16

Likelihood Ratio Tests

The critical values of the likelihood ratio test are usually based on an asymptotic approximation. We derive the asymptotic distribution of the likelihood ratio statistic and investigate its asymptotic quality through its asymptotic power function and its Bahadur efficiency.

16.1 Introduction

Suppose that we observe a sample $X_1, \ldots, X_n$ from a density p_θ, and wish to test the null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative $H_1 : \theta \in \Theta_1$. If both the null and the alternative hypotheses consist of single points, then a most powerful test can be based on the log likelihood ratio, by the Neyman-Pearson theory. If the two points are θ_0 and θ_1, respectively, then the optimal test statistic is given by

$$\log \frac{\prod_{i=1}^n p_{\theta_1}(X_i)}{\prod_{i=1}^n p_{\theta_0}(X_i)}.$$

For certain special models and hypotheses, the most powerful test turns out not to depend on θ_1, and the test is uniformly most powerful for a composite hypothesis Θ_1. Sometimes the null hypothesis can be extended as well, and the testing problem has a fully satisfactory solution. Unfortunately, in many situations there is no single best test, not even in an asymptotic sense (see Chapter 15). A variety of ideas lead to reasonable tests. A sensible extension of the idea behind the Neyman-Pearson theory is to base a test on the log likelihood ratio

$$\tilde{\Lambda}_n = \log \frac{\sup_{\theta \in \Theta_1} \prod_{i=1}^n p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n p_\theta(X_i)}.$$

The single points are replaced by maxima over the hypotheses. As before, the null hypothesis is rejected for large values of the statistic.

Because the distributional properties of $\tilde{\Lambda}_n$ can be somewhat complicated, one usually replaces the supremum in the numerator by a supremum over the whole parameter set $\Theta = \Theta_0 \cup \Theta_1$. This changes the test statistic only if $\tilde{\Lambda}_n \leq 0$, which is inessential, because in most cases the critical value will be positive. We study the asymptotic properties of the (log) *likelihood ratio statistic*

$$\Lambda_n = 2\log\frac{\sup_{\theta\in\Theta}\prod_{i=1}^n p_\theta(X_i)}{\sup_{\theta\in\Theta_0}\prod_{i=1}^n p_\theta(X_i)} = 2(\tilde\Lambda_n \vee 0).$$

The most important conclusion of this chapter is that, under the null hypothesis, the sequence Λ_n is asymptotically chi squared–distributed. The main conditions are that the model is differentiable in θ and that the null hypothesis Θ_0 and the full parameter set Θ are (locally) equal to linear spaces. The number of degrees of freedom is equal to the difference of the (local) dimensions of Θ and Θ_0. Then the test that rejects the null hypothesis if Λ_n exceeds the upper α-quantile of the chi-square distribution is asymptotically of level α. Throughout the chapter we assume that $\Theta \subset \mathbb{R}^k$.

The "local linearity" of the hypotheses is essential for the chi-square approximation, which fails already in a number of simple examples. An open set is certainly locally linear at every of its points, and so is a relatively open subset of an affine subspace. On the other hand, a half line or space, which arises, for instance, if testing a one-sided hypothesis $H_0: \mu_\theta \leq 0$, or a ball $H_0: \|\theta\| \leq 1$, is not locally linear at its boundary points. In that case the asymptotic null distribution of the likelihood ratio statistic is not chi-square, but the distribution of a certain functional of a Gaussian vector.

Besides for testing, the likelihood ratio statistic is often used for constructing confidence regions for a parameter $\psi(0)$. These are defined, as usual, as the values τ for which a null hypothesis $H_0: \psi(\theta) = \tau$ is not rejected. Asymptotic confidence sets obtained by using the chi-square approximation are thought to be of better coverage accuracy than those obtained by other asymptotic methods.

The likelihood ratio test has the desirable property of automatically achieving reduction of the data by sufficiency: The test statistic depends on a minimal sufficient statistic only. This is immediate from its definition as a quotient and the characterization of sufficiency by the factorization theorem. Another property of the test is also immediate: The likelihood ratio statistic is invariant under transformations of the parameter space that leave the null and alternative hypotheses invariant. This requirement is often imposed on test statistics but is not necessarily desirable.

16.1 ***Example (Multinomial vector).*** A vector $N = (N_1, \ldots, N_k)$ that possesses the multinomial distribution with parameters n and $p = (p_1, \ldots, p_k)$ can be viewed as the sum of n independent multinomial vectors with parameters 1 and p. By the sufficiency reduction, the likelihood ratio statistic based on N is the same as the statistic based on the single observations. Thus our asymptotic results apply to the likelihood ratio statistic based on N, if $n \to \infty$.

If the success probabilities are completely unknown, then their maximum likelihood estimator is N/n. Thus, the log likelihood ratio statistic for testing a null hypothesis $H_0: p \in \mathcal{P}_0$ against the alternative $H_1: p \notin \mathcal{P}_0$ is given by

$$2\log\frac{\binom{n}{N_1\cdots N_k}(N_1/n)^{N_1}\cdots(N_k/n)^{N_k}}{\sup_{p\in\mathcal{P}_0}\binom{n}{N_1\cdots N_k}p_1^{N_1}\cdots p_k^{N_k}} = 2\inf_{p\in\mathcal{P}_0}\sum_{i=1}^k N_i\log\left(\frac{N_i}{np_i}\right).$$

The full parameter set can be identified with an open subset of $\mathbb{R}^{k-1}$, if p with zero coordinates are excluded. The null hypothesis may take many forms. For a simple null hypothesis the statistic is asymptotically chi-square distributed with $k-1$ degrees of freedom. This follows from the general results in this chapter.[†]

Multinomial variables arise, among others, in testing goodness-of-fit. Suppose we wish to test that the true distribution of a sample of size n belongs to a certain parametric model $\{P_\theta : \theta \in \Theta\}$. Given a partition of the sample space into sets $\mathcal{X}_1, \dots, \mathcal{X}_k$, define $N_1, \dots, N_k$ as the numbers of observations falling into each of the sets of the partition. Then the vector $N = (N_1, \dots, N_k)$ possesses a multinomial distribution, and the original problem can be translated in testing the null hypothesis that the success probabilities p have the form $\big(P_\theta(\mathcal{X}_1), \dots, P_\theta(\mathcal{X}_k)\big)$ for some θ. □

16.2 ***Example (Exponential families).*** Suppose that the observations are sampled from a density p_θ in the k-dimensional exponential family

$$p_\theta(x) = c(\theta)h(x)e^{\theta^T t(x)}.$$

Let $\Theta \subset \mathbb{R}^k$ be the natural parameter space, and consider testing a null hypothesis $\Theta_0 \subset \mathring{\Theta}$ versus its complement $\Theta - \Theta_0$. The log likelihood ratio statistic is given by

$$\Lambda_n = 2n \sup_{\theta\in\Theta} \inf_{\theta\in\Theta_0} \big[(\theta - \theta_0)^T \bar{t}_n + \log c(\theta) - \log c(\theta_0)\big].$$

