14

Relative Efficiency of Tests

The quality of sequences of tests can be judged from their power at alternatives that become closer and closer to the null hypothesis. This motivates the study of local asymptotic power functions. The relative efficiency of two sequences of tests is the quotient of the numbers of observations needed with the two tests to obtain the same level and power. We discuss several types of asymptotic relative efficiencies.

14.1 Asymptotic Power Functions

Consider the problem of testing a null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative $H_1 : \theta \in \Theta_1$. The power function of a test that rejects the null hypothesis if a test statistic falls into a *critical region* K_n is the function $\theta \mapsto \pi_n(\theta) = \mathrm{P}_\theta(T_n \in K_n)$, which gives the probability of rejecting the null hypothesis. The test is of *level* α if its *size* $\sup\{\pi_n(\theta) : \theta \in \Theta_0\}$ does not exceed α. A sequence of tests is called *asymptotically of level* α if

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_0} \pi_n(\theta) \leq \alpha.$$

(An alternative definition is to drop the supremum and require only that $\limsup \pi_n(\theta) \leq \alpha$ for every $\theta \in \Theta_0$.) A test with power function π_n is better than a test with power function $\underline{\pi}_n$ if both

$$\begin{aligned} \pi_n(\theta) &\leq \underline{\pi}_n(\theta), \qquad && \theta \in \Theta_0, \\ \text{and} \quad \pi_n(\theta) &\geq \underline{\pi}_n(\theta), \qquad && \theta \in \Theta_1. \end{aligned}$$

The aim of this chapter is to compare tests asymptotically. We consider sequences of tests with power functions π_n and $\underline{\pi}_n$ and wish to decide which of the sequences is best as $n \to \infty$. Typically, the tests corresponding to a sequence $\pi_1, \pi_2, \ldots$ are of the same type. For instance, they are all based on a certain U-statistic or rank statistic, and only the number of observations changes with n. Otherwise the comparison would have little relevance.

A first idea is to consider limiting power functions of the form

$$\pi(\theta) = \lim_{n\to\infty} \pi_n(\theta).$$

If this limit exists for all θ, and the same is true for the competing tests $\underline{\pi}_n$, then the sequence π_n is better than the sequence $\underline{\pi}_n$ if the limiting power function π is better than the limiting power function $\underline{\pi}$. It turns out that this approach is too naive. The limiting power functions typically exist, but they are trivial and identical for all reasonable sequences of tests.

14.1 ***Example (Sign test).*** Suppose the observations $X_1, \dots, X_n$ are a random sample from a distribution with unique median θ. The null hypothesis $H_0 : \theta = 0$ can be tested against the alternative $H_1 : \theta > 0$ by means of the *sign statistic* $S_n = n^{-1} \sum_{i=1}^n 1\{X_i > 0\}$. If $F(x - \theta)$ is the distribution function of the observations, then the expectation and variance of S_n are equal to $\mu(\theta) = 1 - F(-\theta)$ and $\sigma^2(\theta)/n = \big(1 - F(-\theta)\big)F(-\theta)/n$, respectively. By the normal approximation to the binomial distribution, the sequence $\sqrt{n}\big(S_n - \mu(\theta)\big)$ is asymptotically normal $N\big(0, \sigma^2(\theta)\big)$. Under the null hypothesis the mean and variance are equal to $\mu(0) = 1/2$ and $\sigma^2(0) = 1/4$, respectively, so that $\sqrt{n}(S_n - 1/2) \overset{0}{\rightsquigarrow} N(0, 1/4)$. The test that rejects the null hypothesis if $\sqrt{n}(S_n - 1/2)$ exceeds the critical value $\frac{1}{2} z_\alpha$ has power function

$$\begin{aligned} \pi_n(\theta) &= \mathrm{P}_\theta\Big(\sqrt{n}\big(S_n - \mu(\theta)\big) > \tfrac{1}{2} z_\alpha - \sqrt{n}\big(\mu(\theta) - \mu(0)\big)\Big) \\ &= 1 - \Phi\left(\frac{\frac{1}{2} z_\alpha - \sqrt{n}\big(F(0) - F(-\theta)\big)}{\sigma(\theta)}\right) + o(1). \end{aligned}$$

Because $F(0) - F(-\theta) > 0$ for every $\theta > 0$, it follows that for $\alpha = \alpha_n \to 0$ sufficiently slowly

$$\pi_n(\theta) \to \begin{cases} 0 & \text{if } \theta = 0, \\ 1 & \text{if } \theta > 0. \end{cases}$$

The limit power function corresponds to the perfect test with all error probabilities equal to zero. □

The example exhibits a sequence of tests whose (pointwise) limiting power function is the perfect power function. This type of behavior is typical for all reasonable tests. The point is that, with arbitrarily many observations, it should be possible to tell the null and alternative hypotheses apart with complete accuracy. The power at every fixed alternative should therefore converge to 1.

14.2 **Definition.** A sequence of tests with power functions $\theta \mapsto \pi_n(\theta)$ is asymptotically *consistent* at level α at (or against) the alternative θ if it is asymptotically of level α and $\pi_n(\theta) \to 1$. If a family of sequences of tests contains for every level $\alpha \in (0, 1)$ a sequence that is consistent against every alternative, then the corresponding tests are simply called consistent.

Consistency is an optimality criterion for tests, but because most sequences of tests are consistent, it is too weak to be really useful. To make an informative comparison between sequences of (consistent) tests, we shall study the performance of the tests in problems that become harder as more observations become available. One way of making a testing problem harder is to choose null and alternative hypotheses closer to each other. In this section we fix the null hypothesis and consider the power at sequences of alternatives that converge to the null hypothesis.

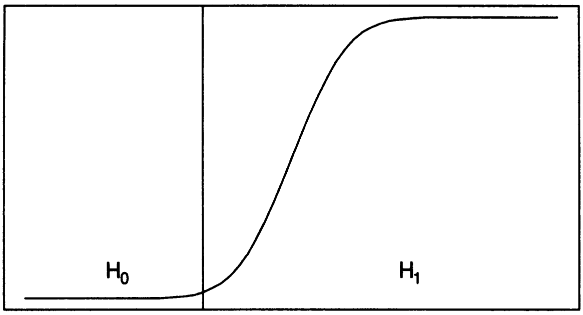

Figure 14.1. Asymptotic power function.

14.3 ***Example (Sign test, continued).*** Consider the power of the sign test at sequences of alternatives $\theta_n \downarrow 0$. Suppose that the null hypothesis $H_0 : \theta = 0$ is rejected if $\sqrt{n}(S_n - \frac{1}{2}) \geq \frac{1}{2} z_\alpha$. Extension of the argument of the preceding example yields

$$\pi_n(\theta_n) = 1 - \Phi\left(\frac{\frac{1}{2}z_\alpha - \sqrt{n}\big(F(0) - F(-\theta_n)\big)}{\sigma(\theta_n)}\right) + o(1).$$

Since $\sigma(0) = \frac{1}{2}$, the levels $\pi_n(0)$ of the tests converge to $\Phi(z_\alpha) = \alpha$. The asymptotic power at θ_n depends on the rate at which $\theta_n \to 0$. If θ_n converges to zero fast enough to ensure that $\sqrt{n}\big(F(0) - F(-\theta_n)\big) \to 0$, then the power $\pi_n(\theta_n)$ converges to α: the sign test is not able to discriminate these alternatives from the null hypothesis. If θ_n converges to zero at a slow rate, then $\sqrt{n}\big(F(0) - F(-\theta_n)\big) \to \infty$, and the asymptotic power is equal to 1: these alternatives are too easy. The intermediate rates, which yield a nontrivial asymptotic power, appear to be of most interest. Suppose that the underlying distribution function F is differentiable at zero with positive derivative $f(0) > 0$. Then

$$\sqrt{n}\big(F(0) - F(-\theta_n)\big) = \sqrt{n}\,\theta_n f(0) + \sqrt{n}\,o(\theta_n).$$

This is bounded away from zero and infinity if θ_n converges to zero at rate $\theta_n = O(n^{-1/2})$. For such rates the power $\pi_n(\theta_n)$ is asymptotically strictly between α and 1. In particular, for every h,

$$\pi_n\left(\frac{h}{\sqrt{n}}\right) \to 1 - \Phi\big(z_\alpha - 2hf(0)\big).$$

The form of the limit power function is shown in Figure 14.1. □

In the preceding example only alternatives θ_n that converge to the null hypothesis at rate $O(1/\sqrt{n})$ lead to a nontrivial asymptotic power. This is typical for parameters that depend "smoothly" on the underlying distribution. In this situation a reasonable method for asymptotic comparison of two sequences of tests for $H_0 : \theta = 0$ versus $H_0 : \theta > 0$ is to consider *local limiting power* functions, defined as

$$\pi(h) = \lim_{n\to\infty} \pi_n\left(\frac{h}{\sqrt{n}}\right), \qquad h \geq 0.$$

These limits typically exist and can be derived by the same method as in the preceding example. A general scheme is as follows.

Let θ be a real parameter and let the tests reject the null hypothesis $H_0 : \theta = 0$ for large values of a test statistic T_n. Assume that the sequence T_n is asymptotically normal in the sense that, for all sequences of the form $\theta_n = h/\sqrt{n}$,

$$\frac{\sqrt{n}\big(T_n - \mu(\theta_n)\big)}{\sigma(\theta_n)} \overset{\theta_n}{\rightsquigarrow} N(0,1). \tag{14.4}$$

Often $\mu(\theta)$ and $\sigma^2(\theta)$ can be taken to be the mean and the variance of T_n, but this is not necessary. Because the convergence (14.4) is under a law indexed by θ_n that changes with n, the convergence is not implied by

$$\frac{\sqrt{n}\big(T_n - \mu(\theta)\big)}{\sigma(\theta)} \overset{\theta}{\rightsquigarrow} N(0,1), \qquad \text{every } \theta. \tag{14.5}$$

On the other hand, this latter convergence uniformly in the parameter θ is more than is needed in (14.4). The convergence (14.4) is sometimes referred to as "locally uniform" asymptotic normality. "Contiguity arguments" can reduce the derivation of asymptotic normality under $\theta_n = h/\sqrt{n}$ to derivation under $\theta = 0$. (See section 14.1.1).

