9 Asymptotic Approximations and Practical Asymptotic Tools

A point estimator is merely an educated guess about the true value of an unknown parameter. The utility of a point estimate without some idea of its accuracy is rather limited. That is why, we considered MSE and risk functions of point estimators in the previous chapters. MSE is itself a summary of the distribution of the point estimator. We would really like to know the distribution of our estimator, for the specific n that we have. Unfortunately, even for the simplest type of estimators, distributions are very hard to find. For example, we cannot find the distribution of the sample mean for a given n except for some special distributions. Distributions are quite simply difficult objects to calculate. As a compromise, we try to find approximations to the exact distribution of an estimator. These approximations are derived by assuming that n is large; more precisely, we use the concept of limiting distributions (see Chapter 1 again, if needed). Remarkably, even though calculation of distributions of estimators for fixed n is usually so difficult, in the limiting (or asymptotic) regime, the story is incredibly structured, and things fall into place like the pieces of a difficult puzzle. Large sample theory of estimators, also often called asymptotic theory, or simply asymptotics, is a central and unifying theme in essentially all of statistical inference. Asymptotic theory is an amazingly powerful and pretty tool for the statistician; theory of inference often cannot go far at all without using asymptotic tools. An introduction to some basic tools and concepts of asymptotic inference is given in this chapter. Some references for this chapter are Serfling (1980), Lehmann (1998), van der Vaart (1998), Bickel and Doksum (2006), DasGupta (2008) and Jiang (2010). Other references are given within the sections.

9.1 Consistency

Generally, consistency is considered to be a minimum positive property that any reasonable sequence of estimates should have. Consistent estimators are not unique; many different reasonable estimates would all be consistent estimates of the same parameter. Consistency simply means that if we have a large number of sample observations, then except for rare occasions of poor luck, our estimator should be very close to the true value of the parameter. The mathematical concept is the same as that of convergence in probability.

Definition 9.1. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ be a sequence of estimators of a (possibly vector valued) parameter $\theta, \theta \in \Theta$. The estimator sequence $\hat{\theta}_n$ is said to be consistent for estimating θ if $\hat{\theta}_n \xrightarrow{P} \theta$, i.e., given any fixed $\epsilon > 0$, and any $\theta \in \Theta$,

$$P_\theta(||\hat{\theta}_n - \theta|| > \epsilon) \to 0,$$

as $n \to \infty$.

Let us see some examples.

Example 9.1. (Consistent Estimators of Binomial p). Suppose $X_1, \cdots, X_n$ are iid Bernoulli with parameter p. Denoting $\sum_{i=1}^n X_i = X$, the MLE as well as the UMVUE of p is $\frac{X}{n}$, which is just the mean of $X_1, \cdots, X_n$. Therefore, by the WLLN (weak law of large numbers; see Chapter 1), $\frac{X}{n}$ is a consistent estimator of p. Coming to Bayes estimates, if $p \sim G(\alpha, \beta)$, then the posterior mean of p is

$$E(p\,|X_1, \cdots, X_n) = \frac{X+\alpha}{n+\alpha+\beta} = \frac{X}{n+\alpha+\beta} + \frac{\alpha}{n+\alpha+\beta}$$

$$= \frac{X}{n}\frac{n}{n+\alpha+\beta} + \frac{\alpha}{n+\alpha+\beta} \xrightarrow{P} p \times 1 + 0 = p.$$

Therefore, all of these Bayes estimates are also consistent estimates of p.

Example 9.2. (Consistent Estimators of Normal Mean). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The sample mean $\bar{X}$ is the MLE as well as the UMVUE of μ. Once again, by the WLLN, it is consistent. Instead of the sample mean, consider the sample median M_n. The median M_n is also unbiased for estimating μ. Its exact variance cannot be found in closed form. But, it may be shown that $\text{Var}_{\mu,\sigma}(M_n) = \frac{\pi}{2}\frac{\sigma^2}{n} + o(\frac{1}{n})$; hence, the MSE of M_n goes to zero under each μ, σ. Therefore, by Markov's inequality,

$$P_{\mu,\sigma}(|M_n - \mu| > \epsilon) \leq \frac{E_{\mu,\sigma}(M_n - \mu)^2}{\epsilon^2} \to 0,$$

as $n \to \infty$, establishing that M_n is also consistent for estimating μ. So would be any linear combination $a\bar{X} + bM_n$, provided $a + b = 1$. We may even use $a_n\bar{X} + b_nM_n$, and get consistent estimates as long as $a_n \to a, b_n \to b$, with $a + b = 1$.

Example 9.3. (Consistent Estimation of Normal Variance). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and consider estimation of σ^2. We will first show that the sample variance s^2 is a consistent estimate of σ^2. Toward this,

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

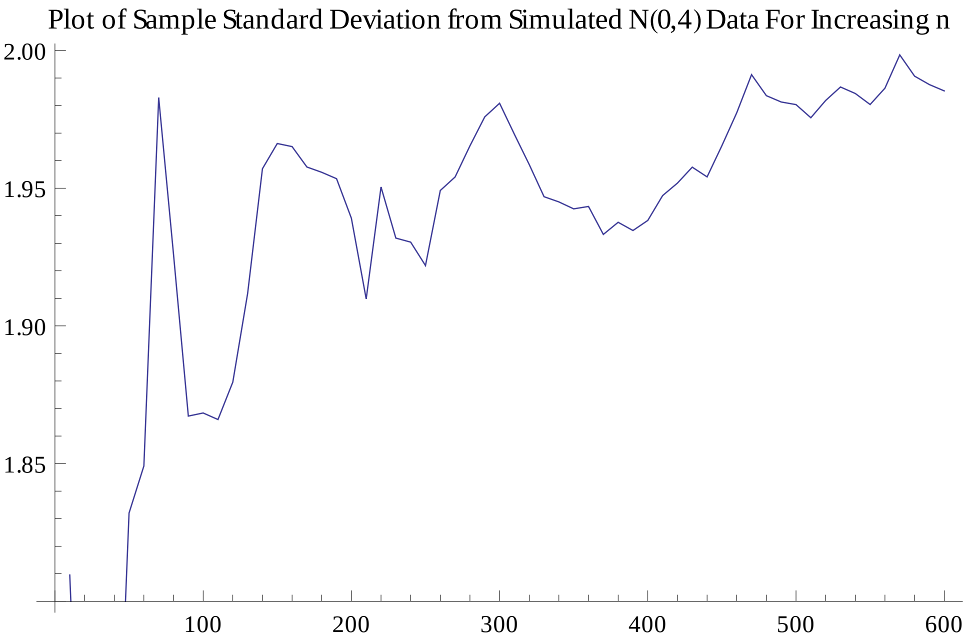
$$= \frac{n}{n-1}\frac{1}{n}\left[\sum_{i=1}^n X_i^2 - n(\bar{X})^2\right] = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X})^2\right].$$

Since our X_i are iid, so are the X_i^2, and therefore, by the WLLN, $\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} E(X_1^2) = \mu^2 + \sigma^2$. On the other hand, by the continuous mapping theorem, $(\bar{X})^2 \xrightarrow{P} \mu^2$ (see Chapter 1 again, if needed), and hence, at the end, we have,

$$s^2 \xrightarrow{P} 1 \times [\mu^2 + \sigma^2 - \mu^2] = \sigma^2.$$

From here, each of the following, as specific examples, follows by using the continuous mapping theorem:

$$s \xrightarrow{P} \sigma; \;\; \frac{1}{s} \xrightarrow{P} \frac{1}{\sigma}; \;\; \log s \xrightarrow{P} \log \sigma.$$



The picture of values of the sample standard deviation s for increasing n hints that it is ultimately getting close to the true $\sigma = 2$.

Example 9.4. (Consistent Estimation in Uniform Distribution). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$. We will first show that $X_{(n)}$, the MLE of θ is a consistent estimator of θ. This is a simple calculation; fix $\epsilon > 0$, and observe

$$P_\theta(|X_{(n)} - \theta| > \epsilon) = P_\theta(X_{(n)} > \theta + \epsilon) + P_\theta(X_{(n)} < \theta - \epsilon)$$

$$= P_\theta(X_{(n)} < \theta - \epsilon) = (\frac{\theta - \epsilon}{\theta})^n \to 0,$$

as $n \to \infty$.

We can write many other consistent estimates of θ; e.g., $2\bar{X}$ (by the usual WLLN), $X_{(n)} - X_{(1)}$, $X_{(n)} + X_{(1)}$, and $X_{(n-1)}$, as only four specific examples.

Example 9.5. (Consistent Estimation in Multinomial Distribution). Suppose $(X_1, \cdots, X_{k+1})$ has just the usual multinomial distribution with cell probabilities $p_1, p_2, \cdots, p_{k+1}$, and the n parameter is assumed known. Marginally, each cell frequency $X_i \sim Bin(n, p_i)$. Therefore, for each specific i, $\frac{X_i}{n}$ (which is the proportion of sampling units that are of type i) is consistent for estimating p_i. As a result, we can conclude, for example, that $\frac{X_i}{n} - \frac{X_j}{n}$ is a consistent estimate of $p_i - p_j$; $\frac{X_i}{X_j}$ is a consistent estimate of $\frac{p_i}{p_j}$; $\max\{\frac{X_i}{n}, \frac{X_j}{n}\}$ is a consistent estimate of $\max\{p_i, p_j\}$, and so on.

Example 9.6. (Unusual Consistency Phenomenon in Cauchy Distribution). Consider estimation of the location parameter of a Cauchy distribution. The heavy tails of the Cauchy distribution cause some unusual consistency phenomena; if you are not careful, you will end up choosing bad estimates.

We will first show that contrary to the normal case, in the Cauchy case, the sample mean is not a consistent estimate of μ. In the Cauchy case, the WLLN does not hold; so the proof that worked in the normal case does not work for Cauchy. Not only that, the sample mean is actually not consistent. To see this, use the fact that if $X_1, \cdots, X_n$ are iid $C(\mu, 1)$, then for every $n, \bar{X}_n$ is also distributed as $C(\mu, 1)$; in other words, as more data accumulate, the distribution of $\bar{X}_n$ does not pile up near μ. You should convince yourself that this fact will follow for general n, if you can prove it for $n = 2$, and that for $n = 2$ you can verify it directly by using the formula for density of a convolution.

Anyway, this implies, for instance,

$$P_\mu(|\bar{X} - \mu| > 1) = 2P_\mu(\bar{X} > \mu + 1) = 2P_\mu(X_1 > \mu + 1)$$
$$= \frac{2}{\pi}\int_1^\infty \frac{1}{1+z^2}dz = \frac{1}{2},$$

and so, obviously it does not go to zero.

However, that does not mean that there aren't any consistent estimates. There are plenty of them. For instance, the sample median M_n is a consistent estimate of μ, and this can be shown by observing that the median is an unbiased estimate of μ (for $n > 2$) and that its variance goes to zero as $n \to \infty$. This allows you to use Markov's inequality, as we did in Example 9.2.

Other consistent estimates can be found easily by using the sample percentiles; see the chapter exercises.

Example 9.7. (Consistency Alone is Not Much). As we remarked above, consistency is a weak large sample positive property, and consistency alone is not a good reason to use an estimate. For example, in the binomial p problem, consider the ridiculous estimate

$$\hat{p} = -10^9, \text{ if } n \leq 10^{5000}, \text{ and } = \frac{X}{n}, \text{ if } n > 10^{5000}.$$

Given any $\epsilon > 0$,

$$\lim_n P_p(|\hat{p} - p| > \epsilon) = \lim_n P_p(|\frac{X}{n} - p| > \epsilon) = 0.$$

So, $\hat{p}$ is consistent, although at any practical instance it estimates p to be -10^9.

9.1.1 Consistency of MLE

Although we can construct all kinds of consistent estimates in a typical problem, including strange ones, interest really lies in knowing whether an estimate that we routinely use is consistent. Maximum likelihood being prime among estimates in daily use, it is natural to ask what can we say about consistency of the MLE in general? Fortunately, we can say good things in fairly good generality. We will first take up the regular Exponential family and then take up distributions which are even more general. The most general such results usually come with clumsy conditions to verify. So, when it comes to practice, you may have to use special techniques, or use an Exponential family connection, rather than verify the clumsy general case conditions.

Theorem 9.1. (Exponential Family). Let $X_1, X_2, \cdots,$ be iid observations from a general density (or pmf) in the k parameter canonical Exponential family

$$f(x\,|\eta) = e^{\sum_{i=1}^k \eta_i T_i(x) - \psi(\eta)} h(x).$$

Assume the family to be regular as well as nonsingular. Then,
(a) For all large n, there is a unique root $\hat{\eta} = \hat{\eta}_n$ of the likelihood equation, which is the unique MLE of $\eta = (\eta_1, \cdots, \eta_k)$.
(b) $\hat{\eta}$ is a consistent estimate of η.
(c) For any continuous function $g(\eta)$, $g(\hat{\eta})$ is a consistent estimate of $g(\eta)$.
Proof: Part (a) was already proved in Chapter 6. That this unique MLE of η is consistent follows by instead first looking at the MLE of the mean vector $\mu_\eta = (E_\eta(T_1), \cdots, E_\eta(T_k))$. The MLE of the mean vector, in the regular Exponential family is the sample mean vector $(\bar{T}_1, \cdots, \bar{T}_k)$, and so, we can use the WLLN to conclude its consistency. Now, in the nonsingular case, η and μ_η are smooth one-to-one functions of each other, which will give us part (b) by applying the continuous mapping theorem. Part (c) is a consequence of part (b).
Remark: One point of technical clarification is that if for the particular n that you have, the likelihood equation does not have a root, then we just define $\hat{\eta}$ in some fixed manner; consistency is not affected by this arbitrariness of the definition of $\hat{\eta}$ for small n.
When we go to regular families that are not in the regular Exponential family, for instance $C(\mu, 1)$, we need to write a fairly long list of conditions that we will need to verify. Under these conditions, there are theorems on the consistency of the MLE. *But, you have to be very very careful about exactly what these general theorems are telling you.* First we give the conditions; we will only present the case of a scalar parameter for simplicity. General multiparameter regular cases are similar, except for the statements being more clumsy.

