Lecture Notes 9
Asymptotic Theory (Chapter 9)

In these notes we look at the large sample properties of estimators, especially the maximum likelihood estimator.

Some Notation: Recall that

$$\mathbb{E}_\theta(g(X)) \equiv \int g(x)p(x;\theta)dx.$$

1 Review of o, O, etc.

1. $a_n = o(1)$ mean $a_n \to 0$ as $n \to \infty$.
2. A random sequence A_n is $o_p(1)$ if $A_n \xrightarrow{\text{P}} 0$ as $n \to \infty$.
3. A random sequence A_n is $o_p(b_n)$ if $A_n/b_n \xrightarrow{\text{P}} 0$ as $n \to \infty$.
4. $n^b\, o_p(1) = o_p(n^b)$, so $\sqrt{n}\, o_p(1/\sqrt{n}) = o_p(1) \xrightarrow{\text{P}} 0$.
5. $o_p(1) \times o_p(1) = o_p(1)$.
6. $a_n = O(1)$ if $|a_n|$ is bounded by a constant as $n \to \infty$.
7. A random sequence Y_n is $O_p(1)$ if for every $\epsilon > 0$ there exists a constant M such that $\lim_{n\to\infty} P(|Y_n| > M) < \epsilon$ as $n \to \infty$.
8. A random sequence Y_n is $O_p(b_n)$ if Y_n/b_n is $O_p(1)$.
9. If $Y_n \rightsquigarrow Y$, then Y_n is $O_p(1)$.
10. If $\sqrt{n}(Y_n - c) \rightsquigarrow Y$ then $Y_n = O_P(1/\sqrt{n})$.
11. $O_p(1) \times O_p(1) = O_p(1)$.
12. $o_p(1) \times O_p(1) = o_p(1)$.

2 Distances Between Probability Distributions

Let P and Q be distributions with densities p and q. We will use the following distances between P and Q.

1. Total variation distance $\text{TV}(P,Q) = \sup_A |P(A) - Q(A)|$.
2. L_1 distance $d_1(P,Q) = \int |p - q|$.
3. Hellinger distance $h(P,Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$.
4. Kullback-Leibler distance $K(P,Q) = \int p \log(p/q)$.

5. L_2 distance $d_2(P,Q) = \int (p-q)^2$.

Here are some properties of these distances:

1. $\mathrm{TV}(P,Q) = \frac{1}{2} d_1(P,Q)$. (prove this!)
2. $h^2(P,Q) = 2(1 - \int \sqrt{pq})$.
3. $\mathrm{TV}(P,Q) \leq h(P,Q) \leq \sqrt{2\mathrm{TV}(P,Q)}$.
4. $h^2(P,Q) \leq K(P,Q)$.
5. $\mathrm{TV}(P,Q) \leq h(P,Q) \leq \sqrt{K(P,Q)}$.
6. $\mathrm{TV}(P,Q) \leq \sqrt{K(P,Q)/2}$.

3 Consistency

An estimator $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ is *consistent* for θ if

$$\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$$

as $n \to \infty$. In other words, $\widehat{\theta}_n - \theta = o_p(1)$. Here are two common ways to prove that $\widehat{\theta}_n$ consistent.

Method 1: Show that, for all $\varepsilon > 0$,

$$\mathbb{P}(|\widehat{\theta}_n - \theta| \geq \varepsilon) \longrightarrow 0.$$

Method 2. Prove convergence in quadratic mean:

$$\mathrm{MSE}(\widehat{\theta}_n) = \mathrm{Bias}^2(\widehat{\theta}_n) + \mathrm{Var}(\widehat{\theta}_n) \longrightarrow 0.$$

If bias $\to 0$ and var $\to 0$ then $\widehat{\theta}_n \xrightarrow{qm} \theta$ which implies that $\widehat{\theta}_n \xrightarrow{p} \theta$.