This is closely related to the Kullback-Leibler divergence of the measures P_{θ_0} and P_θ, which is equal to

$$K(\theta, \theta_0) = P_\theta \log \frac{p_\theta}{p_{\theta_0}} = (\theta - \theta_0)^T P_\theta t + \log c(\theta) - \log c(\theta_0).$$

If the maximum likelihood estimator $\hat{\theta}$ exists and is contained in the interior of Θ, which is the case with probability tending to 1 if the true parameter is contained in $\mathring{\Theta}$, then $\hat{\theta}$ is the moment estimator that solves the equation $P_\theta t = \bar{t}_n$. Comparing the two preceding displays, we see that the likelihood ratio statistic can be written as $\Lambda_n = 2nK(\hat{\theta}, \Theta_0)$, where $K(\theta, \Theta_0)$ is the infimum of $K(\theta, \theta_0)$ over $\theta_0 \in \Theta_0$. This pretty formula can be used to study the asymptotic properties of the likelihood ratio statistic directly. Alternatively, the general results obtained in this chapter are applicable to exponential families. □

*16.2 Taylor Expansion

Write $\hat{\theta}_{n,0}$ and $\hat{\theta}_n$ for the maximum likelihood estimators for θ if the parameter set is taken equal to Θ_0 or Θ, respectively, and set $\ell_\theta = \log p_\theta$. In this section assume that the true value of the parameter ϑ is an inner point of Θ. The likelihood ratio statistic can be rewritten as

$$\Lambda_n = -2\sum_{i=1}^n \big(\ell_{\hat{\theta}_{n,0}}(X_i) - \ell_{\hat{\theta}_n}(X_i)\big).$$

To find the limit behavior of this sequence of random variables, we might replace $\sum \ell_\theta(X_i)$ by its Taylor expansion around the maximum likelihood estimator $\theta = \hat{\theta}_n$. If $\theta \mapsto \ell_\theta(x)$

[†] It is also proved in Chapter 17 by relating the likelihood ratio statistic to the chi-square statistic.

is twice continuously differentiable for every x, then there exists a vector $\tilde{\theta}_n$ between $\hat{\theta}_{n,0}$ and $\hat{\theta}_n$ such that the preceding display is equal to

$$-2(\hat{\theta}_{n,0}-\hat{\theta}_n)\sum_{i=1}^{n}\dot{\ell}_{\hat{\theta}_n}(X_i)-(\hat{\theta}_{n,0}-\hat{\theta}_n)^T\sum\ddot{\ell}_{\tilde{\theta}_n}(X_i)(\hat{\theta}_{n,0}-\hat{\theta}_n).$$

Because $\hat{\theta}_n$ is the maximum likelihood estimator in the unrestrained model, the linear term in this expansion vanishes as soon as $\hat{\theta}_n$ is an inner point of Θ. If the averages $-n^{-1}\sum\ddot{\ell}_{\tilde{\theta}}(X_i)$ converge in probability to the Fisher information matrix I_ϑ and the sequence $\sqrt{n}(\hat{\theta}_{n,0}-\hat{\theta}_n)$ is bounded in probability, then we obtain the approximation

$$\Lambda_n=\sqrt{n}(\hat{\theta}_n-\hat{\theta}_{n,0})^T I_\vartheta\sqrt{n}(\hat{\theta}_n-\hat{\theta}_{n,0})+o_{P_\vartheta}(1). \tag{16.3}$$

In view of the results of Chapter 5, the latter conditions are reasonable if $\vartheta\in\Theta_0$, for then both $\hat{\theta}_n$ and $\hat{\theta}_{n,0}$ can be expected to be $\sqrt{n}$-consistent. The preceding approximation, if it can be justified, sheds some light on the quality of the likelihood ratio test. It shows that, asymptotically, the likelihood ratio test measures a certain distance between the maximum likelihood estimators under the null and the full hypotheses. Such a procedure is intuitively reasonable, even though many other distance measures could be used as well. The use of the likelihood ratio statistic entails a choice as to how to weigh the different "directions" in which the estimators may differ, and thus a choice of weights for "distributing power" over different deviations. This is further studied in section 16.4.

If the null hypothesis is a single point $\Theta_0=\{\theta_0\}$, then $\hat{\theta}_{n,0}=\theta_0$, and the quadratic form in the preceding display reduces under $H_0:\theta=\theta_0$ (i.e., $\vartheta=\theta_0$) to $\hat{h}_n I_\vartheta\hat{h}_n$ for $\hat{h}_n=\sqrt{n}(\hat{\theta}_n-\vartheta)^T$. In view of the results of Chapter 5, the sequence $\hat{h}_n$ can be expected to converge in distribution to a variable $\hat{h}$ with a normal $N(0,I_\vartheta^{-1})$-distribution. Then the sequence Λ_n converges under the null hypothesis in distribution to the quadratic form $\hat{h}^T I_\vartheta\hat{h}$. This is the squared length of the standard normal vector $I_\vartheta^{1/2}\hat{h}$, and possesses a chi-square distribution with k degrees of freedom. Thus the chi-square approximation announced in the introduction follows.

The situation is more complicated if the null hypothesis is composite. If the sequence $\sqrt{n}(\hat{\theta}_{n,0}-\vartheta,\hat{\theta}_n-\vartheta)$ converges jointly to a variable $(\hat{h}_0,\hat{h})$, then the sequence Λ_n is asymptotically distributed as $(\hat{h}-\hat{h}_0)^T I_\vartheta(\hat{h}-\hat{h}_0)$. A null hypothesis Θ_0 that is (a segment of) a lower dimensional affine linear subspace is itself a "regular" parametric model. If it contains ϑ as a relative inner point, then the maximum likelihood estimator $\hat{\theta}_{n,0}$ may be expected to be asymptotically normal within this affine subspace, and the pair $\sqrt{n}(\hat{\theta}_{n,0}-\vartheta,\hat{\theta}_n-\vartheta)$ may be expected to be jointly asymptotically normal. Then the likelihood ratio statistic is asymptotically distributed as a quadratic form in normal variables. Closer inspection shows that this quadratic form possesses a chi-square distribution with $k-l$ degrees of freedom, where k and l are the dimensions of the full and null hypotheses. In comparison with the case of a simple null hypothesis, l degrees of freedom are "lost."