Cramér-Rao Conditions for Consistency of MLE

C1 Identifiability, i.e., $P_{\theta_1} = P_{\theta_2} \Leftrightarrow \theta_1 = \theta_2$.

C2 $\theta \in \Theta =$ an open interval in the real line.

C3 $S = \{x : f(x|\theta) > 0\}$ is free of θ .

C4 $\forall x \in S$, $\frac{d}{d\theta} f(x|\theta)$ exists, i.e., the likelihood function is smooth as a function of the parameter.

Theorem 9.2. (General Regular Families). Suppose $X_1, X_2, \cdots$ be iid from $f(x\,|\theta)$ satisfying the conditions **C1 - C4**, Let $\theta_0 \in \Theta^0$ be the true value of θ. Then there exists a sequence of functions $\hat{\theta_n} = \hat{\theta_n}(X_1, \cdots, X_n)$ such that

(i) $\hat{\theta}_n$ is a root of the likelihood equation for all large n.

(ii) P_{θ_0}(the root $\hat{\theta}_n$ is a local maximum of $l(\theta)) \to 1$ as $n \to \infty$

(iii)$\hat{\theta_n} \xrightarrow{P} \theta_0$.

Remark: This theorem does not say which sequence of roots of the likelihood equation should be chosen to ensure consistency in the case of multiple roots. It does not even guarentee that for any given n, however large, the likelihood function $l(\theta)$ has any local maxima at all. This specific theorem is truly useful *only* in those cases where the likelihood equation has a unique root for all n. *In particular, this general theorem is not directly usable for proving that the MLE of a Cauchy location parameter is consistent, because the Cauchy likelihood equation can have multiple roots for any n. However, it is true that the MLE of the Cauchy location parameter is consistent; it requires a more direct approach than trying to use this above general theorem.*

9.2 Asymptotic Distribution of MLE

Since consistency is regarded as a weak positive property, and since in statistics one usually wants some idea of the distribution of the estimator, it is important to go beyond just consistency and look at results on limiting distributions. For results on limiting distributions, we need *more regularity conditions* than what we need for consistency. Once you have a limiting distribution result, consistency will follow as a corollary to it. So, it makes sense that establishing limiting distribution results requires more regularity conditions than we need for consistency alone. The theorems are going to say that when all the regularity conditions hold, *certain MLE-like estimates are going to be asymptotically normal.* It is really important that you understand exactly what these theorems allow you to conclude.

9.2.1 General Regular Families

As in the case of consistency, for general regular families, one can only assert asymptotic normality of suitable sequences of roots of the likelihood equation. For Exponential families, roots of the likelihood equation are unique, and so the asymptotic normality result is more concrete and useful.

We state here the asymptotic normality result directly for the multiparameter case, from which the one parameter case follows as a special case. For a complete list of the regularity conditions needed to prove the following theorem, and also for a proof, see Lehmann and Casella (1998), or, Bickel and Doksum (2006). We refer to the list of all these assumptions as *Cramér-Rao conditions for asymptotic normality.*

Theorem 9.3. Let $X_1, X_2, \cdots$ be iid observations from a density (pmf) $f(x\,|\theta)$, where θ is a possibly vector valued parameter. Let $\theta_0 \in \Theta^0$ denote the true value of the parameter. Under the *Cramér-Rao conditions for asymptotic normality*, there exists a sequence of roots $\hat{\theta}_n$ of the likelihood equation which is consistent and which satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\mathcal{L}}{\Rightarrow} N(0, I^{-1}(\theta_0)),$$

where $I(\theta)$ is the Fisher information matrix.

See below for a heuristic proof of this theorem.
It is essential that you understand that the variance of the limiting normal distribution is not the same thing as the limit of the variances. In fact, for all finite n, your estimate may have an infinite variance, although the limiting distribution has a finite variance. Never confuse one with the other.
Remark: The theorem applies to any distribution in the regular nonsingular Exponential family. *In the Exponential family, the asymptotic normality of the MLE is nothing but the usual central limit theorem for means of iid random variables with a finite variance.* For more general distributions, the asymptotic normality of the MLE is proved by using the CLT; but it is not a trivial consequence of the CLT. *More work is needed.*
Here is a heuristic explanation for why the asymptotic normality result holds in the one parameter Exponential family case. Consider the score function

$$U(\theta) = \frac{d}{d\theta} \log l(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i\,|\theta).$$

At the MLE, the score function takes the value zero. Hence, by a straightforward Taylor expansion,

$$0 = U(\hat{\theta}) = U(\theta) + (\hat{\theta} - \theta)U'(\theta)$$

$$= \sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i\,|\theta) + (\hat{\theta} - \theta) \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(X_i\,|\theta) \approx \sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i\,|\theta) + (\hat{\theta} - \theta)(-nI(\theta))$$

$$\Rightarrow (\hat{\theta} - \theta) \approx \frac{\sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i\,|\theta)}{nI(\theta)}.$$

By the CLT for iid random variables,

$$\sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i\,|\theta) \approx N(0, nI(\theta))$$

and so,

$$\hat{\theta} - \theta \approx \frac{\sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i\,|\theta)}{nI(\theta)} \approx N(0, \frac{1}{nI(\theta)}).$$

In some sense, Exponential family is the only broad class of distributions for which the MLE asymptotic normality theorem is straightforward to interpret. Beyond the Exponential family, you must be really careful about what the theorem is telling you.

Remark: MLEs are not the only sequence of estimates of a parameter that are asymptotically normal in regular problems. Many other types of estimates, which differ slightly from the MLE, are also asymptotically normal. For example, in regular problems, posterior means with respect to well behaved prior densities are asymptotically normal. The concepts of $\sqrt{n}$-consistency and asymptotic efficiency have been proposed to refer to and compare these various types of asymptotically normal estimates. Here are those two definitions.

Definition 9.2. Suppose the *Cramér-Rao conditions* hold. A sequence of real valued statistics T_n is called *$\sqrt{n}$-consistent* if $\sqrt{n}[T_n - \theta] \overset{\mathcal{L}}{\Rightarrow} N(0, V(\theta))$ for some function $V(\theta), 0 < V(\theta) < \infty$ for all θ.

Definition 9.3. Let $T_{n,i}, i = 1, 2$ be two $\sqrt{n}$-consistent sequences, $\sqrt{n}[T_{n,i} - \theta] \overset{\mathcal{L}}{\Rightarrow} N(0, V_i(\theta))$. The *asymptotic efficiency of $T_{n,2}$ with respect to $T_{n,1}$* is defined as

$$\text{eff}(T_2, T_1) = \frac{V_1(\theta)}{V_2(\theta)}.$$

The definition says that the estimate with a smaller asymptotic variance is more efficient.

Let us now see some examples.

Example 9.8. (Asymptotic Normality of MLE in Binomial). If we define the estimating sequence

$$\hat{p} = c, \text{ if } X = 0, n, \text{ and } = \frac{X}{n}, \text{if } 0 < X < n,$$

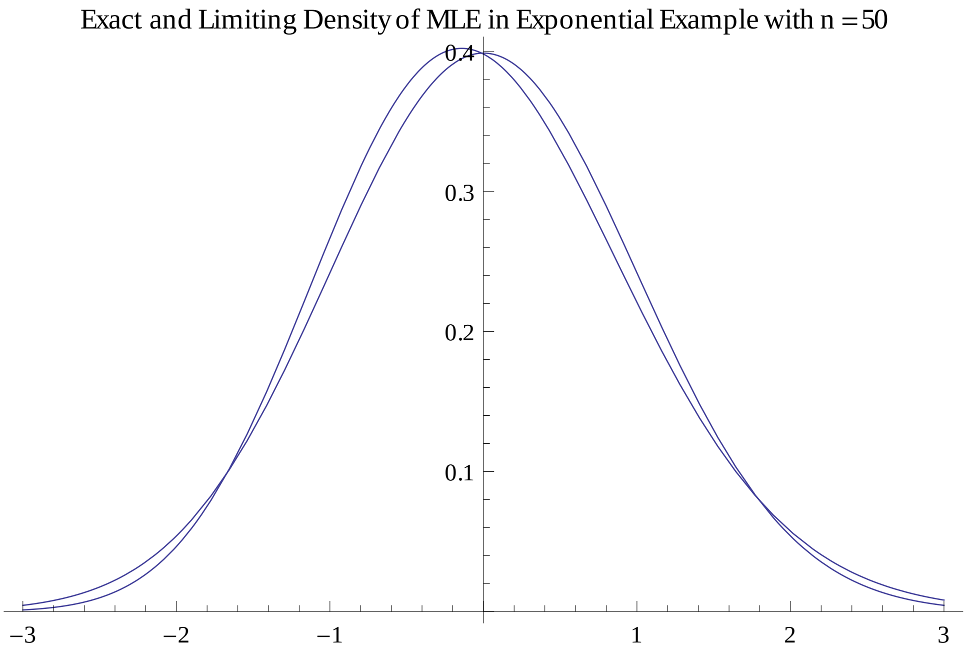
then we can conclude from the above theorem that

$$\sqrt{n}(\hat{p} - p) \overset{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{I(p)})$$

whatever be $p, 0 < p < 1$. Since $I(p) = \frac{1}{p(1-p)}$ in the Bernoulli case, we can say

$$\sqrt{n}(\hat{p} - p) \overset{\mathcal{L}}{\Rightarrow} N(0, p(1-p)).$$

You will recognize that this is really *the same as what we call normal approximation to the binomial.* It is coming back in the form of asymptotic normality of the MLE of p; but it's the same thing as normal approximation to the binomial, i.e., the good old central limit theorem!



Exact and Limiting Density of MLE in Exponential Example with n = 50

Example 9.9. (Asymptotic Normality of MLE in Exponential). Suppose $X_1, X_2, \cdots$ are iid $Exp(\lambda)$. The unique MLE of λ (for any n) is $\hat{\lambda} = \bar{X}$. The Fisher information function in this case is $I(\lambda) = \frac{1}{\lambda^2}$. So, we can conclude from our general theorem above that

$$\sqrt{n}(\hat{\lambda} - \lambda) \overset{\mathcal{L}}{\Rightarrow} N(0, \lambda^2).$$

Once again, you recognize that this is the same as what the CLT tells you, because $\hat{\lambda} = \bar{X}$, the sample mean in this example. A plot helps you compare the true distribution and the asymptotic distribution of the MLE in this exponential example. The accuracy of the asymptotic distribution is reasonable, but not great.

Example 9.10. (Asymptotic Distribution of MLE in Cauchy). Here is an example where the MLE asymptotic normality theorem shows its full force. Thus, suppose $X_1, X_2, \cdots$ are iid $C(\mu, 1)$. We know from Chapter 7 that in this case the MLE $\hat{\mu}$ does not have any formula at all. So, there is no hope of saying what is the exact distribution of the MLE for a given n. But the asymptotic normality theorem will tell us something concrete and incredibly simple.

To see why, recall that in any location parameter case $f(x\,|\mu) = f_0(x - \mu)$, the Fisher information function is a pure constant,

$$I(\mu) = \int_{-\infty}^{\infty} \frac{[f_0'(z)]^2}{f_0(z)} dz.$$

In the Cauchy case,

$$f_0(z) = \frac{1}{\pi(1+z^2)},\ f_0'(z) = -\frac{2z}{\pi(1+z^2)^2},\ \frac{[f_0'(z)]^2}{f_0(z)} = \frac{4z^2}{\pi(1+z^2)^3}.$$

Hence,

$$I(\mu) = \int_{-\infty}^{\infty} \frac{4z^2}{\pi(1+z^2)^3} dz = \frac{8}{\pi} \int_0^{\infty} \frac{z^2}{(1+z^2)^3} dz = \frac{1}{2}.$$

We can now conclude that

$$\sqrt{n}[\hat{\mu} - \mu] \overset{\mathcal{L}}{\Rightarrow} N(0, 2),$$

although we have no formula for $\hat{\mu}$ at all.

Example 9.11. (Asymptotics of MLEs in Two Parameter Beta). Asymptotic normality for the maximum likelihood estimates of the two parameters in the general Beta distribution hold, because the Beta distribution is a member of the two parameter regular Exponential family. However, the MLEs have no analytic formula and can only be numerically computed. Use the notation

$$r_i = \frac{x_i}{1 - x_i}, \overline{\log r} = \frac{1}{n} \sum_{i=1}^n \log r_i, \overline{\log x} = \frac{1}{n} \sum_{i=1}^n \log x_i.$$

Then, the MLE of the two Beta parameters α, β are unique roots of two complicated likelihood equations

$$\overline{\log x} + \psi(\alpha + \psi^{-1}(\psi(\alpha) - \overline{\log r})) - \psi(\alpha) = 0,$$

$$\beta = \psi^{-1}(\psi(\alpha) - \overline{\log r}),$$

where

$$\psi(x) = \frac{d}{dx} \log(\Gamma(x)), \psi'(x) = \frac{d}{dx} \psi(x).$$

The Fisher information matrix is easily seen to be

$$I(\alpha, \beta) = \begin{pmatrix} \psi'(\alpha) - \psi'(\alpha + \beta) & -\psi'(\alpha + \beta) \\ -\psi'(\alpha + \beta) & \psi'(\beta) - \psi'(\alpha + \beta) \end{pmatrix}.$$

For the asymptotic normal distribution of the MLE of (α, β) we need the inverse of the Fisher information matrix, and by straightforward calculation,

$$I^{-1}(\alpha, \beta) = \frac{1}{\psi'(\alpha)\psi'(\beta) - \psi'(\alpha + \beta)[\psi'(\alpha) + \psi'(\beta)]}$$

$$\times \begin{pmatrix} \psi'(\beta) - \psi'(\alpha + \beta) & \psi'(\alpha + \beta) \\ \psi'(\alpha + \beta) & \psi'(\alpha) - \psi'(\alpha + \beta) \end{pmatrix}.$$

The conclusion is that although the MLEs of (α, β) have no formula, and there is no possibility of finding the exact distribution of the MLE under a given (α, β), the asymptotic distribution is explicit. This is another prime example of the force and elegance of asymptotic theory.