Example 1 $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(p)$. *The mle* $\widehat{p}$ *has bias 0 and variance* $p(1-p)/n \to 0$. *Here* $\widehat{p} = \sum_i X_i/n$. *So* $\widehat{p} \xrightarrow{\mathrm{P}} p$ *and is consistent. Now let* $\psi = \log(p/(1-p))$. *Then* $\widehat{\psi} = \log(\widehat{p}/(1-\widehat{p}))$. *Now* $\widehat{\psi} = g(\widehat{p})$ *where* $g(p) = \log(p/(1-p))$. *By the continuous mapping theorem,* $\widehat{\psi} \xrightarrow{\mathrm{P}} \psi$ *so this is consistent. Now consider*

$$\widehat{p} = \frac{\sum_i X_i + 1}{n+1}.$$

Then

$$\text{Bias} = \mathbb{E}(\widehat{p}) - p = -\frac{p-1}{n(1+n)} \to 0$$

and

$$\text{Var} = \frac{p(1-p)}{n} \to 0.$$

So this is consistent.

Example 2 $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. *Let* $\widehat{\theta}_n = X_{(n)}$. *By direct proof (we did it earlier) we have* $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.

Method of moments estimators are typically consistent. Consider one parameter. Recall that $\mu(\widehat{\theta}) = m$ where $m = n^{-1}\sum_{i=1}^n X_i$. Assume that μ^{-1} exists and is continuous. So $\widehat{\theta} = \mu^{-1}(m)$. By the WLLN $m \xrightarrow{\text{P}} \mu(\theta)$. So, by the continuous mapping Theorem,

$$\widehat{\theta}_n = \mu^{-1}(m) \xrightarrow{\text{P}} \mu^{-1}(\mu(\theta)) = \theta.$$

4 Consistency of the MLE

Under regularity conditions, the mle is consistent. Let us prove this in a special case. This will also reveal a connection between the mle and Hellinger distance. Suppose that the model consists of finitely many distinct densities $\mathcal{P} = \{p_0, p_1, \ldots, p_N\}$. The likelihood function is

$$L(p_j) = \prod_{i=1}^n p_j(X_i).$$

The mle $\widehat{p}$ is the density p_j that maximizes $L(p_j)$. Without loss of generality, assume that the true density is p_0.

Theorem 3

$$\mathbb{P}(\widehat{p} \neq p_0) \to 0$$

as $n \to \infty$.

Proof. Let us begin by first proving an inequality. Let $\epsilon_j = h(p_0, p_j)$. Then, for $j \neq 0$,

$$\begin{aligned}
\mathbb{P}\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right) &= \mathbb{P}\left(\prod_{i=1}^n \frac{p_j(X_i)}{p_0(X_i)} > e^{-n\epsilon_j^2/2}\right) = \mathbb{P}\left(\prod_{i=1}^n \sqrt{\frac{p_j(X_i)}{p_0(X_i)}} > e^{-n\epsilon_j^4/2}\right) \\
&\leq e^{n\epsilon_j^2/4}\mathbb{E}\left(\prod_{i=1}^n \sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\right) = e^{n\epsilon_j^2/4}\prod_{i=1}^n \mathbb{E}\left(\sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\right) \\
&= e^{n\epsilon_j^2/4}\left(\int \sqrt{p_j\ p_0}\right)^n = e^{n\epsilon_j^2/4}\left(1 - \frac{h^2(p_0,p_j)}{2}\right)^n = e^{n\epsilon_j^2/4}\left(1 - \frac{\epsilon_j^2}{2}\right)^n \\
&= e^{n\epsilon_j^2/4}\exp\left\{n\log\left(1 - \frac{\epsilon_j^2}{2}\right)\right\} \leq e^{n\epsilon_j^2/4}e^{-n\epsilon_j^2/2} = e^{-n\epsilon_j^2/2}.
\end{aligned}$$

We used the fact that $h^2(p_0, p_j) = 2 - 2\int \sqrt{p_0\, p_j}$ and also that $\log(1-x) \leq -x$ for $x > 0$. Let $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_N\}$. Then

$$\begin{aligned}
\mathbb{P}(\widehat{p} \neq p_0) &\leq \mathbb{P}\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2} \text{ for some } j\right) \\
&\leq \sum_{j=1}^{N} \mathbb{P}\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right) \\
&\leq \sum_{j=1}^{N} e^{-n\epsilon_j^2/2} \leq N e^{-n\epsilon^2/2} \to 0.
\end{aligned}$$

■

We can prove a similar result using Kullback-Leibler distance as follows. Let $X_1, X_2, \ldots$ be iid F_θ. Let θ_0 be the true value of θ and let θ be some other value. We will show that $L(\theta_0)/L(\theta) > 1$ with probability tending to 1. We assume that the model is **identifiable**; this means that $\theta_1 \neq \theta_2$ implies that $K(\theta_1, \theta_2) > 0$ where K is the Kullback-Leibler distance.

