Statistical Asymptotics

G. Alastair Young

Department of Mathematics
Imperial College London

APTS, April 2008

Motivation

Motivation

- To improve on the first-order limit results, to obtain approximations whose asymptotic accuracy is higher by one or two orders.

Motivation

- To improve on the first-order limit results, to obtain approximations whose asymptotic accuracy is higher by one or two orders.
- The Fisherian proposition that inferences on the parameter of interest should be obtained by conditioning on an ancillary statistic, rather than from the original model.

Asymptotic expansion

An asymptotic expansion for a function $g_n(x)$ at some fixed x is expressed as

$$g_n(x) = \gamma_0(x)b_{0,n} + \gamma_1(x)b_{1,n} + \ldots + \gamma_k(x)b_{k,n} + o(b_{k,n}),$$

as $n \longrightarrow \infty$, where $\{b_{r,n}\}_{r=0}^{k}$ is a sequence such as $\{1, n^{-1/2}, n^{-1}, \ldots n^{-k/2}\}$ or $\{1, n^{-1}, n^{-2}, \ldots, n^{-k}\}$.

Asymptotic expansion

An asymptotic expansion for a function $g_n(x)$ at some fixed x is expressed as

$$g_n(x) = \gamma_0(x)b_{0,n} + \gamma_1(x)b_{1,n} + \ldots + \gamma_k(x)b_{k,n} + o(b_{k,n}),$$

as $n \to \infty$, where $\{b_{r,n}\}_{r=0}^{k}$ is a sequence such as $\{1, n^{-1/2}, n^{-1}, \ldots n^{-k/2}\}$ or $\{1, n^{-1}, n^{-2}, \ldots, n^{-k}\}$.

In general the sequence must have the property that $b_{r+1,n} = o(b_{r,n})$ as $n \to \infty$, for each $r = 0, 1, \ldots, k-1$.

Often the function of interest $g_n(x)$ will be the exact density or distribution function of a statistic based on a sample of size n, and $\gamma_0(x)$ will be some simple first-order approximation, such as the normal density or distribution function.

Often the function of interest $g_n(x)$ will be the exact density or distribution function of a statistic based on a sample of size n, and $\gamma_0(x)$ will be some simple first-order approximation, such as the normal density or distribution function.

One important feature of asymptotic expansions is that they are not in general convergent series for $g_n(x)$ for any fixed x: taking successively more terms, letting $k \to \infty$ for fixed n, will not necessarily improve the approximation to $g_n(x)$.

Stochastic asymptotic expansion

For a sequence of random variables $\{Y_n\}$, a stochastic asymptotic expansion is expressed as

$$Y_n = X_0 b_{0,n} + X_1 b_{1,n} + \ldots + X_k b_{k,n} + o_p(b_{k,n}),$$

where $\{b_{k,n}\}$ is a given set of sequences and $\{X_0, X_1, \ldots\}$ have distributions not depending on n.

Stochastic asymptotic expansions are not as well defined as asymptotic expansions, as there is usually considerable arbitrariness in the choice of the coefficient random variables $\{X_0, X_1, \ldots\}$.

A simple application of stochastic asymptotic expansion is the proof of asymptotic normality of the maximum likelihood estimator.

Tools of asymptotic analysis

Tools of asymptotic analysis

- Edgeworth expansions.

Tools of asymptotic analysis

- Edgeworth expansions.
- Saddlepoint approximations.

Tools of asymptotic analysis

- Edgeworth expansions.
- Saddlepoint approximations.
- Laplace's method.

Edgeworth expansion

Let $Y_1, Y_2, \ldots, Y_n$ be IID univariate with cumulant generating function $K_Y(t)$ and cumulants κ_r.

Let $S_n = \sum_1^n Y_i$, $S_n^* = (S_n - n\mu)/\sqrt{n}\sigma$ where $\mu \equiv \kappa_1 = EY_1, \sigma^2 \equiv \kappa_2 = \operatorname{var} Y_1$.

Define the rth standardised cumulant by $\rho_r = \kappa_r/\kappa_2^{r/2}$.

The Edgeworth expansions for the density of S_n^* is:

$$\begin{aligned} f_{S_n^*}(x) = \quad & \phi(x)\left\{1 + \frac{\rho_3}{6\sqrt{n}} H_3(x)\right. \\ + \quad & \left.\frac{1}{n}\left[\frac{\rho_4 H_4(x)}{24} + \frac{\rho_3^2 H_6(x)}{72}\right]\right\} + O(n^{-3/2}). \end{aligned}$$

Here $\phi(x)$ is the standard normal density and $H_r(x)$ is the rth degree Hermite polynomial defined by

$$\begin{aligned} H_r(x) &= (-1)^r \frac{d^r \phi(x)}{dx^r} \Big/ \phi(x) \\ &= (-1)^r \phi^{(r)}(x)/\phi(x), \quad \text{say.} \end{aligned}$$

Here $\phi(x)$ is the standard normal density and $H_r(x)$ is the rth degree Hermite polynomial defined by

$$\begin{aligned} H_r(x) &= (-1)^r \frac{d^r \phi(x)}{dx^r} \Big/ \phi(x) \\ &= (-1)^r \phi^{(r)}(x)/\phi(x), \quad \text{say.} \end{aligned}$$

We have $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$ and $H_6(x) = x^6 - 15x^4 + 45x^2 - 15$.

