (i), first recall (14.38), which implies G_n is tight. Assuming $P_n\{A_n\} \to 0$, we must show $Q_n\{A_n\} \to 0$. Assume that this is not the case. Then, there exists a subsequence n_j and $\epsilon > 0$ such that $Q_{n_j}\{A_{n_j}\} \geq \epsilon$ for all n_j. But, there exists a further subsequence n_{j_k} such that $G_{n_{j_k}}$ converges to some G. Assuming (iii), G has mean 1. By Theorem 14.3.1, $P_{n_{j_k}}$ and $Q_{n_{j_k}}$ are contiguous. Since $Q_{n_{j_k}}\{A_{n_{j_k}}\} \to 0$, this is a contradiction.

Conversely, suppose (i) and that G_n converges weakly to G (or apply the following argument to any convergent subsequence). By Example 11.4.4, it follows that

$$\int x dG(x) \leq \liminf_n E_{P_n}(L_n) \leq 1 ,$$

so it suffices to show $\int x dG(x) \geq 1$. Let t be a continuity point of G. Then, also by Example 11.4.4 (specifically (11.39)),

$$\int x dG(x) \geq \int_{\{x \leq t\}} x dG(x) = \lim_n E_{P_n}(L_n 1\{L_n \leq t\}) = \lim_n Q_n\{L_n \leq t\} .$$

So, it suffices to show that, given any $\epsilon > 0$, there exists a t such that $Q_n\{L_n > t\} < \epsilon$ for all large n. If this fails, then for every j, there exists n_j such that $Q_{n_j}\{L_{n_j} > j\} > \epsilon$. But, by (14.38),

$$P_{n_j}\{L_{n_j} > j\} \leq \frac{1}{j} \to 0$$

as $j \to \infty$, which would contradict (i). ■

As will be seen in many important examples, loglikelihood ratios are typically asymptotically normally distributed, and the following corollary is useful.

Corollary 14.3.1 *Consider a sequence $\{P_n, Q_n\}$ with likelihood ratio L_n defined in (14.36). Assume*

$$\mathcal{L}(L_n|P_n) \xrightarrow{d} \mathcal{L}(e^Z) , \tag{14.41}$$

where Z is distributed as $N(\mu, \sigma^2)$. Then, Q_n and P_n are mutually contiguous if and only if $\mu = -\sigma^2/2$.

PROOF. To show Q_n is contiguous to P_n, apply part (iii) of Theorem 14.3.2 by showing $E(e^Z) = 1$. But, recalling the characteristic function of Z from equation (11.10), it follows that

$$E(e^Z) = \exp(\mu + \frac{1}{2}\sigma^2) ,$$

which equals 1 if and only if $\mu = -\sigma^2/2$. That P_n is contiguous to Q_n follows by Problem 14.23. ■

We may write (14.41) equivalently as

$$\mathcal{L}(\log(L_n)|P_n) \xrightarrow{d} \mathcal{L}(Z) .$$

However, since $P_n\{L_n = 0\}$ may be positive, we may have $\log(L_n) = -\infty$ with positive probability, in which case $\log(L_n)$ is regarded as an extended real-valued random variable taking values in $\mathbb{R} \bigcup \{\pm\infty\}$. If X_n is an extended real-valued random variable and X is a real-valued random variable with c.d.f. F, we say (as in Definition 11.2.1) X_n converges in distribution to X if

$$P_n\{X_n \in (-\infty, t]\} \to F(t)$$

whenever t is a continuity point of F. It follows that if X_n converges in distribution to a random variable that is finite (with probability one), then the probability that X_n is finite must tend to 1.

Example 14.3.5 (**Example** 14.3.2, **continued**). Again, suppose that $P_n = N(0, 1)$ and $Q_n = N(\xi_n, 1)$. In this case,

$$L_n = L_n(X) = \exp(\xi_n X - \frac{1}{2}\xi_n^2) .$$

Thus,

$$\mathcal{L}(\log(L_n)|P_n) = N\left(-\frac{\xi_n^2}{2}, \xi_n^2\right) .$$

Such a sequence of distributions will converge weakly along a subsequence n_j if and only if $\xi_{n_j} \to \xi$ (for some $|\xi| < \infty$), in which case, the limiting distribution is $N(\frac{-\xi^2}{2}, \xi^2)$ and the relationship between the mean and the variance ($\mu = -\sigma^2/2$) is satisfied. Hence, Q_n is contiguous to P_n if and only if ξ_n is bounded. In fact, Q_n and P_n are mutually contiguous under the same condition. ■

Example 14.3.6 (**Example** 14.3.3, **continued**). Suppose $X_1, \ldots, X_n$ are i.i.d. with common distribution $N(\xi, 1)$. Let P_n represent the joint distribution when $\xi = 0$ and let Q_n represent the joint distribution when $\xi = \xi_n$. Then,

$$\log(L_n(X_1, \ldots, X_n)) = \xi_n \sum_{i=1}^{n} X_i - \frac{n\xi_n^2}{2}, \tag{14.42}$$

and so

$$\mathcal{L}(\log(L_n)|P_n) = N\left(-\frac{n\xi_n^2}{2}, n\xi_n^2\right).$$

By an argument similar to that of the previous example, Q_n is contiguous to P_n if and only if $n\xi_n^2$ remains bounded, i.e., $\xi_n = O(n^{-1/2})$; P_n and Q_n are mutually contiguous if and only if the same condition holds. Note that, even if $\xi_n \to 0$, but at a rate slower than $n^{-1/2}$, Q_n is not contiguous to P_n. This is related to the assertion that the problem of testing P_n versus Q_n is degenerate unless $\xi_n \asymp n^{-1/2}$, in the sense that the most powerful level α test ϕ_n has asymptotic power satisfying $E_{\xi_n}(\phi_n) \to 1$ if $n^{1/2}|\xi_n| \to \infty$ and $E_{\xi_n}(\phi_n) \to \alpha$ if $n^{1/2}\xi_n \to 0$.[4] Indeed, suppose without loss of generality that $\xi_n > 0$. Then, the most powerful level α test rejects when $n^{1/2}\bar{X}_n > z_{1-\alpha}$, where $\bar{X}_n = \sum_{i=1} X_i/n$ and $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of the standard normal distribution. The power of ϕ_n against ξ_n is then

$$P_{\xi_n}\{n^{1/2}\bar{X}_n > z_{1-\alpha}\} == P\{Z > z_{1-\alpha} - n^{1/2}\xi_n\},$$

where Z is a standard normal variable. Clearly, the last expression tends to 1 if and only if $n^{1/2}\xi_n \to \infty$; furthermore, it tends to α if and only if $n^{1/2}\xi_n \to 0$. The limiting power is bounded away from α and 1 if and only if $\xi_n \asymp n^{-1/2}$. ■

Example 14.3.7 (**Q.m.d. families**) Let $\{P_\theta,\ \theta \in \Omega\}$ with Ω an open subset of $\mathbb{R}^k$ be q.m.d., with corresponding densities $p_\theta(\cdot)$. By Theorem 14.2.3, under θ_0,

$$\log\left(\frac{dP^n_{\theta_0+hn^{-1/2}}}{dP^n_{\theta_0}}\right) = n^{-1/2}\sum_{i=1}^{n}\langle h, \tilde{\eta}(X_i, \theta_0)\rangle - \frac{1}{2}\langle h, I(\theta_0)h\rangle + o_{P^n_{\theta_0}}(1), \tag{14.43}$$

where $\tilde{\eta}(x, \theta) = 2\eta(x, \theta)/p_\theta^{1/2}(x)$, $\eta(\cdot, \theta)$ is the quadratic mean derivative at θ, and $I(\theta)$ is the Information matrix at θ. Hence, by Corollary 14.3.1, $P^n_{\theta_0+hn^{-1/2}}$ and $P^n_{\theta_0}$ are mutually contiguous. ■

Suppose Q_n is contiguous to P_n. As before, let L_n be the likelihood ratio defined by (14.28). Let T_n be an arbitrary sequence of real-valued statistics. The following theorem allows us to determine the asymptotic behavior of (T_n, L_n) under Q_n from the behavior of (T_n, L_n) under P_n.

Theorem 14.3.3 *Suppose Q_n is contiguous to P_n. Let T_n be a sequence of real-valued random variables. Suppose, under P_n, (T_n, L_n) converges in distribution to*

[4] Two real-valued sequences $\{a_n\}$ and $\{b_n\}$ are said to be of the same order, written $a_n \asymp b_n$ if $|a_n/b_n|$ is bounded away from 0 and ∞.

a limit law $F(\cdot,\cdot)$; that is, for any bounded continuous function f on $(-\infty,\infty)\times[0,\infty)$,

$$E_{P_n}[f(T_n,L_n)]\to\int\int f(t,r)dF(t,r)\ . \tag{14.44}$$

Then, the limiting distribution of (T_n,L_n) under Q_n has density $rdF(t,r)$; that is,

$$E_{Q_n}[f(T_n,L_n)]\to\int\int f(t,r)rdF(t,r) \tag{14.45}$$

for any bounded continuous f. Equivalently, if under P_n $(T_n,\log(L_n))$ converges weakly to a limit law $\bar{F}(\cdot,\cdot)$, then

$$E_{Q_n}[f(T_n,\log(L_n))]\to\int\int f(t,r)e^r d\bar{F}(t,r) \tag{14.46}$$

for any bounded continuous f.

Note that equation (14.45) is simply an asymptotic version of (14.31).

Remark 14.3.1 The result is also true if T_n is vector-valued, and the proof is the same.

PROOF. Let $F_n=\mathcal{L}((T_n,L_n)|P_n)$ and $G_n=\mathcal{L}((T_n,L_n)|Q_n)$. Since L_n converges in distribution under P_n, contiguity and Theorem 14.3.2 (iii) imply that

$$\int rdF(t,r)=1\ .$$

Thus, $rdF(t,r)$ defines a probability distribution on $(-\infty,\infty)\times[0,\infty)$.

Let f be a nonnegative, continuous function on $(-\infty,\infty)\times[0,\infty]$. By the Portmanteau Theorem (11.2.1 (vi)), it suffices to show that

$$\liminf_n\int f(t,r)dG_n(t,r)\ge\int f(t,r)rdF(t,r)\ .$$

Note that

$$\int f(t,r)dG_n(t,r)=E_{Q_n}[f(T_n,L_n)]=\int f(T_n,L_n)dQ_n$$

$$\ge\int_{\{p_n>0\}}f(T_n,L_n)dQ_n=\int f(T_n,L_n)L_ndP_n=\int f(t,r)rdF_n(t,r)\ .$$

So, it suffices to show

$$\liminf_n\int f(t,r)rdF_n(t,r)\ge\int rf(t,r)dF(t,r)\ .$$

But, $rf(t,r)$ is a nonnegative, continuous function, and so the result follows again by the Portmanteau Theorem. ■

The following special case is often referred to as Le Cam's Third Lemma.

Corollary 14.3.2 *Assume that, under P_n, $(T_n, \log(L_n)) \xrightarrow{d} (T,Z)$, where (T,Z) is bivariate normal with $E(T)=\mu_1$, $Var(T)=\sigma_1^2$, $E(Z)=\mu_2$, $Var(Z)=\sigma_2^2$, and $Cov(T,Z)=\sigma_{1,2}$. Assume $\mu_2=-\sigma_2^2/2$, so that Q_n is contiguous to P_n. Then, under Q_n, T_n is asymptotically normal:*

$$\mathcal{L}(T_n|Q_n) \xrightarrow{d} N(\mu_1+\sigma_{1,2},\sigma_1^2) .$$

PROOF. Let $\bar{F}(\cdot,\cdot)$ denote the bivariate normal distribution of (T,Z). By Theorem 14.3.3, the limiting distribution of $\mathcal{L}((T_n,\log(L_n))|Q_n)$ has density $e^r d\bar{F}(x,r)$; let $(\tilde{T},\tilde{Z})$ denote a random variable having this distribution. The characteristic function of $\tilde{T}$ is given by:

$$E(e^{i\lambda\tilde{T}}) = \int e^{i\lambda x} e^r d\bar{F}(x,r) = E(e^{i\lambda T+Z}) , \tag{14.47}$$

which is the characteristic function of (T,Z) evaluated at $t=(t_1,t_2)^\top=(\lambda,-i)^\top$. By Example 11.2.1, this is given by

$$\exp(i\langle\mu,t\rangle - \frac{1}{2}\langle\Sigma t,t\rangle) = \exp(i\mu_1\lambda+\mu_2-\frac{1}{2}\langle\Sigma(\lambda,-i)^\top,(\lambda,-i)^\top\rangle)$$

$$= \exp(i\mu_1\lambda+\mu_2-\frac{1}{2}\lambda^2\sigma_1^2+\lambda i\sigma_{1,2}+\frac{\sigma_2^2}{2}) = \exp[i(\mu_1+\sigma_{1,2})\lambda-\frac{1}{2}\lambda^2\sigma_1^2] ,$$

the last equality following from the fact that $\mu_2=-\sigma_2^2/2$ (by contiguity). But, this last expression is indeed the characteristic function of the normal distribution with mean $\mu_1+\sigma_{1,2}$ and variance σ_1^2. ■

Example 14.3.8 (**Asymptotically Linear Statistic**) Let $\{P_\theta,\ \theta\in\Omega\}$ with Ω an open subset of $\mathbb{R}^k$ be q.m.d., with corresponding densities $p_\theta(\cdot)$. Recall Example 14.3.7, which shows that $P^n_{\theta_0+hn^{-1/2}}$ and $P^n_{\theta_0}$ are mutually contiguous. The expansion (14.43) shows a lot more. For example, suppose an estimator (sequence) $\hat{\theta}_n$ is asymptotically linear in the following sense: under θ_0,

$$n^{1/2}(\hat{\theta}_n-\theta_0) = n^{-1/2}\sum_{i=1}^n \psi_{\theta_0}(X_i) + o_{P^n_{\theta_0}}(1) , \tag{14.48}$$

where $E_{\theta_0}[\psi_{\theta_0}(X_1)]=0$ and $\tau^2\equiv Var_{\theta_0}[\psi_{\theta_0}(X_1)]<\infty$. Thus, under θ_0,

$$n^{1/2}(\hat{\theta}_n-\theta_0) \xrightarrow{d} N(0,\tau^2) .$$

Then, the joint behavior of $\hat{\theta}_n$ with the loglikelihood ratio satisfies

$$(n^{1/2}(\hat{\theta}_n - \theta_0), \log(\frac{dP^n_{\theta_0+hn^{-1/2}}}{dP^n_{\theta_0}})) \tag{14.49}$$

$$= [n^{-1/2}\sum_{i=1}^{n}(\psi_{\theta_0}(X_i), \langle h, \tilde{\eta}(X_i, \theta_0)\rangle)] + (0, -\frac{1}{2}\langle h, I(\theta_0)h\rangle) + o_{P^n_{\theta_0}}(1) .$$