Theorem 4 *Suppose the model is identifiable. Let θ_0 be the true value of the parameter. For any $\theta \neq \theta_0$*

$$\mathbb{P}\left(\frac{L(\theta_0)}{L(\theta)} > 1\right) \to 1$$

as $n \to \infty$.

Proof. We have

$$\begin{aligned}
\frac{1}{n}(\ell(\theta_0) - \ell(\theta)) &= \frac{1}{n}\sum_{i=1}^{n} \log p(X_i; \theta_0) - \frac{1}{n}\sum_{i=1}^{n} \log p(X_i; \theta) \\
&\xrightarrow{p} E(\log p(X; \theta_0)) - E(\log p(X; \theta)) \\
&= \int (\log p(x; \theta_0)) p(x; \theta_0) dx - \int (\log p(x; \theta)) p(x; \theta_0) dx \\
&= \int \left(\log \frac{p(x; \theta_0)}{p(x; \theta)}\right) p(x; \theta_0) dx \\
&= K(\theta_0, \theta) > 0.
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{P}\left(\frac{L(\theta_0)}{L(\theta)} > 1\right) &= \mathbb{P}\left(\ell(\theta_0) - \ell(\theta) > 0\right) \\
&= \mathbb{P}\left(\frac{1}{n}(\ell(\theta_0) - \ell(\theta)) > 0\right) \to 1. \quad \square
\end{aligned}$$

■

This is not quite enough to show that $\widehat{\theta}_n \to \theta_0$.

Example 5 *Inconsistency of an mle. In all examples so far $n \to \infty$, but the number of parameters is fixed. What if the number of parameters also goes to ∞? Let*

$$\begin{aligned} Y_{11}, Y_{12} &\sim N(\mu_1, \sigma^2) \\ Y_{21}, Y_{22} &\sim N(\mu_2, \sigma^2) \\ \vdots &\sim \vdots \\ Y_{n1}, Y_{n2} &\sim N(\mu_n, \sigma^2). \end{aligned}$$

Some calculations show that

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} \sum_{j=1}^{2} \frac{(Y_{ij} - \overline{Y}_i)^2}{2n}.$$

It is easy to show (good test question) that

$$\widehat{\sigma}^2 \xrightarrow{p} \frac{\sigma^2}{2}.$$

Note that the modified estimator $2\widehat{\sigma}^2$ is consistent.

The reason why consistency fails is because the dimension of the parameter space is increasing with n.

Theorem 6 *Under regularity conditions on the model $\{p(x;\theta) : \theta \in \Theta\}$, the mle is consistent.*

The regularity conditions are technical. Basically, we need to assume that (i) the dimension of the parameter space does not change with n and (ii) $p(x;\theta)$ is a smooth function of θ.

5 Score and Fisher Information

The *score function* and *Fisher information* are the key quantities in many aspects of statistical inference. Suppose for now that $\theta \in \mathbb{R}$. The score function is

$$S_n(\theta) \equiv S_n(\theta, X_1, \ldots, X_n) = \ell'(\theta) = \frac{\partial \log p(X_1, \ldots, X_n; \theta)}{\partial \theta} \stackrel{\text{iid}}{=} \sum_i \frac{\partial \log p(X_i; \theta)}{\partial \theta}.$$

The Fisher information is defined to be

$$I_n(\theta) = \text{Var}_\theta(S_n(\theta))$$

that is, the variance of the score function. Later we will see that for the mle, $\text{Var}(\widehat{\theta}) \approx 1/I_n(\theta)$. That is why is $I_n(\theta)$ called "Information."

Theorem 7 *Under regularity conditions,*

$$\mathbb{E}_\theta[S_n(\theta)] = 0.$$

In other words,

$$\int \cdots \int \left(\frac{\partial \log p(x_1, \ldots, x_n; \theta)}{\partial \theta} \right) p(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n = 0.$$

That is, if the expected value is taken at the same θ as we evaluate $S_n(\theta)$, then the expectation is 0. This does not hold when the θ's mismatch: $\mathbb{E}_{\theta_0}[S_n(\theta_1)] \neq 0$. We'll see later that this property is very important.

