Chapter 5

# ASYMPTOTIC APPROXIMATIONS

## 5.1 INTRODUCTION: THE MEANING AND USES OF ASYMPTOTICS

Despite the many simple examples we have dealt with, closed form computation of risks in terms of known functions or simple integrals is the exception rather than the rule. Even if the risk is computable for a specific $P$ by numerical integration in one dimension, the qualitative behavior of the risk as a function of parameter and sample size is hard to ascertain. Worse, computation even at a single point may involve high-dimensional integrals. In particular, consider a sample $X_1, \dots, X_n$ from a distribution $F$, our setting for this section and most of this chapter. If we want to estimate $\mu(F) \equiv E_F X_1$ and use $\bar{X}$ we can write,

$$MSE_F(\bar{X}) = \frac{\sigma^2(F)}{n}. \tag{5.1.1}$$

This is a highly informative formula, telling us exactly how the MSE behaves as a function of $n$, and calculable for any $F$ and all $n$ by a single one-dimensional integration. However, consider $\text{med}(X_1, \dots, X_n)$ as an estimate of the population median $\nu(F)$. If $n$ is odd, $\nu(F) = F^{-1}\left(\frac{1}{2}\right)$, and $F$ has density $f$ we can write

$$MSE_F(\text{med}(X_1, \dots, X_n)) = \int_{-\infty}^{\infty} \left(x - F^{-1}\left(\tfrac{1}{2}\right)\right)^2 g_n(x)dx \tag{5.1.2}$$

where, from Problem (B.2.13), if $n = 2k + 1$,

$$g_n(x) = n \left( \begin{array}{c} 2k \\ k \end{array} \right) F^k(x)(1 - F(x))^k f(x). \tag{5.1.3}$$

Evaluation here requires only evaluation of $F$ and a one-dimensional integration, but a different one for each $n$ (Problem 5.1.1). Worse, the qualitative behavior of the risk as a function of $n$ and simple parameters of $F$ is not discernible easily from (5.1.2) and (5.1.3). To go one step further, consider evaluation of the power function of the one-sided $t$ test of Chapter 4. If $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ we have seen in Section 4.9.2 that $\sqrt{n}\bar{X}/s$ has a noncentral $t$ distribution with parameter $\mu/\sigma$ and $n - 1$ degrees of freedom. This distribution may be evaluated by a two-dimensional integral using classical functions

(Problem 5.1.2) and its qualitative properties are reasonably transparent. But suppose $F$ is not Gaussian. It seems impossible to determine explicitly what happens to the power function because the distribution of $\sqrt{n}\bar{X}/s$ requires the joint distribution of $(\bar{X}, s)$ and in general this is only representable as an $n$-dimensional integral;

$$P\left[\sqrt{n}\frac{\bar{X}}{s} \le t\right] = \int_A f(x_1)\dots f(x_n)d\mathbf{x}$$

where

$$A = \left\{(x_1,\dots,x_n) : \sum_{i=1}^n x_i \le \frac{\sqrt{n}t}{n-1}\left(\sum_{i=1}^n x_i^2 - \frac{(\Sigma x_i)^2}{n}\right)\right\}.$$

There are two complementary approaches to these difficulties. The first, which occupies us for most of this chapter, is to approximate the risk function under study

$$R_n(F) \equiv E_F l(F, \delta(X_1,\dots,X_n)),$$

by a qualitatively simpler to understand and easier to compute function, $\widetilde{R}_n(F)$. The other, which we explore further in later chapters, is to use the Monte Carlo method. In its simplest form, Monte Carlo is described as follows. Draw $B$ independent "samples" of size $n$, $\{X_{1j},\dots,X_{nj}\}$, $1 \le j \le B$ from $F$ using a random number generator and an explicit form for $F$. Approximately evaluate $R_n(F)$ by

$$\widehat{R}_B = \frac{1}{B}\sum_{j=1}^B l(F, \delta(X_{1j},\dots,X_{nj})). \tag{5.1.4}$$

By the law of large numbers as $B \to \infty$, $\widehat{R}_B \xrightarrow{P} R_n(F)$. Thus, save for the possibility of a very unlikely event, just as in numerical integration, we can approximate $R_n(F)$ arbitrarily closely. We now turn to a detailed discussion of asymptotic approximations but will return to describe Monte Carlo and show how it complements asymptotics briefly in Example 5.3.3.

Asymptotics in statistics is usually thought of as the study of the limiting behavior of statistics or, more specifically, of distributions of statistics, based on observing $n$ i.i.d. observations $X_1,\dots,X_n$ as $n \to \infty$. We shall see later that the scope of asymptotics is much greater, but for the time being let's stick to this case as we have until now.

Asymptotics, in this context, always refers to a sequence of statistics

$$\{T_n(X_1,\dots,X_n)\}_{n\ge1},$$

for instance the sequence of means $\{\bar{X}_n\}_{n\ge1}$, where $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^n X_i$, or the sequence of medians, or it refers to the sequence of their distributions

$$\{\mathcal{L}_F(T_n(X_1,\dots,X_n))\}_{n\ge1}.$$

Asymptotic statements are always statements about *the sequence*. The classical examples are, $\bar{X}_n \xrightarrow{P} E_F(X_1)$ or

$$\mathcal{L}_F(\sqrt{n}(\bar{X}_n - E_F(X_1))) \to \mathcal{N}(0, \text{Var}_F(X_1)).$$

In theory these limits say nothing about any particular $T_n(X_1,\dots,X_n)$ but in practice we act as if they do because the $T_n(X_1,\dots,X_n)$ we consider are closely related as functions of $n$ so that we expect the limit to *approximate* $T_n(X_1,\dots,X_n)$ or $\mathcal{L}_F(T_n(X_1,\dots,X_n))$ (in an appropriate sense). For instance, the weak law of large numbers tells us that, if $E_F|X_1| < \infty$, then

$$\bar{X}_n \xrightarrow{P} \mu \equiv E_F(X_1). \tag{5.1.5}$$

That is, (see A.14.1)

$$P_F[|\bar{X}_n - \mu| \geq \epsilon] \to 0 \tag{5.1.6}$$

for all $\epsilon > 0$. We interpret this as saying that, for $n$ sufficiently large, $\bar{X}_n$ is approximately equal to its expectation. The trouble is that for any specified degree of approximation, say, $\epsilon = .01$, $(5.1.6)$ does not tell us how large $n$ has to be for the chance of the approximation not holding to this degree (the left-hand side of $(5.1.6)$) to fall, say, below .01. Is $n \geq 100$ enough or does it have to be $n \geq 100,000$? Similarly, the central limit theorem tells us that if $E_F|X_1^2| < \infty$, $\mu$ is as above and $\sigma^2 \equiv \text{Var}_F(X_1)$, then

$$P_F\left[\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \leq z\right] \to \Phi(z) \tag{5.1.7}$$

where $\Phi$ is the standard normal d.f.

As an approximation, this reads

$$P_F[\bar{X}_n \leq x] \approx \Phi\left(\sqrt{n}\frac{(x-\mu)}{\sigma}\right). \tag{5.1.8}$$

Again we are faced with the questions of how good the approximation is for given $n$, $x$, and $P_F$. What we in principle prefer are bounds, which are available in the classical situations of $(5.1.6)$ and $(5.1.7)$. Thus, by Chebychev's inequality, if $E_F X_1^2 < \infty$,

$$P_F[|\bar{X}_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}. \tag{5.1.9}$$

As a bound this is typically far too conservative. For instance, if $|X_1| \leq 1$, the much more delicate Hoeffding bound (B.9.6) gives

$$P_F[|\bar{X}_n - \mu| \geq \epsilon] \leq 2\exp\left\{-\tfrac{1}{2}n\epsilon^2\right\}. \tag{5.1.10}$$

Because $|X_1| \leq 1$ implies that $\sigma^2 \leq 1$ with $\sigma^2 = 1$ possible (Problem 5.1.3), the right-hand side of $(5.1.9)$ when $\sigma^2$ is unknown becomes $1/n\epsilon^2$. For $\epsilon = .1$, $n = 400$, $(5.1.9)$ is .25 whereas $(5.1.10)$ is .14.

Further qualitative features of these bounds and relations to approximation $(5.1.8)$ are given in Problem 5.1.4. Similarly, the celebrated Berry–Esséen bound (A.15.11) states that if $E_F|X_1|^3 < \infty$,

$$\sup_x \left| P_F\left[\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \leq x\right] - \Phi(x)\right| \leq C\frac{E_F|X_1|^3}{\sigma^3 n^{1/2}} \tag{5.1.11}$$

where $C$ is a universal constant known to be $\leq 33/4$. Although giving us some idea of how much $(5.1.8)$ differs from the truth, $(5.1.11)$ is again much too conservative generally.$^{(1)}$ The approximation $(5.1.8)$ is typically much better than $(5.1.11)$ suggests.

Bounds for the goodness of approximations have been available for $\bar{X}_n$ and its distribution to a much greater extent than for nonlinear statistics such as the median. Yet, as we have seen, even here they are not a very reliable guide. Practically one proceeds as follows:

(a) Asymptotic approximations are derived.

(b) Their validity for the given $n$ and $T_n$ for some plausible values of $F$ is tested by numerical integration if possible or Monte Carlo computation.

If the agreement is satisfactory we use the approximation even though the agreement for the true but unknown $F$ generating the data may not be as good.

Asymptotics has another important function beyond suggesting numerical approximations for specific $n$ and $F$. If they are simple, asymptotic formulae suggest qualitative properties that may hold even if the approximation itself is not adequate. For instance, $(5.1.7)$ says that the behavior of the distribution of $\bar{X}_n$ is for large $n$ governed (approximately) only by $\mu$ and $\sigma^2$ in a precise way, although the actual distribution depends on $P_F$ in a complicated way. It suggests that qualitatively the risk of $\bar{X}_n$ as an estimate of $\mu$, for any loss function of the form $l(F,d) = \lambda(|\mu - d|)$ where $\lambda(0) = 0$, $\lambda'(0) > 0$, behaves like $\lambda'(0)(\sigma/\sqrt{n})(\sqrt{2\pi})$ (Problem 5.1.5) and quite generally that risk increases with $\sigma$ and decreases with $n$, which is reasonable.

As we shall see, quite generally, good estimates $\widehat{\theta}_n$ of parameters $\theta(F)$ will behave like $\bar{X}_n$ does in relation to $\mu$. The estimates $\widehat{\theta}_n$ will be *consistent*, $\widehat{\theta}_n \xrightarrow{P} \theta(F)$, for all $F$ in the model, and *asymptotically normal*,

$$\mathcal{L}_F\left(\frac{\sqrt{n}[\widehat{\theta}_n - \theta(F)]}{\sigma(\theta, F)}\right) \to \mathcal{N}(0,1) \tag{5.1.12}$$

where $\sigma(\theta, F)$ typically is the standard deviation (SD) of $\sqrt{n}\widehat{\theta}_n$ or an approximation to this SD. Consistency will be pursued in Section 5.2 and asymptotic normality via the delta method in Section 5.3. The qualitative implications of results such as these are very important when we consider comparisons between competing procedures. Note that this feature of simple asymptotic approximations using the normal distribution is not replaceable by Monte Carlo.

We now turn to specifics. As we mentioned, Section 5.2 deals with consistency of various estimates including maximum likelihood. The arguments apply to vector-valued estimates of Euclidean parameters. In particular, consistency is proved for the estimates of canonical parameters in exponential families. Section 5.3 begins with asymptotic computation of moments and asymptotic normality of functions of a scalar mean and include as an application asymptotic normality of the maximum likelihood estimate for one-parameter exponential families. The methods are then extended to vector functions of vector means and applied to establish asymptotic normality of the MLE $\widehat{\boldsymbol{\eta}}$ of the canonical parameter $\boldsymbol{\eta}$

in exponential families among other results. Section 5.4 deals with optimality results for likelihood-based procedures in one-dimensional parameter models. Finally in Section 5.5 we examine the asymptotic behavior of Bayes procedures. The notation we shall use in the rest of this chapter conforms closely to that introduced in Sections A.14, A.15, and B.7. We will recall relevant definitions from that appendix as we need them, but we shall use results we need from A.14, A.15, and B.7 without further discussion.

**Summary.** Asymptotic statements refer to the behavior of sequences of procedures as the sequence index tends to $\infty$. In practice, asymptotics are methods of approximating risks, distributions, and other statistical quantities that are not realistically computable in closed form, by quantities that can be so computed. Most asymptotic theory we consider leads to approximations that in the i.i.d. case become increasingly valid as the sample size increases. We also introduce Monte Carlo methods and discuss the interaction of asymptotics, Monte Carlo, and probability bounds.

## 5.2 CONSISTENCY

### 5.2.1 Plug-In Estimates and MLEs in Exponential Family Models

Suppose that we have a sample $X_1, \ldots, X_n$ from $P_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \Theta$ and want to estimate a real or vector $q(\boldsymbol{\theta})$. The least we can ask of our estimate $\widehat{q}_n(X_1, \ldots, X_n)$ is that as $n \to \infty$, $\widehat{q}_n \overset{P_{\boldsymbol{\theta}}}{\to} q(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. That is, in accordance with (A.14.1) and (B.7.1), for all $\boldsymbol{\theta} \in \Theta$, $\epsilon > 0$,

$$P_{\boldsymbol{\theta}}[|\widehat{q}_n(X_1, \ldots, X_n) - q(\boldsymbol{\theta})| \geq \epsilon] \to 0. \tag{5.2.1}$$

where $|\cdot|$ denotes Euclidean distance. A stronger requirement is

$$\sup_{\boldsymbol{\theta}} \left\{ P_{\boldsymbol{\theta}}\left[|\widehat{q}_n(X_1, \ldots, X_n) - q(\boldsymbol{\theta})| \geq \epsilon\right] : \boldsymbol{\theta} \in \Theta \right\} \to 0. \tag{5.2.2}$$

Bounds $b(n, \epsilon)$ for $\sup_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}\left[|\widehat{q}_n - q(\boldsymbol{\theta})| \geq \epsilon\right]$ that yield (5.2.2) are preferable and we shall indicate some of qualitative interest when we can. But, with all the caveats of Section 5.1, (5.2.1), which is called *consistency* of $\widehat{q}_n$ and can be thought of as 0'th order asymptotics, remains central to all asymptotic theory. The stronger statement (5.2.2) is called *uniform consistency*. If $\Theta$ is replaced by a smaller set $K$, we talk of uniform consistency over $K$.

**Example 5.2.1.** *Means.* The simplest example of consistency is that of the mean. If $X_1, \ldots, X_n$ are i.i.d. $P$ where $P$ is unknown but $E_P|X_1| < \infty$ then, by the WLLN,

$$\bar{X} \overset{P}{\to} \mu(P) \equiv E(X_1)$$

and $\mu(\widehat{P}) = \bar{X}$, where $\widehat{P}$ is the empirical distribution, is a consistent estimate of $\mu(P)$. For $\mathcal{P}$ this large it is not uniformly consistent. (See Problem 5.2.2.) However, if, for instance, $\mathcal{P} \equiv \{P : E_P X_1^2 \leq M < \infty\}$, then $\bar{X}$ is uniformly consistent over $\mathcal{P}$ because by Chebyshev's inequality, for all $P \in \mathcal{P}$,

$$P[|\bar{X} - \mu(P)| \geq \epsilon]| \leq \frac{\text{Var}(\bar{X})}{\epsilon^2} \leq \frac{M}{n\epsilon^2}$$

□

**Example 5.2.2.** *Binomial Variance.* Let $X_1, \dots, X_n$ be the indicators of binomial trials with $P[X_1 = 1] = p$. Then $N = \sum X_i$ has a $\mathcal{B}(n, p)$ distribution, $0 \leq p \leq 1$, and $\widehat{p} = \bar{X} = N/n$ is a uniformly consistent estimate of $p$. But further, consider the plug-in estimate $\widehat{p}(1-\widehat{p})/n$ of the variance of $\widehat{p}$, which is $\frac{1}{n}q(\widehat{p})$, where $q(p) = p(1-p)$. Evidently, by A.14.6, $q(\widehat{p})$ is consistent. Other moments of $X_1$ can be consistently estimated in the same way. □

To some extent the plug-in method was justified by consistency considerations and it is not surprising that consistency holds quite generally for frequency plug-in estimates.

**Theorem 5.2.1.** *Suppose that $\mathcal{P} = \mathcal{S} = \{(p_1, \dots, p_k) : 0 \leq p_j \leq 1, 1 \leq j \leq k, \sum_{j=1}^k p_j = 1\}$, the k-dimensional simplex, where $p_j = P[X_1 = x_j]$, $1 \leq j \leq k$, and $\{x_1, \dots, x_k\}$ is the range of $X_1$. Let $N_j \equiv \sum_{i=1}^n 1(X_i = x_j)$ and $\widehat{p}_j \equiv N_j/n$, $\mathbf{p}_n \equiv (\widehat{p}_1, \dots, \widehat{p}_k) \in \mathcal{S}$ be the empirical distribution. Suppose that $q : \mathcal{S} \to R^p$ is continuous. Then $\widehat{q}_n \equiv q(\widehat{\mathbf{p}}_n)$ is a uniformly consistent estimate of $q(\mathbf{p})$.*

***Proof.*** By the weak law of large numbers for all $\mathbf{p}$, $\delta > 0$

$$P_{\mathbf{p}}\left[|\widehat{\mathbf{p}}_n - \mathbf{p}| \geq \delta\right] \to 0.$$

Because $q$ is continuous and $\mathcal{S}$ is compact, it is uniformly continuous on $\mathcal{S}$. Thus, for every $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that $\mathbf{p}, \mathbf{p}' \in \mathcal{S}$, $|\mathbf{p}' - \mathbf{p}| \leq \delta(\epsilon)$, implies $|q(\mathbf{p}') - q(\mathbf{p})| \leq \epsilon$. Then

$$P_{\mathbf{p}}[|\widehat{q}_n - q| \geq \epsilon] \leq P_{\mathbf{p}}[|\widehat{\mathbf{p}}_n - \mathbf{p}| \geq \delta(\epsilon)]$$

But, $\sup\{P_{\mathbf{p}}[|\widehat{\mathbf{p}}_n - \mathbf{p}| \geq \delta] : \mathbf{p} \in \mathcal{S}\} \leq k/4n\delta^2$ (Problem 5.2.1) and the result follows. □

In fact, in this case, we can go further. Suppose the *modulus of continuity* of $q$, $\omega(q, \delta)$ is defined by

$$\omega(q, \delta) = \sup\{|q(\mathbf{p}) - q(\mathbf{p}')| : |\mathbf{p} - \mathbf{p}'| \leq \delta\}. \tag{5.2.3}$$

Evidently, $\omega(q, \cdot)$ is increasing in $\delta$ and has the range $[a, b)$ say. If $q$ is continuous $\omega(q, \delta) \downarrow 0$ as $\delta \downarrow 0$. Let $\omega^{-1} : [a, b] \leq R^+$ be defined as the inverse of $\omega$,

$$\omega^{-1}(\epsilon) = \inf\{\delta : \omega(q, \delta) \geq \epsilon\} \tag{5.2.4}$$

It easily follows (Problem 5.2.3) that

$$\sup\{P[|\widehat{q}_n - q(\mathbf{p})| \geq \epsilon] : P \in \mathcal{P}\} \leq n^{-1}[\omega^{-1}(\epsilon)]^{-2}\frac{k}{4}. \tag{5.2.5}$$

A simple and important result for the case in which $X_1, \dots, X_n$ are i.i.d. with $X_i \in \mathcal{X}$ is the following:

**Proposition 5.2.1.** *Let* $\mathbf{g} \equiv (g_1, \dots, g_d)$ *map* $\mathcal{X}$ *onto* $\mathcal{Y} \subset R^d$. *Suppose* $E_\theta|g_j(X_1)| < \infty$, $1 \le j \le d$, *for all* $\theta$; *let* $m_j(\theta) \equiv E_\theta g_j(X_1)$, $1 \le j \le d$, *and let* $q(\theta) = h(\mathbf{m}(\theta))$, *where* $h : \mathcal{Y} \to R^p$. *Then, if* $h$ *is continuous,*

$$\widehat{q} \equiv h(\bar{\mathbf{g}}) \equiv h\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(X_i)\right)$$

*is a consistent estimate of* $q(\theta)$. *More generally if* $\nu(P) = h(E_P\mathbf{g}(X_1))$ *and* $\mathcal{P} = \{P : E_P|\mathbf{g}(X_1)| < \infty\}$, *then* $\nu(\widehat{P}) \equiv h(\bar{\mathbf{g}})$, *where* $\widehat{P}$ *is the empirical distribution, is consistent for* $\nu(P)$.

***Proof.*** We need only apply the general weak law of large numbers (for vectors) to conclude that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(X_i) \xrightarrow{P} E_P\mathbf{g}(X_1) \tag{5.2.6}$$

if $E_p|\mathbf{g}(X_1)| < \infty$. For consistency of $h(\bar{\mathbf{g}})$ apply Proposition B.7.1: $\mathbf{U}_n \xrightarrow{P} \mathbf{U}$ implies that $h(\mathbf{U}_n) \xrightarrow{P} h(\mathbf{U})$ for all continuous $h$. □

**Example 5.2.3.** *Variances and Correlations.* Let $X_i = (U_i, V_i)$, $1 \le i \le n$ be i.i.d. $\mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, $\sigma_i^2 > 0$, $|\rho| < 1$. Let $\mathbf{g}(u, v) = (u, v, u^2, v^2, uv)$ so that $\sum_{i=1}^n \mathbf{g}(U_i, V_i)$ is the statistic generating this 5-parameter exponential family. If we let $\boldsymbol{\theta} \equiv (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then

$$\mathbf{m}(\boldsymbol{\theta}) = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \rho\sigma_1\sigma_2 + \mu_1\mu_2).$$

If $h = \mathbf{m}^{-1}$, then

$$\begin{aligned} &h(m_1, \dots, m_5) \\ &\quad = (m_1, m_2, m_3 - m_1^2, m_4 - m_2^2, (m_5 - m_1m_2)(m_3 - m_1^2)^{-1/2}(m_4 - m_2^2)^{-1/2}), \end{aligned}$$

which is well defined and continuous at all points of the range of $\mathbf{m}$. We may, thus, conclude by Proposition 5.2.1 that the empirical means, variances, and correlation coefficient are all consistent. Questions of uniform consistency and consistency when $\mathcal{P}$ = { Distributions such that $EU_1^2 < \infty$, $EV_1^2 < \infty$, $\text{Var}(U_1) > 0$, $\text{Var}(V_1) > 0$, $|\,\text{Corr}(U_1, V_1)| < 1$ } are discussed in Problem 5.2.4. □

Here is a general consequence of Proposition 5.2.1 and Theorem 2.3.1.

**Theorem 5.2.2.** *Suppose* $\mathcal{P}$ *is a canonical exponential family of rank* $d$ *generated by* $\mathbf{T}$. *Let* $\boldsymbol{\eta}, \mathcal{E}$ *and* $A(\cdot)$ *correspond to* $\mathcal{P}$ *as in Section* 1.6. *Suppose* $\mathcal{E}$ *is open. Then, if* $X_1, \dots, X_n$ *are a sample from* $P_{\boldsymbol{\eta}} \in \mathcal{P}$,

*(i)* $P_{\boldsymbol{\eta}}$*[The MLE* $\widehat{\boldsymbol{\eta}}$ *exists]* $\to 1$.

*(ii)* $\widehat{\boldsymbol{\eta}}$ *is consistent.*

***Proof.*** Recall from Corollary 2.3.1 to Theorem 2.3.1 that $\widehat{\boldsymbol{\eta}}(X_1, \dots, X_n)$ exists iff $\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i) = \overline{\mathbf{T}}_n$ belongs to the interior $C_{\mathbf{T}}^{\circ}$ of the convex support of the distribution of $\overline{\mathbf{T}}_n$. Note that, if $\boldsymbol{\eta}_0$ is true, $E_{\boldsymbol{\eta}_0}(\mathbf{T}(X_1))$ must by Theorem 2.3.1 belong to the interior of the convex support because the equation $\dot{A}(\boldsymbol{\eta}) = \mathbf{t}_0$, where $\mathbf{t}_0 = \dot{A}(\boldsymbol{\eta}_0) = E_{\boldsymbol{\eta}_0}\mathbf{T}(X_1)$, is solved by $\boldsymbol{\eta}_0$. By definition of the interior of the convex support there exists a ball $S_\delta \equiv \{\mathbf{t} : |\mathbf{t} - E_{\boldsymbol{\eta}_0}\mathbf{T}(X_1)| < \delta\} \subset C_{\mathbf{T}}^{\circ}$. By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i) \overset{P_{\boldsymbol{\eta}_0}}{\to} E_{\boldsymbol{\eta}_0}\mathbf{T}(X_1).$$

Hence,

$$P_{\boldsymbol{\eta}_0}[\frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i) \in C_{\mathbf{T}}^{\circ}] \to 1. \tag{5.2.7}$$

But $\widehat{\boldsymbol{\eta}}$, which solves

$$\dot{A}(\boldsymbol{\eta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{T}(X_i),$$

exists iff the event in (5.2.7) occurs and (i) follows. We showed in Theorem 2.3.1 that on $C_{\mathbf{T}}^{\circ}$ the map $\boldsymbol{\eta} \to \dot{A}(\boldsymbol{\eta})$ is 1-1 and continuous on $\mathcal{E}$. By a classical result, see, for example, Rudin (1987), the inverse $\dot{A}^{-1} : \dot{A}(\mathcal{E}) \to \mathcal{E}$ is continuous on $S_\delta$ and the result follows from Proposition 5.2.1. □

## 5.2.2 Consistency of Minimum Contrast Estimates

The argument of the the previous subsection in which a minimum contrast estimate, the MLE, is a continuous function of a mean of i.i.d. vectors evidently used exponential family properties. A more general argument is given in the following simple theorem whose conditions are hard to check. Let $X_1, \dots, X_n$ be i.i.d. $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset R^d$. Let $\widehat{\boldsymbol{\theta}}$ be a minimum contrast estimate that minimizes

$$\rho_n(\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \rho(X_i, \boldsymbol{\theta})$$

where, as usual, $D(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}_0}\rho(X_1, \boldsymbol{\theta})$ is uniquely minimized at $\boldsymbol{\theta}_0$ for all $\boldsymbol{\theta}_0 \in \Theta$.

**Theorem 5.2.3.** *Suppose*

$$\sup\{|\frac{1}{n}\sum_{i=1}^n [\rho(X_i, \boldsymbol{\theta}) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta})]| : \boldsymbol{\theta} \in \Theta\} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0 \tag{5.2.8}$$

*and*

$$\inf\{D(\boldsymbol{\theta}_0, \boldsymbol{\theta}) : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon\} > D(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \quad \textit{for every } \epsilon > 0. \tag{5.2.9}$$

*Then $\widehat{\boldsymbol{\theta}}$ is consistent.*

***Proof.*** Note that,

$$P_{\boldsymbol{\theta}_0}[|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \geq \epsilon] \leq P_{\boldsymbol{\theta}_0}[\inf\{\frac{1}{n}\sum_{i=1}^{n}[\rho(X_i, \boldsymbol{\theta}) - \rho(X_i, \boldsymbol{\theta}_0)] : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon\} \leq 0] \tag{5.2.10}$$

By hypothesis, for all $\delta > 0$,

$$\begin{aligned} P_{\boldsymbol{\theta}_0}[\inf\{\frac{1}{n}\sum_{i=1}^{n}(\rho(X_i, \boldsymbol{\theta}) - \rho(X_i, \boldsymbol{\theta}_0)) : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon\} \\ - \inf\{D(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)) : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon\} < -\delta] \to 0 \end{aligned} \tag{5.2.11}$$

because the event in (5.2.11) implies that

$$\sup\{|\frac{1}{n}\sum_{i=1}^{n}[\rho(X_i, \boldsymbol{\theta}) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta})]| : \boldsymbol{\theta} \in \Theta\} > \frac{\delta}{2}, \tag{5.2.12}$$

which has probability tending to 0 by (5.2.8). But for $\epsilon > 0$ let

$$\delta = \frac{1}{4}\inf\{D(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \epsilon\}.$$

Then (5.2.11) implies that the right-hand side of (5.2.10) tends to 0. □

A simple and important special case is given by the following.