Assumption (14.4) includes that the sequence $\sqrt{n}\big(T_n - \mu(0)\big)$ converges in distribution to a normal $N\big(0, \sigma^2(0)\big)$-distribution under $\theta = 0$. Thus, the tests that reject the null hypothesis $H_0 : \theta = 0$ if $\sqrt{n}\big(T_n - \mu(0)\big)$ exceeds $\sigma(0)z_\alpha$ are asymptotically of level α. The power functions of these tests can be written

$$\pi_n(\theta_n) = \mathrm{P}_{\theta_n}\Big(\sqrt{n}\big(T_n - \mu(\theta_n)\big) > \sigma(0)z_\alpha - \sqrt{n}\big(\mu(\theta_n) - \mu(0)\big)\Big).$$

For $\theta_n = h/\sqrt{n}$, the sequence $\sqrt{n}\big(\mu(\theta_n) - \mu(0)\big)$ converges to $h\mu'(0)$ if μ is differentiable at zero. If $\sigma(\theta_n) \to \sigma(0)$, then under (14.4)

$$\pi_n\Big(\frac{h}{\sqrt{n}}\Big) \to 1 - \Phi\Big(z_\alpha - h\frac{\mu'(0)}{\sigma(0)}\Big). \tag{14.6}$$

For easy reference we formulate this result as a theorem.

14.7 ***Theorem.*** *Let μ and σ be functions of θ such that (14.4) holds for every sequence $\theta_n = h/\sqrt{n}$. Suppose that μ is differentiable and that σ is continuous at $\theta = 0$. Then the power functions π_n of the tests that reject $H_0 : \theta = 0$ for large values of T_n and are asymptotically of level α satisfy (14.6) for every h.*

The limiting power function depends on the sequence of test statistics only through the quantity $\mu'(0)/\sigma(0)$. This is called the *slope* of the sequence of tests. Two sequences of tests can be asymptotically compared by just comparing the sizes of their slopes. The bigger the slope, the better the test for $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. The size of the slope depends on the rate $\mu'(0)$ of change of the asymptotic mean of the test statistics relative to their asymptotic dispersion $\sigma(0)$. A good quantitative measure of comparison is the square of the quotient of two slopes. This quantity is called the *asymptotic relative efficiency* and is discussed in section 14.3.

If θ is the only unknown parameter in the problem, then the available tests can be ranked in asymptotic quality simply by the value of their slopes. In many problems there are also nuisance parameters (for instance the shape of a density), and the slope is a function of the nuisance parameter rather than a number. This complicates the comparison considerably. For every value of the nuisance parameter a different test may be best, and additional criteria are needed to choose a particular test.

14.8 ***Example (Sign test).*** According to Example 14.3, the sign test has slope $2f(0)$. This can also be obtained from the preceding theorem, in which we can choose $\mu(\theta) = 1-F(-\theta)$ and $\sigma^2(\theta) = \big(1 - F(-\theta)\big)F(-\theta)$. □

14.9 ***Example (t-test).*** Let $X_1, \dots, X_n$ be a random sample from a distribution with mean θ and finite variance. The t-test rejects the null hypothesis for large values of Σ. The sample variance S^2 converges in probability to the variance σ^2 of a single observation. The central limit theorem and Slutsky's lemma give

$$\sqrt{n}\left(\frac{\overline{X}}{S} - \frac{h/\sqrt{n}}{\sigma}\right) = \frac{\sqrt{n}(\overline{X} - h/\sqrt{n})}{S} + h\left(\frac{1}{S} - \frac{1}{\sigma}\right) \overset{h/\sqrt{n}}{\rightsquigarrow} N(0, 1).$$

Thus Theorem 14.7 applies with $\mu(\theta) = \theta/\sigma$ and $\sigma(\theta) = 1$. The slope of the t-test equals $1/\sigma$.[†] □

14.10 ***Example (Sign versus t-test).*** Let $X_1, \dots, X_n$ be a random sample from a density $f(x - \theta)$, where f is symmetric about zero. We shall compare the performance of the sign test and the t-test for testing the hypothesis $H_0 : \theta = 0$ that the observations are symmetrically distributed about zero. Assume that the distribution with density f has a unique median and a finite second moment.

It suffices to compare the slopes of the two tests. By the preceding examples these are $2f(0)$ and $\big(\int x^2 f(x)\,dx\big)^{-1/2}$, respectively. Clearly the outcome of the comparison depends on the shape f. It is interesting that the two slopes depend on the underlying shape in an almost orthogonal manner. The slope of the sign test depends only on the height of f at zero; the slope of the t-test depends mainly on the tails of f. For the standard normal distribution the slopes are $\sqrt{2/\pi}$ and 1. The superiority of the t-test in this case is not surprising, because the t-test is uniformly most powerful for every n. For the Laplace distribution, the ordering is reversed: The slopes are 1 and $\frac{1}{2}\sqrt{2}$. The superiority of the sign test has much to do with the "unsmooth" character of the Laplace density at its mode.

The relative efficiency of the sign test versus the t-test is equal to

$$4f^2(0)\int x^2 f(x)\,dx.$$

Table 14.1 summarizes these numbers for a selection of shapes. For the uniform distribution, the relative efficiency of the sign test with respect to the t-test equals $1/3$. It can be shown that this is the minimal possible value over all densities with mode zero (problem 14.7). On the other hand, it is possible to construct distributions for which this relative efficiency is arbitrarily large, by shifting mass into the tails of the distribution. The sign test is "robust" against heavy tails, the t-test is not. □

The simplicity of comparing slopes is attractive on the one hand, but indicates the potential weakness of asymptotics on the other. For instance, the slope of the sign test was seen to be $f(0)$, but it is clear that this value alone cannot always give an accurate indication

[†] Although (14.4) holds with this choice of μ and σ, it is not true that the sequence $\sqrt{n}\big(\overline{X}/S - \theta/\sigma\big)$ is asymptotically standard normal for every fixed θ. Thus (14.5) is false for this choice of μ and σ. For fixed θ the contribution of $S - \sigma$ to the limit distribution cannot be neglected, but for our present purpose it can.

Table 14.1. *Relative efficiencies of the sign test versus the t-test for some distributions.*

Distribution	Efficiency (sign/t-test)
Logistic	$\pi^2/12$
Normal	$2/\pi$
Laplace	2
Uniform	1/3

of the quality of the sign test. Consider a density that is basically a normal density, but a tiny proportion of $10^{-10}\%$ of its total mass is located under an extremely thin but enormously high peak at zero. The large value $f(0)$ would strongly favor the sign test. However, at moderate sample sizes the observations would not differ significantly from a sample from a normal distribution, so that the t-test is preferable. In this situation the asymptotics are only valid for unrealistically large sample sizes.

Even though asymptotic approximations should always be interpreted with care, in the present situation there is actually little to worry about. Even for $n = 20$, the comparison of slopes of the sign test and the t-test gives the right message for the standard distributions listed in Table 14.1.

14.11 ***Example (Mann-Whitney).*** Suppose we observe two independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from distributions $F(x)$ and $G(y - \theta)$, respectively. The base distributions F and G are fixed, and it is desired to test the null hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta > 0$. Set $N = m + n$ and assume that $m/N \to \lambda \in (0, 1)$. Furthermore, assume that G has a bounded density g.

The Mann-Whitney test rejects the null hypothesis for large numbers of $U = (mn)^{-1} \sum_i \sum_j 1\{X_i \leq Y_j\}$. By the two-sample U-statistic theorem

$$\sqrt{N}\big(U - \mathrm{P}_\theta(X \leq Y)\big) = -\frac{\sqrt{N}}{m} \sum_{i=1}^{m} \big(G(X_i - \theta) - \mathrm{E}G(X_i - \theta)\big) + \frac{\sqrt{N}}{n} \sum_{j=1}^{n} \big(F(Y_i) - \mathrm{E}F(Y_i)\big) + o_{P_\theta}(1).$$

This readily yields the asymptotic normality (14.5) for every fixed θ, with

$$\mu(\theta) = 1 - \int G(x - \theta)\, dF(x), \qquad \sigma^2(\theta) = \frac{1}{\lambda} \operatorname{var} G(X - \theta) + \frac{1}{1 - \lambda} \operatorname{var} F(Y).$$

To obtain the local asymptotic power function, this must be extended to sequences $\theta_N = h/\sqrt{N}$. It can be checked that the U-statistic theorem remains valid and that the Lindeberg central limit theorem applies to the right side of the preceding display with θ_N replacing θ. Thus, we find that (14.4) holds with the same functions μ and σ. (Alternatively we can use contiguity and Le Cam's third lemma.) Hence, the slope of the Mann-Whitney test equals $\mu'(0)/\sigma(0) = \int g\, dF/\sigma(0)$. □

14.12 ***Example (Two-sample t-test).*** In the set-up of the preceding example suppose that the base distributions F and G have equal means and finite variances. Then $\theta = \mathrm{E}(Y - X)$

Table 14.2. *Relative efficiencies of the Mann-Whitney test versus the two-sample t-test if $f = g$ equals a number of distributions.*

Distribution	Efficiency (Mann-Whitney/two-sample t-test)
Logistic	$\pi^2/9$
Normal	$3/\pi$
Laplace	$3/2$
Uniform	1
t_3	1.24
t_5	1.90
$c(1-x^2)\vee 0$	108/125

and the t-test rejects the null hypothesis $H_0: \theta = 0$ for large values of the statistic $(\bar{Y}-\bar{X})/S$, where $S^2/N = S_X^2/m + S_Y^2/n$ is the unbiased estimator of $\operatorname{var}(\bar{Y}-\bar{X})$. The sequence S^2 converges in probability to $\sigma^2 = \operatorname{var} X/\lambda + \operatorname{var} Y/(1-\lambda)$. By Slutsky's lemma and the central limit theorem

$$\sqrt{N}\left(\frac{\bar{Y}-\bar{X}}{S} - \frac{h/\sqrt{N}}{\sigma}\right) \overset{h/\sqrt{N}}{\rightsquigarrow} N(0,1).$$

Thus (14.4) is satisfied and Theorem 14.7 applies with $\mu(\theta) = \theta/\sigma$ and $\sigma(\theta) = 1$. The slope of the t-test equals $\mu'(0)/\sigma(0) = 1/\sigma$. □

14.13 ***Example (t-Test versus Mann-Whitney test).*** Suppose we observe two independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from distributions $F(x)$ and $G(x-\theta)$, respectively. The base distributions F and G are fixed and are assumed to have equal means and bounded densities. It is desired to test the null hypothesis $H_0: \theta = 0$ versus the alternative $H_1: \theta > 0$. Set $N = m+n$ and assume that $m/N \to \lambda \in (0,1)$.