Proof.

$$\begin{aligned}
\mathbb{E}_\theta[S_n(\theta)] &= \int \cdots \int \frac{\partial \ \log \ p(x_1, \ldots, x_n; \theta)}{\partial \theta} \ p(x_1, \ldots, x_n; \theta) \ dx_1 \cdots dx_n \\
&= \int \cdots \int \frac{\frac{\partial}{\partial \theta} \ p(x_1, \ldots, x_n; \theta)}{p(x_1, \ldots, x_n; \theta)} p(x_1, \ldots, x_n; \theta) \ dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial \theta} \underbrace{\int \cdots \int p(x_1, \ldots, x_n; \theta) \ dx_1 \cdots dx_n}_{1} \\
&= 0.
\end{aligned}$$

■

Example 8 *Let* $X_1, \ldots, X_n \sim N(\theta, 1)$. *Then*

$$S_n(\theta) = \sum_{i=1}^{n} (X_i - \theta).$$

Clearly, $\mathbb{E}_\theta[S_n(\theta)] = 0$.

Warning: If the support of p depends on θ, then $\int \cdots \int$ and $\frac{\partial}{\partial \theta}$ cannot be switched.

Now we discuss some properties of the Fisher information. Recall that the Fisher information is defined to be $I_n(\theta) = \mathrm{Var}(S_n(\theta))$. Since the mean of the score is 0, we have that

$$I_n(\theta) = \mathbb{E}_\theta[S_n^2(\theta)].$$

Lemma 9 *For the iid case, we have* $I_n(\theta) = nI(\theta)$ *where* $I(\theta)$ *is the Fisher information for* $n = 1$.

Proof. This follows since the log-likelihood — and hence the score — is the the sum of n, independent terms. ■

The next result gives a very simple formula for calculating the Fisher information.

Lemma 10 *Under regularity conditions,*

$$I_n(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \ell_n(\theta) \right].$$

Proof. For simplicity take $n = 1$. First note that

$$\int p = 1 \quad \Rightarrow \quad \int p' = 0 \quad \Rightarrow \quad \int p'' = 0 \quad \Rightarrow \quad \int \frac{p''}{p} p = 0 \quad \Rightarrow \quad \mathbb{E}\left(\frac{p''}{p}\right) = 0.$$

Let $\ell = \log p$ and $S = \ell' = p'/p$. Then $\ell'' = (p''/p) - (p'/p)^2$ and

$$\begin{aligned} \text{Var}(S) &= \mathbb{E}(S^2) - (\mathbb{E}(S))^2 = \mathbb{E}(S^2) = \mathbb{E}\left(\frac{p'}{p}\right)^2 = \mathbb{E}\left(\frac{p'}{p}\right)^2 - \mathbb{E}\left(\frac{p''}{p}\right) \\ &= -\mathbb{E}\left(\left(\frac{p''}{p}\right) - \left(\frac{p'}{p}\right)^2\right) = -\mathbb{E}(\ell''). \end{aligned}$$

■

The Vector Case. Let $\theta = (\theta_1, \cdots, \theta_k)$. $L_n(\theta)$ and $\ell_n(\theta)$ are defined as before. The score function $S_n(\theta)$ is now a vector of length k and the j^{th} component is $\partial \ell_n(\theta)/\partial \theta_j$. The Fisher information $I_n(\theta)$ is now a $k \times k$ matrix; it is the variance-covraince matrix of the score. We have the identity:

$$I_n(r, s) = -\mathbb{E}_\theta \left[\frac{\partial^2 \ell(\theta)}{\partial \theta_r \, \partial \theta_s} \right].$$

Example 11 *Suppose that* $X_1, \cdots, X_n \sim N(\mu, \gamma)$. *Then:*

$$\begin{aligned} L_n(\mu, \gamma) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\gamma}} \exp\left\{ \frac{-1}{2\gamma}(x_i - \mu)^2 \right\} \propto \gamma^{\frac{-n}{2}} \exp\left\{ \frac{-1}{2\gamma} \Sigma (x_i - \mu)^2 \right\} \\ \ell_n(\mu, \gamma) &= -\frac{n}{2} \log \gamma - \frac{1}{2\gamma} \Sigma (x_i - \mu)^2 \\ S_n(\mu, \gamma) &= \begin{bmatrix} \frac{1}{\gamma} \Sigma (x_i - \mu) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2} \Sigma (x_i - \mu)^2 \end{bmatrix} \end{aligned}$$