**Corollary 5.2.1.** *If $\Theta$ is finite, $\Theta = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d\}$, $E_{\boldsymbol{\theta}_0}|\log p(X_1, \boldsymbol{\theta})| < \infty$ and the parameterization is identifiable, then, if $\widehat{\boldsymbol{\theta}}$ is the MLE, $P_{\boldsymbol{\theta}_j}[\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_j] \to 0$ for all $j$.*

***Proof.*** Note that for some $\epsilon > 0$,

$$P_{\boldsymbol{\theta}_0}[\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_j] = P_{\boldsymbol{\theta}_0}[|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| \geq \epsilon]. \tag{5.2.13}$$

By Shannon's Lemma 2.2.1 we need only check that (5.2.8) and (5.2.9) hold for $\rho(x, \boldsymbol{\theta}) = \log p(x, \boldsymbol{\theta})$. But because $\Theta$ is finite, (5.2.8) follows from the WLLN and

$$\begin{aligned} &P_{\boldsymbol{\theta}_0}[\max\{|\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\rho(X_i, \boldsymbol{\theta}_j) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta}_j)) : 1 \leq j \leq d\} \geq \epsilon] \\ &\leq d\max\{P_{\boldsymbol{\theta}_0}[|\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\rho(X_i, \boldsymbol{\theta}_j) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta}_j))| \geq \epsilon] : 1 \leq j \leq d\} \to 0. \end{aligned}$$

and (5.2.9) follows from Shannon's lemma. □

Condition (5.2.8) can often fail—see Problem 5.2.5. An alternative condition that is readily seen to work more widely is the replacement of (5.2.8) by

(i) For all compact $K \subset \Theta$,

$$\sup\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(\rho(X_i, \boldsymbol{\theta}) - D(\boldsymbol{\theta}_0, \boldsymbol{\theta})) : \boldsymbol{\theta} \in K\right|\right\} \xrightarrow{P_{\boldsymbol{\theta}_0}} 0.$$

(ii) For some compact $K \subset \Theta$, (5.2.14)

$$P_{\boldsymbol{\theta}_0}\left[\inf\left\{\frac{1}{n}\sum_{i=1}^{n}(\rho(X_i, \boldsymbol{\theta}) - \rho(X_i, \boldsymbol{\theta}_0)) : \boldsymbol{\theta} \in K^c\right\} > 0\right] \to 1.$$

We shall see examples in which this modification works in the problems. Unfortunately checking conditions such as (5.2.8) and (5.2.14) is in general difficult. A general approach due to Wald and a similar approach for consistency of generalized estimating equation solutions are left to the problems. When the observations are independent but not identically distributed, consistency of the MLE may fail if the number of parameters tends to infinity, see Problem 5.3.33.

**Summary.** We introduce the minimal property we require of any estimate (strictly speaking, sequence of estimates), consistency. If $\widehat{\theta}_n$ is an estimate of $\theta(P)$, we require that $\widehat{\theta}_n \overset{P}{\to} \theta(P)$ as $n \to \infty$. Uniform consistency for $\mathcal{P}$ requires more, that $\sup\{P[|\widehat{\theta}_n - \theta(P)| \geq \epsilon] : P \in \mathcal{P}\} \to 0$ for all $\epsilon > 0$. We show how consistency holds for continuous functions of vector means as a consequence of the law of large numbers and derive consistency of the MLE in canonical multiparameter exponential families. We conclude by studying consistency of the MLE and more generally MC estimates in the case $\Theta$ finite and $\Theta$ Euclidean. Sufficient conditions are explored in the problems.

## 5.3 FIRST- AND HIGHER-ORDER ASYMPTOTICS: THE DELTA METHOD WITH APPLICATIONS

We have argued in Section 5.1 that the principal use of asymptotics is to provide quantitatively or qualitatively useful approximations to risk.

### 5.3.1 The Delta Method for Moments

We begin this section by deriving approximations to moments of smooth functions of scalar means and even provide crude bounds on the remainders. We then sketch the extension to functions of vector means.

As usual let $X_1, \dots, X_n$ be i.i.d. $\mathcal{X}$ valued and for the moment take $\mathcal{X} = R$. Let $h : R \to R$, let $\|g\|_\infty = \sup\{|g(t)| : t \in R\}$ denote the sup norm, and assume

(i) (a) $h$ is $m$ times differentiable on $R$, $m \geq 2$. We denote the $j$th derivative of $h$ by $h^{(j)}$ and assume

(b) $||h^{(m)}||_\infty \equiv \sup_{\mathcal{X}} |h^{(m)}(x)| \leq M < \infty$

(ii) $E|X_1|^m < \infty$

Let $E(X_1) = \mu$, $\text{Var}(X_1) = \sigma^2$. We have the following.

**Theorem 5.3.1.** *If (i) and (ii) hold, then*

$$Eh(\bar{X}) = h(\mu) + \sum_{j=1}^{m-1} \frac{h^{(j)}(\mu)}{j!} E(\bar{X} - \mu)^j + R_m \tag{5.3.1}$$

*where, for some $C_m > 0$,*

$$|R_m| \leq C_m \frac{E|X_1|^m}{m!} n^{-m/2}.$$

The proof is an immediate consequence of Taylor's expansion,

$$h(\bar{X}) = h(\mu) + \sum_{k=1}^{m-1} \frac{h^{(k)}(\mu)}{k!}(\bar{X}-\mu)^k + \frac{h^{(m)}(\bar{X}^*)}{m!}(\bar{X}-\mu)^m \tag{5.3.2}$$

where $|\bar{X}^* - \mu| \le |\bar{X} - \mu|$, and the following lemma.

**Lemma 5.3.1.** *If* $E|X_1|^j < \infty$, $j \ge 2$, *then there are constants* $C_j > 0$ *and* $D_j > 0$ *such that*

$$E|\bar{X}-\mu|^j \le C_j E|X_1|^j n^{-j/2} \tag{5.3.3}$$

$$|E(\bar{X}-\mu)^j| \le D_j E|X_1|^j n^{-(j+1)/2}, \ j \textit{ odd.} \tag{5.3.4}$$

Note that for $j$ even, $E|\bar{X}-\mu|^j = E(\bar{X}-\mu)^j$.

***Proof.*** We give the proof of (5.3.4) for all $j$ and (5.3.3) for $j$ even. The more difficult argument needed for (5.3.3) and $j$ odd is given in Problem 5.3.2.

Let $\mu = E(X_1) = 0$, then

(a)
$$\begin{aligned} E(\bar{X}^j) &= n^{-j} E(\textstyle\sum_{i=1}^n X_i)^j \\ &= n^{-j} \textstyle\sum_{1 \le i_1, \ldots, i_j \le n} E(X_{i_1} \ldots X_{i_j}) \end{aligned}$$

But $E(X_{i_1} \ldots X_{i_j}) = 0$ unless each integer that appears among $\{i_1, \ldots, i_j\}$ appears at least twice. Moreover,

(b)
$$\sup_{i_1, \ldots, i_j} |E(X_{i_1} \ldots X_{i_j})| = E|X_1|^j$$

by Problem 5.3.5, so the number $d$ of nonzero terms in (a) is

(c)
$$\sum_{r=1}^{[j/2]} \frac{n}{r} \sum_{\substack{i_1+\ldots+i_r=j \\ i_k \ge 2 \text{ all } k}} \frac{j}{i_1, \ldots, i_r}$$

where $\frac{n}{i_1, \ldots, i_r} = \frac{n!}{i_1! \ldots i_r!}$ and $[t]$ denotes the greatest integer $\le t$. The expression in (c) is, for $j \le n/2$, bounded by

(d)
$$\frac{C_j}{[\frac{j}{2}]!} n(n-1) \ldots (n - [j/2] + 1)$$

where

$$C_j = \max_{1 \le r \le [j/2]} \left\{ \sum \left\{ \frac{j}{i_1, \ldots, i_r} : i_1 + \ldots + i_r = j, i_k \ge 2, 1 \le k \le r \right\} \right\}.$$

But

$$\text{(e)} \qquad n^{-j}n(n-1)\dots(n-[j/2]+1) \leq n^{[j/2]-j}$$

and (c), (d), and (e) applied to (a) imply (5.3.4) for $j$ odd and $(5.3.3)$ for $j$ even, if $\mu = 0$. In general by considering $X_i - \mu$ as our basic variables we obtain the lemma but with $E|X_1|^j$ replaced by $E|X_1 - \mu|^j$. By Problem 5.3.6, $E|X_1 - \mu|^j \leq 2^j E|X_1|^j$ and the lemma follows. □

The two most important corollaries of Theorem 5.3.1, respectively, give approximations to the bias of $h(\bar{X})$ as an estimate of $h(\mu)$ and its variance and MSE.

**Corollary 5.3.1.**

*(a) If $E|X_1|^3 < \infty$ and $||h^{(3)}||_\infty < \infty$, then*

$$Eh(\bar{X}) = h(\mu) + \frac{h^{(2)}(\mu)\sigma^2}{2n} + O(n^{-3/2}). \tag{5.3.5}$$

*(b) If $E(X_1^4) < \infty$ and $||h^{(4)}||_\infty < \infty$ then $O(n^{-3/2})$ in (5.3.5) can be replaced by $O(n^{-2})$.*

***Proof.*** For (5.3.5) apply Theorem 5.3.1 with $m = 3$. Because $E(\bar{X} - \mu)^2 = \sigma^2/n$, (5.3.5) follows. If the conditions of (b) hold, apply Theorem 5.3.1 with $m = 4$. Then $R_m = O(n^{-2})$ and also $E(\bar{X} - \mu)^3 = O(n^{-2})$ by (5.3.4). □

**Corollary 5.3.2.** *If*

*(a) $||h^{(j)}||_\infty < \infty$, $0 \leq j \leq 3$ and $E|X_1|^3 < \infty$, then*

$$\operatorname{Var} h(\bar{X}) = \frac{\sigma^2[h^{(1)}(\mu)]^2}{n} + O(n^{-3/2}) \tag{5.3.6}$$

*(b) If $||h^{(j)}||_\infty < \infty$, $0 \leq j \leq 3$, and $EX_1^4 < \infty$, then $O(n^{-3/2})$ in (5.3.6) can be replaced by $O(n^{-2})$.*

***Proof.*** (a) Write

$$\begin{aligned} Eh^2(\bar{X}) &= h^2(\mu) + 2h(\mu)h^{(1)}(\mu)E(\bar{X} - \mu) + \{h^{(2)}(\mu)h(\mu) + [h^{(1)}]^2(\mu)\}E(\bar{X} - \mu)^2 \\ &\quad + \frac{1}{6}E[h^2]^{(3)}(\bar{X}^*)(\bar{X} - \mu)^3 \\ &= h^2(\mu) + \{h^{(2)}(\mu)h(\mu) + [h^{(1)}]^2(\mu)\}\frac{\sigma^2}{n} + O(n^{-3/2}). \end{aligned}$$

(b) Next, using Corollary 5.3.1,

$$\begin{aligned} [Eh(\bar{X})]^2 &= (h(\mu) + \frac{h^{(2)}(\mu)}{2}\frac{\sigma^2}{n} + O(n^{-3/2}))^2 \\ &= h^2(\mu) + h(\mu)h^{(2)}(\mu)\frac{\sigma^2}{n} + O(n^{-3/2}). \end{aligned}$$

Subtracting (a) from (b) we get (5.3.6). To get part (b) we need to expand $Eh^2(\bar{X})$ to four terms and similarly apply the appropriate form of (5.3.5). □

Clearly the statements of the corollaries as well can be turned to expansions as in Theorem 5.3.1 with bounds on the remainders.

Note an important qualitative feature revealed by these approximations. If $h(\bar{X})$ is viewed, as we normally would, as the plug-in estimate of the parameter $h(\mu)$ then, for large $n$, the bias of $h(\bar{X})$ defined by $Eh(\bar{X}) - h(\mu)$ is $O(n^{-1})$, which is neglible compared to the standard deviation of $h(\bar{X})$, which is $O(n^{-1/2})$ unless $h^{(1)}(\mu) = 0$. A qualitatively simple explanation of this important phenonemon will be given in Theorem 5.3.3.

**Example 5.3.1.** If $X_1, \dots, X_n$ are i.i.d. $\mathcal{E}(\lambda)$ the MLE of $\lambda$ is $\bar{X}^{-1}$. If the $X_i$ represent the lifetimes of independent pieces of equipment in hundreds of hours and the warranty replacement period is (say) 200 hours, then we may be interested in the warranty failure probability

$$P_\lambda[X_1 \le 2] = 1 - e^{-2\lambda}. \tag{5.3.7}$$

If $h(t) = 1 - \exp(-2/t)$, then $h(\bar{X})$ is the MLE of $1 - \exp(-2\lambda) = h(\mu)$, where $\mu = E_\lambda X_1 = 1/\lambda$.

We can use the two corollaries to compute asymptotic approximations to the means and variance of $h(\bar{X})$. Thus, by Corollary 5.3.1,

$$\begin{aligned} \text{Bias}_\lambda(h(\bar{X})) &\equiv E_\lambda(h(\bar{X}) - h(\mu)) \\ &= \tfrac{h^{(2)}(\mu)}{2}\tfrac{\sigma^2}{n} + O(n^{-2}) \\ &= 2e^{-2\lambda}\lambda(1-\lambda)/n + O(n^{-2}) \end{aligned} \tag{5.3.8}$$

because $h^{(2)}(t) = 4(t^{-3} - t^{-4})\exp(-2/t)$, $\sigma^2 = 1/\lambda^2$, and, by Corollary 5.3.2 (Problem 5.3.1)

$$\text{Var}_\lambda\, h(\bar{X}) = 4\lambda^2 e^{-4\lambda}/n + O(n^{-2}). \tag{5.3.9}$$

□

Further expansion can be done to increase precision of the approximation to Var $h(\bar{X})$ for large $n$. Thus, by expanding $Eh^2(\bar{X})$ and $Eh(\bar{X})$ to six terms we obtain the approximation

$$\begin{aligned} \text{Var}(h(\bar{X})) = &\tfrac{1}{n}[h^{(1)}(\mu)]^2\sigma^2 + \tfrac{1}{n^2}\left\{h^{(1)}(\mu)h^{(2)}(\mu)\mu_3\right. \\ &\left.+ \tfrac{1}{2}[h^{(2)}(\mu)]^2\sigma^4\right\} + R'_n \end{aligned} \tag{5.3.10}$$

with $R'_n$ tending to zero at the rate $1/n^3$. Here $\mu_k$ denotes the $k$th central moment of $X_i$ and we have used the facts that (see Problem 5.3.3)

$$E(\bar{X} - \mu)^3 = \frac{\mu_3}{n^2}, \quad E(\bar{X} - \mu)^4 = \frac{\mu_4}{n^3} + \frac{3(n-1)\sigma^4}{n^3}. \tag{5.3.11}$$

**Example 5.3.2.** *Bias and Variance of the MLE of the Binomial Variance.* We will compare $E(h(\bar{X}))$ and Var $h(\bar{X})$ with their approximations, when $h(t) = t(1-t)$ and $X_i \sim$

$\mathcal{B}(1, p)$, and will illustrate how accurate (5.3.10) is in a situation in which the approximation can be checked.

First calculate

$$\begin{aligned} Eh(\bar{X}) &= E(\bar{X}) - E(\bar{X}^2) = p - [\mathrm{Var}(\bar{X}) + (E(\bar{X}))^2] \\ &= p(1-p) - \frac{1}{n}p(1-p) = \frac{n-1}{n}p(1-p). \end{aligned}$$

Because $h^{(1)}(t) = 1 - 2t$, $h^{(2)} = -2$, (5.3.5) yields

$$E(h(\bar{X})) = p(1-p) - \frac{1}{n}p(1-p)$$

and in this case (5.3.5) is exact as it should be. Next compute

$$\mathrm{Var}\, h(\bar{X}) = \frac{p(1-p)}{n}\left\{(1-2p)^2 + \frac{2p(1-p)}{n-1}\right\}\left(\frac{n-1}{n}\right)^2.$$

Because $\mu_3 = p(1-p)(1-2p)$, (5.3.10) yields

$$\begin{aligned} \mathrm{Var}\, h(\bar{X}) &= \frac{1}{n}(1-2p)^2 p(1-p) + \frac{1}{n^2}\{-2(1-2p)p(1-p)(1-2p) \\ &\quad +2p^2(1-p)^2\} + R_n' \\ &= \frac{p(1-p)}{n}\{(1-2p)^2 + \frac{1}{n}[2p(1-p) - 2(1-2p)^2]\} + R_n'. \end{aligned}$$

Thus, the error of approximation is

$$\begin{aligned} R_n' &= \frac{p(1-p)}{n^3}[(1-2p)^2 - 2p(1-p)] \\ &= \frac{p(1-p)}{n^3}[1 - 6p(1-p)] = O(n^{-3}). \end{aligned}$$

□

The generalization of this approach to approximation of moments for functions of vector means is formally the same but computationally not much used for $d$ larger than 2.

**Theorem 5.3.2.** *Suppose* $\mathbf{g} : \mathcal{X} \to R^d$ *and let* $\mathbf{Y}_i = \mathbf{g}(X_i) = (g_1(X_i), \dots, g_d(X_i))^T$, $i = 1, \dots, n$, *where* $X_1, \dots, X_n$ *are i.i.d. Let* $h : R^d \to R$, *assume that* $h$ *has continuous partial derivatives of order up to* $m$, *and that*

*(i)* $||D^m(h)||_\infty < \infty$ *where* $D^m h(\mathbf{x})$ *is the array (tensor)*

$$\left\{\frac{\partial^m h}{\partial x_1^{i_1} \dots \partial x_d^{i_d}}(\mathbf{x}) : i_1 + \dots + i_d = m, 0 \le i_j \le m, 1 \le j \le d\right\}$$

*and* $||D^m h||_\infty$ *is the sup over all* $x$ *and* $i_1, \dots, i_d$ *of* $|\frac{\partial^m h}{\partial x_1^{i_1} \dots \partial x_d^{i_d}}(\mathbf{x})|$.

*(ii)* $E|Y_{ij}|^m < \infty$, $1 \le i \le n$, $1 \le j \le d$, *where* $Y_{ij} \equiv g_j(X_i)$.

*Then, if* $\bar{Y}_k = \frac{1}{n}\sum_{i=1}^n Y_{ik}$, $\bar{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i$, *and* $\boldsymbol{\mu} = E\mathbf{Y}_1$, *then*

$$\begin{aligned} Eh(\bar{\mathbf{Y}}) &= h(\boldsymbol{\mu}) + \textstyle\sum_{j=1}^{m-1} \sum \{ \frac{\partial^m h}{\partial x_1^{i_1} \ldots \partial x_d^{i_d}}(\boldsymbol{\mu})(i_1! \ldots i_d!)^{-1} \\ & \qquad E \textstyle\prod_{k=1}^d (\bar{Y}_k - \mu_k)^{i_k} : i_1 + \ldots + i_d = j, 0 \le i_k \le j\} + O(n^{-m/2}). \end{aligned} \tag{5.3.12}$$

This is a consequence of Taylor's expansion in $d$ variables, B.8.11, and the appropriate generalization of Lemma 5.3.1. The proof is outlined in Problem 5.3.4. The most interesting application, as for the case $d = 1$, is to $m = 3$. We get, for $d = 2$, $E|Y_1|^3 < \infty$

$$\begin{aligned} Eh(\bar{\mathbf{Y}}) &= h(\boldsymbol{\mu}) + \frac{1}{n}\left\{ \frac{1}{2}\frac{\partial^2 h}{\partial x_1^2}(\boldsymbol{\mu}) \operatorname{Var}(Y_{11}) + \frac{\partial^2 h}{\partial x_1 \partial x_2}(\boldsymbol{\mu}) \operatorname{Cov}(Y_{11}, Y_{12}) \right. \\ &+ \left. \frac{1}{2}\frac{\partial^2 h}{\partial x_2^2}(\boldsymbol{\mu}) \operatorname{Var}(Y_{12}) \right\} + O(n^{-3/2}). \end{aligned} \tag{5.3.13}$$

Moreover, by (5.3.3), if $E|\mathbf{Y}_1|^4 < \infty$, then $O(n^{-3/2})$ in (5.3.13) can be replaced by $O(n^{-2})$. Similarly, under appropriate conditions (Problem 5.3.12), for $d = 2$,

$$\begin{aligned} \operatorname{Var} h(\bar{\mathbf{Y}}) &= \frac{1}{n}\left[ \left( \frac{\partial h}{\partial x_1}(\boldsymbol{\mu}) \right)^2 \operatorname{Var}(Y_{11}) + 2\frac{\partial h}{\partial x_1}(\boldsymbol{\mu})\frac{\partial h}{\partial x_2}(\boldsymbol{\mu}) \operatorname{Cov}(Y_{11}, Y_{12}) \right. \\ &+ \left. \left( \frac{\partial h}{\partial x_2}(\boldsymbol{\mu}) \right)^2 \operatorname{Var}(Y_{12}) \right] + O(n^{-2}) \end{aligned} \tag{5.3.14}$$

Approximations (5.3.5), (5.3.6), (5.3.13), and (5.3.14) do not help us to approximate risks for loss functions other than quadratic (or some power of $(d - h(\mu))$). The results in the next subsection go much further and "explain" the form of the approximations we already have.

## 5.3.2 The Delta Method for In Law Approximations

As usual we begin with $d = 1$.

**Theorem 5.3.3.** *Suppose that* $\mathcal{X} = R$, $h : R \to R$, $EX_1^2 < \infty$ *and* $h$ *is differentiable at* $\mu = E(X_1)$. *Then*

$$\mathcal{L}(\sqrt{n}(h(\bar{X}) - h(\mu))) \to \mathcal{N}(0, \sigma^2(h)) \tag{5.3.15}$$

*where*

$$\sigma^2(h) = [h^{(1)}(\mu)]^2 \sigma^2$$

*and* $\sigma^2 = \operatorname{Var}(X_1)$.

The result follows from the more generally useful lemma.

**Lemma 5.3.2.** *Suppose* $\{U_n\}$ *are real random variables and that for a sequence* $\{a_n\}$ *of constants with* $a_n \to \infty$ *as* $n \to \infty$, *(i)* $a_n(U_n - u) \xrightarrow{\mathcal{L}} V$ *for some constant* $u$.

*(ii)* $g : R \to R$ *is differentiable at* $u$ *with derivative* $g^{(1)}(u)$.

*Then*

$$a_n(g(U_n) - g(u)) \xrightarrow{\mathcal{L}} g^{(1)}(u)V. \tag{5.3.16}$$

***Proof.*** By definition of the derivative, for every $\epsilon > 0$ there exists a $\delta > 0$ such that

(a) $$|v - u| \le \delta \Rightarrow |g(v) - g(u) - g^{(1)}(u)(v - u)| \le \epsilon|v - u|$$

Note that (i) $\Rightarrow$

(b) $$a_n(U_n - u) = O_p(1)$$

$\Rightarrow$

(c) $$U_n - u = O_p(a_n^{-1}) = o_p(1).$$

Using (c), for every $\delta > 0$

(d) $$P[|U_n - u| \le \delta] \to 1$$

and, hence, from (a), for every $\epsilon > 0$,

(e) $$P[|g(U_n) - g(u) - g^{(1)}(u)(U_n - u)| \le \epsilon|U_n - u|] \to 1.$$

But (e) implies

(f) $$a_n[g(U_n) - g(u) - g^{(1)}(u)(U_n - u)] = o_p(a_n(U_n - u)) = o_p(1)$$

from (b). Therefore,

(g) $$a_n[g(U_n) - g(u)] = g^{(1)}(u)a_n(U_n - u) + o_p(1).$$

But, by hypothesis, $a_n(U_n - u) \xrightarrow{\mathcal{L}} V$ and the result follows. □

The theorem follows from the central limit theorem letting $U_n = \bar{X}$, $a_n = n^{1/2}$, $u = \mu$, $V \sim \mathcal{N}(0, \sigma^2)$. □

Note that (5.3.15) "explains" Lemma 5.3.1. Formally we expect that if $V_n \xrightarrow{\mathcal{L}} V$, then $EV_n^j \to EV^j$ (although this need not be true, see Problems 5.3.32 and B.7.8). Consider $V_n = \sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathcal{L}} V \sim \mathcal{N}(0, \sigma^2)$. Thus, we expect

$$E[\sqrt{n}(\bar{X} - \mu)]^j \to \sigma^j EZ^j \tag{5.3.17}$$

where $Z \sim \mathcal{N}(0,1)$. But if $j$ is even, $EZ^j > 0$, else $EZ^j = 0$. Then (5.3.17) yields

$$\begin{aligned} E(\bar{X}-\mu)^j = O(\sigma^j EZ^j n^{-j/2}) &= O(n^{-j/2}),\ j \text{ even} \\ &= o(n^{-j/2}), j \text{ odd}. \end{aligned}$$

**Example 5.3.3.** *"$t$" Statistics.*

*(a) The One-Sample Case.* Let $X_1, \dots, X_n$ be i.i.d. $F \in \mathcal{F}$ where $E_F(X_1) = \mu$, $\text{Var}_F(X_1) = \sigma^2 < \infty$. A statistic for testing the hypothesis $H : \mu = 0$ versus $K : \mu > 0$ is

$$T_n = \sqrt{n}\frac{\bar{X}}{s}$$

where

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

If $\mathcal{F} = \{\text{Gaussian distributions}\}$, we can obtain the critical value $t_{n-1}(1-\alpha)$ for $T_n$ from the $\mathcal{T}_{n-1}$ distribution. In general we claim that if $F \in \mathcal{F}$ and $H$ is true, then

$$T_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{5.3.18}$$

In particular this implies not only that $t_{n-1}(1-\alpha) \to z_{1-\alpha}$ but that the $t_{n-1}(1-\alpha)$ critical value (or $z_{1-\alpha}$) is approximately correct if $H$ is true and $F$ is not Gaussian. For the proof note that

$$U_n \equiv \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

by the central limit theorem, and

$$s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2\right) \xrightarrow{P} \sigma^2$$

by Proposition 5.2.1 and Slutsky's theorem. Now Slutsky's theorem yields (5.3.18) because $T_n = U_n/(s_n/\sigma) = g(U_n, s_n/\sigma)$, where $g(u,v) = u/v$.

*(b) The Two-Sample Case.* Let $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ be two independent samples with $\mu_1 = E(X_1)$, $\sigma_1^2 = \text{Var}(X_1)$, $\mu_2 = E(Y_1)$ and $\sigma_2^2 = \text{Var}(Y_1)$. Consider testing $H : \mu_1 = \mu_2$ versus $K : \mu_2 > \mu_1$. In Example 4.9.3 we saw that the two sample $t$ statistic

$$S_n = \sqrt{\frac{n_1 n_2}{n}}\left(\frac{\bar{Y}-\bar{X}}{s}\right),\ n = n_1 + n_2$$

has a $\mathcal{T}_{n-2}$ distribution under $H$ when the $X$'s and $Y$'s are normal with $\sigma_1^2 = \sigma_2^2$. Using the central limit theorem, Slutsky's theorem, and the foregoing arguments, we find (Problem 5.3.28) that if $n_1/n \to \lambda$, $0 < \lambda < 1$, then

$$S_n \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{(1-\lambda)\sigma_1^2 + \lambda\sigma_2^2}{\lambda\sigma_1^2 + (1-\lambda)\sigma_2^2}\right).$$

It follows that if $n_1 = n_2$ or $\sigma_1^2 = \sigma_2^2$, then the critical value $t_{n-2}(1-\alpha)$ for $S_n$ is approximately correct if $H$ is true and the $X$'s and $Y$'s are not normal.

### Monte Carlo Simulation

As mentioned in Section 5.1, approximations based on asymptotic results should be checked by Monte Carlo simulations. We illustrate such simulations for the preceding $t$ tests by generating data from the $\chi_d^2$ distribution $M$ times independently, each time computing the value of the $t$ statistics and then giving the proportion of times out of $M$ that the $t$ statistics exceed the critical values from the $t$ table. Here we use the $\chi_d^2$ distribution because for small to moderate $d$ it is quite different from the normal distribution. Other distributions should also be tried. Figure 5.3.1 shows that for the one-sample $t$ test, when $\alpha = 0.05$, the asymptotic result gives a good approximation when $n \geq 10^{1.5} \cong 32$, and the true distribution $F$ is $\chi_d^2$ with $d \geq 10$. The $\chi_2^2$ distribution is extremely skew, and in this case the $t_{n-1}(0.95)$ approximation is only good for $n \geq 10^{2.5} \cong 316$.