9.2.2 In Exponential Families

It is useful to formally record the asymptotic distribution of the MLE in the general Exponential family. This is now easy, because we can simply borrow the formula of the Fisher information function in Exponential family from Chapter 6 (Theorem 6.4). Here is the general theorem on the asymptotic distribution of the MLE in one parameter Exponential families; the multiparameter case is analogous, but the asymptotic variance formula is a little messy.

Theorem 9.4. Let $X_1, X_2, \cdots$ be iid from $f(x\,|\theta) = e^{\eta(\theta)T(x)-\psi(\theta)}h(x), \theta \in \Theta \subseteq \mathcal{R}$. Assume that the family is regular and nonsingular.
(a) Let $\hat{\theta}$ denote the MLE of θ. Then,

$$\sqrt{n}[\hat{\theta} - \theta] \overset{\mathcal{L}}{\Rightarrow} N(0, \frac{\eta'(\theta)}{\psi''(\theta)\eta'(\theta) - \psi'(\theta)\eta''(\theta)}).$$

(b) If the density is written in the canonical form $f(x\,|\eta) = e^{\eta T(x)-\psi(\eta)}h(x), \eta \in \mathcal{T}$, then the result simplifies to

$$\sqrt{n}[\hat{\eta} - \eta] \overset{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{\psi''(\eta)}).$$

Proof: Follows from Theorem 6.14 and Theorem 9.3.

9.2.3 Nonregular Cases

In nonregular cases, the asymptotics of MLEs differ fundamentally in a few ways. Typically, the results in nonregular cases are of the following form:

$$c_n[\hat{\theta} - \theta] \overset{\mathcal{L}}{\Rightarrow} G_\theta,$$

where the limiting distribution G_θ is not normal, and is essentially case-specific. Second, the normalizing sequence c_n in the above is not $\sqrt{n}$ and is also rather case-specific. So, there are still meaningful asymptotics of the MLEs in nonregular cases; but the story is far less unifying than when the usual Cramér-Rao regularity conditions hold. It would be useful to see an example.

Example 9.12. (Asymptotics of MLE of Uniform Endpoint). Suppose $X_1, X_2, \cdots \overset{iid}{\sim} U[0, \theta]$. The unique MLE of θ is the sample maximum $X_{(n)}$. We can write $X_{(n)} = \theta Z_{(n)}$, where $Z_{(n)}$ is the maximum of n iid $U[0, 1]$ random variables. In Example 1.153, we showed that $n(1 - Z_{(n)}) \overset{\mathcal{L}}{\Rightarrow} Exp(1)$. It follows immediately that

$$n(\theta - X_{(n)}) \overset{\mathcal{L}}{\Rightarrow} Exp(\theta),$$

an exponential with mean θ. Notice that the limiting distribution is *not normal*, and that the normalizing sequence is n, *not* $\sqrt{n}$. Indeed, if instead of $n(\theta - X_{(n)})$, we consider

$\sqrt{n}(\theta - X_{(n)})$, we get the uninteresting result $\sqrt{n}(\theta - X_{(n)}) \overset{P}{\Rightarrow} 0$. What we need is convergence in distribution to a nondegenerate distribution; convergence in probability to zero is not useful for followup statistical analysis, such as construction of confidence intervals. The moral is that you must be careful about the normalizing sequence c_n and about the form of the limiting distribution in nonregular cases.

9.3 Asymptotics in Action: Approximate Pivots and Making Confidence Statements

We have remarked several times that point estimation is just a first step for follow up inference. Principal among follow up inferences are confidence intervals (or their Bayesian counterparts, which will be discussed later). A method or trick to calculate confidence intervals is to use functionals $T_n(X_1, \cdots, X_n, \theta)$ such that the distribution of $T_n(X_1, \cdots, X_n, \theta)$ is free of θ. Do not confuse this with the notion of an ancillary statistic; $T_n(X_1, \cdots, X_n, \theta)$ is *a functional, not a statistic.*

Such functionals whose distributions do not depend on the parameter are called *pivots.* Here is an example.

Example 9.13. (Pivot in the Case of Normal Distributions). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, both μ, σ being considered unknown. Let $\theta = (\mu, \sigma)$. Define the functional

$$T_n(X_1, \cdots, X_n, \theta) = \frac{\sqrt{n}(\bar{X} - \mu)}{s},$$

where s is as usual the sample standard deviation. For iid samples from a normal distribution, for any given n, the distribution of T_n is a t distribution with $n-1$ degrees of freedom (no other parameters involved). Hence, in the iid normal case, $T_n = \frac{\sqrt{n}(\bar{X}-\mu)}{s}$ is a genuine pivot.

Here is another example.

Example 9.14. (Pivot in the Case of Uniform). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$. Consider the functional

$$T_n(X_1, \cdots, X_n, \theta) = \frac{n(\theta - X_{(n)})}{\theta}.$$

Then, simple calculations show that for any given n, and under any θ, T_n has the density

$$f_{T_n}(t) = (1 - \frac{t}{n})^{n-1},\ 0 \leq t \leq n.$$

This is a fixed density with no other parameters involved. Hence, $T_n = \frac{n(\theta - X_{(n)})}{\theta}$ is a pivot in the $U[0, \theta]$ case.

If we can find a pivot, then interesting and useful additional calculations can be done. For example, we would be able to find pure constants a_n, b_n (which do not depend on θ) such that

$$P_\theta(a_n \leq T_n(X_1, \cdots, X_n, \theta) \leq b_n) \geq .95 \text{ for all } \theta.$$

Now, with some luck, a little algebraic manipulation will reduce the event $a_n \leq T_n(X_1, \cdots, X_n, \theta) \leq b_n$ to an event in the form $l_n(X_1, \cdots, X_n) \leq \theta \leq u_n(X_1, \cdots, X_n)$; in other words, with some algebra, we would be able to separate θ and the data $(X_1, \cdots, X_n)$ from our pivot $T_n(X_1, \cdots, X_n, \theta)$. Then, we will have

$$P_\theta(l_n \leq \theta \leq u_n) \geq .95 \text{ for all } \theta.$$

This latest statement traps the unknown θ within two limits l_n, u_n that depend only on the data. Because the statement is true at least 95% of the times, the interval of values $l_n \leq \theta \leq u_n$ is called *a 95% confidence interval for θ*. Of course, 95% is only an artifact; we could use 80% or 99.99%, or whatever we want. The point is that *if we can find for our distributions, a pivot T_n then we should be able to construct confidence intervals for the parameter by using that pivot.*

For some distributions that have built-in structure, such pivots can indeed be found. For example, pivots can be found if our distribution is a location-scale parameter distribution. But in general, we won't be able to find pivots for fixed n. And, here comes asymptotic theory. For instance, asymptotics of MLEs will hand us functionals which are *almost pivotal*; in other words, functionals whose asymptotic distributions are free of the parameter θ. Then, we work with that asymptotic distribution, and we are squarely back in the game. We can then calculate asymptotic confidence intervals. A famous example of this idea that derives from our asymptotic theory for MLEs is the *Wald confidence interval.*

9.3.1 Wald Confidence Intervals

Suppose θ is a scalar parameter of some density or pmf $f(x\,|\theta)$. We now know that under enough regularity conditions, the MLE has an asymptotic normal distribution:

$$\sqrt{n}[\hat{\theta} - \theta] \overset{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{I(\theta)}).$$

This, in turn, implies,

$$\sqrt{nI(\theta)}[\hat{\theta} - \theta] \overset{\mathcal{L}}{\Rightarrow} N(0, 1).$$

Consider now the slight alteration of the LHS of the above display

$$\sqrt{nI(\hat{\theta})}[\hat{\theta} - \theta] = \sqrt{nI(\theta)}[\hat{\theta} - \theta] \times \sqrt{\frac{I(\hat{\theta})}{I(\theta)}}.$$

We already have $\sqrt{nI(\theta)}[\hat{\theta} - \theta] \overset{\mathcal{L}}{\Rightarrow} N(0, 1)$, and if the Fisher information function $I(\theta)$ is a continuous function of θ, then by the continuous mapping theorem (see Chapter 1),

$I(\hat{\theta}) \xrightarrow{P} I(\theta)$, and consequently, the ratio $\frac{I(\hat{\theta})}{I(\theta)} \xrightarrow{P} 1$.
Therefore, by Slutsky's theorem (see Chapter 1 again), we have the conclusion

$$\sqrt{nI(\hat{\theta})}[\hat{\theta} - \theta] \overset{\mathcal{L}}{\Rightarrow} N(0, 1).$$

Fix any number $\alpha, 0 < \alpha < 1$. As usual, let $z_{\alpha/2}$ denote the $100(1 - \frac{\alpha}{2})\%$ percentile of $N(0, 1)$. Then, by definition of convergence in distribution, as $n \to \infty$,

$$P(-z_{\alpha/2} \leq \sqrt{nI(\hat{\theta})}[\hat{\theta} - \theta] \leq z_{\alpha/2}) \to 1 - \alpha$$

$$\Rightarrow P(-\frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}} \leq \hat{\theta} - \theta \leq \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}) \to 1 - \alpha$$

$$\Rightarrow P(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}} \leq \theta \leq \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}) \to 1 - \alpha.$$

The confidence interval

$$\hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}$$

is called the *nominal* $100(1 - \alpha)\%$ *Wald confidence interval for* θ. The word *nominal* is used to remind us that only its limiting coverage probability is $1 - \alpha$; the probability need not, in general, be $1 - \alpha$ for given fixed n. Thus, Wald confidence intervals are asymptotic confidence intervals; *notice that in constructing Wald confidence intervals, we are simply using the fact that* $\sqrt{nI(\hat{\theta})}[\hat{\theta} - \theta]$ *is an approximate pivot for large* n.
The idea can be extended to the multiparameter case, where we call it nominal $100(1-\alpha)\%$ Wald confidence set for θ.
Let us see some examples.

Example 9.15. (Wald Confidence Interval for Normal Mean). Consider the case that $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ is known. The Fisher information function $I(\mu) \equiv \frac{1}{\sigma^2}$. For every n,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

exactly. In this case, no asymptotic approximations are involved. Straightforwardly, the interval $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is an *exact* $100(1 - \alpha)\%$ confidence interval for μ. This is called the *z confidence interval* in the literature.

Example 9.16. (Wald Confidence Interval for Poisson Mean). Suppose $X_1, \cdots, X_n \overset{iid}{\sim}$ $Poi(\lambda)$. In this case, $I(\lambda) = \frac{1}{\lambda}$ and the MLE is $\hat{\lambda} = \bar{X}$. Hence the nominal $100(1 - \alpha)\%$ Wald confidence interval for λ is

$$\bar{X} \pm z_{\alpha/2}\frac{\sqrt{\bar{X}}}{\sqrt{n}}.$$

Note that in rare cases, each X_i may be zero, in which case the Wald confidence interval becomes the singleton set $\{0\}$.

Example 9.17. (Wald Confidence Interval for Binomial p). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} Ber(p)$. Then, the Fisher infoormation function $I(p) = \frac{1}{p(1-p)}$ and the MLE is $\hat{p} == \bar{X} = \frac{X}{n}$, if we let $\sum_{i=1}^n X_i = X$. Hence the nominal $100(1-\alpha)\%$ Wald confidence interval for p is

$$\hat{p} \pm z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}},$$

where $\hat{p} = \frac{X}{n}$.

9.4 Asymptotics in Action: Delta Theorem

As we have experienced numerous times, in statistical practice, we often want to estimate a function (transformation) $g(\theta)$ of a parameter θ. Plug-in principle says that if $\hat{\theta}$ is a good estimate of θ (e.g., an MLE), then we may estimate $g(\theta)$ by $g(\hat{\theta})$. But, as always, we would then want to have some idea of the distribution of this estimator, namely $g(\hat{\theta})$. A hugely useful theorem, known as the *delta theorem*, tells us how to find the asymptotic distribution of $g(\hat{\theta})$ by using the asymptotic distribution of $\hat{\theta}$. The theorem is useful in numerous problems of statistics; e.g., to construct confidence intervals for $g(\theta)$, for testing hypotheses, and for finding certain popular transformations known as *variance stabilizing transformations.*

We will present the delta theorem in three parts. The first part is the version that you would most use in regular one parameter examples. The second part is useful for regular multiparameter examples. Finally, the third part is the most general delta theorem that covers one or multiparameter problems, and whether or not the problem is regular. Part (c) will at first look formidable, and you should read it several times to understand the notation as well as what it says.