By the bivariate Central Limit Theorem, this converges under θ_0 to a bivariate normal distribution with covariance

$$\sigma_{1,2} \equiv Cov_{\theta_0}(\psi_{\theta_0}(X_1), \langle h, \tilde{\eta}(X_i, \theta_0)\rangle) . \tag{14.50}$$

Hence, under $P^n_{\theta_0+hn^{-1/2}}$, $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution to $N(\sigma_{1,2}, \tau^2)$, by Corollary 14.3.2. It follows that, under $P^n_{\theta_0+hn^{-1/2}}$,

$$n^{1/2}(\hat{\theta}_n - (\theta_0 + hn^{-1/2})) \xrightarrow{d} N(\sigma_{1,2} - h, \tau^2) . \blacksquare$$

Example 14.3.9 (**t-statistic**) Consider a location model $f(x - \theta)$ for which $f(x)$ has mean 0 and variance σ^2, and which satisfies the assumptions of Corollary 14.2.1, which imply this family is q.m.d. For testing $\theta = \theta_0 = 0$, consider the behavior of the usual t-statistic

$$t_n = \frac{n^{1/2}\bar{X}_n}{S_n} = \frac{n^{1/2}\bar{X}_n}{\sigma} + o_{P_{\theta_0}}(1) .$$

Then, (14.48) holds with $\psi_{\theta_0}(X_i) = X_i/\sigma$. We seek the behavior of t_n under $\theta_n = h/n^{1/2}$. Although this can be obtained by direct means, let us obtain the results by contiguity. Note that (14.43) holds with

$$\tilde{\eta}(X_i, \theta_0) = -\frac{f'(x)}{f(x)} .$$

Thus, $\sigma_{1,2}$ in (14.50) reduces to

$$\sigma_{1,2} = -\frac{h}{\sigma}Cov_{\theta_0=0}\left(X_i, \frac{f'(X_i)}{f(X_i)}\right) = -\frac{h}{\sigma}\int_{-\infty}^{\infty} xf'(x)dx = \frac{h}{\sigma} .$$

Hence, under $\theta_n = h/n^{1/2}$,

$$t_n \xrightarrow{d} N(\frac{h}{\sigma}, 1) . \blacksquare$$

Example 14.3.10 (**Sign Test**) As in the previous example, consider a location model $f(x-\theta)$, where f is a density with respect to Lebesgue measure. Assume the conditions in Corollary 14.2.1, so that the family is q.m.d. Further suppose that $f(x)$ is continuous at $x=0$ and $P_{\theta=0}\{X_i>0\}=1/2$. For testing $\theta=\theta_0=0$, consider the (normalized) sign statistic

$$S_n = n^{-1/2}\sum_{i=1}^{n}[I\{X_i>0\}-\frac{1}{2}]\,,$$

where $I\{X_i>0\}$ is one if $X_i>0$ and is 0 otherwise. Then, (14.48) holds with $\psi_0(X_i)=I\{X_i>0\}-\frac{1}{2}$ and so

$$S_n \xrightarrow{d} N(0,\frac{1}{4})\,.$$

Under $\theta_n=h/n^{1/2}$, $S_n \xrightarrow{d} N(\sigma_{1,2},1/4)$, where $\sigma_{1,2}$ is given by (14.50) and equals

$$\sigma_{1,2} = -hCov_0\left[I\{X_i>0\},\frac{f'(X_i)}{f(X_i)}\right] = -h\int_0^\infty f'(x)dx = hf(0)\,.$$

Hence, under $\theta_n=h/n^{1/2}$,

$$S_n \xrightarrow{d} N(hf(0),\frac{1}{4})\,.\ \blacksquare$$

Example 14.3.11 (**Example** 14.3.1, **continued**). Recall the Wilcoxon signed-rank statistic W_n given by (14.32). For illustration, suppose the underlying density $f(\cdot)$ of the observations is normal with mean θ and variance 1. Under the null hypothesis $\theta=0$, W_n is asymptotically normal $N(0,\frac{1}{3})$. The problem now is to compute the asymptotic power against the sequence of alternatives $\theta_n=h/n^{1/2}$ for some $h>0$. Under the null hypothesis, by (14.35) and (14.42),

$$(W_n,\log(L_n)) = (n^{-1/2}\sum_{i=1}^{n}U_i\text{sign}(X_i),\, hn^{-1/2}\sum_{i=1}^{n}X_i-\frac{h^2}{2}) + o_{P_0^n}(1)\,, \quad (14.51)$$

where $U_i=G(|X_i|)$ and G is the c.d.f. of $|X_i|$. This last expression is asymptotically bivariate normal with covariance under $\theta=0$ equal to

$$\sigma_{1,2} = hCov_0[G(|X_1|)\text{sign}(X_1),X_1] = hE_0[G(|X_1|)|X_1|]\,, \quad (14.52)$$

and thus $\sigma_{1,2}$ is equal to $h/\sqrt{\pi}$ (Problem 14.28). Hence, under $\theta_n=h/n^{1/2}$, W_n is asymptotically normal with mean $h/\sqrt{\pi}$ and variance 1/3. Thus, the asymptotic power of the test that rejects when $W_n>3^{-1/2}z_{1-\alpha}$ is

$$\lim_{n\to\infty} P_{\theta_n}\{W_n - \frac{h}{\sqrt{\pi}} > 3^{-1/2}z_{1-\alpha} - \frac{h}{\sqrt{\pi}}\} = 1 - \Phi(z_{1-\alpha} - (3/\pi)^{1/2}h) \ ,$$

where $\Phi(\cdot)$ is the standard normal c.d.f.

More generally, assume the underlying model is a location model $f(x-\theta)$, where $f(x)$ is assumed symmetric about zero. Assume $f'(x)$ exists for Lebesgue almost all x and

$$0 < I \equiv \int \frac{[f'(x)]^2}{f(x)} dx < \infty \ .$$

Then, by Corollary 14.2.1, this model is q.m.d. and (14.43) holds with

$$\tilde{\eta}(x, 0) = -\frac{f'(x)}{f(x)} \ .$$

Under the null hypothesis $\theta = 0$, $W_n \xrightarrow{d} N(0, 1/3)$, as in the normal case. Under the sequence of alternatives $\theta_n = h/n^{1/2}$,

$$W_n \xrightarrow{d} N(\sigma_{1,2}, \frac{1}{3}) \ ,$$

where $\sigma_{1,2}$ is given by (14.50). In this case,

$$\sigma_{1,2} = Cov_{\theta=0}[U\text{sign}(X), -h\frac{f'(X)}{f(X)}] \ ,$$

where $U = G(|X|)$ and G is the c.d.f. of $|X|$ when X has density $f(\cdot)$. So, $G(x) = 2F(x) - 1$, where F is the c.d.f. of X. By an integration by parts (see Problem 14.29),

$$\sigma_{1,2} = -hE_{\theta=0}[G(|X|)\text{sign}(X)\frac{f'(X)}{f(X)}] = 2h\int_{-\infty}^{\infty} f^2(x)dx \ . \tag{14.53}$$

Thus, under $\theta_n = h/n^{1/2}$,

$$W_n \xrightarrow{d} N(2h\int_{-\infty}^{\infty} f^2(x)dx, \frac{1}{3}) \ .$$

An alternative approach that uses the projection of U-statistics is given in Problem 14.30. ■

Example 14.3.12 (**Neyman–Pearson Statistic**) Assume $\{P_\theta, \ \theta \in \Omega\}$ is q.m.d. at θ_0, where Ω is an open subset of $\mathbb{R}^k$ and $I(\theta_0)$ is nonsingular, so that the assumptions behind Theorem 14.2.3 are in force. Let $p_\theta(\cdot)$ be the corresponding density of P_θ. Consider the likelihood ratio statistic based on n i.i.d. observations $X_1, \ldots, X_n$ given by

$$L_{n,h} = \frac{dP^n_{\theta_0+hn^{-1/2}}}{dP^n_{\theta_0}} = \prod_{i=1}^{n} \frac{p_{\theta_0+hn^{-1/2}}(X_i)}{p_{\theta_0}(X_i)} \,. \tag{14.54}$$

By Theorem 14.2.3, under P_{θ_0},

$$\log(L_{n,h}) \xrightarrow{d} N(-\frac{\sigma_h^2}{2}, \sigma_h^2) \,, \tag{14.55}$$

where $\sigma_h^2 = \langle h, I(\theta_0)h\rangle$. Apply Corollary 14.3.2 with $T_n \equiv \log(L_{n,h})$, so that $T = Z$ and $\sigma_{1,2} = \sigma_h^2$. Then, under $P^n_{\theta_0+hn^{-1/2}}$, $\log(L_{n,h})$ is asymptotically $N(\frac{\sigma_h^2}{2}, \sigma_h^2)$. Hence, the test that rejects when $\log(L_{n,h})$ exceeds $-\frac{1}{2}\sigma_h^2 + z_{1-\alpha}\sigma_h$ is asymptotically level α for testing $\theta = \theta_0$ versus $\theta = \theta_0 + hn^{-1/2}$, where $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of $N(0,1)$. Then, the limiting power of this test sequence for testing $\theta = \theta_0$ versus $\theta = \theta_0 + hn^{-1/2}$ is $1 - \Phi(z_{1-\alpha} - \sigma_h)$ (Problem 14.31). ■

14.4 Likelihood Methods in Parametric Models

The goal of this section is to study some classical large-sample methods based on the likelihood function. The classical likelihood ratio test, as well as the tests of Wald and Rao will be introduced, but optimality of these tests will be deferred until the next chapter. Throughout this section, we will assume that $X_1, \ldots, X_n$ are i.i.d. with common distribution P_θ, where $\theta \in \Omega$ and Ω is an open subset of $\mathbb{R}^k$. We will also assume each P_θ is absolutely continuous with respect to a common σ-finite measure μ, so that p_θ denotes the density of P_θ with respect to μ. The *likelihood function* is defined by

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i) \,. \tag{14.56}$$

It is thus the (joint) probability density of the observations at fixed values of $X_1, \ldots, X_n$, viewed as a function of θ. Note that, for the sake of simplicity, the dependence of $L_n(\theta)$ on $X_1, \ldots, X_n$ has been suppressed. (In the case that $X_1, \ldots, X_n$ are not i.i.d., $L_n(\theta)$ is modified so that the joint density of the X_i's is used rather than the product of the marginal densities.)

14.4.1 Efficient Likelihood Estimation

In preparation for the construction of reasonable large-sample tests and confidence regions, we begin by studying some efficient point estimators of θ which will serve as a basis for such tests. If the likelihood $L_n(\theta)$ has a unique maximum $\hat{\theta}_n$, then

$\hat{\theta}_n$ is called the *maximum likelihood estimator* (MLE) of θ. If, in addition, $L_n(\theta)$ is differentiable in θ, $\hat{\theta}_n$ will be a solution of the *likelihood equations*

$$\frac{\partial}{\partial \theta_j} \log L_n(\theta) = 0 \quad j = 1, \ldots, k \, .$$

Example 14.4.1 (**Normal Family**) Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from $N(\mu, \sigma^2)$, with both parameters unknown, so $\theta = (\mu, \sigma^2)^\top$. In this case, the loglikelihood function is

$$\log L_n(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \, ,$$

and the likelihood equations reduce to

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0$$

and

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - \mu)^2 = 0 \, .$$

These equations have a unique solution, given by the maximum likelihood estimator $(\hat{\mu}_n, \hat{\sigma}_n^2)$, where $\hat{\mu}_n = \bar{X}_n$ is the usual sample mean and $\hat{\sigma}_n^2$ is the biased version of the sample variance given by

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

(Problem 14.36). By the weak law of large numbers, $\bar{X}_n \to \mu$ in probability; by Example 11.3.3, $\hat{\sigma}_n^2 \to \sigma^2$ in probability as well. A direct argument easily establishes the joint limiting distribution of the MLE. First note that

$$n^{1/2}[\hat{\sigma}_n^2 - n^{-1} \sum_{i=1}^{n} (X_i - \mu)^2] = n^{1/2} (\bar{X}_n - \mu)^2 \xrightarrow{P} 0$$

since $n^{1/2}(\bar{X}_n - \mu)$ is $N(0, \sigma^2)$ and $\bar{X}_n - \mu \xrightarrow{P} 0$. Hence, by Slutsky's Theorem, $n^{1/2}((\bar{X}_n, \hat{\sigma}_n^2)^\top - (\mu, \sigma^2)^\top)$ has the same limiting distribution as

$$n^{1/2}[(\bar{X}_n, n^{-1} \sum_{i=1}^{n} (X_i - \mu)^2)^\top - (\mu, \sigma^2)^\top] \, ,$$

which by the multivariate CLT tends in distribution to $N(0, \Sigma)$, where Σ is the 2×2 diagonal matrix with (i, j) entry $\sigma_{i,j}$ given by $\sigma_{1,1} = \sigma^2$ and $\sigma_{2,2} = Var[(X_1 - \mu)^2] = 2\sigma^4$. In fact, $\Sigma = I^{-1}(\theta)$ in this case. ■

Example 14.4.2 (**MLE for a one-parameter exponential family**) Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from a one-parameter exponential family with common density with respect to a σ-finite measure μ given by

$$p_\theta(x) = \exp[\theta T(x) - A(\theta)] .$$

Here, θ is assumed to be an interior point of the natural parameter space. From Problem 2.16, recall that $E_\theta[T(X_i)] = A'(\theta)$ and $Var_\theta[T(X_i)] = A''(\theta)$. To show the maximum likelihood estimator is well-defined and to find an expression for it, we examine the derivative of the log of $L_n(\theta)$, which is equal to

$$\frac{\partial \log L_n(\theta)}{\partial \theta} = \sum_{i=1}^{n} [T(X_i) - A'(\theta)] .$$

The likelihood equation sets this equal to zero, which reduces to the equation $\bar{T}_n = A'(\theta)$, where $\bar{T}_n = n^{-1} \sum_{i=1}^n T(X_i)$. Hence, the MLE is found by equating the sample mean of the $T(X_i)$ values to its expected value. Assuming the equation $\bar{T}_n = A'(\theta)$ can be solved for θ, it must be the maximum likelihood estimator. Indeed, the second derivative of the loglikelihood is $-nA''(\theta) < 0$, which also shows there can be at most one solution to the likelihood equation. Furthermore, by the law of large numbers, $\bar{T}_n \xrightarrow{P} A'(\theta)$, which combined with the fact that $A''(\theta) > 0$ yields that, with probability tending to one, there exists exactly one solution to the likelihood equation. Thus, $\hat{\theta}_n$ is well-defined with probability tending to one. To determine its limiting distribution, first note that

$$n^{1/2}[\bar{T}_n - A'(\theta)] \xrightarrow{d} N(0, A''(\theta)) ,$$

by the Central Limit Theorem. Since A' is strictly increasing, we can define the inverse function B of A', so that $B(A'(\theta)) = \theta$. Then, $\hat{\theta}_n = B(A'(\hat{\theta}_n)) = B(\bar{T}_n)$. By the delta method,

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \tau^2) ,$$

where

$$\tau^2 = A''(\theta)[B'(A'(\theta))]^2 .$$

But using the chain rule to differentiate both sides of the identity $B(A'(\theta)) = \theta$ yields $B'(A'(\theta))A''(\theta) = 1$, so

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{A''(\theta)}\right) .$$

In fact, the asymptotic variance $[A''(\theta)]^{-1}$ is $I^{-1}(\theta)$, where $I(\theta)$ is the Fisher Information. ■

Problem 14.38 generalizes the previous example to multiparameter exponential families.

