Chapter 10

Asymptotic Evaluations

"I know, my dear Watson, that you share my love of all that is bizarre and outside the conventions and humdrum routine of everyday life."

Sherlock Holmes

The Red-headed League

All of the criteria we have considered thus far have been finite-sample criteria. In contrast, we might consider asymptotic properties, properties describing the behavior of a procedure as the sample size becomes infinite. In this section we will look at some of such properties and consider point estimation, hypothesis testing, and interval estimation separately. We will place particular emphasis on the asymptotics of maximum likelihood procedures.

The power of asymptotic evaluations is that, when we let the sample size become infinite, calculations simplify. Evaluations that were impossible in the finite-sample case become routine. This simplification also allows us to examine some other techniques (such as bootstrap and M-estimation) that typically can be evaluated only asymptotically.

Letting the sample size increase without bound (sometimes referred to as "asymptopia") should not be ridiculed as merely a fanciful exercise. Rather, asymptotics uncover the most fundamental properties of a procedure and give us a very powerful and general evaluation tool.

10.1 Point Estimation

10.1.1 Consistency

The property of consistency seems to be quite a fundamental one, requiring that the estimator converges to the "correct" value as the sample size becomes infinite. It is such a fundamental property that the worth of an inconsistent estimator should be questioned (or at least vigorously investigated).

Consistency (as well as all asymptotic properties) concerns a sequence of estimators rather than a single estimator, although it is common to speak of a "consistent estimator." If we observe $X_1, X_2, \ldots$ according to a distribution $f(x|\theta)$, we can construct a sequence of estimators $W_n = W_n(X_1, \ldots, X_n)$ merely by performing the same estimation procedure for each sample size n. For example, $\bar{X}_1 = X_1, \bar{X}_2 = (X_1 + X_2)/2$, $\bar{X}_3 = (X_1 + X_2 + X_3)/3$, etc. We can now define a consistent sequence.

Definition 10.1.1 A sequence of estimators $W_n = W_n(X_1, \ldots, X_n)$ is a *consistent sequence of estimators* of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| < \epsilon) = 1. \tag{10.1.1}$$

Informally, (10.1.1) says that as the sample size becomes infinite (and the sample information becomes better and better), the estimator will be arbitrarily close to the parameter with high probability, an eminently desirable property. Or, turning things around, we can say that the probability that a consistent sequence of estimators misses the true parameter is small. An equivalent statement to (10.1.1) is this: For every $\epsilon > 0$ and every $\theta \in \Theta$, a consistent sequence W_n will satisfy

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| \geq \epsilon) = 0. \tag{10.1.2}$$

Definition 10.1.1 should be compared to Definition 5.5.1, the definition of convergence in probability. Definition 10.1.1 says that a consistent sequence of estimators converges in probability to the parameter θ it is estimating. Whereas Definition 5.5.1 dealt with one sequence of random variables with one probability structure, Definition 10.1.1 deals with an entire family of probability structures, indexed by θ. For each different value of θ, the probability structure associated with the sequence W_n is different. And the definition says that for each value of θ, the probability structure is such that the sequence converges in probability to the true θ. This is the usual difference between a probability definition and a statistics definition. The probability definition deals with one probability structure, but the statistics definition deals with an entire family.

Example 10.1.2 (Consistency of $\bar{X}$) Let $X_1, X_2, \ldots$ be iid $\mathrm{n}(\theta, 1)$, and consider the sequence

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Recall that $\bar{X}_n \sim \mathrm{n}(\theta, 1/n)$, so

$$\begin{aligned}
P_\theta(|\bar{X}_n - \theta| < \epsilon) &= \int_{\theta-\epsilon}^{\theta+\epsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-(n/2)(\bar{x}_n-\theta)^2}\, d\bar{x}_n && \text{(definition)}\\
&= \int_{-\epsilon}^{\epsilon} \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-(n/2)y^2}\, dy && (\text{substitute } y = \bar{x}_n - \theta)\\
&= \int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}} \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} e^{-(1/2)t^2}\, dt && (\text{substitute } t = y\sqrt{n})\\
&= P(-\epsilon\sqrt{n} < Z < \epsilon\sqrt{n}) && (Z \sim \mathrm{n}(0,1))\\
&\to 1 \quad \text{as } n \to \infty,
\end{aligned}$$

and, hence, $\bar{X}_n$ is a consistent sequence of estimators of θ. ||

In general, a detailed calculation, such as the above, is not necessary to verify consistency. Recall that, for an estimator W_n, Chebychev's Inequality states

$$P_\theta(|W_n - \theta| \geq \epsilon) \leq \frac{\mathrm{E}_\theta[(W_n - \theta)^2]}{\epsilon^2},$$

so if, for every $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathrm{E}_\theta[(W_n - \theta)^2] = 0,$$

then the sequence of estimators is consistent. Furthermore, by (7.3.1),

$$\mathrm{E}_\theta[(W_n - \theta)^2] = \mathrm{Var}_\theta\, W_n + [\mathrm{Bias}_\theta W_n]^2. \tag{10.1.3}$$

Putting this all together, we can state the following theorem.

Theorem 10.1.3 *If W_n is a sequence of estimators of a parameter θ satisfying*

i. $\lim_{n\to\infty} \mathrm{Var}_\theta\, W_n = 0$,
ii. $\lim_{n\to\infty} \mathrm{Bias}_\theta W_n = 0$,

for every $\theta \in \Theta$, then W_n is a consistent sequence of estimators of θ.

Example 10.1.4 (Continuation of Example 10.1.2) Since

$$\mathrm{E}_\theta \bar{X}_n = \theta \quad \text{and} \quad \mathrm{Var}_\theta\, \bar{X}_n = \frac{1}{n},$$

the conditions of Theorem 10.1.3 are satisfied and the sequence $\bar{X}_n$ is consistent. Furthermore, from Theorem 5.2.6, if there is iid sampling from any population with mean θ, then $\bar{X}_n$ is consistent for θ as long as the population has a finite variance. ||

At the beginning of this section we commented that the worth of an inconsistent sequence of estimators should be questioned. Part of the basis for this comment is the fact that there are so many consistent sequences, as the next theorem shows. Its proof is left to Exercise 10.2.

Theorem 10.1.5 *Let W_n be a consistent sequence of estimators of a parameter θ. Let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ be sequences of constants satisfying*

i. $\lim_{n\to\infty} a_n = 1$,
ii. $\lim_{n\to\infty} b_n = 0$.

Then the sequence $U_n = a_n W_n + b_n$ is a consistent sequence of estimators of θ.

We close this section with the outline of a more general result concerning the consistency of maximum likelihood estimators. This result shows that MLEs are consistent estimators of their parameters and is the first case we have seen in which a method of finding an estimator guarantees an optimality property.

To have consistency of the MLE, the underlying density (likelihood function) must satisfy certain "regularity conditions" that we will not go into here, but see Miscellanea 10.6.2 for details.

Theorem 10.1.6 (Consistency of MLEs) *Let $X_1, X_2, \ldots,$ be iid $f(x|\theta)$, and let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ be the likelihood function. Let $\hat{\theta}$ denote the MLE of θ. Let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions in Miscellanea 10.6.2 on $f(x|\theta)$ and, hence, $L(\theta|\mathbf{x})$, for every $\epsilon > 0$ and every $\theta \in \Theta$,*

$$\lim_{n\to\infty} P_\theta(|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon) = 0.$$

That is, $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$.

Proof: The proof proceeds by showing that $\frac{1}{n}\log L(\hat{\theta}|\mathbf{x})$ converges almost surely to $\mathrm{E}_\theta(\log f(X|\theta))$ for every $\theta \in \Theta$. Under some conditions on $f(x|\theta)$, this implies that $\hat{\theta}$ converges to θ in probability and, hence, $\tau(\hat{\theta})$ converges to $\tau(\theta)$ in probability. For details see Stuart, Ord, and Arnold (1999, Chapter 18). □

10.1.2 Efficiency

The property of consistency is concerned with the asymptotic accuracy of an estimator: Does it converge to the parameter that it is estimating? In this section we look at a related property, efficiency, which is concerned with the asymptotic variance of an estimator.

In calculating an asymptotic variance, we are, perhaps, tempted to proceed as follows. Given an estimator T_n based on a sample of size n, we calculate the finite-sample variance $\operatorname{Var} T_n$, and then evaluate $\lim_{n\to\infty} k_n \operatorname{Var} T_n$, where k_n is some normalizing constant. (Note that, in many cases, $\operatorname{Var} T_n \to 0$ as $n \to \infty$, so we need a factor k_n to force it to a limit.)

Definition 10.1.7 For an estimator T_n, if $\lim_{n\to\infty} k_n \operatorname{Var} T_n = \tau^2 < \infty$, where $\{k_n\}$ is a sequence of constants, then τ^2 is called the *limiting variance* or *limit of the variances.*

Example 10.1.8 (Limiting variances) For the mean $\bar{X}_n$ of n iid normal observations with $\mathrm{E}X = \mu$ and $\operatorname{Var} X = \sigma^2$, if we take $T_n = \bar{X}_n$, then $\lim \sqrt{n} \operatorname{Var} \bar{X}_n = \sigma^2$ is the limiting variance of T_n.

But a troubling thing happens if, for example, we were instead interested in estimating $1/\mu$ using $1/\bar{X}_n$. If we now take $T_n = 1/\bar{X}_n$, we find that the variance is $\operatorname{Var} T_n = \infty$, so the limit of the variances is infinity. But recall Example 5.5.23, where we said that the "approximate" mean and variance of $1/\bar{X}_n$ are

$$\mathrm{E}\left(\frac{1}{\bar{X}_n}\right) \approx \frac{1}{\mu},$$

$$\operatorname{Var}\left(\frac{1}{\bar{X}_n}\right) \approx \left(\frac{1}{\mu}\right)^4 \operatorname{Var}\bar{X}_n,$$

and thus by this second calculation the variance is $\operatorname{Var} T_n \approx \frac{\sigma^2}{n\mu^4} < \infty$. ||

This example points out the problems of using the limit of the variances as a large sample measure. Of course the exact finite sample variance of $1/\bar{X}$ is ∞. However, if

$\mu \neq 0$, the region where $1/\bar{X}$ gets very large has probability going to 0. So the second approximation in Example 10.1.8 is more realistic (as well as being much more useful). It is this second approach to calculating large sample variances that we adopt.

Definition 10.1.9 For an estimator T_n, suppose that $k_n(T_n - \tau(\theta)) \to \mathrm{n}(0, \sigma^2)$ in distribution. The parameter σ^2 is called the *asymptotic variance* or *variance of the limit distribution* of T_n.

For calculations of the variances of sample means and other types of averages, the limit variance and the asymptotic variance typically have the same value. But in more complicated cases, the limiting variance will sometimes fail us. It is also interesting to note that it is always the case that the asymptotic variance is smaller than the limiting variance (Lehmann and Casella 1998, Section 6.1). Here is an illustration.

Example 10.1.10 (Large-sample mixture variances) The hierarchical model

$$Y_n|W_n = w_n \sim \mathrm{n}\left(0, w_n + (1 - w_n)\sigma_n^2\right),$$
$$W_n \sim \text{Bernoulli}(p_n),$$

can exhibit big discrepancies between the asymptotic and limiting variances. (This is also sometimes described as a mixture model, where we observe $Y_n \sim \mathrm{n}(0,1)$ with probability p_n and $Y_n \sim \mathrm{n}(0, \sigma_n^2)$ with probability $1 - p_n$.)

First, using Theorem 4.4.7 we have

$$\mathrm{Var}(Y_n) = p_n + (1 - p_n)\sigma_n^2.$$

It then follows that the limiting variance of Y_n is finite only if $\lim_{n\to\infty}(1-p_n)\sigma_n^2 < \infty$.

On the other hand, the asymptotic distribution of Y_n can be directly calculated using

$$P(Y_n < a) = p_n P(Z < a) + (1 - p_n)P(Z < a/\sigma_n).$$

Suppose now we let $p_n \to 1$ and $\sigma_n \to \infty$ in such a way that $(1 - p_n)\sigma_n^2 \to \infty$. It then follows that $P(Y_n < a) \to P(Z < a)$, that is, $Y_n \to \mathrm{n}(0,1)$, and we have

$$\begin{aligned} \text{limiting variance} &= \lim_{n\to\infty} p_n + (1 - p_n)\sigma_n^2 = \infty, \\ \text{asymptotic variance} &= 1. \end{aligned}$$

See Exercise 10.6 for more details. ||

In the spirit of the Cramér–Rao Lower Bound (Theorem 7.3.9), there is an optimal asymptotic variance.

Definition 10.1.11 A sequence of estimators W_n is *asymptotically efficient* for a parameter $\tau(\theta)$ if $\sqrt{n}[W_n - \tau(\theta)] \to \mathrm{n}[0, v(\theta)]$ in distribution and

$$v(\theta) = \frac{[\tau'(\theta)]^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)};$$

that is, the asymptotic variance of W_n achieves the Cramér–Rao Lower Bound.

Recall that Theorem 10.1.6 stated that, under general conditions, MLEs are consistent. Under somewhat stronger regularity conditions, the same type of theorem holds with respect to asymptotic efficiency so, in general, we can consider MLEs to be consistent and asymptotically efficient. Again, details on the regularity conditions are in Miscellanea 10.6.2.

Theorem 10.1.12 (Asymptotic efficiency of MLEs) *Let $X_1, X_2, \ldots,$ be iid $f(x|\theta)$, let $\hat{\theta}$ denote the MLE of θ, and let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions in Miscellanea 10.6.2 on $f(x|\theta)$ and, hence, $L(\theta|\mathbf{x})$,*

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \to \mathrm{n}[0, v(\theta)],$$

where $v(\theta)$ is the Cramér–Rao Lower Bound. That is, $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

Proof: The proof of this theorem is interesting for its use of Taylor series and its exploiting of the fact that the MLE is defined as the zero of the likelihood function. We will outline the proof showing that $\hat{\theta}$ is asymptotically efficient; the extension to $\tau(\hat{\theta})$ is left to Exercise 10.7.

Recall that $l(\theta|\mathbf{x}) = \sum \log f(x_i|\theta)$ is the log likelihood function. Denote derivatives (with respect to θ) by $l', l'', \ldots$. Now expand the first derivative of the log likelihood around the true value θ_0,

$$l'(\theta|\mathbf{x}) = l'(\theta_0|\mathbf{x}) + (\theta - \theta_0)l''(\theta_0|\mathbf{x}) + \cdots, \tag{10.1.4}$$

where we are going to ignore the higher-order terms (a justifiable maneuver under the regularity conditions).

Now substitute the MLE $\hat{\theta}$ for θ, and realize that the left-hand side of (10.1.4) is 0. Rearranging and multiplying through by $\sqrt{n}$ gives us

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\frac{-l'(\theta_0|\mathbf{x})}{l''(\theta_0|\mathbf{x})} = \frac{-\frac{1}{\sqrt{n}}l'(\theta_0|\mathbf{x})}{\frac{1}{n}l''(\theta_0|\mathbf{x})}. \tag{10.1.5}$$

If we let $I(\theta_0) = \mathrm{E}[l'(\theta_0|X)]^2 = 1/v(\theta)$ denote the information number for one observation, application of the Central Limit Theorem and the Weak Law of Large Numbers will show (see Exercise 10.8 for details)

$$\begin{aligned} &-\frac{1}{\sqrt{n}}l'(\theta_0|\mathbf{X}) \to \mathrm{n}[0, I(\theta_0)], && \text{(in distribution)} \\ &\frac{1}{n}l''(\theta_0|\mathbf{X}) \to I(\theta_0). && \text{(in probability)} \end{aligned} \tag{10.1.6}$$

Thus, if we let $W \sim \mathrm{n}[0, I(\theta_0)]$, then $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to $W/I(\theta_0) \sim \mathrm{n}[0, 1/I(\theta_0)]$, proving the theorem. □

Example 10.1.13 (Asymptotic normality and consistency) The above theorem shows that it is typically the case that MLEs are efficient and consistent. We want to note that this phrase is somewhat redundant, as efficiency is defined only when the estimator is asymptotically normal and, as we will illustrate, asymptotic normality implies consistency. Suppose that

$$\sqrt{n}\frac{W_n-\mu}{\sigma} \to Z \text{ in distribution,}$$

where $Z \sim \mathrm{n}(0,1)$. By applying Slutsky's Theorem (Theorem 5.5.17) we conclude

$$W_n-\mu = \left(\frac{\sigma}{\sqrt{n}}\right)\left(\sqrt{n}\frac{W_n-\mu}{\sigma}\right) \to \lim_{n\to\infty}\left(\frac{\sigma}{\sqrt{n}}\right)Z = 0,$$

so $W_n - \mu \to 0$ in distribution. From Theorem 5.5.13 we know that convergence in distribution to a point is equivalent to convergence in probability, so W_n is a consistent estimator of μ. ||

10.1.3 Calculations and Comparisons

The asymptotic formulas developed in the previous sections can provide us with approximate variances for large-sample use. Again, we have to be concerned with regularity conditions (Miscellanea 10.6.2), but these are quite general and almost always satisfied in common circumstances. One condition deserves special mention, however, whose violation can lead to complications, as we have already seen in Example 7.3.13. For the following approximations to be valid, it must be the case that the support of the pdf or pmf, hence likelihood function, must be independent of the parameter.

If an MLE is asymptotically efficient, the asymptotic variance in Theorem 10.1.6 is the Delta Method variance of Theorem 5.5.24 (without the $1/n$ term). Thus, we can use the Cramér–Rao Lower Bound as an approximation to the true variance of the MLE. Suppose that $X_1, \ldots, X_n$ are iid $f(x|\theta)$, $\hat{\theta}$ is the MLE of θ, and $I_n(\theta) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})\right)^2$ is the information number of the sample. From the Delta Method and asymptotic efficiency of MLEs, the variance of $h(\hat{\theta})$ can be approximated by

$$\begin{aligned}
(10.1.7)\qquad \mathrm{Var}(h(\hat{\theta})|\theta) &\approx \frac{[h'(\theta)]^2}{I_n(\theta)} \\
&= \frac{[h'(\theta)]^2}{\mathrm{E}_\theta\left(-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{X})\right)} && \left(\begin{matrix}\text{using the identity}\\ \text{of Lemma 7.3.11}\end{matrix}\right) \\
&\approx \frac{[h'(\theta)]^2|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{X})|_{\theta=\hat{\theta}}}. && \left(\begin{matrix}\text{the denominator is } \hat{I}_n(\hat{\theta}) \text{, the}\\ \text{observed information number}\end{matrix}\right)
\end{aligned}$$

Furthermore, it has been shown (Efron and Hinkley 1978) that use of the *observed information number* is superior to the *expected information number*, the information number as it appears in the Cramér–Rao Lower Bound.