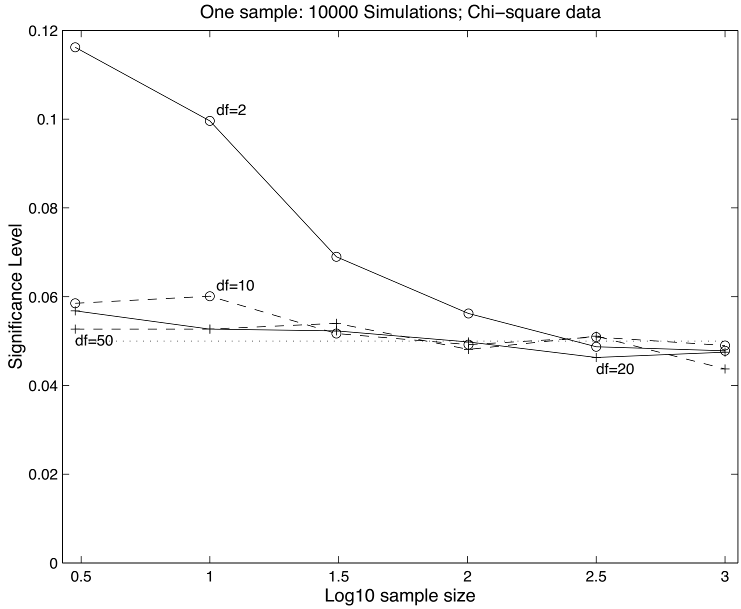

**Figure 5.3.1.** Each plotted point represents the results of 10,000 one-sample $t$ tests using $\chi_d^2$ data, where $d$ is either 2, 10, 20, or 50, as indicated in the plot. The simulations are repeated for different sample sizes and the observed significance levels are plotted.

For the two-sample $t$ tests, Figure 5.3.2 shows that when $\sigma_1^2 = \sigma_2^2$ and $n_1 = n_2$, the $t_{n-2}(1-\alpha)$ critical value is a very good approximation even for small $n$ and for $X, Y \sim \chi_2^2$.

This is because, in this case, $\bar{Y} - \bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1}(Y_i - X_i)$, and $Y_i - X_i$ have a symmetric distribution. Other Monte Carlo runs (not shown) with $\sigma_1^2 \neq \sigma_2^2$ show that as long as $n_1 = n_2$, the $t_{n-2}(0.95)$ approximation is good for $n_1 \geq 100$, even when the $X$'s and $Y$'s have different $\chi_2^2$ distributions, scaled to have the same means, and $\sigma_2^2 = 12\sigma_1^2$. Moreover, the $t_{n-2}(1-\alpha)$ approximation is good when $n_1 \neq n_2$ and $\sigma_1^2 = \sigma_2^2$. However, as we see from the limiting law of $S_n$ and Figure 5.3.3, when both $n_1 \neq n_2$ and $\sigma_1^2 \neq \sigma_2^2$, then the two-sample $t$ tests with critical region $1\{S_n \geq t_{n-2}(1-\alpha)\}$ do not have approximate level $\alpha$. In this case Monte Carlo studies have shown that the test in Section 4.9.4 based on Welch's approximation works well.

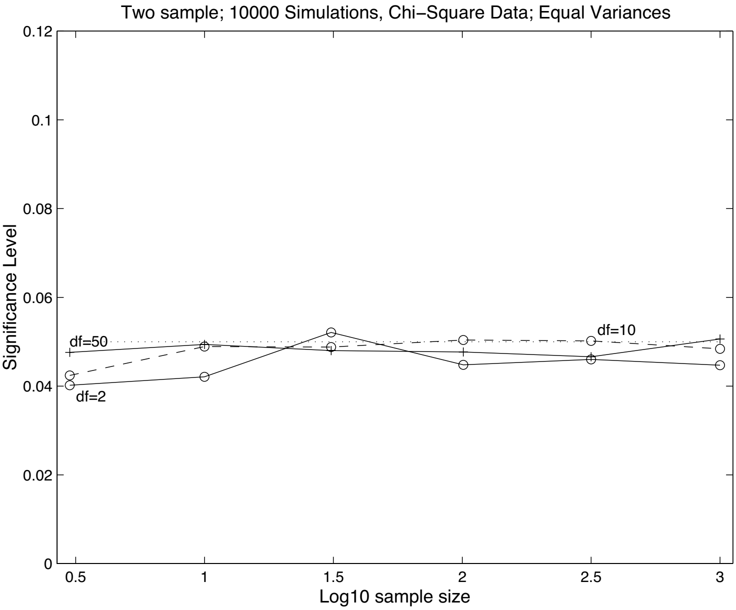

**Figure 5.3.2.** Each plotted point represents the results of 10,000 two-sample $t$ tests. For each simulation the two samples are the same size (the size indicated on the $x$-axis), $\sigma_1^2 = \sigma_2^2$, and the data are $\chi_d^2$ where $d$ is one of 2, 10, or 50.

□

Next, in the one-sample situation, let $h(\bar{X})$ be an estimate of $h(\mu)$ where $h$ is continuously differentiable at $\mu$, $h^{(1)}(\mu) \neq 0$. By Theorem 5.3.3, $\sqrt{n}[h(\bar{X}) - h(\mu)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2[h^{(1)}(\mu)]^2)$. To test the hypothesis $H : h(\mu) = h_0$ versus $K : h(\mu) > h_0$ the natural test statistic is

$$T_n = \frac{\sqrt{n}[h(\bar{X}) - h_0]}{s|h^{(1)}(\bar{X})|}.$$

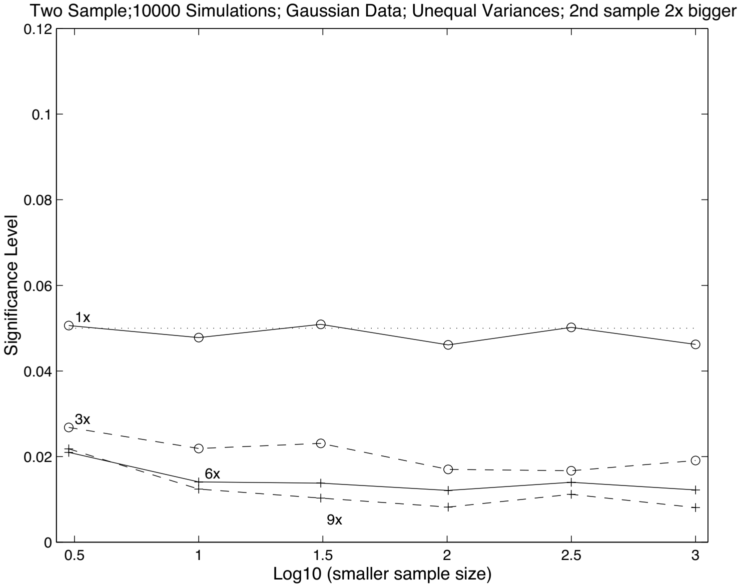

**Figure 5.3.3.** Each plotted point represents the results of 10,000 two-sample $t$ tests. For each simulation the two samples differ in size: The second sample is two times the size of the first. The $x$-axis denotes the size of the smaller of the two samples. The data in the first sample are $\mathcal{N}(0,1)$ and in the second they are $\mathcal{N}(0,\sigma^2)$ where $\sigma^2$ takes on the values 1, 3, 6, and 9, as indicated in the plot.

Combining Theorem 5.3.3 and Slutsky's theorem, we see that here, too, if $H$ is true

$$T_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

so that $z_{1-\alpha}$ is the asymptotic critical value.

## Variance Stabilizing Transformations

**Example 5.3.4.** In Appendices A and B we encounter several important families of distributions, such as the binomial, Poisson, gamma, and beta, which are indexed by one or more parameters. If we take a sample from a member of one of these families, then the sample mean $\bar{X}$ will be approximately normally distributed with variance $\sigma^2/n$ depending on the parameters indexing the family considered. We have seen that smooth transformations $h(\bar{X})$ are also approximately normally distributed. It turns out to be useful to know transformations $h$, called *variance stabilizing*, such that $\operatorname{Var} h(\bar{X})$ is approximately independent of the parameters indexing the family we are considering. From (5.3.6) and

(5.3.13) we see that a first approximation to the variance of $h(\bar{X})$ is $\sigma^2[h^{(1)}(\mu)]^2/n$. Thus, finding a variance stabilizing transformation is equivalent to finding a function $h$ such that

$$\sigma^2[h^{(1)}(\mu)]^2 \equiv c$$

for all $\mu$ and $\sigma$ appropriate to our family. Such a function can usually be found if $\sigma$ depends only on $\mu$, which varies freely. In this case (5.3.19) is an ordinary differential equation. As an example, suppose that $X_1, \dots, X_n$ is a sample from a $\mathcal{P}(\lambda)$ family. In this case $\sigma^2 = \lambda$ and $\text{Var}(\bar{X}) = \lambda/n$. To have $\text{Var}\, h(\bar{X})$ approximately constant in $\lambda$, $h$ must satisfy the differential equation $[h^{(1)}(\lambda)]^2\lambda = c > 0$ for some arbitrary $c > 0$. If we require that $h$ is increasing, this leads to $h^{(1)}(\lambda) = \sqrt{c}/\sqrt{\lambda}$, $\lambda > 0$, which has as its solution $h(\lambda) = 2\sqrt{c\lambda} + d$, where $d$ is arbitrary. Thus, $h(t) = \sqrt{t}$ is a variance stabilizing transformation of $\bar{X}$ for the Poisson family of distributions. Substituting in (5.3.6) we find $\text{Var}(\bar{X})^{\frac{1}{2}} \cong 1/4n$ and $\sqrt{n}((\bar{X})^{\frac{1}{2}} - (\lambda)^{\frac{1}{2}})$ has approximately a $\mathcal{N}(0, 1/4)$ distribution. □

One application of variance stabilizing transformations, by their definition, is to exhibit monotone functions of parameters of interest for which we can give fixed length (independent of the data) confidence intervals. Thus, in the preceding $\mathcal{P}(\lambda)$ case,

$$\sqrt{\bar{X}} \pm \frac{2z(1 - \frac{1}{2}\alpha)}{\sqrt{n}}$$

is an approximate $1 - \alpha$ confidence interval for $\sqrt{\lambda}$. A second application occurs for models where the families of distribution for which variance stabilizing transformations exist are used as building blocks of larger models. Major examples are the generalized linear models of Section 6.5. The comparative roles of variance stabilizing and canonical transformations as link functions are discussed in Volume II. Some further examples of variance stabilizing transformations are given in the problems.

The notion of such transformations can be extended to the following situation. Suppose, $\widehat{\gamma}_n(X_1, \dots, X_n)$ is an estimate of a real parameter $\gamma$ indexing a family of distributions from which $X_1, \dots, X_n$ are an i.i.d. sample. Suppose further that

$$\mathcal{L}_\gamma(\sqrt{n}(\widehat{\gamma}_n - \gamma)) \to \mathcal{N}(0, \sigma^2(\gamma)).$$

Then again, a variance stabilizing transformation $h$ is such that

$$\mathcal{L}_\gamma\left[\sqrt{n}(h(\widehat{\gamma}_n) - h(\gamma))\right] \to \mathcal{N}(0, c) \qquad (5.3.19)$$

for all $\gamma$. See Example 5.3.6. Also closely related but different are so-called normalizing transformations. See Problems 5.3.15 and 5.3.16.

## Edgeworth Approximations

The normal approximation to the distribution of $\bar{X}$ utilizes only the first two moments of $\bar{X}$. Under general conditions (Bhattacharya and Rao, 1976, p. 538) one can improve on the normal approximation by utilizing the third and fourth moments. Let $F_n$ denote the distribution of $T_n = \sqrt{n}(\bar{X} - \mu)/\sigma$ and let $\gamma_{1n}$ and $\gamma_{2n}$ denote the coefficient of skewness and kurtosis of $T_n$. Then under some conditions,$^{(1)}$

$$F_n(x) = \Phi(x) - \varphi(x)[\frac{1}{6}\gamma_{1n}H_2(x) + \frac{1}{24}\gamma_{2n}H_3(x) + \frac{1}{72}\gamma_{1n}^2 H_5(x)] + r_n \qquad (5.3.20)$$

where $r_n$ tends to zero at a rate faster than $1/n$ and $H_2$, $H_3$, and $H_5$ are *Hermite polynomials* defined by

$$H_2(x) = x^2 - 1, H_3(x) = x^3 - 3x, H_5(x) = x^5 - 10x^3 + 15x. \qquad (5.3.21)$$

The expansion (5.3.20) is called the *Edgeworth expansion* for $F_n$.

**Example 5.3.5.** *Edgeworth Approximations to the $\chi^2$ Distribution.* Suppose $V \sim \chi^2_n$. According to Theorem B.3.1, $V$ has the same distribution as $\sum_{i=1}^n X_i^2$, where the $X_i$ are independent and $X_i \sim \mathcal{N}(0,1)$, $i = 1, \dots, n$. It follows from the central limit theorem that $T_n = (\sum_{i=1}^n X_i^2 - n)/\sqrt{2n} = (V - n)/\sqrt{2n}$ has approximately a $\mathcal{N}(0,1)$ distribution. To improve on this approximation, we need only compute $\gamma_{1n}$ and $\gamma_{2n}$. We can use Problem B.2.4 to compute

$$\gamma_{1n} = \frac{E(V-n)^3}{(2n)^{\frac{3}{2}}} = \frac{2\sqrt{2}}{\sqrt{n}}, \gamma_{2n} = \frac{E(V-n)^4}{(2n)^2} - 3 = \frac{12}{n}.$$

Therefore,

$$F_n(x) = \Phi(x) - \varphi(x)\left[\frac{\sqrt{2}}{3\sqrt{n}}(x^2 - 1) + \frac{1}{2n}(x^3 - 3x) + \frac{1}{9n}(x^5 - 10x^3 + 15x)\right] + r_n.$$

Table 5.3.1 gives this approximation together with the exact distribution and the normal approximation when $n = 10$. □

| $x$ | -2.04 | -1.95 | -1.91 | -1.75 | -1.66 | -1.51 | -1.35 | |
|---|---|---|---|---|---|---|---|---|
| Exact | 0.0001 | 0.0005 | 0.0010 | 0.0050 | 0.0100 | 0.0250 | 0.0500 | |
| EA | 0 | 0 | 0 | 0.0032 | 0.0105 | 0.0287 | 0.0553 | |
| NA | 0.0208 | 0.0254 | 0.0284 | 0.0397 | 0.0481 | 0.0655 | 0.0877 | |
| $x$ | -1.15 | -0.85 | -0.61 | -0.38 | -0.15 | 0.11 | 0.40 | 0.77 |
| Exact | 0.1000 | 0.2000 | 0.3000 | 0.4000 | 0.5000 | 0.6000 | 0.7000 | 0.8000 |
| EA | 0.1051 | 0.2024 | 0.3006 | 0.4000 | 0.4999 | 0.5999 | 0.6999 | 0.8008 |
| NA | 0.1254 | 0.1964 | 0.2706 | 0.3513 | 0.4415 | 0.5421 | 0.6548 | 0.7792 |
| $x$ | 1.34 | 1.86 | 2.34 | 2.95 | 3.40 | 4.38 | 4.79 | 5.72 |
| Exact | 0.9000 | 0.9500 | 0.9750 | 0.9900 | 0.9950 | 0.9990 | 0.9995 | 0.9999 |
| EA | 0.9029 | 0.9506 | 0.9724 | 0.9876 | 0.9943 | 0.9996 | 0.9999 | 1.0000 |
| NA | 0.9097 | 0.9684 | 0.9905 | 0.9984 | 0.9997 | 1.0000 | 1.0000 | 1.0000 |

**TABLE 5.3.1.** Edgeworth$^{(2)}$ and normal approximations EA and NA to the $\chi^2_{10}$ distribution, $P(T_n \leq x)$, where $T_n$ is a standardized $\chi^2_{10}$ random variable.

### The Multivariate Case

Lemma $5.3.2$ extends to the $d$-variate case.

**Lemma 5.3.3.** *Suppose* $\{\mathbf{U}_n\}$ *are d-dimensional random vectors and that for some sequence of constants* $\{a_n\}$ *with* $a_n \to \infty$ *as* $n \to \infty$,

(i) $a_n(\mathbf{U}_n - \mathbf{u}) \xrightarrow{\mathcal{L}} \mathbf{V}_{d\times 1}$ *for some* $d \times 1$ *vector of constants* $\mathbf{u}$.

(ii) $\mathbf{g} : R^d \to R^p$ *has a differential* $\mathbf{g}^{(1)}_{p\times d}(\mathbf{u})$ *at* $\mathbf{u}$. *Then*

$$a_n[\mathbf{g}(\mathbf{U}_n) - \mathbf{g}(\mathbf{u})] \xrightarrow{\mathcal{L}} \mathbf{g}^{(1)}(\mathbf{u})\mathbf{V}.$$

***Proof.*** The proof follows from the arguments of the proof of Lemma $5.3.2$.

**Example 5.3.6.** Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. as $(X, Y)$ where $0 < EX^4 < \infty$, $0 < EY^4 < \infty$. Let $\rho^2 = \text{Cov}^2(X, Y)/\sigma_1^2\sigma_2^2$ where $\sigma_1^2 = \text{Var}\, X$, $\sigma_2^2 = \text{Var}\, Y$; and let $r^2 = \widehat{C}^2/\widehat{\sigma}_1^2\widehat{\sigma}_2^2$ where

$$\widehat{C} = n^{-1}\Sigma(X_i - \bar{X})(Y_i - \bar{Y}),\ \widehat{\sigma}_1^2 = n^{-1}\Sigma(X_i - \bar{X})^2,\ \widehat{\sigma}_2^2 = n^{-1}\Sigma(Y_i - \bar{Y})^2.$$

Recall from Section $4.9.5$ that in the bivariate normal case the sample correlation coefficient $r$ is the MLE of the population correlation coefficient $\rho$ and that the likelihood ratio test of $H : \rho = 0$ is based on $|r|$. We can write $r^2 = g(\widehat{C}, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2) : R^3 \to R$, where $g(u_1, u_2, u_3) = u_1^2/u_2u_3$. Because of the location and scale invariance of $\rho$ and $r$, we can use the transformations $\widetilde{X}_i = (X_i - \mu_1)/\sigma_1$, and $\widetilde{Y}_j = (Y_j - \mu_2)/\sigma_2$ to conclude that without loss of generality we may assume $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$, $\rho = E(XY)$. Using the central limit and Slutsky's theorems, we can show (Problem $5.3.9$) that $\sqrt{n}(\widehat{C} - \rho)$, $\sqrt{n}(\widehat{\sigma}_1^2 - 1)$ and $\sqrt{n}(\widehat{\sigma}_2^2 - 1)$ jointly have the same asymptotic distribution as $\sqrt{n}(\mathbf{U}_n - u)$ where

$$\mathbf{U}_n = (n^{-1}\Sigma X_iY_i, n^{-1}\Sigma X_i^2, n^{-1}\Sigma Y_i^2)$$

and $\mathbf{u} = (\rho, 1, 1)$. Let $\tau_{k,j}^2 = \text{Var}(\widetilde{X}^k\widetilde{Y}^j)$ and $\lambda_{k,j,m,l} = \text{Cov}(\widetilde{X}^k\widetilde{Y}^j, \widetilde{X}^m\widetilde{Y}^l)$, then by the central limit theorem

$$\sqrt{n}(\mathbf{U} - \mathbf{u}) \to \mathcal{N}(0, \Sigma),\ \Sigma = \begin{pmatrix} \tau_{1,1}^2 & \lambda_{1,1,2,0} & \lambda_{1,1,0,2} \\ \lambda_{1,1,2,0} & \tau_{2,0}^2 & \lambda_{2,0,2,0} \\ \lambda_{1,1,0,2} & \lambda_{2,0,2,0} & \tau_{0,2}^2 \end{pmatrix}. \tag{5.3.22}$$

Next we compute

$$g^{(1)}(\mathbf{u}) = (2u_1/u_2u_3, -u_1^2/u_2^2u_3, -u_1^2/u_2u_3^2) = (2\rho, -\rho^2, -\rho^2).$$

It follows from Lemma $5.3.3$ and (B.5.6) that $\sqrt{n}(r^2 - \rho^2)$ is asymptotically normal, $\mathcal{N}(0, \sigma_0^2)$, with

$$\begin{aligned} \sigma_0^2 = g^{(1)}(u)\Sigma[g^{(1)}(u)]^T \quad &= \quad 4\rho^2\tau_{11}^2 + \rho^4\tau_{20}^2 + \rho^4\tau_{02}^2 \\ &\quad\ \ +2\{-2\rho^3\lambda_{1,1,2,0} - 2\rho^3\lambda_{1,1,0,2} + \rho^4\lambda_{2,0,2,0}\}. \end{aligned}$$

When $(X, Y) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then $\sigma_0^2 = 4\rho^2(1-\rho^2)^2$, and (Problem 5.3.9) $\sqrt{n}(r - \rho) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1-\rho^2)^2)$.

Referring to (5.3.19), we see (Problem 5.3.10) that in the bivariate normal case a variance stabilizing transformation $h(r)$ with $\sqrt{n}[h(r) - h(\rho)] \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ is achieved by choosing

$$h(\rho) = \frac{1}{2} \log \left( \frac{1+\rho}{1-\rho} \right).$$

The approximation based on this transformation, which is called *Fisher's $z$*, has been studied extensively and it has been shown (e.g., David, 1938) that

$$\mathcal{L}(\sqrt{n-3}(h(r) - h(\rho))$$

is closely approximated by the $\mathcal{N}(0,1)$ distribution, that is,

$$P(r \le c) \approx \Phi(\sqrt{n-3}[h(c) - h(\rho)]), \; c \in (-1, 1).$$

This expression provides approximations to the critical value of tests of $H : \rho = 0$, it gives approximations to the power of these tests, and it provides the approximate $100(1-\alpha)\%$ confidence interval of fixed length,

$$\rho = \tanh \left\{ h(r) \pm z \left(1 - \tfrac{1}{2}\alpha\right) / \sqrt{n-3} \right\}$$

where tanh is the hyperbolic tangent. □

Here is an extension of Theorem 5.3.3.

**Theorem 5.3.4.** *Suppose $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are independent identically distributed $d$ vectors with $E|\mathbf{Y}_1|^2 < \infty$, $E\mathbf{Y}_1 = \mathbf{m}$,* $\mathrm{Var}\, \mathbf{Y}_1 = \mathbf{\Sigma}$ *and $\mathbf{h} : \mathcal{O} \to R^p$ where $\mathcal{O}$ is an open subset of $R^d$, $\mathbf{h} = (h_1, \dots, h_p)$ and $\mathbf{h}$ has a total differential $\mathbf{h}^{(1)}(\mathbf{m}) = \left\| \frac{\partial h_i}{\partial x_j}(\mathbf{m}) \right\|_{p \times d}$. Then*

$$\mathbf{h}(\bar{\mathbf{Y}}) = \mathbf{h}(\mathbf{m}) + \mathbf{h}^{(1)}(\mathbf{m})(\bar{\mathbf{Y}} - \mathbf{m}) + o_p(n^{-1/2}) \tag{5.3.23}$$

$$\sqrt{n}[\mathbf{h}(\bar{\mathbf{Y}}) - \mathbf{h}(\mathbf{m})] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{h}^{(1)}(\mathbf{m})\mathbf{\Sigma}[\mathbf{h}^{(1)}(\mathbf{m})]^T) \tag{5.3.24}$$

***Proof.*** Argue as before using B.8.5

(a) $$\mathbf{h}(\mathbf{y}) = \mathbf{h}(\mathbf{m}) + \mathbf{h}^{(1)}(\mathbf{m})(\mathbf{y} - \mathbf{m}) + o(|\mathbf{y} - \mathbf{m}|)$$

and

(b) $$\sqrt{n}(\bar{\mathbf{Y}} - \mathbf{m}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{\Sigma})$$

so that

(c) $$\sqrt{n}(\mathbf{h}(\bar{\mathbf{Y}}) - \mathbf{h}(\mathbf{m})) = \sqrt{n}\mathbf{h}^{(1)}(\mathbf{m})(\bar{\mathbf{Y}} - \mathbf{m}) + o_p(1).$$

□

**Example 5.3.7.** *$\chi_1^2$ and Normal Approximation to the Distribution of $\mathcal{F}$ Statistics.* Suppose that $X_1, \dots, X_n$ is a sample from a $\mathcal{N}(0,1)$ distribution. Then according to Corollary B.3.1, the $\mathcal{F}$ statistic

$$T_{k,m} = \frac{(1/k)\sum_{i=1}^{k} X_i^2}{(1/m)\sum_{i=k+1}^{k+m} X_i^2} \tag{5.3.25}$$

has an $\mathcal{F}_{k,m}$ distribution, where $k + m = n$. Suppose that $n \geq 60$ so that Table IV cannot be used for the distribution of $T_{k,m}$. When $k$ is fixed and $m$ (or equivalently $n = k + m$) is large, we can use Slutsky's theorem (A.14.9) to find an approximation to the distribution of $T_{k,m}$. To show this, we first note that $(1/m)\sum_{i=k+1}^{k+m} X_i^2$ is the average of $m$ independent $\chi_1^2$ random variables. By Theorem B.3.1, the mean of a $\chi_1^2$ variable is $E(Z^2)$, where $Z \sim \mathcal{N}(0,1)$. But $E(Z^2) = \mathrm{Var}(Z) = 1$. Now the weak law of large numbers (A.15.7) implies that as $m \to \infty$,

$$\frac{1}{m}\sum_{i=k+1}^{k+m} X_i^2 \xrightarrow{P} 1.$$

Using the (b) part of Slutsky's theorem, we conclude that for fixed $k$,

$$T_{k,m} \xrightarrow{\mathcal{L}} \frac{1}{k}\sum_{i=1}^{k} X_i^2$$

as $m \to \infty$. By Theorem B.3.1, $\sum_{i=1}^{k} X_i^2$ has a $\chi_k^2$ distribution. Thus, when the number of degrees of freedom in the denominator is large, the $\mathcal{F}_{k,m}$ distribution can be approximated by the distribution of $V/k$, where $V \sim \chi_k^2$.

To get an idea of the accuracy of this approximation, check the entries of Table IV against the last row. This row, which is labeled $m = \infty$, gives the quantiles of the distribution of $V/k$. For instance, if $k = 5$ and $m = 60$, then $P[T_{5,60} \leq 2.37] = P[(V/k) \leq 2.21] = 0.05$ and the respective $0.05$ quantiles are $2.37$ for the $\mathcal{F}_{5,60}$ distribution and $2.21$ for the distribution of $V/k$. See also Figure B.3.1 in which the density of $V/k$, when $k = 10$, is given as the $\mathcal{F}_{10,\infty}$ density.