Because we shall rigorously derive the limit distribution by a different approach in the next section, we make this argument precise only in the particular case that the null hypothesis Θ_0 consists of all points $(\theta_1,\dots,\theta_l,0,\dots,0)$, if θ ranges over an open subset Θ of $\mathbb{R}^k$. Then the score function for θ under the null hypothesis consists of the first l coordinates of the score function $\dot{\ell}_\vartheta$ for the whole model, and the information matrix under the null hypothesis is equal to the $(l\times l)$ principal submatrix of I_ϑ. Write these as $\dot{\ell}_{\vartheta,\le l}$ and $I_{\vartheta,\le l,\le l}$, respectively, and use a similar partitioning notation for other vectors and matrices.

Under regularity conditions we have the linear approximations (see Theorem 5.39)

$$\sqrt{n}(\hat{\theta}_{n,0,\leq l} - \vartheta_{\leq l}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{\vartheta,\leq l,\leq l}^{-1}\dot{\ell}_{\vartheta,\leq l}(X_i) + o_{P_\vartheta}(1),$$

$$\sqrt{n}(\hat{\theta}_n - \vartheta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_\vartheta^{-1}\dot{\ell}_\vartheta(X_i) + o_{P_\vartheta}(1).$$

Given these approximations, the multivariate central limit theorem and Slutsky's lemma yield the joint asymptotic normality of the maximum likelihood estimators. From the form of the asymptotic covariance matrix we see, after some matrix manipulation,

$$\sqrt{n}(\hat{\theta}_{n,\leq l} - \hat{\theta}_{n,0,\leq l}) = -I_{\vartheta,\leq l,\leq l}^{-1} I_{\vartheta,\leq l,>l}\sqrt{n}\,\hat{\theta}_{n,>l} + o_P(1).$$

(Alternatively, this approximation follows from a Taylor expansion of $0 = \sum_{i=1}^{n}\dot{\ell}_{\hat{\theta}_n,\leq l}$ around $\hat{\theta}_{n,0,\leq l}$.) Substituting this in (16.3) and again carrying out some matrix manipulations, we find that the likelihood ratio statistic is asymptotically equivalent to (see problem 16.5)

$$\sqrt{n}\,\hat{\theta}_{n,>l}^T\Bigl(\bigl(I_\vartheta^{-1}\bigr)_{>l,>l}\Bigr)^{-1}\sqrt{n}\,\hat{\theta}_{n,>l}. \tag{16.4}$$

The matrix $(I_\vartheta^{-1})_{>l,>l}$ is the asymptotic covariance matrix of the sequence $\sqrt{n}\,\hat{\theta}_{n,>l}$, whence we obtain an asymptotic chi-square distribution with $k - l$ degrees of freedom, by the same argument as before.

We close this section by relating the likelihood ratio statistic to two other test statistics. Under the simple null hypothesis $\Theta_0 = \{\theta_0\}$, the likelihood ratio statistic is asymptotically equivalent to both the *maximum likelihood statistic* (or *Wald statistic*) and the *score statistic*. These are given by

$$n(\hat{\theta}_n - \theta_0)^T I_{\theta_0}(\hat{\theta}_n - \theta_0) \quad \text{and} \quad \frac{1}{n}\Biggl[\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i)\Biggr]^T I_{\theta_0}^{-1}\Biggl[\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i)\Biggr].$$

The Wald statistic is a natural statistic, but it is often criticized for necessarily yielding ellipsoidal confidence sets, even if the data are not symmetric. The score statistic has the advantage that calculation of the supremum of the likelihood is unnecessary, but it appears to perform less well for smaller values of n.

In the case of a composite hypothesis, a Wald statistic is given in (16.4) and a score statistic can be obtained by substituting the approximation $n\hat{\theta}_{n,>l} \approx (I_\vartheta^{-1})_{>l,>l}\sum\dot{\ell}_{\hat{\theta}_{n,0,>l}}(X_i)$ in (16.4). (This approximation is obtainable from linearizing $\sum(\dot{\ell}_{\hat{\theta}_n} - \dot{\ell}_{\hat{\theta}_{n,o}})$.) In both cases we also replace the unknown parameter ϑ by an estimator.

16.3 Using Local Asymptotic Normality

An insightful derivation of the asymptotic distribution of the likelihood ratio statistic is based on convergence of experiments. This approach is possible for general experiments, but this section is restricted to the case of local asymptotic normality. The approach applies also in the case that the (local) parameter spaces are not linear.

Introducing the local parameter spaces $H_n = \sqrt{n}(\Theta - \vartheta)$ and $H_{n,0} = \sqrt{n}(\Theta_0 - \vartheta)$, we can write the likelihood ratio statistic in the form

$$\Lambda_n = 2\sup_{h\in H_n}\log\frac{\prod_{i=1}^{n}p_{\vartheta+h/\sqrt{n}}(X_i)}{\prod_{i=1}^{n}p_\vartheta(X_i)} - 2\sup_{h\in H_{n,0}}\log\frac{\prod_{i=1}^{n}p_{\vartheta+h/\sqrt{n}}(X_i)}{\prod_{i=1}^{n}p_\vartheta(X_i)}.$$

In Chapter 7 it is seen that, for large n, the rescaled likelihood ratio process in this display is similar to the likelihood ratio process of the normal experiment $\bigl(N(h, I_\vartheta^{-1}): h \in \mathbb{R}^k\bigr)$. This suggests that, if the sets H_n and $H_{n,0}$ converge in a suitable sense to sets H and H_0, the sequence Λ_n converges in distribution to the random variable Λ obtained by substituting the normal likelihood ratios, given by

$$\Lambda = 2 \sup_{h \in H} \log \frac{dN\bigl(h, I_\vartheta^{-1}\bigr)}{dN\bigl(0, I_\vartheta^{-1}\bigr)}(X) - 2 \sup_{h \in H_0} \log \frac{dN\bigl(h, I_\vartheta^{-1}\bigr)}{dN\bigl(0, I_\vartheta^{-1}\bigr)}(X).$$

This is exactly the likelihood ratio statistic for testing the null hypothesis $H_0 : h \in H_0$ versus the alternative $H_1 : h \in H - H_0$ based on the observation X in the normal experiment. Because the latter experiment is simple, this heuristic is useful not only to derive the asymptotic distribution of the sequence Λ_n, but also to understand the asymptotic quality of the corresponding sequence of tests.

The likelihood ratio statistic for the normal experiment is

$$\begin{aligned}\Lambda &= \inf_{h \in H_0} (X - h)^T I_\vartheta (X - h) - \inf_{h \in H} (X - h)^T I_\vartheta (X - h)\\ &= \bigl\| I_\vartheta^{1/2} X - I_\vartheta^{1/2} H_0 \bigr\|^2 - \bigl\| I_\vartheta^{1/2} X - I_\vartheta^{1/2} H \bigr\|^2. \qquad (16.5)\end{aligned}$$

The distribution of the sequence Λ_n under ϑ corresponds to the distribution of Λ under $h = 0$. Under $h = 0$ the vector $I_\vartheta^{1/2} X$ possesses a standard normal distribution. The following lemma shows that the squared distance of a standard normal variable to a linear subspace is chi square–distributed and hence explains the chi-square limit when H_0 is a linear space.