Thm **(Delta Theorem of Cramér)** (a) Let T_n be a sequence of real valued statistics such that

$$\sqrt{n}(T_n - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, \sigma^2(\theta)), \quad \sigma(\theta) > 0.$$

Let $g : \mathcal{R} \to \mathcal{R}$ be once differentiable at θ with $g'(\theta) \neq 0$. Then,

$$\sqrt{n}[g(T_n) - g(\theta)] \overset{\mathcal{L}}{\Rightarrow} N(0, [g'(\theta)]^2 \sigma^2(\theta)).$$

(b) Let T_n be a sequence of p-dimensional statistics such that for some p-dimensional vector θ, and some $p \times p$ positive definite matrix $\Sigma = \Sigma(\theta)$,

$$\sqrt{n}[T_n - \theta] \overset{\mathcal{L}}{\Rightarrow} N_p(\mathbf{0}, \Sigma(\theta)).$$

Let $g : \mathcal{R}^p \to \mathcal{R}$ be a function which is once partially differentiable with respect to each coordiante at the point θ, and suppose its gradient vector $\nabla g(\theta) \neq \mathbf{0}$. Then,

$$\sqrt{n}[g(T_n) - g(\theta)] \overset{\mathcal{L}}{\Rightarrow} N(0, [\nabla g(\theta)]' \Sigma(\theta) [\nabla g(\theta)]).$$

In the above, the quantity

$$[\nabla g(\theta)]'\Sigma(\theta)[\nabla g(\theta)] = \sum_{i=1}^{p}\sum_{j=1}^{p}(\frac{\partial}{\partial\theta_i}g)(\frac{\partial}{\partial\theta_j}g)\sigma_{ij}.$$

(c) Let $\mathbf{T_n}$ be p-dimensional random vectors. Suppose for some p-dimensional vector θ, and some sequence of reals $c_n \to \infty$,

$$c_n(\mathbf{T_n} - \theta) \overset{\mathcal{L}}{\Rightarrow} G_\theta,$$

where G_θ is a distribution on $\mathcal{R}^p$. Let $g : \mathcal{R}^p \to \mathcal{R}^k$ be a function with each coordinate of g once continuously differentiable with respect to every coordinate of $\mathbf{x}$ at the point $\mathbf{x} = \theta$. Define the $k \times p$ matrix

$$Dg(\mu) = ((\frac{\partial g_i}{\partial x_j}))|_{\mathbf{x}=\mu}.$$

Then,

$$c_n(g(\mathbf{T_n}) - g(\theta)) \overset{\mathcal{L}}{\Rightarrow} H_\theta,$$

where H_θ iis the distribution of the k-dimensional random vector $Dg(\mu)\mathbf{X}$, and $\mathbf{X} \sim G$.

Proof: We prove part (a). The proof of parts (b) and (c) uses the same logic, except for needing more notation.

The basic reason that the delta theorem holds is that a smooth function is locally linear. Take a smooth function $g(x)$ and expand it around a given $x_0 : g(x) \approx g(x_0) + (x - x_0)g'(x_0)$. In the statistical context, the role of x is played by T_n, the role of x_0 by θ. So, locally, near θ, $g(T_n)$ behaves like a linear function of T_n. By hypothesis, T_n is asymptotically normal. Since linear functions of normals are also normals, you would expect $g(T_n)$ also to be asymptotically normal. This is the crux of the delta theorem.

Writing a correct proof takes care. First note that the hypothesis of part (a) that $\sqrt{n}[T_n-\theta]$ has a limiting normal distribution implies as a corollary that T_n converges in probability to θ (verify this easy fact!). Hence, $T_n - \theta = o_p(1)$. The proof of the theorem now follows from a simple application of Taylor's theorem. Here is why.

By Taylor's theorem,

$$g(x_0 + h) = g(x_0) + hg'(x_0) + o(h)$$

if g is differentiable at x_0. Therefore, it follows that

$$g(T_n) = g(\theta) + (T_n - \theta)g'(\theta) + o_p(T_n - \theta).$$

That the remainder term is $o_p(T_n - \theta)$ follows from our previous observation that $T_n - \theta = o_p(1)$. Taking $g(\theta)$ to the left side, and multiplying both sides by $\sqrt{n}$, we obtain

$$\sqrt{n}[g(T_n) - g(\theta)] = \sqrt{n}(T_n - \theta)g'(\theta) + \sqrt{n}\; o_p(T_n - \theta).$$

But, our hypothesis implies that $\sqrt{n}\, o_p(T_n - \theta) = o_p(1)$. Hence, an application of Slutsky's theorem gives

$$\sqrt{n}[g(T_n) - g(\theta)] \overset{\mathcal{L}}{\Rightarrow} N(0, [g'(\theta)]^2\sigma^2(\theta)),$$

and thiis proves part (a).

Remark: There are instances in which $g'(\theta) = 0$ (at least for some special values of θ). In such cases, the limiting distribution of $g(T_n)$ is not determined by the delta theorem. The one term Taylor expansion will not give the correct limiting distribution of $g(T_n)$ if $g'(\theta) = 0$. You must carry the Taylor expansion to the next term. If you do, you will get the result that if $g'(\theta) = 0$, then

$$n[g(T_n) - g(\theta)] \overset{\mathcal{L}}{\Rightarrow} \frac{g''(\theta)\sigma^2(\theta)}{2}\chi^2_1.$$

9.4.1 Examples of the Delta Theorem

Let us see some examples of use of the delta theorem.

Example 9.18. (Simple Example of Delta Theorem). Suppose $X_1, X_2, \cdots$ are iid with mean μ and finite variance σ^2. Let $T_n(X_1, X_2, \cdots, X_n) = \bar{X}$ and let $g(\bar{X}) = (\bar{X})^2$. By the CLT, $\sqrt{n}(\bar{X} - \mu) \overset{\mathcal{L}}{\Rightarrow} N(0, \sigma^2)$. We have $g(\mu) = \mu^2, g'(\mu) = 2\mu$ and $\sigma^2(\mu) = \sigma^2$. If $\mu \neq 0$, then $[g'(\mu)]^2\sigma^2(\mu) = 4\mu^2\sigma^2 > 0$. Therefore, by the delta theorem, for $\mu \neq 0$,

$$\sqrt{n}[\bar{X}^2 - \mu^2] \overset{\mathcal{L}}{\Rightarrow} N(0, 4\mu^2\sigma^2).$$

If $\mu = 0$, then $\sqrt{n}\bar{X} \overset{\mathcal{L}}{\Rightarrow} N(0, \sigma^2)$, and so $\frac{\sqrt{n}\bar{X}}{\sigma} \overset{\mathcal{L}}{\Rightarrow} N(0, 1)$. It follows that if $\mu = 0$, then $\frac{n\bar{X}^2}{\sigma^2} \overset{\mathcal{L}}{\Rightarrow} \chi^2_1$. Notice how the result changed when $\mu = 0$.

Example 9.19. (Applying Delta Theorem to Poisson). Suppose $X_1, X_2, \cdots$ are iid $Poi(\lambda)$. Suppose we want to estimate $P(X = 0) = e^{-\lambda} = g(\lambda)$. We know from Chapter 6 that the MLE of $e^{-\lambda}$ is $e^{-\bar{X}} = g(\bar{X})$. In this case, we have, $g'(\lambda) = -e^{-\lambda}, \sigma^2(\lambda) = \lambda$, and therefore, by the delta theorem,

$$\sqrt{n}[e^{-\bar{X}} - e^{-\lambda}] \overset{\mathcal{L}}{\Rightarrow} N(0, \lambda e^{-2\lambda}).$$

Example 9.20. (Wrong Use of Delta Theorem). The delta theorem is a result on finding limiting distributions of functions of statistics. It cannot be blindly used to approximate the mean or the variance of functions of statistics. Here is an example. Suppose $X_1, X_2, \cdots$ are iid $Exp(\lambda)$. Sometimes, we want to estimate $\frac{1}{\lambda}$; its MLE is $\frac{1}{\bar{X}}$. The derivative of $\frac{1}{\lambda}$ is $-\frac{1}{\lambda^2}$, and $\sqrt{n}[\bar{X} - \lambda] \overset{\mathcal{L}}{\Rightarrow} N(0, \lambda^2)$. Therefore, the delta theorem gives us

$$\sqrt{n}[\frac{1}{\bar{X}} - \frac{1}{\lambda}] \overset{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{\lambda^2}).$$

However, it would be meaningless to use this result to say that $\mathrm{Var}(\frac{1}{\bar{X}}) \approx \frac{1}{\lambda^2}$ (or, $\approx \frac{1}{\bar{X}^2}$), because $\mathrm{Var}(\frac{1}{\bar{X}})$ is always ∞; you cannot talk about approximating a variance which never exists. It would also be meaningless in this example to talk about approximating the mean of $\frac{1}{\bar{X}}$. Do not use the delta theorem blindly for what it does not do.

Example 9.21. (Applying Delta Theorem to Sample Variance). This is an example where we have to apply a bivariate delta theorem, namely part (b) of the delta theorem. Suppose X_i are iid with $E(X^4) < \infty$. The purpose of the example is to derive the limiting distribution of the sample variance $s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$; note that we have taken the divisor to be n instead of $n-1$ and this does not affect the final result.
We denote

$$E(X_i) = \mu;\ \mathrm{Var}(X_i) = \sigma^2;\ E[(X_i - \mu)^4] = \mu_4.$$

To apply the bivariate delta theorem, we first define

$$T_n = \begin{pmatrix} \bar{X} \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \end{pmatrix},\ \theta = \begin{pmatrix} EX_1 \\ EX_1^2 \end{pmatrix},\ \Sigma(\theta) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_1^2) \\ Cov(X_1, X_1^2) & Var(X_1^2) \end{pmatrix}.$$

Thus, in the delta theorem, we identify the vector θ with $\theta = (\mu, \mu^2 + \sigma^2)$. Next define the scalar function of two variables $g(u, v) = v - u^2$; this means $g(T_n) = s^2, g(\theta) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$. Also, the gradient vector of g is

$$\nabla g(u, v) = (-2u, 1)'.$$

Therefore,

$$\nabla g(\theta) = (-2\mu, 1).$$

By the multivariate central limit theorem,

$$\sqrt{n}(T_n - \theta) \stackrel{\mathcal{L}}{\Rightarrow} N_2(\mathbf{0}, \Sigma(\theta)).$$

Therefore, by the bivariate delta theorem,

$$\sqrt{n}[(\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2) - \sigma^2] \stackrel{\mathcal{L}}{\Rightarrow} N(0, [\nabla g(\theta)]'\Sigma(\theta)[\nabla g(\theta)]).$$

A little bit of matrix algebra shows that $[\nabla g(\theta)]'\Sigma(\theta)[\nabla g(\theta)] = \mu_4 - \sigma^4$. Thus, the final result is

$$\sqrt{n}[s^2 - \sigma^2] \stackrel{\mathcal{L}}{\Rightarrow} N(0, \mu_4 - \sigma^4).$$

In the special case where the X_i are iid $N(\mu, \sigma^2)$, we have $\mu_4 = 3\sigma^4$, and so, in that case,

$$\sqrt{n}[s^2 - \sigma^2] \stackrel{\mathcal{L}}{\Rightarrow} N(0, 2\sigma^4).$$

You should be able to prove this special normal case result directly by using the fact that in the normal case $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$.
See DasGupta (2011) for more examples of use of the delta theorem.

9.6 At Instructor's Discretion

9.6.1 Mean and Variance of Complicated Statistics

Complicated nonlinear statistics $T = T_n(X_1, \cdots, X_n)$ are often used in statistical inference for purposes of estimation, testing, and model verification. A few examples of such nonlinear statistics that we frequently use are the sample variance and standard deviation, the sample correlation, the sample coefficients of skewness and kurtosis, etc. A number of useful formulas are put together in this section. These classic formulas have now gotten very hard to locate and are getting forgotten. For even more detailed treatment of the material in this section, see the classic work Cramér (1946).

As a first step of the analysis, we often want to know the mean and the variance of T. As a rule, we cannot find the mean and the variance of complicated statistics exactly. The next best thing is an approximation; and, once again, asymptotic techniques come to our rescue. Let us see an example.

Example 9.24. (Mean and Variance of Sample Variance). Let $X_1, \cdots, X_n$ be an iid sample of size n from a CDF F on the real line with four finite moments, and let $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ be the sample variance. Of course, $E(s^2) = \sigma^2$; but what is $\text{Var}(s^2)$?

Even this is not a really simple calculation. By expanding

$$\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2 = \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)^2$$

$$= (\sum_{i=1}^n X_i^2)^2 + n^2\bar{X}^4 - 2n\bar{X}^2(\sum_{i=1}^n X_i^2),$$

and then by evaluating the expectation of each term carefully, we can ultimately get a formula for the second moment of s^2, and from there, a formula for the variance of s^2. Here is what the exact formula is:

$$\text{Var}(s^2) = \frac{n}{(n-1)^2}[\mu_4 - \sigma^4] - \frac{2}{(n-1)^2}[\mu_4 - 2\sigma^4] + \frac{1}{n(n-1)^2}[\mu_4 - 3\sigma^4],$$

where, we use the notation

$$\mu_r = E(X_1 - \mu)^r,$$

when this exists. You can see how complicated the formula for $\text{Var}(s^2)$ is in general.

On inspecting this exact formula, we notice that

$$\text{Var}(s^2) = \frac{\mu_4 - \sigma^4}{n} + O(n^{-2}).$$

Asymptotic techniques are useful in writing down this leading term (or the first two leading terms) in the mean and variance of complex statistics.

The first result below deals with special but useful statistics, and gives exact formulas. We state the result in terms of the more algebraically convenient central sample moments

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r, r = 1, 2, \cdots.$$

Theorem 9.7. Let $X_1, X_2, \cdots, X_n$ be iid real valued random variables. Let also as usual $\bar{X}$ be the mean of the first n observations. Then,

$$(a)\, E[(\bar{X} - \mu)^3] = \frac{\mu_3}{n^2};\ E[(\bar{X} - \mu)^4] = \frac{3\sigma^4}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}.$$

$$(b)\, \mathrm{Cov}(\bar{X}, m_2) = \frac{n-1}{n^2}\mu_3.$$

$$(c)\, E[m_3] = \frac{(n-1)(n-2)}{n^2}\mu_3;\ E[m_4] = \frac{(n-1)(n^2-3n+3)}{n^3}\mu_4 + \frac{3(n-1)(2n-3)}{n^3}\sigma^4.$$

The second result gives approximate formulas for more general statistics, and the approximations come with mathematically valid error terms.