Next we turn to the normal approximation to the distribution of $T_{k,m}$. Suppose for simplicity that $k = m$ and $k \to \infty$. We write $T_k$ for $T_{k,k}$. The case $m = \lambda k$ for some $\lambda > 0$ is left to the problems. We do not require the $X_i$ to be normal, only that they be i.i.d. with $EX_1 = 0$, $EX_1^2 > 0$ and $EX_1^4 < \infty$. Then, if $\sigma^2 = \mathrm{Var}(X_1)$, we can write,

$$T_k = \frac{1}{k}\sum_{i=1}^{k} Y_{i1} \bigg/ \frac{1}{k}\sum_{i=1}^{k} Y_{i2} \tag{5.3.26}$$

where $Y_{i1} = X_i^2/\sigma^2$ and $Y_{i2} = X_{k+i}^2/\sigma^2$, $i = 1, \dots, k$. Equivalently $T_k = h(\bar{\mathbf{Y}})$ where $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^T$, $E(\mathbf{Y}_i) = (1,1)^T$ and $h(u,v) = \frac{u}{v}$. By Theorem 5.3.4,

$$\sqrt{n}(T_k - 1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, h^{(1)}(\mathbf{1})\boldsymbol{\Sigma}[h^{(1)}(\mathbf{1})]^T) \tag{5.3.27}$$

where $\mathbf{1} = (1,1)^T$, $h^{(1)}(u,v) = (\frac{1}{v}, -\frac{u}{v^2})^T$ and $\boldsymbol{\Sigma} = \operatorname{Var}(Y_{11})\mathbf{J}$, where $\mathbf{J}$ is the $2 \times 2$ identity. We conclude that

$$\sqrt{n}(T_k - 1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\operatorname{Var}(Y_{11})).$$

In particular if $X_1 \sim \mathcal{N}(0, \sigma^2)$, as $k \to \infty$,

$$\sqrt{n}(T_k - 1) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4).$$

In general, when $\min\{k, m\} \to \infty$, the distribution of $\sqrt{\frac{mk}{m+k}}(T_{k,m} - 1)$ can be approximated by a $\mathcal{N}(0,2)$ distribution. Thus (Problem 5.3.7), when $X_i \sim \mathcal{N}(0, \sigma^2)$,

$$\begin{aligned} P[T_{k,m} \le t] &= P[\sqrt{\tfrac{mk}{m+k}}(T_{k,m} - 1) \le \sqrt{\tfrac{mk}{m+k}}(t-1)] \\ &\approx \Phi(\sqrt{\tfrac{mk}{m+k}}(t-1)/\sqrt{2}). \end{aligned} \tag{5.3.28}$$

An interesting and important point (noted by Box, 1953) is that unlike the $t$ test, the $F$ test for equality of variances (Problem 5.3.8(a)) does not have robustness of level. Specifically, if $\operatorname{Var}(X_1^2) \ne 2\sigma^4$, the upper $\mathcal{F}_{k,m}$ critical value $f_{k,m}(1-\alpha)$, which by (5.3.28) satisfies

$$z_{1-\alpha} \approx \sqrt{\frac{mk}{m+k}}(f_{k,m}(1-\alpha) - 1)/\sqrt{2}$$

or

$$f_{k,m}(1-\alpha) \approx 1 + \sqrt{\frac{2(m+k)}{mk}} z_{1-\alpha}$$

is asymptotically incorrect. In general (Problem 5.3.8(c)) one has to use the critical value

$$c_{k,m} = \left(1 + \sqrt{\frac{\kappa(m+k)}{mk}} z_{1-\alpha}\right) \tag{5.3.29}$$

where $\kappa = \operatorname{Var}[(X_1 - \mu_1)/\sigma_1]^2$, $\mu_1 = E(X_1)$, and $\sigma_1^2 = \operatorname{Var}(X_1)$. When $\kappa$ is unknown, it can be estimated by the method of moments (Problem 5.3.8(d)). □

### 5.3.3 Asymptotic Normality of the Maximum Likelihood Estimate in Exponential Families

Our final application of the $\delta$-method follows.

**Theorem 5.3.5.** *Suppose $\mathcal{P}$ is a canonical exponential family of rank $d$ generated by $\mathbf{T}$ with $\mathcal{E}$ open. Then if $X_1, \dots, X_n$ are a sample from $P_{\boldsymbol{\eta}} \in \mathcal{P}$ and $\widehat{\boldsymbol{\eta}}$ is defined as the* MLE *if it exists and equal to $\mathbf{c}$ (some fixed value) otherwise,* *(i)* $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta} + \frac{1}{n}\sum_{i=1}^{n} \ddot{A}^{-1}(\boldsymbol{\eta})(\mathbf{T}(X_i) - \dot{A}(\boldsymbol{\eta})) + o_{P_{\boldsymbol{\eta}}}(n^{-\frac{1}{2}})$

*(ii)* $\mathcal{L}_{\boldsymbol{\eta}}(\sqrt{n}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})) \to \mathcal{N}_d(\mathbf{0}, \ddot{A}^{-1}(\boldsymbol{\eta}))$.

***Proof.*** The result is a consequence of Theorems 5.2.2 and 5.3.4. We showed in the proof of Theorem 5.2.2 that, if $\bar{\mathbf{T}} \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbf{T}(X_i)$, $P_{\boldsymbol{\eta}}[\bar{\mathbf{T}} \in \dot{A}(\mathcal{E})] \to 1$ and, hence, $P_{\boldsymbol{\eta}}[\widehat{\boldsymbol{\eta}} = \dot{A}^{-1}(\bar{\mathbf{T}})] \to 1$. Identify $\mathbf{h}$ in Theorem 5.3.4 with $\dot{A}^{-1}$ and $\mathbf{m}$ with $\dot{A}(\boldsymbol{\eta})$. Note that by B.8.14, if $\mathbf{t} = \dot{A}(\boldsymbol{\eta})$,

$$D\dot{A}^{-1}(\mathbf{t}) = [D\dot{A}(\boldsymbol{\eta})]^{-1}. \tag{5.3.30}$$

But $D\dot{A} = \ddot{A}$ by definition and, thus, in our case,

$$\mathbf{h}^{(1)}(\mathbf{m}) = \ddot{A}^{-1}(\boldsymbol{\eta}). \tag{5.3.31}$$

Thus, (i) follows from (5.3.23). For (ii) simply note that, in our case, by Corollary 1.6.1,

$$\boldsymbol{\Sigma} = \text{Var}(\mathbf{T}(X_1)) = \ddot{A}(\boldsymbol{\eta})$$

and, therefore,

$$\mathbf{h}^{(1)}(\mathbf{m})\boldsymbol{\Sigma}[\mathbf{h}^{(1)}(\mathbf{m})]^T = \ddot{A}^{-1}\ddot{A}\ddot{A}^{-1}(\boldsymbol{\eta}) = \ddot{A}^{-1}(\boldsymbol{\eta}). \tag{5.3.32}$$

Hence, (ii) follows from (5.3.24). □

**Remark 5.3.1.** Recall that

$$\ddot{A}(\boldsymbol{\eta}) = \text{Var}_{\boldsymbol{\eta}}(\mathbf{T}) = I(\boldsymbol{\eta})$$

is the Fisher information. Thus, the asymptotic variance matrix $I^{-1}(\boldsymbol{\eta})$ of $\sqrt{n}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ equals the lower bound (3.4.38) on the variance matrix of $\sqrt{n}(\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta})$ for any unbiased estimator $\widetilde{\boldsymbol{\eta}}$. This is an "asymptotic efficiency" property of the MLE we return to in Section 6.2.1.

**Example 5.3.8.** Let $X_1, \dots, X_n$ be i.i.d. as $X$ with $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $T_1 = \bar{X}$ and $T_2 = n^{-1}\Sigma X_i^2$ are sufficient statistics in the canonical model. Now

$$\sqrt{n}[T_1 - \mu, T_2 - (\mu^2 + \sigma^2)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 0, I(\boldsymbol{\eta})) \tag{5.3.33}$$

where, by Example 2.3.4,

$$I(\boldsymbol{\eta}) = \ddot{A}(\boldsymbol{\eta}) = \frac{1}{2\eta_2^2}\begin{pmatrix} -\eta_2 & \eta_1 \\ \eta_1 & 1 - \eta_1^2(4\eta_2)^{-1} \end{pmatrix}.$$

Here $\eta_1 = \mu/\sigma^2$, $\eta_2 = -1/2\sigma^2$, $\widehat{\eta}_1 = \bar{X}/\widehat{\sigma}^2$, and $\widehat{\eta}_2 = -1/2\widehat{\sigma}^2$ where $\widehat{\sigma}^2 = T_2 - (T_1)^2$. By Theorem 5.3.5,

$$\sqrt{n}(\widehat{\eta}_1 - \eta_1, \widehat{\eta}_2 - \eta_2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 0, I^{-1}(\boldsymbol{\eta})).$$

Because $\bar{X} = T_1$ and $\widehat{\sigma}^2 = T_2 - (T_1)^2$, we can use (5.3.33) and Theorem 5.3.4 to find (Problem 5.3.26)

$$\sqrt{n}(\bar{X} - \mu, \widehat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 0, \Sigma_0)$$

where $\Sigma_0 = \text{diag}(\sigma^2, 2\sigma^4)$.

**Summary.** Consistency is 0th-order asymptotics. First-order asymptotics provides approximations to the difference between a quantity tending to a limit and the limit, for instance, the difference between a consistent estimate and the parameter it estimates. Second-order asymptotics provides approximations to the difference between the error and its first-order approximation, and so on. We begin in Section 5.3.1 by studying approximations to moments and central moments of estimates. Fundamental asymptotic formulae are derived for the bias and variance of an estimate first for smooth function of a scalar mean and then a vector mean. These "$\delta$ method" approximations based on Taylor's formula and elementary results about moments of means of i.i.d. variables are explained in terms of similar stochastic approximations to $h(\bar{\mathbf{Y}}) - h(\boldsymbol{\mu})$ where $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are i.i.d. as $\mathbf{Y}$, $E\mathbf{Y} = \boldsymbol{\mu}$, and $h$ is smooth. These stochastic approximations lead to Gaussian approximations to the laws of important statistics. The moment and in law approximations lead to the definition of variance stabilizing transformations for classical one-dimensional exponential families. Higher-order approximations to distributions (Edgeworth series) are discussed briefly. Finally, stochastic approximations in the case of vector statistics and parameters are developed, which lead to a result on the asymptotic normality of the MLE in multiparameter exponential families.

## 5.4 ASYMPTOTIC THEORY IN ONE DIMENSION

In this section we define and study asymptotic optimality for estimation, testing, and confidence bounds, under i.i.d. sampling, when we are dealing with one-dimensional smooth parametric models. Specifically we shall show that important likelihood based procedures such as MLE's are asymptotically optimal. In Chapter 6 we sketch how these ideas can be extended to multi-dimensional parametric families.

### 5.4.1 Estimation: The Multinomial Case

Following Fisher (1958),[1] we develop the theory first for the case that $X_1, \dots, X_n$ are i.i.d. taking values $\{x_0, \dots, x_k\}$ only so that $P$ is defined by $\mathbf{p} \equiv (p_0, \dots, p_k)$ where

$$p_j \equiv P[X_1 = x_j],\ 0 \le j \le k \tag{5.4.1}$$

and $\mathbf{p} \in \mathcal{S}$, the $(k+1)$-dimensional simplex (see Example 1.6.7). Thus, $\mathbf{N} = (N_0, \dots, N_k)$ where $N_j \equiv \sum_{i=1}^n 1(X_i = x_j)$ is sufficient. We consider one-dimensional parametric submodels of $\mathcal{S}$ defined by $\mathcal{P} = \{(p(x_0, \theta), \dots, p(x_k, \theta)) : \theta \in \Theta\}$, $\Theta$ open $\subset R$ (e.g., see Example 2.1.4 and Problem 2.1.15). We focus first on estimation of $\theta$. Assume

$$A : \theta \to p(x_j, \theta),\ 0 < p_j < 1,\ \text{is twice differentiable for } 0 \le j \le k.$$

Note that $A$ implies that

$$l(X_1,\theta) \equiv \log p(X_1,\theta) = \sum_{j=0}^{k} \log p(x_j,\theta) 1(X_1 = x_j) \tag{5.4.2}$$

is twice differentiable and $\frac{\partial l}{\partial\theta}(X_1,\theta)$ is a well-defined, bounded random variable

$$\frac{\partial l}{\partial\theta}(X_1,\theta) = \sum_{j=0}^{k} \left(\frac{\partial p}{\partial\theta}(x_j,\theta)\right) \frac{1}{p(x_j,\theta)} \cdot 1(X_1 = x_j). \tag{5.4.3}$$

Furthermore (Section 3.4.2),

$$E_\theta \frac{\partial l}{\partial\theta}(X_1,\theta) = 0 \tag{5.4.4}$$

and $\frac{\partial^2 l}{\partial\theta^2}(X_1,\theta)$ is similarly bounded and well defined with

$$I(\theta) \equiv \text{Var}_\theta\left(\frac{\partial l}{\partial\theta}(X_1,\theta)\right) = -E_\theta \frac{\partial^2}{\partial\theta^2} l(X_1,\theta). \tag{5.4.5}$$

As usual we call $I(\theta)$ the *Fisher information.*

Next suppose we are given a plug-in estimator $h\left(\frac{\mathbf{N}}{n}\right)$ (see (2.1.11)) of $\theta$ where

$$h : \mathcal{S} \to R$$

satisfies

$$h(\mathbf{p}(\theta)) = \theta \text{ for all } \theta \in \Theta \tag{5.4.6}$$

where $\mathbf{p}(\theta) = (p(x_0,\theta),\ldots,p(x_k,\theta))^T$. Many such $h$ exist if $k > 1$. Consider Example 2.1.4, for instance. Assume

$$H : h \text{ is differentiable.}$$

Then we have the following theorem.

**Theorem 5.4.1.** *Under $H$, for all $\theta$,*

$$\mathcal{L}_\theta\left(\sqrt{n}\left(h\left(\frac{\mathbf{N}}{n}\right) - \theta\right)\right) \to \mathcal{N}(0,\sigma^2(\theta,h)) \tag{5.4.7}$$

*where $\sigma^2(\theta,h)$ is given by* (5.4.11). *Moreover, if $A$ also holds,*

$$\sigma^2(\theta,h) \geq I^{-1}(\theta) \tag{5.4.8}$$

*with equality if and only if,*

$$\left.\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\right|_{\mathbf{p}(\theta)} = I^{-1}(\theta)\frac{\partial l}{\partial\theta}(x_j,\theta),\ 0 \leq j \leq k. \tag{5.4.9}$$

***Proof.*** Apply Theorem 5.3.2 noting that

$$\sqrt{n}\left(h\left(\frac{\mathbf{N}}{n}\right)-h(\mathbf{p}(\theta))\right)=\sqrt{n}\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\left(\frac{N_j}{n}-p(x_j,\theta)\right)+o_p(1).$$

Note that, using the definition of $N_j$,

$$\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\left(\frac{N_j}{n}-p(x_j,\theta)\right)=n^{-1}\sum_{i=1}^{n}\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))(1(X_i=x_j)-p(x_j,\theta)). \tag{5.4.10}$$

Thus, by (5.4.10), not only is $\sqrt{n}\left\{h\left(\frac{\mathbf{N}}{n}\right)-h(\mathbf{p}(\theta))\right\}$ asymptotically normal with mean 0, but also its asymptotic variance is

$$\begin{aligned}\sigma^2(\theta,h) &= \text{Var}_\theta\left(\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))1(X_1=x_j)\right)\\ &= \sum_{j=0}^{k}\left(\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\right)^2 p(x_j,\theta)-\left(\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))p(x_j,\theta)\right)^2.\end{aligned} \tag{5.4.11}$$

Note that by differentiating (5.4.6), we obtain

$$\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))\frac{\partial p}{\partial\theta}(x_j,\theta)=1 \tag{5.4.12}$$

or equivalently, by noting $\frac{\partial p}{\partial\theta}(x_j,\theta)=\left[\frac{\partial}{\partial\theta}l(x_j,\theta)\right]p(x_j,\theta)$,

$$\text{Cov}_\theta\left(\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))1(X_1=x_j),\frac{\partial l}{\partial\theta}(X_1,\theta)\right)=1. \tag{5.4.13}$$

By (5.4.13), using the correlation inequality (A.11.16) as in the proof of the information inequality (3.4.12), we obtain

$$1\leq\sigma^2(\theta,h)\text{Var}_\theta\frac{\partial l}{\partial\theta}(X_1,\theta)=\sigma^2(\theta,h)I(\theta) \tag{5.4.14}$$

with equality iff,

$$\sum_{j=0}^{k}\frac{\partial h}{\partial p_j}(\mathbf{p}(\theta))(1(X_1=x_j)-p(x_j,\theta))=a(\theta)\frac{\partial l}{\partial\theta}(X_1,\theta)+b(\theta) \tag{5.4.15}$$

for some $a(\theta)\neq 0$ and some $b(\theta)$ with probability 1. Taking expectations we get $b(\theta)=0$. Noting that the covariance of the right- and left-hand sides is $a(\theta)$, while their common variance is $a^2(\theta)I(\theta)=\sigma^2(\theta,h)$, we see that equality in (5.4.8) gives

$$a^2(\theta)I^2(\theta)=1, \tag{5.4.16}$$

which implies (5.4.9). □

We shall see in Section 5.4.3 that the information bound (5.4.8) is, if it exists and under regularity conditions, achieved by $\widehat{\theta} = \widehat{h}\left(\frac{\mathbf{N}}{n}\right)$, the MLE of $\theta$ where $\widehat{h}$ is defined implicitly by: $\widehat{h}(\mathbf{p})$ is the value of $\theta$, which

(i) maximizes $\sum_{j=0}^{k} N_j \log p(x_j, \theta)$

and

(ii) solves $\sum_{j=0}^{k} N_j \frac{\partial l}{\partial \theta}(x_j, \theta) = 0$.

**Example 5.4.1.** *One-Parameter Discrete Exponential Families.* Suppose $p(x,\theta) = \exp\{\theta T(x) - A(\theta)\}h(x)$ where $h(x) = 1(x \in \{x_0, \dots, x_k\})$, $\theta \in \Theta$, is a canonical one-parameter exponential family (supported on $\{x_0, \dots, x_k\}$) and $\Theta$ is open. Then Theorem 5.3.5 applies to the MLE $\widehat{\theta}$ and

$$\mathcal{L}_\theta(\sqrt{n}(\widehat{\theta} - \theta)) \to N\left(0, \frac{1}{I(\theta)}\right) \tag{5.4.17}$$

with the asymptotic variance achieving the information bound $I^{-1}(\theta)$. Note that because $\bar{T} = n^{-1}\sum_{i=1}^{n} T(X_i) = \sum_{j=0}^{k} T(x_j)\frac{N_j}{n}$, then, by (2.3.3)

$$\widehat{\theta} = [\dot{A}]^{-1}(\bar{T}), \tag{5.4.18}$$

and

$$\widehat{h}(\mathbf{p}) = [\dot{A}]^{-1}\left(\sum_{j=0}^{k} T(x_j)p_j\right). \tag{5.4.19}$$

The binomial $(n,p)$ and Hardy–Weinberg models can both be put into this framework with canonical parameters such as $\theta = \log\left(\frac{p}{1-p}\right)$ in the first case. □

Both the asymptotic variance bound and its achievement by the MLE are much more general phenomena. In the next two subsections we consider some more general situations.

### 5.4.2 Asymptotic Normality of Minimum Contrast and $M$-Estimates

We begin with an asymptotic normality theorem for minimum contrast estimates. As in Theorem 5.2.3 we give this result under conditions that are themselves implied by more technical sufficient conditions that are easier to check.

Suppose i.i.d. $X_1, \dots, X_n$ are tentatively modeled to be distributed according to $P_\theta$, $\theta \in \Theta$ open $\subset R$ and corresponding density/frequency functions $p(\cdot, \theta)$. Write $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let $\rho : X \times \Theta \to R$ where

$$D(\theta, \theta_0) = E_{\theta_0}(\rho(X_1, \theta) - \rho(X_1, \theta_0))$$

is uniquely minimized at $\theta_0$. Let $\bar{\theta}_n$ be the minimum contrast estimate

$$\bar{\theta}_n = \text{argmin}\ \frac{1}{n}\sum_{i=1}^{n}\rho(X_i,\theta).$$

Suppose

**A0:** $\psi = \frac{\partial\rho}{\partial\theta}$ is well defined.
Then

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\bar{\theta}_n) = 0. \tag{5.4.20}$$

In what follows we let $P$, rather than $P_\theta$, denote the distribution of $X_i$. This is because, as pointed out later in Remark 5.4.3, under regularity conditions the properties developed in this section are valid for $P \notin \{P_\theta : \theta \in \Theta\}$. We need only that $\theta(P)$ is a well defined parameter as defined in Section 1.1.2. As we saw in Section 2.1, parameters and their estimates can often be extended to larger classes of distributions than they originally were defined for. Suppose

**A1:** The parameter $\theta(P)$ given by the solution of

$$\int \psi(x,\theta)dP(x) = 0 \tag{5.4.21}$$

is well defined on $\mathcal{P}$. That is,

$$\int |\psi(x,\theta)|dP(x) < \infty,\ \theta\in\Theta,\ P\in\mathcal{P}$$

and $\theta(P)$ is the unique solution of (5.4.21) and, hence, $\theta(P_\theta) = \theta$.

**A2:** $E_P\psi^2(X_1,\theta(P)) < \infty$ for all $P \in \mathcal{P}$.

**A3:** $\psi(\cdot,\theta)$ is differentiable, $\frac{\partial\psi}{\partial\theta}(X_1,\theta)$ has a finite expectation and

$$E_P\frac{\partial\psi}{\partial\theta}(X_1,\theta(P)) \neq 0.$$

**A4:** $\sup_t\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\psi}{\partial\theta}(X_i,t) - \frac{\partial\psi}{\partial\theta}(X_i,\theta(P))\right)\right| : |t-\theta(P)| \le \epsilon_n\right\} \overset{P}{\to} 0$ if $\epsilon_n \to 0$.

**A5:** $\bar{\theta}_n \overset{P}{\to} \theta(P)$. That is, $\bar{\theta}_n$ is consistent on $\mathcal{P} = \{P_\theta : \theta\in\Theta\}$.

**Theorem 5.4.2.** *Under* A0–A5,

$$\bar{\theta}_n = \theta(P) + \frac{1}{n}\sum_{i=1}^{n}\widetilde{\psi}(X_i,\theta(P)) + o_p(n^{-1/2}) \tag{5.4.22}$$

*where*

$$\widetilde{\psi}(x,P) = \psi(x,\theta(P)) \Big/ \left(-E_P\frac{\partial\psi}{\partial\theta}(X_1,\theta(P))\right). \tag{5.4.23}$$

*Hence,*

$$\mathcal{L}_P(\sqrt{n}(\bar{\theta}_n - \theta(P))) \to \mathcal{N}(0, \sigma^2(\psi, P))$$

*where*

$$\sigma^2(\psi, P) = \frac{E_P \psi^2(X_1, \theta(P))}{\left(E_P \frac{\partial \psi}{\partial \theta}(X_1, \theta(P))\right)^2}. \tag{5.4.24}$$

***Proof.*** Claim (5.4.24) follows from the central limit theorem and Slutsky's theorem, applied to (5.4.22) because

$$\sqrt{n}(\bar{\theta}_n - \theta(P)) = n^{-1/2} \sum_{i=1}^n \widetilde{\psi}(X_i, P) + o_p(1)$$

and

$$\begin{aligned} E_P \widetilde{\psi}(X_1, P) &= E_P \psi(X_1, \theta(P)) \Big/ \left(-E_P \tfrac{\partial \psi}{\partial \theta}(X_1, \theta(P))\right) \\ &= 0 \end{aligned}$$

while

$$E_P \widetilde{\psi}^2(X_1, P) = \sigma^2(\psi, P) < \infty$$

by A1, A2, and A3. Next we show that (5.4.22) follows by a Taylor expansion of the equations (5.4.20) and (5.4.21). Let $\bar{\theta}_n = \theta(\widehat{P})$ where $\widehat{P}$ denotes the empirical probability. By expanding $n^{-1} \sum_{i=1}^n \psi(X_i, \bar{\theta}_n)$ around $\theta(P)$, we obtain, using (5.4.20),

$$-\frac{1}{n} \sum_{i=1}^n \psi(X_i, \theta(P)) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \psi}{\partial \theta}(X_i, \theta_n^*)(\bar{\theta}_n - \theta(P)) \tag{5.4.25}$$

where $|\theta_n^* - \theta(P)| \leq |\bar{\theta}_n - \theta(P)|$. Apply A5 and A4 to conclude that

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \psi}{\partial \theta}(X_i, \theta_n^*) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \psi(X_i, \theta(P)) + o_p(1) \tag{5.4.26}$$

and A3 and the WLLN to conclude that

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \psi}{\partial \theta}(X_i, \theta(P)) = E_P \frac{\partial \psi}{\partial \theta}(X_1, \theta(P)) + o_p(1). \tag{5.4.27}$$

Combining (5.4.25)–(5.4.27) we get,

$$(\bar{\theta}_n - \theta(P)) \left(-E_P \frac{\partial \psi}{\partial \theta}(X_1, \theta(P)) + o_p(1)\right) = \frac{1}{n} \sum_{i=1}^n \psi(X_i, \theta(P)). \tag{5.4.28}$$

But by the central limit theorem and A1,

$$\frac{1}{n} \sum_{i=1}^n \psi(X_i, \theta(P)) = O_p(n^{-1/2}). \tag{5.4.29}$$

Dividing by the second factor in (5.4.28) we finally obtain

$$\bar{\theta}_n - \theta(P) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\psi}(X_i, \theta(P)) + o_p\left(\frac{1}{n}\sum_{i=1}^{n}\psi(X_i, \theta(P))\right)$$

and (5.4.22) follows from the foregoing and (5.4.29). □

**Remark 5.4.1.** An additional assumption A6 gives a slightly different formula for $E_P \frac{\partial\psi}{\partial\theta}(X_1, \theta(P))$ if $P = P_\theta$.

**A6:** Suppose $P = P_\theta$ so that $\theta(P) = \theta$, and that the model $\mathcal{P}$ is regular and let $l(x, \theta) = \log p(x, \theta)$ where $p(\cdot, \theta)$ is as usual a density or frequency function. Suppose $l$ is differentiable and assume that

$$\begin{aligned} E_\theta \frac{\partial\psi}{\partial\theta}(X_1, \theta(P)) &= -E_\theta \frac{\partial l}{\partial\theta}(X_1, \theta)\psi(X_1, \theta) \\ &= -\text{Cov}_\theta\left(\frac{\partial l}{\partial\theta}(X_1, \theta), \psi(X_1, \theta)\right). \end{aligned} \tag{5.4.30}$$

Note that (5.4.30) is formally obtained by differentiating the equation (5.4.21), written as

$$\int \psi(x, \theta)p(x, \theta)d\mu(x) = 0 \tag{5.4.31}$$

for all $\theta$. If an unbiased estimate $\delta(X_1)$ of $\theta$ exists and we let $\psi(x, \theta) = \delta(x) - \theta$, it is easy to see that A6 is the same as $(3.4.8)$. If further $X_1$ takes on a finite set of values, $\{x_0, \ldots, x_k\}$, and we define $h(\mathbf{p}) = \sum_{j=0}^{k} \delta(x_j)p_j$, we see that A6 corresponds to (5.4.12).

Identity (5.4.30) suggests that if $\mathcal{P}$ is regular the conclusion of Theorem 5.4.2 may hold even if $\psi$ is not differentiable provided that $-E_\theta \frac{\partial\psi}{\partial\theta}(X_1, \theta)$ is replaced by $\text{Cov}_\theta(\psi(X_1, \theta), \frac{\partial l}{\partial\theta}(X_1, \theta))$ and a suitable replacement for A3, A4 is found. This is in fact true—see Problem 5.4.1.

**Remark 5.4.2.** Solutions to $(5.4.20)$ are called $M$-estimates as well as estimating equation estimates—see Section 2.2.1. Our arguments apply to $M$-estimates. Nothing in the arguments require that $\bar{\theta}_n$ be a minimum contrast as well as an $M$-estimate (i.e., that $\psi = \frac{\partial\rho}{\partial\theta}$ for some $\rho$).

**Remark 5.4.3.** Our arguments apply even if $X_1, \ldots, X_n$ are i.i.d. $P$ but $P \notin \mathcal{P} = \{P_\theta : \theta \in \Theta\}$. $\theta(P)$ in A1–A5 is then replaced by

(1) $\theta(P) = \text{argmin } E_P\rho(X_1, \theta)$

or more generally, for $M$-estimates,

(2) $\theta(P)$ solves $E_P\psi(X_1, \theta) = 0$.

Theorem 5.4.2 is valid with $\theta(P)$ as in (1) or (2). This extension will be pursued in Section 6.2.1 and Volume 2.

We conclude by stating some sufficient conditions, essentially due to Cramér (1946), for A4 and A6. Conditions A0, A1, A2, and A3 are readily checkable whereas we have given conditions for A5 in Section 5.2.

**A4$'$:**

(a) $\theta \to \frac{\partial\psi}{\partial\theta}(x_1, \theta)$ is a continuous function of $\theta$ for all $x$.

(b) There exists $\delta(\theta) > 0$ such that

$$\sup\left\{\left|\frac{\partial\psi}{\partial\theta}(X_1,\theta') - \frac{\partial\psi}{\partial\theta}(X_1,\theta)\right| : |\theta - \theta'| \le \delta(\theta)\right\} \le M(X_1,\theta),$$

where $E_\theta M(X_1,\theta) < \infty$.

**A6$'$:** $\frac{\partial\psi}{\partial\theta}(x,\theta')$ is defined for all $x$, $|\theta' - \theta| \le \delta(\theta)$ and $\int_{\theta-\delta}^{\theta+\delta}\int\left|\frac{\partial\psi}{\partial s}(x,s)\right| d\mu(x)ds < \infty$ for some $\delta = \delta(\theta) > 0$, where $\mu(x)$ is the dominating measure for $P(x)$ defined in (A.10.13). That is, "$dP(x) = p(x)d\mu(x)$."