16.6 ***Lemma.*** *Let X be a k-dimensional random vector with a standard normal distribution and let H_0 be an l-dimensional linear subspace of $\mathbb{R}^k$. Then $\|X - H_0\|^2$ is chi square–distributed with $k - l$ degrees of freedom.*

Proof. Take an orthonormal base of $\mathbb{R}^k$ such that the first l elements span H_0. By Pythagoras' theorem, the squared distance of a vector z to the space H_0 equals the sum of squares $\sum_{i>l} z_i^2$ of its last $k - l$ coordinates with respect to this basis. A change of base corresponds to an orthogonal transformation of the coordinates. Because the standard normal distribution is invariant under orthogonal transformations, the coordinates of X with respect to any orthonormal base are independent standard normal variables. Thus $\|X - H_0\|^2 = \sum_{i>l} X_i^2$ is chi square–distributed. ■

If ϑ is an inner point of Θ, then the set H is the full space $\mathbb{R}^k$ and the second term on the right of (16.5) is zero. Thus, if the local null parameter spaces $\sqrt{n}(\Theta_0 - \vartheta)$ converge to a linear subspace of dimension l, then the asymptotic null distribution of the likelihood ratio statistic is chi-square with $k - l$ degrees of freedom.

The following theorem makes the preceding informal derivation rigorous under the same mild conditions employed to obtain the asymptotic normality of the maximum likelihood estimator in Chapter 5. It uses the following notion of *convergence of sets*. Write $H_n \to H$ if H is the set of all limits $\lim h_n$ of converging sequences h_n with $h_n \in H_n$ for every n and, moreover, the limit $h = \lim_i h_{n_i}$ of every converging sequence h_{n_i} with $h_{n_i} \in H_{n_i}$ for every i is contained in H.

16.7 ***Theorem.*** *Let the model $(P_\theta : \theta \in \Theta)$ be differentiable in quadratic mean at ϑ with nonsingular Fisher information matrix, and suppose that for every θ_1 and θ_2 in a neighborhood of ϑ and for a measurable function $\dot\ell$ such that $P_\vartheta \dot\ell^2 < \infty$,*

$$\left|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\right| \le \dot\ell(x)\, \|\theta_1 - \theta_2\|.$$

If the maximum likelihood estimators $\hat\theta_{n,0}$ and $\hat\theta_n$ are consistent under ϑ and the sets $H_{n,0}$ and H_n converge to sets H_0 and H, then the sequence of likelihood ratio statistics Λ_n converges under $\vartheta + h/\sqrt{n}$ in distribution to Λ given in (16.5), for X normally distributed with mean h and covariance matrix I_ϑ^{-1}.

Proof. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_\vartheta)$ be the empirical process, and define stochastic processes $\mathbb{Z}_n$ by

$$\mathbb{Z}_n(h) = n\mathbb{P}_n \log \frac{p_{\vartheta+h/\sqrt{n}}}{p_\vartheta} - h^T \mathbb{G}_n \dot\ell_\vartheta + \frac{1}{2} h^T I_\vartheta h.$$

The differentiability of the model implies that $\mathbb{Z}_n(h) \xrightarrow{\mathrm{P}} 0$ for every h. In the proof of Theorem 7.12 this is strengthened to the uniform convergence

$$\sup_{\|h\| \le M} \left|\mathbb{Z}_n(h)\right| \xrightarrow{\mathrm{P}} 0, \qquad \text{every } M.$$

Furthermore, it follows from this proof that both $\hat\theta_{n,0}$ and $\hat\theta_n$ are $\sqrt{n}$-consistent under ϑ. (These statements can also be proved by elementary arguments, but under stronger regularity conditions.)

The preceding display is also valid for every sequence M_n that increases to ∞ sufficiently slowly. Fix such a sequence. By the $\sqrt{n}$-consistency, the estimators $\hat\theta_{n,0}$ and $\hat\theta_n$ are contained in the ball of radius $M_n/\sqrt{n}$ around ϑ with probability tending to 1. Thus, the limit distribution of Λ_n does not change if we replace the sets H_n and $H_{n,0}$ in its definition by the sets $H_n \cap \mathrm{ball}(0, M_n)$ and $H_{n,0} \cap \mathrm{ball}(0, M_n)$. These "truncated" sequences of sets still converge to H and H_0, respectively. Now, by the uniform convergence to zero of the processes $\mathbb{Z}_n(h)$ on H_n and $H_{n,0}$, and simple algebra,

$$\begin{aligned}
\Lambda_n &= 2 \sup_{h \in H_n} n\mathbb{P}_n \log \frac{p_{\vartheta+h/\sqrt{n}}}{p_\vartheta} - 2 \sup_{h \in H_{n,0}} n\mathbb{P}_n \log \frac{p_{\vartheta+h/\sqrt{n}}}{p_\vartheta} \\
&= 2 \sup_{h \in H_n} \left(h^T \mathbb{G}_n \dot\ell_\vartheta - \tfrac{1}{2} h^T I_\vartheta h\right) - 2 \sup_{h \in H_{n,0}} \left(h^T \mathbb{G}_n \dot\ell_\vartheta - \tfrac{1}{2} h^T I_\vartheta h\right) + o_P(1) \\
&= \left\| I_\vartheta^{-1/2} \mathbb{G}_n \dot\ell_\vartheta - I_\vartheta^{1/2} H_0 \right\|^2 - \left\| I_\vartheta^{-1/2} \mathbb{G}_n \dot\ell_\vartheta - I_\vartheta^{1/2} H \right\|^2 + o_P(1)
\end{aligned}$$

by Lemma 7.13 (ii) and (iii). The theorem follows by the continuous-mapping theorem. ■

16.8 ***Example (Generalized linear models).*** In a generalized linear model a typical observation (X, Y), consisting of a "covariate vector" X and a "response" Y, possesses a density of the form

$$p_\beta(x, y) = e^{yk(\beta^T x)\phi - b \circ k(\beta^T x)\phi} c_\phi(y) p_X(x).$$

(It may be more natural to model the covariates as (observed) constants, but to fit the model into our i.i.d. setup, we consider them to be a random sample from a density p_X.) Thus, given

X, the variable Y follows an exponential family density $e^{y\theta\phi - b(\theta)\phi}c_\phi(y)$ with parameters $\theta = k(\beta^T X)$ and ϕ. Using the identities for exponential families based on Lemma 4.5, we obtain

$$\mathrm{E}_\beta(Y \mid X) = b' \circ k(\beta^T X), \qquad \mathrm{var}_{\beta,\phi}(Y \mid X) = \frac{b'' \circ k(\beta^T X)}{\phi}.$$

The function $(b' \circ k)^{-1}$ is called the *link function* of the model and is assumed known. To make the parameter β identifiable, we assume that the matrix $\mathrm{E}XX^T$ exists and is nonsingular.