Notice that the variance estimation process is a two-step procedure, a fact that is somewhat masked by (10.1.7). To estimate $\mathrm{Var}_\theta\, h(\hat{\theta})$, first we *approximate* $\mathrm{Var}_\theta\, h(\hat{\theta})$; then we *estimate* the resulting approximation, usually by substituting $\hat{\theta}$ for θ. The resulting estimate can be denoted by $\mathrm{Var}_{\hat{\theta}}\, h(\hat{\theta})$ or $\widehat{\mathrm{Var}}_\theta\, h(\hat{\theta})$.

It follows from Theorem 10.1.6 that $-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log L\left(\theta|\mathbf{X}\right)|_{\theta=\hat{\theta}}$ is a consistent estimator of $I(\theta)$, so it follows that $\operatorname{Var}_{\hat{\theta}} h(\hat{\theta})$ is a consistent estimator of $\operatorname{Var}_\theta h(\hat{\theta})$.

Example 10.1.14 (Approximate binomial variance) In Example 7.2.7 we saw that $\hat{p} = \sum X_i/n$ is the MLE of p, where we have a random sample $X_1, \ldots, X_n$ from a Bernoulli(p) population. We also know by direct calculation that

$$\operatorname{Var}_p \hat{p} = \frac{p(1-p)}{n},$$

and a reasonable estimate of $\operatorname{Var}_p \hat{p}$ is

$$\widehat{\operatorname{Var}}_p \hat{p} = \frac{\hat{p}(1-\hat{p})}{n}. \tag{10.1.8}$$

If we apply the approximation in (10.1.7), with $h(p) = p$, we get as an estimate of $\operatorname{Var}_p \hat{p}$,

$$\widehat{\operatorname{Var}}_p \hat{p} \approx \frac{1}{-\frac{\partial^2}{\partial p^2}\log L(p|\mathbf{x})|_{p=\hat{p}}}.$$

Recall that

$$\log L(p|\mathbf{x}) = n\hat{p}\log(p) + n(1-\hat{p})\log(1-p),$$

and so

$$\frac{\partial^2}{\partial p^2}\log L(p|\mathbf{x}) = -\frac{n\hat{p}}{p^2} - \frac{n(1-\hat{p})}{(1-p)^2}.$$

Evaluating the second derivative at $p = \hat{p}$ yields

$$\left.\frac{\partial^2}{\partial p^2}\log L(p|\mathbf{x})\right|_{p=\hat{p}} = -\frac{n\hat{p}}{\hat{p}^2} - \frac{n(1-\hat{p})}{(1-\hat{p})^2} = -\frac{n}{\hat{p}(1-\hat{p})},$$

which gives a variance approximation identical to (10.1.8). We now can apply Theorem 10.1.6 to assert the asymptotic efficiency of $\hat{p}$ and, in particular, that

$$\sqrt{n}(\hat{p} - p) \to \mathrm{n}[0, p(1-p)]$$

in distribution. If we also employ Theorem 5.5.17 (Slutsky's Theorem) we can conclude that

$$\sqrt{n}\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}} \to \mathrm{n}[0, 1].$$

Estimating the variance of $\hat{p}$ is not really that difficult, and it is not necessary to bring in all of the machinery of these approximations. If we move to a slightly more complicated function, however, things can get a bit tricky. Recall that in Exercise 5.5.22 we used the Delta Method to approximate the variance of $\hat{p}/(1-\hat{p})$, an estimate of the odds $p/(1-p)$. Now we see that this estimator is, in fact, the MLE of the odds, and we can estimate its variance by

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right) &= \frac{\left[\frac{\partial}{\partial p}\left(\frac{p}{1-p}\right)\right]^2\Big|_{p=\hat{p}}}{-\frac{\partial^2}{\partial p^2}\log L(p|\mathbf{x})\Big|_{p=\hat{p}}} \\ &= \frac{\left[\frac{(1-p)+p}{(1-p)^2}\right]^2\Big|_{p=\hat{p}}}{\frac{n}{p(1-p)}\Big|_{p=\hat{p}}} \\ &= \frac{\hat{p}}{n(1-\hat{p})^3}.\end{aligned}$$

Moreover, we also know that the estimator is asymptotically efficient. ||

The MLE variance approximation works well in many cases, but it is not infallible. In particular, we must be careful when the function $h(\hat{\theta})$ is not monotone. In such cases, the derivative h' will have a sign change, and that may lead to an underestimated variance approximation. Realize that, since the approximation is based on the Cramér–Rao Lower Bound, it is probably an underestimate. However, nonmonotone functions can make this problem worse.

Example 10.1.15 (Continuation of Example 10.1.14) Suppose now that we want to estimate the variance of the Bernoulli distribution, $p(1-p)$. The MLE of this variance is given by $\hat{p}(1-\hat{p})$, and an estimate of the variance of this estimator can be obtained by applying the approximation of (10.1.7). We have

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{p}(1-\hat{p})\right) &= \frac{\left[\frac{\partial}{\partial p}(p(1-p))\right]^2\Big|_{p=\hat{p}}}{-\frac{\partial^2}{\partial p^2}\log L(p|\mathbf{x})\Big|_{p=\hat{p}}} \\ &= \frac{(1-2p)^2|_{p=\hat{p}}}{\frac{n}{p(1-p)}\Big|_{p=\hat{p}}} \\ &= \frac{\hat{p}(1-\hat{p})(1-2\hat{p})^2}{n},\end{aligned}$$

which can be 0 if $\hat{p}=\frac{1}{2}$, a clear underestimate of the variance of $\hat{p}(1-\hat{p})$. The fact that the function $p(1-p)$ is not monotone is a cause of this problem.

Using Theorem 10.1.6, we can conclude that our estimator is asymptotically efficient as long as $p \neq 1/2$. If $p = 1/2$ we need to use a second-order approximation as given in Theorem 5.5.26 (see Exercise 10.10). ||

The property of asymptotic efficiency gives us a benchmark for what we can hope to attain in asymptotic variance (although see Miscellanea 10.6.1). We also can use the asymptotic variance as a means of comparing estimators, through the idea of *asymptotic relative efficiency.*

Definition 10.1.16 If two estimators W_n and V_n satisfy

$$\sqrt{n}[W_n - \tau(\theta)] \rightarrow \mathrm{n}[0, \sigma_W^2]$$
$$\sqrt{n}[V_n - \tau(\theta)] \rightarrow \mathrm{n}[0, \sigma_V^2]$$

in distribution, the *asymptotic relative efficiency* (ARE) of V_n with respect to W_n is

$$\mathrm{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

Example 10.1.17 (AREs of Poisson estimators) Suppose that $X_1, X_2, \ldots$ are iid Poisson(λ), and we are interested in estimating the 0 probability. For example, the number of customers that come into a bank in a given time period is sometimes modeled as a Poisson random variable, and the 0 probability is the probability that no one will enter the bank in one time period. If $X \sim$ Poisson(λ), then $P(X = 0) = e^{-\lambda}$, and a natural (but somewhat naive) estimator comes from defining $Y_i = I(X_i = 0)$ and using

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

The Y_is are Bernoulli($e^{-\lambda}$), and hence it follows that

$$\mathrm{E}(\hat{\tau}) = e^{-\lambda} \quad \text{and} \quad \mathrm{Var}(\hat{\tau}) = \frac{e^{-\lambda}(1 - e^{-\lambda})}{n}.$$

Alternatively, the MLE of $e^{-\lambda}$ is $e^{-\hat{\lambda}}$, where $\hat{\lambda} = \sum_i X_i/n$ is the MLE of λ. Using Delta Method approximations, we have that

$$\mathrm{E}(e^{-\hat{\lambda}}) \approx e^{-\lambda} \quad \text{and} \quad \mathrm{Var}(e^{-\hat{\lambda}}) \approx \frac{\lambda e^{-2\lambda}}{n}.$$

Since

$$\sqrt{n}(\hat{\tau} - e^{-\lambda}) \rightarrow \mathrm{n}[0, e^{-\lambda}(1 - e^{-\lambda})]$$
$$\sqrt{n}(e^{-\hat{\lambda}} - e^{-\lambda}) \rightarrow \mathrm{n}[0, \lambda e^{-2\lambda}]$$

in distribution, the ARE of $\hat{\tau}$ with respect to the MLE $e^{-\hat{\lambda}}$ is

$$\mathrm{ARE}(\hat{\tau}, e^{-\hat{\lambda}}) = \frac{\lambda e^{-2\lambda}}{e^{-\lambda}(1 - e^{-\lambda})} = \frac{\lambda}{e^{\lambda} - 1}.$$

Examination of this function shows that it is strictly decreasing with a maximum of 1 (the best that $\hat{\tau}$ could hope to do) attained at $\lambda = 0$ and tailing off rapidly (being less than 10% when $\lambda = 4$) to asymptote to 0 as $\lambda \rightarrow \infty$. (See Exercise 10.9.) ||

Since the MLE is typically asymptotically efficient, another estimator cannot hope to beat its asymptotic variance. However, other estimators may have other desirable properties (ease of calculation, robustness to underlying assumptions) that make them desirable. In such situations, the efficiency of the MLE becomes important in calibrating what we are giving up if we use an alternative estimator.

We will look at one last example, contrasting ease of calculation with optimal variance. In the next section the robustness issue will be addressed.

Example 10.1.18 (Estimating a gamma mean) Difficult as it may seem to believe, estimation of the mean of a gamma distribution is not an easy task. Recall that the gamma pdf $f(x|\alpha,\beta)$ is given by

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}.$$

The mean of this distribution is $\alpha\beta$, and to compute the maximum likelihood estimator we have to deal with the derivative the of the gamma function (called the *digamma* function), which is never pleasant. In contrast, the method of moments gives us an easily computable estimate.

To be specific, suppose we have a random sample $X_1, X_2, \ldots, X_n$ from the gamma density above, but reparameterized so the mean, denoted by $\mu = \alpha\beta$, is explicit. This gives

$$f(x|\mu,\beta) = \frac{1}{\Gamma(\mu/\beta)\beta^{\mu/\beta}} x^{\mu/\beta-1} e^{-x/\beta},$$

and the method of moments estimator of μ is $\bar{X}$, with variance $\beta\mu/n$.

To calculate the MLE, we use the log likelihood

$$l(\mu,\beta|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\mu,\beta).$$

To ease the computations, assume that β is known so we solve $\frac{d}{d\mu} l(\mu,\beta|\mathbf{x}) = 0$ to get the MLE $\hat{\mu}$. There is no explicit solution, so we proceed numerically.

By Theorem 10.1.6 we know that $\hat{\mu}$ is asymptotically efficient. The question of interest is how much do we lose by using the easier-to-calculate method of moments estimator. To compare, we calculate the asymptotic relative efficiency,

$$\mathrm{ARE}(\bar{X},\hat{\mu}) = \frac{\mathrm{E}\left(-\frac{d^2}{d\mu^2} l(\mu,\beta|X)\right)}{\beta\mu}$$

and display it in Figure 10.1.1 for a selection of values of β. Of course, we know that the ARE must be greater than 1, but we see from the figure that for larger values of β it pays to do the more complex calculation and use the MLE. (See Exercise 10.11 for an extension, and Example A.0.7 for details on the calculations.) ||

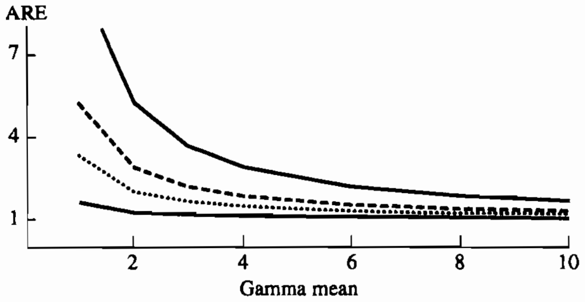

Figure 10.1.1. *Asymptotic relative efficiency of the method of moments estimator versus the MLE of a gamma mean. The four curves correspond to scale parameter values of (1,3,5,10), with the higher curves corresponding to the higher values of the scale parameter.*

10.1.4 Bootstrap Standard Errors

The *bootstrap*, which we first saw in Example 1.2.20, provides an alternative means of calculating standard errors. (It can also provide much more; see Miscellanea 10.6.3.)

The bootstrap is based on a simple, yet powerful, idea (whose mathematics can get quite involved).[1] In statistics, we learn about the characteristics of the population by taking samples. As the sample represents the population, analogous characteristics of the sample should give us information about the population characteristics. The bootstrap helps us learn about the sample characteristics by taking *resamples* (that is, we retake samples from the original sample) and use this information to infer to the population. The bootstrap was developed by Efron in the late 1970s, with the original ideas appearing in Efron (1979a, b) and the monograph by Efron (1982). See also Efron (1998) for more recent thoughts and developments.

Let us first look at a simple example where the bootstrap really is not needed.

Example 10.1.19 (Bootstrapping a variance) In Example 1.2.20 we calculated all possible averages of four numbers selected from

$$2, 4, 9, 12,$$

where we drew the numbers *with replacement*. This is the simplest form of the bootstrap, sometimes referred to as the *nonparametric bootstrap*. Figure 1.2.2 displays these values in a histogram.

What we have created is a resample of possible values of the sample mean. We saw that there are $\binom{4+4-1}{4} = 35$ distinct possible values, but these values are not equiprobable (and thus cannot be treated like a random sample). The $4^4 = 256$ (nondistinct) resamples are all equally likely, and they can be treated as a random sample. For the ith resample, we let $\bar{x}_i^*$ be the mean of that resample. We can then

[1] See Lehmann (1999, Section 6.5) for a most readable introduction.

estimate the variance of the sample mean $\bar{X}$ by

$$\text{Var}^*(\bar{X}) = \frac{1}{n^n - 1}\sum_{i=1}^{n^n}(\bar{x}_i^* - \bar{\bar{x}}^*)^2, \tag{10.1.9}$$

where $\bar{\bar{x}}^* = \frac{1}{n^n}\sum_{i=1}^{n^n}\bar{x}_i^*$, the mean of the resamples. (It is standard to let the notation $*$ denote a bootstrapped, or resampled, value.)

For our example we have that the bootstrap mean and variance are $\bar{\bar{x}}^* = 6.75$ and $\text{Var}^*(\bar{X}) = 3.94$. It turns out that, as far as means and variances are concerned, the bootstrap estimates are almost the same as the usual ones (see Exercise 10.13). $\|$

We have now seen how to calculate a bootstrap standard error, but in a problem where it is really not needed. However, the real advantage of the bootstrap is that, like the Delta Method, the variance formula (10.1.9) is applicable to virtually any estimator. Thus, for any estimator $\hat{\theta}(\mathbf{x}) = \hat{\theta}$, we can write

$$\text{Var}^*(\hat{\theta}) = \frac{1}{n^n - 1}\sum_{i=1}^{n^n}(\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2, \tag{10.1.10}$$

where $\hat{\theta}_i^*$ is the estimator calculated from the ith resample and $\bar{\hat{\theta}}^* = \frac{1}{n^n}\sum_{i=1}^{n^n}\hat{\theta}_i^*$, the mean of the resampled values.

Example 10.1.20 (Bootstrapping a binomial variance) In Example 10.1.15, we used the Delta Method to estimate the variance of $\hat{p}(1-\hat{p})$. Based on a sample of size n, we could alternatively estimate this variance by

$$\text{Var}^*(\hat{p}(1-\hat{p})) = \frac{1}{n^n - 1}\sum_{i=1}^{n^n}(\hat{p}(1-\hat{p})_i^* - \overline{\hat{p}(1-\hat{p})^*})^2. \qquad \|$$

But now a problem pops up. For our Example 10.1.19, with $n = 4$, there were 256 terms in the bootstrap sum. In more typical sample sizes, this number grows so large as to be uncomputable. (Enumerating all the possible resamples when $n > 15$ is virtually impossible, certainly for the authors.) But now we remember that we are statisticians – we take a sample of the resamples.

Thus, for a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and an estimate $\hat{\theta}(x_1, x_2, \ldots, x_n) = \hat{\theta}$, select B resamples (or *bootstrap samples*) and calculate

$$\text{Var}_B^*(\hat{\theta}) = \frac{1}{B - 1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2. \tag{10.1.11}$$

Example 10.1.21 (Conclusion of Example 10.1.20) For a sample of size $n = 24$, we compute the Delta Method variance estimate and the bootstrap variance estimate of $\hat{p}(1-\hat{p})$ using $B = 1000$. For $\hat{p} \neq 1/2$, we use the first-order Delta Method variance of Example 10.1.15, while for $\hat{p} = 1/2$, we use the second-order variance estimate of Theorem 5.5.26 (see Exercise 10.16). We see in Table 10.1.1 that in all cases the

Table 10.1.1. *Bootstrap and Delta Method variances of $\hat{p}(1-\hat{p})$. The second-order Delta Method (see Theorem 5.5.26) is used when $\hat{p} = 1/2$. The true variance is calculated numerically assuming that $\hat{p} = p$.*

	$\hat{p} = 1/4$	$\hat{p} = 1/2$	$\hat{p} = 2/3$
Bootstrap	.00508	.00555	.00561
Delta Method	.00195	.00022	.00102
True	.00484	.00531	.00519

bootstrap variance estimate is closer to the true variance, while the Delta Method variance is an underestimate. (This should not be a surprise, based on (10.1.7), which shows that the Delta Method variance estimate is based on a lower bound.)

The Delta Method is a "first-order" approximation, in that it is based on the first term of a Taylor series expansion. When that term is zeroed out (as when $\hat{p} = 1/2$), we must use the second-order Delta Method. In contrast, the bootstrap can often have "second-order" accuracy, getting more than the first term in an expansion correct (see Miscellanea 10.6.3). So here, the bootstrap automatically corrects for the case $\hat{p} = 1/2$. (Note that $24^{24} \approx 1.33 \times 10^{13}$, an enormous number, so enumerating the bootstrap samples is not feasible.) ||

The type of bootstrapping that we have been talking about so far is called the *nonparametric bootstrap*, as we have assumed no functional form for the population pdf or cdf. In contrast, we may also have a *parametric bootstrap*.