Theorem 9.8. For any r, s,

$$(a) E[m_r] = \mu_r + O(n^{-1});\ \mathrm{Var}(m_r) = \frac{\mu_{2r} + r^2\sigma^2\mu_{r-1}^2 - 2r\mu_{r-1}\mu_{r+1} - \mu_r^2}{n} + O(n^{-2}).$$

$$(b)\, \mathrm{Cov}(m_r, m_s) = \frac{\mu_{r+s} - r\mu_{r-1}\mu_{s+1} - s\mu_{r+1}\mu_{s-1} - \mu_r\mu_s + rs\sigma^2\mu_{r-1}\mu_{s-1}}{n} + O(n^{-2}).$$

The third result treats coefficients of sample skewness and kurtosis, which are quite popular in statistical analysis. We define the necessary notation:

$$\beta = \frac{\mu_3}{\sigma^3};\ \gamma = \frac{\mu_4}{\sigma^4} - 3;\ b_1 = \frac{m_3}{s^3};\ b_2 = \frac{m_4}{s^4} - 3.$$

Theorem 9.9. For any r, s,

$$(a)\, E[b_1] = \beta + O(n^{-1});\ E[b_2] = \gamma + O(n^{-1}).$$

$$(b)\, \mathrm{Var}(b_1) = \frac{4\mu_2^2\mu_6 - 12\mu_2\mu_3\mu_5 - 24\mu_2^3\mu_4 + 9\mu_3^2\mu_4 + 35\mu_2^2\mu_3^2 + 36\mu_2^5}{4\mu_2^5 n} + O(n^{-2}).$$

$$(c)\, \mathrm{Var}(b_2) = \frac{\mu_2^2\mu_8 - 4\mu_2\mu_4\mu_6 - 8\mu_2^2\mu_3\mu_5 + 4\mu_4^3 - \mu_2^2\mu_4^2 + 16\mu_2\mu_3^2\mu_4 + 16\mu_2^3\mu_3^2}{\mu_2^6 n} + O(n^{-2}).$$

The final result treats very general smooth statistics, but the approximations are only formal. To get mathematically valid error terms, one has to assume stringent conditions on the derivatives and mixed derivatives of the statistic. We cannot call this a theorem. For notational simplicity, we present only the case of statistics $T(X_1, \cdots, X_n)$ where $X_1, \cdots, X_n$ are iid and real valued. With messier notation, these can be extended to X_i

which are vector valued, and not necessarily iid or even independent.

Let

$$T_i = T_i(X_1, \cdots, X_n) = \frac{\partial T}{\partial X_i},\ T_{ij} = T_{ij}(X_1, \cdots, X_n) = \frac{\partial^2 T}{\partial X_i \partial X_j}.$$

Then, we have formal approximations as below; these come out of Taylor expansions of T around the mean vector of $(X_1, \cdots, X_n)$.

Formal Approximation

$$E[T] \approx T(\mu, \cdots, \mu) + \frac{\sigma^2}{2} \sum_{i=1}^{n} T_{ii}(\mu, \cdots, \mu).$$

Formal Approximation Let $T(X_1, \cdots, X_n)$ be a function of the sample mean, $T(X_1, \cdots, X_n) = t(\bar{X})$. Then,

$$\text{Var}(T) \approx \frac{\sigma^2 [t'(\mu)]^2}{n} + \frac{t'(\mu)t''(\mu)\mu_3 + [t''(\mu)]^2 \sigma^4/2}{n^2}.$$

9.6.2 Observed and Expected Fisher Information

Consider the one parameter problem with iid observations from a density $f(x|\theta)$ that is sufficiently regular. According to the theorem on asymptotic normality of MLEs, $\hat{\theta}_n$ is asymptotically normal with mean θ and variance $\frac{1}{nI(\theta)}$.

Therefore, an immediate plug-in estimate of the variance of the MLE (provided the variance is finite) is $\frac{1}{nI(\hat{\theta})}$.

On the other hand, from Chapter 7,

$$I(\theta) = -E_\theta(\frac{\partial^2}{\partial \theta^2} \log f(X|\theta)).$$

Since the observations X_i are iid, by the usual SLLN, the average $\frac{1}{n} \sum_{i=1}^{n} -\frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta) \overset{a.s.}{\Rightarrow} I(\theta)$. Thus, it is also very reasonable to provide the variance estimate

$$\frac{1}{\sum_{i=1}^{n} -\frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta)|_{\theta=\hat{\theta}}}.$$

The quantity $\frac{1}{n} \sum_{i=1}^{n} -\frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta)$ is called the *Observed Fisher Information.* Its expectation, which is just the Fisher information function $I(\theta)$, is also known as the *Expected Fisher Information.* It is natural to ask which gives a better estimate of the true variance of the MLE: $\frac{1}{nI(\hat{\theta})}$, or $\frac{1}{\sum_{i=1}^{n} -\frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta)|_{\theta=\hat{\theta}}}$?

We present two illustrative examples.

Example 9.25. (Observed vs Expected Information in Exponential Family). Suppose $X_1, \ldots, X_n$ are iid from a distribution in the canonical one parameter Exponential family with density $f(x|\theta) = e^{\theta T(x) - \psi(\theta)} h(x)$. Then, $\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) = -\psi''(\theta)$. Thus, $I(\theta)$ and $\frac{1}{n} \sum_{i=1}^{n} -\frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta)$ are both equal to $\psi''(\theta)$, and so use of the observed or the expected Fisher information lead to the same estimate for the variance of $\hat{\theta}_n$. This is a neat general fact.

Example 9.26. (Observed vs Expected Information in Cauchy). Suppose $X_1, \ldots, X_n$ are iid from the Cauchy distribution $C(\theta, 1)$. Then, $f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$, and $\frac{\partial^2}{\partial\theta^2} \log f(x|\theta) = \frac{2((x-\theta)^2-1)}{(1+(x-\theta)^2)^2}$. We already know that $I(\theta) \equiv \frac{1}{2}$ in the Cauchy case. Thus the estimate of the variance of the MLE based on the expected information is $\frac{2}{n}$.
However, this time, the observed information method would produce an estimated variance that depends on the actual observed data. It is not clear which variance estimate is better. The relative comparison of the two estimates of the true variance of the MLE of θ can be examined by simulation. For $n = 20$, and the true θ value equal to 0, a simulation of size 500 was conducted to enquire into the performance of the two variance estimates. The estimate based on the expected Fisher information is $\frac{2}{n} = .1$. The true variance of the MLE when $n = 20$ is .1225, according to the same simulation.
Thus the expected Fisher information method produces an underestimate of the true variance. The variance estimate produced by the observed information method gives an average estimate of .1071 in the simulations. Thus, the underestimation is not as bad as it was for the expected Fisher information method.
However, there is need for caution. Over the 500 different simulations, the smallest variance estimate produced by the observed information method is .0443, while the largest variance estimate produced by this method was .9014. The variance estimate produced by the observed information method is too erratic. To summarize, the variance estimate produced by the observed information method has a smaller bias, but a high variability. It is not clear which variance estimate you should prefer, the expected information estimate or the observed information estimate. Both have drawbacks. This example illustrates the care needed in assessing the accuracy of maximum likelihood estimates; the problem is harder than it first seems to be.

9.6.3 Some Inconsistent MLEs

MLEs can sometimes be inconsistent. This typically happens when there are too many parameters to estimate compared to the volume of the data. The first example of an MLE being inconsistent was provided by Neyman and Scott (1948). It is by now a classic example and is known as the Neyman-Scott example. That first example shocked everyone at the time and sparked a flurry of new examples of inconsistent MLEs including those offered by LeCam (1953) and Basu (1955). We will see that what makes the Neyman-Scott example work is that, compared to the number of parameters, there isn't enough data to kill the bias of the MLE. It is possible to find adjustments to the MLE or suitable Bayesian estimates in many of these problems which do have the consistency property; see Ghosh

(1994) for examples and also some general techniques.

Example 9.27. (Neyman-Scott Example). Let $X_{ij} \quad i = 1, 2, ..., n$ and $j = 1, 2, ..., k$ be independent with $X_{ij} \sim N(\mu_i, \sigma^2)$. We want to estimate the common variance of the groups. By routine calculus, the MLEs are

$$\hat{\mu}_i = \bar{X}_i \text{ and } \hat{\sigma}^2 = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \bar{X}_i)^2.$$

It is the MLE of σ^2 that is inconsistent. Indeed,

$$\hat{\sigma}^2 = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \bar{X}_i)^2 = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{n}\left(\sum_{j=1}^{k}(X_{ij} - \bar{X}_i)^2\right) = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{n}\sigma^2 W_i$$

where the W_i are independent χ^2_{k-1}. By the WLLN,

$$\frac{\sigma^2}{k}\frac{1}{n}\sum_{i=1}^{n}W_i \xrightarrow{P} \frac{\sigma^2}{k}(k-1)$$

Hence, the MLE for σ^2 does not converge to σ^2!
It is the bias of the MLE that is making the estimate inconsistent; if we kill the bias by multiplying by $\frac{k}{k-1}$ the new estimator becomes consistent, i.e., if we "adjust" the MLE and use

$$\frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \bar{X}_i)^2$$

then we recover consistency. In these sorts of problems, where the number of observations and the number of free parameters grow at the same rate, maximum likelihood often runs into problems. However, these problems are hard for any school of thought.

9.6.4 Rao's Score Intervals

The Wald confidence interval is constructed by using the fact that under enough regularity conditions, $\frac{\sqrt{n}(\hat{\theta}-\theta)}{I(\hat{\theta})} \overset{\mathcal{L}}{\Rightarrow} N(0,1)$. In contrast, the score confidence interval of Rao (Rao (1948)) is constructed by using the fact that $\frac{\sqrt{n}(\hat{\theta}-\theta)}{I(\theta)} \overset{\mathcal{L}}{\Rightarrow} N(0,1)$. Thus, the Wald intervals replace $I(\theta)$ by $I(\hat{\theta})$ at the final step. This requires use of Slutsky's theorem, and the score interval avoids that. In exchange, the Wald interval is easier to compute than the score interval. Both intervals are asymptotically correct; i.e., their limiting coverage probabilities are equal to the nominal level that you used. It was shown in Wilks (1938) that among the confidence intervals centered at sufficiently nice point estimators, Rao score intervals are asymptotically the shortest subject to a nominal coverage probability. Here is an example that illustrates the score confidence interval.

Example 9.28. (Score Confidence Interval for Poisson Mean). For ease of notation, we show the case $n = 1$. The case of a general n is exactly the same, and the final result for a general n will be given at the end of this example. Suppose then $X \sim Poi(\lambda)$. If we apply a normal approximation, we get $\frac{X-\lambda}{\sqrt{\lambda}} \approx N(0,1)$. Recall that a standard normal random variable Z has the property $P(-1.96 \leq Z \leq 1.96) = .95$. Since $\frac{X-\lambda}{\sqrt{\lambda}} \approx N(0,1)$, we have

$$P(-1.96 \leq \frac{X-\lambda}{\sqrt{\lambda}} \leq 1.96) \approx .95$$

$$\Leftrightarrow P(\frac{(X-\lambda)^2}{\lambda} \leq 1.96^2) \approx .95$$

$$\Leftrightarrow P((X-\lambda)^2 - 1.96^2\lambda \leq 0) \approx .95$$

$$\Leftrightarrow P(\lambda^2 - \lambda(2X + 1.96^2) + X^2 \leq 0) \approx .95. \ (*)$$

Now the quadratic equation

$$\lambda^2 - \lambda(2X + 1.96^2) + X^2 = 0$$

has the roots

$$\lambda = \lambda_\pm = \frac{(2X + 1.96^2) \pm \sqrt{(2X + 1.96^2)^2 - 4X^2}}{2}$$

$$= \frac{(2X + 1.96^2) \pm \sqrt{14.76 + 15.37X}}{2}$$

$$= (X + 1.92) \pm \sqrt{3.69 + 3.84X}.$$

The quadratic $\lambda^2 - \lambda(2X + 1.96^2) + X^2$ is ≤ 0 when λ is between these two values $\lambda_\pm$. So we can rewrite $(*)$ as

$$P((X + 1.92) - \sqrt{3.69 + 3.84X} \leq \lambda \leq (X + 1.92) + \sqrt{3.69 + 3.84X}) \approx .95 \ (**).$$

The statement $(**)$ says that with approximately 95% probability, λ will fall inside the interval of values

$$(X + 1.92) - \sqrt{3.69 + 3.84X} \leq \lambda \leq (X + 1.92) + \sqrt{3.69 + 3.84X},$$

and so the interval

$$[(X + 1.92) - \sqrt{3.69 + 3.84X}, (X + 1.92) + \sqrt{3.69 + 3.84X}]$$

is an approximate 95% confidence interval for λ. This is the score interval for λ when $n = 1$.

For a general n, you follow exactly the same mathematical steps, but start with the asymptotic result $\frac{\sqrt{n}(\bar{X}-\lambda)}{\lambda} \overset{\mathcal{L}}{\Rightarrow} N(0,1)$. This gives, for a general $\alpha, 0 < \alpha < 1$, at the end, the score interval

$$(\bar{X} + \frac{z_{\alpha/2}^2}{2n}) \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\bar{X} + \frac{z_{\alpha/2}^2}{4n}}.$$

In contrast, the Wald interval is

$$\bar{X} \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\bar{X}}.$$

The formulas are very similar.

In general, the score method *may not produce an interval of values of the parameter*, but a union of some disjoint intervals. In common applications, this does not happen. The endpoints of the score interval in a general regular problem are found by finding two real roots of the equation

$$(\theta - \hat{\theta})^2 - \frac{z_{\alpha/2}^2}{nI(\theta)} = 0.$$

If the roots are $\theta = l_n, u_n$, then the score interval is $l_n \leq \theta \leq u_n$. Work on comparison between the Wald and the score confidence interval may be found in Peers (1971), Bera and Jarque (1981), Boos (1992), Chandra and Mukerjee (1985), and Brown, Cai, and DasGupta (2001, 2002, 2003).