To judge the goodness-of-fit of a generalized linear model to a given set of data $(X_1, Y_1), \ldots, (X_n, Y_n)$, it is customary to calculate, for fixed ϕ, the log likelihood ratio statistic for testing the model as described previously within the model in which each Y_i, given X_i, still follows the given exponential family density, but in which the parameters θ (and hence the conditional means $\mathrm{E}(Y_i \mid X_i)$) are allowed to be arbitrary values θ_i, unrelated across the n observations (X_i, Y_i). This statistic, with the parameter ϕ set to 1, is known as the *deviance*, and takes the form, with $\hat\beta_n$ the maximum likelihood estimator for β,[†]

$$\begin{aligned} D(\vec{Y}_n, \hat\mu) &= -2\log \frac{\sup_\beta \prod_{i=1}^n e^{Y_i k(\beta^T X_i) - b \circ k(\beta^T X_i)}}{\sup_{\theta_1,\ldots,\theta_n} \prod_{i=1}^n e^{Y_i\theta_i - b(\theta_i)}} \\ &= -2\sum_{i=1}^n \Big[Y_i\Big(k\big(\hat\beta_n^T X_i\big) - (b')^{-1}(Y_i)\Big) - b \circ k\big(\hat\beta_n^T X_i\big) + b \circ (b')^{-1}(Y_i)\Big]. \end{aligned}$$

In our present setup, the codimension of the null hypothesis within the "full model" is equal to $n - k$, if β is k-dimensional, and hence the preceding theory does not apply to the deviance. (This could be different if there are multiple responses for every given covariate and the asymptotics are relative to the number of responses.) On the other hand, the preceding theory allows an "analysis of deviance" to test nested sequences of regression models corresponding to inclusion or exclusion of a given covariate (i.e., column of the regression matrix). For instance, if $D_i(\vec{Y}_n, \hat\mu_{(i)})$ is the deviance of the model in which the $i+1, i+2, \ldots, k$th coordinates of β are a priori set to zero, then the difference $D_{i-1}(\vec{Y}_n, \hat\mu_{(i-1)}) - D_i(\vec{Y}_n, \hat\mu_{(i)})$ is the log likelihood ratio statistic for testing that the ith coordinate of β is zero within the model in which all higher coordinates are zero. According to the theory of this chapter, ϕ times this statistic is asymptotically chi square–distributed with one degree of freedom under the smaller of the two models.

To see this formally, it suffices to verify the conditions of the preceding theorem. Using the identities for exponential families based on Lemma 4.5, the score function and Fisher information matrix can be computed to be

$$\begin{aligned} \dot\ell_\beta(x, y) &= \big(y - b' \circ k(\beta^T x)\big)k'(\beta^T x)x, \\ I_\beta &= \mathrm{E}b'' \circ k(\beta^T X)k'(\beta^T X)^2 XX^T. \end{aligned}$$

Depending on the function k, these are very well-behaved functions of β, because b is a strictly convex, analytic function on the interior of the natural parameter space of the family, as is seen in section 4.2. Under reasonable conditions the function $\sup_{\beta\in U} \|\dot\ell_\beta\|$ is

[†] The arguments $\vec{Y}_n$ and $\hat\mu$ of D are the vectors of estimated (conditional) means of Y given the full model and the generalized linear model, respectively. Thus $\hat\mu_i = b' \circ k(\hat\beta_n^T X_i)$.

square-integrable, for every small neighborhood U, and the Fisher information is continuous. Thus, the local conditions on the model are easily satisfied.

Proving the consistency of the maximum likelihood estimator may be more involved, depending on the link function. If the parameter β is restricted to a compact set, then most approaches to proving consistency apply without further work, including Wald's method, Theorem 5.7, and the classical approach of section 5.7. The last is particularly attractive in the case of *canonical link functions*, which correspond to setting k equal to the identity. Then the second-derivative matrix $\ddot{\ell}_\beta$ is equal to $-b''(\beta^T x)xx^T$, whence the likelihood is a strictly concave function of β whenever the observed covariate vectors are of full rank. Consequently, the point of maximum of the likelihood function is unique and hence consistent under the conditions of Theorem 5.14.[†] □

16.9 ***Example (Location scale).*** Suppose we observe a sample from the density $f\big((x-\mu)/\sigma\big)/\sigma$ for a given probability density f, and a location-scale parameter $\theta = (\mu, \sigma)$ ranging over the set $\Theta = \mathbb{R} \times \mathbb{R}^+$. We consider two testing problems.

(i). Testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ corresponds to setting $\Theta_0 = \{0\} \times \mathbb{R}^+$. For a given point $\vartheta = (0, \sigma)$ from the null hypothesis the set $\sqrt{n}(\Theta_0 - \vartheta)$ equals $\{0\} \times (-\sqrt{n}\sigma, \infty)$ and converges to the linear space $\{0\} \times \mathbb{R}$. Under regularity conditions on f, the sequence of likelihood ratio statistics is asymptotically chi square–distributed with 1 degree of freedom.

(ii). Testing $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ corresponds to setting $\Theta_0 = (-\infty, 0] \times \mathbb{R}^+$. For a given point $\vartheta = (0, \sigma)$ on the boundary of the null hypothesis, the sets $\sqrt{n}(\Theta_0 - \vartheta)$ converge to $H_0 = (-\infty, 0] \times \mathbb{R}$. In this case, the limit distribution of the likelihood ratio statistics is not chi-square but equals the distribution of the square distance of a standard normal vector to the set $I_\vartheta^{1/2} H_0 = \big\{h : \langle h, I_\vartheta^{-1/2} e_1 \rangle \leq 0\big\}$. The latter is a half-space with boundary line through the origin. Because a standard normal vector is rotationally symmetric, the distribution of its distance to a half-space of this type does not depend on the orientation of the half-space. Thus the limit distribution is equal to the distribution of the squared distance of a standard normal vector to the half-space $\{h : h_2 \leq 0\}$: the distribution of $(Z \vee 0)^2$ for a standard normal variable Z. Because $\mathrm{P}\big((Z \vee 0)^2 > c\big) = \frac{1}{2}\mathrm{P}(Z^2 > c)$ for every $c > 0$, we must choose the critical value of the test equal to the upper 2α-quantile of the chi-square distribution with 1 degree of freedom. Then the asymptotic level of the test is α for every ϑ on the boundary of the null hypothesis (provided $\alpha < 1/2$).