$$I_n(\mu, \gamma) = -\mathbb{E} \begin{bmatrix} \frac{-n}{\gamma} & \frac{-1}{\gamma^2} \Sigma (x_i - \mu) \\ \frac{-1}{\gamma^2} \Sigma (x_i - \mu) & \frac{n}{2\gamma^2} - \frac{1}{\gamma^3} \Sigma (x_i - \mu)^2 \end{bmatrix} = \begin{bmatrix} \frac{n}{\gamma} & 0 \\ 0 & \frac{n}{2\gamma^2} \end{bmatrix}$$

You can check that $\mathbb{E}_\theta(S) = (0, 0)^T$.

6 Asymptotic Normality of the MLE

In this section we prove that the mle satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, I^{-1}(\theta)).$$

In other words,

$$\widehat{\theta}_n \approx N\left(\theta, \frac{1}{nI(\theta)}\right).$$

In fact we will show that

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \psi^*(X_i) + o_P(n^{-1/2}) \tag{1}$$

where

$$\psi^*(x) = \frac{S(\theta, X_i)}{I(\theta)}$$

is called the *influence function*. In the next section we shall see that any well-behaved estimator $\widehat{\theta}$ can also be written as (1) for some ψ and that $\text{Var}(\psi) \geq \text{Var}(\psi^*)$

The regularity conditions we need to prove the asymptotic Normality of the mle are stronger than those needed for consistency. We need to assume that (i) the dimension of the parameter space does not change with n, (ii) $p(x;\theta)$ is a smooth function of θ, (iii) we can interchange differentition w.r.t. θ and integration over x, and (iv) the range of X does not depend on θ.

Theorem 12

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N\left(0, \frac{1}{I(\theta)}\right).$$

Hence, $\widehat{\theta}_n = \theta + O_P\left(\frac{1}{\sqrt{n}}\right)$.

Proof. By Taylor's theorem

$$0 = \ell'(\widehat{\theta}) = \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\theta) + \cdots.$$

Hence

$$\sqrt{n}(\widehat{\theta} - \theta) \approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \equiv \frac{A}{B}.$$

Now

$$A = \frac{1}{\sqrt{n}}\ell'(\theta) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^{n} S(\theta, X_i) = \sqrt{n}(\overline{S} - 0)$$

where $S(\theta, X_i)$ is the score function based on X_i. Recall that $E(S(\theta, X_i)) = 0$ and $\text{Var}(S(\theta, X_i)) = I(\theta)$. By the central limit theorem, $A \rightsquigarrow N(0, I(\theta)) = \sqrt{I(\theta)}Z$ where $Z \sim N(0,1)$. By the WLLN,

$$B \xrightarrow{\text{P}} -E(\ell'') = I(\theta).$$

By Slutsky's theorem

$$\frac{A}{B} \rightsquigarrow \frac{\sqrt{I(\theta)}Z}{I(\theta)} = \frac{Z}{\sqrt{I(\theta)}} = N\left(0, \frac{1}{I(\theta)}\right).$$

So

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N\left(0, \frac{1}{I(\theta)}\right). \quad \square$$

■

A small modification of the above proof yields:

Theorem 13 *We have*

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \psi^*(X_i) + o_P(n^{-1/2}) \tag{2}$$

where

$$\psi^*(x) = \frac{S(\theta, X_i)}{I(\theta)}$$

Now suppose we want to estimate $\tau(\theta)$. By the delta method we have:

Theorem 14 *Let τ be a smooth function of θ. Then*

$$\sqrt{n}(\tau(\widehat{\theta}_n) - \tau(\theta)) \rightsquigarrow N(0, (\tau'(\theta))^2/I(\theta)).$$

From all the above, we see that the approximate standard error of $\widehat{\theta}$ is

$$se = \sqrt{\frac{1}{nI(\theta)}} = \sqrt{\frac{1}{I_n(\theta)}}.$$

The estimated standard error is

$$\widehat{se} = \sqrt{\frac{1}{I_n(\widehat{\theta})}}.$$

The standard error of $\widehat{\tau} = \tau(\widehat{\theta})$ is

$$se = \sqrt{\frac{|\tau'(\theta)|}{nI(\theta)}} = \sqrt{\frac{|\tau'(\theta)|}{I_n(\theta)}}.$$