The slopes of the Mann-Whitney test and the t-test depend on the nuisance parameters F and G. According to the preceding examples the relative efficiency of the two sequences of tests equals

$$\frac{\big((1-\lambda)\operatorname{var} X + \lambda \operatorname{var} Y\big)\left(\int g\,dF\right)^2}{(1-\lambda)\operatorname{var}_0 G(X) + \lambda \operatorname{var}_0 F(Y)}.$$

In the important case that $F = G$, this expression simplifies. Then the variables $G(X)$ and $F(Y)$ are uniformly distributed on $[0,1]$. Hence they have variance $1/12$ and the relative efficiency reduces to $12 \operatorname{var} X\big(\int f^2(y)\,dy\big)^2$. Table 14.2 gives the relative efficiency if $F = G$ are both equal to a number of standard distributions. The Mann-Whitney test is inferior to the t-test if $F = G$ equals the normal distribution, but better for the logistic, Laplace, and t-distribution. Even for the normal distribution the Mann-Whitney test does remarkably well, with a relative efficiency of $3/\pi \approx 95\%$. The density that is proportional to $(1-x^2)\vee 0$ (and any member of its scale family) is least favorable for the Mann-Whitney test. This density yields the lowest possible relative efficiency, which is still equal to $108/125 \approx 86\%$ (problem 14.8). On the other hand, the relative efficiency of the Mann-Whitney test is large for heavy-tailed distributions; the supremum value is infinite. Together with the fact that the Mann-Whitney test is distribution-free under the null hypothesis, this makes the Mann-Whitney test a strong competitor to the t-test, even in situations in which the underlying distribution is thought to be approximately normal. □

*14.1.1 *Using Le Cam's Third Lemma*

In the preceding examples the asymptotic normality of sequences of test statistics was established by direct methods. For more complicated test statistics the validity of (14.4) is easier checked by means of Le Cam's third lemma. This is illustrated by the following example.

14.14 ***Example (Median test).*** In the two-sample set-up of Example 14.11, suppose that $F = G$ is a continuous distribution function with finite Fisher information for location I_g. The median test rejects the null hypothesis $H_0 : \theta = 0$ for large values of the rank statistic $T_N = N^{-1}\sum_{i=m+1}^N 1\{R_{Ni} \leq (N+1)/2\}$. By the rank central limit theorem, Theorem 13.5, under the null hypothesis,

$$\sqrt{N}\left(T_N - \frac{n}{2N}\right) = -\frac{n}{N\sqrt{N}}\sum_{i=1}^m 1\{F(X_i) \leq 1/2\} + \frac{m}{N\sqrt{N}}\sum_{j=1}^n 1\{F(Y_j) \leq 1/2\} + o_P(1).$$

Under the null hypothesis the sequence of variables on the right side is asymptotically normal with mean zero and variance $\sigma^2(0) = \lambda(1-\lambda)/4$. By Theorem 7.2, for every $\theta_N = h/\sqrt{N}$,

$$\log\frac{\prod_i f(X_i)\prod_j g(Y_j - \theta_N)}{\prod_i f(X_i)\prod_j g(Y_j)} = -\frac{h\sqrt{1-\lambda}}{\sqrt{n}}\sum_{j=1}^n \frac{g'}{g}(Y_i) - \frac{1}{2}h^2(1-\lambda)I_g + o_P(1).$$

By the multivariate central limit theorem, the linear approximations on the right sides of the two preceding displays are jointly asymptotically normal. By Slutsky's lemma the same is true for the left sides. Consequently, by Le Cam's third lemma the sequence $\sqrt{N}\big(T_N - n/(2N)\big)$ converges under the alternatives $\theta_N = h/\sqrt{N}$ in distribution to a normal distribution with variance $\sigma^2(0)$ and mean the asymptotic covariance $\tau(h)$ of the linear approximations. This is given by

$$\tau(h) = -h\lambda(1-\lambda)\int_{F(y)\leq 1/2}\frac{f'}{f}(y)\,dF(y).$$

Conclude that (14.4) is valid with $\mu(\theta) = \tau(\theta)$ and $\sigma(\theta) = \sigma(0)$. (Use the test statistics $T_N - n/(2N)$ rather than T_N.) The slope of the median test is given by $-2\sqrt{\lambda(1-\lambda)}\int_0^{1/2} (f'/f)\big(F^{-1}(u)\big)\,du$. □

14.2 Consistency

After noting that the power at fixed alternatives typically tends to 1, we focused attention on the performance of tests at alternatives converging to the null hypothesis. The comparison of local power functions is only of interest if the sequences of tests are consistent at fixed alternatives. Fortunately, establishing consistency is rarely a problem. The following lemmas describe two basic methods.

14.15 ***Lemma.*** *Let T_n be a sequence of statistics such that $T_n \overset{P_\theta}{\rightarrow} \mu(\theta)$ for every θ. Then the family of tests that reject the null hypothesis $H_0 : \theta = 0$ for large values of T_n is consistent against every θ such that $\mu(\theta) > \mu(0)$.*

14.16 ***Lemma.*** *Let μ and σ be functions of θ such that (14.4) holds for every sequence $\theta_n = h/\sqrt{n}$. Suppose that μ is differentiable and that σ is continuous at zero, with $\mu'(0) > 0$ and $\sigma(0) > 0$. Suppose that the tests that reject the null hypothesis for large values of T_n possess nondecreasing power functions $\theta \mapsto \pi_n(\theta)$. Then this family of tests is consistent against every alternative $\theta > 0$. Moreover, if $\pi_n(0) \to \alpha$, then $\pi_n(\theta_n) \to \alpha$ or $\pi_n(\theta_n) \to 1$ when $\sqrt{n}\,\theta_n \to 0$ or $\sqrt{n}\,\theta_n \to \infty$, respectively.*

Proofs. For the first lemma, suppose that the tests reject the null hypothesis if T_n exceeds the critical value c_n. By assumption, the probability under $\theta = 0$ that T_n is outside the interval $\big(\mu(0) - \varepsilon, \mu(0) + \varepsilon\big)$ converges to zero as $n \to \infty$, for every fixed $\varepsilon > 0$. If the asymptotic level $\lim \mathrm{P}_0(T_n > c_n)$ is positive, then it follows that $c_n < \mu(0) + \varepsilon$ eventually. On the other hand, under θ the probability that T_n is in $\big(\mu(\theta) - \varepsilon, \mu(\theta) + \varepsilon\big)$ converges to 1. For sufficiently small ε and $\mu(\theta) > \mu(0)$, this interval is to the right of $\mu(0) + \varepsilon$. Thus for sufficiently large n, the power $\mathrm{P}_\theta(T_n > c_n)$ can be bounded below by $\mathrm{P}_\theta\big(T_n \in \big(\mu(\theta) - \varepsilon, \mu(\theta) + \varepsilon\big)\big) \to 1$.

For the proof of the second lemma, first note that by Theorem 14.7 the sequence of local power functions $\pi_n(h/\sqrt{n})$ converges to $\pi(h) = 1 - \Phi\big(z_\alpha - h\mu'(0)/\sigma(0)\big)$, for every h, if the asymptotic level is α. If $\sqrt{n}\,\theta_n \to 0$, then eventually $\theta_n < h/\sqrt{n}$ for every given $h > 0$. By the monotonicity of the power functions, $\pi_n(\theta_n) \leq \pi_n(h/\sqrt{n})$ for sufficiently large n. Thus $\limsup \pi_n(\theta_n) \leq \pi(h)$ for every $h > 0$. For $h \downarrow 0$ the right side converges to $\pi(0) = \alpha$. Combination with the inequality $\pi_n(\theta_n) \geq \pi_n(0) \to \alpha$ gives $\pi_n(\theta_n) \to \alpha$. The case that $\sqrt{n}\,\theta_n \to \infty$ can be handled similarly. Finally, the power $\pi_n(\theta)$ at fixed alternatives is bounded below by $\pi_n(\theta_n)$ eventually, for every sequence $\theta_n \downarrow 0$. Thus $\pi_n(\theta) \to 1$, and the sequence of tests is consistent at θ. ■

The following examples show that the t-test and Mann-Whitney test are both consistent against large sets of alternatives, albeit not exactly the same sets. They are both tests to compare the locations of two samples, but the pertaining definitions of "location" are not the same. The t-test can be considered a test to detect a difference in mean; the Mann-Whitney test is designed to find a difference of $\mathrm{P}(X \leq Y)$ from its value $1/2$ under the null hypothesis. This evaluation is justified by the following examples and is further underscored by the consideration of asymptotic efficiency in nonparametric models. It is shown in Section 25.6 that the tests are asymptotically efficient for testing the parameters $\mathrm{E}Y - \mathrm{E}X$ or $\mathrm{P}(X \leq Y)$ if the underlying distributions F and G are completely unknown.

14.17 ***Example (t-test).*** The two-sample t-statistic $(\overline{Y} - \overline{X})/S$ converges in probability to $\mathrm{E}(Y - X)/\sigma$, where $\sigma^2 = \lim \mathrm{var}(\overline{Y} - \overline{X})$. If the null hypothesis postulates that $\mathrm{E}Y = \mathrm{E}X$, then the test that rejects the null hypothesis for large values of the t-statistic is consistent against every alternative for which $\mathrm{E}Y > \mathrm{E}X$. □

14.18 ***Example (Mann-Whitney test).*** The Mann-Whitney statistic U converges in probability to $\mathrm{P}(X \leq Y)$, by the two-sample U-statistic theorem. The probability $\mathrm{P}(X \leq Y)$ is equal to $1/2$ if the two samples are equal in distribution and possess a continuous distribution function. If the null hypothesis postulates that $\mathrm{P}(X \leq Y) = 1/2$, then the test that rejects for large values of U is consistent against any alternative for which $\mathrm{P}(X \leq Y) > 1/2$. □

14.3 Asymptotic Relative Efficiency

Sequences of tests can be ranked in quality by comparing their asymptotic power functions. For the test statistics we have considered so far, this comparison only involves the "slopes" of the tests. The concept of relative efficiency yields a method to quantify the interpretation of the slopes.