For a point ϑ in the interior of the null hypothesis $H_0 : \mu \leq 0$ the sets $\sqrt{n}(\Theta_0 - \vartheta)$ converge to $\mathbb{R} \times \mathbb{R}$ and the sequence of likelihood ratio statistics converges in distribution to the squared distance to the whole space, which is zero. This means that the probability of an error of the first kind converges to zero for every ϑ in the interior of the null hypothesis. □

16.10 ***Example (Testing a ball).*** Suppose we wish to test the null hypothesis $H_0 : \|\theta\| \leq 1$ that the parameter belongs to the unit ball versus the alternative $H_1 : \|\theta\| > 1$ that this is not case.

If the true parameter ϑ belongs to the interior of the null hypothesis, then the sets $\sqrt{n}(\Theta_0 - \vartheta)$ converge to the whole space, whence the sequence of likelihood ratio statistics converges in distribution to zero.

[†] For a detailed study of sufficient conditions for consistency see [45].

For ϑ on the boundary of the unit ball, the sets $\sqrt{n}(\Theta_0 - \vartheta)$ grow to the half-space $H_0 = \{h : \langle h, \vartheta\rangle \le 0\}$. The sequence of likelihood ratio statistics converges in distribution to the distribution of the square distance of a standard normal vector to the half-space $I_\vartheta^{1/2} H_0 = \{h : \langle h, I_\vartheta^{-1/2}\vartheta\rangle \le 0\}$. By the same argument as in the preceding example, this is the distribution of $(Z \vee 0)^2$ for a standard normal variable Z. Once again we find an asymptotic level-α test by using a 2α-quantile. □

16.11 ***Example (Testing a range).*** Suppose that the null hypothesis is equal to the image $\Theta_0 = g(T)$ of an open subset T of a Euclidean space of dimension $l \le k$. If g is a homeomorphism, continuously differentiable, and of full rank, then the sets $\sqrt{n}\big(\Theta_0 - g(\tau)\big)$ converge to the range of the derivative of g at τ, which is a subspace of dimension l.

Indeed, for any $\eta \in \mathbb{R}^l$ the vectors $\tau + \eta/\sqrt{n}$ are contained in T for sufficiently large n, and the sequence $\sqrt{n}\big(g(\tau + \eta/\sqrt{n}) - g(\tau)\big)$ converges to $g'_\tau \eta$. Furthermore, if a subsequence of $\sqrt{n}\big(g(t_n) - g(\tau)\big)$ converges to a point h for a given sequence t_n in T, then the corresponding subsequence of $\sqrt{n}(t_n - \tau)$ converges to $\eta = (g^{-1})'_{g(\tau)} h$ by the differentiability of the inverse mapping g^{-1} and hence $\sqrt{n}\big(g(t_n) - g(\tau)\big) \to g'_\tau \eta$. (We can use the rank theorem to give a precise definition of the differentiability of the map g^{-1} on the manifold $g(T)$.) □

16.4 Asymptotic Power Functions

Because the sequence of likelihood ratio statistics converges to the likelihood ratio statistic in the Gaussian limit experiment, the likelihood ratio test is "asymptotically efficient" in the same way as the likelihood ratio statistic in the limit experiment is "efficient." If the local limit parameter set H_0 is a half-space or a hyperplane, then the latter test is uniformly most powerful, and hence the likelihood ratio tests are asymptotically optimal (see Proposition 15.2).This is the case, in particular, for testing a simple null hypothesis in a one-dimensional parametric model. On the other hand, if the hypotheses are higher-dimensional, then there is often no single best test, not even under reasonable restrictions on the class of admitted tests. For different (one-dimensional) deviations of the null hypothesis, different tests are optimal (see the discussion in Chapter 15). The likelihood ratio test is an omnibus test that gives reasonable power in all directions. In this section we study its local asymptotic power function more closely.

We assume that the parameter ϑ is an inner point of the parameter set and denote the true parameter by $\vartheta + h/\sqrt{n}$. Under the conditions of Theorem 16.7, the sequence of likelihood ratio statistics is asymptotically distributed as

$$\Lambda = \big\| Z + I_\vartheta^{1/2} h - I_\vartheta^{1/2} H_0 \big\|^2$$

for a standard normal vector Z. Suppose that the limiting local parameter set H_0 is a linear subspace of dimension l, and that the null hypothesis is rejected for values of Λ_n exceeding the critical value $\chi^2_{k-l,\alpha}$. Then the local power functions of the resulting tests satisfy

$$\pi_n\left(\vartheta + \frac{h}{\sqrt{n}}\right) = \mathrm{P}_{\vartheta + h/\sqrt{n}}\big(\Lambda_n > \chi^2_{k-l,\alpha}\big) \to \mathrm{P}_h\big(\Lambda > \chi^2_{k-l,\alpha}\big) =: \pi(h).$$

The variable Λ is the squared distance of the vector Z to the affine subspace $-I_\vartheta^{1/2} h + I_\vartheta^{1/2} H_0$. By the rotational invariance of the normal distribution, the distribution of Λ does not depend on the orientation of the affine subspace, but only on its codimension and its distance

$\delta = \|I_\vartheta^{1/2}h - I_\vartheta^{1/2}H_0\|$ to the origin. This distribution is known as the *noncentral chi-square distribution* with noncentrality parameter δ. Thus

$$\pi(h) = \mathrm{P}\Big(\chi^2_{k-l}\Big(\big\|I_\vartheta^{1/2}h - I_\vartheta^{1/2}H_0\big\|\Big) > \chi^2_{k-l,\alpha}\Big).$$

The noncentral chi-square distributions are stochastically increasing in the noncentrality parameter. It follows that the likelihood ratio test has good (local) power at h that yield a large value of the noncentrality parameter.

The shape of the asymptotic power function is easiest to understand in the case of a simple null hypothesis. Then $H_0 = \{0\}$, and the noncentrality parameter reduces to the square root of $h^T I_\vartheta h$. For $h = \mu h_e$ equal to a multiple of an eigenvector h_e (of unit norm) of I_ϑ with eigenvalue λ_e, the noncentrality parameter equals $\sqrt{\lambda_e}\mu$. The asymptotic power function in the direction of h_e equals

$$\pi(\mu h_e) = \mathrm{P}\big(\chi^2_k(\sqrt{\lambda_e}\mu) > \chi^2_{k,\alpha}\big).$$

The test performs best for departures from the null hypothesis in the direction of the eigenvector corresponding to the largest eigenvalue. Even though the likelihood ratio test gives power in all directions, it does not treat the directions equally. This may be worrisome if the eigenvalues are very inhomogeneous.