9.6.5 Variance Stabilization

A major use of the Delta theorem is construction of variance stabilizing transformations(VST), a technique that is of fundamental use in many statistical inference problems. In particular, VSTs are useful tools for constructing confidence intervals for unknown parameters.

The general idea is the following. Suppose we want to find a confidence interval for some parameter $\theta \in \mathcal{R}$. If $T_n = T_n(X_1, \cdots, X_n)$ is some natural estimate for θ, e.g., sample mean as an estimate of a population mean, then often the CLT, or some generalization of the CLT, will tell us that

$$\sqrt{n}(T_n - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, \sigma^2(\theta)),$$

for some suitable function $\sigma^2(\theta)$. Therefore, by the Delta theorem, if $g(T_n)$ is a smooth transformation of T_n, then

$$\sqrt{n}\,(g(T_n) - g(\theta)) \quad \overset{\mathcal{L}}{\Rightarrow} N(0, [g'(\theta)]^2\sigma^2(\theta))$$

Consequently, if we choose g to make

$$[g'(\theta)]^2\sigma^2(\theta) \quad = \quad k^2$$

for some constant k, then

$$\sqrt{n}\,(g(T_n) - g(\theta)) \overset{\mathcal{L}}{\Rightarrow} N(0, k^2),$$

the important point being the variance in the limiting distribution is now a constant, k^2, and *does not depend on* θ. Hence, by using our usual arguments, we will get an approximate confidence interval for $g(\theta)$:

$$g(T_n) \pm z_{\frac{\alpha}{2}} \frac{k}{\sqrt{n}}.$$

By *retransforming* back these endpoints to θ, we get a *new type* of confidence interval for θ:

$$g^{-1}(g(T_n) - z_{\frac{\alpha}{2}} \frac{k}{\sqrt{n}}) \leq \theta \leq g^{-1}(g(T_n) + z_{\frac{\alpha}{2}} \frac{k}{\sqrt{n}}).$$

The reason that this one is sometimes preferred to the Wald confidence interval is that unlike the latter, we do not need to use a variance estimate at the final stage. In the VST method, the asymptotic variance is already a constant and does not need to be estimated. The transformation $g(T_n)$ is obtained from its defining property

$$[g'(\theta)]^2\sigma^2(\theta) = k^2.$$

Simple manipulation of this equation leads to

$$g(\theta) = k \int \frac{1}{\sigma(\theta)} d\theta,$$

where the integral is to be interpreted as a primitive (indefinite integral). The constant k can be chosen as any nonzero real number, and $g(T_n)$ is called a *variance stabilizing transformation* (VST). Some recent references are DiCcicio, Monti and Young (2006), Politis (2003), and DasGupta (2008).

Let us see some examples.

Example 9.29. (VST in Binomial Case). Suppose $X_n \sim Bin(n, p)$. Then $\sqrt{n}(X_n/n - p) \overset{\mathcal{L}}{\Rightarrow} N(0, p(1-p))$. So using the notation used above, $\sigma(p) = \sqrt{p(1-p)}$ and consequently, on taking $k = \frac{1}{2}$,

$$g(p) \;=\; \int \frac{1/2}{\sqrt{p(1-p)}} dp \;=\; \arcsin(\sqrt{p})$$

Hence, $g(X_n) = \arcsin(\sqrt{X_n/n})$ is a variance stabilizing transformation and indeed,

$$\sqrt{n}\left(\arcsin\left(\sqrt{\frac{X_n}{n}}\right) - \arcsin\left(\sqrt{p}\right)\right) \overset{\mathcal{L}}{\Rightarrow} N\left(0, \tfrac{1}{4}\right)$$

Thus, a confidence interval for p is

$$\sin^2\left(\arcsin\left(\sqrt{\frac{X_n}{n}}\right) \mp \frac{z_{\alpha/2}}{2\sqrt{n}}\right)$$

Example 9.30. (VST in Poisson Distribution). Suppose $X_1, X_2, \ldots$ are iid $Poisson(\lambda)$. Since $\sqrt{n}(\bar{X} - \lambda) \overset{\mathcal{L}}{\Rightarrow} N(0, \lambda)$, we have $\sigma^2(\lambda) = \lambda$, and therefore, a variance stabilizing transformation is

$$g(\lambda) = \int \frac{k}{\sqrt{\lambda}} d\lambda = 2k\sqrt{\lambda}$$

Taking $k = 1/2$ gives that $g(\lambda) = \sqrt{\lambda}$ is a variance stabilization transformation in the Poisson case. The result is:

$$\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \overset{\mathcal{L}}{\Rightarrow} N(0, \frac{1}{4}).$$

Thus, a VST based confidence interval for $\sqrt{\lambda}$ is

$$\sqrt{\bar{X}} \quad \pm \quad \frac{z_{\alpha/2}}{2\sqrt{n}}.$$

Retransforming these limits, a VST based confidence interval for λ is

$$\left(\left(\sqrt{\bar{X}} - \frac{z_{\alpha/2}}{2\sqrt{n}} \right)^2, \left(\sqrt{\bar{X}} + \frac{z_{\alpha/2}}{2\sqrt{n}} \right)^2 \right)$$

If $\sqrt{\bar{X}} - z_{\alpha/2}/(2\sqrt{n}) < 0$, that expression should be replaced by 0, because λ cannot be negative.

This confidence interval is different from the Wald as well as the score confidence interval for λ. The coverage probabilities of the interval based on the VST are significantly better than those of the Wald interval; the comparison with the score interval is less clear. See Brown, Cai, and DasGupta (2003).

Example 9.31. (Fisher's z). Suppose (X_i, Y_i), $i = 1, \ldots, n$, are iid bivariate normal with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$. Then, it may be shown by a tedious use of the multivariate delta theorem that $\sqrt{n}(r_n - \rho) \overset{\mathcal{L}}{\Rightarrow} N(0, (1-\rho^2)^2)$, r_n being the sample correlation coefficient. Therefore,

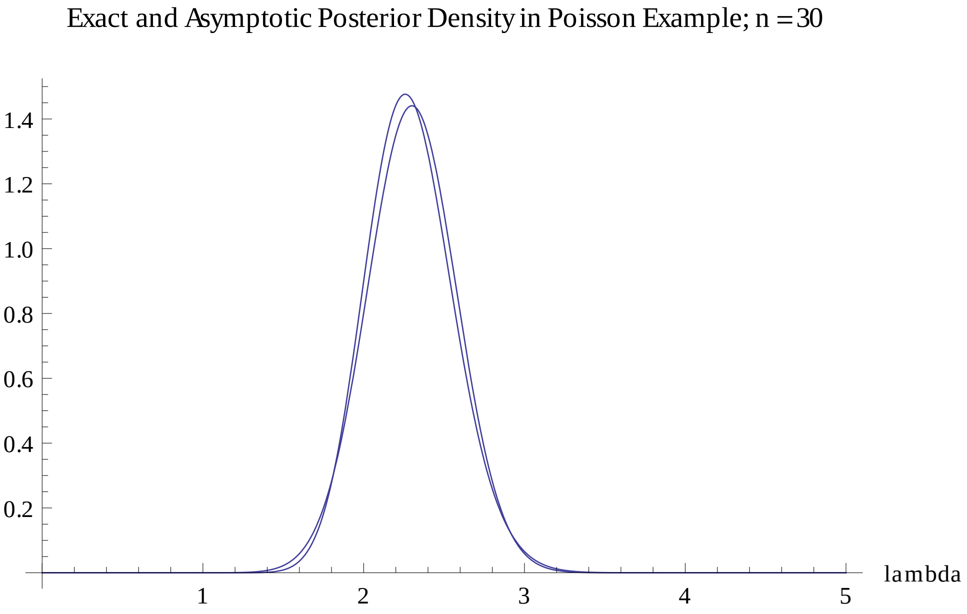
$$g(\rho) \quad = \quad \int \frac{1}{(1-\rho)^2} d\rho = \frac{1}{2} \log \frac{1+\rho}{1-\rho} = \operatorname{arctanh}(\rho)$$

provides a variance stabilizing transformation for r_n.

This is the famous arctanh transformation of Fisher, popularly known as *Fisher's z*. It follows directly from its construction that $\sqrt{n}(\operatorname{arctanh}(r_n) - \operatorname{arctanh}(\rho))$ converges in distribution to the $N(0, 1)$ distribution. And so, we can calculate confidence intervals for ρ:

$$\tanh\left(\operatorname{arctanh}(r_n) \pm \frac{z_{\alpha/2}}{\sqrt{n}} \right).$$

This confidence interval for ρ is preferred to the Wald confidence interval for ρ, because the VST based interval's coverage probabilities are quite a bit better than the coverage probabilities of the Wald interval. Fisher's z is a classic in statistics.



Exact and Asymptotic Posterior Density in Poisson Example; n = 30

$\bar{X} = 2.3$. The exact posterior density of λ is a Gamma, namely $G(71, \frac{1}{31})$. On the other hand, the posterior asymptotic normality theorem suggests that the posterior density is approximately $N(\hat{\lambda}, \frac{1}{nb^2(\hat{\lambda})})$. Recall that in any Exponential family case, the expected and the observed Fisher information coincide, so that

$$b^2(\hat{\lambda}) = I(\bar{X}) = \frac{1}{\bar{X}}.$$

This gives the normal approximation

$$\lambda \approx N(2.3, .2769).$$

The exact posterior density and the normal approximation are quite close, as you can see from the plot.

9.6.9 Asymptotic Optimality of MLEs and Work of Le Cam

It was first believed as a folklore that the MLE under regularity conditions on the underlying distribution is asymptotically the best for every value of θ, i.e. if a MLE $\hat{\theta}$ exists and , $\sqrt{n}(\hat{\theta} - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, I^{-1}(\theta))$, and if another competing sequence T_n satisfies $\sqrt{n}(T_n - \theta) \overset{\mathcal{L}}{\Rightarrow} N(0, V(\theta))$, then for every θ, $V(\theta) \geq \frac{1}{I(\theta)}$.

The classic Hodges' estimator (Hodges, 1951, unpublished) of a one dimensional normal mean demolished this statistical folklore that maximum likelihood estimates are asymptotically uniformly optimal provided the family of underlying densities satisfies enough regularity conditions. Hodges produced an estimate T_n for which the variance function $V(\theta)$ satisfies $V(0) < \frac{1}{I(0)}$, and $V(\theta) = \frac{1}{I(\theta)}$ at any $\theta \neq 0$. Thus, Hodges' estimate T_n is *superefficient* at $\theta = 0$.

Hodges' original estimate is

$$T_n(X_1, \cdots, X_n) = \begin{cases} \bar{X}_n & \text{if} \quad |\bar{X}_n| > n^{-1/4} \\ 0 & \text{if} \quad |\bar{X}_n| \le n^{-1/4} \end{cases}$$

A more general version is

$$S_n(X_1, \ldots, X_n) = \begin{cases} \bar{X}_n & \text{if} \quad |\bar{X}_n| > c_n \\ a_n\bar{X}_n & \text{if} \quad |\bar{X}_n| \le c_n \end{cases}$$

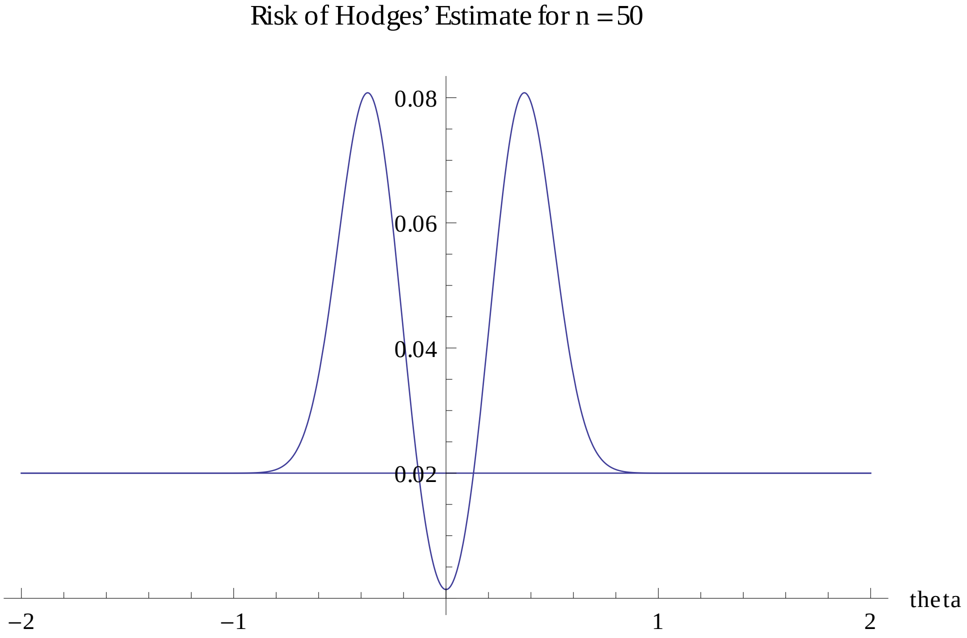
It is clear however that T_n has certain undesirable features. As a function of $X_1, ..., X_n$, T_n is not smooth. Nevertheless, with squared error as the loss function, the risk of $\bar{X}_n$, the unique MLE, satisfies $nR(\theta, \bar{X}_n) \equiv 1$, and Hodges' estimate satisfies