Suppose we have a sample $X_1, X_2, \ldots, X_n$ from a distribution with pdf $f(x|\theta)$, where θ may be a vector of parameters. We can estimate θ with $\hat{\theta}$, the MLE, and draw samples

$$X_1^*, X_2^*, \ldots, X_n^* \sim f(x|\hat{\theta}).$$

If we take B such samples, we can estimate the variance of $\hat{\theta}$ using (10.1.11). Note that these samples are not resamples of the data, but actual random samples drawn from $f(x|\hat{\theta})$, which is sometimes called the *plug-in distribution*.

Example 10.1.22 (Parametric bootstrap) Suppose that we have a sample

$$-1.81, 0.63, 2.22, 2.41, 2.95, 4.16, 4.24, 4.53, 5.09$$

with $\bar{x} = 2.71$ and $s^2 = 4.82$. If we assume that the underlying distribution is normal, then a parametric bootstrap would take samples

$$X_1^*, X_2^*, \ldots, X_n^* \sim \mathrm{n}(2.71, 4.82).$$

Based on $B = 1000$ samples, we calculate $\mathrm{Var}_B^*(S^2) = 4.33$. Based on normal theory, the variance of S^2 is $2(\sigma^2)^2/8$, which we could estimate with the MLE $2(4.82)^2/8 = 5.81$. The data values were actually generated from a normal distribution with variance 4, so $\mathrm{Var}\, S^2 = 4.00$. The parametric bootstrap is a better estimate here. (In Example 5.6.6 we estimated the distribution of S^2 using what we now know is the parametric bootstrap.) ||

Now that we have an all-purpose method for computing standard errors, how do we know it is a good method? In Example 10.1.21 it seems to do better than the Delta Method, which we know has some good properties. In particular, we know that the Delta Method, which is based on maximum likelihood estimation, will typically produce consistent estimators. Can we say the same for the bootstrap? Although we cannot answer this question in great generality, we say that, in many cases, the bootstrap does provide us with a reasonable estimator that is consistent.

To be a bit more precise, we separate the two distinct pieces in calculating a bootstrap estimator.

a. Establish that (10.1.11) converges to (10.1.10) as $B \to \infty$, that is,

$$\mathrm{Var}^*_B(\hat{\theta}) \overset{B\to\infty}{\to} \mathrm{Var}^*(\hat{\theta}).$$

b. Establish the consistency of the estimator (10.1.10), which uses the entire bootstrap sample, that is,

$$\mathrm{Var}^*(\hat{\theta}) \overset{n\to\infty}{\to} \mathrm{Var}(\hat{\theta}).$$

Part (a) can be established using the Law of Large Numbers (Exercise 10.15). Also notice that all of part (a) takes place in the sample. (Lehmann 1999, Section 6.5, calls $\mathrm{Var}^*_B(\hat{\theta})$ an *approximator* rather than an estimator.)

Establishing part (b) is a bit delicate, and this is where consistency is established. Typically consistency will be obtained in iid sampling, but in more general situations it may not occur. (Lehmann 1999, Section 6.5, gives an example.) For more details on consistency (necessarily at a more advanced level), see Shao and Tu (1995, Section 3.2.2) or Shao (1999, Section 5.5.3).

10.2 Robustness

Thus far, we have evaluated the performance of estimators assuming that the underlying model is the correct one. Under this assumption, we have derived estimators that are optimal in some sense. However, if the underlying model is not correct, then we cannot be guaranteed of the optimality of our estimator.

We cannot guard against all possible situations and, moreover, if our model is arrived at through some careful considerations, we shouldn't have to. But we may be concerned about small or medium-sized deviations from our assumed model. This may lead us to the consideration of *robust estimators.* Such estimators will give up optimality at the assumed model in exchange for reasonable performance if the assumed model is not the true model. Thus we have a trade-off, and the more important criterion, optimality or robustness, is probably best decided on a case-by-case basis.

The term "robustness" can have many interpretations, but perhaps it is best summarized by Huber (1981, Section 1.2), who noted:

> ...any statistical procedure should possess the following desirable features:
>
> (1) It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.

(2) It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly....

(3) Somewhat larger deviations from the model should not cause a catastrophe.

We first look at some simple examples to understand these items better; then we proceed to look at more general robust estimators and measures of robustness.

10.2.1 The Mean and the Median

Is the sample mean a robust estimator? It may depend on exactly how we formalize measures of robustness.

Example 10.2.1 (Robustness of the sample mean) Let $X_1, X_2, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$. We know that $\bar{X}$ has variance $\mathrm{Var}(\bar{X}) = \sigma^2/n$, which is the Cramér–Rao Lower Bound. Hence, $\bar{X}$ satisfies (1) in that it attains the best variance at the assumed model.

To investigate (2), the performance of $\bar{X}$ under small deviations from the model, we first need to decide on what this means. A common interpretation is to use an *δ-contamination* model; that is, for small δ, assume that we observe

$$X_i \sim \begin{cases} \mathrm{n}(\mu, \sigma^2) & \text{with probability } 1-\delta \\ f(x) & \text{with probability } \delta, \end{cases}$$

where $f(x)$ is some other distribution.

Suppose that we take $f(x)$ to be any density with mean θ and variance τ^2. Then

$$\mathrm{Var}(\bar{X}) = (1-\delta)\frac{\sigma^2}{n} + \delta\frac{\tau^2}{n} + \frac{\delta(1-\delta)(\theta-\mu)^2}{n}.$$

This actually looks pretty good for $\bar{X}$, since if $\theta \approx \mu$ and $\sigma \approx \tau$, $\bar{X}$ will be near optimal. We can perturb the model a little more, however, and make things quite bad. Consider what happens if $f(x)$ is a Cauchy pdf. Then it immediately follows that $\mathrm{Var}(\bar{X}) = \infty$. (See Exercises 10.18 for details and 10.19 for another situation.) ||

Turning to item (3), we ask what happens if there is an usually aberrant observation. Envision a particular set of sample values and then consider the effect of increasing the largest observation. For example, suppose that $X_{(n)} = x$, where $x \to \infty$. The effect of such an observation could be considered "catastrophic." Although none of the distributional properties of $\bar{X}$ are affected, the observed value would be "meaningless." This illustrates the *breakdown value*, an idea attributable to Hampel (1974).

Definition 10.2.2 Let $X_{(1)} < \cdots < X_{(n)}$ be an ordered sample of size n, and let T_n be a statistic based on this sample. T_n has *breakdown value* $b, 0 \le b \le 1$, if, for every $\epsilon > 0$,

$$\lim_{X_{(\{(1-b)n\})} \to \infty} T_n < \infty \quad \text{and} \quad \lim_{X_{(\{(1-(b+\epsilon))n\})} \to \infty} T_n = \infty.$$

(Recall Definition 5.4.2 on percentile notation.)

It is easy to see that the breakdown value of $\bar{X}$ is 0; that is, if any fraction of the sample is driven to infinity, so is the value of $\bar{X}$. In stark contrast, the sample median is unchanged by this change of the sample values. This insensitivity to extreme observations is sometimes considered an asset of the sample median, which has a breakdown value of 50%. (See Exercise 10.20 for more about breakdown values.)

Since the median is improving on the robustness of the mean, we might ask if we are losing anything by switching to a more robust estimator (of course we must!). For example, in the simple normal model of Example 10.2.1, the mean is the best unbiased estimator if the model is true. Therefore it follows that at the normal model (and close to it), the mean is a better estimator. But, the key question is, just how much better is the mean at the normal model? If we can answer this, we can make an informative choice on which estimator to use–and which criterion (optimality or robustness) we consider more important. To answer this question in some generality we call on the criterion of asymptotic relative efficiency.

To compute the ARE of the median with respect to the mean, we must first establish the asymptotic normality of the median and calculate the variance of the asymptotic distribution.

Example 10.2.3 (Asymptotic normality of the median) To find the limiting distribution of the median, we resort to an argument similar to that in the proof of Theorems 5.4.3 and 5.4.4, that is, an argument based on the binomial distribution.

Let $X_1, \ldots, X_n$ be a sample from a population with pdf f and cdf F (assumed to be differentiable), with $P(X_i \leq \mu) = 1/2$, so μ is the population median. Let M_n be the sample median, and consider computing

$$\lim_{n\to\infty} P\left(\sqrt{n}(M_n - \mu) \leq a\right)$$

for some a. If we define the random variables Y_i by

$$Y_i = \begin{cases} 1 & \text{if } X_i \leq \mu + a/\sqrt{n} \\ 0 & \text{otherwise,} \end{cases}$$

it follows that Y_i is a Bernoulli random variable with success probability $p_n = F(\mu + a/\sqrt{n})$. To avoid complications, we will assume that n is odd and thus the event $\{M_n \leq \mu + a/\sqrt{n}\}$ is equivalent to the event $\{\sum_i Y_i \geq (n+1)/2\}$.

Some algebra then yields

$$P\left(\sqrt{n}(M_n - \mu) \leq a\right) = P\left(\frac{\sum_i Y_i - np_n}{\sqrt{np_n(1-p_n)}} \geq \frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}\right).$$

Now $p_n \to p = F(\mu) = 1/2$, so we expect that an application of the Central Limit Theorem will show that $\frac{\sum_i Y_i - np_n}{\sqrt{np_n(1-p_n)}}$ converges in distribution to Z, a standard normal random variable. A straightforward limit calculation will also show that

$$\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} \to -2aF'(\mu) = -2af(\mu).$$

Putting this all together yields that

$$P\left(\sqrt{n}(M_n - \mu) \le a\right) \to P\left(Z \ge -2af(\mu)\right).$$

and thus $\sqrt{n}(M_n - \mu)$ is asymptotically normal with mean 0 and variance $1/[2f(\mu)]^2$. (For details, see Exercise 10.22, and for a rigorous, and more general, development of this result, see Shao 1999, Section 5.3.) ||

Example 10.2.4 (AREs of the median to the mean) As there are simple expressions for the asymptotic variances of the mean and the median, the ARE is easily computed. The following table gives the AREs for three symmetric distributions. We find, as might be expected, that as the tails of the distribution get heavier, the ARE gets bigger. That is, the performance of the median improves in distributions with heavy tails. See Exercise 10.23 for more comparisons.

Median/mean asymptotic relative efficiencies

Normal	Logistic	Double exponential
.64	.82	2

||

10.2.2 M-Estimators

Many of the estimators that we use are the result of minimizing a particular criterion. For example, if $X_1, X_2, \ldots, X_n$ are iid from $f(x|\theta)$, possible estimators are the mean, the minimizer of $\sum(x_i - a)^2$; the median, the minimizer of $\sum|x_i - a|$; and the MLE, the maximizer of $\prod_{i=1}^n f(x_i|\theta)$ (or the minimizer of the negative likelihood). As a systematic way of obtaining a robust estimator, we might attempt to write down a criterion function whose minimum would result in an estimator with desirable robustness properties.

In an attempt at defining a robust criterion, Huber (1964) considered a compromise between the mean and the median. The mean criterion is a square, which gives it sensitivity, but in the "tails" the square gives too much weight to big observations. In contrast, the absolute value criterion of the median does not overweight big or small observations. The compromise is to minimize a criterion function

$$\sum_{i=1}^{n} \rho(x_i - a), \tag{10.2.1}$$

where ρ is given by

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le k \\ k|x| - \frac{1}{2}k^2 & \text{if } |x| \ge k. \end{cases} \tag{10.2.2}$$

The function $\rho(x)$ acts like x^2 for $|x| \le k$ and like $|x|$ for $|x| > k$. Moreover, since $\frac{1}{2}k^2 = k|k| - \frac{1}{2}k^2$, the function is continuous (see Exercise 10.28). In fact ρ is differentiable. The constant k, which can also be called a *tuning parameter*, controls the mix, with small values of k yielding a more "median-like" estimator.

Table 10.2.1. *Huber estimators*

k	0	1	2	3	4	5	6	8	10
Estimate	$-.21$	.03	$-.04$	.29	.41	.52	.87	.97	1.33

Example 10.2.5 (Huber estimator) The estimator defined as the minimizer of (10.2.1) and (10.2.2) is called a *Huber estimator* . To see how the estimator works, and how the choice of k matters, consider the following data set consisting of eight standard normal deviates and three "outliers":

$$\mathbf{x} = -1.28, -.96, -.46, -.44, -.26, -.21, -.063, .39, 3, 6, 9$$

For these data the mean is 1.33 and the median is $-.21$. As k varies, we get the range of Huber estimates given in Table 10.2.1. We see that as k increases, the Huber estimate varies between the median and the mean, so we interpret increasing k as decreasing robustness to outliers. ||

The estimator minimizing (10.2.2) is a special case of the estimators studied by Huber. For a general function ρ, we call the estimator minimizing $\sum_i \rho(x_i - \theta)$ an *M-estimator*, a name that is to remind us that these are *maximum-likelihood-type* estimators. Note that if we choose ρ to be the negative log likelihood $-l(\theta|x)$, then the M-estimator is the usual MLE. But with more flexibility in choosing the function to be minimized, estimators with different properties can be derived.

Since minimization of a function is typically done by solving for the zeros of the derivative (when we can take a derivative), defining $\psi = \rho'$, we see that an M-estimator is the solution to

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0. \tag{10.2.3}$$

Characterizing an estimator as the root of an equation is particularly useful for getting properties of the estimator, for arguments like those used for likelihood estimators can be extended. In particular, look at Section 10.1.2, especially the proof of Theorem 10.1.12. We assume that the function $\rho(x)$ is symmetric, and its derivative $\psi(x)$ is monotone increasing (which ensures that the root of (10.2.3) is the unique minimum). Then, as in the proof of Theorem 10.1.12, we write a Taylor expansion for ψ as

$$\sum_{i=1}^{n} \psi(x_i - \theta) = \sum_{i=1}^{n} \psi(x_i - \theta_0) + (\theta - \theta_0) \sum_{i=1}^{n} \psi'(x_i - \theta_0) + \cdots,$$

where θ_0 is the true value, and we ignore the higher-order terms. Let $\hat{\theta}_M$ be the solution to (10.2.3) and substitute this for θ to obtain

$$0 = \sum_{i=1}^{n} \psi(x_i - \theta_0) + (\hat{\theta}_M - \theta_0) \sum_{i=1}^{n} \psi'(x_i - \theta_0) + \cdots,$$

where the left-hand side is 0 because $\hat{\theta}_M$ is the solution. Now, again analogous to the proof of Theorem 10.1.12, we rearrange terms, divide through by $\sqrt{n}$, and ignore the remainder terms to get

$$\sqrt{n}(\hat{\theta}_M - \theta_0) = \frac{\frac{-1}{\sqrt{n}}\sum_{i=1}^n \psi(x_i - \theta_0)}{\frac{1}{n}\sum_{i=1}^n \psi'(x_i - \theta_0)}.$$

Now we assume that θ_0 satisfies $\mathrm{E}_{\theta_0}\psi(X - \theta_0) = 0$ (which is usually taken as the definition of θ_0). It follows that

$$(10.2.4)\quad \frac{-1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i - \theta_0) = \sqrt{n}\left[\frac{-1}{n}\sum_{i=1}^n \psi(X_i - \theta_0)\right] \to \mathrm{n}\left(0, \mathrm{E}_{\theta_0}\psi(X - \theta_0)^2\right)$$

in distribution, and the Law of Large Numbers yields

$$(10.2.5)\qquad \frac{1}{n}\sum_{i=1}^n \psi'(x_i - \theta_0) \to \mathrm{E}_{\theta_0}\psi'(X - \theta_0)$$

in probability. Putting this all together we have

$$(10.2.6)\qquad \sqrt{n}(\hat{\theta}_M - \theta_0) \to \mathrm{n}\left(0, \frac{\mathrm{E}_{\theta_0}\psi(X - \theta_0)^2}{[\mathrm{E}_{\theta_0}\psi'(X - \theta_0)]^2}\right).$$

Example 10.2.6 (Limit distribution of the Huber estimator) If $X_1, \ldots, X_n$ are iid from a pdf $f(x-\theta)$, where f is symmetric around 0, then for ρ given by (10.2.2) we have

$$(10.2.7)\qquad \psi(x) = \begin{cases} x & \text{if } |x| \le k \\ k & \text{if } x > k \\ -k & \text{if } x < -k \end{cases}$$

and thus

$$(10.2.8)\qquad \begin{aligned} E_\theta\psi(X - \theta) &= \int_{\theta-k}^{\theta+k} (x-\theta)f(x-\theta)\,dx \\ &\quad - k\int_{-\infty}^{\theta-k} f(x-\theta)\,dx + k\int_{\theta+k}^{\infty} f(x-\theta)\,dx \\ &= \int_{-k}^{k} yf(y)\,dy - k\int_{-\infty}^{-k} f(y)\,dy + k\int_{k}^{\infty} f(y)\,dy = 0, \end{aligned}$$

where we substitute $y = x - \theta$. The integrals add to 0 by the symmetry of f. Thus, the Huber estimator has the correct mean (see Exercise 10.25).

To calculate the variance we need the expected value of ψ'. While ψ is not differentiable, beyond the points of nondifferentiability ($x = \pm k$) ψ' will be 0. Thus, we only need deal with the expectation for $|x| \le k$, and we have

$$E_\theta\psi'(X-\theta) = \int_{\theta-k}^{\theta+k} f(x-\theta)\,dx = P_0(|X|\le k),$$

$$\begin{aligned} E_\theta\psi(X-\theta)^2 &= \int_{\theta-k}^{\theta+k}(x-\theta)^2 f(x-\theta)\,dx + k^2\int_{\theta+k}^{\infty} f(x-\theta)\,dx + k^2\int_{-\infty}^{\theta-k} f(x-\theta)\,dx \\ &= \int_{-k}^{k} x^2 f(x)\,dx + 2k^2\int_{k}^{\infty} f(x)\,dx. \end{aligned}$$

Thus we can conclude that the Huber estimator is asymptotically normal with mean θ and asymptotic variance

$$\frac{\int_{-k}^{k} x^2 f(x)\,dx + 2k^2 P_0(|X|>k)}{[P_0(|X|\le k)]^2}. \qquad \|$$

As we did in Example 10.2.4, we now examine the ARE of the Huber estimator for a variety of distributions.