The general theory of asymptotic normality of the MLE is much more difficult and we shall here only give a heuristic treatment. For precise conditions and rigorous proofs, see Lehmann and Casella (1998), Chapter 6 and Ibragimov and Has'minskii (1981), Section 3.3. Let $X_1, \ldots, X_n$ be i.i.d. according to a family $\{P_\theta\}$ which is q.m.d. at θ_0 with nonsingular Fisher Information matrix $I(\theta_0)$ and quadratic mean derivative $\eta(\cdot, \theta_0)$. Define

$$L_{n,h} = \frac{L_n(\theta_0 + hn^{-1/2})}{L_n(\theta_0)} \,. \tag{14.57}$$

By Theorem 14.2.3,

$$\log(L_{n,h}) = \langle h, Z_n \rangle - \frac{1}{2}\langle h, I(\theta_0)h \rangle + o_{P_{\theta_0}^n}(1) \,, \tag{14.58}$$

where Z_n is the normalized score vector

$$Z_n = Z_n(\theta_0) = 2n^{-1/2} \sum_{i=1}^{n} [\eta(X_i, \theta_0)/p_{\theta_0}^{1/2}(X_i)] \tag{14.59}$$

and satisfies, under θ_0,

$$Z_n \xrightarrow{d} N(0, I(\theta_0)) \,.$$

Note that $Z_n = Z_n(\theta_0)$ depends on θ_0, but we will usually omit this dependence in the notation.

If the MLE $\hat{\theta}_n$ is well-defined, then $\hat{\theta}_n = \theta_0 + \hat{h}_n n^{-1/2}$, where $\hat{h}_n$ is the value of h maximizing $L_{n,h}$. The result (14.58) suggests that, if θ_0 is the true value, $\hat{h}_n$ is approximately equal to $\tilde{h}_n$ which maximizes

$$\log(\tilde{L}_{n,h}) \equiv \langle h, Z_n \rangle - \frac{1}{2}\langle h, I(\theta_0)h \rangle \,. \tag{14.60}$$

Since $\log(\tilde{L}_{n,h})$ is a simple (quadratic) function of h, it is easily checked (Problem 14.46) that

$$\tilde{h}_n = I^{-1}(\theta_0) Z_n \,. \tag{14.61}$$

It then follows that

$$n^{1/2}(\hat{\theta}_n - \theta_0) = \hat{h}_n \approx \tilde{h}_n = I^{-1}(\theta_0) Z_n \xrightarrow{d} N(0, I^{-1}(\theta_0)) \,.$$

The symbol $\approx$ is used to indicate an approximation based on heuristic considerations. Unfortunately, the above approximation is not rigorous without further conditions. In fact, without further conditions, the maximum likelihood estimator may not even be consistent. Indeed, an example of Le Cam (presented in Example 4.1 of Chapter 6 in Lehmann and Casella (1998)) shows that the maximum likelihood estimator $\hat{\theta}_n$ may exist and be unique but does not converge to the true value θ in probability (i.e., it is inconsistent). Moreover, the example shows this can happen even in very smooth families in which good estimators do exist. Rigorous conditions for the MLE to be consistent were given by Wald (1949), and have since then been weakened (for a survey, see Perlman (1972)). Cramér (1943) derived good asymptotic behavior of the maximum likelihood estimator under just certain smoothness conditions, often known as *Cramér type conditions*. Furthermore, he gave conditions under which there exists a consistent sequence of roots $\hat{\theta}_n$ of the likelihood equations (not necessarily the MLE) satisfying

$$n^{1/2}(\hat{\theta}_n - \theta_0) = I^{-1}(\theta_0) Z_n + o_{P^n_{\theta_0}}(1) \,, \tag{14.62}$$

from which asymptotic normality follows. Cramér's conditions required that the underlying family of densities was three times differentiable with respect to θ, as well as further technical assumptions on differentiability inside the integral signs; see Chapter 6 of Lehmann and Casella (1998). Estimators satisfying (14.62) are called *efficient*. In the case where $\hat{\theta}_n$ is a solution to the likelihood equations, it is called an *efficient likelihood estimator* (ELE) sequence.

Determination of an efficient sequence of roots of the likelihood equations tends to be difficult when the equations have multiple roots. Asymptotically equivalent estimators can be constructed by starting with any estimator $\tilde{\theta}_n$ that is $n^{1/2}$-consistent, i.e., for which $n^{1/2}(\tilde{\theta}_n - \theta)$ is bounded in probability. The resulting estimator can be taken to be the root closest to $\tilde{\theta}_n$, or an approximation to it based on a Newton–Raphson linearization method; for more details, see Section 6.4 of Lehmann and Casella (1998), Gan and Jiang (1999) and Small, Wang and, Yang (2000). A similar, but distinct, approach based on discretization of an initial estimator, leads to Le Cam's (1956, 1969) *one-step maximum likelihood estimator*, which satisfies (14.62) under fairly weak conditions.

If $\hat{\theta}_n$ is any estimator sequence (not necessarily the MLE or an ELE) which satisfies (14.62), it follows that, under θ_0,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0)) \,.$$

For the remainder of this section, we will assume such an estimator sequence $\hat{\theta}_n$ is available, by means of verification of Cramér type assumptions presented in Lehmann and Casella (1998), or by direct verification as in the case of exponential families of Example 14.4.2 and Problem 14.38. For testing applications, it is also important to study the behavior of the estimator under contiguous alternatives. The following

theorem assumes the expansion (14.62) (which is only assumed to hold under θ_0) in order to derive the limiting behavior of $\hat{\theta}_n$ under contiguous sequences θ_n.

Theorem 14.4.1 *Assume $X_1, \ldots, X_n$ are i.i.d. according to a q.m.d. model $\{P_\theta,\ \theta \in \Omega\}$ with nonsingular Information matrix $I(\theta)$, $\theta \in \Omega$, an open subset of $\mathbb{R}^k$. Suppose an estimator $\hat{\theta}_n$ has the expansion (14.62) when $\theta = \theta_0$. Let $\theta_n = \theta_0 + h_n n^{-1/2}$, where $h_n \to h \in \mathbb{R}^k$. Then, under $P^n_{\theta_n}$,*

$$n^{1/2}(\hat{\theta}_n - \theta_n) \xrightarrow{d} N(0, I^{-1}(\theta_0)) \; ; \tag{14.63}$$

equivalently, under $P^n_{\theta_n}$,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(h, I^{-1}(\theta_0)) \; . \tag{14.64}$$

Furthermore, if $g(\theta)$ is a differentiable map from Ω to $\mathbb{R}$ with nonzero gradient $\dot{g}(\theta)$ of dimension $1 \times k$, then under $P^n_{\theta_n}$,

$$n^{1/2}(g(\hat{\theta}_n) - g(\theta_n)) \xrightarrow{d} N(0, \sigma^2_{\theta_0}) \; , \tag{14.65}$$

where

$$\sigma^2_{\theta_0} = \dot{g}(\theta_0) I^{-1}(\theta_0) \dot{g}(\theta_0)^\top \; . \tag{14.66}$$

PROOF. We prove the result in the case $h_n = h$, the more general case deferred to Problem 15.13. We will first show (14.64). By the Cramér–Wold Device, it is enough to show that, for any $t \in \mathbb{R}^k$, under $P^n_{\theta_n}$,

$$\langle n^{1/2}(\hat{\theta}_n - \theta_0), t \rangle \xrightarrow{d} N(\langle h, t \rangle, \langle t, I^{-1}(\theta_0) t \rangle) \; .$$

By the assumption (14.62), we only need to show that, under $P^n_{\theta_n}$,

$$\langle I^{-1}(\theta_0) Z_n, t \rangle \xrightarrow{d} N(\langle h, t \rangle, \langle t, I^{-1}(\theta_0) t \rangle) \; .$$

By Example 14.3.7, $P^n_{\theta_n}$ is contiguous to $P^n_{\theta_0}$, so we can apply Corollary 14.3.2 with $T_n = \langle I^{-1}(\theta_0) Z_n, t \rangle$. Then,

$$(T_n, \log(L_{n,h})) = (\langle I^{-1}(\theta_0) Z_n, t \rangle, \langle h, Z_n \rangle - \frac{1}{2} \langle h, I(\theta_0) h \rangle) + o_{P^n_{\theta_0}}(1) \; .$$

But, under θ_0, Z_n converges in law to Z, where Z is distributed as $N(0, I(\theta_0))$. By Slutsky's Theorem and the Continuous Mapping Theorem (or the bivariate Central Limit Theorem), under θ_0,

$$(T_n, \log(L_{n,h})) \xrightarrow{d} (\langle I^{-1}(\theta_0) Z, t \rangle, \langle h, Z \rangle - \frac{1}{2} \langle h, I(\theta_0) h \rangle) \; .$$

This limiting distribution is bivariate normal with covariance

$$\sigma_{1,2} = Cov(\langle I^{-1}(\theta_0)Z, t\rangle, \langle h, Z\rangle) = E[(h^\top Z)(I^{-1}(\theta_0)Z)^\top t]$$

$$= h^\top E(Z_1 Z_1^\top) I^{-1}(\theta_0)t = h^\top I(\theta_0) I^{-1}(\theta_0)t = \langle h, t\rangle \ .$$

The result (14.64) follows from Corollary 14.3.2. The assertion (14.65) follows from (14.63) and the delta method. ■

Under the conditions of the previous theorem, the estimator sequence $g(\hat{\theta}_n)$ possesses a weak robustness property in the sense that its limiting distribution is unchanged by small perturbations of the parameter values. In the literature, such estimator sequences are sometimes called *regular*.

Corollary 14.4.1 *Assume $X_1, \ldots, X_n$ are i.i.d. according to a q.m.d. model $\{P_\theta,\ \theta \in \Omega\}$ with normalized score vector Z_n given by (14.59) and nonsingular Information matrix $I(\theta_0)$. Let $\theta_n = \theta_0 + h_n n^{-1/2}$, where $h_n \to h \in \mathbb{R}^k$. Then, under $P^n_{\theta_n}$,*

$$Z_n \xrightarrow{d} N(I(\theta_0)h, I(\theta_0)) \ . \tag{14.67}$$

The proof is left as an exercise (Problem 14.39).

14.4.2 Wald Tests and Confidence Regions

Wald proposed tests and confidence regions based on the asymptotic distribution of the maximum likelihood estimator. In this section, we introduce these methods and study their large-sample behavior; some optimality properties will be discussed in Sections 15.3 and 15.4. We assume $\hat{\theta}_n$ is any estimator satisfying (14.62). Let $g(\theta)$ be a mapping from Ω to the real line, assumed differentiable with nonzero gradient vector $\dot{g}(\theta)$ of dimension $1 \times k$. Suppose the problem is to test the null hypothesis $g(\theta) = 0$ versus the alternative $g(\theta) > 0$. Let θ_0 denote the true value of θ. Under the assumptions of Theorem 14.4.1, under θ_0,

$$n^{1/2}[g(\hat{\theta}_n) - g(\theta_0)] \xrightarrow{d} N(0, \sigma^2_{\theta_0}) \ ,$$

where

$$\sigma^2_{\theta_0} = \dot{g}(\theta_0) I^{-1}(\theta_0) \dot{g}(\theta_0)^\top \ .$$

Assuming that $\dot{g}(\cdot)$ and $I(\cdot)$ are continuous, the asymptotic variance can be consistently estimated by

$$\hat{\sigma}^2_n \equiv \dot{g}(\hat{\theta}_n) I^{-1}(\hat{\theta}_n) \dot{g}(\hat{\theta}_n)^\top \ .$$

Hence, the test that rejects when

$$n^{1/2} g(\hat{\theta}_n) > \hat{\sigma}_n z_{1-\alpha}$$

is pointwise asymptotically level α.