Details of how A4$'$ (with A0–A3) implies A4 and A6$'$ implies A6 are given in the problems. We also indicate by example in the problems that some conditions are needed (Problem 5.4.4) but A4$'$ and A6$'$ are not necessary (Problem 5.4.1).

## 5.4.3 Asymptotic Normality and Efficiency of the MLE

The most important special case of (5.4.30) occurs when $\rho(x,\theta) = -l(x,\theta) \equiv -\log p(x,\theta)$ and $\psi(x,\theta) \equiv \frac{\partial l}{\partial\theta}(x,\theta)$ obeys A0–A6. In this case $\bar{\theta}_n$ is the MLE $\widehat{\theta}_n$ and we obtain an identity of Fisher's,

$$\begin{aligned} -E_\theta\frac{\partial^2 l}{\partial\theta^2}(X_1,\theta) &= E_\theta\left(\frac{\partial l}{\partial\theta}(X_1,\theta)\right)^2 \\ &= \text{Var}_\theta\left(\frac{\partial l}{\partial\theta}(X_1,\theta)\right) \equiv I(\theta), \end{aligned} \tag{5.4.32}$$

where $I(\theta)$ is the Fisher information introduced in Section 3.4. We can now state the basic result on asymptotic normality and efficiency of the MLE.

**Theorem 5.4.3.** *If* A0–A6 *apply to* $\rho(x,\theta) = -l(x,\theta)$ *and* $P = P_\theta$*, then the MLE* $\widehat{\theta}_n$ *satisfies*

$$\widehat{\theta}_n = \theta + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{I(\theta)}\frac{\partial l}{\partial\theta}(X_1,\theta) + o_p(n^{-1/2}) \tag{5.4.33}$$

*so that*

$$\mathcal{L}_\theta(\sqrt{n}(\widehat{\theta}_n - \theta)) \to N\left(0, \frac{1}{I(\theta)}\right). \tag{5.4.34}$$

*Furthermore, if* $\bar{\theta}_n$ *is a minimum contrast estimate whose corresponding* $\rho$ *and* $\psi$ *satisfy* A0–A6, *then*

$$\sigma^2(\psi, P_\theta) \ge \frac{1}{I(\theta)} \tag{5.4.35}$$

*with equality iff* $\psi = a(\theta)\frac{\partial l}{\partial\theta}$ *for some* $a \neq 0$.

***Proof.*** Claims (5.4.33) and (5.4.34) follow directly by Theorem 5.4.2. By (5.4.30) and (5.4.35), claim (5.4.35) is equivalent to

$$\frac{E_\theta \psi^2(X_1,\theta)}{\left[\mathrm{Cov}_\theta\left(\psi(X_1,\theta), \frac{\partial l}{\partial \theta}(X_1,\theta)\right)\right]^2} \geq \frac{1}{\mathrm{Var}_\theta\left(\frac{\partial l}{\partial \theta}(X_1,\theta)\right)}. \tag{5.4.36}$$

Because $E_\theta \psi(X_1,\theta) = 0$, cross multiplication shows that (5.4.36) is just the correlation inequality and the theorem follows because equality holds iff $\psi$ is a nonzero multiple $a(\theta)$ of $\frac{\partial l}{\partial \theta}(X_1,\theta)$. □

Note that Theorem 5.4.3 generalizes Example 5.4.1 once we identify $\psi(x,\theta)$ with $T(x) - A'(\theta)$.

The optimality part of Theorem 5.4.3 is not valid without some conditions on the estimates being considered.

**Example 5.4.2.** *Hodges's Example.* Let $X_1,\dots,X_n$ be i.i.d. $\mathcal{N}(\theta,1)$. Then $\bar{X}$ is the MLE of $\theta$ and it is trivial to calculate $I(\theta) \equiv 1$.

Consider the following competitor to $\bar{X}$:

$$\begin{aligned} \widetilde{\theta}_n &= 0 \text{ if } |\bar{X}| \leq n^{-1/4} \\ &= \bar{X} \text{ if } |\bar{X}| > n^{-1/4}. \end{aligned} \tag{5.4.37}$$

We can interpret this estimate as first testing $H : \theta = 0$ using the test "Reject iff $|\bar{X}| \geq n^{-1/4}$" and using $\bar{X}$ as our estimate if the test rejects and 0 as our estimate otherwise. We next compute the limiting distribution of $\sqrt{n}(\widetilde{\theta}_n - \theta)$. Let $Z \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned} P_\theta[|\bar{X}| \leq n^{-1/4}] &= P[|Z + \sqrt{n}\theta| \leq n^{1/4}] \\ &= \Phi(n^{1/4} - \sqrt{n}\theta) - \Phi(-n^{1/4} - \sqrt{n}\theta). \end{aligned} \tag{5.4.38}$$

Therefore, if $\theta \neq 0$, $P_\theta[|\bar{X}| \leq n^{-1/4}] \to 0$ because $n^{1/4} - \sqrt{n}\theta \to -\infty$, and, thus, $P_\theta[\widetilde{\theta}_n = \bar{X}] \to 1$. If $\theta = 0$, $P_\theta[|\bar{X}| \leq n^{1/4}] \to 1$, and $P_\theta[\widetilde{\theta}_n = 0] \to 1$. Therefore,

$$\mathcal{L}_\theta(\sqrt{n}(\widetilde{\theta}_n - \theta)) \to \mathcal{N}(0, \sigma^2(\theta)) \tag{5.4.39}$$

where $\sigma^2(\theta) = 1 = \frac{1}{I(\theta)}$, $\theta \neq 0$, $\sigma^2(0) = 0 < \frac{1}{I(\theta)}$.

The phenomenon (5.4.39) with $\sigma^2(\theta) \leq I^{-1}(\theta)$ for all $\theta \in \Theta$ and $\sigma^2(\theta_0) < I^{-1}(\theta_0)$, for some $\theta_0 \in \Theta$ is known as *superefficiency*. For this estimate superefficiency implies poor behavior of $\widehat{\theta}_n$ at values close to 0, see Lehmann and Casella, 1998, p. 442. However, for higher-dimensional $\theta$, the phenomenon becomes more disturbing and has important practical consequences. We discuss this further in Volume II. □

## 5.4.4 Testing

The major testing problem if $\theta$ is one-dimensional is $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$. If $p(\cdot,\theta)$ is an MLR family in $T(X)$, we know that all likelihood ratio tests for simple $\theta_1$

versus simple $\theta_2$, $\theta_1 < \theta_2$, as well as the likelihood ratio test for $H$ versus $K$, are of the form "Reject $H$ for $T(X)$ large" with the critical value specified by making the probability of type I error $\alpha$ at $\theta_0$. If $p(\cdot, \theta)$ is a one-parameter exponential family in $\theta$ generated by $T(X)$, this test can also be interpreted as a test of $H : \lambda \le \lambda_0$ versus $K : \lambda > \lambda_0$, where $\lambda = \dot{A}(\theta)$ because $\dot{A}$ is strictly increasing. The test is then precisely, "Reject $H$ for large values of the MLE $T(X)$ of $\lambda$." It seems natural in general to study the behavior of the test, "Reject $H$ if $\widehat{\theta}_n \ge c(\alpha, \theta_0)$" where $P_{\theta_0}[\widehat{\theta}_n \ge c(\alpha, \theta_0)] = \alpha$ and $\widehat{\theta}_n$ is the MLE of $\theta$. We will use asymptotic theory to study the behavior of this test when we observe i.i.d. $X_1, \dots, X_n$ distributed according to $P_\theta$, $\theta \in (a, b)$, $a < \theta_0 < b$, derive an optimality property, and then directly and through problems exhibit other tests with the same behavior.

Let $c_n(\alpha, \theta_0)$ denote the critical value of the test using the MLE $\widehat{\theta}_n$ based on $n$ observations.

**Theorem 5.4.4.** *Suppose the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is such that the conditions of Theorem* 5.4.2 *apply to $\psi = \frac{\partial l}{\partial \theta}$ and $\widehat{\theta}_n$, the MLE. That is,*

$$\mathcal{L}_\theta(\sqrt{n}(\widehat{\theta}_n - \theta)) \to \mathcal{N}(0, I^{-1}(\theta)) \tag{5.4.40}$$

*where $I(\theta) > 0$ for all $\theta$. Then*

$$c_n(\alpha, \theta_0) = \theta_0 + z_{1-\alpha}/\sqrt{nI(\theta_0)} + o(n^{-1/2}) \tag{5.4.41}$$

*where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the $\mathcal{N}(0, 1)$ distribution.*

*Suppose* (A4′) *holds as well as* (A6) *and $I(\theta) < \infty$ for all $\theta$. Then*
*If $\theta > \theta_0$,*

$$P_\theta[\widehat{\theta}_n > c_n(\alpha, \theta_0)] \to 1. \tag{5.4.42}$$

*If $\theta < \theta_0$,*

$$P_\theta[\widehat{\theta}_n > c_n(\alpha, \theta_0)] \to 0. \tag{5.4.43}$$

Property (5.4.42) is sometimes called *consistency* of the test against a fixed alternative.

***Proof.*** The proof is straightforward:

$$P_{\theta_0}[\sqrt{nI(\theta_0)}(\widehat{\theta}_n - \theta_0) \ge z] \to 1 - \Phi(z)$$

by (5.4.40). Thus,

$$P_{\theta_0}[\widehat{\theta}_n \ge \theta_0 + z_{1-\alpha}/\sqrt{nI(\theta_0)}] = P_{\theta_0}[\sqrt{nI(\theta_0)}(\widehat{\theta}_n - \theta_0) \ge z_{1-\alpha}] \to \alpha. \tag{5.4.44}$$

But Polya's theorem (A.14.22) guarantees that

$$\sup_z |P_{\theta_0}[\sqrt{n}(\widehat{\theta}_n - \theta_0) \ge z] - (1 - \Phi(z))| \to 0, \tag{5.4.45}$$

which implies that $\sqrt{nI(\theta_0)}(c_n(\alpha, \theta_0) - \theta_0) - z_{1-\alpha} \to 0$, and (5.4.41) follows. On the other hand,

$$P_\theta[\widehat{\theta}_n \ge c_n(\alpha, \theta_0)] = P_\theta[\sqrt{nI(\theta)}(\widehat{\theta}_n - \theta) \ge \sqrt{nI(\theta)}(c_n(\alpha, \theta_0) - \theta)]. \tag{5.4.46}$$

By (5.4.41),

$$\begin{aligned}\sqrt{nI(\theta)}(c_n(\alpha,\theta_0)-\theta) &= \sqrt{nI(\theta)}(\theta_0-\theta+z_{1-\alpha}/\sqrt{nI(\theta_0)}+o(n^{-1/2}))\\ &= \sqrt{nI(\theta)}(\theta_0-\theta)+O(1)\to-\infty \text{ if } \theta>\theta_0\end{aligned}$$

and $\to\infty$ if $\theta<\theta_0$. Claims (5.4.42) and (5.4.43) follow. □

Theorem 5.4.4 tells us that the test under discussion is consistent and that for $n$ large the power function of the test rises steeply to $\alpha$ from the left at $\theta_0$ and continues rising steeply to 1 to the right of $\theta_0$. Optimality claims rest on a more refined analysis involving a reparametrization from $\theta$ to $\gamma\equiv\sqrt{n}(\theta-\theta_0)$.

**Theorem 5.4.5.** *Suppose the conditions of Theorem* 5.4.2 *and* (5.4.40) *hold uniformly for $\theta$ in a neighborhood of $\theta_0$. That is, assume*

$$\sup\{|P_\theta[\sqrt{nI(\theta)}(\widehat{\theta}_n-\theta)\le z]-(1-\Phi(z))|:|\theta-\theta_0|\le\epsilon(\theta_0)\}\to 0, \tag{5.4.47}$$

*for some $\epsilon(\theta_0)>0$. Let $Q_\gamma\equiv P_\theta$, $\gamma=\sqrt{n}(\theta-\theta_0)$, then*

$$Q_\gamma[\widehat{\theta}_n\ge c_n(\alpha,\theta_0)]\to 1-\Phi(z_{1-\alpha}-\gamma\sqrt{I(\theta_0)}) \tag{5.4.48}$$

*uniformly in $\gamma$. Furthermore, if $\varphi_n(X_1,\dots,X_n)$ is any sequence of (possibly randomized) critical (test) functions such that*

$$E_{\theta_0}\varphi_n(X_1,\dots,X_n)\to\alpha, \tag{5.4.49}$$

*then*

$$\begin{aligned}\overline{\lim}_n E_{\theta_0+\frac{\gamma}{\sqrt{n}}}\varphi_n(X_1,\dots,X_n) &\le 1-\Phi(z_{1-\alpha}-\gamma\sqrt{I(\theta_0)}) \textit{ if } \gamma>0\\ &\ge 1-\Phi(z_{1-\alpha}-\gamma\sqrt{I(\theta_0)}) \textit{ if } \gamma<0.\end{aligned} \tag{5.4.50}$$

Note that (5.4.48) and (5.4.50) can be interpreted as saying that among all tests that are asymptotically level $\alpha$ (obey (5.4.49)) the test based on rejecting for large values of $\widehat{\theta}_n$ is asymptotically uniformly most powerful (obey (5.4.50)) and has asymptotically smallest probability of type I error for $\theta\le\theta_0$. In fact, these statements can only be interpreted as valid in a small neighborhood of $\theta_0$ because $\gamma$ fixed means $\theta\to\theta_0$. On the other hand, if $\sqrt{n}(\theta-\theta_0)$ tends to zero, then by (5.4.50), the power of tests with asymptotic level $\alpha$ tend to $\alpha$. If $\sqrt{n}(\theta-\theta_0)$ tends to infinity, the power of the test based on $\widehat{\theta}_n$ tends to 1 by (5.4.48). In either case, the test based on $\widehat{\theta}_n$ is still asymptotically MP.

***Proof.*** Write

$$\begin{aligned}P_\theta[\widehat{\theta}_n\ge c_n(\alpha,\theta_0)] &= P_\theta[\sqrt{nI(\theta)}(\widehat{\theta}_n-\theta)\ge\sqrt{nI(\theta)}(c_n(\alpha,\theta_0)-\theta)]\\ &= P_\theta[\sqrt{nI(\theta)}(\widehat{\theta}_n-\theta)\ge\sqrt{nI(\theta)}(\theta_0-\theta+z_{1-\alpha}/\sqrt{nI(\theta_0)}\\ &\qquad +o(n^{-1/2}))].\end{aligned} \tag{5.4.51}$$

If $\gamma = \sqrt{n}(\theta - \theta_0)$ is fixed, $I(\theta) = I\left(\theta_0 + \frac{\gamma}{\sqrt{n}}\right) \to I(\theta_0)$ because our uniformity assumption implies that $\theta \to I(\theta)$ is continuous (Problem 5.4.7). Thus,

$$\begin{aligned} Q_\gamma[\widehat{\theta}_n \geq c_n(\alpha, \theta_0)] &= 1 - \Phi(z_{1-\alpha}(1 + o(1)) + \sqrt{n(I(\theta_0) + o(1))}(\theta_0 - \theta) + o(1)) \\ &= 1 - \Phi(z_{1-\alpha} - \gamma\sqrt{I(\theta_0)}) + o(1)) \end{aligned} \tag{5.4.52}$$

and (5.4.48) follows.

To prove (5.4.50) note that by the Neyman–Pearson lemma, if $\gamma > 0$,

$$\begin{aligned} E_{\theta_0 + \frac{\gamma}{\sqrt{n}}} \varphi_n(X_1, \dots, X_n) \;\; \leq \;\; & P_{\theta_0 + \frac{\gamma}{\sqrt{n}}} \left[ \sum_{i=1}^n \log \frac{p\left(X_i, \theta_0 + \frac{\gamma}{\sqrt{n}}\right)}{p(X_i, \theta_0)} \geq d_n(\alpha, \theta_0) \right] \\ & + \epsilon_n P_{\theta_0 + \frac{\gamma}{\sqrt{n}}} \left[ \sum_{i=1}^n \log \frac{p\left(X_i, \theta_0 + \frac{\gamma}{\sqrt{n}}\right)}{p(X_i, \theta_0)} = d_n(\alpha, \theta_0) \right], \end{aligned} \tag{5.4.53}$$

where $p(x, \theta)$ denotes the density of $X_i$ and $d_n$, $\epsilon_n$ are uniquely chosen so that the right-hand side of (5.4.53) is $\alpha$ if $\gamma$ is 0.

Further Taylor expansion and probabilistic arguments of the type we have used show that the right-hand side of (5.4.53) tends to the right-hand side of (5.4.50) for all $\gamma$. The details are in Problem 5.4.5. □

The asymptotic results we have just established do *not* establish that the test that rejects for large values of $\widehat{\theta}_n$ is necessarily good for all alternatives for any $n$.

The test $1[\widehat{\theta}_n \geq c_n(\alpha, \theta_0)]$ of Theorems 5.4.4 and 5.4.5 in the future will be referred to as a *Wald* test. There are two other types of test that have the same asymptotic behavior. These are the likelihood ratio test and the *score* or *Rao* test.

It is easy to see that the likelihood ratio test for testing $H : \theta \leq \theta_0$ versus $K : \theta > \theta_0$ is of the form

$$\text{“Reject if } \sum_{i=1}^n \log[p(X_i, \widehat{\theta}_n)/p(X_i, \theta_0)] 1(\widehat{\theta}_n > \theta_0) \geq k_n(\theta_0, \alpha).\text{”}$$

It may be shown (Problem 5.4.8) that, for $\alpha \leq \frac{1}{2}$, $k_n(\theta_0, \alpha) = z_{1-\alpha}^2 + o(1)$ and that if $\delta_{Wn}^*(X_1, \dots, X_n)$ is the critical function of the Wald test and $\delta_{Ln}^*(X_1, \dots, X_n)$ is the critical function of the LR test then, for all $\gamma$,

$$P_{\theta_0 + \frac{\gamma}{\sqrt{n}}} [\delta_{Ln}^*(X_1, \dots, X_n) = \delta_{Wn}^*(X_1, \dots, X_n)] \to 1. \tag{5.4.54}$$

Assertion (5.4.54) establishes that the test $\delta_{Ln}^*$ yields equality in (5.4.50) and, hence, is asymptotically most powerful as well.

Finally, note that the Neyman Pearson LR test for $H : \theta = \theta_0$ versus $K : \theta_0 + \epsilon$, $\epsilon > 0$ rejects for large values of

$$\frac{1}{\epsilon}[\log p_n(X_1, \dots, X_n, \theta_0 + \epsilon) - \log p_n(X_1, \dots, X_n, \theta_0)]$$

where $p_n(X_1,\dots,X_n,\theta)$ is the joint density of $X_1,\dots,X_n$. For $\epsilon$ small, $n$ fixed, this is approximately the same as rejecting for large values of $\frac{\partial}{\partial\theta_0}\log p_n(X_1,\dots,X_n,\theta_0)$.

The preceding argument doesn't depend on the fact that $X_1,\dots,X_n$ are i.i.d. with common density or frequency function $p(x,\theta)$ and the test that rejects $H$ for large values of $\frac{\partial}{\partial\theta_0}\log p_n(X_1,\dots,X_n,\theta_0)$ is, in general, called the *score* or *Rao* test. For the case we are considering it simplifies, becoming

$$\text{“Reject } H \text{ iff } \sum_{i=1}^{n}\frac{\partial}{\partial\theta_0}\log p(X_i,\theta_0) \geq r_n(\alpha,\theta_0)\text{.”}$$

It is easy to see (Problem 5.4.15) that

$$r_n(\alpha,\theta_0) = z_{1-\alpha}\sqrt{nI(\theta_0)} + o(n^{1/2}) \tag{5.4.55}$$

and that again if $\delta^*_{Rn}(X_1,\dots,X_n)$ is the critical function of the Rao test then

$$P_{\theta_0+\frac{\gamma}{\sqrt{n}}}[\delta^*_{Rn}(X_1,\dots,X_n) = \delta^*_{Wn}(X_1,\dots,X_n)] \to 1, \tag{5.4.56}$$

(Problem 5.4.8) and the Rao test is asymptotically optimal.

Note that for all these tests and the confidence bounds of Section 5.4.5, $I(\theta_0)$, which may require numerical integration, can be replaced by $-n^{-1}\frac{d^2}{d\theta^2}l_n(\widehat{\theta}_n)$ (Problem 5.4.10).

### 5.4.5 Confidence Bounds

We define an *asymptotic level* $1-\alpha$ *lower confidence bound (LCB)* $\underline{\theta}_n$ by the requirement that

$$P_\theta[\underline{\theta}_n \leq \theta] \to 1-\alpha \tag{5.4.57}$$

for all $\theta$ and similarly define asymptotic level $1-\alpha$ UCBs and confidence intervals.

We can approach obtaining asymptotically optimal confidence bounds in two ways:

(i) By using a natural pivot.

(ii) By inverting the testing regions derived in Section 5.4.4.

Method (i) is easier: If the assumptions of Theorem 5.4.4 hold, that is, (A0)–(A6), (A4′), and $I(\theta)$ finite for all $\theta$, it follows (Problem 5.4.9) that

$$\mathcal{L}_\theta\left(\sqrt{nI(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta)\right) \to \mathcal{N}(0,1) \tag{5.4.58}$$

for all $\theta$ and, hence, an asymptotic level $1-\alpha$ lower confidence bound is given by

$$\underline{\theta}^*_n = \widehat{\theta}_n - z_{1-\alpha}/\sqrt{nI(\widehat{\theta}_n)}. \tag{5.4.59}$$

Turning to method (ii), inversion of $\delta^*_{Wn}$ gives formally

$$\underline{\theta}^*_{n1} = \inf\{\theta : c_n(\alpha,\theta) \geq \widehat{\theta}_n\} \tag{5.4.60}$$

or if we use the approximation $\widetilde{c}_n(\alpha, \theta) = \theta + z_{1-\alpha}/\sqrt{nI(\theta)}$, (5.4.41),

$$\underline{\theta}^*_{n2} = \inf\{\theta : \widetilde{c}_n(\alpha, \theta) \geq \widehat{\theta}_n\}. \tag{5.4.61}$$

In fact neither $\underline{\theta}^*_{n1}$, or $\underline{\theta}^*_{n2}$ properly inverts the tests unless $c_n(\alpha, \theta)$ and $\widetilde{c}_n(\alpha, \theta)$ are increasing in $\theta$. The three bounds are different as illustrated by Examples 4.4.3 and 4.5.2.

If it applies and can be computed, $\underline{\theta}^*_{n1}$ is preferable because this bound is not only approximately but genuinely level $1 - \alpha$. But computationally it is often hard to implement because $c_n(\alpha, \theta)$ needs, in general, to be computed by simulation for a grid of $\theta$ values. Typically, (5.4.59) or some equivalent alternatives (Problem 5.4.10) are preferred but can be quite inadequate (Problem 5.4.11).

These bounds $\underline{\theta}^*_n, \underline{\theta}^*_{n1}, \underline{\theta}^*_{n2}$, are in fact asymptotically equivalent and optimal in a suitable sense (Problems 5.4.12 and 5.4.13).

**Summary.** We have defined asymptotic optimality for estimates in one-parameter models. In particular, we developed an asymptotic analogue of the information inequality of Chapter 3 for estimates of $\theta$ in a one-dimensional subfamily of the multinomial distributions, showed that the MLE formally achieves this bound, and made the latter result sharp in the context of one-parameter discrete exponential families. In Section 5.4.2 we developed the theory of minimum contrast and $M$-estimates, generalizations of the MLE, along the lines of Huber (1967). The asymptotic formulae we derived are applied to the MLE both under the model that led to it and under an arbitrary $P$. We also delineated the limitations of the optimality theory for estimation through Hodges's example. We studied the optimality results parallel to estimation in testing and confidence bounds. Results on asymptotic properties of statistical procedures can also be found in Ferguson (1996), Le Cam and Yang (1990), Lehmann (1999), Rao (1973), and Serfling (1980).

## 5.5 ASYMPTOTIC BEHAVIOR AND OPTIMALITY OF THE POSTERIOR DISTRIBUTION

Bayesian and frequentist inferences merge as $n \to \infty$ in a sense we now describe. The framework we consider is the one considered in Sections 5.2 and 5.4, i.i.d. observations from a regular model in which $\Theta$ is open $\subset R$ or $\Theta = \{\theta_1, \ldots, \theta_k\}$ finite, and $\theta$ is identifiable.

Most of the questions we address and answer are under the assumption that $\boldsymbol{\theta} = \theta$, an arbitrary specified value, or in frequentist terms, that $\theta$ is true.

### Consistency

The first natural question is whether the Bayes posterior distribution as $n \to \infty$ concentrates all mass more and more tightly around $\theta$. Intuitively this means that the data that are coming from $P_\theta$ eventually wipe out any prior belief that parameter values not close to $\theta$ are likely.

Formalizing this statement about the posterior distribution, $\Pi(\cdot \mid X_1, \ldots, X_n)$, which is a function-valued statistic, is somewhat subtle in general. But for $\Theta = \{\theta_1, \ldots, \theta_k\}$ it is straightforward. Let

$$\pi(\theta \mid X_1, \dots, X_n) \equiv P[\boldsymbol{\theta} = \theta \mid X_1, \dots, X_n]. \tag{5.5.1}$$

Then we say that $\Pi(\cdot \mid X_1, \dots, X_n)$ is *consistent* iff for all $\theta \in \Theta$,

$$P_\theta[|\pi(\theta \mid X_1, \dots, X_n) - 1| \geq \epsilon] \to 0 \tag{5.5.2}$$

for all $\epsilon > 0$. There is a slightly stronger definition: $\Pi(\cdot \mid X_1, \dots, X_n)$ is *a.s. consistent* iff for all $\theta \in \Theta$,

$$\pi(\theta \mid X_1, \dots, X_n) \to 1 \text{ a.s. } P_\theta. \tag{5.5.3}$$

General a.s. consistency is not hard to formulate:

$$\pi(\cdot \mid X_1, \dots, X_n) \Rightarrow \delta_{\{\theta\}} \text{ a.s. } P_\theta \tag{5.5.4}$$

where $\Rightarrow$ denotes convergence in law and $\delta_{\{\theta\}}$ is point mass at $\theta$. There is a completely satisfactory result for $\Theta$ finite.

**Theorem 5.5.1.** *Let $\pi_j \equiv P[\boldsymbol{\theta} = \theta_j]$, $j = 1, \dots, k$ denote the prior distribution of $\boldsymbol{\theta}$. Then $\Pi(\cdot \mid X_1, \dots, X_n)$ is consistent (a.s. consistent) iff $\pi_j > 0$ for $j = 1, \dots, k$.*

***Proof.*** Let $p(\cdot, \theta)$ denote the frequency or density function of $X$. The necessity of the condition is immediate because $\pi_j = 0$ for some $j$ implies that $\pi(\theta_j \mid X_1, \dots, X_n) = 0$ for all $X_1, \dots, X_n$ because, by (1.2.8),

$$\begin{aligned} \pi(\theta_j \mid X_1, \dots, X_n) &= P[\boldsymbol{\theta} = \theta_j \mid X_1, \dots, X_n] \\ &= \frac{\pi_j \prod_{i=1}^n p(X_i, \theta_j)}{\sum_{a=1}^k \pi_a \prod_{i=1}^n p(X_i, \theta_a)}. \end{aligned} \tag{5.5.5}$$

Intuitively, no amount of data can convince a Bayesian who has decided a priori that $\theta_j$ is impossible.

