Asymptotic of Likelihood-Based Inference

In the following arguments, we adopt the notations $O_p(1), o_p(1)$. Specifically, we say a sequence of random variables $\{Y_n\}$ is $O_p(1)$, or stochastically bounded, if and only if, $\forall \epsilon > 0$, there exists $M_\epsilon > 0$ such that $P(|Y_n| \leq M_\epsilon) \geq 1 - \epsilon$. Note that convergence in distribution implies stochastically bounded. We say $\{Y_n\}_{n \geq 1}$ is $o_p(1)$ if and only if $Y_n \xrightarrow{p} 0$. One can see A+L or Shao for references.

65. (a) Let $\mu$ be the sum of two measures, one is the Lebesgue measure on $(1, \infty)$, the other is the point mass measure at 0. Then the R-N derivative of the probability measure of $X$ with respect to $\mu$ is $\lambda e^{-\lambda x}I\{x \neq 0\} + (1 - e^{-\lambda})I\{x = 0\}$ almost surely. Adopt the notation in the class, the likelihood-function is $f(X^n) = \lambda \sum_i \delta_i e^{-\lambda \sum_i \delta_i X_i} (1 - e^{-\lambda})^n - \sum_i \delta_i$, where $\delta_i = I\{X_i \neq 0\}$. Then under the event $M = n - \sum_i \delta_i = n$, the likelihood function becomes $(1 - e^{-\lambda})^n$, which has no maximizer for $\lambda$ on $(0, \infty)$. In all the other cases, at least one of $\delta_i$ is 1, notice that $\log f(X^n) = \sum_i \delta_i \log(\lambda) - \lambda \sum_i \delta_i X_i + (n - \sum_i \delta_i) \log(1 - e^{-\lambda})$. Further, $\frac{\partial}{\partial \lambda} f(X^n) = \sum_i \delta_i - \sum_i \delta_i X_i + (n - \sum_i \delta_i)/(1 - e^{-\lambda})$. As we could see, when $\lambda \to 0$, the score function converges to $+\infty$, and when $\lambda \to \infty$, the score function converges to $-\sum_i \delta_i X_i$. By the intermediate value theorem, the likelihood equation has at least one root on $(0, \infty)$ in this case. Therefore, $P_\lambda$(MLE exists) $\geq P_\lambda(M < n) = 1 - P_\lambda(X_i = 0, \forall i) = 1 - (1 - e^{-\lambda})^n \to 1$ as $n \to \infty$, for any $\lambda > 0$.

(b) Let $T_n = -\log(1 - M/n)$, then by SLLN, $M/n \to 1 - e^{-\lambda}$, w.p.1., which implies $T_n$ converges to $\lambda$ almost surely $P_\lambda$. A “1-step Newton improvement” would then be $\hat{\lambda}_n = T_n - \frac{\sum_i \delta_i}{T_n - \sum_i \delta_i X_i + (n - \sum_i \delta_i)/nM}$. 

(c) In the event that the MLE exists, we can use the Newton-Raphson method with $T_n$ as the starting value. Basically, we are repeating the procedure as done in $(b)$ and update each step with a “1-step newton improvement” adjusted value.

(d) For a $(1 - \alpha)100\%$ asymptotic confidence interval, we can use $\hat{\lambda}_n \pm z_{1 - \alpha/2} \{n I_1(\hat{\lambda}_n)\}^{-1/2}$ and $\tilde{\lambda}_n \pm z_{1 - \alpha/2} \{-L''(\tilde{\lambda}_n)\}^{-1/2}$, where $L_n$ is the log-likelihood function. More specifically, $L_n(\lambda) = -\sum_i \delta_i / \lambda^2 - (n - \sum_i \delta_i) e^{-\lambda} / (1 - e^{-\lambda})^2, n I_1(\lambda) = E_\lambda(-L''(\lambda)) = ne^{-\lambda}/\lambda^2 + ne^{-\lambda}/(1 - e^{-\lambda})$.