Example 10.2.7 (ARE of the Huber estimator) As the Huber estimator is, in a sense, a mean/median compromise, we'll look at its relative efficiency with respect to both of these estimators.

Huber estimator asymptotic relative efficiencies, $k = 1.5$

	Normal	Logistic	Double exponential
vs. mean	.96	1.08	1.37
vs. median	1.51	1.31	.68

The Huber estimator behaves similarly to the mean for the normal and logistic distributions and is an improvement on the median. For the double exponential it is an improvement over the mean but not as good as the median. Recall that the mean is the MLE for the normal, and the median is the MLE for the double exponential (so AREs < 1 are expected). The Huber estimator has performance similar to the MLEs for these distributions but also seems to maintain reasonable performance in other cases. ‖

We see that an M-estimator is a compromise between robustness and efficiency. We now look a bit more closely at what we may be giving up, in terms of efficiency, to gain robustness.

Let us look more closely at the asymptotic variance in (10.2.6). The denominator of the variance contains the term $\mathrm{E}_{\theta_0}\psi'(X-\theta_0)$, which we can write as

$$\mathrm{E}_\theta\psi'(X-\theta) = \int \psi'(x-\theta)f(x-\theta)\,dx = -\int\left[\frac{\partial}{\partial\theta}\psi(x-\theta)\right]f(x-\theta)\,dx.$$

Now we use the differentiation product rule to get

$$\frac{d}{d\theta}\int\psi(x-\theta)f(x-\theta)\,dx = \int\left[\frac{d}{d\theta}\psi(x-\theta)\right]f(x-\theta)\,dx + \int\psi(x-\theta)\left[\frac{d}{d\theta}f(x-\theta)\right]dx.$$

The left hand side is 0 because $E_\theta\psi(x-\theta)=0$, so we have

$$-\int\left[\frac{d}{d\theta}\psi(x-\theta)\right]f(x-\theta)\,dx=\int\psi(x-\theta)\left[\frac{d}{d\theta}f(x-\theta)\right]dx$$
$$=\int\psi(x-\theta)\left[\frac{d}{d\theta}\log f(x-\theta)\right]f(x-\theta)\,dx,$$

where we use the fact that $\frac{d}{dy}g(y)/g(y)=\frac{d}{dy}\log g(y)$. This last expression can be written $\mathrm{E}_\theta[\psi(X-\theta)l'(\theta|X)]$, where $l(\theta|X)$ is the log likelihood, yielding the identity

$$\mathrm{E}_\theta\psi'(X-\theta)=-\mathrm{E}_\theta\left[\frac{d}{d\theta}\psi(X-\theta)\right]=\mathrm{E}_\theta[\psi(X-\theta)l'(\theta|X)]$$

(which, when we choose $\psi=l'$, yields the (we hope) familiar equation $-\mathrm{E}_\theta[l''(\theta|X)]=\mathrm{E}_\theta l'(\theta|X)^2$; see Lemma 7.3.11).

It is now a simple matter to compare the asymptotic variance of an M-estimator to that of the MLE. Recall that the asymptotic variance of the MLE, $\hat\theta$, is given by $1/\mathrm{E}_\theta l'(\theta|X)^2$, so we have

$$\text{(10.2.9)}\qquad \mathrm{ARE}(\hat\theta_M,\hat\theta)=\frac{[\mathrm{E}_\theta\psi(X-\theta_0)l'(\theta|X)]^2}{\mathrm{E}_\theta\psi(X-\theta)^2\mathrm{E}_\theta l'(\theta|X)^2}\le 1$$

by virtue of the Cauchy-Swartz Inequality. Thus, an M-estimator is always less efficient than the MLE, and matches its efficiency only if ψ is proportional to l' (see Exercise 10.29).

In this section we did not try to classify all types of robust estimators, but rather we were content with some examples. There are many good books that treat robustness in detail; the interested reader might try Staudte and Sheather (1990) or Hettmansperger and McKean (1998).

10.3 Hypothesis Testing

As in Section 10.1, this section describes a few methods for deriving *some* tests in complicated problems. We are thinking of problems in which no optimal test, as defined in earlier sections, exists (for example, no UMP unbiased test exists) or is known. In such situations, the derivation of any reasonable test might be of use. In two subsections, we will discuss large-sample properties of likelihood ratio tests and other approximate large-sample tests.

10.3.1 Asymptotic Distribution of LRTs

One of the most useful methods for complicated models is the likelihood ratio method of test construction because it gives an explicit definition of the test statistic,

$$\lambda(\mathbf{x})=\frac{\sup_{\Theta_0}L(\theta|\mathbf{x})}{\sup_{\Theta}L(\theta|\mathbf{x})},$$

and an explicit form for the rejection region, $\{\mathbf{x}: \lambda(\mathbf{x}) \leq c\}$. After the data $\mathbf{X} = \mathbf{x}$ are observed, the likelihood function, $L(\theta|\mathbf{x})$, is a completely defined function of the variable θ. Even if the two suprema of $L(\theta|\mathbf{x})$, over the sets Θ_0 and Θ, cannot be analytically obtained, they can usually be computed numerically. Thus, the test statistic $\lambda(\mathbf{x})$ can be obtained for the observed data point even if no convenient formula defining $\lambda(\mathbf{x})$ is available.

To define a level α test, the constant c must be chosen so that

$$\sup_{\theta\in\Theta_0} P_\theta\left(\lambda(\mathbf{X}) \leq c\right) \leq \alpha. \tag{10.3.1}$$

If we cannot derive a simple formula for $\lambda(\mathbf{x})$, it might seem that it is hopeless to derive the sampling distribution of $\lambda(\mathbf{X})$ and thus know how to pick c to ensure (10.3.1). However, if we appeal to asymptotics, we can get an approximate answer.

Analogous to Theorem 10.1.12, we have the following result.

Theorem 10.3.1 (Asymptotic distribution of the LRT—simple H_0) *For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, suppose $X_1, \ldots, X_n$ are iid $f(x|\theta)$, $\hat{\theta}$ is the MLE of θ, and $f(x|\theta)$ satisfies the regularity conditions in Miscellanea 10.6.2. Then under H_0, as $n \to \infty$,*

$$-2\log\lambda(\mathbf{X}) \to \chi_1^2 \text{ in distribution,}$$

where χ_1^2 is a χ^2 random variable with 1 degree of freedom.

Proof: First expand $\log L(\theta|\mathbf{x}) = l(\theta|\mathbf{x})$ in a Taylor series around $\hat{\theta}$, giving

$$l(\theta|\mathbf{x}) = l(\hat{\theta}|\mathbf{x}) + l'(\hat{\theta}|\mathbf{x})(\theta - \hat{\theta}) + l''(\hat{\theta}|\mathbf{x})\frac{(\theta-\hat{\theta})^2}{2!} + \cdots.$$

Now substitute the expansion for $l(\theta_0|\mathbf{x})$ in $-2\log\lambda(\mathbf{x}) = -2l(\theta_0|\mathbf{x}) + 2l(\hat{\theta}|\mathbf{x})$, and get

$$-2\log\lambda(\mathbf{x}) \approx \frac{(\theta-\hat{\theta})^2}{-l''(\hat{\theta}|\mathbf{x})},$$

where we use the fact that $l'(\hat{\theta}|\mathbf{x}) = 0$. Since the denominator is the observed information $\hat{I}_n(\hat{\theta})$ and $\frac{1}{n}\hat{I}_n(\hat{\theta}) \to I(\theta_0)$ it follows from Theorem 10.1.12 and Slutsky's Theorem (Theorem 5.5.17) that $-2\log\lambda(\mathbf{X}) \to \chi_1^2$. □

Example 10.3.2 (Poisson LRT) For testing $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$ based on observing $X_1, \ldots, X_n$ iid Poisson(λ), we have

$$-2\log\lambda(\mathbf{x}) = -2\log\left(\frac{e^{-n\lambda_0}\lambda_0^{\Sigma x_i}}{e^{-n\hat{\lambda}}\hat{\lambda}^{\Sigma x_i}}\right) = 2n\left[(\lambda_0 - \hat{\lambda}) - \hat{\lambda}\log(\lambda_0/\hat{\lambda})\right],$$

where $\hat{\lambda} = \Sigma x_i/n$ is the MLE of λ. Applying Theorem 10.3.1, we would reject H_0 at level α if $-2\log\lambda(\mathbf{x}) > \chi_{1,\alpha}^2$.

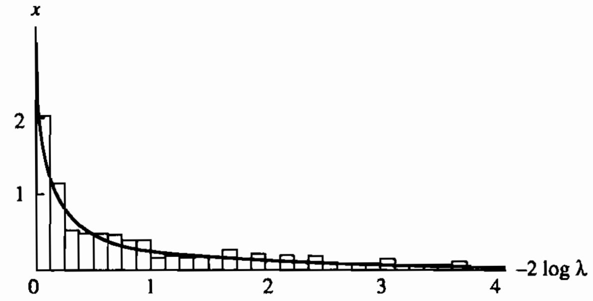

Figure 10.3.1. *Histogram of* 10,000 *values of* $-2\log\lambda(\mathbf{x})$ *along with the pdf of a* χ_1^2, $\lambda_0 = 5$ *and* $n = 25$

To get some idea of the accuracy of the asymptotics, here is a small simulation of the test statistic. For $\lambda_0 = 5$ and $n = 25$, Figure 10.3.1 shows a histogram of 10,000 values of $-2\log\lambda(\mathbf{x})$ along with the pdf of a χ_1^2. The match seems to be reasonable. Moreover, a comparison of the simulated (exact) and χ_1^2 (approximate) cutoff points in the following table shows that the cutoffs are remarkably similar.

Simulated (exact) and approximate percentiles of the Poisson LRT statistic

Percentile	.80	.90	.95	.99
Simulated	1.630	2.726	3.744	6.304
χ^2	1.642	2.706	3.841	6.635

||

Theorem 10.3.1 can be extended to the cases where the null hypothesis concerns a vector of parameters. The following generalization, which we state without proof, allows us to ensure (10.3.1) is true, at least for large samples. A complete discussion of this topic may be found in Stuart, Ord, and Arnold (1999, Chapter 22).

Theorem 10.3.3 *Let* $X_1, \ldots, X_n$ *be a random sample from a pdf or pmf* $f(x|\theta)$. *Under the regularity conditions in Miscellanea 10.6.2, if* $\theta \in \Theta_0$, *then the distribution of the statistic* $-2\log\lambda(\mathbf{X})$ *converges to a chi squared distribution as the sample size* $n \to \infty$. *The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by* $\theta \in \Theta_0$ *and the number of free parameters specified by* $\theta \in \Theta$.

Rejection of $H_0\colon \theta \in \Theta_0$ for small values of $\lambda(\mathbf{X})$ is equivalent to rejection for large values of $-2\log\lambda(\mathbf{X})$. Thus,

$$H_0 \text{ is rejected if and only if } -2\log\lambda(\mathbf{X}) \geq \chi^2_{\nu,\alpha},$$

where ν is the degrees of freedom specified in Theorem 10.3.3. The Type I Error probability will be approximately α if $\theta \in \Theta_0$ and the sample size is large. In this way, (10.3.1) will be approximately satisfied for large sample sizes and an *asymptotic size* α *test* has been defined. Note that the theorem will actually imply only that

$$\lim_{n\to\infty} P_\theta(\text{reject } H_0) = \alpha \quad \text{for each } \theta \in \Theta_0,$$

not that the $\sup_{\theta\in\Theta_0} P_\theta(\text{reject}H_0)$ converges to α. This is usually the case for asymptotic size α tests.

The computation of the degrees of freedom for the test statistic is usually straightforward. Most often, Θ can be represented as a subset of q-dimensional Euclidian space that contains an open subset in $\Re^q$, and Θ_0 can be represented as a subset of p-dimensional Euclidian space that contains an open subset in $\Re^p$, where $p < q$. Then $q - p = \nu$ is the degrees of freedom for the test statistic.

Example 10.3.4 (Multinomial LRT) Let $\theta = (p_1, p_2, p_3, p_4, p_5)$, where the p_js are nonnegative and sum to 1. Suppose $X_1, \ldots, X_n$ are iid discrete random variables and $P_\theta(X_i = j) = p_j, j = 1, \ldots, 5$. Thus the pmf of X_i is $f(j|\theta) = p_j$ and the likelihood function is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) = p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} p_5^{y_5},$$

where y_j = number of $x_1, \ldots, x_n$ equal to j. Consider testing

$$H_0\colon p_1 = p_2 = p_3 \text{ and } p_4 = p_5 \qquad \text{versus} \qquad H_1\colon H_0 \text{ is not true.}$$

The full parameter space, Θ, is really a four-dimensional set. Since $p_5 = 1 - p_1 - p_2 - p_3 - p_4$, there are only four free parameters. The parameter set is defined by

$$\sum_{j=1}^{4} p_j \leq 1 \quad \text{and} \quad p_j \geq 0, \quad j = 1, \ldots, 4,$$

a subset of $\Re^4$ containing an open subset of $\Re^4$. Thus $q = 4$. There is only one free parameter in the set specified by H_0 because, once $p_1, 0 \leq p_1 \leq \frac{1}{3}$, is fixed, $p_2 = p_3$ must equal p_1 and $p_4 = p_5$ must equal $\frac{1-3p_1}{2}$. Thus $p = 1$, and the degrees of freedom is $\nu = 4 - 1 = 3$.

To calculate $\lambda(\mathbf{x})$, the MLE of θ under both Θ_0 and Θ must be determined. By setting

$$\frac{\partial}{\partial p_j} \log L(\theta|\mathbf{x}) = 0 \quad \text{for each of } j = 1, \ldots, 4,$$

and using the facts that $p_5 = 1 - p_1 - p_2 - p_3 - p_4$ and $y_5 = n - y_1 - y_2 - y_3 - y_4$, we can verify that the MLE of p_j under Θ is $\hat{p}_j = y_j/n$. Under H_0, the likelihood function reduces to

$$L(\theta|\mathbf{x}) = p_1^{y_1+y_2+y_3} \left(\frac{1-3p_1}{2}\right)^{y_4+y_5}.$$

Again, the usual method of setting the derivative equal to 0 shows that the MLE of p_1 under H_0 is $\hat{p}_{10} = (y_1 + y_2 + y_3)/(3n)$. Then $\hat{p}_{10} = \hat{p}_{20} = \hat{p}_{30}$ and $\hat{p}_{40} = \hat{p}_{50} = (1 - 3\hat{p}_{10})/2$. Substituting these values and the $\hat{p}_j$ values into $L(\theta|\mathbf{x})$ and combining terms with the same exponent yield

$$\lambda(\mathbf{x}) =$$

$$\left(\frac{y_1+y_2+y_3}{3y_1}\right)^{y_1} \left(\frac{y_1+y_2+y_3}{3y_2}\right)^{y_2} \left(\frac{y_1+y_2+y_3}{3y_3}\right)^{y_3} \left(\frac{y_4+y_5}{2y_4}\right)^{y_4} \left(\frac{y_4+y_5}{2y_5}\right)^{y_5}.$$

Thus the test statistic is

$$-2\log\lambda(\mathbf{x}) = 2\sum_{i=1}^{5} y_i \log\left(\frac{y_i}{m_i}\right), \tag{10.3.2}$$

where $m_1 = m_2 = m_3 = (y_1+y_2+y_3)/3$ and $m_4 = m_5 = (y_4+y_5)/2$. The asymptotic size α test rejects H_0 if $-2\log\lambda(\mathbf{x}) \geq \chi^2_{3,\alpha}$. This example is one of a large class of testing problems for which the asymptotic theory of the likelihood ratio test is extensively used. ||

10.3.2 Other Large-Sample Tests

Another common method of constructing a large-sample test statistic is based on an estimator that has an asymptotic normal distribution. Suppose we wish to test a hypothesis about a real-valued parameter θ, and $W_n = W(X_1,\ldots,X_n)$ is a point estimator of θ, based on a sample of size n, that has been derived by some method. For example, W_n might be the MLE of θ. An approximate test, based on a normal approximation, can be justified in the following way. If σ_n^2 denotes the variance of W_n and if we can use some form of the Central Limit Theorem to show that, as $n \to \infty$, $(W_n-\theta)/\sigma_n$ converges in distribution to a standard normal random variable, then $(W_n-\theta)/\sigma_n$ can be compared to a $\mathrm{n}(0,1)$ distribution. We therefore have the basis for an approximate test.

There are, of course, many details to be verified in the argument of the previous paragraph, but this idea does have application in many situations. For example, if W_n is an MLE, Theorem 10.1.12 can be used to validate the above arguments. Note that the distribution of W_n and, perhaps, the value of σ_n depend on the value of θ. The convergence, therefore, more formally says that for each fixed value of $\theta \in \Theta$, if we use the corresponding distribution for W_n and the corresponding value for σ_n, $(W_n-\theta)/\sigma_n$ converges to a standard normal. If, for each n, σ_n is a calculable constant (which may depend on θ but not any other unknown parameters), then a test based on $(W_n-\theta)/\sigma_n$ might be derived.

In some instances, σ_n also depends on unknown parameters. In such a case, we look for an estimate S_n of σ_n with the property that σ_n/S_n converges in probability to 1. Then, using Slutsky's Theorem (as in Example 5.5.18) we can deduce that $(W_n-\theta)/S_n$ also converges in distribution to a standard normal distribution. A large-sample test may be based on this fact.

Suppose we wish to test the two-sided hypothesis $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$. An approximate test can be based on the statistic $Z_n = (W_n - \theta_0)/S_n$ and would reject H_0 if and only if $Z_n < -z_{\alpha/2}$ or $Z_n > z_{\alpha/2}$. If H_0 is true, then $\theta = \theta_0$ and Z_n converges in distribution to $Z \sim \mathrm{n}(0,1)$. Thus, the Type I Error probability,

$$P_{\theta_0}(Z_n < -z_{\alpha/2} \text{ or } Z_n > z_{\alpha/2}) \to P(Z < -z_{\alpha/2} \text{ or } Z > z_{\alpha/2}) = \alpha,$$

and this is an asymptotically size α test.