On the other hand, suppose all $\pi_j$ are positive. If the true $\theta$ is $\theta_j$ or equivalently $\boldsymbol{\theta} = \theta_j$, then

$$\log \frac{\pi(\theta_a \mid X_1, \dots, X_n)}{\pi(\theta_j \mid X_1, \dots, X_n)} = n \left( \frac{1}{n} \log \frac{\pi_a}{\pi_j} + \frac{1}{n} \sum_{i=1}^n \log \frac{p(X_i, \theta_a)}{p(X_i, \theta_j)} \right).$$

By the weak (respectively strong) LLN, under $P_{\theta_j}$,

$$\frac{1}{n} \sum_{i=1}^n \log \frac{p(X_i, \theta_a)}{p(X_i, \theta_j)} \to E_{\theta_j} \left( \log \frac{p(X_1, \theta_a)}{p(X_1, \theta_j)} \right)$$

in probability (respectively a.s.). But $E_{\theta_j} \left( \log \frac{p(X_1, \theta_a)}{p(X_1, \theta_j)} \right) < 0$, by Shannon's inequality, if $\theta_a \neq \theta_j$. Therefore,

$$\log \frac{\pi(\theta_a \mid X_1, \dots, X_n)}{\pi(\theta_j \mid X_1, \dots, X_n)} \to -\infty,$$

in the appropriate sense, and the theorem follows. □

**Remark 5.5.1.** We have proved more than is stated. Namely, that for each $\theta \in \Theta$, $P_\theta[\boldsymbol{\theta} \neq \theta \mid X_1, \dots, X_n] \to 0$ exponentially. □

As this proof suggests, consistency of the posterior distribution is very much akin to consistency of the MLE. The appropriate analogues of Theorem 5.2.3 are valid. Next we give a much stronger connection that has inferential implications:

### Asymptotic normality of the posterior distribution

Under conditions A0–A6 for $\rho(x,\theta) = l(x,\theta) \equiv \log p(x,\theta)$, we showed in Section 5.4 that if $\widehat{\theta}$ is the MLE,

$$\mathcal{L}_\theta(\sqrt{n}(\widehat{\theta} - \theta)) \to \mathcal{N}(0, I^{-1}(\theta)). \tag{5.5.6}$$

Consider $\mathcal{L}(\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}) \mid X_1, \dots, X_n)$, the posterior probability distribution of $\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}(X_1, \dots, X_n))$, where we emphasize that $\widehat{\theta}$ depends only on the data and is a constant given $X_1, \dots, X_n$. For conceptual ease we consider A4(a.s.) and A5(a.s.), assumptions that strengthen A4 and A5 by replacing convergence in $P_\theta$ probability by convergence a.s. $P_\theta$. We also add,

**A7:** For all $\theta$, and all $\delta > 0$ there exists $\epsilon(\delta,\theta) > 0$ such that

$$P_\theta\left[\sup\left\{\frac{1}{n}\sum_{i=1}^{n}[l(X_i,\theta') - l(X_i,\theta)] : |\theta' - \theta| \geq \delta\right\} \leq -\epsilon(\delta,\theta)\right] \to 1.$$

**A8:** The prior distribution has a density $\pi(\cdot)$ on $\Theta$ such that $\pi(\cdot)$ is continuous and positive at all $\theta$.

Remarkably,

**Theorem 5.5.2 ("Bernstein/von Mises").** *If conditions* A0–A3, A4(a.s.), A5(a.s.), A6, A7, *and* A8 *hold, then*

$$\mathcal{L}(\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}) \mid X_1, \dots, X_n) \to \mathcal{N}(0, I^{-1}(\theta)) \tag{5.5.7}$$

*a.s. under $P_\theta$ for all $\theta$.*

We can rewrite (5.5.7) more usefully as

$$\sup_x |P[\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}) \leq x \mid X_1, \dots, X_n] - \Phi(x\sqrt{I(\theta)})| \to 0 \tag{5.5.8}$$

for all $\theta$ a.s. $P_\theta$ and, of course, the statement holds for our usual and weaker convergence in $P_\theta$ probability also. From this restatement we obtain the important corollary.

**Corollary 5.5.1.** *Under the conditions of Theorem* 5.5.2,

$$\sup_x |P[\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}) \leq x \mid X_1, \dots, X_n] - \Phi(x\sqrt{I(\widehat{\theta})})| \to 0 \tag{5.5.9}$$

*a.s. $P_\theta$ for all $\theta$.*

**Remarks**

(1) Statements (5.5.4) and (5.5.7)–(5.5.9) are, in fact, frequentist statements about the asymptotic behavior of certain function-valued statistics.

(2) Claims (5.5.8) and (5.5.9) hold with a.s. replaced by in $P_\theta$ probability if A4 and A5 are used rather than their strong forms—see Problem 5.5.7.

(3) Condition A7 is essentially equivalent to (5.2.8), which coupled with (5.2.9) and identifiability guarantees consistency of $\widehat{\theta}$ in a regular model.

***Proof.*** We compute the posterior density of $\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta})$ as

$$q_n(t) = c_n^{-1}\pi\left(\widehat{\theta} + \frac{t}{\sqrt{n}}\right)\prod_{i=1}^{n} p\left(X_i, \widehat{\theta} + \frac{t}{\sqrt{n}}\right) \tag{5.5.10}$$

where $c_n = c_n(X_1, \dots, X_n)$ is given by

$$c_n(X_1, \dots, X_n) = \int_{-\infty}^{\infty} \pi\left(\widehat{\theta} + \frac{s}{\sqrt{n}}\right)\prod_{i=1}^{n} p\left(X_i, \widehat{\theta} + \frac{s}{\sqrt{n}}\right) ds.$$

Divide top and bottom of (5.5.10) by $\prod_{i=1}^{n} p(X_i, \widehat{\theta})$ to obtain

$$q_n(t) = d_n^{-1}\pi\left(\widehat{\theta} + \frac{t}{\sqrt{n}}\right)\exp\left\{\sum_{i=1}^{n}\left(l\left(X_i, \widehat{\theta} + \frac{t}{\sqrt{n}}\right) - l(X_i, \widehat{\theta})\right)\right\} \tag{5.5.11}$$

where $l(x, \theta) = \log p(x, \theta)$ and

$$d_n = \int_{-\infty}^{\infty} \pi\left(\widehat{\theta} + \frac{s}{\sqrt{n}}\right)\exp\left\{\sum_{i=1}^{n}\left(l\left(X_i, \widehat{\theta} + \frac{s}{\sqrt{n}}\right) - l(X_i, \widehat{\theta})\right)\right\} ds.$$

We claim that

$$P_\theta\left[d_n q_n(t) \to \pi(\theta)\exp\left\{-\frac{t^2 I(\theta)}{2}\right\} \text{ for all } t\right] = 1 \tag{5.5.12}$$

for all $\theta$. To establish this note that

(a) $\sup\left\{\left|\pi\left(\widehat{\theta} + \frac{t}{\sqrt{n}}\right) - \pi(\theta)\right| : |t| \le M\right\} \to 0$ a.s. for all $M$ because $\widehat{\theta}$ is a.s. consistent and $\pi$ is continuous.

(b) Expanding,

$$\sum_{i=1}^{n}\left(l\left(X_i, \widehat{\theta} + \frac{t}{\sqrt{n}}\right) - l(X_i, \widehat{\theta})\right) = \frac{t^2}{2}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l}{\partial\theta^2}(X_i, \theta^*(t)) \tag{5.5.13}$$

where $|\widehat{\theta} - \theta^*_{(t)}| \leq \frac{t}{\sqrt{n}}$. We use $\sum_{i=1}^n \frac{\partial l}{\partial \theta}(X_i, \widehat{\theta}) = 0$ here. By A4(a.s.), A5(a.s.),

$$\sup\left\{\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 l}{\partial \theta^2}(X_i, \theta^*(t)) - \frac{1}{n}\sum_{i=1}^n \frac{\partial^2 l}{\partial \theta^2}(X_i, \theta)\right| : |t| \leq M\right\} \to 0,$$

for all $M$, a.s. $P_\theta$. Using $(5.5.13)$, the strong law of large numbers (SLLN) and A8, we obtain (Problem $5.5.3$),

$$P_\theta\left[d_n q_n(t) \to \pi(\theta)\exp\left\{E_\theta \frac{\partial^2 l}{\partial \theta^2}(X_1, \theta)\frac{t^2}{2}\right\} \text{ for all } t\right] = 1. \tag{5.5.14}$$

Using A6 we obtain $(5.5.12)$.

Now consider

$$\begin{aligned} d_n &= \int_{-\infty}^{\infty} \pi\left(\widehat{\theta} + \frac{s}{\sqrt{n}}\right)\exp\left\{\sum_{i=1}^n l\left(X_i, \widehat{\theta} + \frac{s}{\sqrt{n}}\right) - l(X_i, \widehat{\theta})\right\} ds \\ &= \int_{|s| \leq \delta\sqrt{n}} d_n q_n(s) ds \\ &\quad + \sqrt{n}\int \pi(t)\exp\left\{\sum_{i=1}^n (l(X_i, t) - l(X_i, \widehat{\theta}))\right\} 1(|t - \widehat{\theta}| > \delta) dt \end{aligned} \tag{5.5.15}$$

By A5 and A7,

$$P_\theta\left[\sup\left\{\exp\left\{\sum_{i=1}^n (l(X_i, t) - l(X_i, \widehat{\theta}))\right\} : |t - \widehat{\theta}| > \delta\right\} \leq e^{-n\epsilon(\delta, \theta)}\right] \to 1 \tag{5.5.16}$$

for all $\delta$ so that the second term in $(5.5.14)$ is bounded by $\sqrt{n}e^{-n\epsilon(\delta, \theta)} \to 0$ a.s. $P_\theta$ for all $\delta > 0$. Finally note that (Problem $5.5.4$) by arguing as for $(5.5.14)$, there exists $\delta(\theta) > 0$ such that

$$P_\theta\left[d_n q_n(t) \leq 2\pi(\theta)\exp\left\{\frac{1}{4}E_\theta\left(\frac{\partial^2 l}{\partial \theta^2}(X_1, \theta)\right)\frac{t^2}{2}\right\} \text{ for all } |t| \leq \delta(\theta)\sqrt{n}\right] \to 1. \tag{5.5.17}$$

By $(5.5.15)$ and $(5.5.16)$, for all $\delta > 0$,

$$P_\theta\left[d_n - \int_{|s| \leq \delta\sqrt{n}} d_n q_n(s) ds \to 0\right] = 1. \tag{5.5.18}$$

Finally, apply the dominated convergence theorem, Theorem B.7.5, to $d_n q_n(s1(|s| \leq \delta(\theta)\sqrt{n}))$, using $(5.5.14)$ and $(5.5.17)$ to conclude that, a.s. $P_\theta$,

$$d_n \to \pi(\theta)\int_{-\infty}^{\infty} \exp\left\{-\frac{s^2 I(\theta)}{2}\right\} ds = \frac{\pi(\theta)\sqrt{2\pi}}{\sqrt{I(\theta)}}. \tag{5.5.19}$$

Hence, a.s. $P_\theta$,

$$q_n(t) \to \sqrt{I(\theta)}\varphi(t\sqrt{I(\theta)})$$

where $\varphi$ is the standard Gaussian density and the theorem follows from Scheffé's Theorem B.7.6 and Proposition B.7.2. □

**Example 5.5.1.** *Posterior Behavior in the Normal Translation Model with Normal Prior.* (Example 3.2.1 continued). Suppose as in Example 3.2.1 we have observations from a $\mathcal{N}(\theta, \sigma^2)$ distribution with $\sigma^2$ known and we put a $\mathcal{N}(\eta, \tau^2)$ prior on $\boldsymbol{\theta}$. Then the posterior distribution of $\boldsymbol{\theta}$ is $\mathcal{N}\left(\omega_{1n}\eta + \omega_{2n}\bar{X}, \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$ where

$$\omega_{1n} = \frac{\sigma^2}{n\tau^2 + \sigma^2}, \; \omega_{2n} = 1 - \omega_{1n}. \tag{5.5.20}$$

Evidently, as $n \to \infty$, $\omega_{1n} \to 0$, $\bar{X} \to \theta$, a.s., if $\boldsymbol{\theta} = \theta$, and $\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \to 0$. That is, the posterior distribution has mean approximately $\theta$ and variance approximately 0, for $n$ large, or equivalently the posterior is close to point mass at $\theta$ as we expect from Theorem 5.5.1. Because $\widehat{\theta} = \bar{X}$, $\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta})$ has posterior distribution $\mathcal{N}\left(\sqrt{n}\omega_{1n}(\eta - \bar{X}), n\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$. Now, $\sqrt{n}\omega_{1n} = O(n^{-1/2}) = o(1)$ and $n\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \to \sigma^2 = I^{-1}(\theta)$ and we have directly established the conclusion of Theorem 5.5.2. □

**Example 5.5.2.** *Posterior Behavior in the Binomial-Beta Model.* (Example 3.2.3 continued). If we observe $S_n$ with a binomial, $B(n, \theta)$, distribution, or equivalently we observe $X_1, \dots, X_n$ i.i.d. Bernoulli $(1, \theta)$ and put a beta, $\beta(r, s)$ prior on $\theta$, then, as in Example 3.2.3, $\boldsymbol{\theta}$ has posterior $\beta(S_n + r, n + s - S_n)$. We have shown in Problem 5.3.20 that if $U_{a,b}$ has a $\beta(a, b)$ distribution, then as $a \to \infty$, $b \to \infty$,

$$\left[\frac{(a+b)^3}{ab}\right]^{\frac{1}{2}} \left(U_{a,b} - \frac{a}{a+b}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \tag{5.5.21}$$

If $0 < \theta < 1$ is true, $S_n/n \overset{\text{a.s.}}{\to} \theta$ so that $S_n + r \to \infty$, $n + s - S_n \to \infty$ a.s. $P_\theta$. By identifying $a$ with $S_n + r$ and $b$ with $n + s - S_n$ we conclude after some algebra that because $\widehat{\theta} = \bar{X}$,

$$\sqrt{n}(\boldsymbol{\theta} - \bar{X}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1 - \theta))$$

a.s. $P_\theta$, as claimed by Theorem 5.5.2. □

### Bayesian optimality of optimal frequentist procedures and frequentist optimality of Bayesian procedures

Theorem 5.5.2 has two surprising consequences.

(a) Bayes estimates for a wide variety of loss functions and priors are asymptotically efficient in the sense of the previous section.

(b) The maximum likelihood estimate is asymptotically equivalent in a Bayesian sense to the Bayes estimate for a variety of priors and loss functions.

As an example of this phenomenon consider the following.

**Theorem 5.5.3.** *Suppose the conditions of Theorem* 5.5.2 *are satisfied. Let* $\widehat{\theta}$ *be the* MLE *of* $\theta$ *and let* $\widehat{\theta}^*$ *be the median of the posterior distribution of* $\boldsymbol{\theta}$*. Then*

(i)

$$\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) \to 0 \tag{5.5.22}$$

*a.s.* $P_\theta$ *for all* $\theta$*. Consequently,*

$$\widehat{\theta}^* = \theta + \frac{1}{n}\sum_{i=1}^{n} I^{-1}(\theta)\frac{\partial l}{\partial \theta}(X_i, \theta) + o_{P_\theta}(n^{-1/2}) \tag{5.5.23}$$

*and* $\mathcal{L}_\theta(\sqrt{n}(\widehat{\theta}^* - \theta)) \to \mathcal{N}(0, I^{-1}(\theta))$.

(ii)

$$E(\sqrt{n}(|\boldsymbol{\theta} - \widehat{\theta}| - |\boldsymbol{\theta} - \widehat{\theta}^*|) \mid X_1, \dots, X_n) = o_P(1) \tag{5.5.24}$$

$$E(\sqrt{n}(|\boldsymbol{\theta} - \widehat{\theta}| - |\boldsymbol{\theta}| \mid X_1, \dots, X_n) = \min_d E(\sqrt{n}(|\boldsymbol{\theta} - d| - |\boldsymbol{\theta}|) \mid X_1, \dots, X_n) + o_P(1). \tag{5.5.25}$$

Thus, (i) corresponds to claim (a) whereas (ii) corresponds to claim (b) for the loss functions $l_n(\theta, d) = \sqrt{n}(|\theta - d| - |\theta|)$. But the Bayes estimates for $l_n$ and for $l(\theta, d) = |\theta - d|$ must agree whenever $E(|\boldsymbol{\theta}| \mid X_1, \dots, X_n) < \infty$. (Note that if $E(|\boldsymbol{\theta}| \mid X_1, \dots, X_n) = \infty$, then the posterior Bayes risk under $l$ is infinite and all estimates are equally poor.) Hence, (5.5.25) follows. The proof of a corresponding claim for quadratic loss is sketched in Problem 5.5.5.

***Proof.*** By Theorem 5.5.2 and Polya's theorem (A.14.22)

$$\sup_x |P[\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}) \le x \mid X_1, \dots, X_n) - \Phi(x\sqrt{I(\theta)})| \to 0 \text{ a.s. } P_\theta. \tag{5.5.26}$$

But uniform convergence of distribution functions implies convergence of quantiles that are unique for the limit distribution (Problem B.7.11). Thus, any median of the posterior distribution of $\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta})$ tends to 0, the median of $\mathcal{N}(0, I^{-1}(\theta))$, a.s. $P_\theta$. But the median of the posterior of $\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta})$ is $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$, and (5.5.22) follows. To prove (5.5.24) note that

$$|\sqrt{n}(|\boldsymbol{\theta} - \widehat{\theta}| - |\boldsymbol{\theta} - \widehat{\theta}^*|)| \le \sqrt{n}|\widehat{\theta} - \widehat{\theta}^*|$$

and, hence, that

$$E(\sqrt{n}(|\boldsymbol{\theta} - \widehat{\theta}| - |\boldsymbol{\theta} - \widehat{\theta}^*|) \mid X_1, \dots, X_n) \le \sqrt{n}|\widehat{\theta} - \widehat{\theta}^*| \to 0 \tag{5.5.27}$$

a.s. $P_\theta$, for all $\theta$. Because a.s. convergence $P_\theta$ for all $\theta$ implies a.s. convergence $P$ (B.?), claim (5.5.24) follows and, hence,

$$E(\sqrt{n}(|\boldsymbol{\theta} - \widehat{\theta}| - |\boldsymbol{\theta}|) \mid X_1, \dots, X_n) = E(\sqrt{n}(|\boldsymbol{\theta} - \widehat{\theta}^*| - |\boldsymbol{\theta}|) \mid X_1, \dots, X_n) + o_P(1). \tag{5.5.28}$$

Because by Problem 1.4.7 and Proposition 3.2.1, $\widehat{\theta}^*$ is the Bayes estimate for $l_n(\theta, d)$, (5.5.25) and the theorem follows. □

**Remark.** In fact, Bayes procedures can be efficient in the sense of Sections 5.4.3 and 6.2.3 even if MLEs do not exist. See Le Cam and Yang (1990).

### Bayes credible regions

There is another result illustrating that the frequentist inferential procedures based on $\widehat{\theta}$ agree with Bayesian procedures to first order.

**Theorem 5.5.4.** *Suppose the conditions of Theorem* 5.5.2 *are satisfied. Let*

$$C_n(X_1, \dots, X_n) = \{\theta : \pi(\theta \mid X_1, \dots, X_n) \geq c_n\},$$

*where $c_n$ is chosen so that $\pi(C_n \mid X_1, \dots, X_n) = 1 - \alpha$, be the Bayes credible region defined in Section* 4.7. *Let $I_n(\gamma)$ be the asymptotically level $1 - \gamma$ optimal interval based on $\widehat{\theta}$, given by*

$$I_n(\gamma) = [\widehat{\theta} - d_n(\gamma), \widehat{\theta} + d_n(\gamma)]$$

*where $d_n(\gamma) = z\left(1 - \frac{\gamma}{2}\right)\sqrt{\frac{I(\widehat{\theta})}{n}}$. Then, for every $\epsilon > 0$, $\theta$,*

$$P_\theta[I_n(\alpha + \epsilon) \subset C_n(X_1, \dots, X_n) \subset I_n(\alpha - \epsilon)] \to 1. \tag{5.5.29}$$

The proof, which uses a strengthened version of Theorem 5.5.2 by which the posterior density of $\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta})$ converges to the $\mathcal{N}(0, I^{-1}(\theta))$ density uniformly over compact neighborhoods of $\theta$ for each fixed $\theta$, is sketched in Problem 5.5.6. The message of the theorem should be clear. Bayesian and frequentist coverage statements are equivalent to first order. A finer analysis both in this case and in estimation reveals that any approximations to Bayes procedures on a scale finer than $n^{-1/2}$ do involve the prior. A particular choice, the Jeffrey's prior, makes agreement between frequentist and Bayesian confidence procedures valid even to the higher $n^{-1}$ order (see Schervisch, 1995).

## Testing

Bayes and frequentist inferences diverge when we consider testing a point hypothesis. For instance, in Problem 5.5.1, the posterior probability of $\theta_0$ given $X_1, \dots, X_n$ if $H$ is false is of a different magnitude than the $p$-value for the same data. For more on this so-called Lindley paradox see Berger (1985), Schervisch (1995) and Problem 5.5.1. However, if instead of considering hypotheses specifying one points $\theta_0$ we consider indifference regions where $H$ specifies $[\theta_0 + \Delta)$ or $(\theta_0 - \Delta, \theta_0 + \Delta)$, then Bayes and frequentist testing procedures agree in the limit. See Problem 5.5.2.

**Summary.** Here we established the frequentist consistency of Bayes estimates in the finite parameter case, if all parameter values are a prior possible. Second, we established the so-called Bernstein–von Mises theorem actually dating back to Laplace (see Le Cam and Yang, 1990), which establishes frequentist optimality of Bayes estimates and Bayes optimality of the MLE for large samples and priors that do not rule out any region of the parameter space. Finally, the connection between the behavior of the posterior given by the so-called Bernstein–von Mises theorem and frequentist confidence regions is developed.

## 5.6 PROBLEMS AND COMPLEMENTS

### Problems for Section 5.1

**1.** Suppose $X_1, \dots, X_n$ are i.i.d. as $X \sim F$, where $F$ has median $F^{-1}\left(\frac{1}{2}\right)$ and a continuous case density.

**(a)** Show that, if $n = 2k+1$,

$$\begin{aligned} E_F \text{med}(X_1, \dots, X_n) &= n \begin{pmatrix} 2k \\ k \end{pmatrix} \int_0^1 F^{-1}(t) t^k (1-t)^k dt \\ E_F \text{med}^2(X_1, \dots, X_n) &= n \begin{pmatrix} 2k \\ k \end{pmatrix} \int_0^1 [F^{-1}(t)]^2 t^k (1-t)^k dt \end{aligned}$$

**(b)** Suppose $F$ is uniform, $\mathcal{U}(0,1)$. Find the MSE of the sample median for $n = 1$, $3$, and $5$.

**2.** Suppose $Z \sim \mathcal{N}(\mu, 1)$ and $V$ is independent of $Z$ with distribution $\chi^2_m$. Then $T \equiv Z \Big/ \left(\frac{V}{m}\right)^{\frac{1}{2}}$ is said to have a noncentral $t$ distribution with noncentrality $\mu$ and $m$ degrees of freedom. See Section $4.9.2$.

**(a)** Show that

$$P[T \le t] = 2m \int_0^\infty \Phi(tw - \mu) f_m(mw^2) w dw$$

where $f_m(w)$ is the $\chi^2_m$ density, and $\Phi$ is the normal distribution function.

**(b)** If $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ show that $\sqrt{n}\bar{X} \Big/ \left(\frac{1}{n-1}\sum(X_i - \bar{X})^2\right)^{\frac{1}{2}}$ has a noncentral $t$ distribution with noncentrality parameter $\sqrt{n}\mu/\sigma$ and $n-1$ degrees of freedom.

**(c)** Show that $T^2$ in (a) has a noncentral $\mathcal{F}_{1,m}$ distribution with noncentrality parameter $\mu^2$. Deduce that the density of $T$ is

$$p(t) = 2\sum_{i=0}^\infty P[R = i] \cdot f_{2i+1}(t^2)[\varphi(t-\mu)1(t>0) + \varphi(t+\mu)1(t<0)]$$

where $R$ is given in Problem B.3.12.

*Hint:* Condition on $|T|$.

**3.** Show that if $P[|X| \leq 1] = 1$, then $\text{Var}(X) \leq 1$ with equality iff $X = \pm 1$ with probability $\frac{1}{2}$.
*Hint:* $\text{Var}(X) \leq EX^2$.

**4.** *Comparison of Bounds:* Both the Hoeffding and Chebychev bounds are functions of $n$ and $\epsilon$ through $\sqrt{n}\epsilon$.

**(a)** Show that the ratio of the Hoeffding function $h(\sqrt{n}\epsilon)$ to the Chebychev function $c(\sqrt{n}\epsilon)$ tends to 0 as $\sqrt{n}\epsilon \to \infty$ so that $h(\cdot)$ is arbitrarily better than $c(\cdot)$ in the tails.

**(b)** Show that the normal approximation $2\Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right) - 1$ gives lower results than $h$ in the tails if $P[|X| \leq 1] = 1$ because, if $\sigma^2 \leq 1$, $1 - \Phi(t) \sim \varphi(t)/t$ as $t \to \infty$.
*Note:* Hoeffding (1963) exhibits better bounds for known $\sigma^2$.

**5.** Suppose $\lambda : R \to R$ has $\lambda(0) = 0$, is bounded, and has a bounded second derivative $\lambda''$. Show that if $X_1, \ldots, X_n$ are i.i.d., $EX_1 = \mu$ and Var $X_1 = \sigma^2 < \infty$, then

$$E\lambda(\bar{X} - \mu) = \lambda'(0)\frac{\sigma}{\sqrt{n}}\sqrt{\frac{2}{\pi}} + O\left(\frac{1}{n}\right) \text{ as } n \to \infty.$$

*Hint:* $\sqrt{n}E(\lambda(|\bar{X} - \mu|) - \lambda(0)) = E\lambda'(0)\sqrt{n}|\bar{X} - \mu| + E\left(\frac{\lambda''}{2}(\widetilde{X} - \mu)(\bar{X} - \mu)^2\right)$ where $|\widetilde{X} - \mu| \leq |\bar{X} - \mu|$. The last term is $\leq \sup_x |\lambda''(x)|\sigma^2/2n$ and the first tends to $\lambda'(0)\sigma\int_{-\infty}^{\infty} |z|\varphi(z)dz$ by Remark B.7.1.

### Problems for Section 5.2

**1.** Using the notation of Theorem 5.2.1, show that

$$\sup\{P_{\mathbf{p}}(|\widehat{p}_n - \mathbf{p}| \geq \delta) : \mathbf{p} \in \mathcal{S}\} \leq k/4n\delta^2.$$

**2.** Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Show that for all $n \geq 1$, all $\epsilon > 0$

$$\sup_{\sigma} P_{(\mu,\sigma)}[|\bar{X} - \mu| \geq \epsilon] = 1.$$

*Hint:* Let $\sigma \to \infty$.

**3.** Establish (5.2.5).
*Hint:* $|\widehat{q}_n - q(\mathbf{p})| \geq \epsilon \Rightarrow |\widehat{p}_n - \mathbf{p}| \geq \omega^{-1}(\epsilon)$.

**4.** Let $(U_i, V_i)$, $1 \leq i \leq n$, be i.i.d. $\sim P \in \mathcal{P}$.

**(a)** Let $\gamma(P) = P[U_1 > 0, V_1 > 0]$. Show that if $P = \mathcal{N}(0, 0, 1, 1, \rho)$, then

$$\rho = \sin 2\pi\left(\gamma(P) - \frac{1}{4}\right).$$

**(b)** Deduce that if $P$ is the bivariate normal distribution, then

$$\widetilde{\rho} \equiv \sin\left\{2\pi\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i > \bar{X})1(Y_i > \bar{Y})\right)\right\}$$

is a consistent estimate of $\rho$.

**(c)** Suppose $\rho(P)$ is defined generally as $\text{Cov}_P(U, V)/\sqrt{\text{Var}_P U \ \text{Var}_P V}$ for $P \in \mathcal{P} = \{P : E_P U^2 + E_P V^2 < \infty, \ \text{Var}_P U \ \text{Var}_P V > 0\}$. Show that the sample correlation coefficient continues to be a consistent estimate of $\rho(P)$ but $\widetilde{\rho}$ is no longer consistent.

**5.** Suppose $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$ where $\sigma_0$ is known and $\rho(\mathbf{x}, \mu) = -\log p(\mathbf{x}, \mu)$.

**(a)** Show that condition $(5.2.8)$ fails even in this simplest case in which $\bar{X} \xrightarrow{P} \mu$ is clear.
*Hint:* $\sup_\mu \left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(X_i-\mu)^2}{\sigma_0^2} - \left(1 + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\right)\right| = \infty$.

**(b)** Show that condition $(5.2.14)$(i),(ii) holds.
*Hint:* $K$ can be taken as $[-A, A]$, where $A$ is an arbitrary positive and finite constant.

**6.** Prove that $(5.2.14)$(i) and (ii) suffice for consistency.

**7.** (Wald) Suppose $\theta \to \rho(X, \theta)$ is continuous, $\theta \in R$ and

(i) For some $\epsilon(\theta_0) > 0$

$$E_{\theta_0} \sup\{|\rho(X, \theta') - \rho(X, \theta)| : |\theta - \theta'| \leq \epsilon(\theta_0)\} < \infty.$$

(ii) $E_{\theta_0} \inf\{\rho(X, \theta) - \rho(X, \theta_0) : |\theta - \theta_0| \geq A\} > 0$ for some $A < \infty$.