The estimated standard error is

$$\widehat{se} = \sqrt{\frac{|\tau'(\widehat{\theta})|}{I_n(\widehat{\theta})}}.$$

Example 15 $X_1, \cdots, X_n \sim \text{Exponential}(\theta)$. *Now* $p(x;\theta) = \theta e^{-\theta x}$ *and* $L(\theta) = e^{-n\theta\overline{X}+n\ln\theta}$. *Hence,* $\ell(\theta) = -n\theta\overline{X} + n \log \theta$ *and* $S(\theta) = \frac{n}{\theta} - n\overline{X}$. *The mle is* $\widehat{\theta} = \frac{1}{\overline{X}}$. *Now* $\ell''(\theta) = \frac{-n}{\theta^2}$ *so that* $I(\theta) = E[-\ell''(\theta)] = \frac{n}{\theta^2}$. *Thus,* $\widehat{\theta} \approx N\left(\theta, \frac{\theta^2}{n}\right)$.

Example 16 $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *The mle is* $\widehat{p} = \overline{X}$. *The Fisher information for* $n = 1$ *is*

$$I(p) = \frac{1}{p(1-p)}.$$

So

$$\sqrt{n}(\widehat{p} - p) \rightsquigarrow N(0, p(1-p)).$$

Informally,

$$\widehat{p} \approx N\left(p, \frac{p(1-p)}{n}\right).$$

The asymptotic variance is $p(1-p)/n$. *This can be estimated by* $\widehat{p}(1-\widehat{p})/n$. *That is, the estimated standard error of the mle is*

$$\widehat{se} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Now suppose we want to estimate $\tau = p/(1-p)$. *The mle is* $\widehat{\tau} = \widehat{p}/(1-\widehat{p})$. *Now*

$$\frac{\partial}{\partial p}\frac{p}{1-p} = \frac{1}{(1-p)^2}$$

The estimated standard error is

$$\widehat{se}(\widehat{\tau}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \times \frac{1}{(1-\widehat{p})^2} = \sqrt{\frac{\widehat{p}}{n(1-\widehat{p})^3}}.$$

7 Efficiency

Under the previously mentioned regularity conditions, the mle is optimal. The details are very complicated so we will only discuss this heuristically.[1] Recall that the mle satisfies In fact we will show that

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \psi^*(X_i) + o_P(n^{-1/2}). \tag{3}$$

It can be shown that any well-behaved estimator $\widetilde{\theta}$ satisfies

$$\widetilde{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i) + o_P(n^{-1/2}) \tag{4}$$

[1]For details see the book *Asymptotic Statistics* by Aad van der Vaart.

for some function ψ. Moreover, it can be shown that $\text{Var}(\psi) \geq \text{Var}(\psi^*)$. It follows that:

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N(0, 1/I(\theta)), \quad \sqrt{n}(\widetilde{\theta} - \theta) \rightsquigarrow N(0, v(\theta))$$

where $v(\theta) \geq 1/I(\theta)$. We say that the mle is *efficient*.

8 Relative Efficiency

Given two asymptotically Normal estimators, we can compare thier performance by looking at the ratio of the asymptotic variances. Specifically, if

$$\begin{aligned} \sqrt{n}(W_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_W^2) \\ \sqrt{n}(V_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_V^2) \end{aligned}$$

then the *asymptotic relative efficiency (ARE)* is

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

Example 17 *Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. The mle of λ is $\overline{X}$. Let*

$$\tau = \mathbb{P}(X_i = 0).$$

So $\tau = e^{-\lambda}$. Define $Y_i = I(X_i = 0)$. This suggests the estimator

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

Another estimator is the mle

$$V_n = e^{-\widehat{\lambda}}.$$

The delta method gives

$$\text{Var}(V_n) \approx \frac{\lambda e^{-2\lambda}}{n}.$$

We have

$$\begin{aligned} \sqrt{n}(W_n - \tau) &\rightsquigarrow N(0, e^{-\lambda}(1 - e^{-\lambda})) \\ \sqrt{n}(V_n - \tau) &\rightsquigarrow N(0, \lambda e^{-2\lambda}). \end{aligned}$$

So

$$\text{ARE}(W_n, V_n) = \frac{\lambda}{e^{\lambda} - 1} \leq 1. \quad \square$$

Since the mle is efficient, we know that, in general, $\text{ARE}(W_n, \text{mle}) \leq 1$.

