Overview of important results

BS2 Statistical Inference, Lecture 16
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; December 3, 2004

Unbiased estimators

An estimator $\hat{g}(\theta) = t(x)$ is *unbiased for a function* $g(\theta)$ if

$$\mathbf{E}_\theta\{t(X)\} = g(\theta).$$

The *bias* of an estimator of $g(\theta)$ is

$$\text{bias}\{\hat{g}(\theta)\} = \mathbf{E}\{t(X) - g(\theta)\}.$$

Even if $\hat{\theta}$ is an unbiased estimator of θ, $g(\hat{\theta})$ will generally *not* be an unbiased estimator of $g(\theta)$.

If $\hat{g}(\theta)$ has mimimum variance among all unbiased estimators of $g(\theta)$ it is a *minimum variance unbiased estimator* (MVUE). *An MVUE is unique*.

Mean Square Error

The *mean square error* (MSE) of an estimator of θ is:

$$\mathrm{mse}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)^2.$$

The MSE can be decomposed as

$$\mathrm{mse}(\hat{\theta}) = \mathbf{V}(\hat{\theta} - \theta) + \{\mathbf{E}(\hat{\theta} - \theta)\}^2 = \mathbf{V}(\hat{\theta}) + \{\mathrm{bias}(\theta)\}^2.$$

For *unbiased* estimators, the MSE is equal to the variance, $\mathrm{mse}(\hat{\theta}) = \mathbf{V}(\hat{\theta})$.

Score and Fisher information

The *score statistic* is define as

$$s(X;\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta) = \frac{f'(X;\theta)}{f(X;\theta)}.$$

Under regularity conditions we have

$$\mathbf{E}\{S(\theta)\} = 0$$

and the variance of $S(\theta)$ is the *Fisher information*:

$$i(\theta) = \mathbf{V}\{S(\theta)\} = \mathbf{E}\{S(\theta)^2\},$$

and further that

$$i(\theta) = -\mathbf{E}\left\{\frac{\partial}{\partial\theta}S(\theta)\right\} = -\mathbf{E}\left\{\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right\},$$

Cramér–Rao's inequality

Under the usual regularity conditions *it holds for any unbiased estimator* $T = t(X)$ *of* $g(\theta)$ *that*

$$\mathbf{V}(T) = \mathbf{V}(\hat{g}(\theta)) \geq \{g'(\theta)\}^2 / i(\theta).$$

An unbiased estimator is said to be *efficient* if it attains the lower bound.

this happens only if the score statistic has the form

$$s(x;\theta) = \frac{i(\theta)\{t(x) - g(\theta)\}}{g'(\theta)},$$

essentially implying that $g(\theta)$ is the mean value parameter in a linear exponential family.

Sufficiency

A statistic $T = t(X)$ is *sufficient* for θ if $P_\theta\{X = x \,|\, T = t\}$ does not depend on θ.

A statistic is *minimal sufficient* if it is sufficient and it can be calculated from any other sufficient statistic.

A statistic $T = t(X)$ is sufficient for θ if and only if the family of densities can be factorized as

$$f(x;\theta) = h(x)k\{t(x);\theta\}, \quad x \in \mathcal{X}, \theta \in \Theta, \qquad (1)$$

i.e. into a function which does not depend on θ and one which only depends on x through $t(x)$.

The Rao–Blackwell theorem

If $U = u(X)$ is an unbiased estimator of a function $g(\theta)$ and $T = t(X)$ is sufficient for θ then $U^* = u^*(X)$ where $u^*(x) = \mathbf{E}_\theta\{U \,|\, T = t(x)\}$ is also unbiased for $g(\theta)$ and

$$\mathbf{V}(U^*) \leq \mathbf{V}(U),$$

The process of modifying U to the improved estimator U^* by taking conditional expectation w.r.t. a sufficient statistic T, is known as *Rao–Blackwellization*.

It follows that *an MVUE must be a function of any minimal sufficient statistic*.

Likelihood

The function $L_x(\theta) = f(x \,|\, \theta) = f(x; \theta)$ is known as the *likelihood function*, obtained from $f(x; \theta)$ by considering x fixed and θ varying.

Its logarithm is traditionally denoted $l_x(\theta) = \log L_x(\theta)$.

The likelihood function is only well-defined up to a multiplicative constant so *proportional likelihood functions are equivalent*.

The likelihood function is always minimal sufficient.

