

Given a model for  $Y$  which assumes  $Y$  has a density  $f(y; \theta)$ ,  $\theta \in \Theta \subset \mathbb{R}^d$ , we have the following definitions:

observed likelihood function	$L(\theta; y) = c(y)f(y; \theta)$
log-likelihood function	$\ell(\theta; y) = \log L(\theta; y) = \log f(y; \theta) + a(y)$
score function	$U(\theta) = \partial \ell(\theta; y) / \partial \theta$
observed information function	$j(\theta) = -\partial^2 \ell(\theta; y) / \partial \theta \partial \theta^T$
expected information (in one observation)	$i(\theta) = E_{\theta} U(\theta) U(\theta)^T$ (called $i_1(\theta)$ in CH)

When we have  $Y_i$  independent, identically distributed from  $f(y_i; \theta)$ , then, denoting the observed sample  $y = (y_1, \dots, y_n)$  we have:

log-likelihood function	$\ell(\theta) = \ell(\theta; y) + a(y)$	$O_p(n)$
maximum likelihood estimate	$\hat{\theta} = \hat{\theta}(y) = \arg \sup_{\theta} \ell(\theta)$	$\theta + O_p(n^{-1/2})$
score function	$U(\theta) = \ell'(\theta) = \sum U_i(\theta) = U_+(\theta)$	$O_p(n^{1/2})$
observed information function	$j(\theta) = -\ell''(\theta) = -\ell(\theta; Y)$	$O_p(n)$
observed (Fisher) information	$j(\hat{\theta})$	
expected (Fisher) information	$i(\theta) = E_{\theta} \{U(\theta) U(\theta)^T\} = n i_1(\theta)$	$O(n)$ ,

where with the risk of some confusion we use the same notation. Sometimes the expected Fisher information is defined instead as  $i(\theta) = E_{\theta} \{-\partial U(\theta; Y) / \partial \theta^T\}$  (e.g. in BNC). In models for which we can interchange differentiation and integration in  $\int f(y; \theta) dy = 1$ , these are the same due to the Bartlett identities:

$$\begin{aligned} E_{\theta} \{U(\theta)\} &= 0, \\ E_{\theta} \{U'(\theta)\} + E_{\theta} \{U^2(\theta)\} &= 0, \\ E_{\theta} \{U''(\theta)\} + 3E_{\theta} \{U(\theta)U'(\theta)\} + E_{\theta} \{U^3(\theta)\} &= 0, \end{aligned}$$

and so on, where the result applies to vector  $\theta$ , but as presented here is for scalar  $\theta$ . (In the vector setting the second derivative of  $U$  is a  $d \times d \times d$  array.)

## First order asymptotic theory

The following results are used for approximate inference based on the likelihood function:

1.  $\theta$  is a scalar

$\frac{1}{\sqrt{n}}U(\theta)/i_1^{1/2}(\theta) \xrightarrow{d} N(0, 1)$	by the central limit theorem
standardized score statistic	$r_u = U(\theta)/j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, 1)$
$\sqrt{n}(\hat{\theta} - \theta)i_1^{1/2}(\theta) =$	$\frac{1}{\sqrt{n}} \frac{U(\theta)}{i_1^{1/2}(\theta)} \{1 + o_p(1)\}$
standardized m.l.e.	$r_e = (\hat{\theta} - \theta)j^{1/2}(\hat{\theta}) \xrightarrow{d} N(0, 1)$
(log) likelihood ratio statistic	$w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} = (\hat{\theta} - \theta)^2 i(\theta) \{1 + o_p(1)\}$
	$w(\theta) \xrightarrow{d} \chi_1^2$
likelihood root	$r(\theta) = \text{sign}(\theta - \hat{\theta})\{w(\theta)\}^{1/2}$
	$r(\theta) \xrightarrow{d} N(0, 1)$

2.  $\theta$  a vector of length  $d$

$\frac{1}{\sqrt{n}}\{U(\theta)\} \xrightarrow{d} N_d\{0, i_1(\theta)\}$	by the central limit theorem
standardized score statistic	$w_u = U(\theta)^T \{i(\theta)\}^{-1} U(\theta)$
$\sqrt{n}(\hat{\theta} - \theta) =$	$\frac{1}{\sqrt{n}} i_1^{-1}(\theta) U(\theta) \{1 + o_p(1)\}$
standardized m.l.e.	$w_e = (\hat{\theta} - \theta)^T i(\theta) (\hat{\theta} - \theta)$
likelihood ratio statistic	$w = 2\{\ell(\hat{\theta}) - \ell(\theta)\} = (\hat{\theta} - \theta)^T i(\theta) (\hat{\theta} - \theta) \{1 + o_p(1)\}$
	$w(\theta) \xrightarrow{d} \chi_d^2$

3.  $\theta = (\psi, \lambda) = (\psi_1, \dots, \psi_q, \lambda_1, \dots, \lambda_{d-q})$  We partition the information matrices compatibly and write

$$U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix},$$

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}$$

and

$$i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$$

The constrained maximum likelihood estimator of  $\lambda$  is denoted by  $\hat{\lambda}_\psi$ , which in regular models satisfies  $U_\lambda(\psi, \hat{\lambda}_\psi) = 0$ .

Note that

$$i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta)i_{\lambda\lambda}^{-1}(\theta)i_{\lambda\psi}(\theta)\}^{-1}, \quad (1)$$

using the formula for the determinant of a partitioned matrix. A similar result holds for  $j$ .