Consider a sequence of testing problems consisting of testing a null hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta = \theta_\nu$. We use the parameter ν to describe the asymptotics; thus $\nu \to \infty$. We require a priori that our tests attain asymptotically level α and power $\gamma \in (\alpha, 1)$. Usually we can meet this requirement by choosing an appropriate number of observations at "time" ν. A larger number of observations allows smaller level and higher power. If π_n is the power function of a test if n observations are available, then we define n_ν to be the minimal number of observations such that both

$$\pi_{n_\nu}(0) \leq \alpha, \quad \text{and} \quad \pi_{n_\nu}(\theta_\nu) \geq \gamma.$$

If two sequences of tests are available, then we prefer the sequence for which the numbers n_ν are smallest. Suppose that $n_{\nu,1}$ and $n_{\nu,2}$ observations are needed for two given sequences of tests. Then, if it exists, the limit

$$\lim_{\nu\to\infty} \frac{n_{\nu,2}}{n_{\nu,1}}$$

is called the (asymptotic) *relative efficiency* or *Pitman efficiency* of the first with respect to the second sequence of tests. A relative efficiency larger than 1 indicates that fewer observations are needed with the first sequence of tests, which may then be considered the better one.

In principle, the relative efficiency may depend on α, γ and the sequence of alternatives θ_ν. The concept is mostly of interest if the relative efficiency is the same for all possible choices of these parameters. This is often the case. In particular, in the situations considered previously, the relative efficiency turns out to be the square of the quotient of the slopes.

14.19 ***Theorem.*** *Consider statistical models $(P_{n,\theta} : \theta \geq 0)$ such that $\|P_{n,\theta} - P_{n,0}\| \to 0$ as $\theta \to 0$, for every n. Let $T_{n,1}$ and $T_{n,2}$ be sequences of statistics that satisfy (14.4) for every sequence $\theta_n \downarrow 0$ and functions μ_i and σ_i such that μ_i is differentiable at zero and σ_i is continuous at zero, with $\mu_i'(0) > 0$ and $\sigma_i(0) > 0$. Then the relative efficiency of the tests that reject the null hypothesis $H_0 : \theta = 0$ for large values of $T_{n,i}$ is equal to*

$$\left(\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\right)^2,$$

for every sequence of alternatives $\theta_\nu \downarrow 0$, independently of $\alpha > 0$ and $\gamma \in (\alpha, 1)$. If the power functions of the tests based on $T_{n,i}$ are nondecreasing for every n, then the assumption *of asymptotic normality of $T_{n,i}$ can be relaxed to asymptotic normality under every sequence $\theta_n = O(1/\sqrt{n})$ only.*

Proof. Fix α and γ as in the introduction and, given alternatives $\theta_\nu \downarrow 0$, let $n_{\nu,i}$ observations be used with each of the two tests. The assumption that $\|P_{n,\theta_\nu} - P_{n,0}\| \to 0$ as $\nu \to \infty$ for each fixed n forces $n_{\nu,i} \to \infty$. Indeed, the sum of the probabilities of the first and second kind of the test with critical region K_n equals

$$\int_{K_n} dP_{n,0} + \int_{K_n^c} dP_{n,\theta_\nu} = 1 + \int_{K_n} (p_{n,0} - p_{n,\theta_\nu})\, d\mu_n.$$

This sum is minimized for the critical region $K_n = \{p_{n,0} - p_{n,\theta_\nu} < 0\}$, and then equals $1 - \frac{1}{2}\|P_{n,\theta_\nu} - P_{n,0}\|$. By assumption, this converges to 1 as $\nu \to \infty$ uniformly in every finite set of n. Thus, for every bounded sequence $n = n_\nu$ and any sequence of tests, the sum of the error probabilities is asymptotically bounded below by 1 and cannot be bounded above by $\alpha + 1 - \gamma < 1$, as required.

Now that we have ascertained that $n_{\nu,i} \to \infty$ as $\nu \to \infty$, we can use the asymptotic normality of the test statistics $T_{n,i}$. The convergence to a continuous distribution implies that the asymptotic level and power attained for the minimal numbers of observations (minimal for obtaining at most level α and at least power γ) is exactly α and γ. In order to obtain asymptotic level α the tests must reject H_0 if $\sqrt{n_\nu}\big(T_{n_\nu,i} - \mu_i(0)\big) > \sigma_i(0) z_\alpha + o(1)$. The powers of these tests are equal to

$$\pi_{n_{\nu,i}}(\theta_\nu) = 1 - \Phi\bigg(z_\alpha + o(1) - \sqrt{n_{\nu,i}}\,\theta_\nu \frac{\mu_i'(0)}{\sigma_i(0)}\big(1 + o(1)\big)\bigg) + o(1).$$

This sequence of powers tends to $\gamma < 1$ if and only if the argument of Φ tends to z_γ. Thus the relative efficiency of the two sequences of tests equals

$$\lim_{\nu\to\infty} \frac{n_{\nu,2}}{n_{\nu,1}} = \lim_{\nu\to\infty} \frac{n_{\nu,2}\theta_\nu^2}{n_{\nu,1}\theta_\nu^2} = \frac{(z_\alpha - z_\gamma)^2}{\big(\mu_2'(0)/\sigma_2(0)\big)^2} \Bigg/ \frac{(z_\alpha - z_\gamma)^2}{\big(\mu_1'(0)/\sigma_1(0)\big)^2}.$$

This proves the first assertion of the theorem.

If the power functions of the tests are monotone and the test statistics are asymptotically normal for every sequence $\theta_n = O(1/\sqrt{n})$, then $\pi_{n,i}(\theta_n) \to \alpha$ or 1 if $\sqrt{n}\,\theta_n \to 0$ or ∞, respectively (see Lemma 14.16). In that case the sequences of tests can only meet the (α, γ) requirement for testing alternatives θ_ν such that $\sqrt{n_{\nu,i}}\,\theta_\nu = O(1)$. For such sequences the preceding argument is valid and gives the asserted relative efficiency. ■

*14.4 Other Relative Efficiencies

The asymptotic relative efficiency defined in the preceding section is known as the *Pitman relative efficiency*. In this section we discuss some other types of relative efficiencies. Define $n_i(\alpha, \gamma, \theta)$ as the minimal numbers of observations needed, with $i \in \{1, 2\}$ for two given sequences of tests, to test a null hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta = \theta$ at level α and with power at least γ. Then the Pitman efficiency against a sequence of alternatives $\theta_\nu \to 0$ is defined as (if the limits exists)

$$\lim_{\nu\to\infty} \frac{n_2(\alpha, \gamma, \theta_\nu)}{n_1(\alpha, \gamma, \theta_\nu)}.$$

The device to let the alternatives θ_ν tend to the null hypothesis was introduced to make the testing problems harder and harder, so that the required numbers of observations tend to infinity, and the comparison becomes an asymptotic one. There are other possibilities that can serve the same end. The testing problem is harder as α is smaller, as γ is larger, and (typically) as θ is closer to the null hypothesis. Thus, we could also let α tend to zero, or γ tend to one, keeping the other parameters fixed, or even let two or all three of the parameters vary. For each possible method we could define the relative efficiency of two sequences of tests as the limit of the quotient of the minimal numbers of observations that are needed. Most of these possibilities have been studied in the literature. Next to the Pitman efficiency the most popular efficiency measure appears to be the *Bahadur efficiency*, which is defined as

$$\lim_{\nu\to\infty} \frac{n_2(\alpha_\nu, \gamma, \theta)}{n_1(\alpha_\nu, \gamma, \theta)}.$$

Here α_ν tends to zero, but γ and θ are fixed. Typically, the Bahadur efficiency depends on θ, but not on γ, and not on the particular sequence $\alpha_\nu \downarrow 0$ that is used.

Whereas the calculation of Pitman efficiencies is most often based on distributional limit theorems, Bahadur efficiencies are derived from large deviations results. The reason is that the probabilities of first or second kind for testing a fixed null hypothesis against a fixed alternative usually tend to zero at an exponential speed. Large deviations theorems quantify this speed. Suppose that the null hypothesis $H_0 : \theta = 0$ is rejected for large values of a test statistic T_n, and that

$$-\frac{2}{n} \log \mathrm{P}_0(T_n \geq t) \to e(t), \qquad \text{every } t, \tag{14.20}$$

$$T_n \xrightarrow{P_\theta} \mu(\theta). \tag{14.21}$$

The first result is a *large deviation* type result, and the second a "law of large numbers." The *observed significance level* of the test is defined as $\mathrm{P}_0(T_n \geq t)_{|t=T_n}$. Under the null hypothesis, this random variable is uniformly distributed if T_n possesses a continuous distribution function. For a fixed alternative θ, it typically converges to zero at an exponential rate. For instance, under the preceding conditions, if e is continuous at $\mu(\theta)$, then (because e is necessarily monotone) it is immediate that

$$-\frac{2}{n} \log \mathrm{P}_0(T_n \geq t)_{|t=T_n} \xrightarrow{P_\theta} e\bigl(\mu(\theta)\bigr).$$

The quantity $e\bigl(\mu(\theta)\bigr)$ is called the *Bahadur slope* of the test (or rather the limit in probability of the left side if it exists). The quotient of the slopes of two sequences of test statistics gives the *Bahadur relative efficiency*.

14.22 ***Theorem.*** *Let $T_{n,1}$ and $T_{n,2}$ be sequences of statistics in statistical models ($P_{n,0}$, $P_{n,\theta}$) that satisfy (14.20) and (14.21) for functions e_i and numbers $\mu_i(\theta)$ such that e_i is continuous at $\mu_i(\theta)$. Then the Bahadur relative efficiency of the sequences of tests that reject for large values of $T_{n,i}$ is equal to $e_1\bigl(\mu_1(\theta)\bigr)/e_2\bigl(\mu_2(\theta)\bigr)$, for every $\alpha_\nu \downarrow 0$ and every $1 > \gamma > \sup_n P_{n,\theta}(p_{n,0} = 0)$.*

Proof. For simplicity of notation, we drop the index $i \in \{1, 2\}$ and write n_ν for the minimal numbers of observations needed to obtain level α_ν and power γ with the test statistics T_n.

The sample sizes n_ν necessarily converge to ∞ as $\nu \to \infty$. If not, then there would exist a fixed value n and a (sub)sequence of tests with levels tending to 0 and powers at least γ. However, for any fixed n, and any sequence of measurable sets K_m with $P_{n,0}(K_m) \to 0$ as $m \to \infty$, the probabilities $P_{n,\theta}(K_m) = P_{n,\theta}(K_m \cap p_{n,0} = 0) + o(1)$ are eventually strictly smaller than γ, by assumption.