66. The log-likelihood function for the uncensored data (suppose the dominating measure is the Lebesgue measure on $\mathbb{R}^+$) is

$$\ell_n(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n X_i$$ (1)
Figure 1: Log-likelihood function for the uncensored and censored data.

For the censored data, the R-N derivative of $X \{ X > 1 \}$, with respect to the $\sigma$-finite measure that is the sum of Lebesgue measure on $(0, \infty)$ and point mass measure at 0, is (see problem 35)

$$\ell_n^*(\lambda) = \sum_i \delta_i \log \lambda - \lambda \sum_i \delta_i X_i + (n - \sum_i \delta_i) \log(1 - e^{-\lambda}),$$

(2)

where $\delta_i = I\{X_i > 1\}$. For the data given in this problem, their likelihood function is shown in the Figure 1.

(a) Given the uncensored data, the 90% two-sided confidence set for $\lambda$ is $\{\lambda \mid \chi^2_{1,0.05} \leq 2(\ell_n(\hat{\lambda}) - \ell_n(\lambda)) \leq \chi^2_{1,0.95} \}$. In this case, $\hat{\lambda} = 1/X_n = 0.9341, \ell_n(\hat{\lambda}) = -21.3625, \chi^2_{1,0.05} = 0.00393, \chi^2_{1,0.95} = 3.84146$. Then one can solve some nonlinear equations and get the confidence interval of $\lambda$ (OK, a huge literature that is not the focus of this course!).

(b) Their relationship can be seen in Figure 1.

(c) The Fisher-information in $\lambda$ about $X \{ X > 1 \}$ evaluated at $\lambda = \lambda_0$ is

$$I_1^*(\lambda_0) = E_{\lambda_0} \left\{ \delta/\lambda_0^2 + (1 - \delta)e^{\lambda_0}/(e^{\lambda_0} - 1)^2 \right\} = e^{-\lambda_0}/\lambda_0^2 + e^{-\lambda_0}/(1 - e^{-\lambda_0}).$$

(3)
Two different approximate (Wald) 90% confidence sets would be \( \tilde{\lambda} \pm z_{0.05} \sqrt{n I_1(\lambda)} \) and
\( \tilde{\lambda} \pm z_{0.05} \sqrt{-\ell''(\tilde{\lambda})} \), where \( \tilde{\lambda} \) is the maximum likelihood estimator for censored data.

67. So \( \theta = (p_1, p_2, p_3, p_4)' \) and the parameter space \( \Theta \) is \([0, 1]^4\). The hypothesis testing problem here is \( H_0 : \theta \in \Theta_0 \) vs \( H_1 : \theta \notin \Theta_0 \), where \( \Theta_0 = \{(p_1, p_2, p_3, p_4) \in \Theta \mid p_1 = p_2, p_3 = p_4\} \).

(a) The likelihood function is
\[
f_{\theta}^{n}(X^{n}) = \prod_{i=1}^{4} \left\{ \binom{n}{X_i} p_i^{X_i} (1 - p_i)^{n-X_i} \right\},
\]
(4)

Then the LRT corresponding to this hypothesis testing problem is
\[
LRT = \sup_{\theta \in \Theta} f_{\theta}^{n}(X^{n}) / \sup_{\theta \in \Theta_0} f_{\theta}^{n}(X^{n}) = \prod_{i=1}^{4} \left( \frac{2X_i}{X_i[i/3]+1 + X_i[i/3]+2} \right)^{X_i} \left( \frac{2n - 2X_i}{2n - X_i[i/3]+1 - X_i[i/3]+2} \right)^{n-X_i},
\]
(5)

where \( [i/3] \) stands for the integer part of \( i/3 \), e.g. \( [1/3] = [2/3] = 0, [3/3] = [4/3] = 1 \).