Completeness and sufficiency

A statistic T is *complete* w.r.t. θ if for all functions h

$$\mathbf{E}_\theta\{h(T)\} = 0 \text{ for all } \theta \implies h(t) = 0 \text{ a.s.}$$

Any estimator of the form $U = h(T)$ of a complete and sufficient statistic T is the unique unbiased estimator based on T of its expectation.

In fact, *if T is complete and sufficient, it is also minimal sufficient.*

Hence, *if T is complete and sufficient, $U = h(T)$ is the MVUE of its expectation.*

Method of moments

Consider a sample $X = (X_1, \ldots, X_n)$ from $f(x;\theta)$ where $\theta \in \Theta$ is unknown.

Estimate the first p moments of the f with corresponding empirical moments:

$$\mu_k(\theta) = \int x^k f(x;\theta)\,dx, \quad m_k = \frac{1}{n}\sum x_i^k, \quad k = 1, \ldots, p.$$

Equivalently, use *central* moments where $\mu = \mu_1(\theta)$:

$$\bar{\mu}_k(\theta) = \int (x-\mu)^k f(x;\theta)\,dx, \quad \bar{m}_k = \frac{1}{n}\sum (x_i - \bar{x})^k$$

for $k = 1, \ldots, p$.

Use as many as necessary to ensure unique solution in θ of the equations

$$\bar{\mu}_k(\theta) = \bar{m}_k, \quad k = 1, 2, \ldots, p$$

Moment estimators are *unbiased for moments* (but not for central moments).

The method of moments is generally not very good.

Method of maximum likelihood

The MLE $\hat{\theta}$ is given by

$$\hat{\theta} = \hat{\theta}(x) = \arg\max_{\theta} L_x(\theta) = \arg\max_{\theta} f(x;\theta).$$

If g is a one-to-one function, and $\hat{\theta}$ is the MLE of θ then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

If $\hat{\theta}$ is the unique MLE and $T = t(X)$ is sufficient, then $\hat{\theta}$ is a function of t.

If the MLE is itself sufficient, it is minimal sufficient.

Rao–Blackwellization is never needed for the MLE.

Ancillarity and Basu's Theorem

A statistic $A = a(X)$ is *ancillary* if the distribution of A does not depend on θ.

Intuitively, A then carries no information about θ.

Basu's Theorem says:

If $T = t(X)$ is complete and sufficient and A is ancillary, then T and A are independent.

Exponential Families

A family $\mathcal{F} = \{f(\cdot;\theta), \theta \in \Theta\}$ is said to be (curved) *exponential* if the densities have the form

$$f(x;\theta) = b(x)e^{a(\theta)^\top t(x) - c(\theta)},$$

where $b(x)$ is known, $t(x)^\top = (t_1(x), \ldots, t_k(x))$ is a vector of known real-valued functions, $a(\theta) = (a_1(\theta), \ldots, a_k(\theta))$ are twice continuously differentiable functions of $\theta \in \Theta$, and Θ is an open and connected subset of $\mathcal{R}^d$ with $d \leq k$. Also, the Jacobian $J(\theta)$ of $a(\theta)$ is assumed to have full rank d.

The representation is *minimal* if $(1, t_1, \ldots, t_k)$ are linearly independent. Then the *dimension* of the family is equal to d. The family is called *linear* if $d = k$ and *canonical* if $a(\theta) = \theta$.

Linear and canonical exponential families

If the family is linear, $T = t(X)$ is complete and sufficient.

For a canonical family it holds that

$$\mathbf{E}_\theta\{t_r(X)\} = \frac{\partial}{\partial\theta_r}c(\theta), \quad \mathrm{Cov}_\theta\{t_r(X), t_s(X)\} = \frac{\partial^2}{\partial\theta_r\theta_s}c(\theta),$$

so the *score and Fisher information is*

$$S_r(\theta) = t_r(X) - \mathbf{E}_\theta\{t_r(X)\}\frac{\partial}{\partial\theta_r}c(\theta)$$

and

$$i_{rs}(\theta) = \mathrm{Cov}\{t_r(X), t_s(X)\} = \frac{\partial^2}{\partial\theta_r\theta_s}c(\theta).$$

Maximum likelihood in linear exponential families

In a linear exponential families the log-likelihood function has at most one local maximum within Θ. This is then equal to the global maximum and determined by the unique solution to the equation

$$\mathbf{E}_\theta\{t(X)\} = t(x).$$

In this sense the method of MLE for linear exponential families is similar to the method of moments, just that general functions $t_1(X), t_2(X), \ldots, t_d(X)$ are used rather than the powers $X, X^2, \ldots, X^d$.