Show that the minimum contrast estimate $\widehat{\theta}$ is consistent.
*Hint:* From continuity of $\rho$, (i), and the dominated convergence theorem,

$$\lim_{\delta \to 0} E_{\theta_0} \sup\{|\rho(X, \theta') - \rho(X, \theta)| : \theta' \in S(\theta, \delta)\} = 0$$

where $S(\theta, \delta)$ is the $\delta$ ball about $\theta$. Therefore, by the basic property of minimum contrast estimates, for each $\theta \neq \theta_0$, and $\epsilon > 0$ there is $\delta(\theta) > 0$ such that

$$E_{\theta_0} \inf\{\rho(X, \theta') - \rho(X, \theta_0) : \theta' \in S(\theta, \delta(\theta))\} > \epsilon.$$

By compactness there is a finite number $\theta_1, \dots, \theta_r$ of sphere centers such that

$$K \cap \{\theta : |\theta - \theta_0| \geq \lambda\} \subset \bigcup_{j=1}^{r} S(\theta_j, \delta(\theta_j)).$$

Now

$$\inf\left\{\frac{1}{n}\sum_{i=1}^{n}\{\rho(X_i, \theta) - \rho(X_i, \theta_0)\} : \theta \in K \cap \{\theta : |\theta - \theta_0| \geq \lambda\right\}$$

$$\geq \min_{1\leq j\leq r}\left\{\frac{1}{n}\sum_{i=1}^{n}\inf\{\rho(X_i,\theta')-\rho(X_i,\theta_0)\} : \theta'\in S(\theta_j,\delta(\theta_j))\right\}.$$

For $r$ fixed apply the law of large numbers.

**8.** The condition of Problem 7(ii) can also fail. Let $X_i$ be i.i.d. $\mathcal{N}(\mu,\sigma^2)$. Compact sets $K$ can be taken of the form $\{|\mu|\leq A, \epsilon\leq\sigma\leq 1/\epsilon,\ \epsilon>0\}$. Show that the log likelihood tends to $\infty$ as $\sigma\to 0$ and the condition fails.

**9.** Indicate how the conditions of Problem 7 have to be changed to ensure uniform consistency on $K$.

**10.** Extend the result of Problem 7 to the case $\theta\in R^p$, $p>1$.

## Problems for Section 5.3

**1.** Establish $(5.3.9)$ in the exponential model of Example $5.3.1$.

**2.** Establish $(5.3.3)$ for $j$ odd as follows:

(i) Suppose $X_1',\dots,X_n'$ are i.i.d. with the same distribution as $X_1,\dots,X_n$ but independent of them, and let $\bar{X}'=n^{-1}\Sigma X_i'$. Then $E|\bar{X}-\mu|^j\leq E|\bar{X}-\bar{X}'|^j$.

(ii) If $\epsilon_i$ are i.i.d. and take the values $\pm 1$ with probability $\frac{1}{2}$, and if $c_1,\dots,c_n$ are constants, then by Jensen's inequality, for some constants $M_j$,

$$E\left|\sum_{i=1}^{n}c_i\epsilon_i\right|^j\leq E^{\frac{j}{j+1}}\left(\sum_{i=1}^{n}c_i\epsilon_i\right)^{j+1}\leq M_j\left(\sum_{i=1}^{n}c_i^2\right)^{\frac{j}{2}}.$$

(iii) Condition on $|X_i-X_i'|$, $i=1,\dots,n$, in (i) and apply (ii) to get

(iv) $E\left|\sum_{i=1}^n(X_i-X_i')\right|^j\leq M_jE\left[\sum_{i=1}^n(X_i-X_i')^2\right]^{\frac{j}{2}}\leq$ $M_jn^{\frac{j}{2}}E\left[\frac{1}{n}\sum_{i=1}^n(X_i-X_i')^2\right]^{\frac{j}{2}}\leq M_jn^{\frac{j}{2}}E\left(\frac{1}{n}\sum|X_i-X_i'|^j\right)\leq M_jn^{\frac{j}{2}}E|X_1-\mu|^j$.

**3.** Establish $(5.3.11)$.
*Hint:* See part (a) of the proof of Lemma 5.3.1.

**4.** Establish Theorem 5.3.2. *Hint:* Taylor expand and note that if $i_1+\cdots+i_d=m$

$$\begin{aligned}E\left|\prod_{k=1}^{d}(\bar{Y}_k-\mu_k)^{i_k}\right| &\leq m^{m+1}\sum_{k=1}^{d}E|\bar{Y}_k-\mu_k|^m\\ &\leq C_mn^{-m/2}.\end{aligned}$$

Suppose $a_j \geq 0$, $1 \leq j \leq d$, $\sum_{j=1}^{d} i_j = m$, then

$$\begin{aligned} a_1^{i_1}, \ldots, a_d^{i_d} &\leq [\max(a_1, \ldots, a_d)]^m \leq \left(\sum_{j=1}^{d} a_j\right)^m \\ &\leq m^{m-1} \sum_{j=1}^{d} a_j^m. \end{aligned}$$

**5.** Let $X_1, \ldots, X_n$ be i.i.d. $R$ valued with $EX_1 = 0$ and $E|X_1|^j < \infty$. Show that

$$\sup\{|E(X_{i_1}, \ldots, X_{i_j})| : 1 \leq i_k \leq n; k = 1, \ldots, n\} = E|X_1|^j.$$

**6.** Show that if $E|X_1|^j < \infty$, $j \geq 2$, then $E|X_1 - \mu|^j \leq 2^j E|X_1|^j$.
*Hint:* By the iterated expectation theorem

$$\begin{aligned} E|X_1 - \mu|^j &= E\{|X_1 - \mu|^j \mid |X_1| \geq |\mu|\} P(|X_1| \geq |\mu|) \\ &\quad + E\{|X_1 - \mu|^j \mid |X_1| < |\mu|\} P(|X_1| < |\mu|). \end{aligned}$$

**7.** Establish 5.3.28.

**8.** Let $X_1, \ldots, X_{n_1}$ be i.i.d. $F$ and $Y_1, \ldots, Y_{n_2}$ be i.i.d. $G$, and suppose the $X$'s and $Y$'s are independent.

**(a)** Show that if $F$ and $G$ are $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, respectively, then the LR test of $H : \sigma_1^2 = \sigma_2^2$ versus $K : \sigma_1^2 \neq \sigma_2^2$ is based on the statistic $s_1^2/s_2^2$, where $s_1^2 = (n_1 - 1)^{-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$, $s_2^2 = (n_2 - 1)^{-1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$.

**(b)** Show that when $F$ and $G$ are normal as in part (a), then $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ has an $\mathcal{F}_{k,m}$ distribution with $k = n_1 - 1$ and $m = n_2 - 1$.

**(c)** Now suppose that $F$ and $G$ are not necessarily normal but that

$$G \in \mathcal{G} = \left\{ F\left(\frac{\cdot - a}{b}\right) : a \in R,\ b > 0 \right\}$$

and that $0 < \text{Var}(X_1^2) < \infty$. Show that if $m = \lambda k$ for some $\lambda > 0$ and

$$c_{k,m} = 1 + \sqrt{\frac{\kappa(k+m)}{km}} z_{1-\alpha},\ \kappa = \text{Var}[(X_1 - \mu_1)/\sigma_1]^2,\ \mu_1 = E(X_1),\ \sigma_1^2 = \text{Var}(X_1).$$

Then, under $H : \text{Var}(X_1) = \text{Var}(Y_1)$, $P(s_1^2/s_2^2 \leq c_{k,m}) \to 1 - \alpha$ as $k \to \infty$.

**(d)** Let $\widehat{c}_{k,m}$ be $c_{k,m}$ with $\kappa$ replaced by its method of moments estimate. Show that under the assumptions of part (c), if $0 < EX_1^8 < \infty$, $P_H(s_1^2/s_2^2 \leq \widehat{c}_{k,m}) \to 1 - \alpha$ as $k \to \infty$.

**(e)** Next drop the assumption that $G \in \mathcal{G}$. Instead assume that $0 < \text{Var}(Y_1^2) < \infty$. Under the assumptions of part (c), use a normal approximation to find an approximate critical value $q_{k,m}$ (depending on $\kappa_1 = \text{Var}[(X_1 - \mu_1)/\sigma_1]^2$ and $\kappa_2 = \text{Var}[(X_2 - \mu_2)/\sigma_2]^2$ such that $P_H(s_1^2/s_2^2 \le q_{k,m}) \to 1 - \alpha$ as $k \to \infty$.

**(f)** Let $\widehat{q}_{k,m}$ be $q_{k,m}$ with $\kappa_1$ and $\kappa_2$ replaced by their method of moment estimates. Show that under the assumptions of part (e), if $0 < EX_1^8 < \infty$ and $0 < EY_1^8 < \infty$, then $P(s_1^2/s_2^2 \le \widehat{q}_{k,m}) \to 1 - \alpha$ as $k \to \infty$.

**9.** In Example 5.3.6, show that

**(a)** If $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\sqrt{n}((\widehat{C} - \rho), (\widehat{\sigma}_1^2 - 1), (\widehat{\sigma}_2^2 - 1))^T$ has the same asymptotic distribution as $n^{\frac{1}{2}}[n^{-1}\Sigma X_i Y_i - \rho, n^{-1}\Sigma X_i^2 - 1, n^{-1}\Sigma Y_i^2 - 1]^T$.

**(b)** If $(X, Y) \sim \mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then $\sqrt{n}(r^2 - \rho^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\rho^2(1 - \rho^2)^2)$ and, if $\rho \ne 0$, then $\sqrt{n}(r - \rho) \to \mathcal{N}(0, (1 - \rho^2)^2)$.

**(c)** Show that if $\rho = 0$, $\sqrt{n}(r - \rho) \to \mathcal{N}(0, 1)$.

*Hint:* Use the central limit theorem and Slutsky's theorem. Without loss of generality, $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$.

**10.** Show that $\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$ is the variance stabilizing transformation for the correlation coefficient in Example 5.3.6.

*Hint:* Write $\frac{1}{(1-\rho)^2} = \frac{1}{2}\left(\frac{1}{1-\rho} + \frac{1}{1+\rho}\right)$.

**11.** In survey sampling the *model-based* approach postulates that the population $\{x_1, \dots, x_N\}$ or $\{(u_1, x_1), \dots, (u_N, x_N)\}$ we are interested in is itself a sample from a superpopulation that is known up to parameters; that is, there exists $T_1, \dots, T_N$ i.i.d. $P_\theta$, $\theta \in \Theta$ such that $T_i = t_i$ where $t_i \equiv (u_i, x_i)$, $i = 1, \dots, N$. In particular, suppose in the context of Example 3.4.1 that we use $T_{i_1}, \dots, T_{i_n}$, which we have sampled at random from $\{t_1, \dots, t_N\}$, to estimate $\bar{x} \equiv \frac{1}{N}\sum_{i=1}^N x_i$. Without loss of generality, suppose $i_j = j$, $1 \le j \le n$. Consider as estimates

(i) $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ when $T_i \equiv (U_i, X_i)$.

(ii) $\widehat{\bar{X}}_R = \bar{X} - b_{\text{opt}}(\bar{U} - \bar{u})$ as in Example 3.4.1.

Show that, if $\frac{n}{N} \to \lambda$ as $N \to \infty$, $0 < \lambda < 1$, and if $EX_1^2 < \infty$ (in the supermodel), then

**(a)** $\sqrt{n}(\bar{X} - \bar{x}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau^2(1 - \lambda))$ where $\tau^2 = \text{Var}(X_1)$.

**(b)** Suppose $P_\theta$ is such that $X_i = bU_i + \epsilon_i$, $i = 1, \dots, N$ where the $\epsilon_i$ are i.i.d. and independent of the $U_i$, $E\epsilon_i = 0$, $\text{Var}(\epsilon_i) = \sigma^2 < \infty$ and $\text{Var}(U_i) > 0$. Show that $\sqrt{n}(\bar{X}_R - \bar{x}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1 - \lambda)\sigma^2)$, $\sigma^2 < \tau^2$.

*Hint:* (a) $\bar{X} - \bar{x} = \left(1 - \frac{n}{N}\right)(\bar{X} - \bar{X}^c)$ where $\bar{X}^c = \frac{1}{N-n}\sum_{i=n+1}^N X_i$.

(b) Use the multivariate delta method and note $(\widehat{b}_{\text{opt}} - b)(\bar{U} - \bar{u}) = o_p(n^{-\frac{1}{2}})$.

**12.** **(a)** Let $\bar{Y}_k$ be as defined in Theorem 5.3.2. Suppose that $E|\mathbf{Y}_1|^3 < \infty$. Show that $|E(\bar{Y}_a - \mu_a)(\bar{Y}_b - \mu_b)(\bar{Y}_c - \mu_c)| \le Mn^{-2}$; $a, b, c \in \{1, \dots, d\}$.

**(b)** Assume $EY_1^4 < \infty$. Deduce formula (5.3.14).

*Hint:* **(a)** If $U$ is independent of $(V, W)$, $EU = 0$, $E(WV) < \infty$, then $E(UVW) = 0$.

**13.** Let $S_n$ have a $\chi_n^2$ distribution.

**(a)** Show that if $n$ is large, $\sqrt{S_n} - \sqrt{n}$ has approximately a $\mathcal{N}(0, \frac{1}{2})$ distribution. This is known as *Fisher's approximation*.

**(b)** From (a) deduce the approximation $P[S_n \le x] \approx \Phi(\sqrt{2x} - \sqrt{2n})$.

**(c)** Compare the approximation of (b) with the central limit approximation $P[S_n \le x] = \Phi((x - n)/\sqrt{2n})$ and the exact values of $P[S_n \le x]$ from the $\chi^2$ table for $x = x_{0.90}$, $x = x_{0.99}$, $n = 5, 10, 25$. Here $x_q$ denotes the $q$th quantile of the $\chi_n^2$ distribution.

**14.** Suppose $X_1, \dots, X_n$ is a sample from a population with mean $\mu$, variance $\sigma^2$, and third central moment $\mu_3$. Justify formally

$$E[h(\bar{X}) - E(h(\bar{X}))]^3 = \frac{1}{n^2}[h'(\mu)]^3\mu_3 + \frac{3}{n^2}h''(\mu)[h'(\mu)]^2\sigma^4 + O(n^{-3}).$$

**15.** It can be shown (under suitable conditions) that the normal approximation to the distribution of $h(\bar{X})$ improves as the coefficient of skewness $\gamma_{1n}$ of $h(\bar{X})$ diminishes.

**(a)** Use this fact and Problem 5.3.14 to explain the numerical results of Problem 5.3.13(c).

**(b)** Let $S_n \sim \chi_n^2$. The following approximation to the distribution of $S_n$ (due to Wilson and Hilferty, 1931) is found to be excellent

$$P[S_n \le x] \approx \Phi\left\{\left[\left(\frac{x}{n}\right)^{1/3} - 1 + \frac{2}{9n}\right]\sqrt{\frac{9n}{2}}\right\}.$$

Use (5.3.6) to explain why.

**16.** *Normalizing Transformation for the Poisson Distribution.* Suppose $X_1, \dots, X_n$ is a sample from a $\mathcal{P}(\lambda)$ distribution.

**(a)** Show that the only transformations $h$ that make $E[h(\bar{X}) - E(h(\bar{X}))]^3 = 0$ to terms up to order $1/n^2$ for all $\lambda > 0$ are of the form $h(t) = ct^{2/3} + d$.

**(b)** Use (a) to justify the approximation

$$P\left[\bar{X} \le \frac{k}{n}\right] \approx \Phi\left\{\sqrt{n}\left[\left(\frac{k + \frac{1}{2}}{n}\right)^{2/3} - \lambda^{2/3}\right]\Big/\frac{2}{3}\lambda^{1/6}\right\}.$$

**17.** Suppose $X_1, \dots, X_n$ are independent, each with Hardy–Weinberg frequency function $f$ given by

| $x$ | $0$ | $1$ | $2$ |
|---|---|---|---|
| $f(x)$ | $\theta^2$ | $2\theta(1-\theta)$ | $(1-\theta)^2$ |

where $0 < \theta < 1$.

**(a)** Find an approximation to $P[\bar{X} \leq t]$ in terms of $\theta$ and $t$.

**(b)** Find an approximation to $P[\sqrt{\bar{X}} \leq t]$ in terms of $\theta$ and $t$.

**(c)** What is the approximate distribution of $\sqrt{n}(\bar{X} - \mu) + \bar{X}^2$, where $\mu = E(X_1)$?

**18.** *Variance Stabilizing Transformation for the Binomial Distribution.* Let $X_1, \dots, X_n$ be the indicators of $n$ binomial trials with probability of success $\theta$. Show that the only variance stabilizing transformation $h$ such that $h(0) = 0$, $h(1) = 1$, and $h'(t) \geq 0$ for all $t$, is given by $h(t) = (2/\pi)\sin^{-1}(\sqrt{t})$.

**19.** Justify formally the following expressions for the moments of $h(\bar{X}, \bar{Y})$ where $(X_1, Y_1), \dots, (X_n, Y_n)$ is a sample from a bivariate population with $E(X) = \mu_1$, $E(Y) = \mu_2$, Var$(X) = \sigma_1^2$, Var$(Y) = \sigma_2^2$, Cov$(X, Y) = \rho\sigma_1\sigma_2$.

**(a)**

$$E(h(\bar{X}, \bar{Y})) = h(\mu_1, \mu_2) + O(n^{-1}).$$

**(b)**

$$\begin{aligned}\text{Var}(h(\bar{X}, \bar{Y})) \cong\ & \frac{1}{n}\{[h_1(\mu_1, \mu_2)]^2\sigma_1^2 \\ & +2h_1(\mu_1, \mu_2)h_2(\mu_1, \mu_2)\rho\sigma_1\sigma_2 + [h_2(\mu_1, \mu_2)]^2\sigma_2^2\} + 0(n^{-2})\end{aligned}$$

where

$$h_1(x, y) = \frac{\partial}{\partial x}h(x, y),\ h_2(x, y) = \frac{\partial}{\partial y}h(x, y).$$

*Hint:* $h(\bar{X}, \bar{Y}) - h(\mu_1, \mu_2) = h_1(\mu_1, \mu_2)(\bar{X} - \mu_1) + h_2(\mu_1, \mu_2)(\bar{Y} - \mu_2) + 0(n^{-1})$.

**20.** Let $B_{m,n}$ have a beta distribution with parameters $m$ and $n$, which are integers. Show that if $m$ and $n$ are both tending to $\infty$ in such a way that $m/(m + n) \to \alpha$, $0 < \alpha < 1$, then

$$P\left[\sqrt{m+n}\frac{(B_{m,n} - m/(m+n))}{\sqrt{\alpha(1-\alpha)}} \leq x\right] \to \Phi(x).$$

*Hint:* Use $B_{m,n} = (m\bar{X}/n\bar{Y})[1 + (m\bar{X}/n\bar{Y})]^{-1}$ where $X_1, \dots, X_m$, $Y_1, \dots, Y_n$ are independent standard exponentials.

**21.** Show directly using Problem B.2.5 that under the conditions of the previous problem, if $m/(m + n) - \alpha$ tends to zero at the rate $1/(m + n)^2$, then

$$E(B_{m,n}) = \frac{m}{m+n},\ \text{Var}\ B_{m,n} = \frac{\alpha(1-\alpha)}{m+n} + O((m+n)^{-2}).$$

**22.** Let $S_n \sim \chi^2_n$. Use Stirling's approximation and Problem B.2.4 to give a direct justification of

$$E(\sqrt{S_n}) = \sqrt{n} + R_n$$

where $R_n/\sqrt{n} \to 0$ as in $n \to \infty$. Recall *Stirling's approximation:*

$$\Gamma(p+1)/(\sqrt{2\pi}e^{-p}p^{p+\frac{1}{2}}) \to 1 \text{ as } p \to \infty.$$

(It may be shown but is not required that $|\sqrt{n}R_n|$ is bounded.)

**23.** Suppose that $X_1, \dots, X_n$ is a sample from a population and that $h$ is a real-valued function of $\bar{X}$ whose derivatives of order $k$ are denoted by $h^{(k)}$, $k > 1$. Suppose $|h^{(4)}(x)| \le M$ for all $x$ and some constant $M$ and suppose that $\mu_4$ is finite. Show that $Eh(\bar{X}) = h(\mu) + \frac{1}{2}h^{(2)}(\mu)\frac{\sigma^2}{n} + R_n$ where $|R_n| \le h^{(3)}(\mu)|\mu_3|/6n^2 + M(\mu_4 + 3\sigma^2)/24n^2$.

*Hint:*

$$\left| h(x) - h(\mu) - h^{(1)}(\mu)(x-\mu) - \frac{h^{(2)}(\mu)}{2}(x-\mu)^2 - \frac{h^{(3)}(\mu)}{6}(x-\mu)^3 \right| \le \frac{M}{24}(x-\mu)^4.$$

Therefore,

$$\begin{aligned} &\left| Eh(\bar{X}) - h(\mu) - h^{(1)}(\mu)E(\bar{X}-\mu) - \frac{h^{(2)}(\mu)}{2}E(\bar{X}-\mu)^2 \right| \\ &\le \frac{|h^{(3)}(\mu)|}{6}|E(\bar{X}-\mu)^3| + \frac{M}{24}E(\bar{X}-\mu)^4 \\ &\le \frac{|h^{(3)}(\mu)|}{6}\frac{|\mu_3|}{n^2} + \frac{M}{24}\frac{(\mu_4 + 3\sigma^4)}{n^2}. \end{aligned}$$

**24.** Let $X_1, \dots, X_n$ be a sample from a population with mean $\mu$ and variance $\sigma^2 < \infty$. Suppose $h$ has a second derivative $h^{(2)}$ continuous at $\mu$ and that $h^{(1)}(\mu) = 0$.

**(a)** Show that $\sqrt{n}[h(\bar{X}) - h(\mu)] \to 0$ while $n[h(\bar{X} - h(\mu)]$ is asymptotically distributed as $\frac{1}{2}h^{(2)}(\mu)\sigma^2 V$ where $V \sim \chi^2_1$.

**(b)** Use part (a) to show that when $\mu = \frac{1}{2}$, $n[\bar{X}(1-\bar{X}) - \mu(1-\mu)] \xrightarrow{\mathcal{L}} -\sigma^2 V$ with $V \sim \chi^2_1$. Give an approximation to the distribution of $\bar{X}(1-\bar{X})$ in terms of the $\chi^2_1$ distribution function when $\mu = \frac{1}{2}$.

**25.** Let $X_1, \dots, X_n$ be a sample from a population with $\sigma^2 = \text{Var}(X) < \infty$, $\mu = E(X)$ and let $T = \bar{X}^2$ be an estimate of $\mu^2$.

**(a)** When $\mu \ne 0$, find the asymptotic distribution of $\sqrt{n}(T - \mu^2)$ using the delta method.

**(b)** When $\mu = 0$, find the asymptotic distriution of $nT$ using $P(nT \le t) = P(-\sqrt{t} \le \sqrt{n}\bar{X} \le \sqrt{t})$. Compare your answer to the answer in part (a).

**(c)** Find the limiting laws of $\sqrt{n}(\bar{X} - \mu)^2$ and $n(\bar{X} - \mu)^2$.

**26.** Show that if $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then

$$\sqrt{n}(\bar{X} - \mu, \widehat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 0, \Sigma_0)$$

where $\Sigma_0 = \text{diag}(\sigma^2, 2\sigma^4)$.

*Hint:* Use (5.3.33) and Theorem 5.3.4.

**27.** Suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ are $n$ sets of control and treatment responses in a matched pair experiment. Assume that the observations have a common $\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution. We want to obtain confidence intervals on $\mu_2 - \mu_1 = \Delta$. Suppose that instead of using the one-sample $t$ intervals based on the differences $Y_i - X_i$ we treat $X_1, \dots, X_n$, $Y_1, \dots, Y_n$ as separate samples and use the two-sample $t$ intervals (4.9.3). What happens? Analysis for fixed $n$ is difficult because $T(\Delta)$ no longer has a $\mathcal{T}_{2n-2}$ distribution. Let $n \to \infty$ and

**(a)** Show that $P[T(\Delta) \le t] \to \Phi\left(t\left[1 - \frac{2\sigma_1\sigma_2\rho}{(\sigma_1^2+\sigma_2^2)}\right]^{-\frac{1}{2}}\right)$.

**(b)** Deduce that if $\rho > 0$ and $I_n$ is given by (4.9.3), then $\lim_n P[\Delta \in I_n] > 1 - \alpha$.

**(c)** Show that if $|I_n|$ is the length of the interval $I_n$,

$$\sqrt{n}|I_n| \to 2\sqrt{\sigma_1^2 + \sigma_2^2}\, z(1 - \frac{1}{2}\alpha) > 2(\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2})z(1 - \frac{1}{2}\alpha)$$

where the right-hand side is the limit of $\sqrt{n}$ times the length of the one-sample $t$ interval based on the differences.

*Hint:* (a), (c) Apply Slutsky's theorem.

**28.** Suppose $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ are as in Section 4.9.3 independent samples with $\mu_1 = E(X_1)$, $\sigma_1^2 = \text{Var}(X_1)$, $\mu_2 = E(Y_1)$, and $\sigma_2^2 = \text{Var}(Y_1)$. We want to study the behavior of the two-sample pivot $T(\Delta)$ of Example 4.9.3, if $n_1, n_2 \to \infty$, so that $n_1/n \to \lambda$, $0 < \lambda < 1$.

**(a)** Show that $P[T(\Delta) \le t] \to \Phi(t[(\lambda\sigma_1^2 + (1-\lambda)\sigma_2^2)/((1-\lambda)\sigma_1^2 + \lambda\sigma_2^2)]^{\frac{1}{2}})$.

**(b)** Deduce that if $\lambda = \frac{1}{2}$ or $\sigma_1 = \sigma_2$, the intervals (4.9.3) have correct asymptotic probability of coverage.

**(c)** Show that if $\sigma_2^2 > \sigma_1^2$ and $\lambda > 1 - \lambda$, the interval (4.9.3) has asymptotic probability of coverage $< 1 - \alpha$, whereas the situation is reversed if the sample size inequalities and variance inequalities agree.

**(d)** Make a comparison of the asymptotic length of (4.9.3) and the intervals based on the pivot $|D - \Delta|/s_D$ where $D$ and $s_D$ are as in Section 4.9.4.

**29.** Let $T = (D - \Delta)/s_D$ where $D$, $\Delta$ and $s_D$ are as defined in Section 4.9.4. Suppose that $E(X_i^4) < \infty$ and $E(Y_j^4) < \infty$.

**(a)** Show that $T$ has asymptotically a standard normal distribution as $n_1 \to \infty$ and $n_2 \to \infty$.

**(b)** Let $k$ be the Welch degrees of freedom defined in Section 4.9.4. Show that $k \xrightarrow{P} \infty$ as $n_1 \to \infty$ and $n_2 \to \infty$.

**(c)** Show using parts (a) and (b) that the tests that reject $H : \mu_1 = \mu_2$ in favor of $K : \mu_2 > \mu_1$ when $T \geq t_k(1-\alpha)$, where $t_k(1-\alpha)$ is the critical value using the Welch approximation, has asymptotic level $\alpha$.

**(d)** Find or write a computer program that carries out the Welch test. Carry out a Monte Carlo study such as the one that led to Figure 5.3.3 using the Welch test based on $T$ rather than the two-sample $t$ test based on $S_n$. Plot your results.

**30.** Generalize Lemma 5.3.1 by showing that if $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \in R^d$ are i.i.d. vectors with zero means and $E|\mathbf{Y}_1|^k < \infty$, where $|\cdot|$ is the Euclidean norm, then for all integers $k \geq 2$:

$$E|\bar{\mathbf{Y}}|^k \leq Cn^{-k/2}$$

where $C$ depends on $d$, $E|\mathbf{Y}_1|^k$ and $k$ only.

*Hint:* If $|\mathbf{x}|_1 = \sum_{j=1}^d |x_j|$, $\mathbf{x} = (x_1, \ldots, x_d)^T$ and $|\mathbf{x}|$ is Euclidean distance, then there exist universal constants $0 < c_d < C_d < \infty$ such that $c_d|\mathbf{x}|_1 \leq |\mathbf{x}| \leq C_d|\mathbf{x}|_1$.