The log-likelihood ratio would then be the sum of the following two terms
\[
\sum_{i=1}^{2} X_i \log\left\{ \frac{2X_i}{X_i + X_2} \right\} + (n - X_i) \log\left\{ \frac{(2n - 2X_i)}{(2n - X_1 - X_2)} \right\},
\]
(6)

\[
\sum_{j=3}^{4} X_j \log\left\{ \frac{2X_j}{X_3 + X_4} \right\} + (n - X_j) \log\left\{ \frac{(2n - 2X_j)}{(2n - X_3 - X_4)} \right\}.
\]
(7)

They correspond to the logs of independent LRT statistics for \( H_0 : p = p_2 \) and \( H_0 : p_3 = p_4 \) respectively.

(b) \( h(\theta) = (p_1 - p_2, p_3 - p_4) \) and \( H(\theta) = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \) and FI is \( I(\theta) = \text{diag}(n/\{p_1(1-p_1)\}, n/\{p_2(1-p_2)\}, n/\{p_3(1-p_3)\}, n/\{p_4(1-p_4)\}) \) Then the Wald statistic is
\[
W_n = \begin{pmatrix} \frac{X_1 - X_2}{n} & \frac{X_3 - X_4}{n} \\ \frac{X_1(n-X_1)+X_2(n-X_2)}{n} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \frac{n^2}{X_3(n-X_3)+X_4(n-X_4)} \end{pmatrix} \begin{pmatrix} \frac{X_1 - X_2}{n} \\ \frac{X_3 - X_4}{n} \end{pmatrix}.
\]
(8)

A simplification gives us
\[
W_n = \frac{\{\sqrt{n}(X_1 - X_2)/n\}^2}{X_1(n-X_1)/n + X_2(n-X_2)/n} + \frac{\{\sqrt{n}(X_3 - X_4)/n\}^2}{X_3(n-X_3)/n + X_4(n-X_4)/n}.
\]
(9)

Then we can see that, under \( H_0 \), \( W_n \xrightarrow{d} \chi_2^2 \).

(c) Under \( H_0 \), \( \tilde{\theta} = (\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4)' \), where \( \tilde{p}_1 = \tilde{p}_2 = (X_1 + X_2)/(2n), \tilde{p}_3 = \tilde{p}_4 = (X_3 + X_4)/(2n) \).

Then the score test statistic is
\[
V_n = \frac{(X_1 - n\tilde{p}_1)^2}{n\tilde{p}_1(1 - \tilde{p}_1)} + \frac{(X_2 - n\tilde{p}_2)^2}{n\tilde{p}_2(1 - \tilde{p}_2)} + \frac{(X_3 - n\tilde{p}_3)^2}{n\tilde{p}_3(1 - \tilde{p}_3)} + \frac{(X_4 - n\tilde{p}_4)^2}{n\tilde{p}_4(1 - \tilde{p}_4)}.
\]
(10)
After some algebra, it becomes
\[
V_n = \frac{(X_1 - X_2)^2}{2np_1(1 - p_1)} + \frac{(X_3 - X_4)^2}{2np_3(1 - p_3)},
\] (11)
which converges to $\chi^2_2$ under $H_0$.

68. (a) $1/P_\theta(\{0\}) = 1 + e^\theta + e^{2\theta}$, thus a moment estimator of $\theta$ based on $n_0$ is
\[
\hat{\theta} = \log \left\{ -1 + \sqrt{4n/n_0 - 3} \right\}. \tag{12}
\]
Denote $g(t) = \log\{-1 + \sqrt{4/t - 3}/2\}$. As $\sqrt{n}(n_0/n - P_\theta(\{0\}))$ is asymptotically normal, and $g'(P_\theta(\{0\})) \neq 0$. By delta method, $\sqrt{n}(\hat{\theta} - \theta) = O_p(1)$, specifically, it is asymptotically normally distributed.