We can also calculate the limiting power against a sequence of alternatives $\theta_n = \theta_0 + hn^{-1/2}$. Assume $g(\theta_0) = 0$. Then,

$$P_{\theta_n}\{n^{1/2} g(\hat{\theta}_n) > \hat{\sigma}_n z_{1-\alpha}\} = P_{\theta_n}\{n^{1/2}[g(\hat{\theta}_n) - g(\theta_n)] > \hat{\sigma}_n z_{1-\alpha} - n^{1/2} g(\theta_n)\} .$$

By Theorem 14.4.1, $n^{1/2}[g(\hat{\theta}_n) - g(\theta_n)]$ is asymptotically $N(0, \sigma^2_{\theta_0})$, under θ_n. Also, $\hat{\sigma}_n \to \sigma_{\theta_0}$ in probability under θ_n (since this convergence holds under θ_0 and therefore under θ_n by contiguity). Finally, $n^{1/2} g(\theta_n) \to \dot{g}(\theta_0)h$. Hence, the limiting power is

$$\lim_{n\to\infty} P_{\theta_n}\{n^{1/2} g(\hat{\theta}_n) > \hat{\sigma}_n z_{1-\alpha}\} = 1 - \Phi(z_{1-\alpha} - \sigma^{-1}_{\theta_0} \dot{g}(\theta_0)h) . \tag{14.68}$$

Similarly, a pointwise asymptotically level $1 - \alpha$ confidence interval for $g(\theta)$ is given by

$$g(\hat{\theta}_n) \pm z_{1-\frac{\alpha}{2}} n^{-1/2} \hat{\sigma}_n .$$

Example 14.4.3 (**Normal Coefficient of Variation**) Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with both parameters unknown, as in Example 14.4.1. Consider inferences for $g((\mu, \sigma^2)^\top) = \mu/\sigma$, the coefficient of variation. Recall that a uniformly most accurate invariant one-sided confidence bound exists for μ/σ; however, it is quite complicated to compute since it involves the noncentral t-distribution and no explicit formula is available. However, a normal approximation leads to an interval that is asymptotically valid. Note that

$$\dot{g}((\mu, \sigma^2)^\top) = (\frac{1}{\sigma}, -\frac{\mu}{2\sigma^3}) .$$

By Example 14.4.1, $n^{1/2}[(\bar{X}_n, S_n^2)^\top - (\mu, \sigma^2)^\top]$ is asymptotically bivariate normal with asymptotic covariance matrix Σ, where Σ is the diagonal matrix with $(1, 1)$ entry σ^2 and $(2, 2)$ entry $2\sigma^4$. Then, the delta method implies that

$$n^{1/2}(\frac{\bar{X}_n}{S_n} - \frac{\mu}{\sigma}) \xrightarrow{d} N(0, 1 + \frac{\mu^2}{2\sigma^2}) .$$

Thus, the interval

$$\frac{\bar{X}_n}{S_n} \pm n^{-1/2}(1 + \frac{\bar{X}_n^2}{2S_n^2}) z_{1-\frac{\alpha}{2}}$$

is asymptotically pointwise level $1 - \alpha$. ■

Consider now the general problem of constructing a confidence region for θ, under the assumptions of Theorem 14.4.1. The convergence

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I^{-1}(\theta)) \tag{14.69}$$

implies that

$$I^{1/2}(\theta) n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I_k) ,$$

the multivariate normal distribution in $\mathbb{R}^k$ with mean 0 and identity covariance matrix I_k. Hence, by the Continuous Mapping Theorem 11.2.10 and Example 11.2.5,

$$n(\hat{\theta}_n - \theta)^\top I(\theta)(\hat{\theta}_n - \theta) \xrightarrow{d} \chi^2_k ,$$

the Chi-squared distribution with k degrees of freedom. Thus, a pointwise asymptotic level $1 - \alpha$ confidence region for θ is

$$\{\theta : n(\hat{\theta}_n - \theta)^\top I(\theta)(\hat{\theta}_n - \theta) \le c_{k,1-\alpha}\} , \tag{14.70}$$

where $c_{k,1-\alpha}$ is the $1 - \alpha$ quantile of χ^2_k. In (14.70), $I(\theta)$ is often replaced by a consistent estimator, such as $I(\hat{\theta}_n)$ (assuming $I(\cdot)$ is continuous), and the resulting confidence region is known as Wald's confidence ellipsoid.

By the duality between confidence regions and tests, this leads to an asymptotic level α test of $\theta = \theta_0$ versus $\theta \neq \theta_0$, known as Wald tests. Specifically, for testing $\theta = \theta_0$ versus $\theta \neq \theta_0$, Wald's test rejects if

$$n(\hat{\theta}_n - \theta_0) I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) > c_{k,1-\alpha} . \tag{14.71}$$

Alternatively, $I(\hat{\theta}_n)$ may be replaced by $I(\theta_0)$ or any consistent estimator of $I(\theta_0)$. Under $\theta_n = \theta_0 + hn^{-1/2}$, the limiting distribution of the Wald statistic given by the left side of (14.71) is $\chi^2_k(|I^{1/2}(\theta_0)h|^2)$, the noncentral Chi-squared distribution with k degrees of freedom and noncentrality parameter $|I^{1/2}(\theta_0)h|^2$ (Problem 14.50).

More generally, consider inference for $g(\theta)$, where $g = (g_1, \ldots, g_p)^\top$ is a mapping from $\mathbb{R}^k$ to $\mathbb{R}^p$. Assume g_i is differentiable and let $D = D(\theta)$ denote the $p \times k$ matrix with (i, j) entry given by

$$D_{i,j}(\theta) = \partial g_i(\theta_1, \ldots, \theta_k)/\partial\theta_j . \tag{14.72}$$

Then, the Delta Method and (14.69) imply that

$$n^{1/2}[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{d} N(0, V(\theta)) , \tag{14.73}$$

where $V(\theta) = D(\theta) I^{-1}(\theta) D^\top(\theta)$. Assume $V(\theta)$ is positive definite and continuous in θ. By the Continuous Mapping Theorem,

$$n[g(\hat{\theta}_n) - g(\theta)]^\top V^{-1}(\theta)[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{d} \chi^2_p .$$

Hence, a pointwise asymptotically level $1-\alpha$ confidence region for $g(\theta)$ is

$$\{\theta:\ n[g(\hat{\theta}_n)-g(\theta)]^\top V^{-1}(\hat{\theta}_n)[g(\hat{\theta}_n)-g(\theta)]\le\chi^2_p(1-\alpha)\}\ .$$

Next, suppose it is desired to test $g(\theta)=0$. The Wald test rejects when

$$W_n = ng(\hat{\theta}_n)V^{-1}(\hat{\theta}_n)g^\top(\hat{\theta}_n)$$

exceeds $\chi^2_p(1-\alpha)$, and it is pointwise asymptotically level α.

14.4.3 Rao Score Tests

Instead of the Wald tests, it is possible to construct tests based directly on Z_n in (14.59), which have the advantage of not requiring computation of a maximum likelihood estimator. Assume q.m.d. holds at θ_0, with derivative $\eta(\cdot,\theta_0)$ and, as usual, set

$$\tilde{\eta}(x,\theta_0)=2\eta(x,\theta_0)/p^{1/2}_{\theta_0}(x)\ .$$

Under the assumptions of Theorem 14.2.2, the quadratic mean derivative $\eta(\cdot,\theta_0)$ is given by (14.9) and $n^{1/2}Z_n$ can then be computed by

$$n^{1/2}Z_n=\sum_{i=1}^n\tilde{\eta}(X_i,\theta_0)=$$

$$\sum_{i=1}^n\frac{\dot{p}_{\theta_0}(X_i)}{p_{\theta_0}(X_i)}=\left(\frac{\partial}{\partial\theta_1}\log L_n(\theta),\ldots,\frac{\partial}{\partial\theta_k}\log L_n(\theta)\right)\Big|_{\theta=\theta_0}\ . \tag{14.74}$$

As mentioned earlier, the statistic Z_n is known as the normalized *score* vector. Its use stems from the fact that inference can be based on Z_n, which involves differentiating the loglikelihood at a single point θ_0, avoiding the problem of maximizing the likelihood. Even if the ordinary differentiability conditions assumed in Theorem 14.2.2 fail, inference can be based on Z_n, as we will now see.

Suppose for the moment that θ is real-valued and consider testing $\theta=\theta_0$ versus $\theta>\theta_0$. For a given test $\phi=\phi(X_1,\ldots,X_n)$, let

$$\beta_\phi(\theta)=E_\theta[\phi(X_1,\ldots,X_n)]$$

denote its power function. By Problem 14.17, assuming q.m.d., $\beta_\phi(\theta)$ is differentiable at θ_0 with

$$\beta'_\phi(\theta_0) = \int \cdots \int \phi(x_1, \ldots, x_n) \sum_{i=1}^{n} \tilde{\eta}(x_i, \theta_0) \prod_{i=1}^{n} p_{\theta_0}(x_i) \mu(dx_1) \cdots \mu(dx_n) \, .$$

Consider the problem of finding the level α test ϕ that maximizes $\beta'_\phi(\theta_0)$. By the general form of the Neyman–Pearson Lemma, the optimal test rejects for large values of $\sum_i \tilde{\eta}(X_i, \theta_0)$, or equivalently, large values of Z_n. By Problem 8.4, if this is the unique test maximizing the slope of the power function at θ_0, then it is also locally most powerful. Thus, tests based on Z_n are appealing from this point of view.

We turn now to the asymptotic behavior of tests based on Z_n. Assume the assumptions of quadratic mean differentiability hold for general k, so that under θ_0,

$$Z_n \xrightarrow{d} N(0, I(\theta_0)) \, .$$

By Corollary 14.4.1, under $\theta_n = \theta_0 + hn^{-1/2}$,

$$Z_n \xrightarrow{d} N(I(\theta_0)h, I(\theta_0)) \, .$$

It follows that, under $\theta_n = \theta_0 + hn^{-1/2}$,

$$I^{-1/2}(\theta_0) Z_n \xrightarrow{d} N(I^{1/2}(\theta_0)h, I_k) \, . \tag{14.75}$$

Now, suppose $k = 1$ and the problem is to test $\theta = \theta_0$ versus $\theta > \theta_0$. Rao's score test rejects when the one-sided *score statistic* $I^{-1/2}(\theta_0) Z_n$ exceeds $z_{1-\alpha}$ and is asymptotically level α. In this case, the Wald test that rejects when $I^{1/2}(\theta_0) n^{1/2} (\hat{\theta}_n - \theta_0)$ exceeds $z_{1-\alpha}$ and the score test are asymptotically equivalent, in the sense that the probability that the two tests yield the same decision tends to one, both under the null hypothesis $\theta = \theta_0$ and under a sequence of alternatives $\theta_0 + hn^{-1/2}$. The equivalence follows from contiguity, the expansion (14.62), and the fact that $I(\hat{\theta}_n) \to I(\theta_0)$ in probability under θ_0 and under $\theta_0 + hn^{-1/2}$. Note that the two tests may differ greatly for alternatives far from θ_0; see Example 15.3.3.

Example 14.4.4 (**Bivariate Normal Correlation**) Assume $X_i = (U_i, V_i)$ are i.i.d. according to the bivariate normal distribution with means zero and variances one, so that the only unknown parameter is ρ, the correlation. In this case,

$$\log L_n(\rho) = -n \log(2\pi) - \frac{n}{2} \log(1 - \rho^2) - \sum_{i=1}^{n} [\frac{1}{2(1 - \rho^2)} (U_i^2 - 2\rho U_i V_i + V_i^2)]$$

and so

$$\frac{\partial}{\partial \rho} \log L_n(\rho) = \frac{n\rho}{1 - \rho^2} + \frac{1}{1 - \rho^2} \sum_{i=1}^{n} U_i V_i - \frac{\rho}{(1 - \rho^2)^2} \sum_{i=1}^{n} (U_i^2 - 2\rho U_i V_i + V_i^2) \, .$$

In the special case $\theta_0 = \rho_0 = 0$,

$$Z_n = n^{-1/2} \sum_{i=1}^{n} U_i V_i \xrightarrow{d} N(0, 1) .$$

For other values of ρ_0, the statistic is more complicated; however, we have bypassed maximizing the likelihood function which may have multiple roots in this example. ■

For general k, consider testing a simple null hypothesis $\theta = \theta_0$ versus a multi-sided alternative $\theta \neq \theta_0$. Then, assuming the expansion (14.62), we can replace $n^{1/2}(\hat{\theta}_n - \theta_0)$ in the Wald statistic (14.70) by $I^{-1}(\theta_0) Z_n$. In this case, the *score test* rejects the null hypothesis when the multi-sided *score statistic* $Z_n^\top I^{-1}(\theta_0) Z_n$ exceeds $c_{k,1-\alpha}$, and is asymptotically level α. Again, the Wald test and Rao's score test are asymptotically equivalent in the sense described above.

Next, we consider a composite null hypothesis. Interest focuses on the first p components of θ, $\theta_1, \ldots, \theta_p$, with the remaining $k - p$ components viewed as nuisance parameters. Let $\theta_{1,0}, \ldots, \theta_{p,0}$ be fixed and consider testing the null hypothesis $\theta_i = \theta_{i,0}$ for $i = 1, \ldots, p$. The Wald test is based on the limit

$$n^{1/2}(\hat{\theta}_{n,1} - \theta_1, \ldots, \hat{\theta}_{n,p} - \theta_p) \xrightarrow{d} N\left(0, \Sigma^{(p)}(\theta)\right) ,$$

where $\Sigma(\theta) = I^{-1}(\theta)$ and $\Sigma^{(p)}(\theta)$ is the $p \times p$ matrix formed by the intersection of the first p rows and columns of $\Sigma(\theta)$. Similarly, define $I^{(p)}(\theta)$ as the $p \times p$ matrix formed by the intersection of the first p rows and columns of $I(\theta)$. Partition $I(\theta)$ as

$$I(\theta) = \begin{pmatrix} I^{(p)}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix} . \tag{14.76}$$

Note that (Problem 14.51)

$$[\Sigma^{(p)}(\theta)]^{-1} = [I^{(p)}(\theta)] - I_{12}(\theta) I_{22}^{-1}(\theta) I_{21}(\theta) . \tag{14.77}$$

The score test is based on $Z_n^{(p)}(\theta)$, the p-vector obtained as the first p components of $Z_n(\theta)$, where $Z_n(\theta)$ is defined in (14.59). Under q.m.d. at θ,

$$Z_n^{(p)}(\theta) \xrightarrow{d} N\left(0, I^{(p)}(\theta)\right) ,$$

and so

$$T_n(\theta) = [Z_n^{(p)}(\theta)]^\top [I^{(p)}(\theta)]^{-1} [Z_n^{(p)}(\theta)] \xrightarrow{d} \chi_p^2 .$$

However, when the null hypothesis is not completely specified, the Rao score test statistic is $T_n(\hat{\theta}_{n,0})$, where

$$\hat{\theta}_{n,0} = (\theta_{1,0}, \ldots, \theta_{p,0}, \hat{\theta}_{p+1,0}, \ldots, \hat{\theta}_{k,0})$$

is the constrained maximum likelihood estimator of θ, that is, the maximum likelihood estimator under the restricted parameter space satisfying the constraints of the null hypothesis. In fact, as argued by Hall and Mathiason (1990), any $\sqrt{n}$-consistent estimator can be used in the score statistic.