Now consider an alternative parameter value $\theta \neq \theta_0$. We can write

$$Z_n = \frac{W_n-\theta_0}{S_n} = \frac{W_n-\theta}{S_n} + \frac{\theta-\theta_0}{S_n}. \tag{10.3.3}$$

No matter what the value of θ, the term $(W_n - \theta)/S_n \to \mathrm{n}(0,1)$. Typically, it is also the case that $\sigma_n \to 0$ as $n \to \infty$. (Recall, $\sigma_n = \operatorname{Var} W_n$, and estimators typically become more precise as $n \to \infty$.) Thus, S_n will converge in probability to 0 and the term $(\theta - \theta_0)/S_n$ will converge to $+\infty$ or $-\infty$ in probability, depending on whether $(\theta - \theta_0)$ is positive or negative. Thus, Z_n will converge to $+\infty$ or $-\infty$ in probability and

$$P_\theta(\text{reject } H_0) = P_\theta(Z_n < -z_{\alpha/2} \text{ or } Z_n > z_{\alpha/2}) \to 1 \quad \text{as } n \to \infty.$$

In this way, a test with asymptotic size α and asymptotic power 1 can be constructed.

If we wish to test the one-sided hypothesis $H_0\colon \theta \le \theta_0$ versus $H_1\colon \theta > \theta_0$, a similar test might be constructed. Again, the test statistic $Z_n = (W_n - \theta_0)/S_n$ would be used and the test would reject H_0 if and only if $Z_n > z_\alpha$. Using reasoning similar to the above, we could conclude that the power function of this test converges to 0, α, or 1 according as $\theta < \theta_0, \theta = \theta_0$, or $\theta > \theta_0$. Thus this test too has reasonable asymptotic power properties.

In general, a *Wald test* is a test based on a statistic of the form

$$Z_n = \frac{W_n - \theta_0}{S_n},$$

where θ_0 is a hypothesized value of the parameter θ, W_n is an estimator of θ, and S_n is a standard error for W_n, an estimate of the standard deviation of W_n. If W_n is the MLE of θ, then, as discussed in Section 10.1.3, $1/\sqrt{I_n(W_n)}$ is a reasonable standard error for W_n. Alternatively, $1/\sqrt{\hat{I}_n(W_n)}$, where

$$\hat{I}_n(W_n) = -\left.\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{X})\right|_{\theta = W_n}$$

is the observed information number, is often used (see (10.1.7)).

Example 10.3.5 (Large-sample binomial tests) Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli(p) population. Consider testing $H_0 : p \le p_0$ versus $H_1 : p > p_0$, where $0 < p_0 < 1$ is a specified value. The MLE of p, based on a sample of size n, is $\hat{p}_n = \sum_{i=1}^n X_i/n$. Since $\hat{p}_n$ is just a sample mean, the Central Limit Theorem applies and states that for any p, $0 < p < 1, (\hat{p}_n - p)/\sigma_n$ converges to a standard normal random variable. Here $\sigma_n = \sqrt{p(1-p)/n}$, a value that depends on the unknown parameter p. A reasonable estimate of σ_n is $S_n = \sqrt{\hat{p}_n(1-\hat{p}_n)/n}$, and it can be shown (see Exercise 5.32) that σ_n/S_n converges in probability to 1. Thus, for any p, $0 < p < 1$,

$$\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \to \mathrm{n}(0,1).$$

The Wald test statistic Z_n is defined by replacing p by p_0, and the large-sample Wald test rejects H_0 if $Z_n > z_\alpha$. As an alternative estimate of σ_n, it is easily checked that

$1/I_n(\hat{p}_n) = \hat{p}_n(1-\hat{p}_n)/n$. So, the same statistic Z_n obtains if we use the information number to derive a standard error for $\hat{p}_n$.

If there was interest in testing the two-sided hypothesis $H_0: p = p_0$ versus $H_1: p \neq p_0$, where $0 < p_0 < 1$ is a specified value, the above strategy is again applicable. However, in this case, there is an alternative approximate test. By the Central Limit Theorem, for any p, $0 < p < 1$,

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \to \mathrm{n}(0,1).$$

Therefore, if the null hypothesis is true, the statistic

$$Z_n' = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathrm{n}(0,1) \qquad \text{(approximately)}. \tag{10.3.4}$$

The approximate level α test rejects H_0 if $|Z_n'| > z_{\alpha/2}$.

In cases where both tests are applicable, for example, when testing $H_0: p = p_0$, it is not clear which test is to be preferred. The power functions (actual, not approximate) cross one another, so each test is more powerful in a certain portion of the parameter space. (Ghosh 1979) gives some insights into this problem. A related binomial controversy, that of the two-sample problem, is discussed by Robbins 1977 and Eberhardt and Fligner 1977. Two different test statistics for this problem are given in Exercise 10.31.)

Of course, any comparison of power functions is confounded by the fact that these are *approximate* tests and do not necessarily maintain level α. The use of a continuity correction (see Example 3.3.2) can help in this problem. In many cases, approximate procedures that use the continuity correction turn out to be *conservative*; that is, they maintain their nominal α level (see Example 10.4.6). ||

Equation (10.3.4) is a special case of another useful large-sample test, the *score test*. The *score statistic* is defined to be

$$S(\theta) = \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta) = \frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X}).$$

From (7.3.8) we know that, for all θ, $\mathrm{E}_\theta\, S(\theta) = 0$. In particular, if we are testing $H_0: \theta = \theta_0$ and if H_0 is true, then $S(\theta_0)$ has mean 0. Furthermore, from (7.3.10),

$$\mathrm{Var}_\theta\, S(\theta) = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})\right)^2\right) = -\,\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{X})\right) = I_n(\theta);$$

the information number is the variance of the score statistic. The test statistic for the score test is

$$Z_S = S(\theta_0)/\sqrt{I_n(\theta_0)}.$$

If H_0 is true, Z_S has mean 0 and variance 1. From Theorem 10.1.12 it follows that Z_S converges to a standard normal random variable if H_0 is true. Thus, the approximate level α score test rejects H_0 if $|Z_S| > z_{\alpha/2}$. If H_0 is composite, then $\hat{\theta}_0$, an estimate of θ assuming H_0 is true, replaces θ_0 in Z_S. If $\hat{\theta}_0$ is the restricted MLE, the restricted maximization might be accomplished using Lagrange multipliers. Thus, the score test is sometimes called the *Lagrange multiplier test.*

Example 10.3.6 (Binomial score test) Consider again the Bernoulli model from Example 10.3.5, and consider testing $H_0\colon p = p_0$ versus $H_1\colon p \neq p_0$. Straightforward calculations yield

$$S(p) = \frac{\hat{p}_n - p}{p(1-p)/n} \quad \text{and} \quad I_n(p) = \frac{n}{p(1-p)}.$$

Hence, the score statistic is

$$Z_S = \frac{S(p_0)}{\sqrt{I_n(p_0)}} = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}},$$

the same as (10.3.4). ||

One last class of approximate tests to be considered are robust tests (see Miscellanea 10.6.6). From Section 10.2, we saw that if $X_1, \ldots, X_n$ are iid from a location family and $\hat{\theta}_M$ is an M-estimator, then

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \to \mathrm{n}\left(0, \mathrm{Var}_{\theta_0}(\hat{\theta}_M)\right), \tag{10.3.5}$$

where $\mathrm{Var}_{\theta_0}(\hat{\theta}_M) = \frac{\mathrm{E}_{\theta_0}\psi(X-\theta_0)^2}{[\mathrm{E}_{\theta_0}\psi'(X-\theta_0)]^2}$ is the asymptotic variance. Thus, we can construct a "generalized" score statistic,

$$Z_{GS} = \sqrt{n}\frac{\hat{\theta}_M - \theta_0}{\sqrt{\mathrm{Var}_{\theta_0}(\hat{\theta}_M)}},$$

or a generalized Wald statistic,

$$Z_{GW} = \sqrt{n}\frac{\hat{\theta}_M - \theta_0}{\sqrt{\widehat{\mathrm{Var}_{\theta_0}}(\hat{\theta}_M)}},$$

where $\widehat{\mathrm{Var}_{\theta_0}}(\hat{\theta}_M)$ can be any consistent estimator. For example, we could use a bootstrap estimate of standard error, or simply substitute an estimator into (10.2.6) and use

$$\widehat{\mathrm{Var}_1}(\hat{\theta}_M) = \frac{\frac{1}{n}\sum_{i=1}^n[\psi(x_i - \hat{\theta}_M)]^2}{\left[\frac{1}{n}\sum_{i=1}^n \psi'(x_i - \hat{\theta}_M)\right]^2}. \tag{10.3.6}$$

The choice of variance estimate can be important; see Boos (1992) or Carroll, Ruppert, and Stefanski (1995, Appendix A.3) for guidance.

Example 10.3.7 (Tests based on the Huber estimator) If $X_1, \dots, X_n$ are iid from a pdf $f(x-\theta)$, where f is symmetric around 0, then for the Huber M-estimator using the ρ function in (10.2.2) and the ψ function (10.2.7), we have an asymptotic variance

$$\frac{\int_{-k}^{k} x^2 f(x)dx + k^2 P_0(|X|>k)}{[P_0(|X|\le k)]^2}. \tag{10.3.7}$$

Therefore, based on the asymptotic normality of the M-estimator, we can (for example) test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ at level α by rejecting H_0 if $|Z_{GS}| > z_{\alpha/2}$. To be a bit more practical, we will look at the approximate tests that use an estimated standard error. We will use the statistic Z_{GW}, but we will base our variance estimate on (10.3.7), that is

$$\widehat{\mathrm{Var}}_2(\hat{\theta}_M) =$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\theta}_M)^2 I(|x_i-\hat{\theta}_M|<k) + k^2\left(\frac{1}{n}\sum_{i=1}^{n} I(|x_i-\hat{\theta}_M|>k)\right)}{\left(1-\frac{1}{n}\sum_{i=1}^{n} I(|x_i-\hat{\theta}_M|<k)\right)^2}. \tag{10.3.8}$$

Also, we added a "naive" test, Z_N, that uses a simple variance estimate

$$\widehat{\mathrm{Var}}_3(\hat{\theta}_M) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\theta}_M)^2. \tag{10.3.9}$$

How do these tests fare? Analytical evaluation is difficult, but the small simulation in Table 10.3.1 shows that the $z_{\alpha/2}$ cutoffs are generally too small (neglecting to account for variation in the variance estimates), as the actual size is typically greater than the nominal size. However, there is consistency across a range of distributions, with the double exponential being the best case. (This last occurrence is not totally surprising, as the Huber estimator enjoys an optimality property against distributions with exponential tails; see Huber 1981, Chapter 4.) ||

10.4 Interval Estimation

As we have done in the previous two sections, we now explore some approximate and asymptotic versions of confidence sets. Our purpose is, as before, to illustrate some methods that will be of use in more complicated situations, methods that will get *some* answer. The answers obtained here are almost certainly not the best but are certainly not the worst. In many cases, however, they are the best that we can do.

We start, as previously, with approximations based on MLEs.

10.4.1 Approximate Maximum Likelihood Intervals

From the discussion in Section 10.1.2, and using Theorem 10.1.12, we have a general method to get an asymptotic distribution for a MLE. Hence, we have a general method to construct a confidence interval.

Table 10.3.1. *Power, at specified parameter values, of nominal $\alpha = .1$ tests based on Z_{GW} and Z_N, for a sample of size $n = 15$ (10,000 simulations)*

	Underlying pdf							
	Normal		t_5		Logistic		Double exponential	
	Z_{GW}	Z_N	Z_{GW}	Z_N	Z_{GW}	Z_N	Z_{GW}	Z_N
θ_0	.16	.16	.14	.13	.15	.15	.11	.09
$\theta_0 + .25\sigma$	.27	.29	.29	.27	.27	.27	.31	.26
$\theta_0 + .5\sigma$	.58	.60	.65	.63	.59	.60	.70	.64
$\theta_0 + .75\sigma$	.85	.87	.89	.89	.85	.87	.92	.90
$\theta_0 + 1\sigma$	.96	.97	.98	.97	.96	.97	.98	.98
$\theta_0 + 2\sigma$	1.	1.	1.	1.	1.	1.	1.	1.

If $X_1, \ldots, X_n$ are iid $f(x|\theta)$ and $\hat{\theta}$ is the MLE of θ, then from (10.1.7) the variance of a function $h(\hat{\theta})$ can be approximated by

$$\widehat{\text{Var}}(h(\hat{\theta})|\theta) \approx \frac{[h'(\theta)]^2|_{\theta=\hat{\theta}}}{-\frac{\partial^2}{\partial\theta^2} \log L(\theta|\mathbf{x})|_{\theta=\hat{\theta}}}.$$

Now, for a fixed but arbitrary value of θ, we are interested in the asymptotic distribution of

$$\frac{h(\hat{\theta}) - h(\theta)}{\sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)}}.$$

It follows from Theorem 10.1.12 and Slutsky's Theorem (Theorem 5.5.17) (see Exercise 10.33) that

$$\frac{h(\hat{\theta}) - h(\theta)}{\sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)}} \to \text{n}(0,1),$$

giving the approximate confidence interval

$$h(\hat{\theta}) - z_{\alpha/2}\sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)} \le h(\theta) \le h(\hat{\theta}) + z_{\alpha/2}\sqrt{\widehat{\text{Var}}(h(\hat{\theta})|\theta)}.$$

Example 10.4.1 (Continuation of Example 10.1.14) We have a random sample $X_1, \ldots, X_n$ from a Bernoulli(p) population. We saw that we could estimate the odds ratio $p/(1-p)$ by its MLE $\hat{p}/(1-\hat{p})$ and that this estimate has approximate variance

$$\widehat{\text{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right) \approx \frac{\hat{p}}{n(1-\hat{p})^3}.$$

We therefore can construct the approximate confidence interval

$$\frac{\hat{p}}{1-\hat{p}} - z_{\alpha/2}\sqrt{\widehat{\text{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)} \le \frac{p}{1-p} \le \frac{\hat{p}}{1-\hat{p}} + z_{\alpha/2}\sqrt{\widehat{\text{Var}}\left(\frac{\hat{p}}{1-\hat{p}}\right)}. \quad \|$$

A more restrictive form of the likelihood approximation, but one that, when applicable, gives better intervals, is based on the score statistic (see Section 10.3.2). The random quantity

$$Q(\mathbf{X}|\theta)=\frac{\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})}{\sqrt{-\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{X})\right)}} \tag{10.4.1}$$

has a $\mathrm{n}(0,1)$ distribution asymptotically as $n\to\infty$. Thus, the set

$$\left\{\theta\colon |Q(\mathbf{x}|\theta)|\le z_{\alpha/2}\right\} \tag{10.4.2}$$

is an approximate $1-\alpha$ confidence set. Notice that, applying results from Section 7.3.2, we have

$$\mathrm{E}_\theta(Q(\mathbf{X}|\theta))=\frac{\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})\right)}{\sqrt{-\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{X})\right)}}=0$$

and

$$\mathrm{Var}_\theta(Q(\mathbf{X}|\theta))=\frac{\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{X})\right)}{-\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta|\mathbf{X})\right)}=1, \tag{10.4.3}$$

and so this approximation exactly matches the first two moments of a $\mathrm{n}(0,1)$ random variable. Wilks (1938) proved that these intervals have an asymptotic optimality property; they are, asymptotically, the shortest in a certain class of intervals.

Of course, these intervals are not totally general and may not always be applicable to a function $h(\theta)$. We must be able to express (10.4.2) as a function of $h(\theta)$.

Example 10.4.2 (Binomial score interval) Again using a binomial example, if $Y=\sum_{i=1}^n X_i$, where each X_i is an independent Bernoulli(p) random variable, we have

$$\begin{aligned}Q(Y|p)&=\frac{\frac{\partial}{\partial p}\log L(p|Y)}{\sqrt{-\mathrm{E}_p\left(\frac{\partial^2}{\partial p^2}\log L(p|Y)\right)}}\\&=\frac{\frac{y}{p}-\frac{n-y}{1-p}}{\sqrt{\frac{n}{p(1-p)}}}\\&=\frac{\hat{p}-p}{\sqrt{p(1-p)/n}},\end{aligned}$$

where $\hat{p}=y/n$. From (10.4.2), an approximate $1-\alpha$ confidence interval is given by

$$\left\{p\colon\left|\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right|\le z_{\alpha/2}\right\}. \tag{10.4.4}$$

This is the interval that results from inverting the score statistic (see Example 10.3.6). To calculate this interval we need to solve a quadratic in p; see Example 10.4.6 for details. ||

In Section 10.3 we derived another likelihood test based on the fact that $-2\log\lambda(\mathbf{X})$ has an asymptotic chi squared distribution. This suggests that if $X_1,\ldots,X_n$ are iid $f(x|\theta)$ and $\hat\theta$ is the MLE of θ, then the set

$$\left\{\theta : -2\log\left(\frac{L(\theta|\mathbf{x})}{L(\hat\theta|\mathbf{x})}\right) \le \chi^2_{1,\alpha}\right\} \tag{10.4.5}$$

is an approximate $1-\alpha$ confidence interval. This is indeed the case and gives us yet another approximate likelihood interval.

Of course, (10.4.5) is just the highest likelihood region (9.2.7) that we originally derived by inverting the LRT statistic. However, we now have an automatic way of attaching an approximate confidence level.

Example 10.4.3 (Binomial LRT interval) For $Y=\sum_{i=1}^n X_i$, where each X_i is an independent Bernoulli(p) random variable, we have the approximate $1-\alpha$ confidence set

$$\left\{p : -2\log\left(\frac{p^y(1-p)^{n-y}}{\hat p^y(1-\hat p)^{n-y}}\right) \le \chi^2_{1,\alpha}\right\}.$$

This confidence set, along with the intervals based on the score and Wald tests, are compared in Example 10.4.7. ||

10.4.2 Other Large-Sample Intervals

Most approximate confidence intervals are based on either finding approximate (or asymptotic) pivots or inverting approximate level α test statistics. If we have any statistics W and V and a parameter θ such that, as $n\to\infty$,

$$\frac{W-\theta}{V}\to \mathrm{n}(0,1),$$

then we can form the approximate confidence interval for θ given by

$$W - z_{\alpha/2}V \le \theta \le W + z_{\alpha/2}V,$$

which is essentially a Wald-type interval. Direct application of the Central Limit Theorem, together with Slutsky's Theorem, will usually give an approximate confidence interval. (Note that the approximate maximum likelihood intervals of the previous section all reflect this strategy.)