(b) Note that $\sum_i I\{X_i = 1\}\theta + 2\sum_i I\{X_i = 2\} = \sum_i X_i$. The log-likelihood is
\[
L_n(\theta) = \theta \sum_i X_i - n \log(1 + e^\theta + e^{2\theta}) \tag{13}
\]
\[
L'_n(\theta) = \sum_i X_i - \frac{ne^\theta + 2ne^{2\theta}}{1 + e^\theta + e^{2\theta}} \tag{14}
\]
\[
L''_n(\theta) = \frac{(ne^\theta + 4ne^{2\theta})(1 + e^\theta + e^{2\theta}) - n(e^\theta + 2e^{2\theta})^2}{(1 + e^\theta + e^{2\theta})^2} \tag{15}
\]
If we write
\[
L'_n(\hat{\theta}_n) = L'_n(\theta) + (\hat{\theta}_n - \theta)L''_n(\theta) + (\hat{\theta}_n - \theta)^2 L'''_n(\theta)/2, \tag{16}
\]
where $\theta^*_n$ is some estimator between $\hat{\theta}_n$ and $\theta_0$. Then we have
\[
\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\sqrt{n}L'_n(\theta)}{-L''_n(\hat{\theta}_n)} + \sqrt{n}(\hat{\theta}_n - \theta) \left\{ 1 - \frac{L''_n(\theta)}{L''_n(\hat{\theta}_n)} \right\} - \frac{\sqrt{n}(\hat{\theta}_n - \theta)^2 L'''_n(\theta^*_n)}{2L''_n(\hat{\theta}_n)}. \tag{17}
\]
It can be seen that the last two terms of the above expression is $o_p(1)$ under the conditions given. Moreover, it can be verified that $\sqrt{n}\{L'_n(\theta)/n\} \xrightarrow{d} N(0, I_1(\theta))$ and $-L''_n(\hat{\theta}_n)/n \xrightarrow{p} I_1(\theta)$, thus the result follows.

(c) The likelihood equation is
\[
(2 - \bar{x})e^{2\theta} + (1 - \bar{x})e^\theta - \bar{x} = 0. \tag{18}
\]
When $\bar{x} \in (0, 2)$, the solution is
\[
\hat{\theta} = \log \left\{ \bar{x} - 1 + \sqrt{(1 - \bar{x})^2 + 4\bar{x}(2 - \bar{x})} \right\}. \tag{19}
\]
To show that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, 1/I_1(\theta))$ in $\theta$ probability. It suffices to verify the conditions of Theorem 173 (which would be easier than proving directly). Conditions 1-5 can be easily verified to be true. Then $P_\theta(\hat{\theta}_n$ is a root of likelihood equation) $\geq P_\theta(0 < \bar{X} < 2) \to 1$, as $n \to \infty$. Moreover, as $\bar{x}$ converges to $(e^\theta + 2e^{2\theta})/(1 + e^\theta + e^{2\theta})$ almost surely. Plug the limit back into the above equation, we get $\hat{\theta}$ converges to $\theta$ almost surely.
(d) The FI is \( I_1(\theta) = \text{Var}_\theta(X) = e^\theta (1 + 4e^\theta + e^{2\theta})/(1 + e^\theta + e^{2\theta})^2 \). Then any consistent estimator of \( \theta \) say \( \hat{\theta}_n \) such that \( \hat{\theta}_n \xrightarrow{p} \theta \), \( I_1(\hat{\theta}_n) \xrightarrow{p} I_1(\theta) \). And the observed fisher information is \( -L''(\theta)/n = e^\theta (1 + 4e^\theta + e^{2\theta})/(1 + e^\theta + e^{2\theta})^2 \), which is exactly \( I_1(\theta) \). Thus their approximations are the same. If we use \( \widetilde{\theta}_n \), then

\[
I_1(\widetilde{\theta}_n) = \bar{x} \left\{ 1 - \bar{x} + 2 \frac{(\bar{x} - 1) + \sqrt{-3\bar{x}^2 + 6\bar{x} + 1}}{\bar{x} + 2\sqrt{-3\bar{x}^2 + 6\bar{x} + 1}} \right\}.
\]

(20)

69. Without loss of generality, assume that \( h(x) = 1 \) (otherwise consider \( \bar{\mu}(A) = \int_A h d\mu \)). Then for this one-parameter density, the R-N derivative has the form

\[
f_\eta(x) = K(\eta) \exp\{\eta T(x)\}, \text{ a.e.} (\mu),
\]

where \( K(\eta) \) is the normalizing constant. The parameter space is \( \Gamma = \\{ \eta \mid \int \exp\{\eta T(x)\} d\mu(x) < \infty \} \).