In order to determine the constrained maximum likelihood estimator $\tilde{\theta}_n$, one typically introduces the Lagrangian function and maximizes

$$\log L_n(\theta) - \sum_{i=1}^{p} \lambda_i(\theta_i - \theta_{i,0})$$

over θ and the so-called Lagrange multipliers $\lambda_1, \ldots, \lambda_p$. Assuming differentiability, the first-order conditions require that

$$\frac{\partial}{\partial \theta_i} \log L_n(\theta)|_{\theta=\tilde{\theta}_n} = \tilde{\lambda}_i \quad i = 1, \ldots, p \, .$$

Since the left-hand side represents the components of $\sqrt{n} Z_n^{(p)}(\tilde{\theta}_n)$, tests based on $Z_n^{(p)}(\tilde{\theta}_n)$ are equivalent to tests based on $\tilde{\lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p)^\top$. For this reason, score tests are sometimes referred to as Lagrange multiplier tests. Such terminology is particularly popular in econometrics.

More generally, suppose the null hypothesis H_0 specifies $g_i(\theta) = c_i$, for $i = 1, \ldots, p$, where $p \leq k$ and the c_i are fixed. Let $\tilde{\theta}_n$ be the restricted maximum likelihood estimator for the null hypothesis parameter space, assuming it exists. The score test is based on

$$Z_n(\tilde{\theta}_n)^\top I^{-1}(\tilde{\theta}_n) Z_n(\tilde{\theta}_n) \, ,$$

which, under smoothness assumptions on the g_j, is asymptotically Chi-squared with p degrees of freedom under H_0. Note $Z_n(\tilde{\theta}_n)$ can also be represented in terms of Lagrange multipliers. To see how, let $S_n(\theta)$ be the $k \times 1$ score vector with ith component

$$S_{n,i}(\theta) = \frac{\partial}{\partial \theta_i} \log L_n(\theta) \, .$$

(So, $Z_n(\theta) = n^{-1/2} S_n(\theta)$ under usual differentiability; see Theorem 14.2.2 and (14.74)). The Lagrangian function is

$$\log L_n(\theta) - \sum_{i=1}^{p} \lambda_i [g_i(\theta) - c_i] \, .$$

Let $D_n(\theta)$ be the $p \times k$ matrix with (i, j) entry given in (14.72). Then, $\tilde{\theta}_n$ satisfies the first-order conditions if

$$g_i(\tilde{\theta}_n) = c_i \quad \text{for } i = 1, \ldots, p$$

and

$$S_n(\tilde{\theta}_n) - D(\tilde{\theta}_n)^\top \tilde{\lambda} = 0 \,,$$

or $Z_n(\tilde{\theta}_n) = \sqrt{n} D(\tilde{\theta}_n)^\top \tilde{\lambda}$.

14.4.4 Likelihood Ratio Tests

In addition to that Wald and Rao scores tests of Sections 14.4.2 and 14.4.3, let us now consider a third test of $\theta \in \Omega_0$ versus $\theta \notin \Omega_0$, based on the *likelihood ratio statistic* $2\log(R_n)$, where

$$R_n = \frac{\sup_{\theta \in \Omega} L_n(\theta)}{\sup_{\theta \in \Omega_0} L_n(\theta)} \,. \tag{14.78}$$

The *likelihood ratio test* rejects for large values of $2\log(R_n)$. If $\hat{\theta}_n$ and $\hat{\theta}_{n,0}$ are MLEs for θ as θ varies in Ω and Ω_0, respectively, then

$$R_n = L_n(\hat{\theta}_n)/L_n(\hat{\theta}_{n,0}) \,. \tag{14.79}$$

In the real-valued case when testing the simple null hypothesis $\theta = \theta_0$, Figure 14.1 plots the logarithm of the likelihood function $L_n(\theta)$ as a function of θ. The difference between the MLE $\hat{\theta}_n$ and θ_0 can easily be seen on the horizontal θ-axis, and serves as a basis for a Wald test. The difference between the loglikelihood function at $\hat{\theta}_n$ and θ_0 is depicted as $\log(R_n)$. Rao's score test derives from the tangent (in green) to the loglikelihood when $\theta = \theta_0$.

Example 14.4.5 (**Multivariate Normal Mean**) Suppose $X = (X_1, \ldots, X_k)^\top$ is multivariate normal with unknown mean vector θ and known positive definite covariance matrix Σ. The likelihood function is given by

$$\frac{|\Sigma|^{-1/2}}{(2\pi)^{k/2}} \exp\left[-\frac{1}{2}(X-\theta)^\top \Sigma^{-1}(X-\theta)\right] \,.$$

Assume $\theta \in \mathbb{R}^k$ and that the null hypothesis asserts $\theta_i = 0$ for $i = 1, \ldots, k$. Then,

$$2\log(R_1) = -\inf_{\theta}(X-\theta)^\top \Sigma^{-1}(X-\theta) + X^\top \Sigma^{-1} X = X^\top \Sigma^{-1} X = |\Sigma^{-1/2} X|^2 \,.$$

Under the null hypothesis, $\Sigma^{-1/2} X$ is exactly standard multivariate normal, and so the null distribution of $2\log(R_1)$ is exactly χ^2_k in this case.

Now, consider testing the composite hypothesis $\theta_i = 0$ for $i = 1, \ldots, p$, with the remaining parameters $\theta_{p+1}, \ldots, \theta_k$ regarded as nuisance parameters. More generally, suppose

$$\Omega_0 = \{\theta = (\theta_1, \ldots, \theta_k) : \; A(\theta - a) = 0\} \,, \tag{14.80}$$

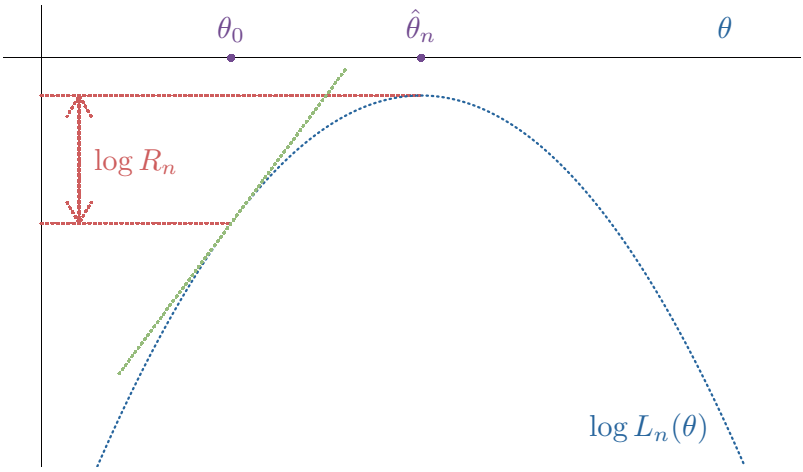

Figure 14.1 Loglikelihood Function

where A is a $p \times k$ matrix of rank p and a is some fixed $k \times 1$ vector. Then,

$$2\log(R_1) = -\inf_{\theta \in \mathbb{R}^k} (X-\theta)^\top \Sigma^{-1}(X-\theta) + \inf_{\theta \in \Omega_0} (X-\theta)^\top \Sigma^{-1}(X-\theta)$$

$$= \inf_{\theta \in \Omega_0} (X-\theta)^\top \Sigma^{-1}(X-\theta)\,. \tag{14.81}$$

The null distribution of (14.81) is χ^2_p (Problem 14.52). ■

Let us now consider the large-sample behavior of the likelihood ratio test in greater generality. First, suppose $\Omega_0 = \{\theta_0\}$ is simple. Then,

$$\log(R_n) = \sup_h [\log(L_{n,h})]\,,$$

where $L_{n,h}$ is defined in (14.57). If the family is q.m.d. at θ_0, then

$$\log(R_n) = \sup_h [\langle h, Z_n\rangle - \frac{1}{2}\langle h, I(\theta_0)h\rangle + o_{P^n_{\theta_0}}(1)]\,.$$

It is plausible that $\log(R_n)$ should behave like

$$\log \tilde{R}_n \equiv \sup_h [\log(\tilde{L}_{n,h})]\,,$$

where $\tilde{L}_{n,h}$ is defined by (14.60). But $\tilde{L}_{n,h}$ is maximized at $\tilde{h}_n = I^{-1}(\theta_0)Z_n$ and so

$$\log(R_n) \approx \log(\tilde{R}_n) = \log(\tilde{L}_{n,\tilde{h}_n}) = \frac{1}{2} Z_n^\top I^{-1}(\theta_0) Z_n \ .$$

Since, $2\log(\tilde{R}_n) \xrightarrow{d} \chi^2_k$, the heuristics suggest that $2\log(R_n) \xrightarrow{d} \chi^2_k$ as well. In fact, $2\log(\tilde{R}_n)$ is Rao's score test statistic, and so these heuristics also suggest that Rao's score test, the likelihood ratio test, and Wald's test are all asymptotically equivalent in the sense described earlier in comparing the Wald test and the score test. Note, however, that the tests are not always asymptotically equivalent; some striking differences will be presented in Section 15.3.

These heuristics can be made rigorous under stronger assumptions, such as Cramér type differentiability conditions used in proving asymptotic normality of the MLE or an ELE; see Theorem 7.7.2 in Lehmann (1999). Alternatively, once the general heuristics point toward the limiting behavior, the approximations may be made rigorous by direct calculation in a particular situation. A general theorem based on the existence of efficient likelihood estimators will be presented following the next example.

Example 14.4.6 (**Multinomial Goodness of Fit**) Consider a sequence of n independent trials, each resulting in one of $k+1$ outcomes $1, \ldots, k+1$. Outcome j occurs with probability p_j on any given trial. Let Y_j be the number of trials resulting in outcome j. Consider testing the simple null hypothesis $p_j = \pi_j$ for $j = 1, \ldots, k+1$. The parameter space Ω is

$$\Omega = \{(p_1, \ldots, p_k) \in \mathbb{R}^k : \ p_i \geq 0, \ \sum_{j=1}^{k} p_j \leq 1\} \tag{14.82}$$

since p_{k+1} is determined as $1 - \sum_{j=1}^{k} p_j$. In this case, the likelihood can be written as

$$L_n(p_1, \ldots, p_k) = \frac{n!}{Y_1! \cdots Y_{k+1}!} p_1^{Y_1} \cdots p_{k+1}^{Y_{k+1}} \ .$$

By solving the likelihood equations, it is easily checked that the unique MLE is given by $\hat{p}_j = Y_j/n$ (Problem 14.57 (i)). Hence, the likelihood ratio statistic is

$$R_n = \frac{L_n(Y_1/n, \ldots, Y_k/n)}{L_n(\pi_1, \ldots, \pi_k)} \ ,$$

and so (Problem 14.57 (ii))

$$\log(R_n) = n \sum_{j=1}^{k+1} \hat{p}_j \log(\frac{\hat{p}_j}{\pi_j}) \ . \tag{14.83}$$

The previous heuristics suggest that $2\log(R_n)$ converges in distribution to χ^2_k, which will be proved in Theorem 14.4.2 below. Note that the Taylor expansion

$$f(x) = x\log(x/x_0) = (x - x_0) + \frac{1}{2x_0}(x - x_0)^2 + o[(x - x_0)^2]$$

as $x \to x_0$ implies $2\log(R_n) \approx Q_n$, where Q_n is Pearson's Chi-squared statistic given by

$$Q_n = \sum_{j=1}^{k+1} \frac{(Y_j - n\pi_j)^2}{n\pi_j} \,. \tag{14.84}$$

Indeed, $2\log(R_n) - Q_n \xrightarrow{P} 0$, under the null hypothesis (Problem 14.59) and so they have the same limiting distribution. Moreover, it can be checked (Problem 14.58) that Rao's Score test statistic is exactly Q_n. The Chi-squared test will be treated more fully in Section 16.3. ■

Next, we present a fairly general result on the asymptotic distribution of the likelihood ratio statistic. Actually, we consider a generalization of the likelihood ratio statistic. Rather than having to compute the maximum likelihood estimators $\hat{\theta}_n$ and $\hat{\theta}_{n,0}$ in (14.79), we assume these estimators satisfy (14.62) under the models with parameter spaces Ω and Ω_0, respectively.

Theorem 14.4.2 *Assume $X_1, \ldots, X_n$ are i.i.d. according to q.m.d. family $\{P_\theta,\ \theta \in \Omega\}$, where Ω is an open subset of $\mathbb{R}^k$ and $I(\theta)$ is positive definite. Further assume, for θ in a neighborhood of θ_0 and a (measurable) function $M(x)$ satisfying $E_{\theta_0}[M(X_i)] < \infty$,*

$$|\log p_\theta(x) - \log p_{\theta_0}(x) - (\theta - \theta_0)\tilde{\eta}_{\theta_0}(x)| \le M(x)|\theta - \theta_0|^2 \,. \tag{14.85}$$

(i) Consider testing the simple null hypothesis $\theta = \theta_0$. Suppose $\hat{\theta}_n$ is an efficient estimator for θ assuming $\theta \in \Omega$ in the sense that it satisfies (14.62) when $\theta = \theta_0$. Then, the likelihood ratio $R_n = L_n(\hat{\theta}_n)/L_n(\theta_0)$ satisfies, under θ_0,

$$2\log(R_n) \xrightarrow{d} \chi^2_k \,.$$

(ii) Consider testing the composite null hypothesis $\theta \in \Omega_0$, where

$$\Omega_0 = \{\theta = (\theta_1, \ldots, \theta_k) :\ A(\theta - a) = 0\} \,,$$

and A is a $p \times k$ matrix of rank p and a is a fixed $k \times 1$ vector. Let $\hat{\theta}_{n,0}$ denote an efficient estimator of θ assuming $\theta \in \Omega_0$; that is, assume the expansion (14.62) holds based on the model $\{P_\theta,\ \theta \in \Omega_0\}$ and any $\theta \in \Omega_0$. Then, the likelihood ratio $R_n = L_n(\hat{\theta}_n)/L_n(\hat{\theta}_{n,0})$ satisfies, under any $\theta_0 \in \Omega_0$,

$$2\log(R_n) \xrightarrow{d} \chi^2_p \,.$$

(iii) More generally, suppose Ω_0 is represented as

$$\Omega_0 = \{\theta : \ g = (g_1(\theta), \ldots, g_p(\theta))^\top = 0\} ,$$

where $g_i(\theta)$ is a continuously differentiable function from $\mathbb{R}^k$ to $\mathbb{R}$. Let $D = D(\theta)$ be the $p \times k$ matrix with (i, j) entry $\partial g_i(\theta)/\partial\theta_j$, assumed to have rank p. Then, $2\log(R_n) \xrightarrow{d} \chi^2_p$.