The most powerful level α_ν-test that rejects for large values of T_n has critical region $\{T_n \geq c_n\}$ or $\{T_n > c_n\}$ for $c_n = \inf\{c : \mathrm{P}_0(T_n \geq c) \leq \alpha_\nu\}$, where we use $\geq$ if $\mathrm{P}_0(T_n \geq c_n) \leq \alpha_\nu$ and $>$ otherwise. Equivalently, with the notation $L_n = \mathrm{P}_0(T_n \geq t)_{|t=T_n}$, this is the test with critical region $\{L_n \leq \alpha_\nu\}$. By the definition of n_ν we conclude that

$$P_{n,\theta}\Bigl(-\frac{2}{n}\log L_n \geq -\frac{2}{n}\log\alpha_\nu\Bigr)\begin{cases}\geq \gamma & \text{for } n = n_\nu,\\ < \gamma & \text{for } n = n_\nu - 1.\end{cases}$$

By (14.20) and (14.21), the random variable inside the probability converges in probability to the number $e\bigl(\mu(\theta)\bigr)$ as $n \to \infty$. Thus, the probability converges to 0 or 1 if $-(2/n)\log\alpha_\nu$ is asymptotically strictly bigger or smaller than $e\bigl(\mu(\theta)\bigr)$, respectively. Conclude that

$$\limsup_{\nu\to\infty} -\frac{2}{n_\nu}\log\alpha_\nu \leq e\bigl(\mu(\theta)\bigr)$$
$$\liminf_{\nu\to\infty} -\frac{2}{n_\nu - 1}\log\alpha_\nu \geq e\bigl(\mu(\theta)\bigr).$$

Combined, this yields the asymptotic equivalence $n_\nu \sim -2\log\alpha_\nu / e\bigl(\mu(\theta)\bigr)$. Applying this for both $n_{\nu,1}$ and $n_{\nu,2}$ and taking the quotient, we obtain the theorem. ∎

Bahadur and Pitman efficiencies do not always yield the same ordering of sequences of tests. In numerical comparisons, the Pitman efficiencies appear to be more relevant for moderate sample sizes. This is explained by their method of calculation. By the preceding theorem, Bahadur efficiencies follow from a large deviations result under the null hypothesis and a law of large numbers under the alternative. A law of large numbers is of less accuracy than a distributional limit result. Furthermore, large deviation results, while mathematically interesting, often yield poor approximations for the probabilities of interest. For instance, condition (14.20) shows that $\mathrm{P}_0(T_n \geq t) = \exp\bigl(-\frac{1}{2}ne(t)\bigr)\exp o(n)$. Nothing guarantees that the term $\exp o(n)$ is close to 1.

On the other hand, often the Bahadur efficiencies as a function of θ are more informative than Pitman efficiencies. The Pitman slopes are obtained under the condition that the sequence $\sqrt{n}\bigl(T_n - \mu(0)\bigr)$ is asymptotically normal with mean zero and variance $\sigma^2(0)$. Suppose, for the present argument, that T_n is normally distributed for every finite n, with the parameters $\mu(0)$ and $\sigma^2(0)/n$. Then, because $1 - \Phi(t) \sim \phi(t)/t$ as $t \to \infty$,

$$-\frac{2}{n}\log\mathrm{P}_0\bigl(T_n \geq \mu(0) + t\bigr) = -\frac{2}{n}\log\biggl(1 - \Phi\Bigl(\frac{t\sqrt{n}}{\sigma(0)}\Bigr)\biggr) \to \frac{t^2}{\sigma^2(0)}, \qquad \text{every } t.$$

The Bahadur slope would be equal to $\bigl(\mu(\theta) - \mu(0)\bigr)^2/\sigma^2(0)$. For $\theta \to 0$, this is approximately equal to θ^2 times the square of the Pitman slope $\mu'(0)^2/\sigma^2(0)$. Consequently, the limit of the Bahadur efficiencies as $\theta \to 0$ would yield the Pitman efficiency.

Now, the preceding argument is completely false if T_n is only approximately normally distributed: Departures from normality that are negligible in the sense of weak convergence need not be so for large-deviation probabilities. The difference between the "approximate

Bahadur slopes" just obtained and the true slopes is often substantial. However, the argument tends to be "more correct" as t approaches $\mu(0)$, and the conclusion that limiting Bahadur efficiencies are equal to Pitman efficiencies is often correct.[†]

The main tool needed to evaluate Bahadur efficiencies is the large-deviation result (14.20). For averages T_n, this follows from the Cramér-Chernoff theorem, which can be thought of as the analogue of the central limit theorem for large deviations. It is a refinement of the weak law of large numbers that yields exponential convergence of probabilities of deviations from the mean.

The *cumulant generating function* of a random variable Y is the function $u \mapsto K(u) = \log \mathrm{E}e^{uY}$. If we allow the value ∞, then this is well-defined for every $u \in \mathbb{R}$. The set of u such that $K(u)$ is finite is an interval that may or may not contain its boundary points and may be just the point $\{0\}$.

14.23 ***Proposition (Cramér-Chernoff theorem).*** *Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with cumulant generating function K. Then, for every t,*

$$\frac{1}{n} \log \mathrm{P}(\overline{Y} \geq t) \to \inf_{u \geq 0} (K(u) - tu).$$

Proof. The cumulant generating function of the variables $Y_i - t$ is equal to $u \mapsto K(u) - ut$. Therefore, we can restrict ourselves to the case $t = 0$. The proof consists of separate upper and lower bounds on the probabilities $\mathrm{P}(\overline{Y} \geq 0)$.

The upper bound is easy and is valid for every n. By Markov's inequality, for every $u \geq 0$,

$$\mathrm{P}(\overline{Y} \geq 0) = \mathrm{P}\big(e^{un\overline{Y}_n} \geq 1\big) \leq \mathrm{E}e^{un\overline{Y}_n} = e^{nK(u)}.$$

Take logarithms, divide by n, and take the infimum over $u \geq 0$ to find one half of the proposition.

For the proof of the lower bound, first consider the cases that Y_i is nonnegative or nonpositive. If $\mathrm{P}(Y_i < 0) = 0$, then the function $u \mapsto K(u)$ is monotonely increasing on $\mathbb{R}$ and its infimum on $u \geq 0$ is equal to 0 (attained at $u = 0$); this is equal to $n^{-1} \log \mathrm{P}(\overline{Y} \geq 0)$ for every n. Second, if $\mathrm{P}(Y_i > 0) = 0$, then the function $u \mapsto K(u)$ is monotonely decreasing on $\mathbb{R}$ with $K(\infty) = \log \mathrm{P}(Y_1 = 0)$; this is equal to $n^{-1} \log \mathrm{P}(\overline{Y} \geq 0)$ for every n. Thus, the theorem is valid in both cases, and we may exclude them from now on.

First, assume that $K(u)$ is finite for every $u \in \mathbb{R}$. Then the function $u \mapsto K(u)$ is analytic on $\mathbb{R}$, and, by differentiating under the expectation, we see that $K'(0) = \mathrm{E}Y_1$. Because Y_i takes both negative and positive values, $K(u) \to \infty$ as $u \to \pm\infty$. Thus, the infimum of the function $u \mapsto K(u)$ over $u \in \mathbb{R}$ is attained at a point u_0 such that $K'(u_0) = 0$.

The case that $u_0 < 0$ is trivial, but requires an argument. By the convexity of the function $u \mapsto K(u)$, K is nondecreasing on $[u_0, \infty)$. If $u_0 < 0$, then it attains its minimum value over $u \geq 0$ at $u = 0$, which is $K(0) = 0$. Furthermore, in this case $\mathrm{E}Y_1 = K'(0) > K'(u_0) = 0$ (strict inequality under our restrictions, for instance because $K''(0) = \operatorname{var} Y_1 > 0$) and hence $\mathrm{P}(\overline{Y} \geq 0) \to 1$ by the law of large numbers. Thus, the limit of the left side of the proposition (with $t = 0$) is 0 as well.

[†] In [85] a precise argument is given.

For $u_0 \geq 0$, let $Z_1, Z_2, \dots$ be i.i.d. random variables with the distribution given by

$$dP_Z(z) = e^{-K(u_0)}\, e^{u_0 z}\, dP_Y(z).$$

Then Z_1 has cumulant generating function $u \mapsto K(u_0 + u) - K(u_0)$, and, as before, its mean can be found by differentiating this function at $u = 0$: $\mathrm{E}Z_1 = K'(u_0) = 0$. For every $\varepsilon > 0$,

$$\begin{aligned} \mathrm{P}(\bar{Y} \geq 0) &= \mathrm{E}1\{\bar{Z}_n \geq 0\} e^{-u_0 n \bar{Z}_n}\, e^{nK(u_0)} \\ &\geq \mathrm{P}(0 \leq \bar{Z}_n \leq \varepsilon)\, e^{-u_0 n\varepsilon}\, e^{nK(u_0)}. \end{aligned}$$

Because $\bar{Z}_n$ has mean 0, the sequence $\mathrm{P}(0 \leq \bar{Z}_n \leq \varepsilon)$ is bounded away from 0, by the central limit theorem. Conclude that n^{-1} times the limit inferior of the logarithm of the left side is bounded below by $-u_0\varepsilon + K(u_0)$. This is true for every $\varepsilon > 0$ and hence also for $\varepsilon = 0$.