Example 10.4.4 (Approximate interval) If $X_1,\ldots,X_n$ are iid with mean μ and variance σ^2, then, from the Central Limit Theorem,

$$\frac{\bar X-\mu}{\sigma/\sqrt{n}}\to \mathrm{n}(0,1).$$

Table 10.4.1. *Confidence coefficient for the pivotal interval (10.4.6), $n = 15$, based on 10,000 simulations*

	Underlying pdf			
Nominal level	Normal	t_5	Logistic	Double Exponential
$1-\alpha = .90$	.879	.864	.880	.876
$1-\alpha = .95$	.931	.924	.931	.933

Moreover, from Slutsky's Theorem, if $S^2 \to \sigma^2$ in probability, then

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \to \mathrm{n}(0,1),$$

giving the approximate $1-\alpha$ confidence interval

$$\bar{x} - z_{\alpha/2}s/\sqrt{n} \le \mu \le \bar{x} + z_{\alpha/2}s/\sqrt{n}. \tag{10.4.6}$$

To see how good the approximation is, we present a small simulation to calculate the exact coverage probability of the approximate interval for a variety of pdfs. Note that, since the interval is pivotal, the coverage probability does not depend on the parameter value; it is constant and hence is the confidence coefficient. We see from Table 10.4.1 that even for a sample size as small as $n = 15$, the pivotal confidence interval does a reasonable job, but clearly does not achieve the nominal confidence coefficient. This is, no doubt, due to the optimism of using the $z_{\alpha/2}$ cutoff, which does not account for the variability in S. As the sample size increases, the approximation will improve. ||

In the above example, we could get an approximate confidence interval without specifying the form of the sampling distribution. We should be able to do better when we do specify the form.

Example 10.4.5 (Approximate Poisson interval) If $X_1, \ldots, X_n$ are iid Poisson(λ), then we know that

$$\frac{\bar{X}-\lambda}{S/\sqrt{n}} \to \mathrm{n}(0,1).$$

However, this is true even if we did not sample from a Poisson population. Using the Poisson assumption, we know that $\mathrm{Var}(X) = \lambda = E\bar{X}$ and $\bar{X}$ is a good estimator of λ (see Example 7.3.12). Thus, using the Poisson assumption, we could also get an approximate confidence interval from the fact that

$$\frac{\bar{X}-\lambda}{\sqrt{\bar{X}/n}} \to \mathrm{n}(0,1),$$

which is the interval that results from inverting the Wald test. We can use the Poisson assumption in another way. Since $\mathrm{Var}(X) = \lambda$, it follows that

$$\frac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \to \mathrm{n}(0,1),$$

resulting in the interval corresponding to the score test, which is also the likelihood interval of (10.4.2) and is best according to Wilks (1938) (see Exercise 10.40). ||

Generally speaking, a reasonable rule of thumb is to use as few estimates and as many parameters as possible in an approximation. This is sensible for a very simple reason. Parameters are fixed and do not introduce any added variability into an approximation, while each statistic brings more variability along with it.

Example 10.4.6 (More on the binomial score interval) For a random sample $X_1, \ldots, X_n$ from a Bernoulli(p) population, we saw in Example 10.3.5 that, as $n \to \infty$, both

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \quad \text{and} \quad \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$$

converge in distribution to a standard normal random variable, where $\hat{p} = \sum x_i/n$. In Example 10.3.5 we saw that we could base tests on either approximation, with the former being the Wald test and the latter the score test. We also know that we can use either approximation to form a confidence interval for p. However, the score test approximation (with fewer statistics and more parameter values) will give the interval (10.4.4) from Example 10.4.2, which is the asymptotically optimal one; that is,

$$\left\{ p\colon \left| \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \right| \le z_{\alpha/2} \right\}$$

is the better approximate interval.

It is not immediately clear what this interval looks like, but we can explicitly solve for the set of values. If we square both sides and rearrange terms, we are looking for the set of values of p that satisfy

$$\left\{ p\colon (\hat{p}-p)^2 \le z_{\alpha/2}^2 \frac{p(1-p)}{n} \right\}.$$

This inequality is a quadratic in p, which can be put in a more familiar form through some further rearrangement:

$$\left\{ p\colon \left(1 + \frac{z_{\alpha/2}^2}{n}\right) p^2 - \left(2\hat{p} + \frac{z_{\alpha/2}^2}{n}\right) p + \hat{p}^2 \le 0 \right\}.$$

Since the coefficient of p^2 in the quadratic is positive, the quadratic opens upward and, thus, the inequality is satisfied if p lies between the two roots of the quadratic. These two roots are

$$\text{(10.4.7)} \qquad \frac{2\hat{p} + z_{\alpha/2}^2/n \pm \sqrt{(2\hat{p} + z_{\alpha/2}^2/n)^2 - 4\hat{p}^2(1 + z_{\alpha/2}^2/n)}}{2(1 + z_{\alpha/2}^2/n)},$$

and the roots define the endpoints of the confidence interval for p. Although the expressions for the roots are somewhat nasty, the interval is, in fact, a very good

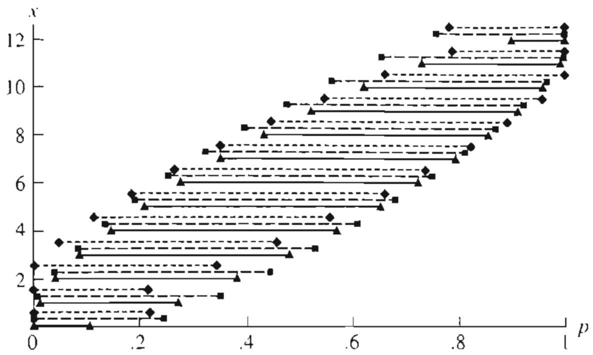

Figure 10.4.1. *Intervals for a binomial proportion from the LRT procedure (solid lines), the score procedure (long dashes), and the modified Wald procedure (short dashes)*

interval for p. The interval can be further improved, however, by using a continuity correction (see Example 3.3.2). To do this, we would solve two separate quadratics (see Exercise 10.45),

$$\left|\frac{\hat{p}+\frac{1}{2n}-p}{\sqrt{p(1-p)/n}}\right| \le z_{\alpha/2}, \quad \text{(larger root = upper interval endpoint)}$$

$$\left|\frac{\hat{p}-\frac{1}{2n}-p}{\sqrt{p(1-p)/n}}\right| \le z_{\alpha/2}. \quad \text{(smaller root = lower interval endpoint)}$$

At the endpoints there are obvious modifications. If $\sum x_i = 0$, then the lower interval endpoint is taken to be 0, while, if $\sum x_i = n$, then the upper interval endpoint is taken to be 1. See Blyth (1986) for some good approximations. ||

We now have seen three intervals for a binomial proportion: those based on the Wald and score statistics and the LRT interval of Example 10.4.3. Typically the Wald interval is least preferred, but it would be interesting to compare all three.

Example 10.4.7 (Comparison of binomial intervals) For $Y = \sum_{i=1}^n X_i$, $X_1, \ldots, X_n$ iid from a Bernoulli(p) population, the Wald interval is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{10.4.8}$$

the score interval (with continuity correction) is described in Example 10.4.6, and the approximate LRT is given in Example 10.4.3. To compare them, we look at an example.

For $n = 12$, Figure 10.4.1 shows the realized intervals for the three procedures. The LRT procedure produces the shortest intervals, and the score procedure the longest. For this picture we have made two modifications to the Wald interval. First, at $y = 0$ the unmodified interval is $(0, 0)$, so we have changed the upper endpoint to

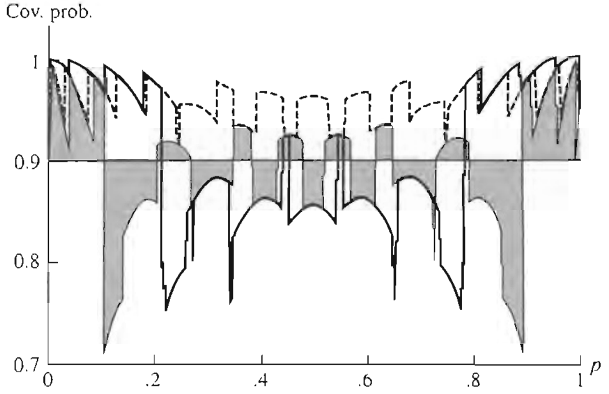

Figure 10.4.2. *Coverage probabilities for nominal .9 confidence procedures for a binomial proportion: the LRT procedure (thin solid lines, shaded grey), the score procedure (dashes), and the modified Wald procedure (thick solid lines)*

$1-(\alpha/2)^{1/n}$, with a similar modification to the lower interval at $y=n$. Also, there are some instances where the endpoints of the Wald interval go outside $[0,1]$; these have been truncated.

In Figure 10.4.2 we see that the longer length of the score interval is reflected in its higher coverage probability. Indeed, the score interval is the only one (of the three) that maintains a coverage probability above .9, and hence is the only interval with confidence coefficient .9. The LRT and Wald intervals appear to be too short, and their coverage probabilities are too far below .9 for them to be acceptable. Of course, their performance will improve with increasing n.

So it appears that the continuity corrected score interval, although longer, is the interval of choice for small n (but see Exercise 10.44 for another option). The LRT and Wald procedures produce intervals that are just too short for small n, with the Wald interval also suffering from endpoint maladies. ||

As we did in Section 10.3.2, we briefly look at intervals based on robust estimators.

Example 10.4.8 (Intervals based on the Huber estimator) In a development similar to Example 10.3.7, we can form asymptotic confidence intervals based on the Huber M-estimator. If $X_1,\ldots,X_n$ are iid from a pdf $f(x-\theta)$, where f is symmetric around 0, we have the approximate interval for θ,

$$\hat{\theta}_M \pm z_{\alpha/2}\sqrt{\frac{\operatorname{Var}(\hat{\theta}_M)}{n}},$$

where $\operatorname{Var}(\hat{\theta}_M)$ is given by (10.3.7). Now we replace $\operatorname{Var}(\hat{\theta}_M)$ by the estimates (10.3.8) and (10.3.9) to get Wald-type intervals. To evaluate these intervals, we produce a table similar to Table 10.4.1. It is interesting that, with the exception of the double exponential distribution, the intervals in Table 10.4.2 fare worse than those in Table

Table 10.4.2. *Confidence coefficients for nominal* $1-\alpha = .9$ *intervals based on Huber's* M*-estimator,* $n = 15$*, based on 10,000 simulations*

	Underlying pdf			
Nominal level	Normal	t_5	Logistic	Double exponential
Variance estimate (10.3.8)	.844	.856	.855	.889
Variance estimate (10.3.9)	.837	.867	.855	.910

10.4.1, which are based on the usual mean and variance. We do not have a good explanation for this, except to once again blame it on the overoptimism of the $z_{\alpha/2}$ cutoff. ||

Thus far, all of the approximations mentioned have been based on letting $n \to \infty$. However, there are other situations where we might use approximate intervals. In Example 9.2.17 we needed approximations as the parameter went to infinity. In another situation, in Example 2.3.13 we saw that for certain parameter configurations, the Poisson distribution can be used to approximate the binomial. This suggests that, if such a parameter configuration is believed to be likely, then an approximate binomial interval can be based on the Poisson distribution. In that spirit we illustrate the following somewhat unusual case.

Example 10.4.9 (Negative binomial interval) Let $X_1, \ldots, X_n$ be iid negative binomial(r,p). We assume that r is known and we are interested in a confidence interval for p. Using the fact that $Y = \sum X_i \sim$ negative binomial(nr,p), we can form intervals in a number of ways. Using a variation of the binomial–F distribution relationship, we can form an exact confidence interval (see Exercise 9.22) or we can use a normal approximation (see Exercise 10.41). There is another approximation that does not rely on large n, but rather small p.

In Exercise 2.38 it is established that, as $p \to 0$,

$$2pY \to \chi^2_{2nr} \quad \text{in distribution.}$$

So, for small p, $2pY$ is a pivot! Using this fact, we can construct a pivotal $1-\alpha$ confidence interval, valid for small p:

$$\left\{p\colon \frac{\chi^2_{2nr,1-\alpha/2}}{2y} \le p \le \frac{\chi^2_{2nr,\alpha/2}}{2y}\right\}.$$

Details are in Exercise 10.47. ||

10.5 Exercises

10.1 A random sample $X_1, \ldots, X_n$ is drawn from a population with pdf

$$f(x|\theta) = \frac{1}{2}(1+\theta x), \quad -1 < x < 1, \quad -1 < \theta < 1.$$

Find a consistent estimator of θ and show that it is consistent.

10.2 Prove Theorem 10.1.5.

10.3 A random sample $X_1, \ldots, X_n$ is drawn from a population that is $\mathrm{n}(\theta, \theta)$, where $\theta > 0$.

(a) Show that the MLE of $\theta, \hat{\theta}$, is a root of the quadratic equation $\theta^2 + \theta - W = 0$, where $W = (1/n)\sum_{i=1}^n X_i^2$, and determine which root equals the MLE.

(b) Find the approximate variance of $\hat{\theta}$ using the techniques of Section 10.1.3.

10.4 A variation of the model in Exercise 7.19 is to let the random variables $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \beta X_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $X_1, \ldots, X_n$ are independent $\mathrm{n}(\mu, \tau^2)$ random variables, $\epsilon_1, \ldots, \epsilon_n$ are iid $\mathrm{n}(0, \sigma^2)$, and the Xs and ϵs are independent. Exact variance calculations become quite difficult, so we might resort to approximations. In terms of μ, τ^2, and σ^2, find approximate means and variances for

(a) $\sum X_i Y_i / \sum X_i^2$.

(b) $\sum Y_i / \sum X_i$.

(c) $\sum (Y_i / X_i)/n$.

10.5 For the situation of Example 10.1.8 show that for $T_n = \sqrt{n}/\bar{X}_n$:

(a) $\mathrm{Var}\,(T_n) = \infty$.

(b) If $\mu \neq 0$ and we delete the interval $(-\delta, \delta)$ from the sample space, thenVar $(T_n) < \infty$.

(c) If $\mu \neq 0$, the probability content of the interval $(-\delta, \delta)$ approaches 0.

10.6 For the situation of Example 10.1.10 show that

(a) $\mathrm{E}Y_n = 0$ and $\mathrm{Var}(Y_n) = p_n + (1 - p_n)\sigma_n^2$.

(b) $P(Y_n < a) \to P(Z < a)$, and hence $Y_n \to \mathrm{n}(0, 1)$ (recall that $p_n \to 1$, $\sigma_n \to \infty$, and $(1 - p_n)\sigma_n^2 \to \infty$).

10.7 In the proof of Theorem 10.1.6 it was shown that the MLE $\hat{\theta}$ is an asymptotically efficient estimator of θ. Show that if $\tau(\theta)$ is a continuous function of θ, then $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

10.8 Finish the proof of Theorem 10.1.6 by establishing the two convergence results in (10.1.6).

(a) Show that

$$\frac{1}{\sqrt{n}} l'(\theta_0|\mathbf{X}) = \sqrt{n}\left[\frac{1}{n}\sum_i W_i\right],$$

where $W_i = \frac{\frac{d}{d\theta} f(X_i|\theta)}{f(X_i|\theta)}$ has mean 0 and variance $I(\theta_0)$. Now use the Central Limit Theorem to establish the convergence to $\mathrm{n}[0, I(\theta_0)]$.

(b) Show that

$$-\frac{1}{n} l''(\theta_0|\mathbf{X}) = \frac{1}{n}\sum_i W_i^2 - \frac{1}{n}\sum_i \frac{\frac{d^2}{d\theta^2} f(X_i|\theta)}{f(X_i|\theta)}$$

and that the mean of the first piece is $I(\theta_0)$ and the mean of the second piece is 0. Apply the WLLN.

10.9 Suppose that $X_1, \ldots, X_n$ are iid Poisson(λ). Find the best unbiased estimator of

(a) $e^{-\lambda}$, the probability that $X = 0$.

(b) $\lambda e^{-\lambda}$, the probability that $X = 1$.

(c) For the best unbiased estimators of parts (a) and (b), calculate the asymptotic relative efficiency with respect to the MLE. Which estimators do you prefer? Why?

(d) A preliminary test of a possible carcinogenic compound can be performed by measuring the mutation rate of microorganisms exposed to the compound. An experimenter places the compound in 15 petri dishes and records the following number of mutant colonies:

$$10,\ 7,\ 8,\ 13,\ 8,\ 9,\ 5,\ 7,\ 6,\ 8,\ 3,\ 6,\ 6,\ 3,\ 5.$$

Estimate $e^{-\lambda}$, the probability that no mutant colonies emerge, and $\lambda e^{-\lambda}$, the probability that one mutant colony will emerge. Calculate both the best unbiased estimator and the MLE.

10.10 Continue the calculations of Example 10.1.15, where the properties of the estimator of $p(1-p)$ were examined.

(a) Show that, if $p \neq 1/2$, the MLE $\hat{p}(1-\hat{p})$ is asymptotically efficient.

(b) If $p = 1/2$, use Theorem 5.5.26 to find a limiting distribution of $\hat{p}(1-\hat{p})$.

(c) Calculate the exact expression for $\text{Var}[\hat{p}(1-\hat{p})]$. Is the reason for the failure of the approximations any clearer?

10.11 This problem will look at some details and extensions of the calculation in Example 10.1.18.

(a) Reproduce Figure 10.1.1, calculating the ARE for known β. (You can follow the calculations in Example A.0.7, or do your own programming.)

(b) Verify that the ARE$(\bar{X}, \hat{\mu})$ comparison is the same whether β is known or unknown.

(c) For estimation of β with known μ, show that the method of moment estimate and MLEs are the same. (It may be easier to use the (α, β) parameterization.)

(d) For estimation of β with unknown μ, the method of moment estimate and MLEs are not the same. Compare these estimates using asymptotic relative efficiency, and produce a figure like Figure 10.1.1, where the different curves correspond to different values of μ.

10.12 Verify that the superefficient estimator d_n of Miscellanea 10.6.1 is asymptotically normal with variance $v(\theta) = 1$ when $\theta \neq 0$ and $v(\theta) = a^2$ when $\theta = 0$. (See Lehmann and Casella 1998, Section 6.2, for more on superefficient estimators.)