PROOF. First, consider (i). Let $\hat{h}_n = n^{1/2}(\hat{\theta}_n - \theta_0)$ so that $2\log(R_n) = 2\log(L_{n,\hat{h}_n})$. Fix any $c > 0$ and define

$$\epsilon_{n,c} = \sup_{|h|\le c} |\log(L_{n,h}) - [\langle h, Z_n\rangle - \frac{1}{2}\langle h, I(\theta_0)h\rangle]| ;$$

by Problem 15.12, $\epsilon_{n,c} \to 0$ in probability under θ_0. By the triangle inequality,

$$2\log(L_{n,\hat{h}_n}) \le 2[\langle \hat{h}_n, Z_n\rangle - \frac{1}{2}\langle \hat{h}_n, I(\theta_0)\hat{h}_n\rangle + \epsilon_{n,c}]$$

if $|\hat{h}_n| \le c$. But, using (14.62),

$$2[\langle \hat{h}_n, Z_n\rangle - \frac{1}{2}\langle \hat{h}_n, I(\theta_0)\hat{h}_n\rangle] = Z_n^\top I^{-1}(\theta_0)Z_n + o_{P_{\theta_0}}(1) ;$$

so,

$$2\log(L_{n,\hat{h}_n}) \le Z_n^\top I^{-1}(\theta_0)Z_n + \tilde{\epsilon}_{n,c}$$

if $|\hat{h}_n| \le c$, where $\tilde{\epsilon}_{n,c} \to 0$ in probability under θ_0 for any $c > 0$. Therefore,

$$P\{2\log(L_{n,\hat{h}_n}) \ge x\} \le P\{Z_n^\top I^{-1}(\theta_0)Z_n + \tilde{\epsilon}_{n,c} \ge x, \ |\hat{h}_n| \le c\} + P\{|\hat{h}_n| > c\}$$

$$\le P\{Z_n^\top I^{-1}(\theta_0)Z_n + \tilde{\epsilon}_{n,c} \ge x\} + P\{|\hat{h}_n| > c\} . \tag{14.86}$$

But, under θ_0 , $Z_n^\top I^{-1}(\theta_0)Z_n$ is asymptotically χ^2_k and $\hat{h}_n \xrightarrow{d} Z$ where Z is $N(0, I^{-1}(\theta_0))$, so (14.86) tends to

$$P\{\chi^2_k \ge x\} + P\{|Z| > c\} .$$

Let $c \to \infty$ to conclude

$$\limsup_n P\{2\log(L_{n,\hat{h}_n}) \ge x\} \le P\{\chi^2_k \ge x\} .$$

A similar argument yields

$$\liminf_n P\{2\log(L_{n,\hat{h}_n}) \geq x\} \geq P\{\chi_k^2 \geq x\}\,, \tag{14.87}$$

and (i) is proved.

The proof of (ii) is based on a similar argument, combined with the results of Example 14.4.5 for testing a composite null hypothesis about a multivariate normal mean vector. The proof of (iii) is left as an exercise (Problem 14.62). ■

Example 14.4.7 (**One-Sample Normal Mean**) Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with both parameters unknown. Consider testing $\mu = 0$ versus $\mu \neq 0$. Then (Problem 14.47),

$$2\log(R_n) = n\log(1 + \frac{t_n^2}{n-1})\,, \tag{14.88}$$

where $t_n^2 = n\bar{X}_n^2/S_n^2$ is the one-sample t-statistic. By Problem 13.28, one can deduce the following Edgeworth expansion for $2\log(R_n)$ (Problem 14.48):

$$P\{2\log(R_n) \leq r\} = 1 - 2[\Phi(-z) + \frac{3}{4n}z\phi(z)] + O(n^{-2})\,, \tag{14.89}$$

where $z = \sqrt{r}$, Φ is the standard normal c.d.f. and $\Phi' = \phi$. This implies that the test that rejects when $2\log(R_n) > z_{1-\frac{\alpha}{2}}$ has rejection probability equal to $\alpha + O(n^{-1})$. But, a simple correction, known as a *Bartlett correction*, can improve the χ_1^2 approximation. Indeed, (14.89) and a Taylor expansion implies

$$P\{2\log(R_n)(1 + \frac{b}{n}) > z_{1-\frac{\alpha}{2}}\} = \alpha + O(n^{-2})\,, \tag{14.90}$$

if we take $b = 3/2$. Thus, the error in rejection probability of the Bartlett-corrected test is $O(n^{-2})$. Of course, in this example, the exact two-sided t-test is available. ■

It is worth knowing that, quite generally, a simple multiplicative correction to the likelihood ratio statistic greatly improves the quality of the approximation. Specifically, for an appropriate choice of b, comparing $2\log(R_n)(1 + \frac{b}{n})$ to the usual limiting χ_p^2 reduces the error in rejection probability from $O(n^{-1})$ to $O(n^{-2})$. In practice, b can be derived by analytical means or estimated. The idea for such a Bartlett correction originated in Bartlett (1937). For appropriate regularity conditions that imply a Bartlett correction works, see Barndorff-Nielsen and Hall (1988), Bickel and Ghosh (1990), Jensen (1993) and DiCiccio and Stern (1994).

For hypotheses of the form assumed in Theorem 14.4.2, the degrees of freedom can be remembered as the dimension of Ω minus the dimension of Ω_0. In the special case where the null hypothesis is specified by $\theta_i = \theta_{i,0}$ for $i = 1, \ldots, p$ and θ_j regarded as a nuisance parameter for $j = p+1, \ldots, k$, then the dimension of Ω is k and the dimension of Ω_0 is $k - p$; the degrees of freedom reduces to the number of parameters p with values specified under the null hypothesis.

However, even for very smooth models, the limiting distribution of the likelihood ratio test need not be Chi-squared under the null hypothesis when the null parameter

space takes a different form than assumed in Theorem 14.4.2. The problem of testing moment inequalities is an important example, as illustrated next.

Example 14.4.8 (**Moment Inequalities**) The problem of testing moment inequalities was first considered in Example 8.7.3. Assume $X_1, \ldots, X_n$ are i.i.d. multivariate normal with unknown mean vector $\theta \in \Omega = \mathbb{R}^k$ and known invertible covariance matrix Σ. The problem is to test the null hypothesis $\theta \in \Omega_0$, where

$$\Omega_0 = \{\theta : \theta_i \leq 0 \quad \text{for all } i = 1, \ldots k\} .$$

Of course, one can reduce by sufficiency to the sample mean vector $\bar{X}_n$. Then, the argument leading to (14.81) shows that the likelihood ratio statistic in this case becomes

$$2\log(R_n) = n \inf_{\theta \in \Omega_0} (\bar{X}_n - \theta)^\top \Sigma^{-1} (\bar{X}_n - \theta) .$$

Assume Σ is the identity matrix. Then, R_n reduces to

$$2\log(R_n) = n \sum_{i=1}^{k} \max^2(\bar{X}_{n,i}, 0) .$$

If $\theta = 0$, then the distribution of $2\log(R_n)$ can be represented (exactly for any n) as that of

$$\sum_{i=1}^{k} \max^2(Z_i, 0) ,$$

where the Z_i are i.i.d. standard normal. Let $c_{k,1-\alpha}$ be the $1-\alpha$ quantile of this distribution. Then, by monotonicity (as in Example 8.7.3), the test that rejects when $2\log(R_n)$ exceeds $c_{k,1-\alpha}$ has size α. On the other hand, if θ is such that exactly p components are 0 and the remaining $k-p$ components are negative, then

$$2\log(R_n) \xrightarrow{d} \sum_{i=1}^{p} \max(Z_i^2, 0) .$$

Note the family of distributions on the right-hand side is stochastically increasing in p. Finally, if θ lies in the interior of Ω_0, $2\log(R_n) \xrightarrow{P} 0$.

Suppose that $\theta_n = \theta_0 + hn^{-1/2}$, where $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0.k})$ lies on the boundary of Ω. Let I be the set of indices i for which $\theta_{0,i} = 0$. Then, the limiting power of the likelihood ratio test can be represented as (Problem 14.67)

$$P\{\sum_{i \in I} \max^2(Z_i + h_i, 0) > c_{k,1-\alpha}\} . \tag{14.91}$$

If $p = |I|$ is the number of indices of θ_0 that are zero and $p < k$, then since $c_{p,1-\alpha} < c_{k,1-\alpha}$, (14.92) is bounded above by

$$P\{\sum_{i\in I} \max^2(Z_i + h_i, 0) > c_{p,1-\alpha}\} . \tag{14.92}$$

Moment selection methods allow one to achieve power (14.92); see Problem 14.68. ■

14.5 Problems

Section 14.2

Problem 14.1 Generalize Example 14.2.1 to the case where X is multivariate normal with mean vector θ and nonsingular covariance matrix Σ.

Problem 14.2 Generalize Example 14.2.2 to the case of a multiparameter exponential family. Compare with the result of Problem 14.1.

Problem 14.3 Suppose g_n is a sequence of functions in $L^2(\mu)$; that is, $\int g_n^2 d\mu < \infty$. Assume, for some function g, $\int (g_n - g)^2 d\mu \to 0$. Prove that $\int g^2 d\mu < \infty$.

Problem 14.4 Suppose g_n is a sequence of functions in $L^2(\mu)$ and, for some function g, $\int (g_n - g)^2 d\mu \to 0$. If $\int h^2 d\mu < \infty$, show that $\int h g_n d\mu \to \int h g d\mu$.

Problem 14.5 Suppose X and Y are independent, with X distributed as P_θ and Y as $\bar{P}_\theta$, as θ varies in a common index set Ω. Assume the families $\{P_\theta\}$ and $\{\bar{P}_\theta\}$ are q.m.d. with Fisher Information matrices $I_X(\theta)$ and $I_Y(\theta)$, respectively. Show that the model based on the joint data (X, Y) is q.m.d. and its Fisher Information matrix is given by $I_X(\theta) + I_Y(\theta)$.

Problem 14.6 Fix a probability P. Let $u(x)$ satisfy

$$\int u(x) dP(x) = 0 .$$

(i) Assume $\sup_x |u(x)| < \infty$, so that

$$p_\theta(x) = [1 + \theta u(x)]$$

defines a family of densities (with respect to P) for all small $|\theta|$. Show this family is q.m.d. at $\theta = 0$. Calculate the quadratic mean derivative, score function, and $I(0)$.
(ii) Alternatively, if u is unbounded, define $p_\theta(x) = C(\theta) \exp(\theta u(x))$, assuming $\int \exp(\theta u(x)) dP(x)$ exists for all small $|\theta|$. For this family, argue the family is q.m.d. at $\theta = 0$, and calculate the score function and $I(0)$.

(iii) Suppose $\int u^2(x)dP(x) < \infty$. Define

$$p_\theta(x) = C(\theta)2[1 + \exp(-2\theta u(x))]^{-1} .$$

Show this family is q.m.d. at $\theta = 0$, and calculate the score function and $I(0)$. [The constructions in this problem are important for nonparametric applications, used later in Chapters 15 and 16. The last construction is given in van der Vaart (1998).]

Problem 14.7 Fix a probability P on S and functions $u_i(x)$ such that $\int u_i(x)dP(x) = 0$ and $\int u_i^2(x)dP(x) < \infty$, for $i = 1, 2$. Adapt Problem 14.6 to construct a family of distributions P_θ with $\theta \in \mathbb{R}^2$, defined for all small $|\theta|$, such that $P_{0,0} = P$, the family is q.m.d. at $\theta = (0, 0)$ with score vector at $\theta = (0, 0)$ given by $(u_1(x), u_2(x))$. If S is the real line, construct the P_θ that works even if P_θ is required to be smooth if P and the u_i are smooth (i.e., having differentiable densities) or subject to moment constraints (i.e., having finite pth moments).

Problem 14.8 Show that the definition of $I(\theta)$ in Definition 14.2.2 does not depend on the choice of dominating measure μ.

Problem 14.9 In Examples 14.2.3 and 14.2.4, find the quadratic mean derivative and $I(\theta)$.

Problem 14.10 In Example 14.2.5, show that $\int\{[f'(x)]^2/f(x)\}dx$ is finite iff $\beta > 1/2$.

Problem 14.11 Prove Theorem 14.2.2 using an argument similar to the proof of Theorem 14.2.1.

Problem 14.12 Suppose $\{P_\theta\}$ is q.m.d. at θ_0 with derivative $\eta(\cdot, \theta_0)$. Show that, on $\{x : p_{\theta_0}(x) = 0\}$, we must have $\eta(x, \theta_0) = 0$, except possibly on a μ-null set. *Hint*: On $\{p_{\theta_0}(x) = 0\}$, write

$$0 \le n^{1/2} p^{1/2}_{\theta_0+hn^{-1/2}}(x) = \langle h, \eta(x, \theta_0)\rangle + r_{n,h}(x) ,$$

where $\int r^2_{n,h}(x)\mu(dx) \to 0$. This implies, with h fixed, that $r_{n,h}(x) \to 0$ except for x in μ-null set, at least along some subsequence.

Problem 14.13 Suppose $\{P_\theta\}$ is q.m.d. at θ_0. Show

$$P_{\theta_0+h}\{x : p_{\theta_0}(x) = 0\} = o(|h|^2)$$

as $|h| \to 0$. Hence, if $X_1, \ldots, X_n$ are i.i.d. with likelihood ratio $L_{n,h}$ defined by (14.12), show that

$$P^n_{\theta_0+hn^{-1/2}}\{L_{n,h} = \infty\} \to 0 .$$

Problem 14.14 To see what might happen when the parameter space is not open, let

$$f_0(x) = xI\{0 \le x \le 1\} + (2 - x)I\{1 < x \le 2\} .$$

Consider the family of densities indexed by $\theta \in [0, 1)$ defined by

$$p_\theta(x) = (1 - \theta^2) f_0(x) + \theta^2 f_0(x - 2) .$$

Show that the condition (14.5) holds when $\theta_0 = 0$, if it is only required that h tends to 0 through positive values. Investigate the behavior of the likelihood ratio (14.12) for such a family. (For a more general treatment, consult Pollard (1997)).