Finally, we remove the restriction that $K(u)$ is finite for every u, by a truncation argument. For a fixed, large M, let $Y_1^M, Y_2^M, \dots$ be distributed as the variables $Y_1, Y_2, \dots$ given that $|Y_i| \leq M$ for every i, that is, they are i.i.d. according to the conditional distribution of Y_1 given $|Y_1| \leq M$. Then, with $u \mapsto K_M(u) = \log \mathrm{E}e^{uY_1}1\{|Y_1| \leq M\}$,

$$\begin{aligned} \liminf \frac{1}{n} \log \mathrm{P}\big(\bar{Y} \geq 0\big) &\geq \frac{1}{n} \log\Big(\mathrm{P}\big(\bar{Y}_n^M \geq 0\big)\mathrm{P}\big(|Y_i^M| \leq M\big)^n\Big) \\ &\geq \inf_{u \geq 0} K_M(u), \end{aligned}$$

by the preceding argument applied to the truncated variables. Let s be the limit of the right side as $M \to \infty$, and let A_M be the set $\{u \geq 0 : K_M(u) \leq s\}$. Then the sets A_M are nonempty and compact for sufficiently large M (as soon as $K_M(u) \to \infty$ as $u \to \pm\infty$), with $A_1 \supset A_2 \supset \cdots$, whence $\cap A_M$ is nonempty as well. Because K_M converges pointwise to K as $M \to \infty$, any point $u_1 \in \cap A_M$ satisfies $K(u_1) = \lim K_M(u_1) \leq s$. Conclude that s is bigger than the right side of the proposition (with $t = 0$). ∎

14.24 ***Example (Sign statistic).*** The cumulant generating function of a variable Y that is -1 and 1, each with probability $\frac{1}{2}$, is equal to $K(u) = \log\cosh u$. Its derivative is $K'(u) = \tanh u$ and hence the infimum of $K(u) - tu$ over $u \in \mathbb{R}$ is attained for $u = \operatorname{arctanh} t$. By the Cramér-Chernoff theorem, for $0 < t < 1$,

$$-\frac{2}{n} \log \mathrm{P}(\bar{Y} \geq t) \to e(t) := -2 \log\cosh \operatorname{arctanh} t + 2t \operatorname{arctanh} t.$$

We can apply this result to find the Bahadur slope of the sign statistic $T_n = n^{-1}\sum_{i=1}^n \operatorname{sign}(X_i)$. If the null distribution of the random variables $X_1, \dots, X_n$ is continuous and symmetric about zero, then (14.20) is valid with $e(t)$ as in the preceding display and with $\mu(\theta) = \mathrm{E}_\theta \operatorname{sign}(X_1)$. Figure 14.2 shows the slopes of the sign statistic and the sample mean for testing the location of the Laplace distribution. The local optimality of the sign statistic is reflected in the Bahadur slopes, but for detecting large differences of location the mean is better than the sign statistic. However, it should be noted that the power of the sign test in this range is so close to 1 that improvement may be irrelevant; for example, the power is 0.999 at level 0.007 for $n = 25$ at $\theta = 2$. □

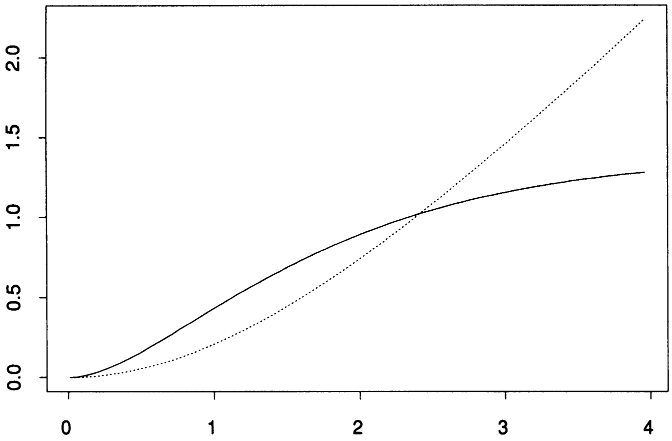

Figure 14.2. Bahadur slopes of the sign statistic (*solid line*) and the sample mean (*dotted line*) for testing that a random sample from the Laplace distribution has mean zero versus the alternative that the mean is θ, as a function of θ.

14.25 ***Example (Student statistic).*** Suppose that $X_1, \dots, X_n$ are a random sample from a normal distribution with mean μ and variance σ^2. We shall consider σ known and compare the slopes of the sample mean and the Student statistic $\overline{X}_n/S_n$ for testing $H_0: \mu = 0$.

The cumulant generating function of the normal distribution is equal to $K(u) = u\mu + \frac{1}{2}u^2\sigma^2$. By the Cramér-Chernoff theorem, for $t > 0$,

$$-\frac{2}{n}\log \mathrm{P}_0(\overline{X}_n \geq t) \to e(t) := \frac{t^2}{\sigma^2}.$$

Thus, the Bahadur slope of the sample mean is equal to μ^2/σ^2, for every $\mu > 0$.

Under the null hypothesis, the statistic $\sqrt{n}\,\overline{X}_n/S_n$ possesses the t-distribution with $(n-1)$ degrees of freedom. Thus, for a random sample $Z_0, Z_1, \dots$ of standard normal variables, for every $t > 0$,

$$\mathrm{P}_0\left(\sqrt{\frac{n}{n-1}}\frac{\overline{X}_n}{S_n} \geq t\right) = \frac{1}{2}\mathrm{P}\left(\frac{t_{n-1}^2}{n-1} \geq t^2\right) = \frac{1}{2}\mathrm{P}\left(Z_0^2 - t^2\sum_{i=1}^{n-1} Z_i^2 \geq 0\right).$$

This probability is not of the same form as in the Cramér-Chernoff theorem, but it concerns almost an average, and we can obtain the large deviation probabilities from the cumulant generating function in an analogous way. The cumulant generating function of a square of a standard normal variable is equal to $u \mapsto -\frac{1}{2}\log(1-2u)$, and hence the cumulant generating function of the variable $Z_0^2 - t^2\sum_{i=1}^{n-1} Z_i^2$ is equal to

$$K_n(u) = -\tfrac{1}{2}\log(1-2u) - \tfrac{1}{2}(n-1)\log(1+2t^2u).$$

This function is nicely differentiable and, by straightforward calculus, its minimum value can be found to be

$$\inf_u K_n(u) = -\frac{1}{2}\log\left(\frac{t^2+1}{t^2 n}\right) - \frac{1}{2}(n-1)\log\left(\frac{(n-1)(t^2+1)}{n}\right).$$

The minimum is achieved on $[0, \infty)$ for $t^2 \geq (n-1)^{-1}$. This expression divided by n is the analogue of $\inf_u K(u)$ in the Cramér-Chernoff theorem. By an extension of this theorem, for every $t > 0$,

$$-\frac{2}{n}\log \mathrm{P}_0\left(\sqrt{\frac{n}{n-1}}\frac{\overline{X}_n}{S_n} \geq t\right) \to e(t) = \log(t^2+1).$$

Thus, the Bahadur slope of the Student statistic is equal to $\log(1+\mu^2/\sigma^2)$.

For μ/σ close to zero, the Bahadur slopes of the sample mean and the Student statistic are close, but for large μ/σ the slope of the sample mean is much bigger. This suggests that the loss in efficiency incurred by unnecessarily estimating the standard deviation σ can be substantial. This suggestion appears to be unrealistic and also contradicts the fact that the Pitman efficiencies of the two sequences of statistics are equal. □

14.26 ***Example (Neyman-Pearson statistic).*** The sequence of Neyman-Pearson statistics $\prod_{i=1}^n (p_\theta/p_{\theta_0})(X_i)$ has Bahadur slope $-2P_\theta \log(p_{\theta_0}/p_\theta)$. This is twice the Kullback-Leibler divergence of the measures P_{θ_0} and P_θ and shows an important connection between large deviations and the Kullback-Leibler divergence.

In regular cases this result is a consequence of the Cramér-Chernoff theorem. The variable $Y = \log p_\theta/p_{\theta_0}$ has cumulant generating function $K(u) = \log\int p_\theta^u p_{\theta_0}^{1-u}\,d\mu$ under P_{θ_0}. The function $K(u)$ is finite for $0 \leq u \leq 1$, and, at least by formal calculus, $K'(1) = P_\theta \log(p_\theta/p_{\theta_0}) = \mu(\theta)$, where $\mu(\theta)$ is the asymptotic mean of the sequence $n^{-1}\sum \log(p_\theta/p_{\theta_0})(X_i)$. Thus the infimum of the function $u \mapsto K(u) - u\mu(\theta)$ is attained at $u = 1$ and the Bahadur slope is given by

$$e\big(\mu(\theta)\big) = -2\big(K(1) - \mu(\theta)\big) = 2P_\theta \log\frac{p_\theta}{p_{\theta_0}}.$$

In section 16.6 we obtain this result by a direct, and rigorous, argument. □

For statistics that are not means, the Cramér-Chernoff theorem is not applicable, and we need other methods to compute the Bahadur efficiencies. An important approach applies to functions of means and is based on more general versions of Cramér's theorem. A first generalization asserts that, for certain sets B, not necessarily of the form $[t,\infty)$,

$$\frac{1}{n}\log \mathrm{P}(\overline{Y} \in B) \to -\inf\nolimits_{y\in B} I(y), \qquad I(y) = \sup_u\big(uy - K(u)\big).$$

For a given statistic of the form $\phi(\overline{Y})$, the large deviation probabilities of interest $\mathrm{P}(\phi(\overline{Y}) \geq t)$ can be written in the form $\mathrm{P}(\overline{Y} \in B_t)$ for the inverse images $B_t = \phi^{-1}[t,\infty)$. If B_t is an eligible set in the preceding display, then the desired large deviations result follows, although we shall still have to evaluate the repeated "inf sup" on the right side. Now, according to Cramér's theorem, the display is valid for every set such that the right side does not change if B is replaced by its interior or its closure. In particular, if ϕ is continuous, then B_t is closed and its interior $\overset{\circ}{B}_t$ contains the set $\phi^{-1}(t,\infty)$. Then we obtain a large deviations result if the difference set $\phi^{-1}\{t\}$ is "small" in that it does not play a role when evaluating the right side of the display.

Transforming a univariate mean $\overline{Y}$ into a statistic $\phi(\overline{Y})$ can be of interest (for example, to study the two-sided test statistics $|\overline{Y}|$), but the real promise of this approach is in its applications to multivariate and infinite-dimensional means. Cramér's theorem has been generalized to these situations. General large deviation theorems can best be formulated as separate upper and lower bounds. A sequence of random maps $X_n : \Omega \mapsto \mathbb{D}$ from a probability space $(\Omega, \mathcal{U}, \mathrm{P})$ into a topological space $\mathbb{D}$ is said to satisfy the *large deviation principle with rate function I* if, for every closed set F and for every open set G,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathrm{P}^*(X_n \in F) \leq - \inf_{y\in F} I(y),$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathrm{P}_*(X_n \in G) \geq - \inf_{y\in G} I(y).$$

The rate function $I : \mathbb{D} \mapsto [0, \infty]$ is assumed to be lower semicontinuous and is called a *good rate function* if the sublevel sets $\{y : I(y) \leq M\}$ are compact, for every $M \in \mathbb{R}$. The inner and outer probabilities that X_n belongs to a general set B is sandwiched between the probabilities that it belongs to the interior $\mathring{B}$ and the closure $\overline{B}$. Thus, we obtain a large deviation result with equality for every set B such that $\inf\{I(y) : y \in \overline{B}\} = \inf\{I(y) : y \in \mathring{B}\}$. An implication for the slopes of test statistics of the form $\phi(X_n)$ is as follows.