**31.** Let $X_1, \ldots, X_n$ be i.i.d. as $X \sim F$ and let $\mu = E(X)$, $\sigma^2 = \text{Var}(X)$, $\kappa = \text{Var}[(X-\mu)/\sigma]^2$, $s^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Then by Theorem B.3.1, $V_n = (n-1)s^2/\sigma^2$ has a $\chi^2_{n-1}$ distribution when $F$ is the $\mathcal{N}(\mu, \sigma^2)$ distribution.

**(a)** Suppose $E(X^4) < \infty$.

**(b)** Let $x_{n-1}(\alpha)$ be the $\alpha$th quantile of $\chi^2_{n-1}$. Find approximations to $P(V_n \leq x_{n-1}(\alpha))$ and $P(V_n \leq x_{n-1}(1-\alpha))$ and evaluate the approximations when $F$ is $\mathcal{T}_5$.

*Hint:* See Problems B.3.9 and 4.4.16.

**(c)** Let $\widehat{\kappa}$ be the method of moment estimate of $\kappa$ and let

$$\widehat{v}_\alpha = (n-1) + \sqrt{\widehat{\kappa}(n-1)}z(\alpha).$$

Show that if $0 < EX^8 < \infty$, then $P(V_n \leq \widehat{v}_\alpha) \to \alpha$ as $n \to \infty$.

**32.** It may be shown that if $T_n$ is any sequence of random variables such that $T_n \xrightarrow{\mathcal{L}} T$ and if the variances of $T$ and $T_n$ exist, then $\liminf_n \text{Var}(T_n) \geq \text{Var}(T)$. Let

$$T_n = X1[|X| \leq 1 - n^{-1}] + n1[|X| > 1 - n^{-1}]$$

where $X$ is uniform, $\mathcal{U}(-1, 1)$. Show that as $n \to \infty$, $T_n \xrightarrow{\mathcal{L}} X$, but $\text{Var}(T_n) \to \infty$.

**33.** Let $X_{ij} (i = 1, \ldots, p;\ j = 1, \ldots, k)$ be independent with $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$.

**(a)** Show that the MLEs of $\mu_i$ and $\sigma^2$ are

$$\widehat{\mu}_i = k^{-1}\sum_{j=1}^k X_{ij} \text{ and } \widehat{\sigma}^2 = (kp)^{-1}\sum_{i=1}^p \sum_{j=1}^k (X_{ij} - \widehat{\mu}_i)^2.$$

**(b)** Show that if $k$ is fixed and $p \to \infty$, then $\widehat{\sigma}^2 \xrightarrow{P} (k-1)\sigma^2/k$. That is the MLE $\widehat{\mu}^2$ is not consistent (Neyman and Scott, 1948).

**(c)** Give a consistent estimate of $\sigma^2$.

## Problems for Section 5.4

**1.** Let $X_1, \dots, X_n$ be i.i.d. random variables distributed according to $P \in \mathcal{P}$. Suppose $\psi : \boldsymbol{R} \to \boldsymbol{R}$

(i) is monotone nondecreasing

(ii) $\psi(-\infty) < 0 < \psi(\infty)$

(iii) $|\psi|(x) \le M < \infty$ for all $x$.

**(a)** Show that (i), (ii), and (iii) imply that $\theta(P)$ defined (not uniquely) by

$$E_P\psi(X_1 - \theta(P)) \ge 0 \ge E_P\psi(X_1 - \theta'), \text{ all } \theta' > \theta(P)$$

is finite.

**(b)** Suppose that for all $P \in \mathcal{P}$, $\theta(P)$ is the unique solution of $E_P\psi(X_1 - \theta) = 0$. Let $\widehat{\theta}_n = \theta(\widehat{P})$, where $\widehat{P}$ is the empirical distribution of $X_1, \dots, X_n$. Show that $\widehat{\theta}_n$ is consistent for $\theta(P)$ over $\mathcal{P}$. Deduce that the sample median is a consistent estimate of the population median if the latter is unique. (Use $\psi(x) = \text{sgn}(x)$.)

*Hint:* Show that $E_P\psi(X_1 - \theta)$ is nonincreasing in $\theta$. Use the bounded convergence theorem applied to $\psi(X_1 - \theta) \xrightarrow{P} \psi(-\infty)$ as $\theta \to \infty$.

**(c)** Assume the conditions in (a) and (b). Set $\lambda(\theta) = E_P\psi(X_1 - \theta)$ and $\tau^2(\theta) = \text{Var}_P\psi(X_1 - \theta)$. Assume that $\lambda'(\theta) < 0$ exists and that

$$\frac{1}{\sqrt{n}\tau(\theta)}\sum_{i=1}^{n}[\psi(X_i - \theta_n) - \lambda(\theta_n)] \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

for every sequence $\{\theta_n\}$ with $\theta_n = \theta + t/\sqrt{n}$ for $t \in R$. Show that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2(\theta)}{[\lambda'(\theta)]^2}\right).$$

*Hint:* $P(\sqrt{n}(\widehat{\theta}_n - \theta)) < t) = P(\widehat{\theta}_n < \theta_n) = P\left(-\sum_{i=1}^{n}\psi(X_i - \theta_n) < 0\right)$.

**(d)** Assume part (c) and A6. Show that if $f'(x) = F''(x)$ exists, then $\lambda'(\theta) = \text{Cov}(\psi(X_1 - \theta), \frac{\partial}{\partial\theta}\log f(X_1 - \theta))$.

**(e)** Suppose that the d.f. $F(x)$ of $X_1$ is continuous and that $f(\theta) = F'(\theta) > 0$ exists. Let $\widehat{X}$ denote the sample median. Show that, under the conditions of (c), $\sqrt{n}(\widehat{X} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1/4f^2(\theta))$.

**(f)** For two estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ with $\sqrt{n}(\widehat{\theta}_j - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_j^2)$, $j = 1, 2$, the *asymptotic relative efficiency* of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$ is defined as $e_P(\widehat{\theta}_1, \widehat{\theta}_2) = \sigma_2^2/\sigma_1^2$. Show that if $P$ is $\mathcal{N}(\mu, \sigma^2)$, then $e_P(\bar{X}, \widehat{X}) = \pi/2$.

**(g)** Suppose $X_1$ has the gross error density $f_\epsilon(x - \theta)$ (see Section 3.5) where

$$f_\epsilon(x) = (1-\epsilon)\varphi_\sigma(x) + \epsilon\varphi_\tau(x),\ 0 \le \epsilon \le 0.5$$

and $\varphi_\sigma$ denotes the $\mathcal{N}(0, \sigma^2)$ density. Find the efficiency $e_P(\bar{X}, \widehat{X})$ as defined in (f). If $\sigma = 1$, $\tau = 4$, evaluate the efficiency for $\epsilon = .05$, $0.10$, $0.15$ and $0.20$ and note that $\widehat{X}$ is more efficient than $\bar{X}$ for these gross error cases.

**(h)** Suppose that $X_1$ has the $t$ distribution with $k \ge 3$ degrees of freedom. Find $e_P(\bar{X}, \widehat{X})$.

**2.** Show that assumption A4$'$ of this section coupled with A0–A3 implies assumption A4.

*Hint:*

$$\begin{aligned} &E \sup\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\psi}{\partial\psi}(X_i, t) - \frac{\partial\psi}{\partial\theta}(X_i, \theta(p)\right)\right| : |t - \theta(p)| \le \epsilon_n\right\} \\ &\le E\sup\left\{\left|\frac{\partial\psi}{\partial\theta}(X_1, t) - \frac{\partial\psi}{\partial\theta}(X_1, \theta(p))\right| : |t - \theta(p)| \le \epsilon_n\right\}. \end{aligned}$$

Apply A4 and the dominated convergence theorem B.7.5.

**3.** Show that A6$'$ implies A6.

*Hint:* $\frac{\partial}{\partial\theta}\int \psi(x,\theta)p(x,\theta)d\mu(x) = \int \frac{\partial}{\partial\theta}(\psi(x,\theta)p(x,\theta))d\mu(x)$ if for all $-\infty < a < b < \infty$,

$$\int_a^b \int \frac{\partial}{\partial\theta}(\psi(x,\theta)p(x,\theta)d\mu(x)) = \int \psi(x,b)p(x,b)d\mu(x) - \int \psi(x,a)p(x,a)d\mu(x).$$

Condition A6$'$ permits interchange of the order of integration by Fubini's theorem (Billingsley, 1979) which you may assume.

**4.** Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{U}(0, \theta)$, $\theta > 0$.

**(a)** Show that $\frac{\partial l}{\partial\theta}(x, \theta) = -\frac{1}{\theta}$ for $\theta > x$ and is undefined for $\theta \le x$. Conclude that $\frac{\partial l}{\partial\theta}(X, \theta)$ is defined with $P_\theta$ probability 1 but

$$E_\theta\frac{\partial l}{\partial\theta}(X, \theta) = -\frac{1}{\theta} \ne 0$$

**(b)** Show that if $\widehat{\theta} = \max(X_1, \ldots, X_n)$ is the MLE, then $\mathcal{L}_\theta(n(\theta - \widehat{\theta})) \to \mathcal{E}\,(1/\theta)$. Thus, not only does asymptotic normality not hold but $\widehat{\theta}$ converges to $\theta$ faster than at rate $n^{-1/2}$. See also Problem 3.4.22.

*Hint:* $P_\theta[n(\theta - \widehat{\theta}) \le x] = 1 - \left(1 - \frac{x}{n\theta}\right)^n \to 1 - \exp(-x/\theta)$.

**5. (a)** Show that in Theorem 5.4.5

$$\sum_{i=1}^{n} \log \frac{p\left(X_i, \theta_0 + \frac{\gamma}{\sqrt{n}}\right)}{p(X_i, \theta_0)} = \frac{\gamma}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p(X_i, \theta_0)$$

$$+ \frac{\gamma^2}{2} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p(X_i, \theta_0) + o_p(1)$$

under $P_{\theta_0}$, and conclude that

$$\mathcal{L}_{\theta_0} \left( \sum_{i=1}^{n} \log \frac{p\left(X_i, \theta_0 + \frac{\gamma}{\sqrt{n}}\right)}{p(X_i, \theta_0)} \right) \to \mathcal{N} \left( -\frac{\gamma^2}{2} I(\theta_0), \gamma^2 I(\theta_0) \right).$$

**(b)** Show that

$$\mathcal{L}_{\theta_0 + \frac{\gamma}{\sqrt{n}}} \left( \sum_{i=1}^{n} \log \frac{p\left(X_i, \theta_0 + \frac{\gamma}{\sqrt{n}}\right)}{p(X_i, \theta_0)} \right) \to \mathcal{N} \left( \frac{\gamma^2}{2} I(\theta_0), \gamma^2 I(\theta) \right).$$

**(c)** Prove (5.4.50).
*Hint:* (b) Expand as in (a) but around $\theta_0 + \frac{\gamma}{\sqrt{n}}$.

**(d)** Show that $P_{\theta_0 + \frac{\gamma}{\sqrt{n}}} \left[ \sum_{i=1}^{n} \log \frac{p\left(X_i, \theta_0 + \frac{\gamma}{\sqrt{n}}\right)}{p(X_i, \theta_0)} = c_n \right] \to 0$ for any sequence $\{c_n\}$ by using (b) and Polyà's theorem (A.14.22).

**6.** Show that for the log likelihood ratio statistic

$$\log \Lambda_n = \sum_{i=1}^{n} \log \frac{p(X_i, \widehat{\theta}_n)}{p(X_i, \theta_0)} 1(\widehat{\theta}_n > \theta_0),$$

**(a)** $\mathcal{L}_{\theta_0}(\log \Lambda_n) \to \mathcal{L}(U)$ where $U \sim Z^2 1\ (Z > 0)$ with $Z \sim \mathcal{N}(0, 1)$.

**(b)** $\mathcal{L}_{\theta_0 + \frac{\gamma}{\sqrt{n}}}(\log \Lambda_n) \to \mathcal{L}\left( \frac{1}{2} \left( Z + \gamma I^{\frac{1}{2}}(\theta_0) \right)^2 1[Z > -\gamma I^{\frac{1}{2}}(\theta_0)] \right)$

**(c)** Show that the asymptotic power function of the level $\alpha$ likelihood ratio test achieves equality in (5.4.50).

**7.** Suppose A4$'$, A2, and A6 hold for $\psi = \partial l / \partial \theta$ so that $E_\theta \frac{\partial^2 l}{\partial \theta^2}(X, \theta) = -I(\theta)$ and $I(\theta) < \infty$. Show that $\theta \to I(\theta)$ is continuous.

*Hint:* $\theta \to \frac{\partial^2 l}{\partial \theta^2}(X, \theta)$ is continuous and

$$\sup \left\{ \frac{\partial^2 l}{\partial \theta^2}(X, \theta') : |\theta - \theta'| \leq \epsilon_n \right\} \xrightarrow{P} 0$$

if $\epsilon_n \to 0$. Apply the dominated convergence theorem (B.7.5) to $\frac{\partial^2 l}{\partial \theta^2}(x, \theta) p(x, \theta)$.

**8.** Establish (5.4.54) and (5.4.56).
*Hint:* Use Problems 5.4.5 and 5.4.6.

**9. (a)** Establish (5.4.58).
*Hint:* Use Problem 5.4.7, A.14.6 and Slutsky's theorem.

**(b)** Suppose the conditions of Theorem 5.4.5 hold. Let $\underline{\theta}_n^*$ be as in (5.4.59). Then (5.4.57) can be strengthened to: For each $\theta_0 \in \Theta$, there is a neighborhood $V(\theta_0)$ of $\theta_0$ such that $\overline{\lim}_n \sup\{P_\theta[\underline{\theta}_n^* \leq \theta] : \theta \in V(\theta_0)\} \to 1 - \alpha$.

**10.** Let

$$\widehat{I} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l}{\partial\theta^2}(X_i, \widehat{\theta}).$$

**(a)** Show that under assumptions (A0)–(A6) for $\psi = \frac{\partial l}{\partial\theta}$ at all $\theta$ and (A4′), $\widehat{I}$ is a consistent estimate of $I(\theta)$.

**(b)** Deduce that

$$\widehat{\underline{\theta}}^{**} = \widehat{\theta} - z_{1-\alpha}/\sqrt{n\widehat{I}}$$

is an asymptotic lower confidence bound for $\theta$.

**(c)** Show that if $P_\theta$ is a one parameter exponential family the bound of (b) and (5.4.59) coincide.

**11.** Consider Example 4.4.3, setting a lower confidence bound for binomial $\theta$.

**(a)** Show that, if $\bar{X} = 0$ or 1, the bound (5.4.59), which is just (4.4.7), gives $\widehat{\underline{\theta}}^* = X$. Compare with the exact bound of Example 4.5.2.

**(b)** Compare the bounds in (a) with the bound (5.4.61), which agrees with (4.4.3), and give the behavior of (5.4.61) for $\bar{X} = 0$ and 1.

**12. (a)** Show that under assumptions (A0)–(A6) for all $\theta$ and (A4′),

$$\underline{\theta}_{nj}^* = \underline{\theta}_n^* + o_p(n^{-1/2})$$

for $j = 1, 2$.

**13.** Let $\widetilde{\underline{\theta}}_{n1}, \widetilde{\underline{\theta}}_{n2}$ be two asymptotic level $1 - \alpha$ lower confidence bounds. We say that $\widetilde{\underline{\theta}}_{n1}$ is asymptotically at least as good as $\widetilde{\underline{\theta}}_{n2}$ if, for all $\gamma > 0$

$$\overline{\lim}_n P_\theta\left[\widetilde{\underline{\theta}}_{n1} \leq \theta - \frac{\gamma}{\sqrt{n}}\right] \leq \underline{\lim}_n P_\theta\left[\widetilde{\underline{\theta}}_{n2} \leq \theta - \frac{\gamma}{\sqrt{n}}\right].$$

Show that $\underline{\theta}_{n1}^*$ and, hence, all the $\underline{\theta}_{nj}^*$ are at least as good as any competitors.
*Hint:* Compare Theorem 4.4.2.

**14.** Suppose that $X_1, \ldots, X_n$ are i.i.d. inverse Gaussian with parameters $\mu$ and $\lambda$, where $\mu$ is known. That is, each $X_i$ has density

$$\left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\frac{\lambda x}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2x}\right\};\ x > 0;\ \mu > 0;\ \lambda > 0.$$

**(a)** Find the Neyman–Pearson (NP) test for testing $H : \lambda = \lambda_0$ versus $K : \lambda < \lambda_0$.

**(b)** Show that the NP test is UMP for testing $H : \lambda \geq \lambda_0$ versus $K : \lambda < \lambda_0$.

**(c)** Find the approximate critical value of the Neyman–Pearson test using a normal approximation.

**(d)** Find the Wald test for testing $H : \lambda = \lambda_0$ versus $K : \lambda < \lambda_0$.

**(e)** Find the Rao score test for testing $H : \lambda = \lambda_0$ versus $K : \lambda < \lambda_0$.

**15.** Establish $(5.4.55)$.

*Hint:* By $(3.4.10)$ and $(3.4.11)$, the test statistic is a sum of i.i.d. variables with mean zero and variance $I(\theta)$. Now use the central limit theorem.

## Problems for Section 5.5

**1.** Consider testing $H : \mu = 0$ versus $K : \mu \neq 0$ given $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, 1)$. Consider the Bayes test when $\boldsymbol{\mu}$ is distributed according to $\pi$ such that

$$1 > \pi(\{0\}) = \lambda > 0,\ \pi(\boldsymbol{\mu} \neq 0) = 1 - \lambda$$

and given $\boldsymbol{\mu} \neq 0$, $\boldsymbol{\mu}$ has a $\mathcal{N}(0, \tau^2)$ distribution with density $\varphi_\tau(\mu)$.

**(a)** Show that the posterior probability of $\{0\}$ is

$$\widetilde{\beta} \equiv \lambda\varphi(\sqrt{n}\bar{X})(\lambda\varphi(\sqrt{n}\bar{X}) + (1-\lambda)m_n(\sqrt{n}\bar{X}))^{-1}$$

where $m_n(\sqrt{n}\bar{X}) = (1 + n\tau^2)^{-1/2}\varphi\left(\frac{\sqrt{n}\bar{X}}{(1+n\tau^2)^{1/2}}\right)$.

*Hint:* Set $T = \bar{X}$. We want $\pi(\{0\}|t) = \lambda p_T(t|0)/p_T(t)$ where $p_T(t) = \lambda p_T(t|0) + (1-\lambda)\int_{-\infty}^{\infty} \varphi_\tau(\mu)p_T(t|\mu)d\mu$.

**(b)** Suppose that $\mu = 0$. By Problem $4.1.5$, the $p$-value $\widehat{\beta} \equiv 2[1 - \Phi(\sqrt{n}|\bar{X}|)]$ has a $\mathcal{U}(0, 1)$ distribution. Show that $\widetilde{\beta} \xrightarrow{P} 1$ as $n \to \infty$.

**(c)** Suppose that $\mu = \delta > 0$. Show that $\widetilde{\beta}/\widehat{\beta} \xrightarrow{P} \infty$. That is, if $H$ is false, the evidence against $H$ as measured by the smallness of the $p$-value is much greater than the evidence measured by the smallness of the posterior probability of the hypothesis (Lindley's "paradox").

**2.** Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, 1)$. Consider the problem of testing $H : \mu \in [0, \Delta]$ versus $K : \mu > \Delta$, where $\Delta$ is a given number.

**(a)** Show that the test that rejects $H$ for large values of $\sqrt{n}(\bar{X} - \Delta)$ has $p$-value $\widehat{p} = \Phi(-\sqrt{n}(\bar{X} - \Delta))$ and that when $\mu = \Delta$, $\widehat{p}$ has a $\mathcal{U}(0,1)$ distribution.

**(b)** Suppose that $\boldsymbol{\mu}$ has a $\mathcal{N}(0,1)$ prior. Show that the posterior probability of $H$ is

$$\widetilde{p} = \Phi\left(\frac{-\sqrt{n}(a_n\bar{X} - \Delta)}{\sqrt{a_n}}\right) - \Phi\left(\frac{-\sqrt{n}a_n\bar{X}}{\sqrt{a_n}}\right)$$

where $a_n = n/(n+1)$.

**(c)** Show that when $\mu = \Delta$, $-\sqrt{n}(a_n\bar{X} - \Delta)/\sqrt{a_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ and $\widetilde{p} \xrightarrow{\mathcal{L}} \mathcal{U}(0,1)$. (Lindley's "paradox" of Problem 5.1.1 is not in effect.)

**(d)** Compute $p\lim_{n\to\infty} \widetilde{p}/\widehat{p}$ for $\mu \neq \Delta$.

**(e)** Verify the following table giving posterior probabilities of $[0,\Delta]$ when $\sqrt{n}\bar{X} = 1.645$ and $\widehat{p} = 0.05$.

| $n$ | 10 | 20 | 50 | 100 |
|---|---|---|---|---|
| $\Delta = 0.1$ | .029 | .034 | .042 | .046 |
| $\Delta = 1.0$ | .058 | .054 | .052 | .050 |

**3.** Establish (5.5.14).
*Hint:* By (5.5.13) and the SLLN,

$$\log d_n q_n(t) =$$
$$-\frac{t^2}{2}\left\{I(\theta) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 l}{\partial\theta^2}(X_i,\theta^*(t)) - \frac{\partial^2 l}{\partial\theta^2}(X_i,\theta_0)\right) + \log\pi\left(\widehat{\theta} + \frac{t}{\sqrt{n}}\right)\right\}.$$

Apply the argument used for Theorem 5.4.2 and the continuity of $\pi(\theta)$.

**4.** Establish (5.5.17).
*Hint:*

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l}{\partial\theta^2}(X_i,\theta^*(t))\right| \leq \frac{1}{n}\sum_{i=1}^{n}\sup\left\{\left|\frac{\partial^2}{\partial\theta^2}l(X_i,\theta')\right| : |\theta' - \theta| \leq \delta\right\}$$

if $|t| \leq \delta\sqrt{n}$. Apply the SLLN and $\delta \to E_\theta \sup\left\{\left|\frac{\partial^2}{\partial\theta^2}l(X_i,\theta')\right| : |\theta - \theta'| \leq \delta\right\}$ continuous at $\delta = 0$.

**5.** Suppose that in addition to the conditions of Theorem 5.5.2, $\int \theta^2\pi(\theta)d\theta < \infty$. Then $\sqrt{n}(E(\boldsymbol{\theta} \mid \mathbf{X}) - \widehat{\theta}) \to 0$ a.s. $P_\theta$.
*Hint:* In view of Theorem 5.5.2 it is equivalent to show that

$$d_n \int tq_n(t)dt \to 0$$

a.s. $P_\theta$. By Theorem 5.5.2, $\int_{-M}^{M} tq_n(t)dt \to 0$ a.s. for all $M < \infty$. By (5.5.17), $\int_{M}^{\delta(\theta)\sqrt{n}} |t|q_n(t)dt \le \int_{M}^{\delta(\theta)\sqrt{n}} |t| \exp\left\{-\frac{1}{4}I(\theta)\frac{t^2}{2}\right\} dt \le \epsilon$ for $M(\epsilon)$ sufficiently large, all $\epsilon > 0$. Finally,

$$d_n \int_{\delta(\theta)\sqrt{n}}^{\infty} tq_n(t)dt = \int_{\widehat{\theta}+\delta(\theta)}^{\infty} \sqrt{n}(t - \widehat{\theta}) \exp\left\{\sum_{i=1}^{n}(l(X_i, t) - l(X_i, \widehat{\theta}))\right\} \pi(t)dt.$$

Apply (5.5.16) noting that $\sqrt{n}e^{-n\epsilon(\delta,\theta)} \to 0$ and $\int |t|\pi(t)dt < \infty$.

**6. (a)** Show that $\sup\{|q_n(t) - I^{\frac{1}{2}}(\theta)\varphi(tI^{\frac{1}{2}}(\theta))| : |t| \le M\} \to 0$ a.s. for all $\theta$.

**(b)** Deduce (5.5.29).

*Hint:* $\{t : \sqrt{I(\theta)}\varphi(t\sqrt{I(\theta)}) \ge c(d)\} = [-d, d]$ for some $c(d)$, all $d$ and $c(d) \nearrow$ in $d$. The sets $C_n(c) \equiv \{t : q_n(t) \ge c\}$ are monotone increasing in $c$. Finally, to obtain

$$P[\boldsymbol{\theta} \in C_n(X_1, \dots, X_n) \mid X_1, \dots, X_n] = 1 - \alpha$$

we must have $c_n = c(z_{1-\frac{\alpha}{2}}[I(\widehat{\theta})n]^{-1/2})(1 + o_p(1))$ by Theorem 5.5.1.

**7.** Suppose that in Theorem 5.5.2 we replace the assumptions A4(a.s.) and A5(a.s.) by A4 and A5. Show that (5.5.8) and (5.5.9) hold with a.s. convergence replaced by convergence in $P_\theta$ probability.

## 5.7 NOTES

### Notes for Section 5.1

(1) The bound is actually known to be essentially attained for $X_i = 0$ with probability $p_n$ and 1 with probability $1 - p_n$ where $p_n \to 0$ or 1. For $n$ large these do not correspond to distributions one typically faces. See Bhattacharya and Ranga Rao (1976) for further discussion.

### Notes for Section 5.3

(1) If the right-hand side is negative for some $x$, $F_n(x)$ is taken to be 0.

(2) Computed by Winston Chow.

### Notes for Section 5.4

(1) This result was first stated by R. A. Fisher (1925). A proof was given by Cramér (1946).

### Notes for Section 5.5

(1) This famous result appears in Laplace's work and was rediscovered by S. Bernstein and R. von Mises—see Stigler (1986) and Le Cam and Yang (1990).

## 5.8 REFERENCES

BERGER, J., *Statistical Decision Theory and Bayesian Analysis* New York: Springer–Verlag, 1985.

BHATTACHARYA, R. H. AND R. RANGA RAO, *Normal Approximation and Asymptotic Expansions* New York: Wiley, 1976.

BILLINGSLEY, P., *Probability and Measure* New York: Wiley, 1979.

BOX, G. E. P., "Non-normality and tests on variances," *Biometrika, 40*, 318–324 (1953).

CRAMÉR, H., *Mathematical Methods of Statistics* Princeton, NJ: Princeton University Press, 1946.

DAVID, F. N., *Tables of the Correlation Coefficient*, Cambridge University Press, reprinted in *Biometrika Tables for Statisticians* (1966), Vol. I, 3rd ed., H. O. Hartley and E. S. Pearson, Editors Cambridge: Cambridge University Press, 1938.

FERGUSON, T. S., *A Course in Large Sample Theory* New York: Chapman and Hall, 1996.

FISHER, R. A., "Theory of statistical estimation," *Proc. Camb. Phil. Soc., 22*, 700–725 (1925).

FISHER, R. A., *Statistical Inference and Scientific Method*, Vth Berkeley Symposium, 1958.

HAMMERSLEY, J. M. AND D. C. HANSCOMB, *Monte Carlo Methods* London: Methuen & Co., 1964.

HOEFFDING, W., "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc., 58*, 13–80 (1963).

HUBER, P. J., *The Behavior of the Maximum Likelihood Estimator Under Non-Standard Conditions*, Proc. Vth Berk. Symp. Math. Statist. Prob., Vol. 1 Berkeley, CA: University of California Press, 1967.

LE CAM, L. AND G. L. YANG, *Asymptotics in Statistics, Some Basic Concepts* New York: Springer, 1990.

LEHMANN, E. L., *Elements of Large-Sample Theory* New York: Springer–Verlag, 1999.

LEHMANN, E. L. AND G. CASELLA, *Theory of Point Estimation* New York: Springer–Verlag, 1998.

NEYMAN, J. AND E. L. SCOTT, "Consistent estimates based on partially consistent observations," *Econometrica, 16*, 1–32 (1948).

RAO, C. R., *Linear Statistical Inference and Its Applications*, 2nd ed. New York: J. Wiley & Sons, 1973.

RUDIN, W., *Mathematical Analysis*, 3rd ed. New York: McGraw Hill, 1987.

SCHERVISCH, M., *Theory of Statistics* New York: Springer, 1995.

SERFLING, R. J., *Approximation Theorems of Mathematical Statistics* New York: J. Wiley & Sons, 1980.

STIGLER, S., *The History of Statistics: The Measurement of Uncertainty Before 1900* Cambridge, MA: Harvard University Press, 1986.

WILSON, E. B. AND M. M. HILFERTY, "The distribution of chi square," *Proc. Nat. Acad. Sci., U.S.A., 17*, p. 684 (1931).

This page intentionally left blank