9 Multivariate Case

Now let $\theta = (\theta_1, \ldots, \theta_k)$. In this case we have

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N(0, I^{-1}(\theta))$$

where $I^{-1}(\theta)$ is the inverse of the Fisher information matrix. The approximate standard error of $\widehat{\theta}_j$ is $\sqrt{I^{-1}(j,j)/n}$. If $\tau = g(\theta)$ with $g : \mathbb{R}^k \to \mathbb{R}$ then by the delta method,

$$\sqrt{n}(\widehat{\tau} - \tau) \rightsquigarrow N(0, (g')^T I^{-1} g')$$

where g' is the gradient of g evaluated at θ.

10 Exponential Families

A density of the form

$$p(x;\theta) = c(\theta)h(x)e^{\theta^T t(x)}$$

is called an exponential family. Examples include: the Normal, the binomial, the Poisson. The Uniform$(0,\theta)$ is not an exponential family.

For an exponential family, the mle is obtained by solving the equations

$$\mathbb{E}_\theta[t(X)] = \frac{1}{n}\sum_{i=1}^n t(X_i).$$

Also, the vector $\frac{1}{n}\sum_{i=1}^n t(X_i)$ is minimal sufficient. Furthermore, the Fisher information is given by $I(\theta) = a''(\theta)$ where $a(\theta) = -\log c(\theta)$.

Exercise: Prove the above facts.

11 Robustness

The mle is efficient only if the model is right. The mle can be bad if the model is wrong. That is why we should consider using nonparametric methods. One can also replace the mle with estimators that are more *robust*.

Suppose we assume that $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. The mle is $\widehat{\theta}_n = \overline{X}_n$. Suppose, however that we have a perturbed model X_i is $N(\theta, \sigma^2)$ with probability $1-\delta$ and X_i is Cauchy with probability δ. Then, $\text{Var}(\overline{X}_n) = \infty$.

Consider the median M_n. We will show that

$$ARE(\text{median}, \text{mle}) = .64.$$

But, under the perturbed model the median still performs well while the mle is terrible. In other words, we can trade efficiency for robustness. Let us now find the limiting distribution of M_n.

Let $Y_i = I(X_i \leq \mu + a/\sqrt{n})$. Then $Y_i \sim \text{Bernoulli}(p_n)$ where

$$p_n = F(\mu + a/\sqrt{n}) = F(\mu) + \frac{a}{\sqrt{n}}p(\mu) + o(n^{-1/2}) = \frac{1}{2} + \frac{a}{\sqrt{n}}p(\mu) + o(n^{-1/2}).$$

Also, $\sum_i Y_i$ has mean np_n and standard deviation

$$\sigma_n = \sqrt{np_n(1-p_n)}.$$

Note that,

$$M_n \leq \mu + \frac{a}{\sqrt{n}} \quad \text{if and only if} \quad \sum_i Y_i \geq \frac{n+1}{2}.$$

Then,

$$\begin{aligned}
\mathbb{P}(\sqrt{n}(M_n - \mu) \leq a) &= \mathbb{P}\left(M_n \leq \mu + \frac{a}{\sqrt{n}}\right) = \mathbb{P}\left(\sum_i Y_i \geq \frac{n+1}{2}\right) \\
&= \mathbb{P}\left(\frac{\sum_i Y_i - np_n}{\sigma_n} \geq \frac{\frac{n+1}{2} - np_n}{\sigma_n}\right).
\end{aligned}$$

Now,

$$\frac{\frac{n+1}{2} - np_n}{\sigma_n} \to -2ap(\mu)$$

and hence

$$\mathbb{P}(\sqrt{n}(M_n - \mu) \leq a) \to \mathbb{P}(Z \geq -2ap(\mu)) = \mathbb{P}\left(-\frac{Z}{2p(\mu)} \leq a\right) = \mathbb{P}\left(\frac{Z}{2p(\mu)} \leq a\right)$$

so that

$$\sqrt{n}(M_n - \mu) \rightsquigarrow N\left(0, \frac{1}{(2p(\mu))^2}\right).$$

For a standard Normal, $(2p(0))^2 = .64$.