Problem 14.15 Suppose $X_1, \ldots, X_n$ are i.i.d. and uniformly distributed on $(0, \theta)$. Let $p_\theta(x) = \theta^{-1} I\{0 < x < \theta\}$ and $L_n(\theta) = \prod_i p_\theta(X_i)$. Fix p and θ_0. Determine the limiting behavior of $L_n(\theta_0 + hn^{-p})/L_n(\theta_0)$ under θ_0. For what p and h is the limiting distribution nondegenerate?

Problem 14.16 Suppose $\{P_\theta, \theta \in \Omega\}$ is a model with Ω an open subset of $\mathbb{R}^k$, and having densities $p_\theta(x)$ with respect to μ. Define the model to be L_1-differentiable at θ_0 if there exists a vector of real-valued functions $\zeta(\cdot, \theta_0)$ such that

$$\int |p_{\theta_0+h}(x) - p_{\theta_0}(x) - \langle \zeta(x, \theta_0), h\rangle| d\mu(x) = o(|h|) \tag{14.93}$$

as $|h| \to 0$. Show that, if the family is q.m.d. at θ_0 with q.m. derivative $\eta(\cdot, \theta_0)$, then it is L_1-differentiable with

$$\zeta(x, \theta_0) = 2\eta(x, \theta_0) p_{\theta_0}^{1/2}(x) ,$$

but the converse is false.

Problem 14.17 Assume $\{P_\theta, \theta \in \Omega\}$ is L_1-differentiable, so that (14.93) holds. For simplicity, assume $k = 1$ (but the problem generalizes). Let $\phi(\cdot)$ be uniformly bounded and set $\beta(\theta) = E_\theta[\phi(X)]$. Show $\beta'(\theta)$ exists at θ_0 and

$$\beta'(\theta_0) = \int \phi(x)\zeta(x, \theta_0)\mu(dx) . \tag{14.94}$$

Hence, if $\{P_\theta\}$ is q.m.d. at θ_0 with derivative $\eta(\cdot, \theta_0)$, then

$$\beta'(\theta_0) = \int \phi(x)\tilde{\eta}(x, \theta_0) p_{\theta_0}(x)\mu(dx) , \tag{14.95}$$

where $\tilde{\eta}(x, \theta_0) = 2\eta(x, \theta_0)/p_{\theta_0}^{1/2}(x)$. More generally, if $X_1, \ldots, X_n$ are i.i.d. P_θ and $\phi(X_1, \ldots, X_n)$ is uniformly bounded, then $\beta(\theta) = E_\theta[\phi(X_1, \ldots, X_n)]$ is differentiable at θ_0 with

$$\beta'(\theta_0) = \int \cdots \int \phi(x_1, \ldots, x_n) \sum_{i=1}^{n} \tilde{\eta}(x_i, \theta_0) \prod_{i=1}^{n} p_{\theta_0}(x_i) \mu(dx_1) \cdots \mu(dx_n) \,. \tag{14.96}$$

Section 14.3

Problem 14.18 Prove (14.31).

Problem 14.19 Show the convergence (14.35).

Problem 14.20 Fix two probabilities P and Q and let $P_n = P$ and $Q_n = Q$. Show that $\{P_n\}$ and $\{Q_n\}$ are contiguous iff P and Q are absolutely continuous.

Problem 14.21 Fix two probabilities P and Q and let $P_n = P^n$ and $Q_n = Q^n$. Show that $\{P_n\}$ and $\{Q_n\}$ are contiguous iff $P = Q$.

Problem 14.22 Suppose Q_n is contiguous to P_n and let L_n be the likelihood ratio defined by (14.36). Show that $E_{P_n}(L_n) \to 1$. Is the converse true?

Problem 14.23 Consider a sequence $\{P_n, Q_n\}$ with likelihood ratio L_n defined in (14.36). Assume $\mathcal{L}(L_n|P_n) \xrightarrow{d} W$, where $P\{W = 0\} = 0$; show P_n is contiguous to Q_n. Also, under (14.41), deduce that P_n is contiguous to Q_n and hence P_n and Q_n are mutually contiguous if and only if $\mu = -\sigma^2/2$.

Problem 14.24 Suppose, under P_n, $X_n = Y_n + o_{P_n}(1)$; that is, $X_n - Y_n \to 0$ in P_n-probability. Suppose Q_n is contiguous to P_n. Show that $X_n = Y_n + o_{Q_n}(1)$.

Problem 14.25 Suppose X_n has distribution P_n or Q_n and $T_n = T_n(X_n)$ is sufficient. Let P_n^T and Q_n^T denote the distribution of T_n under P_n and Q_n, respectively. Prove or disprove: Q_n is contiguous to P_n if and only if Q_n^T is contiguous to P_n^T.

Problem 14.26 Suppose Q is absolutely continuous with respect to P. If $P\{E_n\} \to 0$, then $Q\{E_n\} \to 0$.

Problem 14.27 Prove the convergence (14.40).

Problem 14.28 Show that $\sigma_{1,2}$ in (14.52) reduces to $h/\sqrt{\pi}$.

Problem 14.29 Verify (14.53) and evaluate it in the case where $f(x) = \exp(-|x|)/2$ is the double exponential density.

Problem 14.30 Reconsider Example 14.3.11. Rather than finding the limiting distribution of W_n under contiguous alternatives, find the limiting distribution of U_n (properly normalized) under the same set of alternatives, where U_n is the U-statistic introduced in Example 12.3.6. First, find the projection of U_n under the null hypothesis, which represents U_n as an asymptotically linear statistic. Then, relating U_n and W_n, check that your solution agrees with the solution in Example 14.3.11.

Problem 14.31 Suppose $X_1, \ldots, X_n$ are i.i.d. according to a model which is q.m.d. at θ_0. For testing $\theta = \theta_0$ versus $\theta = \theta_0 + hn^{-1/2}$, consider the test ψ_n that rejects H if $\log(L_{n,h})$ exceeds $z_{1-\alpha}\sigma_h - \frac{1}{2}\sigma_h^2$, where $L_{n,h}$ is defined by (14.54) and $\sigma_h^2 = \langle h, I(\theta_0)h\rangle$. Find the limiting value of $E_{\theta_0+hn^{-1/2}}(\psi_n)$.

Problem 14.32 Suppose P_θ is the uniform distribution on $(0, \theta)$. Fix h and determine whether or not P_1^n and $P_{1+h/n}^n$ are mutually contiguous. Consider both $h > 0$ and $h < 0$.

Problem 14.33 Generalize Corollary 14.3.2 in the following way. Suppose $T_n = (T_{n,1}, \ldots, T_{n,k}) \in \mathbb{R}^k$. Assume that, under P_n,

$$(T_{n,1}, \ldots, T_{n,k}, \log(L_n)) \xrightarrow{d} (T_1, \ldots, T_k, Z) ,$$

where $(T_1, \ldots, T_k, Z)$ is multivariate normal with $Cov(T_i, Z) = c_i$. Then, under Q_n,

$$(T_{n,1}, \ldots, T_{n,k}) \xrightarrow{d} (T_1 + c_1, \ldots, T_k + c_k) .$$

Problem 14.34 Suppose $X_1, \ldots, X_n$ are i.i.d. according to a model $\{P_\theta : \theta \in \Omega\}$, where Ω is an open subset of $\mathbf{R}^k$. Assume that the model is q.m.d. Show that there cannot exist an estimator sequence T_n satisfying

$$\lim_{n\to\infty} \sup_{|\theta-\theta_0|\le n^{-1/2}} P_\theta^n(n^{1/2}|T_n - \theta| > \epsilon) = 0 \tag{14.97}$$

for every $\epsilon > 0$ and any θ_0. (Here P_θ^n means the joint probability distribution of $(X_1, \ldots, X_n)$ under θ.) Suppose the above condition (14.97) only holds for some $\epsilon > 0$. Does the same conclusion hold?

Problem 14.35 Assume X_i are independent, normally distributed with $E(X_i) = \mu_i$. Let P_n be the distribution of $(X_1, \ldots, X_n)$ when $\mu_i = 0$ for all i. Let Q_n be the distribution of $(X_1, \ldots, X_n)$ when the μ_i are arbitrary constants. Find a necessary and sufficient condition on $\mu_1, \mu_2, \ldots$ so that P_n and Q_n are mutually contiguous.

Section 14.4

Problem 14.36 In Example 14.4.1, show that the likelihood equations have a unique solution which corresponds to a global maximum of the likelihood function.

Problem 14.37 Suppose $X_1, \ldots, X_n$ are i.i.d. P_θ according to the lognormal model of Example 14.2.7. Write down the likelihood function and show that it is unbounded.

Problem 14.38 Generalize Example 14.4.2 to multiparameter exponential families.

Problem 14.39 Prove Corollary 14.4.1. *Hint*: Simply define $\hat{\theta}_n = \theta_0 + n^{-1/2} I^{-1}(\theta_0) Z_n$ and apply Theorem 14.4.1.

Problem 14.40 Let $(X_i, Y_i), i = 1 \ldots n$ be i.i.d. such that X_i and Y_i are independent and normally distributed, X_i has variance σ^2, Y_i has variance τ^2 and both have common mean μ.
(i) If σ and τ are known, determine an efficient likelihood estimator (ELE) $\hat{\mu}$ of μ and find the limit distribution of $n^{1/2}(\hat{\mu} - \mu)$.
(ii) If σ and τ are unknown, provide an estimator $\bar{\mu}$ for which $n^{1/2}(\bar{\mu} - \mu)$ has the same limit distribution as $n^{1/2}(\hat{\mu} - \mu)$.
(iii) What can you infer from your results (i) and (ii) regarding the Information matrix $I(\theta)$, $\theta = (\mu, \sigma, \tau)$?

Problem 14.41 Let $X_1, \ldots, X_n$ be a sample from a Cauchy location model with density $f(x - \theta)$, where

$$f(z) = \frac{1}{\pi(1 + z^2)}.$$

Compare the limiting distribution of the sample median with that of an efficient likelihood estimator.

Problem 14.42 Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, \theta^2)$. Compare the asymptotic distribution of $\bar{X}_n^2$ with that of an efficient likelihood estimator sequence.

Problem 14.43 Let $X_1, \cdots, X_n$ be i.i.d. with density

$$f(x, \theta) = [1 + \theta \cos(x)]/2\pi,$$

where the parameter θ satisfies $|\theta| < 1$ and x ranges between 0 and 2π. (The observations X_i may be interpreted as directional data. The case $\theta = 0$ corresponds to the uniform distribution on the circle.) Construct an efficient likelihood estimator of θ, as explicitly as possible.

Problem 14.44 Suppose $X_1, \ldots, X_n$ are i.i.d. with common density function

$$p_\theta(x) = \frac{\theta c^\theta}{x^{\theta+1}}\ ,\ 0 < c < x,\ \theta > 0\ .$$

Here, c is fixed and known and θ is unknown.
(i) Show that the maximum likelihood estimator $\hat{\theta}_n$ is well-defined and determine the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ under θ.
(ii) What is the score test for testing the null hypothesis $\theta = \theta_0$ vs. $\theta \neq \theta_0$?

Problem 14.45 Suppose $X_1, \ldots, X_n$ are i.i.d., uniformly distributed on $[0, \theta]$. Find the maximum likelihood estimator $\hat{\theta}_n$ of θ. Determine a sequence τ_n such that $\tau_n(\hat{\theta}_n - \theta)$ has a limiting distribution, and determine the limit law.

Problem 14.46 Verify that $\tilde{h}_n$ in (14.61) maximizes $\tilde{L}_{n,h}$.

Problem 14.47 Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with both parameters unknown. Consider testing $\mu = 0$ versus $\mu \neq 0$. Find the likelihood ratio test statistic, and determine its limiting distribution under the null hypothesis. Calculate the limiting power of the test against the sequence of alternatives $(\mu, \sigma^2) = (h_1 n^{-1/2}, \sigma^2 + h_2 n^{-1/2})$.

Problem 14.48 In Example 14.4.7, verify (14.89) and (14.90).

Problem 14.49 Suppose a time series $X_0, X_1, X_2, \ldots$ evolves in the following way. The process starts at 0, so $X_0 = 0$. For any $i \geq 1$, conditional on $X_0, \ldots, X_{i-1}$, $X_i = \rho X_{i-1} + \epsilon_i$, where the ϵ_i are i.i.d. standard normal. You observe $X_0, X_1, X_2, \ldots, X_n$. For testing the null hypothesis $\rho = 0$ versus $\rho > 0$, determine both Wald and Rao score tests as well as appropriate critical values. (Compare with Problems 3.35 and 14.49).

Problem 14.50 Suppose $X_1, \ldots, X_n$ are i.i.d. P_θ, with $\theta \in \Omega$, an open subset of $\mathbb{R}^k$. Assume the family is q.m.d. at θ_0 and consider testing the simple null hypothesis $\theta = \theta_0$. Suppose $\hat{\theta}_n$ is an estimator sequence satisfying (14.62), and consider the Wald test statistic $n(\hat{\theta}_n - \theta_0)^\top I(\theta_0)(\hat{\theta}_n - \theta_0)$. Find its limiting distribution against the sequence of alternatives $\theta_0 + hn^{-1/2}$, as well as an expression for its limiting power against such a sequence of alternatives.

Problem 14.51 Prove (14.77). Then, show that

$$[\Sigma^{(p)}(\theta)]^{-1} \leq [I^{(p)}(\theta)] \, .$$

What is the statistical interpretation of this inequality?

Problem 14.52 In Example 14.4.5, consider the case of a composite null hypothesis with Ω_0 given by (14.80). Show that the null distribution of the likelihood ratio statistic given by (14.81) is χ^2_p. *Hint*: First consider the case $a = 0$, so that Ω_0 is a linear subspace of dimension $k - p$. Let $Z = \Sigma^{-1/2} X$, so

$$2\log(R_n) = \inf_{\theta \in \Omega_0} |Z - \Sigma^{-1/2}\theta|^2 \, .$$

As θ varies in Ω_0, $\Sigma^{-1/2}\theta$ varies in a subspace L of dimension $k - p$. If P is the projection matrix onto L and I is the identity matrix, then $2\log(R_n) = |(I - P)Z|^2$.