The profile log-likelihood function is  $\ell_{\mathbf{P}}(\psi) = \ell(\psi, \hat{\lambda}_{\psi})$ , and the (observed) profile information function is  $j_{\mathbf{P}}(\psi) = -\ell''_{\mathbf{P}}(\psi)$ , a  $q \times q$  matrix.

The limiting results above can be used to derive the following

$$\begin{aligned} w_u(\psi) &= U_{\psi}(\psi, \hat{\lambda}_{\psi})^T \{i^{\psi\psi}(\psi, \hat{\lambda}_{\psi})\} U_{\psi}(\psi, \hat{\lambda}_{\psi}) \sim \chi_q^2 \\ w_e(\psi) &= (\hat{\psi} - \psi) \{i^{\psi\psi}(\hat{\psi}, \hat{\lambda})\}^{-1} (\hat{\psi} - \psi) \sim \chi_q^2 \\ w(\psi) &= 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_{\psi})\} = 2\{\ell_{\mathbf{P}}(\hat{\psi}) - \ell_{\mathbf{P}}(\psi)\} \sim \chi_q^2; \end{aligned}$$

see (52), (54) and (56) in CH §9.3.

This determines the following first-order pivotal quantities, for scalar  $\psi$ :

$$\begin{aligned} r_e(\psi) &= (\hat{\psi} - \psi) j_{\mathbf{P}}^{1/2}(\hat{\psi}) \sim N(0, 1), \\ r_u(\psi) &= \ell'_{\mathbf{P}}(\psi) j_{\mathbf{P}}^{-1/2}(\hat{\psi}) \sim N(0, 1), \\ r(\psi) &= \text{sign}(\hat{\psi} - \psi) \sqrt{2\{\ell_{\mathbf{P}}(\hat{\psi}) - \ell_{\mathbf{P}}(\psi)\}} \sim N(0, 1) \\ w(\psi) &= 2\{\ell_{\mathbf{P}}(\hat{\psi}) - \ell_{\mathbf{P}}(\psi)\} \sim \chi_1^2, \end{aligned}$$

where the third form follows from the fourth.

## Measure theory

The likelihood function is defined as (proportional to) the density function, and this is a density with respect to some dominating measure. Since  $\theta$  varies in  $\Theta$ , we need  $f$  to be a density function with respect to the same dominating measure for each value of  $\theta$ . Schervish (p.13) states it this way:

Let  $(S, \mathcal{A}, \mu)$  be a probability space, and let  $(\mathcal{X}, \mathcal{B})$  be a Borel space. Let  $X : S \rightarrow \mathcal{X}$  be measurable. The parametric family of distributions for  $X$  is the set

$$\{P_{\theta} : \forall A \in \mathcal{B}, P_{\theta}(A) = \Pr(X \in A), \theta \in \Theta\}.$$

Assume that each  $P_{\theta}$ , considered as a measure on  $(\mathcal{X}, \mathcal{B})$  is absolutely continuous with respect to a measure  $\nu$  on  $(\mathcal{X}, \mathcal{B})$ . We write

$$f(x; \theta) = \frac{dP_{\theta}}{d\nu}(x);$$

this is the likelihood function for  $\theta$ .

Some books describe the likelihood function as the Radon-Nikodym derivative of the probability measure with respect to a dominating measure. Sometimes the dominating measure is taken to be  $P_{\theta_0}$  for a fixed value  $\theta_0 \in \Theta$ . When we consider probability spaces and/or parameter spaces that are infinite dimensional, it is not obvious what to use as a dominating measure. For counting processes, this is done rigorously in Ch.II of Andersen et al. The result is Jacod's formula for the likelihood ratio:

Suppose we have a counting process  $N(\cdot)$  on  $[0, \tau]$ , and a filtration  $\mathcal{F}_t = \mathcal{F}_0 \cup \sigma\{N(s); s \leq t\}$ , with  $\mathcal{F} = \mathcal{F}_\tau$ . A counting process is a piecewise constant, non-decreasing, stochastic process with jumps of size +1. It can be shown to be a local submartingale, with compensator  $\Lambda$ . Suppose  $P$  and  $\tilde{P}$  are two probability measures on  $\mathcal{F}$ , for which the two compensators are  $\Lambda$  and  $\tilde{\Lambda}$ . Suppose  $\tilde{P}$  is absolutely continuous with respect to  $P$ . If  $\Lambda$  and  $\tilde{\Lambda}$  are absolutely continuous a.s.  $P$ , then

$$\frac{d\tilde{P}}{dP} = \frac{d\tilde{P}}{dP} \Big|_{\mathcal{F}_0} \frac{\prod_t \tilde{\lambda}(t)^{\Delta N(t)} \exp\{-\tilde{\Lambda}(\tau)\}}{\prod_t \lambda(t)^{\Delta N(t)} \exp\{-\Lambda(\tau)\}}.$$

Except for the somewhat unfamiliar notation, this is identical to the likelihood function for the non-homogeneous Poisson process (SM, Ex.6.5),

$$\prod_{j=1}^n \lambda(t_j) \exp\left\{-\int_0^\tau \lambda(u) du\right\}, \quad 0 < t_1 < \dots < t_n < \tau.$$

## References

- Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- [BNC] Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.
- [SM] Davison, A.C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.
- Schervish, M.J. (1995). *Theory of Statistics*. Springer, New York.