It is less trivial to identify when a solution to the likelihoood equation exists.

The mean value parameter

The *mean value mapping* τ is defined as

$$\tau(\theta) = \mathbf{E}_\theta\{t(X)\}.$$

The likelihood equation can then be compactly written as

$$\tau(\theta) = t(x).$$

The parameter $\eta = \tau(\theta)$ is the *mean value parameter* whereas the parameter θ in a canonical exponential family the *canonical* parameter.

the MLE of the mean value parameter is unbiased and efficient in the sense that it attains the Cramér–Rao bound.

Slutsky's theorem and the delta method

Let $Y_1, Y_2 \ldots$ and $Z_1, Z_2, \ldots$ be sequences of random variables so that $Y_n \xrightarrow{D} Y$ and $Z_n \xrightarrow{P} c$.

If $g(y, z)$ is continuous at all points (y, c), it holds that

$$g(Y_n, Z_n) \xrightarrow{D} g(Y, c).$$

If $Y_n = (X_n - b)/a_n \xrightarrow{D} Y$ for a scaling sequence $a_n > 0$ with $a_n \to 0$ and g is continuously differentiable at b

$$\{g(X_n) - g(b)\}/a_n \xrightarrow{D} g'(b)Y.$$

In particular *if $X_n \overset{\text{a}}{\sim} \mathcal{N}(b, a_n^2)$ for $a_n \to 0$ then $g(X_n) \overset{\text{a}}{\sim} \mathcal{N}\{g(b), a_n^2 g'(b)^2\}$.*

Cramér's conditions

The conditions below imply everything with score and Fisher information to be well defined, and are also the conditions needed fot the MLE to have good and simple asymptotic properties:

1. Θ is an open subset of the real line;
2. $A = \{x \mid f(x;\theta) > 0\}$ does not depend on θ:
3. The log-likelihood function is three times continuously differentiable so that for some $\delta > 0$ and all t with $|\theta - t| < \delta$, we have $l^{(i)}(x;t) < M_i(x)$, where $\mathbf{E}_\theta\{M_i(X)\} < \infty$;
4. $i(\theta) = -\mathbf{E}\{l''(\theta)\}$ is positive.

Asymptotic properties of the MLE

Cramér's conditions imply that the MLE is consistent, more precisely *that there is at least one consistent root $\hat{\theta}$ to the likelihood equation.*

Additional conditions ensure that the root is indeed the MLE so that MLE itself is consistent.

Under Cramér's conditions, the consistent root is also *asymptotically normal and efficient* in the sense that

$$\hat{\theta}_n \overset{\text{a}}{\sim} \mathcal{N}\{\theta, i_n(\theta)^{-1}\}$$

where $i_n(\theta) = ni(\theta)$ is the information in the full sample.

Variants of the asymptotics

The quantity

$$j_n(\hat{\theta}) = -\sum l''(X_i; \hat{\theta})$$

is the *observed Fisher information.*

It holds that any of

$$\sqrt{ni(\theta)}(\hat{\theta} - \theta), \quad \sqrt{ni(\hat{\theta})}(\hat{\theta} - \theta), \quad \sqrt{j_n(\hat{\theta})}(\theta_n - \theta)$$

converge in distribution to $\mathcal{N}(0, 1)$

Newton–Raphson and the method of scoring

The *Newton–Raphson* iteration calculates the MLE by repeating

$$\theta \leftarrow \theta + j_n(\theta)^{-1} S(\theta).$$

R. A. Fisher introduced the *method of scoring* which replaces the observed second derivative with its expectation to yield the iteration

$$\theta \leftarrow \theta + i_n(\theta)^{-1} S(\theta)$$

In many cases, $i(\theta)$ is easier to calculate.

In *canonical* exponential families we get

$$j(\theta) = \frac{\partial^2}{\partial \theta^2}\{c(\theta) - \theta t(X)\} = c''(\theta) = i(\theta)$$

so *for canonical exponential families the method of scoring and the method of Newton–Raphson coincide.*

The identity of Newton–Raphson and the method of scoring *only holds for the canonical parameter.*

The score test

The *locally most powerful test* (LMP) for a simple hypothesis $H_0 : \theta = \theta_0$ has critical region

$$S(x; \theta_0) > K.$$

The constant K can be determined by Monte–Carlo methods, or by large sample theory.

