STA2212: Inference and Likelihood

A. Notation

One random variable: Given a model for X which assumes X has a density $f(x;\theta), \quad \theta \in \Theta \subset \mathbb{R}^k$, we have the following definitions:

likelihood function	$L(\theta;x) = c(x)f(x;\theta)$	$\mathcal{L}(\theta)$
log-likelihood function	$\ell(\theta;x) = \log L(\theta;x) = \log f(x;\theta) + a(x)$	
score function	$u(\theta) = \partial\ell(\theta;x)/\partial\theta$	$\ell'(x;\theta)$
observed information function	$j(\theta) = -\partial^2\ell(\theta;x)/\partial\theta\partial\theta^T$	$J(\theta) = \mathrm{E}_\theta\{j(\theta)\}$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\}$[1]	$I(\theta)$ (p.245)

Independent observations: When we have X_i independent, identically distributed from $f(x_i;\theta)$, then, denoting the observed sample $\boldsymbol{x} = (x_1,\ldots,x_n)$ we have:

likelihood function	$L(\theta;\boldsymbol{x}) = \prod_{i=1}^n f(x_i;\theta)$	$\mathcal{L}(\theta)$
log-likelihood function	$\ell(\theta) = \ell(\theta;\boldsymbol{x}) = \sum_{i=1}^n \ell(\theta;x_i)$	$\ell(\theta)$
maximum likelihood estimate	$\hat{\theta} = \hat{\theta}(\boldsymbol{x}) = \arg\sup_\theta \ell(\theta)$	$S(\boldsymbol{X})$
score function	$U(\theta) = \ell'(\theta) = \sum U_i(\theta)$	$\boldsymbol{S}(\theta)$ (p.273)
observed information function	$j(\theta) = -\ell''(\theta) = -\ell''(\theta;\boldsymbol{x})$	$nJ(\theta) = \mathrm{E}_\theta\{-\ell''(x;\theta)\}$
observed (Fisher) information	$j(\hat{\theta})$	$n\widehat{I(\theta)}$ (p.254)
expected (Fisher) information	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\} = ni_1(\theta)$	$I_n(\theta) = nI(\theta)$

Comments:

1. the maximum likelihood estimate $\hat{\theta}_n$ is usually obtained by solving the *score equation* $\ell'(\theta;\boldsymbol{x}) = 0$. Lazy notation is $\hat{\theta}$, but for asymptotics $\hat{\theta}_n$ is preferred.

2. It doesn't really matter for the definitions above if the observations are independent and identically distributed (i.i.d.), or only independent, but the theorems that are proved in MS Ch. 5 and AoS Ch. 9 assume i.i.d..

3. There are important distinctions to be careful about in the notation for likelihood and its quantities:

   (a) Are we working with a single observation x, X or n observations $\boldsymbol{x}, \boldsymbol{X}$?

   (b) Do we want to find the distribution of something; so $\ell(\theta;X)$ or calculate data summaries; $\ell(\theta;x)$?

[1] $U(\theta) = u(\theta;X)$

B. First order asymptotic theory MS §5.4

1. θ is a scalar

If the components of $\boldsymbol{X}$ are i.i.d., then the score function $U(\theta; \boldsymbol{X})$ is a sum of i.i.d. random variables, and we can show that it has expected value 0 and variance $I_n(\theta)$ (or $i(\theta)$ in my notation). Under some regularity conditions on the density $f(x_i; \theta)$ (MS A1-A6, p.245), the central limit theorem gives

$$\frac{U(\theta)}{I_n^{1/2}(\theta)} \xrightarrow{d} N(0,1), \text{ equivalently } \frac{1}{\sqrt{n}} U(\theta) \xrightarrow{d} N\{0, I(\theta)\}. \tag{1}$$

Almost everything else follows from this result and Slutsky's theorem. For example, we can show that

$$(\hat{\theta} - \theta) I_n^{1/2}(\theta) = U(\theta)/I_n^{1/2}(\theta) + o_p(1),$$

where $o_p(1)$ means a remainder term that goes to 0 in probability as $n \to \infty$, so we have the second result

$$(\hat{\theta} - \theta) I_n^{1/2}(\theta) \xrightarrow{d} N(0,1). \tag{2}$$

These limit theorems give us two corresponding approximations to use with n fixed:

$$U(\theta) \dot{\sim} N\left(0, I_n(\theta)\right), \tag{3}$$

and

$$\hat{\theta} - \theta \dot{\sim} N\left(0, 1/I_n(\theta)\right). \tag{4}$$

The notation $\dot{\sim}$ is read as "is approximately distributed as".

The proof of MS Theorem 5.3 allows that $I(\theta) = \text{var}\{\ell'(\theta; X_i)\}$ and $J(\theta) = \text{E}\{-\ell''(\theta); X_i\}$ might be different, which is handy later for the study of misspecified models.

Having the unknown quantity θ in the variance in (3) and (4) is inconvenient, but to the same order of approximation, we can replace $I_n(\theta)$ by $I_n(\hat{\theta})$ or by $j(\hat{\theta})$; the latter is denoted $\widehat{I_n(\theta)}$ in MS, p. 254. In AoS, $I_n^{-1/2}(\theta)$ is called $\mathtt{se}$ and $I_n^{-1/2}(\hat{\theta})$ is called $\widehat{\mathtt{se}}$, but the use of $j(\hat{\theta}) = -\ell''(\hat{\theta}; \boldsymbol{x})$ is not mentioned.