(a) For any \( \eta_0 \) in the interior of \( \Gamma \), consider the open neighborhood of \( \eta_0 \), \( \mathcal{O} = (\eta_0 - \epsilon, \eta_0 + \epsilon) \). First of all, \( f_\eta(x) > 0 \forall x \in \mathcal{X} \) and \( \eta \in \mathcal{O} \). Second, by DCT, \( f_\eta \) has first partial derivative at \( \eta_0 \), especially, when taking derivative on both sides of \( 1 = \int f_\theta(x) d\mu(x) \), we get \( 0 = \int \partial f_\eta(x)/\partial \eta|_{\eta=\eta_0} d\mu(x) \). Thus \( \mathcal{P} \) is FI regular at every \( \eta_0 \) in the interior of \( \Gamma \). (The detail for applying DCT is like this. Let \( g(\eta) = \int \exp\{\eta T(x)\} d\mu(x) \), then for \( \eta \in \mathcal{O} \), \( g(\eta) < \infty \). Also \( \{g(\eta+h) - g(\eta)\}/h = \int \exp\{\eta T(x)\} \cdot \{(\exp(\eta h)-1)/h\} d\mu(x) \). Then notice that \( (\exp(\eta h)-1)/h \) is actually bounded for small enough \( h \) and all \( \eta \in \mathcal{O} \).)

(b) Because of (a), by definition of Fisher Information, the fisher information in \( \mathcal{X} \) about \( \eta \) evaluated at \( \eta_0 \) is

\[
I(\eta_0) = \int \left[ \frac{\partial \{ \log \{ K(\eta) \} + \eta T(x) \} }{\partial \eta} \right]_{\eta=\eta_0}^2 K(\eta_0) \exp\{\eta_0 T(x)\} d\mu(x).
\]

(22)

As \( 0 = \int \partial f_\eta(x)/\partial \eta|_{\eta=\eta_0} d\mu(x) \), some algebra work will give us \( K'(\eta_0)/K(\eta_0) = -E_{\eta_0}\{T(X)\} \). Thus the above expression simplifies to \( I(\eta_0) = \text{Var}_{\eta_0}\{T(X)\} \).

(c)

\[
I(\eta_0, \eta) = \int f_{\eta_0}(x) \log \frac{f_{\eta}(x)}{f_{\eta_0}(x)} d\mu(x) = \int \left\{ \log \frac{K(\eta_0)}{K(\eta)} + (\eta_0 - \eta) T(x) \right\} f_{\eta_0}(x) d\mu(x),
\]

(23)

which is \( \log\{K(\eta_0)/K(\eta)\} + (\eta_0 - \eta) E_{\eta_0}\{T(X)\} \).

(d) The likelihood equation is

\[
nK'(\eta)/K(\eta) + \sum_i T(X_i) = 0
\]

(24)

(e) \( E_{\eta_0}|\log f_\eta(X)\| \leq |\log K(\eta)| + |\eta_0|E_{\eta_0}|T(X)| \). Therefore, all conditions of Theorem 170 are met.

(f) The general form of “one-step Newton improvement” would be

\[
\bar{\eta} - \frac{K'(\bar{\eta})}{K(\bar{\eta})} + \frac{1}{n} \sum_i T(X_i) \left\{ \frac{K'(\bar{\eta})}{K(\bar{\eta})} \right\}^2 - \frac{K''(\bar{\eta})}{K(\bar{\eta})}
\]

(25)
To rigorously argue this, define the following set

\[ A_n = \{ \hat{\theta}_n \text{ is a root of likelihood equation} \} \cap \{ L_n(\theta) \text{ is unimodal} \}, \quad (26) \]

then \( \lim_{n \to \infty} P_{\theta_0}(A_n) = 1 \). Then for any random variable \( T_n \), we can claim that \( |T_n I_{A_n} - T_n| = o_p(1) \), in \( \theta_0 \) probability. This is because, for any \( \epsilon > 0 \),

\[ P_{\theta_0}\{ |T_n I_{A_n} - T_n| > \epsilon \} \leq P_{\theta_0}\{ A_n^c \} \to 0, \text{ as } n \to \infty. \quad (27) \]