14.27 ***Lemma.*** *Suppose that $\phi : \mathbb{D} \mapsto \mathbb{R}$ is continuous at every y such that $I(y) < \infty$ and suppose that $\inf\{I(y) : \phi(y) > t\} = \inf\{I(y) : \phi(y) \geq t\}$. If the sequence X_n satisfies the large-deviation principle with the rate function I under P_0, then $T_n = \phi(X_n)$ satisfies (14.20) with $e(t) = 2\inf\{I(y) : \phi(y) \geq t\}$. Furthermore, if I is a good rate function, then e is continuous at t.*

Proof. Define sets $A_t = \phi^{-1}(t, \infty)$ and $B_t = \phi^{-1}[t, \infty)$, and let $\mathbb{D}_0$ be the set where I is finite. By the continuity of ϕ, $\overline{B}_t \cap \mathbb{D}_0 = B_t \cap \mathbb{D}_0$ and $\mathring{B}_t \cap \mathbb{D}_0 \supset A_t \cap \mathbb{D}_0$. (If $y \notin \mathring{B}_t$, then there is a net $y_n \in B_t^c$ with $y_n \to y$; if also $y \in \mathbb{D}_0$, then $\phi(y) = \lim \phi(y_n) \leq t$ and hence $y \notin A_t$.) Consequently, the infimum of I over $\mathring{B}_t$ is at least the infimum over A_t, which is the infimum over B_t by assumption, and also the infimum over $\overline{B}_t$. Condition (14.20) follows upon applying the large deviation principle to $\mathring{B}_t$ and $\overline{B}_t$.

The function e is nondecreasing. The condition on the pair (I, ϕ) is exactly that e is right-continuous, because $e(t+) = \inf\{I(y) : \phi(y) > t\}$. To prove the left-continuity of e, let $t_m \uparrow t$. Then $e(t_m) \uparrow a$ for some $a \leq e(t)$. If $a = \infty$, then $e(t) = \infty$ and e is left-continuous. If $a < \infty$, then there exists a sequence y_m with $\phi(y_m) \geq t_m$ and $2I(y_m) \leq a + 1/m$. By the goodness of I, this sequence has a converging subnet $y_{m'} \to y$. Then $2I(y) \leq \liminf 2I(y_{m'}) \leq a$ by the lower semicontinuity of I, and $\phi(y) \geq t$ by the continuity of ϕ. Thus $e(t) \leq 2I(y) \leq a$. ∎

Empirical distributions can be viewed as means (of Dirac measures), and are therefore potential candidates for a large-deviation theorem. Cramér's theorem for empirical distributions is known as *Sanov's theorem.* Let $\mathbb{L}_1(\mathcal{X}, \mathcal{A})$ be the set of all probability measures on the measurable space $(\mathcal{X}, \mathcal{A})$, which we assume to be a complete, separable metric space with its Borel σ-field. The *τ-topology* on $\mathbb{L}_1(\mathcal{X}, \mathcal{A})$ is defined as the weak topology generated by the collection of all maps $P \mapsto Pf$ for f ranging over the set of all bounded, measurable functions on $f : \mathcal{X} \mapsto \mathbb{R}$.[†]

14.28 ***Theorem (Sanov's theorem).*** *Let $\mathbb{P}_n$ be the empirical measure of a random sample of size n from a fixed measure P. Then the sequence $\mathbb{P}_n$ viewed as maps into $\mathbb{L}_1(\mathcal{X}, \mathcal{A})$*

[†] For a proof of the following theorem, see [31], [32], or [65].

satisfies the large deviation principle relative to the τ-topology, with the good rate function $I(Q) = -Q\log p/q$.

For $\mathcal{X}$ equal to the real line, $L_1(\mathcal{X}, \mathcal{A})$ can be identified with the set of cumulative distribution functions. The τ-topology is stronger than the topology obtained from the uniform norm on the distribution functions. This follows from the fact that if both $F_n(x) \to F(x)$ and $F_n\{x\} \to F\{x\}$ for every $x \in \mathbb{R}$, then $\|F_n - F\|_\infty \to 0$. (see problem 19.9). Thus any function ϕ that is continuous with respect to the uniform norm is also continuous with respect to the τ-topology, and we obtain a large collection of functions to which we can apply the preceding lemma. Trimmed means are just one example.

14.29 ***Example (Trimmed means).*** Let $\mathbb{F}_n$ be the empirical distribution function of a random sample of size n from the distribution function F, and let $\mathbb{F}_n^{-1}$ be the corresponding quantile function. The function $\phi(\mathbb{F}_n) = (1-2\alpha)^{-1}\int_\alpha^{1-\alpha}\mathbb{F}_n^{-1}(s)\,ds$ yields a version of the α-trimmed mean (see Chapter 22). We assume that $0 < \alpha < \frac{1}{2}$ and (partly for simplicity) that the null distribution F_0 is continuous.

If we show that the conditions of Lemma 14.27 are fulfilled, then we can conclude, by Sanov's theorem,

$$-\frac{2}{n}\log \mathrm{P}_{F_0}\big(\phi(\mathbb{F}_n) \geq t\big) \to e(t) := 2\inf_{G:\phi(G)\geq t} -G\log\frac{f_0}{g}.$$

Because $\mathbb{F}_n \xrightarrow{\mathrm{P}} F$ uniformly by the Glivenko-Cantelli theorem, Theorem 19.1, and ϕz is continuous, $\phi(\mathbb{F}_n) \xrightarrow{\mathrm{P}} \phi(F)$, and the Bahadur slope of the α-trimmed mean at an alternative F is equal to $e\big(\phi(F)\big)$.

Finally, we show that ϕ is continuous with respect to the uniform topology and that the function $t \mapsto \inf\big\{-G\log(f_0/g)) : \phi(G) \geq t\big\}$ is right-continuous at t if F_0 is continuous at t. The map ϕ is even continuous with respect to the weak topology on the set of distribution functions: If a sequence of measures G_m converges weakly to a measure G, then the corresponding quantile functions G_m^{-1} converge weakly to the quantile function G^{-1} (see Lemma 21.2) and hence $\phi(G_m) \to \phi(G)$ by the dominated convergence theorem.

The function $t \mapsto \inf\big\{-G\log(f_0/g) : \phi(G) \geq t\big\}$ is right-continuous at t if for every G with $\phi(G) = t$ there exists a sequence G_m with $\phi(G_m) > t$ and $G_m\log(f_0/g_m) \to G\log(f_0/g)$. If $G\log(f_0/g) = -\infty$, then this is easy, for we can choose any fixed G_m that is singular with respect to F_0 and has a trimmed mean bigger than t. Thus, we may assume that $G\big|\log(f_0/g)\big| < \infty$, that $G \ll F_0$ and hence that G is continuous. Then there exists a point c such that $\alpha < G(c) < 1-\alpha$. Define

$$\frac{dG_m}{dG}(x) = \begin{cases} 1 - \frac{1}{m} & \text{if } x \leq c, \\ 1 + \varepsilon_m & \text{if } x > c. \end{cases}$$

Then G_m is a probability distribution for suitably chosen $\varepsilon_m > 0$, and, by the dominated convergence $G_m\log(f_0/g_m) \to G\log(f_0/g)$ as $m \to \infty$. Because $G_m(x) \leq G(x)$ for all x, with strict inequality (at least) for all $x \leq c$ such that $G(x) > 0$, we have that $G_m^{-1}(s) \geq G^{-1}(s)$ for all s, with strict inequality for all $s \in (0, G(c)]$. Hence the trimmed mean $\phi(G_m)$ is strictly bigger than the trimmed mean $\phi(G)$, for every m. □

*14.5 Rescaling Rates

The asymptotic power functions considered earlier in this chapter are the limits of "local power functions" of the form $h \mapsto \pi_n(h/\sqrt{n})$. The rescaling rate $\sqrt{n}$ is typical for testing smooth parameters of the model. In this section we have a closer look at the rescaling rate and discuss some nonregular situations.

Suppose that in a given sequence of models $(\mathcal{X}_n, \mathcal{A}_n, P_{n,\theta} : \theta \in \Theta)$ it is desired to test the null hypothesis $H_0 : \theta = \theta_0$ versus the alternatives $H_1 : \theta = \theta_n$. For probability measures P and Q define the *total variation distance* $\|P - Q\|$ as the L_1-distance $\int |p - q|\, d\mu$ between two densities of P and Q.

14.30 ***Lemma.*** *The power function π_n of any test in $(\mathcal{X}_n, \mathcal{A}_n, P_{n,\theta} : \theta \in \Theta)$ satisfies*

$$\pi_n(\theta) - \pi_n(\theta_0) \le \tfrac{1}{2} \|P_{n,\theta} - P_{n,\theta_0}\|.$$

For any θ and θ_0 there exists a test whose power function attains equality.

Proof. If π_n is the power function of the test ϕ_n, then the difference on the left side can be written as $\int \phi_n (p_{n,\theta} - p_{n,\theta_0})\, d\mu_n$. This expression is maximized for the test function $\phi_n = 1\{p_{n,\theta} > p_{n,\theta_0}\}$. Next, for any pair of probability densities p and q we have $\int_{q>p} (q - p)\, d\mu = \frac{1}{2} \int |p - q|\, d\mu$, since $\int (p - q)\, d\mu = 0$. ■

This lemma implies that for any sequence of alternatives θ_n:

(i) If $\|P_{n,\theta_n} - P_{n,\theta_0}\| \to 2$, then there exists a sequence of tests with power $\pi_n(\theta_n)$ tending to 1 and size $\pi_n(\theta_0)$ tending to 0 (a *perfect* sequence of tests).

(ii) If $\|P_{n,\theta_n} - P_{n,\theta_0}\| \to 0$, then the power of any sequence of tests is asymptotically less than the level (every sequence of tests is worthless).

(iii) If $\|P_{n,\theta_n} - P_{n,\theta_0}\|$ is bounded away from 0 and 2, then there exists no perfect sequence of tests, but not every test is worthless.

The rescaling rate $h/\sqrt{n}$ used earlier sections corresponds to the third possibility. These examples concern models with independent observations. Because the total variation distance between product measures cannot be easily expressed in the distances for the individual factors, we translate the results into the Hellinger distance and next study the implications for product experiments.