Further insight is gained by comparing the likelihood ratio test to tests that are designed to be optimal in given directions. Let X be an observation in the limit experiment, having a $N(h, I_\vartheta^{-1})$-distribution. The test that rejects the null hypothesis $H_0 = \{0\}$ if $|\sqrt{\lambda_e}\, h_e^T X| > z_{\alpha/2}$ has level α and power function

$$\pi_{h_e}(\mu h_e) = \mathrm{P}\big(\chi^2_1(\sqrt{\lambda_e}\mu) > \chi^2_{1,\alpha}\big).$$

For large k this is a considerably better power function than the power function of the likelihood ratio test (Figure 16.1), but the forms of the power functions are similar. In particular, the optimal power functions show a similar dependence on the eigenvalues of

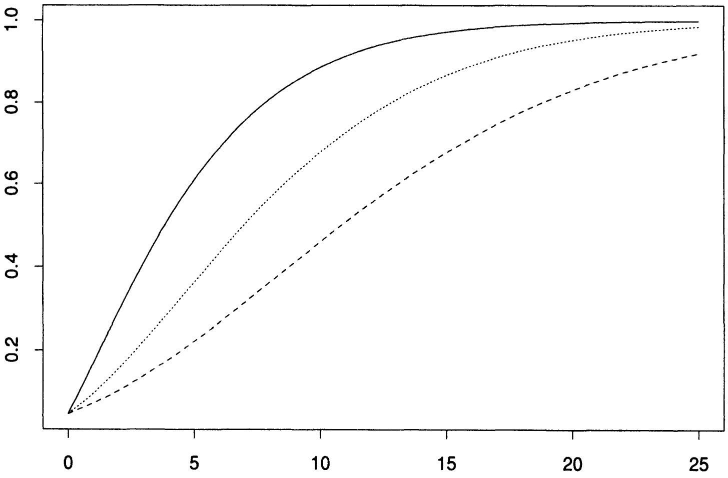

Figure 16.1. The functions $\mu^2 \to \mathrm{P}\big(\chi^2_k(\mu) > \chi^2_{k,\alpha}\big)$ for $k = 1$ (*solid*), $k = 5$ (*dotted*) and $k = 15$ (*dashed*), respectively, for $\alpha = 0.05$.

the covariance matrix. In this sense, the apparently unequal distribution of power over the different directions is not unfair in that it reflects the intrinsic difficulty of detecting changes in different directions. This is not to say that we should never change the (automatic) emphasis given by the likelihood ratio test.

16.5 Bartlett Correction

The chi-square approximation to the distribution of the likelihood ratio statistic is relatively accurate but can be much improved by a correction. This was first noted in the example of testing for inequality of the variances in the one-way layout by Bartlett and has since been generalized. Although every approximation can be improved, the *Bartlett correction* appears to enjoy a particular popularity.

The correction takes the form of a correction of the (asymptotic) mean of the likelihood ratio statistic. In regular cases the distribution of the likelihood ratio statistic is asymptotically chi-square with, say, r degrees of freedom, whence its mean ought to be approximately equal to r. Bartlett's correction is intended to make the mean exactly equal to r, by replacing the likelihood ratio statistic Λ_n by

$$\frac{r\Lambda_n}{\mathrm{E}_{\theta_0}\Lambda_n}.$$

The distribution of this statistic is next approximated by a chi-square distribution with r degrees of freedom. Unfortunately, the mean $\mathrm{E}_{\theta_0}\Lambda_n$ may be hard to calculate, and may depend on an unknown null parameter θ_0. Therefore, one first obtains an expression for the mean of the form

$$\mathrm{E}_{\theta_0}\Lambda_n = 1 + \frac{b(\theta_0)}{n} + \cdots.$$

Next, with $\hat{b}_n$ an appropriate estimator for the parameter $b(\theta_0)$, the corrected statistic takes the form

$$\frac{r\Lambda_n}{1 + \hat{b}_n/n}.$$

The surprising fact is that this recipe works in some generality. Ordinarily, improved approximations would be obtained by writing down and next inverting an Edgeworth expansion of the probabilities $\mathrm{P}(\Lambda_n \leq x)$; the correction would depend on x. In the present case this is equivalent to a simple correction of the mean, independent of x. The technical reason is that the polynomial in x in the $(1/n)$-term of the Edgeworth expansion is of degree 1.[†]

*16.6 Bahadur Efficiency

The claim in the Section 16.4 that in many situations "asymptotically optimal" tests do not exist refers to the study of efficiency relative to the local Gaussian approximations described

[†] For a further discussion, see [5], [9], and [83], and the references cited there.

in Chapter 7. The purpose of this section is to show that, under regularity conditions, the likelihood ratio test is asymptotically optimal in a different setting, the one of Bahadur efficiency.

For simplicity we restrict ourselves to the testing of finite hypotheses. Given finite sets $\mathcal{P}_0$ and $\mathcal{P}_1$ of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and a random sample $X_1, \ldots, X_n$, we study the log likelihood ratio statistic

$$\tilde{\Lambda}_n = \log \frac{\sup_{Q \in \mathcal{P}_1} \prod_{i=1}^n q(X_i)}{\sup_{P \in \mathcal{P}_0} \prod_{i=1}^n p(X_i)}.$$

More general hypotheses can be treated, under regularity conditions, by finite approximation (see e.g., Section 10 of [4]).

The *observed level* of a test that rejects for large values of a statistic T_n is defined as

$$L_n = \sup_{P \in \mathcal{P}_0} \mathrm{P}_P(T_n \geq t)_{|t=T_n}.$$

The test that rejects the null hypothesis if $L_n \leq \alpha$ has level α. The power of this test is maximal if L_n is "minimal" under the alternative (in a stochastic sense). The *Bahadur slope* under the alternative Q is defined as the limit in probability under Q (if it exists) of the sequence $(-2/n) \log L_n$. If this is "large," then L_n is small and hence we prefer sequences of test statistics that have a large slope. The same conclusion is reached in section 14.4 by considering the asymptotic relative Bahadur efficiencies. It is indicated there that the Neyman-Pearson tests for testing the simple null and alternative hypotheses P and Q have Bahadur slope $-2Q \log(p/q)$. Because these are the most powerful tests, this is the maximal slope for testing P versus Q. (We give a precise proof in the following theorem.) Consequently, the slope for a general null hypothesis cannot be bigger than $\inf_{P \in \mathcal{P}_0} -2Q \log(p/q)$. The sequence of likelihood ratio statistics attains equality, even if the alternative hypothesis is composite.