Problem 14.53 In Example 14.4.5, determine the distribution of the likelihood ratio statistic against an alternative, both for the simple and composite null hypotheses.

Problem 14.54 Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with both parameters unknown. Consider testing the simple null hypothesis $(\mu, \sigma^2) = (0, 1)$. Find and compare the Wald test, Rao's Score test, and the likelihood ratio test.

Problem 14.55 Suppose $X_1, \ldots, X_n$ are i.i.d. with the gamma $\Gamma(g, b)$ density

$$f(x) = \frac{1}{\Gamma(g)b^g} x^{g-1} e^{-x/b} \quad x > 0\,,$$

with both parameters unknown (and positive). Consider testing the null hypothesis that $g = 1$, i.e., under the null hypothesis the underlying density is exponential. Determine the likelihood ratio test statistic and find its limiting distribution.

Problem 14.56 Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d., with X_i also independent of Y_i. Further suppose X_i is normal with mean μ_1 and variance 1, and Y_i is normal with mean μ_2 and variance 1. It is known that $\mu_i \geq 0$ for $i = 1, 2$. The problem is to test the null hypothesis that at most one μ_i is positive versus the alternative that both μ_1 and μ_2 are positive.
(i) Determine the likelihood ratio statistic for this problem.
(ii) In order to carry out the test, how would you choose the critical value (sequence) so that the size of the test is α?

Problem 14.57 (i) In Example 14.4.6, check that the MLE is given by $\hat{p}_j = Y_j/n$.
(ii) Show (14.83).

Problem 14.58 In Example 14.4.6, show that Rao's Score test is exactly Pearson's Chi-squared test.

Problem 14.59 In Example 14.4.6, show that $2\log(R_n) - Q_n \xrightarrow{P} 0$ under the null hypothesis.

Problem 14.60 Prove (14.87).

Problem 14.61 Provide the details of the proof to part (ii) of Theorem 14.4.2.

Problem 14.62 Prove (iii) of Theorem 14.4.2. *Hint*: If θ_0 satisfies the null hypothesis $g(\theta_0) = 0$, then testing Ω_0 behaves asymptotically like testing the null hypothesis $D(\theta_0)(\theta - \theta_0) = 0$, which is a hypothesis of the form considered in part (ii) of the theorem.

Problem 14.63 The problem is to test independence in a contingency table. Specifically, suppose $X_1, \ldots, X_n$ are i.i.d., where each X_i is cross-classified, so that $X_i = (r, s)$ with probability $p_{r,s}, r = 1, \ldots, R, s = 1, \ldots, S$. Under the full model, the $p_{r,s}$ vary freely, except they are nonnegative and sum to 1. Let $p_{r\cdot} = \sum_s p_{r,s}$ and $p_{\cdot s} = \sum_r p_{r,s}$. The null hypothesis asserts $p_{r,s} = p_{r\cdot} p_{\cdot s}$ for all r and s. Determine the likelihood ratio test and its limiting null distribution.

Problem 14.64 Consider the following model which therefore generalizes model (iii) of Section 4.7. A sample of n_i subjects is obtained from class $A_i (i = 1, \ldots, a)$, the samples from different classes being independent. If $Y_{i,j}$ is the number of subjects from the ith sample belonging to $B_j (j = 1, \ldots, b)$, the joint distribution of $(Y_{i,1}, \ldots, Y_{i,b})$ is multinomial, say,

$$M(n_i;\, p_{1|i}, \ldots, p_{b|i})\ .$$

Determine the likelihood ratio statistic for testing the hypothesis of *homogeneity* that the vector $(p_{1|i}, \ldots, p_{b|i})$ is independent of i, and specify its asymptotic distribution.

Problem 14.65 The *hypothesis of symmetry* in a square two-way contingency table arises when one of the responses $A_1, \ldots, A_a$ is observed for each of n subjects on two occasions (e.g., before and after some intervention). If $Y_{i,j}$ is the number of subjects whose responses on the two occasions are (A_i, A_j), the joint distribution of the $Y_{i,j}$ is multinomial, with the probability of a subject response of (A_i, A_j) denoted by $p_{i,j}$. The hypothesis H of *symmetry* states that $p_{i,j} = p_{j,i}$ for all i and j; that is, that the intervention has not changed the probabilities. Determine the likelihood ratio statistic for testing H, and specify its asymptotic distribution. [Bowker (1948)].

Problem 14.66 In the situation of Problem 14.65, consider the *hypothesis of marginal homogeneity* $H' : p_{i+} = p_{+i}$ for all i, where $p_{i+} = \sum_{j=1}^{a} p_{iij}$, $p_{+i} = \sum_{j=1}^{a} p_{jii}$.

(i) The maximum-likelihood estimates of the p_{iij} under H' are given by $\hat{\hat{p}}_{ij} = Y_{ij}/(1 + \lambda_i - \lambda_j)$, where the λ's are the solutions of the equations $\sum_j Y_{ij}/(1 + \lambda_i - \lambda_j) = \sum_j Y_{ij}/(1 + \lambda_j - \lambda_i)$. (These equations have no explicit solutions.)
(ii) Determine the number of degrees of freedom for the limiting χ^2-distribution of the likelihood ratio criterion.

Problem 14.67 In Example 14.4.8, show (14.91).

Problem 14.68 Consider testing moment inequalities under the setting of Example 14.4.8. Rather than the likelihood ratio procedure discussed there, consider the following moment selection procedure. Let $J = \{j : \sqrt{n}\bar{X}_{n,j} > -\log(n)\}$. Then, reject if the likelihood ratio statistic $2\log(R_n) > c_{|J|,1-\alpha}$, where $|J|$ is the cardinality of J.
(i) For any fixed $\theta \in \Omega_0$, show that the probability of a Type 1 error under θ is asymptotically no bigger than α. Does the size of the test tend to α?
(ii) Show that this procedure asymptotically achieves the power in (14.92). Can you think of any criticism of the procedure?
(iii) Consider the test that rejects H_0 when

$$M_n = \sqrt{n} \max_{1 \le i \le k} (\bar{X}_{n,i}) > d_{k,1-\alpha}\ ,$$

where $d_{k,1-\alpha}$ is the distribution of $\max_{1 \le i \le k} Z_i$ when the Z_i are i.i.d. standard normal. Compute the limiting rejection probability under $\theta = \theta_0 + hn^{-1/2}$ for any θ_0 on the boundary of Ω_0 and any $h \in \mathbb{R}^k$.
(iv) As in (ii) above, apply a moment selection procedure based on the test statistic M_n, and repeat (iii) for the procedure. [Moment selection methods for testing moment inequalities are discussed in Andrews and Barwick (2012) and Romano, Shaikh and

Wolf (2014). Special emphasis is placed on error control that is uniform in the underlying distribution.]

Problem 14.69 Consider the third of the three sampling schemes for a $2 \times 2 \times K$ table discussed in Section 4.8, and the two hypotheses

$$H_1 : \Delta_1 = \cdots = \Delta_K = 1 \quad \text{and} \quad H_2 : \Delta_1 = \cdots = \Delta_K.$$

(i) Obtain the likelihood ratio test statistic for testing H_1.
(ii) Obtain equations that determine the maximum likelihood estimates of the parameters under H_2. (These equations cannot be solved explicitly.)
(iii) Determine the number of degrees of freedom of the limiting χ^2-distribution of the likelihood ratio test for testing (a) H_1, (b) H_2.
[For a discussion of these and related hypotheses, see for example Shaffer (1973), Plackett (1981), or Bishop, Fienberg, and Holland (1975), and the recent study by Liang and Self (1985).]

Problem 14.70 Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$. Consider Hodges' superefficient estimator of θ (unpublished, but cited in Le Cam (1953)), defined as follows. Let $\hat{\theta}_n$ be 0 if $|\bar{X}_n| \le n^{-1/4}$; otherwise, let $\hat{\theta}_n = \bar{X}_n$. For any fixed θ, determine the limiting distribution of $n^{1/2}(\hat{\theta}_n - \theta)$. Next, determine the limiting distribution of $n^{1/2}(\hat{\theta}_n - \theta_n)$ under $\theta_n = hn^{-1/2}$. Is $\hat{\theta}_n$ regular?

Problem 14.71 Suppose $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbf{R}^k$ having the multivariate normal distribution with unknown mean vector μ and identity covariance matrix. Fix a constant $c > 0$ and consider the shrinkage estimator of μ defined by

$$\hat{\mu}_n = \left(1 - \frac{c}{n\|\bar{X}_n\|^2}\right) \bar{X}_n \,,$$

where $\bar{X}_n$ is the sample mean vector and $\|\cdot\|$ is Euclidean norm. Determine whether or not $\hat{\mu}_n$ is regular at $\mu = 0$ by deriving the limiting distribution of $\sqrt{n}(\hat{\mu}_n - \mu)$ under $\mu_n = h/\sqrt{n}$.

Problem 14.72 Suppose $X_1, \ldots, X_n$ are i.i.d. according to a quadratic mean differentiable model $\{P_\theta, \theta \in \Omega\}$, where Ω is an open subset of the real line. Suppose an estimator sequence $\hat{\theta}_n$ is asymptotically linear in the sense that,

$$n^{1/2}(\hat{\theta}_n - \theta_0) = n^{-1/2} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) + o_{P^n_{\theta_0}}(1)$$

where $E_{\theta_0}[\psi_{\theta_0}(X_i)] = 0$ and $\tau^2 = Var_{\theta_0}[\psi_{\theta_0}(X_i)] < \infty$.
(i) Find the joint limiting behavior of $(n^{1/2}(\hat{\theta}_n - \theta_0), Z_n)$ under $\theta_0 + hn^{-1/2}$, where Z_n is the normalized score statistic given by

$$Z_n = n^{-1/2} \sum_{i=1}^{n} \tilde{\eta}(X_i, \theta_0)$$

and $\tilde{\eta}(\cdot, \theta_0)$ is the usual score function.
(ii) Find a simple if and only if condition for $\hat{\theta}_n$ to be regular (at θ_0). (Your answer should depend on something about the functions $\psi_{\theta_0}(\cdot)$ and $\tilde{\eta}(\cdot, \theta_0)$.)
(iii) Find a simple if and only if condition for the statistic

$$D_n \equiv n^{1/2}(\hat{\theta}_n - \theta_0) - I^{-1}(\theta_0) Z_n$$

to be asymptotically ancillary at θ_0, in the sense that its limiting distribution under $\theta_0 + hn^{-1/2}$ does not depend on h.
(iv) Find a simple if and only if condition for D_n and Z_n to be asymptotically independent.
(v) Under the additional assumption of regularity, find a simple if and only if condition for the statistic sequence D_n defined above to tend in probability under θ_0 to 0 (and hence $\hat{\theta}_n$ is efficient under this condition).

Problem 14.73 Let $(X_{j,1}, X_{j,2})$, $j = 1, \ldots, n$ be independent pairs of independent exponentially distributed random variables with $E(X_{j,1}) = \theta\lambda_j$ and $E(X_{j,2}) = \lambda_j$. Here, θ and the λ_j are all unknown. The problem is to test $\theta = 1$ against $\theta > 1$. Compare the Rao, Wald, and likelihood ratio tests for this problem. Without appealing to any general results, find the limiting distribution of your statistics, as well as the limiting power against suitable local alternatives. (Note: the number of parameters is increasing with n so you can't directly appeal to our previous large-sample results.)

14.6 Notes

According to Le Cam and Yang (2000), the notion of quadratic mean differentiability was initiated in conversations between Hájek and Le Cam in 1962. Hájek (1962) appears to be the first publication making use of this notion. The importance of q.m.d. was prominent in the fundamental works of Le Cam (1969, 1970) and Hájek (1972), and has been used extensively ever since.

The notion of (mutual) contiguity is due to Le Cam (1960). Its usefulness was soon recognized by Hájek (1962), who first considered the one-sided version. Three of Le Cam's fundamental lemmas concerning contiguity became known as Le Cam's three lemmas, largely due to their prominence in Hájek and Sidák (1967). Further results can be found in Roussas (1972), Le Cam (1990), Chapter 6, Hájek et al. (1999), and Le Cam and Yang (2000), Chapter 3.

The methods studied in Section 14.4 are based on the notion of *likelihood*, whose general importance was recognized in Fisher (1922, 1925a, 1925b). Rigorous approaches were developed by Wald (1939, 1943) and Cramér (1943). Cramér defined the asymptotic efficiency of an asymptotically normal estimator to be the

ratio of its asymptotic variance to the Fisher Information; that such a definition is flawed even for asymptotically normal estimators was made clear by Hodges superefficient estimator (Problem 14.70). Le Cam (1956) introduced the *one-step maximum likelihood estimator*, which is based on a discretization trick coupled with a Newton–Raphson approximation. Such estimators satisfy (14.62) under weak assumptions and enjoy other optimality properties; for example, see Section 7.3 of Millar (1983). The notion of a *regular* estimator sequence introduced at the end of Section 14.4.1 plays an important role in the theory of efficient estimation and the Hajék-Inagaki Convolution Theorem; see Hajék (1970), Le Cam (1979), Beran (1999), Millar (1985), and van der Vaart (1988).

The asymptotic behavior of the likelihood ratio statistic was studied in Wilks (1938) and Chernoff (1954). Pearson's Chi-squared statistic was introduced in Pearson (1900) and the Rao score tests by Rao (1947). In fact, the Rao score test was actually introduced in the univariate case by Wald (1941b). One-sided score tests are studied in Silvapulle and Silvapulle (1995). In econometrics, score tests are more commonly known as Lagrange multiplier tests; see Silvey (1959) and Bera and Bilias (2001). The asymptotic equivalence of many of the classical tests is explored in Hall and Mathiason (1990). Methods based on integrated likelihoods are reviewed in Berger et al. (1999). Caveats about the finite-sample behavior of Rao and Wald tests are given in Le Cam (1990); also see Fears et al. (1996) and Pawitan (2000). The behavior of likelihood ratio tests under nonstandard conditions is studied in Vu and Zhou (1997). Extensions of likelihood methods to semiparametric and nonparametric models are developed in Murphy and van der Vaart (1997), Owen (1988, 2001), and Fan et al. (2001). Robust version of the Wald, likelihood, and score tests are given in Heritier and Ronchetti (1994).