Therefore, we only need to consider \( 2I_{A_n}\{ L_n(\hat{\theta}_n) - \max_{\theta < \theta_0} L_n(\theta) \} \) here, which equals to \( 2\{\theta_0 < \hat{\theta}_n\} \{ L_n(\hat{\theta}_n) - L_n(\theta_0) \} \). Also, Theorem 173 says \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, I^{-1}_1(\theta_0)) \) in \( \theta_0 \) probability. Further, under conditions of Theorem 173, Taylor expansion says \( L_n(\theta_0) - L_n(\hat{\theta}_n) = L'_n(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + \frac{1}{2} L''_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^2 \), where \( \hat{\theta}_n \) is between \( \hat{\theta}_n \) and \( \theta_0 \). Therefore, in \( \theta_0 \) probability, \( 2\{\theta_0 < \hat{\theta}_n\} \{ L_n(\hat{\theta}_n) - L_n(\theta_0) \} = -\frac{L''_n(\theta_0)}{n L'_n(\theta_0)} I\{ \sqrt{n I_1(\theta_0)}(\hat{\theta}_n - \theta_0) > 0 \} \{ \sqrt{n I_1(\theta_0)}(\hat{\theta}_n - \theta_0) \}^2 \overset{d}{\to} I\{ Z > 0 \} Z^2, \) where \( Z \sim N(0, 1) \).

By slutsky’s theorem, it suffices to prove that, in \( \theta_0 \), \(-n^{-1} L''_n(\hat{\theta}_n) \overset{d}{\to} I_1(\theta_0) \). Note that \( n^{-1} |L''_n(\hat{\theta}_n) - L''(\theta_0)| \leq |\hat{\theta}_n - \theta_0| n^{-1} \sum_i M(X_i) = o_p(1) \). Therefore, \(-n^{-1} L''_n(\hat{\theta}_n) = -n^{-1} \{ L''_n(\hat{\theta}_n) - L''(\theta_0) \} - n^{-1} L''(\theta_0) = o_p(1) - n^{-1} L''(\theta_0) \overset{d}{\to} I_1(\theta_0) \).

The hypothesis 1-5 of Theorem 173 guarantee that \( L'_n(T_n) = L'_n(\theta_0) + L''(\theta_0^*)(T_n - \theta_0) \), where \( \theta_0^* \) is between \( T_n \) and \( \theta_0 \). Plug this into the one-step Newton improvement formula, we have

\[ \sqrt{n}(\tilde{\theta}_n - \theta_0) = \sqrt{n}(T_n - \theta_0) \left\{ 1 - \frac{L''(\theta_0^*)}{L''(T_n)} \right\} = \sqrt{n} \frac{L''(\theta_0)}{L''(T_n)}. \quad (28) \]

Notice that \( T_n = \theta_0 + O_p(n^{-1/2}) \) (a more general condition than \( \sqrt{n}(T_n - \theta_0) \) converges in distribution). Then \( \theta_0^* = \theta_0 + O_p(n^{-1/2}) \). One can use the same argument as in problem 72, and see that \( L''(T_n) = L''(\theta_0) + o_p(1), L''(\theta_0^*) = L''(\theta_0) + o_p(1) \), which means \( \sqrt{n}(\tilde{\theta}_n - \theta_0) = \sqrt{n}(T_n - \theta_0) \{ 1 - L''(\theta_0^*)/L''(T_n) \} = o_p(1) \). The result then follows, by slutsky’s theorem, from the facts that \(-n^{-1} L''(T_n) \overset{d}{\to} I_1(\theta_0), \sqrt{n}\{ n^{-1} L''(\theta_0) - 0 \} \overset{d}{\to} N(0, I_1(\theta_0)) \), in \( \theta_0 \) probability.