16.12 ***Theorem.*** *The Bahadur slope of any sequence of test statistics for testing an arbitrary null hypothesis $H_0 : P \in \mathcal{P}_0$ versus a simple alternative $H_1 : P = Q$ is bounded above by $\inf_{P \in \mathcal{P}_0} -2Q \log(p/q)$, for any probability measure Q. If $\mathcal{P}_0$ and $\mathcal{P}_1$ are finite sets of probability measures, then the sequence of likelihood ratio statistics for testing $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$ attains equality for every $Q \in \mathcal{P}_1$.*

Proof. Because the observed level is a supremum over $\mathcal{P}_0$, it suffices to prove the upper bound of the theorem for a simple null hypothesis $\mathcal{P}_0 = \{P\}$. If $-2Q \log(p/q) = \infty$, then there is nothing to prove. Thus, we can assume without loss of generality that Q is absolutely continuous with respect to P. Write Λ_n for $\log \prod_{i=1}^n (q/p)(X_i)$. Then, for any constants $B > A > Q \log(q/p)$,

$$\begin{aligned} \mathrm{P}_Q(L_n < e^{-nB}, \Lambda_n < nA) &= \mathrm{E}_P 1\{L_n < e^{-nB}, \Lambda_n < nA\} e^{\Lambda_n} \\ &\leq e^{nA} \mathrm{P}_P(L_n < e^{-nB}). \end{aligned}$$

Because L_n is superuniformly distributed under the null hypothesis, the last expression is bounded above by $\exp -n(B - A)$. Thus, the sum of the probabilities on the left side over $n \in \mathbb{N}$ is finite, whence $-(2/n) \log L_n \leq 2B$ or $\Lambda_n \geq nA$ for all sufficiently large n, almost surely under Q, by the Borel-Cantelli lemma. Because the sequence $n^{-1}\Lambda_n$

converges almost surely under Q to $Q\log(q/p) < A$, by the strong law of large numbers, the second possibility can occur only finitely many times. It follows that $-(2/n)\log L_n \leq 2B$ eventually, almost surely under Q. This having been established for any $B > Q\log(q/p)$, the proof of the first assertion is complete.

To prove that the likelihood ratio statistic attains equality, it suffices to prove that its slope is bigger than the upper bound. Write $\tilde{\Lambda}_n$ for the log likelihood ratio statistic, and write $\sup_P$ and $\sup_Q$ for suprema over the null and alternative hypotheses. Because $(1/n)\tilde{\Lambda}_n$ is bounded above by $\sup_Q \mathbb{P}_n \log(q/p)$, we have, by Markov's inequality,

$$\mathrm{P}_P\left(\frac{1}{n}\tilde{\Lambda}_n \geq t\right) \leq \sum_Q \mathrm{P}_P\left(\mathbb{P}_n \log\frac{q}{p} \geq t\right) \leq |\mathcal{P}_1| \max_Q e^{-nt}\mathrm{E}_P e^{n\mathbb{P}_n \log(q/p)}.$$

The expectation on the right side is the nth power of the integral $\int (q/p)\,dP = Q(p > 0) \leq 1$. Take logarithms left and right and multiply with $-(2/n)$ to find that

$$-\frac{2}{n}\log \mathrm{P}_P\left(\frac{1}{n}\tilde{\Lambda}_n \geq t\right) \geq 2t - \frac{2\log|\mathcal{P}_1|}{n}.$$

Because this is valid uniformly in t and P, we can take the infimum over P on the left side; next evaluate the left and right sides at $t = (1/n)\tilde{\Lambda}_n$. By the law of large numbers, $\mathbb{P}_n \log(q/p) \to Q\log(q/p)$ almost surely under Q, and this remains valid if we first add the infimum over the (finite) set $\mathcal{P}_0$ on both sides. Thus, the limit inferior of the sequence $(1/n)\tilde{\Lambda}_n \geq \inf_P \mathbb{P}_n \log(q/p)$ is bounded below by $\inf_P Q\log(q/p)$ almost surely under Q, where we interpret $Q\log(q/p)$ as ∞ if $Q(p = 0) > 0$. Insert this lower bound in the preceding display to conclude that the Bahadur slope of the likelihood ratio statistics is bounded below by $2\inf_P Q\log(q/p)$. ■

Notes

The classical references on the asymptotic null distribution of likelihood ratio statistic are papers by Chernoff [21] and Wilks [150]. Our main theorem appears to be better than Chernoff's, who uses the "classical regularity conditions" and a different notion of approximation of sets, but is not essentially different. Wilks' treatment would not be acceptable to present-day referees but maybe is not so different either. He appears to be saying that we can replace the original likelihood by the likelihood for having observed only the maximum likelihood estimator (the error is asymptotically negligible), next refers to work by Doob to infer that this is a Gaussian likelihood, and continues to compute the likelihood ratio statistic for a Gaussian likelihood, which is easy, as we have seen. The approach using a Taylor expansion and the asymptotic distributions of both likelihood estimators is one way to make the argument rigorous, but it seems to hide the original intuition.

Bahadur [3] presented the efficiency of the likelihood ratio statistic at the fifth Berkeley symposium. Kallenberg [84] shows that the likelihood ratio statistic remains asymptotically optimal in the setting in which both the desired level and the alternative tend to zero, at least in exponential families. As the proof of Theorem 16.12 shows, the composite nature of the alternative hypothesis "disappears" elegantly by taking $(1/n)$ log of the error probabilities – too elegantly to attach much value to this type of optimality?

PROBLEMS

1. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from the bivariate normal distribution with mean vector (μ, ν) and covariance matrix the diagonal matrix with entries σ^2 and τ^2. Calculate (or characterize) the likelihood ratio statistic for testing $H_0 : \mu = \nu$ versus $H_1 : \mu \neq \nu$.

2. Let N be a kr-dimensional multinomial variable written as a $(k \times r)$ matrix (N_{ij}). Calculate the likelihood ratio statistic for testing the null hypothesis of independence $H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}$ for every i and j. Here the dot denotes summation over all columns and rows, respectively. What is the limit distribution under the null hypothesis?

3. Calculate the likelihood ratio statistic for testing $H_0 : \mu = \nu$ based on independent samples of size n from multivariate normal distributions $N_r(\mu, \Sigma)$ and $N_r(\nu, \Sigma)$. The matrix Σ is unknown. What is the limit distribution under the null hypothesis?

4. Calculate the likelihood ratio statistic for testing $H_0 : \mu_1 = \cdots = \mu_k$ based on k independent samples of size n from $N(\mu_j, \sigma^2)$-distributions. What is the asymptotic distribution under the null hypothesis?

5. Show that $(I_\vartheta^{-1})_{>l,>l}$ is the inverse of the matrix $I_{\vartheta,>l,>l} - I_{\vartheta,>l,\leq l} I_{\vartheta,\leq l,\leq l}^{-1} I_{\vartheta,\leq l,>l}$.

6. Study the asymptotic distribution of the sequence $\tilde{\Lambda}_n$ if the true parameter is contained in both the null and alternative hypotheses.

7. Study the asymptotic distribution of the likelihood ratio statistics for testing the hypothesis $H_0 : \sigma = -\tau$ based on a sample of size n from the uniform distribution on $[\sigma, \tau]$. Does the asymptotic distribution correspond to a likelihood ratio statistic in a limit experiment?