10.13 Refer to Example 10.1.19.

(a) Verify that the bootstrap mean and variance of the sample 2, 4, 9, 12 are 6.75 and 3.94, respectively.

(b) Verify that 6.75 is the mean of the original sample.

(c) Verify that, when we divide by n instead of $n-1$, the bootstrap variance of the mean, and the usual estimate of the variance of the mean are the same.

(d) Show how to calculate the bootstrap mean and standard error using the $\binom{4+4-1}{4} =$ 35 distinct possible resamples.

(e) Establish parts (b) and (c) for a general sample $X_1, X_2, \ldots, X_n$.

10.14 In each of the following situations we will look at the parametric and nonparametric bootstrap. Compare the estimates, and discuss advantages and disadvantages of the methods.

(a) Referring to Example 10.1.22, estimate the variance of S^2 using a nonparametric bootstrap.

(b) In Example 5.6.6 we essentially did a parametric bootstrap of the distribution of S^2 from a Poisson sample. Use the nonparametric bootstrap to provide an alternative histogram of the distribution.

(c) In Example 10.1.18 we looked at the problem of estimating a gamma mean. Suppose that we have a random sample

$$0.28, 0.98, 1.36, 1.38, 2.4, 7.42$$

from a gamma(α, β) distribution. Estimate the mean and variance of the distribution using maximum likelihood and bootstrapping.

10.15 Use the Law of Large Numbers to show that $\text{Var}^*_B(\hat{\theta})$ of (10.1.11) converges to $\text{Var}^*(\hat{\theta})$ of (10.1.10) as $B \to \infty$.

10.16 For the situation of Example 10.1.21, if we observed that $\hat{p} = 1/2$, we might use a variance estimate from Theorem 5.5.26. Show that this variance estimate would be equal to $2[\text{Var}(\hat{p})]^2$.

(a) If we observe $\hat{p} = 11/24$, verify that this variance estimate is .00007.

(b) Using simulation, calculate the "exact variance" of $\hat{p}(1 - \hat{p})$ when $n = 24$ and $p = 11/24$. Verify that it is equal to .00529.

(c) Why do you think the Delta Method is so bad in this case? Might the second-order Delta Method do any better? What about the bootstrap estimate?

10.17 Efron (1982) analyzes data on law school admission, with the object being to examine the correlation between the LSAT (Law School Admission Test) score and the first-year GPA (grade point average). For each of 15 law schools, we have the pair of data points (average LSAT, average GPA):

(576, 3.39)	(635, 3.30)	(558, 2.81)	(578, 3.03)	(666, 3.44)
(580, 3.07)	(555, 3.00)	(661, 3.43)	(651, 3.36)	(605, 3.13)
(653, 3.12)	(575, 2.74)	(545, 2.76)	(572, 2.88)	(594, 2.96)

(a) Calculate the correlation coefficient between LSAT score and GPA.

(b) Use the nonparametric bootstrap to estimate the standard deviation of the correlation coefficient. Use $B = 1000$ resamples, and also plot them in a histogram.

(c) Use the parametric bootstrap to estimate the standard deviation of the correlation coefficient. Assume that (LSAT, GRE) has a bivariate normal distribution, and estimate the five parameters. Then generate 1000 samples of 15 pairs from this bivariate normal distribution.

(d) If (X, Y) are bivariate normal with correlation coefficient ρ and sample correlation r, then the Delta Method can be used to show that

$$\sqrt{n}(r - \rho) \to \text{n}(0, (1 - \rho^2)^2).$$

Use this fact to estimate the standard deviation of r. How does it compare to the bootstrap estimates? Draw an approximate pdf of r.

(e) Fisher's z-transformation is a *variance-stabilizing* transformation for the correlation coefficient (see Exercise 11.4). If (X, Y) are bivariate normal with correlation coefficient ϱ and sample correlation r, then

$$\frac{1}{2}\left[\log\left(\frac{1+r}{1-r}\right) - \log\left(\frac{1+\rho}{1-\rho}\right)\right]$$

is approximately normal. Use this fact to draw an approximate pdf of r. (Establishing the normality result in part (q)d involves some tedious matrix calculations; see Lehmann and Casella 1998, Example 6.5). The z-transformation of part (q)e yields faster convergence to normality that the Delta Method of part (q)d. Diaconis and Holmes 1994 do an exhaustive bootstrap for this problem, enumerating all 77, 558, 760 correlation coefficients.)

10.18 For the situation of Exercise 10.2.1, that is, if $X_1, X_2, \dots, X_n$ are iid, where $X_i \sim n(\mu, \sigma^2)$ with probability $1-\delta$ and $X_i \sim f(x)$ with probability δ, where $f(x)$ is any density with mean θ and variance τ^2, show that

$$\mathrm{Var}(\bar{X}) = (1-\delta)\frac{\sigma^2}{n} + \delta\frac{\tau^2}{n} + \frac{\delta(1-\delta)(\theta-\mu)^2}{n}.$$

Also deduce that contamination with a Cauchy pdf will always result in an infinite variance. (*Hint*: Write this mixture model as a hierarchical model. Let $Y = 0$ with probability $1-\delta$ and $Y = 1$ with probability δ. Then $\mathrm{Var}(X_i) = \mathrm{E}[\mathrm{Var}(X_i)|Y] + \mathrm{Var}(\mathrm{E}[X_i|Y])$.)

10.19 Another way in which underlying assumptions can be violated is if there is correlation in the sampling, which can seriously affect the properties of the sample mean. Suppose we introduce correlation in the case discussed in Exercise 10.2.1; that is, we observe $X_1, \dots, X_n$, where $X_i \sim \mathrm{n}(\theta, \sigma^2)$, but the X_is are no longer independent.

(a) For the equicorrelated case, that is, $\mathrm{Corr}(X_i, X_j) = \rho$ for $i \neq j$, show that

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} + \frac{n-1}{n}\rho\sigma^2,$$

so $\mathrm{Var}(\bar{X}) \not\to 0$ as $n \to \infty$.

(b) If the X_is are observed through time (or distance), it is sometimes assumed that the correlation decreases with time (or distance), with one specific model being $\mathrm{Corr}(X_i, X_j) = \rho^{|i-j|}$. Show that in this case

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

so $\mathrm{Var}(\bar{X}) \to 0$ as $n \to \infty$. (See Miscellanea 5.8.2 for another effect of correlation.)

(c) The correlation structure in part (b) arises in an *autoregressive AR(1) model*, where we assume that $X_{i+1} = \rho X_i + \delta_i$, with δ_i iid $\mathrm{n}(0,1)$. If $|\rho| < 1$ and we define $\sigma^2 = 1/(1-\rho^2)$, show that $\mathrm{Corr}(X_1, X_i) = \rho^{i-1}$.

10.20 Refer to Definition 10.2.2 about breakdown values.

(a) If $T_n = \bar{X}_n$, the sample mean, show that $b = 0$.

(b) If $T_n = M_n$, the sample median, show that $b = .5$.

An estimator that "splits the difference" between the mean and the median in terms of sensitivity is the α-*trimmed* mean, $0 < \alpha < \frac{1}{2}$, defined as follows. $\bar{X}_n^\alpha$, the α-trimmed mean, is computed by deleting the αn smallest observations and the αn largest observations, and taking the arithmetic mean of the remaining observations.

(c) If $T_n = \bar{X}_n^\alpha$, the α-trimmed mean of the sample, $0 < \alpha < \frac{1}{2}$, show that $0 < b < \frac{1}{2}$.

10.21 The breakdown performance of the mean and the median continues with their scale estimate counterparts. For a sample $X_1, \dots, X_n$:

(a) Show that the breakdown value of the sample variance $S^2 = \sum(X_i - \bar{X})^2/(n-1)$ is 0.

(b) A robust alternative is the *median absolute deviation*, or MAD, the median of $|X_1 - \mathrm{M}|, |X_2 - \mathrm{M}|, \ldots, |X_n - \mathrm{M}|$, where M is the sample median. Show that this estimator has a breakdown value of 50%.

10.22 This exercise will look at some of the details of Example 10.2.3.

(a) Verify that, if n is odd, then

$$P\left(\sqrt{n}(M_n - \mu) \le a\right) = P\left(\frac{\sum_i Y_i - np_n}{\sqrt{np_n(1-p_n)}} \ge \frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}\right).$$

(b) Verify that $p_n \to p = F(\mu) = 1/2$ and

$$\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} \to -2aF'(\mu) = -2af(\mu).$$

(*Hint*: Establish that $\frac{(n+1)/2-np_n}{\sqrt{n}}$ is the limit form of a derivative.)

(c) Explain how to go from the statement that

$$P\left(\sqrt{n}(M_n - \mu) \le a\right) \to P(Z \ge -2af(\mu))$$

to the conclusion that $\sqrt{n}(M_n - \mu)$ is asymptotically normal with mean 0 and variance $1/[2f(\mu)]^2$.

(Note that the CLT would directly apply only if p_n did not depend on n. As it does, more work needs to be done to rigorously conclude limiting normality. When the work is done, the result is as expected.)

10.23 In this exercise we will further explore the ARE of the median to the mean, $\mathrm{ARE}(M_n, \bar{X})$.

(a) Verify the three AREs given in Example 10.2.4.

(b) Show that $\mathrm{ARE}(M_n, \bar{X})$ is unaffected by scale changes. That is, it doesn't matter whether the underlying pdf is $f(x)$ or $(1/\sigma)f(x/\sigma)$.

(c) Calculate $\mathrm{ARE}(M_n, \bar{X})$ when the underlying distribution is Student's t with ν degrees of freedom, for $\nu = 3, 5, 10, 25, 50, \infty$. What can you conclude about the ARE and the tails of the distribution?

(d) Calculate $\mathrm{ARE}(M_n, \bar{X})$ when the underlying pdf is the *Tukey model*

$$X \sim \begin{cases} \mathrm{n}(0,1) & \text{with probability } 1-\delta \\ \mathrm{n}(0,\sigma^2) & \text{with probability } \delta. \end{cases}$$

Calculate the ARE for a range of δ and σ. What can you conclude about the relative performance of the mean and the median?

10.24 Assuming that θ_0 satisfies $\mathrm{E}_{\theta_0}\psi(X - \theta_0) = 0$, show that (10.2.4) and (10.2.5) imply (10.2.6).

10.25 If $f(x)$ is a pdf symmetric around 0 and ρ is a symmetric function, show that $\int \psi(x-\theta)f(x-\theta)\,dx = 0$, where $\psi = \rho'$. Show that this then implies that if $X_1, \ldots, X_n$ are iid from $f(x-\theta)$ and $\hat{\theta}_M$ is the minimizer of $\sum_i \rho(x_i - \theta)$, then $\hat{\theta}_M$ is asymptotically normal with mean equal to the true value of θ.

10.26 Here we look at some details in the calculations in Example 10.2.6.

(a) Verify the expressions for $\mathrm{E}_\theta\psi'(X-\theta)$ and $\mathrm{E}_\theta[\psi(X-\theta)]^2$, and hence verify the formula for the variance of $\hat{\theta}_M$.

(b) When calculating the expected value of ψ', we noted that ψ was not differentiable, but we could work with the differentiable portion. Another approach is to realize that the expected value of ψ is differentiable, and that in (10.2.5) we could write

$$\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\theta_0)\to\frac{d}{d\theta}\mathrm{E}\theta_0\psi(X-\theta)\Big|_{\theta=\theta_0}.$$

Show that this is the same limit as in (10.2.5).

10.27 Consider the situation of Example 10.6.2.

(a) Verify that $IF(\bar{X},x)=x-\mu$.

(b) For the median we have $T(F)=m$ if $P(X\le m)=1/2$ or $m=F^{-1}(1/2)$. If $X\sim F_\delta$, show that

$$P(X\le a)=\begin{cases}(1-\delta)F(a) & \text{if } x>a\\ (1-\delta)F(a)+\delta & \text{otherwise}\end{cases}$$

and thus

$$T(F_\delta)=\begin{cases}F^{-1}\left(\frac{1}{2(1-\delta)}\right) & \text{if } x>F^{-1}\left(\frac{1}{2(1-\delta)}\right)\\ F^{-1}\left(\frac{1/2-\delta}{1-\delta}\right) & \text{otherwise.}\end{cases}$$

(c) Show that

$$\frac{1}{\delta}\left[F^{-1}\left(\frac{1}{2(1-\delta)}\right)-F^{-1}\left(\frac{1}{2}\right)\right]\to\frac{1}{2f(m)},$$

and complete the argument to calculate $IF(M,x)$.

(*Hint*: Write $a_\delta=F^{-1}\left(\frac{1}{2(1-\delta)}\right)$, and argue that the limit is $a'_\delta|_{\delta=0}$. This latter quantity can be calculated using implicit differentiation and the fact that $(1-\delta)^{-1}=2F(a_\delta)$.)

10.28 Show that if ρ is defined by (10.2.2), then both ρ and ρ' are continuous.

10.29 From (10.2.9) we know that an M-estimator can never be more efficient than a maximum likelihood estimator. However, we also know when it can be as efficient.

(a) Show that (10.2.9) is an equality if we choose $\psi(x-\theta)=cl'(\theta|x)$, where l is the log likelihood and c is a constant.

(b) For each of the following distributions, verify that the corresponding ψ functions give asymptotically efficient M-estimators.

(i) Normal: $f(x)=e^{-x^2/2}/(\sqrt{2\pi})$, $\psi(x)=x$

(ii) Logistic: $f(x)=e^{-x}/(1+e^{-x})^2$, $\psi(x)=\tanh(x)$, where $\tanh(x)$ is the *hyperbolic tangent*

(iii) Cauchy: $f(x)=[\pi(1+x^2)]^{-1}$, $\psi(x)=2x/(1+x^2)$

(iv) Least informative distribution:

$$f(x)=\begin{cases}Ce^{-x^2/2} & |x|\le c\\ Ce^{-c|x|+c^2/2} & |x|>c\end{cases}$$

with $\psi(x)=\max\{-c,\min(c,x)\}$ and C and c are constants.

(See Huber 1981, Section 3.5, for more details.)

10.30 For M-estimators there is a connection between the ψ function and the breakdown value. The details are rather involved (Huber 1981, Section 3.2) but they can be summarized as follows: If ψ is a bounded function, then the breakdown value of the associated M-estimator is given by

$$b^* = \frac{\eta}{1+\eta}, \text{ where } \eta = \min\left\{-\frac{\psi(-\infty)}{\psi(\infty)}, -\frac{\psi(\infty)}{\psi(-\infty)}\right\}.$$

(a) Calculate the breakdown value of the efficient M-estimators of Exercise 10.29. Which ones are both efficient and robust?

(b) Calculate the breakdown value of these other M-estimators

(i) The Huber estimator given by (10.2.1)

(ii) Tukey's biweight: $\psi(x) = x(c^2 - x^2)$ for $|x| \le c$ and 0 otherwise, where c is a constant

(iii) Andrew's sine wave: $\psi(x) = c\sin(x/c)$ for $|x| \le c\pi$ and 0 otherwise

(c) Evaluate the AREs of the estimators in part (b) with respect to the MLE when the underlying distribution is (i) normal and (ii) double exponential.

10.31 Binomial data gathered from more than one population are often presented in a *contingency table*. For the case of two populations, the table might look like this:

	Population 1	2	Total
Successes	S_1	S_2	$S = S_1 + S_2$
Failures	F_1	F_2	$F = F_1 + F_2$
Total	n_1	n_2	$n = n_1 + n_2$

where Population 1 is binomial(n_1, p_1), with S_1 successes and F_1 failures, and Population 2 is binomial(n_2, p_2), with S_2 successes and F_2 failures. A hypothesis that is usually of interest is

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \ne p_2.$$

(a) Show that a test can be based on the statistic

$$T = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(\hat{p}(1-\hat{p}))},$$

where $\hat{p}_1 = S_1/n_1$, $\hat{p}_2 = S_2/n_2$, and $\hat{p} = (S_1 + S_2)/(n_1 + n_2)$. Also, show that as $n_1, n_2 \to \infty$, the distribution of T approaches χ^2_1. (This is a special case of a test known as a *chi squared test of independence*.)

(b) Another way of measuring departure from H_0 is by calculating an *expected frequency table*. This table is constructed by conditioning on the marginal totals and filling in the table according to $H_0 : p_1 = p_2$, that is,

	Expected frequencies 1	2	Total
Successes	$\frac{n_1 S}{n_1 + n_2}$	$\frac{n_2 S}{n_1 + n_2}$	$S = S_1 + S_2$
Failures	$\frac{n_1 F}{n_1 + n_2}$	$\frac{n_2 F}{n_1 + n_2}$	$F = F_1 + F_2$
Total	n_1	n_2	$n = n_1 + n_2$

Using the expected frequency table, a statistic T^* is computed by going through the cells of the tables and computing

$$T^* = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{\left(S_1 - \frac{n_1 S}{n_1+n_2}\right)^2}{\frac{n_1 S}{n_1+n_2}} + \cdots + \frac{\left(F_2 - \frac{n_2 F}{n_1+n_2}\right)^2}{\frac{n_2 F}{n_1+n_2}}.$$

Show, algebraically, that $T^* = T$ and hence that T^* is asymptotically chi squared.

(c) Another statistic that could be used to test equality of p_1 and p_2 is

$$T^{**} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

Show that, under H_0, T^{**} is asymptotically n$(0,1)$, and hence its square is asymptotically χ^2_1. Furthermore, show that $(T^{**})^2 \neq T^*$.

(d) Under what circumstances is one statistic preferable to the other?

(e) A famous medical experiment was conducted by Joseph Lister in the late 1800s. Mortality associated with surgery was quite high, and Lister conjectured that the use of a disinfectant, carbolic acid, would help. Over a period of several years Lister performed 75 amputations with and without using carbolic acid. The data are given here:

		Carbolic acid used?	
		Yes	No
Patient lived?	Yes	34	19
	No	6	16

Use these data to test whether the use of carbolic acid is associated with patient mortality.

10.32 (a) Let $(X_1, \ldots, X_n) \sim$ multinomial$(m, p_1, \ldots, p_n)$. Consider testing $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$. A test that is often used, called *McNemar's Test*, rejects H_0 if

$$\frac{(X_1 - X_2)^2}{X_1 + X_2} > \chi^2_{1,\alpha}.$$

Show that this test statistic has the form (as in Exercise 10.31)

$$\sum_1^n \frac{(\text{observed} - \text{expected})^2}{\text{expected}},$$

where the X_is are the observed cell frequencies and the expected cell frequencies are the MLEs of mp_i, under the assumption that $p_1 = p_2$.