The *Hellinger distance* $H(P, Q)$ between two probability measures is the L_2-distance between the square roots of the corresponding densities. Thus, its square $H^2(P, Q)$ is equal to $\int (\sqrt{p} - \sqrt{q})^2\, d\mu$. The distance is convenient if considering product measures. First, the Hellinger distance can be expressed in the *Hellinger affinity* $A(P, Q) = \int \sqrt{p}\sqrt{q}\, d\mu$, through the formula

$$H^2(P, Q) = 2 - 2A(P, Q).$$

Next, by Fubini's theorem, the affinity of two product measures is the product of the affinities. Thus we arrive at the formula

$$H^2(P^n, Q^n) = 2 - 2\big(1 - \tfrac{1}{2} H^2(P, Q)\big)^n.$$

14.31 ***Lemma.*** *Given a statistical model $(P_\theta : \theta \geq \theta_0)$ set $P_{n,\theta} = P_\theta^n$. Then the possibilities (i), (ii), and (iii) arise when $nH^2(P_{\theta_n}, P_{\theta_0})$ converges to ∞, converges to 0, or is bounded away from 0 and ∞, respectively. In particular, if $H^2(P_\theta, P_{\theta_0}) = O\big(|\theta - \theta_0|^\alpha\big)$ as $\theta \to \theta_0$, then the possibilities (i), (ii), and (iii) are valid when $n^{1/\alpha}|\theta_n - \theta_0|$ converges to ∞, converges to 0, or is bounded away from 0 and ∞, respectively.*

Proof. The possibilities (i), (ii), and (iii) can equivalently be described by replacing the total variation distance $\|P_{\theta_n}^n - P_{\theta_0}^n\|$ by the squared Hellinger distance $H^2(P_{\theta_n}^n, P_{\theta_0}^n)$. This follows from the inequalities, for any probability measures P and Q,

$$H^2(P, Q) \leq \|P - Q\| \leq \big(2 - A^2(P, Q)\big) \wedge 2H(P, Q).$$

The inequality on the left is immediate from the inequality $|\sqrt{p} - \sqrt{q}|^2 \leq |p - q|$, valid for any nonnegative numbers p and q. For the inequality on the right, first note that $pq = (p \vee q)(p \wedge q) \leq (p + q)(p \wedge q)$, whence $A^2(P, Q) \leq 2\int (p \wedge q)\, d\mu$, by the Cauchy-Schwarz inequality. Now $\int (p \wedge q)\, d\mu$ is equal to $1 - \frac{1}{2}\|P - Q\|$, as can be seen by splitting the domains of both integrals in the sets $p < q$ and $p \geq q$. This shows that $\|P - Q\| \leq 2 - A^2(P, Q)$. That $\|P - Q\| \leq 2H(P, Q)$ is a direct consequence of the Cauchy-Schwarz inequality.

We now express the Hellinger distance of the product measures in the Hellinger distance of P_{θ_n} and P_{θ_0} and manipulate the nth power function to conclude the proof. ■

14.32 ***Example (Smooth models).*** If the model $(\mathcal{X}, \mathcal{A}, P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at θ_0, then $H^2(P_\theta, P_{\theta_0}) = O\big(|\theta - \theta_0|^2\big)$. The intermediate rate of convergence (case (iii)) is $\sqrt{n}$. □

14.33 ***Example (Uniform law).*** If P_θ is the uniform measure on $[0, \theta]$, then $H^2(P_\theta, P_{\theta_0}) = O\big(|\theta - \theta_0|\big)$. The intermediate rate of convergence is n. In this case we would study asymptotic power functions defined as the limits of the local power functions of the form $h \mapsto \pi_n(\theta_0 + h/n)$. For instance, the level α tests that reject the null hypothesis $H_0 : \theta = \theta_0$ for large values of the maximum $X_{(n)}$ of the observations have power functions

$$\pi_n\left(\theta_0 + \frac{h}{n}\right) = \mathrm{P}_{\theta_0 + h/n}\big(X_{(n)} \geq \theta_0(1 - \alpha)^{1/n}\big) \to 1 - (1 - \alpha)e^{-h/\theta_0}.$$

Relative to this rescaling rate, the level α tests that reject the null hypothesis for large values of the mean $\bar{X}_n$ have asymptotic power function α (no power). □

14.34 ***Example (Triangular law).*** Let P_θ be the probability distribution with density $x \mapsto \big(1 - |x - \theta|\big)^+$ on the real line. Some clever integrations show that $H^2(P_\theta, P_0) = \frac{1}{2}\theta^2 \log(1/\theta) + O(\theta^2)$ as $\theta \to 0$. (It appears easiest to compute the affinity first.) This leads to the intermediate rate of convergence $\sqrt{n \log n}$. □

The preceding lemmas concern testing a given simple null hypothesis against a simple alternative hypothesis. In many cases the rate obtained from considering simple hypotheses does not depend on the hypotheses and is also globally attainable at every parameter in the parameter space. If not, then the global problems have to be taken into account from the beginning. One possibility is discussed within the context of density estimation in section 24.3.

Lemma 14.31 gives rescaling rates for problems with independent observations. In models with dependent observations quite different rates may pertain.

14.35 ***Example (Branching).*** Consider the Galton-Watson branching process, discussed in Example 9.10. If the offspring distribution has mean $\mu(\theta)$ larger than 1, then the parameter is estimable at the exponential rate $\mu(\theta)^n$. This is also the right rescaling rate for defining asymptotic power functions. □

Notes

Apparently, E.J.G. Pitman introduced the efficiencies that are named for him in an unpublished set of lecture notes in 1949. A published proof of a slightly more general result can be found in [109].

Cramér [26] was interested in preciser approximations to probabilities of large deviations than are presented in this chapter and obtained the theorem under the condition that the moment-generating function is finite on $\mathbb{R}$. Chernoff [20] proved the theorem as presented here, by a different argument. Chernoff used it to study the minimum weighted sums of error probabilities of tests that reject for large values of a mean and showed that, for any $0 < \pi < 1$,

$$\frac{1}{n}\log\inf_t\bigl(\pi \mathrm{P}_0(\overline{Y} > t) + (1-\pi)\mathrm{P}_1(\overline{Y} \le t)\bigr)$$
$$\to \inf_{\mathrm{E}_0 Y_1 < t < \mathrm{E}_1 Y_1} \inf_u\bigl(K_0(u) - ut\bigr) \vee \inf_u\bigl(K_1(u) - ut\bigr).$$

Furthermore, for $\overline{Y}$ the likelihood ratio statistic for testing P_0 versus P_1, the right side of this display can be expressed in the *Hellinger integral* of the experiment (P_0, P_1) as

$$\inf_{0<u<1} \log \int dP_0^u \, dP_1^{1-u}.$$

Thus, this expression is a lower bound for the $\liminf_{n\to\infty} n^{-1}\log(\alpha_n + \beta_n)$ for α_n and β_n the error probabilities of any test of P_0 versus P_1. That the Bahadur slope of Neyman-Pearson tests is twice the Kullback-Leibler divergence (Example 14.26) is essentially known as *Stein's lemma* and is apparently among those results by Stein that he never cared to publish.

A first version of Sanov's theorem was proved by Sanov in 1957. Subsequently, many authors contributed to strengthening the result, the version presented here being given in [65]. Large-deviation theorems are subject of current research by probabilists, particularly with extensions to more complicated objects than sums of independent variables. See [31] and [32]. For further information and references concerning applications in statistics, we refer to [4] and [61], as well as to Chapters 8, 16, and 17.

For applications and extensions of the results on rescaling rates, see [37].

PROBLEMS

1. Show that the power function of the Wilcoxon two sample test is monotone under shift of location.
2. Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$-distribution, where σ^2 is known. A test for $H_0: \mu = 0$ against $H_1: \mu > 0$ can be based on either $\overline{X}/\sigma$ or $\overline{X}/S$. Show that the asymptotic

relative efficiency of the two sequences of tests is 1. Does it make a difference whether normal or t-critical values are used?

3. Let $X_1, \ldots, X_n$ be a random sample from a density $f(x-\theta)$ where f is symmetric about zero. Calculate the relative efficiency of the t-test and the test that rejects for large values of $\sum\sum_{i<j} 1\{X_i + X_j > 0\}$ for f equal to the logistic, normal, Laplace, and uniform shapes.

4. Calculate the relative efficiency of the van der Waerden test with respect to the t-test in the two-sample problem.

5. Calculate the relative efficiency of the tests based on Kendall's τ and the sample correlation coefficient to test independence for bivariate normal pairs of observations.

6. Suppose $\phi : \mathcal{F} \mapsto \mathbb{R}$ and $\psi : \mathcal{F} \mapsto \mathbb{R}^k$ are arbitrary maps on an arbitrary set $\mathcal{F}$ and we wish to find the minimum value of ϕ over the set $\{f \in \mathcal{F} : \psi(f) = 0\}$. If the map $f \mapsto \phi(f) + a^T\psi(f)$ attains its minimum over $\mathcal{F}$ at f_a, for each fixed a in an arbitrary set A, and there exists $a_0 \in A$ such that $\psi(f_{a_0}) = 0$, then the desired minimum value is $\phi(f_{a_0})$. This is a rather trivial use of Lagrange multipliers, but it is helpful to solve the next problems. ($\phi(f_{a_0}) = \phi(f_{a_0}) + a_0^T\psi(f_{a_0})$ is the minimum of $\phi(f) + a_0^T\psi(f)$ over $\mathcal{F}$ and hence smaller than the minimum of $\phi(f) + a_0^T\psi(f)$ over $\{f \in \mathcal{F} : \psi(f) = 0\}$.)

7. Show that $4f(0)^2 \int y^2 f(y)\,dy \geq 1/3$ for every probability density f that has its mode at 0. (The minimum is equal to the minimum of $4\int y^2 f(y)\,dy$ over all probability densities f that are bounded by 1.)

8. Show that $12\bigl(\int f^2(y)\,dy\bigr)^2 \int y^2 f(y)\,dy \geq 108/125$ for every probability density f with mean zero. (The minimum is equal to 12 times the minimum of the square of $\phi(f) = \int f^2(y)\,dy$ over all probability densities with mean 0 and variance 1.)

9. Study the asymptotic power function of the sign test if the observations are a sample from a distribution that has a positive mass at its median. Is it good or bad to have a nonsmooth distribution?

10. Calculate the Hellinger and total variation distance between two uniform $U[0,\theta]$ measures.

11. Calculate the Hellinger and total variation distance between two normal $N(\mu,\sigma^2)$ measures.

12. Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on $[-\theta, \theta]$.

(i) Calculate the asymptotic power functions of the tests that reject $H_0 : \theta = \theta_0$ for large values of $X_{(n)}$, $X_{(n)} \vee (-X_{(1)})$ and $X_{(n)} - X_{(1)}$.

(ii) Calculate the asymptotic relative efficiencies of these tests.

13. If two sequences of test statistics satisfied (14.4) for every $\theta_n \downarrow 0$, but with norming rate n^α instead of $\sqrt{n}$, how would Theorem 14.19 have to be modified to find the Pitman relative efficiency?