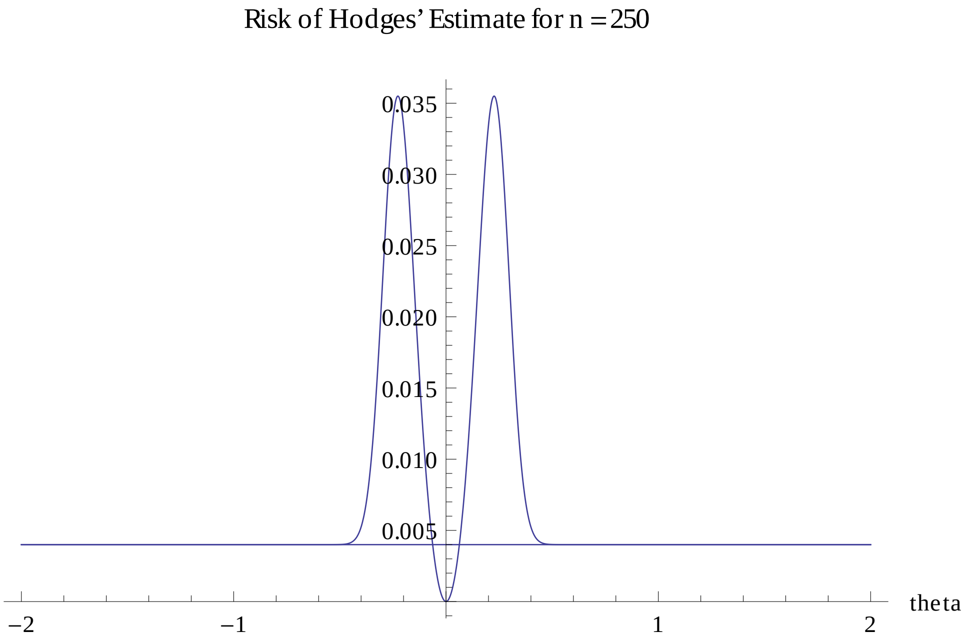
$$\lim_{n\to\infty} n^{\beta} R(0, T_n) = 0 \; \forall \, \beta > 0,$$

while

$$\lim_{n\to\infty} \sup_{\theta} \; nR(\theta, T_n) = \infty.$$

Thus, at $\theta = 0$, Hodges' estimate is asymptotically infinitely superior to the MLE, while globally its peak risk is infinitely more relative to that of the MLE. *Superefficiency at $\theta = 0$ is purchased at a price of infinite asymptotic inflation in risk away from zero.* Indeed, a plot of the risk function of Hodges' estimate nicely illustrates these interesting and intricate risk phenomena.





Example 9.34. (Maximum Risk of Hodges' Estimate). We compute the maximum risk of Hodges' original estimate T_n for some selected n. We can see in the tabulated maximum risks that they are attained at $\theta \approx c_n = n^{-1/4}$.

n	Exact Maximum	$R(c_n, T_n)$
100	.0558	.0550
2500	.0126	.0102
100000	.0025	.0016
250000	.0016	.0010
10^6	.0008	.0005

We can do more than numerical work. The precise behavior of the risk function of the MLE and Hodges' estimate was investigated in DasGupta and Johnstone (2012). Among their various results, they proved the following theorem.

Theorem 9.12. In the iid $N(\theta, 1)$ case with a squared error loss function, the risk function $R(\theta, T_n)$ of Hodges' estimate satisfies

$$(a)\ R(0, T_n) = \sqrt{\frac{2}{\pi}} n^{-3/4} e^{-\frac{\sqrt{n}}{2}} + o(n^{-3/4} e^{-\frac{\sqrt{n}}{2}}).$$

$$(b)\ \sup_{\theta} R(\theta, T_n) \geq \frac{1}{\sqrt{n}} - \frac{2\sqrt{\log n}}{n^{3/4}}.$$

$$(c)\ \text{argmax} R(\theta, T_n) \approx n^{-1/4} - \sqrt{\frac{\log n}{2n}}.$$

The interpretation of the three parts of this theorem is that at $\theta = 0$, Hodges' estimate has a relatively much lower risk than the risk $\frac{1}{n}$ of the MLE; at $\theta \approx n^{-1/4}$, it has a relatively much higher risk of about $\frac{1}{\sqrt{n}}$ in comparison to the MLE, and $\frac{1}{\sqrt{n}}$ is roughly the maximum risk of Hodges' estimate, which of course is relatively much larger than $\frac{1}{n}$, the maximum risk of the MLE.

So, again, Hodges' example showed that the claim of the uniform asymptotic optimality of the MLE is false even in the normal case, and it seeded the development of such fundamental concepts as regular estimates. It culminated in the celebrated *Hájek-Le Cam convolution theorem* (Le Cam (1953, 1973), Hájek (1970)).

In an extremely deep and insightful result, Le Cam (Le Cam (1953)) showed that *superefficiency can only occur on null sets of θ values.* That is, the set of all θ values at which an estimate T_n can have strictly smaller asymptotic variance than the MLE must be a set of *Lebesgue measure zero.* If, in addition, we insist on using only such estimates T_n that have a certain smoothness property, then there cannot be any points of superefficiency at all. Thus, *in regular problems, in the class of sufficiently smooth estimates, the MLE is asymptotically the best at every possible θ value. This is what people really mean when they say informally that MLEs are asymptotically optimal.*

So, to justify the folklore that MLEs are asymptotically the best, one not only needs regularity conditions on $f(x|\theta)$, but one also must restrict attention to only those estimates that are adequately nice (and Hodges' estimate is not). Further details of Le Cam's results are available in van der Vaart (1998) and DasGupta (2008).

9.7 Exercises

Exercise 9.1. (Consistent Estimates of Binomial p). Let $X \sim Bin(n, p)$. Prove that each of the following estimates is a consistent estimate of p;

$$\frac{(n+1)X}{n^2+n+4}; \ \frac{X}{n+2} + \frac{1}{n}; \ \frac{X-100}{n+5}; \ \frac{X-100}{n+\sqrt{n}}.$$

Exercise 9.2. Let $X_1, X_2, \cdots$ be iid $Exp(\lambda)$. Give two consistent estimates of $e^{-\lambda} - e^{-2\lambda}$.

Exercise 9.3. (Consistent Estimation of Magnitude). Let $X_1, X_2, \cdots$ be iid $N(\mu, 1)$. Prove or disprove that $|\bar{X}|$ is a consistent estimate of $|\mu|$.

Exercise 9.4. (Numerical Work). Simulate 200 observations from a standard Cauchy, and compute the mean for the first m observations, $m = 10, 20, 30, \cdots, 200$. Plot these means against m. What do you see in the plot?

Exercise 9.5. (Consistent Estimation of Right Tail Probability). Survival times of patients afflicted with a disease have a Gamma distribution with parameters $\alpha = 2$ and an unknown λ. Give a consistent estimate of each of:

$$\lambda;\ P_\lambda(X > 1).$$

Exercise 9.6. (Consistent Estimation of Density at a Point). Let $X_1, X_2, \cdots$ be iid $N(\mu, 1)$. Find a consistent estimate of the density of X_1 at the point $x = 0$.

Exercise 9.7. (Consistent Estimates of Binomial Variance). Let $X \sim Bin(n, p)$. Give a consistent estimate of $p(1-p)$.

Exercise 9.8. (Consistent Estimation in Hardy-Weinberg Model). According to Mendel's law, the genotypes aa, Aa, and AA in a population with genetic equilibrium with respect to a single gene having two alleles have proportions $\theta^2, 2\theta(1-\theta), (1-\theta)^2$. Suppose n individuals are sampled from the population and the number of observed individuals of each genotype are n_1, n_2, n_3 respectively. Find the MLE of θ and show that it is consistent.

Exercise 9.9. (A Little Hard). Let $X_1, X_2, \cdots$ be iid $Exp(\lambda)$. Show that $\frac{X_{(n)}}{\log n}$ is a consistent estimate of λ.

Exercise 9.10. (Consistency of Moment Estimates).
(a) Find expressions for the moment estimates of the two parameters of a Gamma distribution.
(b) Show that each is consistent.

Exercise 9.11. (Consistency in Nonregular Case). Let $X_1, X_2, \cdots$ be iid $U[\mu-\sigma, \mu+\sigma]$. Find two consistent estimates each of each of μ, σ.

Exercise 9.12. (Consistent Estimation of Correlation Coefficient). Suppose $(X_i, Y_i), i = 1, 2\cdots$ are iid observations from a general bivariate distribution, for which $\text{Var}(X), \text{Var}(Y)$ are finite. Find a consistent estimate for the correlation between X and Y.

Exercise 9.13. (Conceptual). Let $X_1, X_2, \cdots$ be iid $N(\mu, 1)$. But what is recorded are $Y_1, Y_2. \cdots$, where $Y_j = |X_j|$. Which of the following can be consistently estimated?
μ; $|\mu|$; $P_\mu(-1 \leq X_1 \leq 1)$.

Exercise 9.14. (Consistency of UMVUE). Let $X_1, X_2, \cdots$ be iid $N(\mu, \sigma^2)$, both μ, σ being unknown. Show that the UMVUE of μ^2 is consistent.

Exercise 9.15. (Consistency of UMVUE). Let $X_1, X_2, \cdots$ be iid $N(\mu, \sigma^2)$, both μ, σ being unknown. Show that the UMVUE of $\mu + \sigma$ is consistent.

Exercise 9.16. (A Little Difficult). Let $X_1, X_2, \cdots$ be iid $Poi(\lambda)$. Show that the UMVUE of $e^{-\lambda}$ is consistent.

Exercise 9.17. (Numerical Work). For $n = 20$, plot the exact density and the density of the asymptotic distribution of the MLE of σ in the $N(0, \sigma^2)$ case.

Exercise 9.18. (General Result). Suppose T_n is a sequence of estimates of some parameter θ such that $E_\theta(T_n) - \theta$ and $\text{Var}_\theta(T_n)$ both go to zero for any θ. Show that T_n is consistent.

Exercise 9.19. (An Example of Basu). Suppose X_i are iid $N(\mu, 1)$ where μ is known to be a positive integer. Let $g : \mathcal{R} \to \mathcal{R}$ be the function

$$g(x) = \begin{cases} x & \text{if } x \text{ is a prime} \\ -x & \text{if } x \text{ is not a prime} \end{cases}$$

(a) Is $\bar{X}$ consistent for μ?

(b) Is $g(\bar{X})$ consistent for $g(\mu)$?

Exercise 9.20. (Consistency Lost). Suppose X_i are iid $N(\mu, 1)$, but the collector rounds the X_i to Y_i, the nearest integer. Is $\bar{Y}$ consistent for μ?

Exercise 9.21. (Consistency in Multinomial). Let $(X_1, \cdots, X_{k+1})$ have a general multinomial distribution with parameters $p_1, \cdots, p_k$, and $p_{k+1} = 1 - p_1 - \cdots - p_k$. Find a consistent estimate of $\max\{p_1, \cdots, p_k, p_{k+1}\}$.

Exercise 9.22. (Variance of Sample Variance). Use the exact formula of Example 9.21 to find the variance of the sample variance when observations are iid from a χ^2 distribution with m degrees of freedom.

Exercise 9.23. (Correlation Between Mean and Variance). Use the previous exercise and part (b) of Theorem 9.5 to find the correlation between the sample mean and the sample variance when observations are iid from a χ^2 distribution with 4 degrees of freedom and $n = 15$.

Exercise 9.24. (Variance of Skewness Coefficient). Use the exact formula in part (b) of Theorem 9.7 to find the approximate variance of the sample skewness coefficient when observations are iid from a standard double exponential distribution.

Exercise 9.25. (Variance of Sample Variance in Poisson). Use the exact formula of Example 9.21 to find the variance of the sample variance when observations are iid from a Poisson distribution with mean λ.

Exercise 9.26. (Asymptotic Distribution of MLE). Suppose X_i are iid $N(\theta, \theta)$.
(a) Carefully derive the asymptotic distribution of the MLE of θ.
(b) Convert the asymptotic distribution into the Wald confidence interval for θ.
(c) Convert the asymptotic distribution into the Rao score interval for θ.

Exercise 9.27. (Asymptotic Distribution of MLE). Suppose X_i are iid $Beta(\alpha, \alpha)$.
(a) Derive the asymptotic distribution of the MLE of α.
(b) Convert the asymptotic distribution into the Wald confidence interval for α.
(c) Convert the asymptotic distribution into the Rao score interval for α.

Exercise 9.28. (Asymptotic Distribution of MLE in Nonregular Problem). Suppose $X_1, X_2, \cdots$ are p - vectors uniformly distributed in the ball $B_r = \{x : ||x||_2 \leq r\}; r > 0$ is an unknown parameter. Find the MLE of r and its asymptotic distribution.

Exercise 9.29. (Truncated Data Asymptotics). The number of fires reported in a week to a city fire station is Poisson with some mean λ. The city station is supposed to report the number each week to the central state office. But they do not bother to report it if their number of reports is less than 2.
Suppose you are employed at the state central office and want to estimate λ. Model the problem.
(a) Find the asymptotic distribution of the MLE of λ.
(b) Is this MLE computable in closed form?

Exercise 9.30. (Asymptotic Distribution of MLE of Common Mean). $X_1, X_2, \cdots, X_m$ are iid $N(\mu, \sigma_1^2)$ and $Y_1, Y_2, \cdots, Y_n$ are iid $N(\mu, \sigma_2^2)$, and all $m + n$ observations are independent. All three parameters are unknown.
(a) Find the Fisher information matrix.
(b) Find its inverse.
(c) Find the asymptotic distribution of the MLE of μ.
(d) Convert it into a confidence interval for μ.

Exercise 9.31. (Asymptotic Distribution in Double Exponential). Let $X_1, X_2, \cdots,$ be iid from a location parameter double exponential density.
(a) Find the asymptotic distribution of the MLE of the location parameter.
(b) Convert it into a confidence interval.