(b) McNemar's Test is often used in the following type of problem. Subjects are asked if they agree or disagree with a statement. Then they read some information about the statement and are asked again if they agree or disagree. The numbers of responses in each category are summarized in a 2×2 table like this:

		Before	
		Agree	Disagree
After	Agree	X_3	X_2
	Disagree	X_1	X_4

The hypothesis $H_0: p_1 = p_2$ states that the proportion of people who change from agree to disagree is the same as the proportion of people who change from disagree to agree. Another hypothesis that might be tested is that the proportion of those who initially agree and then change is the same as the proportion of those who initially disagree and then change. Express this hypothesis in terms of conditional probabilities and show that it is different from the above H_0. (This hypothesis can be tested with a χ^2 test like those in Exercise 10.31.)

10.33 Fill in the gap in Theorem 10.3.1. Use Theorem 10.1.12 and Slutsky's Theorem (Theorem 5.5.17) to show that $(\theta - \hat{\theta})/\sqrt{-l''(\hat{\theta}|\mathbf{x})} \to \mathrm{n}(0,1)$, and therefore $-2\log\lambda(\mathbf{X}) \to \chi^2_1$.

10.34 For testing $H_0: p = p_0$ versus $H_1: p \neq p_0$, suppose we observe $X_1, \ldots, X_n$ iid Bernoulli(p).

(a) Derive an expression for $-2\log\lambda(\mathbf{x})$, where $\lambda(\mathbf{x})$ is the LRT statistic.

(b) As in Example 10.3.2, simulate the distribution of $-2\log\lambda(\mathbf{x})$ and compare it to the χ^2 approximation.

10.35 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population.

(a) If μ is unknown and σ^2 is known, show that $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$ is a Wald statistic for testing $H_0: \mu = \mu_0$.

(b) If σ^2 is unknown and μ is known, find a Wald statistic for testing $H_0: \sigma = \sigma_0$.

10.36 Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population. Assume α is known and β is unknown. Consider testing $H_0: \beta = \beta_0$.

(a) What is the MLE of β?

(b) Derive a Wald statistic for testing H_0, using the MLE in both the numerator and denominator of the statistic.

(c) Repeat part (b) but using the sample standard deviation in the standard error.

10.37 Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population.

(a) If μ is unknown and σ^2 is known, show that $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$ is a score statistic for testing $H_0: \mu = \mu_0$.

(b) If σ^2 is unknown and μ is known, find a score statistic for testing $H_0: \sigma = \sigma_0$.

10.38 Let $X_1, \ldots, X_n$ be a random sample from a gamma(α, β) population. Assume α is known and β is unknown. Consider testing $H_0: \beta = \beta_0$. Derive a score statistic for testing H_0.

10.39 Expand the comparisons made in Example 10.3.7.

(a) Another test based on Huber's M-estimator would be one that used a variance estimate, based on (10.3.6). Examine the performance of such a test statistic, and comment on its desirability (or lack of) as an alternative to either (10.3.8) or (10.3.9).

(b) Another test based on Huber's M-estimator would be one that used a variance from a bootstrap calculation. Examine the performance of such a test statistic.

(c) A robust competitor to $\hat{\theta}_M$ is the median. Examine the performance of tests of a location parameter based on the median.

10.40 In Example 10.4.5 we saw that the Poisson assumption, together with the Central Limit Theorem, could be used to form an approximate interval based on the fact that

$$\frac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \to \mathrm{n}(0,1).$$

Show that this approximation is optimal according to Wilks (1938). That is, show that

$$\frac{\bar{X}-\lambda}{\sqrt{\lambda/n}} = \frac{\frac{\partial}{\partial\lambda}\log L(\lambda|\mathbf{X})}{\sqrt{-\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log L(\lambda|\mathbf{X})\right)}}.$$

10.41 Let $X_1,\ldots,X_n$ be iid negative binomial(r,p). We want to construct some approximate confidence intervals for the negative binomial parameters.

(a) Calculate Wilks' approximation (10.4.3) and show how to form confidence intervals with this expression.

(b) Find an approximate $1-\alpha$ confidence interval for the *mean* of the negative binomial distribution. Show how to incorporate the continuity correction into your interval.

(c) The aphid data of Exercise 9.23 can also be modeled using the negative binomial distribution. Construct an approximate 90% confidence interval for the aphid data using the results of part (b). Compare the interval to the Poisson-based intervals of Exercise 9.23.

10.42 Show that (10.4.5) is equivalent to the highest likelihood region (9.2.7) in that for any fixed α level, they will produce the same confidence set.

10.43 In Example 10.4.7, two modifications were made to the Wald interval.

(a) At $y=0$ the upper interval endpoint was changed to $1-(\alpha/2)^{1/n}$, and at $y=n$ the lower interval endpoint was changed to $(\alpha/2)^{1/n}$. Justify the choice of these endpoints. (*Hint*: see Section 9.2.3.)

(b) The second modification was to truncate all intervals to be within $[0,1]$. Show that this change, together with the one in part (a), results in an improvement over the original Wald interval.

10.44 Agresti and Coull (1998) "strongly recommend" the score interval for a binomial parameter but are concerned that a formula such as (10.4.7) might be a bit formidable for an elementary course in statistics. To produce a reasonable binomial interval with an easier formula, they suggest the following modification to the Wald interval: Add 2 successes and 2 failures; then use the original Wald formula (10.4.8). That is, use $\hat{p}=(y+2)/(n+4)$ instead of $\hat{p}=y/n$. Using both length and coverage probability, compare this interval to the binomial score interval. Do you agree that it is a reasonable alternative to the score interval?

(Samuels and Lu 1992 suggest another modification to the Wald interval based on sample sizes. Agresti and Caffo 2000 extend these improved approximate intervals to the two sample problem.)

10.45 Solve for the endpoints of the approximate binomial confidence interval, with continuity correction, given in Example 10.4.6. Show that this interval is wider than the corresponding interval without continuity correction, and that the continuity corrected interval has a uniformly higher coverage probability. (In fact, the coverage probability of the uncorrected interval does not maintain $1 - \alpha$; it dips below this level for some parameter values. The corrected interval does maintain a coverage probability greater than $1 - \alpha$ for all parameter values.)

10.46 Expand the comparisons made in Example 10.4.8.

(a) Produce a table similar to Table 10.4.2 that examines the robustness of intervals for a location parameter based on the median. (Intervals based on the mean are done in Table 10.4.1.)

(b) Another interval based on Huber's M-estimator would be one that used a variance from a bootstrap calculation. Examine the robustness of such an interval.

10.47 Let $X_1, \dots, X_n$ be iid negative binomial(r, p).

(a) Complete the details of Example 10.4.9; that is, show that for small p, the interval

$$\left\{p\colon \frac{\chi^2_{2nr,1-\alpha/2}}{2\sum x} \le p \le \frac{\chi^2_{2nr,\alpha/2}}{2\sum x}\right\}$$

is an approximate $1 - \alpha$ confidence interval.

(b) Show how to choose the endpoints in order to obtain a minimum length $1 - \alpha$ interval.

10.48 For the case of Fieller's confidence set (see Miscellanea 9.5.3), that is, given a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from a bivariate normal distribution with parameters $(\mu_X, \mu_Y, \sigma^2_X, \sigma^2_Y, \rho)$, find an approximate confidence interval for $\theta = \mu_Y/\mu_X$. Use the approximate moment calculations in Example 5.5.27 and apply the Central Limit Theorem.

10.6 Miscellanea

10.6.1 Superefficiency

Although the Cramér–Rao Lower Bound of Theorem 7.3.9 is a bona fide lower bound on the variance, the lower bound of Definition 10.1.11 and Theorem 10.1.6, which refers to the asymptotic variance, can be violated. An example of an estimator that beats the bound of Definition 10.1.11 was given by Hodges (see LeCam 1953).

If $X_1, \dots, X_n$ are iid $\mathrm{n}(\theta, 1)$, the Cramér–Rao Lower Bound for unbiased estimators of θ is $v(\theta) = 1/n$. The estimator

$$d_n = \begin{cases} \bar{X} & \text{if } |\bar{X}| \ge 1/n^{1/4} \\ a\bar{X} & \text{if } |\bar{X}| < 1/n^{1/4} \end{cases}$$

satisfies

$$\sqrt{n}(d_n - \theta) \to \mathrm{n}[0, v(\theta)],$$

in distribution, where $v(\theta) = 1$ when $\theta \ne 0$ and $v(\theta) = a^2$ when $\theta = 0$. If $a < 1$, inequality (7.2.5) is therefore violated at $\theta = 0$.

Although estimators such as d_n, called *superefficient*, can be constructed in some generality, they are more of a theoretical oddity than a practical concern. This is because the values of θ for which the variance goes below the bound are a set of Lebesgue measure 0. However, the existence of superefficient estimators serves to remind us to always be careful in our examination of assumptions for establishing properties of estimators (and to be careful in general!).

10.6.2 Suitable Regularity Conditions

The phrase "under suitable regularity conditions" is a somewhat abused phrase, as with enough assumptions we can probably prove whatever we want. However, "regularity conditions" are typically very technical, rather boring, and usually satisfied in most reasonable problems. But they are a necessary evil, so we should deal with them. To be complete, we present a set of regularity conditions that suffice to rigorously establish Theorems 10.1.6 and 10.1.12. These are not the most general conditions but are sufficiently general for many applications (with a notable exception being if the MLE is on the boundary of the parameter space). Be forewarned, the following is not for the fainthearted and can be skipped without sacrificing much in the way of understanding.

These conditions mainly relate to differentiability of the density and the ability to interchange differentiation and integration (as in the conditions for Theorem 7.3.9). For more details and generality, see Stuart, Ord, and Arnold (1999, Chapter 18), Ferguson (1996, Part 4), or Lehmann and Casella (1998, Section 6.3).

The following four assumptions are sufficient to prove Theorem 10.1.6, consistency of MLEs:

($A1$) We observe $X_1, \ldots, X_n$, where $X_i \sim f(x|\theta)$ are iid.

($A2$) The parameter is *identifiable*; that is, if $\theta \neq \theta'$, then $f(x|\theta) \neq f(x|\theta')$.

($A3$) The densities $f(x|\theta)$ have common support, and $f(x|\theta)$ is differentiable in θ.

($A4$) The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.

The next two assumptions, together with ($A1$)–($A4$) are sufficient to prove Theorem 10.1.12, asymptotic normality and efficiency of MLEs.

($A5$) For every $x \in \mathcal{X}$, the density $f(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ, and $\int f(x|\theta)\,dx$ can be differentiated three times under the integral sign.

($A6$) For any $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x|\theta)\right| \leq M(x) \quad \text{for all } x \in \mathcal{X}, \quad \theta_0 - c < \theta < \theta_0 + c,$$

with $\mathrm{E}_{\theta_0}[M(X)] < \infty$.

10.6.3 More on the Bootstrap

Theory

The theory behind the bootstrap is quite sophisticated, being based on *Edgeworth expansions.* These are expansions (in the spirit of Taylor series expansions) of distribution functions around a normal distribution. As an example, for $X_1, \ldots, X_n$ iid with density f with mean and variance μ and σ^2, an Edgeworth expansion of the cdf of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ is (Hall 1992, Equation 2.17)

$$P\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq w\right) = \Phi(w) + \phi(w)\left[\frac{-1}{6\sqrt{n}}\kappa(w^2-1) + Rn\right]$$

where nRn is bounded, Φ and ϕ are, respectively, the distribution and density function of a standard normal and $\kappa = E(X_1 - \mu)^3$ is the skewness. The first term in the expansion is the "usual" normal approximation, and as we add more terms, the expansion becomes more accurate.

The amazing thing about the bootstrap is that in some cases it automatically gets the second term in the expansion correct (hence achieving "second-order" accuracy). This does not happen in all cases, but one case in which it does occur is in bootstrapping a pivotal quantity. The Edgeworth theory of bootstrap is given a thorough treatment by Hall (1992); see also Shao and Tu (1995).

Practice

We have used the bootstrap only to calculate standard errors, but it has many other uses, with perhaps the most popular being the construction of confidence intervals. There are also many variations of the bootstrap developed for different situations. In particular, dealing with dependent data is somewhat delicate. For an introduction to the many uses of the bootstrap and much more, see Efron and Tibshirani (1993).

Limitations

Although the bootstrap is perhaps the single most important development in statistical methodology in recent times, it is not without its limitations and detractors. Outside of the cases of iid sampling and pivotal quantities, the bootstrap is less automatic but still can be extremely useful. For an interesting treatment of these issues, see LePage and Billard (1992) or Young (1994).

10.6.4 Influence Functions

A measure of catastrophic occurrences that does consider distributional properties is the *influence function*, which also measures the effect of an aberrant observation. The influence function has an interpretation as a derivative, which also turns out to have some interesting consequences.

The influence function of a statistic is actually calculated using its population counterpart. For example, the influence function of the sample mean is calculated using the population mean, as it seeks to measure the influence of perturbing the population. Similarly, the influence function of the sample median is calculated using the population median. To treat this idea in a consistent manner, it makes sense to think of an estimator as a function that operates on the cdf F or its sample counterpart, the empirical cdf (Definition 1.5.1) F_n. Such functions, that actually have other functions as arguments are known as *functionals*.

Note that for a sample $X_1, X_2, \ldots, X_n$, knowledge of the sample is equivalent to knowledge of the empirical cdf F_n, as F_n has a jump of size $1/n$ at each X_i. Thus, a statistic $T = T(X_1, X_2, \ldots, X_n)$ can equivalently be written $T(F_n)$. In doing so, we can then denote its population counterpart as $T(F)$.

Definition 10.6.1 For a sample $X_1, X_2, \ldots, X_n$ from a population with cdf F, the *influence function* of a statistic $T = T(F_n)$ at a point x is

$$IF(T, x) = \lim_{\delta \to 0} \frac{1}{\delta} \left[T(F_\delta) - T(F) \right],$$

where $X \sim F_\delta$ if

$$X \sim \begin{cases} F & \text{with probability } 1 - \delta \\ x & \text{with probability } \delta, \end{cases}$$

that is, F_δ is a mixture of F and a point x.

Example 10.6.2 (Influence functions of the mean and median) Suppose that we have a population with continuous cdf F and pdf f. Let μ denote the population mean and $\bar{X}$ the sample mean, and let $T(\cdot)$ be the functional that calculates the mean of a population. Thus $T(F_n) = \bar{X}$, $T(F) = \mu$, and

$$T(F_\delta) = (1 - \delta)\mu + \delta x,$$

so $IF(\bar{X}, x) = x - \mu$, and as x gets larger, its influence on $\bar{X}$ becomes increasingly large.

For the median M, we have (see Exercise 10.27)

$$IF(M, x) = \begin{cases} \frac{1}{2f(m)} & \text{if } x > m \\ -\frac{1}{2f(m)} & \text{otherwise.} \end{cases}$$

So, in contrast to the mean, the median has a bounded influence function. ||

Why is a bounded influence function important? To answer that, we look at the influence function of an M-estimator, of which the mean and median are special cases.

Let $\hat{\theta}_M$ be the M-estimator that is the solution to $\sum_i \psi(x_i - \theta) = 0$, where $X_1, \ldots, X_n$ are iid with cdf F. In Section 10.2.2 we saw that $\hat{\theta}_M$ will be a consistent estimator of the value θ_0 that satisfies $\mathrm{E}_{\theta_0} \psi(X - \theta_0) = 0$. The influence function of $\hat{\theta}_M$ is

$$IF(\hat{\theta}_M, x) = \frac{\psi(x - \theta_0)}{-\int \psi'(t - \theta_0) f(t)\, dt} = \frac{\psi(x - \theta_0)}{-\mathrm{E}_0(\psi'(X - \theta_0))}.$$

Now if we recall (10.2.6), we see that the expected square of the influence function gives the asymptotic variance of $\hat{\theta}_M$, that is,

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \to \mathrm{n}\left(0, \mathrm{E}_{\theta_0}[IF(\hat{\theta}_M, X)]^2\right)$$

in distribution. Thus, the influence function is directly related to the asymptotic variance.

10.6.5 Bootstrap Intervals

In Section 10.1.4 we saw the bootstrap to be a simple, general technique for obtaining a standard error of any statistic. In calculating these standard errors, we actually construct a distribution of a statistic, the *bootstrap* distribution. Then, a natural question arises. Is there a simple, general method of using the bootstrap distribution to make a confidence statement? The bootstrap can indeed be used to construct very good confidence intervals but, alas, the simplicity of application that it enjoys in calculating standard errors does not carry over into confidence intervals.

Methods based on using percentiles of the bootstrap distribution, or on bootstrapping a t-statistic (pivot), would seem to have potential for being generally applicable. However, Efron and Tibshirani (1993, Section 13.4) note that "neither of these intervals works well in general." Hall (1992, Chapter 3) prefers the t-statistic method and points out that bootstrapping a pivot is a superior technique in general.

Percentile and percentile-t intervals are only the tip of a vast development of bootstrap confidence intervals, many of which are excellent performers. However, we cannot summarize these procedures in one simple recipe; different problems will require different techniques.

10.6.6 Robust Intervals

Although we went into some detail about robustness of point estimators in Section 10.2, aside from Examples 10.3.7 and 10.4.8, we did not give much detail about robust tests and confidence intervals. This is not a comment on the importance of the subject but has more to do with space.

When we examined point estimators for robustness properties, the main concerns had to do with performance under deviations (both small and large) from the underlying assumptions. The same concerns are carried over to tests and intervals, with the expectation that robust point estimators will lead to robust tests and intervals. In particular, we would want robust tests to maintain power and robust intervals to maintain coverage over a range of deviations from the underlying model. That this is the case is indicated by the fact (see Staudte and Sheather 1990, Section 5.3.3) that the power function of a test can be related to the influence function of the point estimate on which it is based. Of course, this immediately implies that coverage properties of a related interval estimate can also be related to the influence function.

A nice introduction to robust tests, through estimating equations and score tests, is given by Boos (1992). The books by Staudte and Sheather (1990) and Hettmansperger and McKean (1998) are also excellent sources, as is the now-classic book by Huber (1981).