The *score test* for the hypothesis has critical region

$$S(x; \theta_0) > \sqrt{ni(\theta_0)} z_{1-\alpha}.$$

The *two-sided score test* has critical region

$$\{S(x; \theta_0)\}^2 > ni(\theta_0) \chi^2(1)_{1-\alpha}.$$

χ^2 and Wald's test

The test with critical region

$$X^2 = ni(\theta_0)(\hat{\theta} - \theta_0)^2 > \chi^2(1)_{1-\alpha}$$

is the *χ^2-test*.

The test with critical region

$$W = ni(\hat{\theta})(\hat{\theta} - \theta_0)^2 > \chi^2(1)_{1-\alpha},$$

is the *Wald test*.

The maximized likelihood ratio test

The MLRT (or LRT for short), has critical region of the form

$$\Lambda = \lambda(X) = -2 \log \frac{L(\theta_0; X)}{L(\hat{\theta}; X)} > K.$$

It immediately extends to the multiparameter caseand to the case where the null hypothesis $H_0 : \theta \in \Theta_0$ is *composite*, i.e. where Θ_0 has more than one value. Then

$$\lambda(x) = -2 \log \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)} = -2 \log \frac{L(\hat{\hat{\theta}}; x)}{L(\hat{\theta}; x)},$$

where $\hat{\theta} = \arg\max_{\theta \in \Theta}$ and $\hat{\hat{\theta}} = \arg\max_{\theta \in \Theta_0}$.

Asymptotic properties

Suppose Θ is an open and connected subset of $\mathcal{R}^d$ and Θ_0, specified as

$$\theta \in \Theta_0 \iff h(\theta) = 0$$

where $h(\theta) = (h_1(\theta), \ldots, h_k(\theta))$ is twice continuously differentiable with Jacobian having constant and full rank k for $\theta \in \Theta_0$.

If further Cramér's conditions are fulfilled, it holds that

$$\Lambda \xrightarrow{D} Y$$

where Y follows a χ^2-distribution with k degrees of freedom.

Approximate likelihood ratio tests

The proof of the asymptotic result for the LRT relies on the approximation

$$\Lambda \approx n(\hat{\theta} - \hat{\hat{\theta}})^\top C(\hat{\theta} - \hat{\hat{\theta}})$$

where C is a consistent estimate of $i(\theta_0)$. There are essentially four possibilities for the choice of C:

$$C_1 = i(\hat{\theta}), \quad C_2 = i(\hat{\hat{\theta}}), \quad C_3 = j_n(\hat{\theta})/n, \quad C_4 = j_n(\hat{\hat{\theta}})/n,$$

all leading to test statistics which are asymptotically $\chi^2(k)$.

This leads to the *Wald statistics*

$$W = n(\hat{\theta} - \hat{\hat{\theta}})^\top i(\hat{\theta})(\hat{\theta} - \hat{\hat{\theta}}), \quad \tilde{W} = (\hat{\theta} - \hat{\hat{\theta}})^\top j_n(\hat{\theta})(\hat{\theta} - \hat{\hat{\theta}})$$

or the χ^2 *statistics*

$$X^2 = n(\hat{\theta} - \hat{\hat{\theta}})^\top i(\hat{\hat{\theta}})(\hat{\theta} - \hat{\hat{\theta}}), \quad \tilde{X}^2 = (\hat{\theta} - \hat{\hat{\theta}})^\top j_n(\hat{\hat{\theta}})(\hat{\theta} - \hat{\hat{\theta}}).$$

The sequential probability ratio test

The SPRT for a simple hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ has the following form:

- If $\Lambda_n \geq B$, decide that H_1 is true and stop;
- If $\Lambda_n \leq A$, decide that H_0 is true and stop;
- If $A < \Lambda_n < B$, collect another observation to obtain Λ_{n+1},

where Λ_n is the log-likelihood ratio

$$\Lambda_n = \lambda(X_1, \ldots, X_n) = \log \frac{L(\theta_1; X_1, \ldots, X_n)}{L(\theta_0; X_1, \ldots, X_n)}.$$

Limits and error probabilities

The approximate relation between the decision limits A and B and the error probabilities

$$\alpha = P(D_1 \,|\, H_0), \quad \beta = P(D_0 \,|\, H_1),$$

where $P(D_i \,|\, H_j)$ denotes the probability of deciding that H_i is true when in fact H_j is, is given as

$$B \approx \log \frac{1-\beta}{\alpha}, \quad A \approx \log \frac{\beta}{1-\alpha}.$$

The only error in this approximation is that we have ignored the 'overshoot.