Exercise 9.32. (Asymptotic Distribution in Cauchy). Let $X_1, X_2, \cdots,$ be iid from a scale parameter Cauchy distribution.
(a) Find the asymptotic distribution of the MLE of the scale parameter.
(b) Convert it into a confidence interval.

Exercise 9.33. (Difficulties in Constrained Parameter Problems). Let $X_1, X_2, \cdots,$ be iid from $N(\mu, 1)$. Suppose it is known that $\mu \leq 1$. Show that the MLE of μ does not have an asymptotically normal distribution if the true $\mu = 1$.

Exercise 9.34. (Calculation of Asymptotic Efficiency). Find the asymptotic efficiency of the sample mean with respect to the MLE in the location parameter double exponential case.

Exercise 9.35. (Asymptotic Efficiency of Bayes Estimates).
Let $X_1, X_2, \cdots,$ be iid from $N(\mu, 1)$. Consider the posterior mean with respect to a general $N(\eta, \tau^2)$ prior. Find its asymptotic efficiency with respect to the MLE of μ.

Exercise 9.36. (Conceptual). Let $X_1, X_2, \cdots,$ be iid from $N(\mu, 1)$. Consider the unusual estimate of μ that equals the sample mean if n is even and equals the sample median if n is odd.
(a) Is this estimate consistent?
(b) Is this estimate asymptotically normal?

Exercise 9.37. (Calculation of Asymptotic Efficiency).
Find the asymptotic efficiency of the sample median with respect to the MLE in the location parameter Cauchy case.

Exercise 9.38. (Asymptotic Efficiency of Moment Estimates).
Let X_i be iid from $f(x\,|\theta) = \theta x^{\theta-1}, 0 < x < 1, \theta > 0$.
(a) Find a moment estimate of θ.
(b) Find the asymptotic efficiency of the moment estimate with respect to the MLE of θ.

Exercise 9.39. (Asymptotic Efficiency and Basu's Theorem).
Let X_i be iid from $N(\mu, 1)$. Let $\bar{X}$ be the mean and M_n the median of the first n observations.
(a) Find the asymptotic variance of $\frac{1}{2}\bar{X} + \frac{1}{2}M_n$ by using Basu's theorem.
b) Find the asymptotic efficiency of this estimate of μ with respect to $\bar{X}$.

Exercise 9.40. (Finding Pivots). Let X_i be iid from $U[-\theta, \theta]$.
(a) Show that $\frac{X_{(n)} - X_{(1)}}{\theta}$ is an exact pivot.
(b) Find the MLE of θ and make a pivot out of it.
(c) Use this pivot in part (b) to make a confidence interval for θ.

Exercise 9.41. (Finding Pivots). Let X_i be iid from $C(0, \sigma)$.
(a) Find an exact pivot.
(c) Use this pivot to make a confidence interval for σ.

Exercise 9.42. (Confidence Interval for Geometric Parameter). Let X_i be iid from a geometric distribution with pmf $f(x\,|\theta) = \theta(1-\theta)^{x-1}, x = 1, 2, \cdots$.
(a) Calculate the Fisher information function.
(b) Hence, find the Wald confidence interval for θ.
(c) Try to also find the score confidence interval for θ.

Exercise 9.43. (Score Confidence Interval for Binomial p). Let $X \sim Bin(n, p)$.
(a) Find the score confidence interval for p.

(b) Suppose $n = 100, X = 35$. Compute both the Wald confidence interval and the score confidence interval for p.
(c) Comment on how the two computed intervals differ and how much they differ.

Exercise 9.44. (Delta Theorem). Suppose $X_1, X_2, \ldots$ are iid with mean μ and variance σ^2, a finite fourth moment, and let $Z \sim N(0, 1)$.

(a) Show that $\sqrt{n}(\overline{X}^2 - \mu^2) \overset{\mathcal{L}}{\Rightarrow} 2\mu\sigma Z$.

(b) Show that $\sqrt{n}(e^{\overline{X}} - e^{\mu}) \overset{\mathcal{L}}{\Rightarrow} e^{\mu} Z$.

Exercise 9.45. (Delta Theorem). Suppose $X_1, X_2, \cdots$ are iid $Poi(\lambda)$. Find the limiting distribution of $\bar{X}e^{-\bar{X}}$. Why would you at all be interested in $\bar{X}e^{-\bar{X}}$?

Exercise 9.46. (Delta Theorem). Suppose $X_1, X_2, \cdots$ are iid $N(\mu, 1)$.
(a) Find the MLE of $P_\mu(X_1 > 1)$.
(b) Find its limiting distribution.
(c) Convert the answer in part (b) to a confidence interval for $P_\mu(X_1 > 1)$.

Exercise 9.47. (Asymptotic Distribution of Sample Skewness). Suppose $X_1, X_2, \cdots$ are iid $N(\mu, \sigma^2)$. Let b_1 be the sample skewness coefficient:

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3}{s^3}.$$

By using the multivariate delta theorem, prove that $\sqrt{n}b_1 \overset{\mathcal{L}}{\Rightarrow} N(0, 6)$.

Exercise 9.48. (Asymptotic Distribution of Sample Kurtosis). Suppose $X_1, X_2, \cdots$ are iid $N(\mu, \sigma^2)$. Let b_2 be the sample kurtosis coefficient:

$$b_2 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^4}{s^4} - 3.$$

By using the multivariate delta theorem, prove that $\sqrt{n}b_2 \overset{\mathcal{L}}{\Rightarrow} N(0, 24)$.

Exercise 9.49. (Delta Theorem in a Nonregular Case). Suppose $X_1, X_2, \cdots$ are iid $U[0, \theta]$. Let $T_n = X_{(n)}$ and $g(T_n) = \log T_n$. Find the limiting distribution of $g(T_n)$.

Exercise 9.50. (Anscombe Transformation in Poisson).
Let $X \sim Poi(\lambda)$ and b a fixed positive number.
(a) Show that if λ is large, then

$$E[\sqrt{X + b}] \approx \sqrt{\lambda + b} - \frac{1}{8\sqrt{\lambda}} + \frac{24b - 7}{128\,\lambda^{3/2}}.$$

(b) Now replace X by $\sum_{i=1}^n X_i$, λ by $n\lambda$, and then eventually divide by $\sqrt{n}$ to write an approximation for $E[\sqrt{\bar{X} + \frac{b}{n}}]$ when you have n iid Poissons with a fixed mean λ and b is still a fixed positive number.
(c) Rediscover the bias-corrected Poisson VST from part (b).

Exercise 9.51. (Anscombe Transformation in Poisson).
(a) Similar to the previous exercise, show that if $X \sim Poi(\lambda)$ and b a fixed positive number, then for large λ,

$$\mathrm{Var}[\sqrt{X+b}] \approx \frac{1}{4}\left[1 + \frac{3-8b}{8\,\lambda} + \frac{32b^2 - 52b + 17}{32\,\lambda^2}\right].$$

(b) Hence, rediscover the Anscombe transformation in the case of Poisson.

Exercise 9.52. (Posterior Mean Approximation in Binomial). Suppose $X \sim Bin(n,p)$ and p has a prior density $\rho(p) = c\sin(\pi\,p)$. Use the Johnson expansions in Theorem 9.8 to approximate the posterior mean of p.

Exercise 9.53. (Posterior Mean under t-Priors). Suppose $X_1, X_2, \cdots$ are iid $N(\mu, 1)$ and that μ has a t-prior with center at zero, scale parameter one, and $\alpha > 2$ degrees of freedom. Calculate the Johnson expansions in Theorem 9.8 to approximate the posterior mean of μ.

Exercise 9.54. (Asymptotic Posterior Normality). Suppose $X_1, X_2, \cdots$ are iid $N(\mu, 1)$ and that it is known that $\mu > 0$. Suppose μ has a standard Exponential prior.
(a) Generate $n = 15, 30$ observations from $N(.5, 1)$.
(b) Plot the exact posterior density corresponding to the simulated data.
(c) Plot the normal approximation to the posterior density.
(d) Comment.

Exercise 9.55. (Formula for the Risk Function of Hodges' Estimate). Let $X_1, X_2, \cdots$ be iid $N(\mu, 1)$ and let T_n be the Hodges' estimator as defined in text.
(a) Show that the MSE of T_n has the formula

$$R(\theta, T_n) = \frac{1}{n} + e_n(\theta),$$

where

$$e_n(\theta) = \left[\theta^2 - \frac{1}{n}\right]\Big(\Phi(\sqrt{n}(c_n - \theta)) + \Phi(\sqrt{n}(c_n + \theta)) - 1\Big)$$
$$+\frac{1}{\sqrt{n}}\Big((c_n + \theta)\phi(\sqrt{n}(c_n + \theta)) + (c_n - \theta)\phi(\sqrt{n}(c_n - \theta))\Big),$$

where $c_n = n^{-1/4}$.
(b) Use this exact formula to compute and plot $R(\theta, T_n)$ for $n = 25, 100$.
(c) Visually identify the interval of θ-values for which $R(\theta, T_n) \leq \frac{1}{n}$.
(d) Is this interval smaller when $n = 25$ or $n = 100$?

9.8 References

Anscombe, F. (1948). Transformation of Poisson,Binomial and Negative Binomial Data, Biometrika, 35, 246-254.

Banerjee, M. (2008). Estimating monotone, unimodal, and U-shaped hazards by using asymptotic pivots, Statist. Sinica, 18, 467-492.

Basu, D. (1955). An inconsistency of the method of maximum likelihood, Ann. Math. Statist., 26, 144–145.

Bera, A. and Jarque, C. (1981). An efficient large sample test for normality and regression residuals, Working Papers in Economics and Econometrics, 40, ANU, Canberra, Australia.

Bickel, P.J. and Doksum, K. (2006).Mathematical Statistics, Basic Ideas and Selected Topics, Prentice Hall, Upper Saddle River, NJ.

Boos, D. (1992). On generalized score tests, Amer. Statist., 46, 327-333.

Brown, L., Cai, T., and DasGupta, A. (2001). Interval estimation for a binomial proportion, Statist.Sc., 16, 101-133.

Brown, L., Cai, T., and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions, Ann. Statist., 30, 160-201.

Brown, L., Cai, T., and DasGupta, A. (2003). Interval estimation in exponential families, Statist. Sinica, 13, 19-49.

Brown, L., Cai, T., and DasGupta, A. (2006). On selecting an optimal transformation, Preprint.

Brown, L., Cai, T., and Zhou, H. (2010). Nonparametric regression in exponential families, Ann. Statist., 38, 2005-2046.

Chandra, T. and Mukerjee, R. (1985). Comparison of the likelihood ratio, Wald, and Rao's tests, Sankhya, Ser. A, 47, 271-284.

Cramér, H. (1946). Mathematical Methods of Statistics, Princeton University Press, NJ.

DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability, Springer, New York.

DasGupta, A. and Johnstone, I.M. (2012). Risks and Bayes risks of thresholding and superefficient estimates and optimal thresholding, Preprint.

DiCiccio, T., Monti, A., and Young, G.A. (2006). Variance stabilization for a scalar parameter, JRSS, Ser. B, 281-303.

Ghosal, S. and Samanta, T. (1997). Asymptotic expansions of posterior distributions in nonregular cases, Ann. Inst. Statist. Math., 49, 181-197.

Ghosh, J.K. (1994). Higher Order Asymptotics, IMS, Beachwood, OH.

Ghosh, M. (1994). On some Bayesian solutions of Neyman-Scott problems, Stat. Dec. Theory Rel. Topics, S.S. Gupta and J. Berger Eds., 267-276.

Gordon, I. and Hall, P. (2009). Estimating a parameter when it is known that the parameter exceeds a given value, Austr. N. Z. J. Stat., 51, 449-460.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates, Z. Wahr. verw. Geb.,14, 323-330.

Hall, P. (1992). On the removal of skewness by transformation, JRSS, B, 54, 221-228.

Hartigan, J. (1983). Bayes theory, Springer-Verlag, New York.

Jiang, J. (2010). Large Sample Techniques for Statistics, Springer, New York.

Johnson, R.A. (1970). Asymptotic expansions associated with posterior distributions, Ann. Math. Statist., 41, 851-864.

Lawley, D. (1956). A general method of approximating distribution of likelihood ratio criteria, Biometrika, 43, 295-303.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, Univ. Calif. Publ., 1, 277-330.

Le Cam, L. (1973). Sur les contraintes imposees par les passages a la limite usuels en statistique, Proc. 39th session International Statistical Institute, XLV, 169-177.

LeCam, L. (1986). Asymptotic Methods in Statistical Decision Theory, Springer, New York.

Lehmann, E.L. (1998). Elements of Large Sample Theory, Springer, New York.

Lehmann, E.L. and Casella, G. (1998). Theory of Point Estimation, Springer, New York.

Mukhopadhyay, S. and DasGupta, A. (1997). Uniform approximations of Bayes solutions and posteriors: Frequentistly valid Bayes inference, Stat. Decisions, 15, 51-73.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistet observations, Econometrica, 16, 1-32.

Peers, H. (1971). Likelihood ratio and associated tests criteria, Biometrika, 58, 577-587.

Politis, D. (2003). A normalizing and variance stabilizing transformation for financial time series, Recent Advances and Trends in Nonparametric Statistics, 335-347, Elsevier, Amsterdam.

Rao, C.R. (1948). Large sample tests of statistical hypotheses and applications to problems of estimation, Proc. Cambridge Phil. Soc., 44, 50-57.

Serfling, R. (1980). Approximation Theorems of Mathematical Statistics, Wiley, New York.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities, JASA, 81, 82-86.

van der Vaart, Aad (1998). Asymptotic Statistics, Cambridge University Press, Cambridge.

Wilks, S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Statist., 9, 60-62.

Wilks, S. (1938). Shortest average confidence intervals for large samples, Ann. Math. Statist., 9, 166-175.