Comments

The leading term in the expansion is the standard normal density, as is appropriate from CLT.

Comments

The leading term in the expansion is the standard normal density, as is appropriate from CLT.

The $n^{-1/2}$ term is an adjustment for skewness, via the standardised skewness ρ_3.

Comments

The leading term in the expansion is the standard normal density, as is appropriate from CLT.

The $n^{-1/2}$ term is an adjustment for skewness, via the standardised skewness ρ_3.

The n^{-1} term is a simultaneous adjustment for skewness and kurtosis.

If the density of Y_1 is symmetric, $\rho_3 = 0$ and the normal approximation is accurate to order n^{-1}, rather than the usual $n^{-1/2}$ for $\rho_3 \neq 0$.

If the density of Y_1 is symmetric, $\rho_3 = 0$ and the normal approximation is accurate to order n^{-1}, rather than the usual $n^{-1/2}$ for $\rho_3 \neq 0$.

The accuracy of the Edgeworth approximation, which truncates the expansion, will depend on the value of x. Edgeworth approximations tend to be poor, and may even be negative, in the tails of the distribution, as $|x|$ increases.

Distribution function

Integrating the Edgeworth expansion using the properties of the Hermite polynomials, gives an expansion for the distribution function of S_n^*:

$$F_{S_n^*}(x) = \Phi(x) - \phi(x)\left\{\frac{\rho_3}{6\sqrt{n}}H_2(x) + \frac{\rho_4}{24n}H_3(x) + \frac{\rho_3^2}{72n}H_5(x)\right\} + O(n^{-3/2}).$$

Also, if T_n is a sufficiently smooth function of S_n^*, then a formal Edgeworth expansion can be obtained for the density of T_n.

Cornish-Fisher expansion

Might wish to determine x, as x_α say, so that $F_{S_n^*}(x_\alpha) = \alpha$, to the order considered in the Edgeworth approximation to the distribution function of S_n^*.

The solution is known as the Cornish-Fisher expansion and the formula is

$$\begin{aligned} x_\alpha &= z_\alpha + \frac{1}{6\sqrt{n}}(z_\alpha^2 - 1)\rho_3 + \frac{1}{24n}(z_\alpha^3 - 3z_\alpha)\rho_4 \\ &- \frac{1}{36n}(2z_\alpha^3 - 5z_\alpha)\rho_3^2 + O(n^{-3/2}), \end{aligned}$$

where $\Phi(z_\alpha) = \alpha$.

Derivation*

The density of a random variable can be obtained by inversion of its characteristic function.

Derivation*

The density of a random variable can be obtained by inversion of its characteristic function.

In particular, the density for $\bar{X}$, the mean of a set of IID random variables $X_1, \ldots, X_n$, can be obtained as

$$f_{\bar{X}}(\bar{x}) = \frac{n}{2\pi i} \int_{\tau - i\infty}^{\tau + i\infty} \exp\big[n\{K(\phi) - \phi\bar{x}\}\big] d\phi,$$

where K is the cumulant generating function of X, and τ is any point in the open interval around 0 in which the moment generating function M exists.

Edgeworth expansions are obtained by expanding the cumulant generating function in a Taylor series around 0, exponentiating and inverting term by term.

Saddlepoint expansion

The saddlepoint expansion for the density of S_n is

$$\begin{aligned} f_{S_n}(x) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\{nK''_Y(\hat{\phi})\}^{1/2}} \\ &\times \exp\{nK_Y(\hat{\phi}) - \hat{\phi}x\}\{1 + O(n^{-1})\} \end{aligned}$$

where $\hat{\phi} \equiv \hat{\phi}(x)$ satisfies $nK'_Y(\hat{\phi}) = x$.

The $O(n^{-1})$ term is actually $(3\hat{\rho}_4 - 5\hat{\rho}_3^2)/(24n)$, where $\hat{\rho}_j \equiv \hat{\rho}_j(\hat{\phi}) = K_Y^{(j)}(\hat{\phi})/\{K_Y''(\hat{\phi})\}^{j/2}$ is the jth standardised derivative of the cumulant generating function for Y_1 evaluated at $\hat{\phi}$.

A change of variable gives an expansion for the density of $\bar{Y}_n = S_n/n$:

$$\begin{aligned} f_{\bar{Y}_n}(s) &= (2\pi)^{-1/2}\{n/K_Y''(\hat{\phi})\}^{1/2} \\ &\times \exp\big\{n[K_Y(\hat{\phi}) - \hat{\phi}s]\big\}\big(1 + O(n^{-1})\big), \end{aligned}$$

where now $K_Y'(\hat{\phi}) = s$.

Comparison with Edgeworth expansion

To use the saddlepoint expansion to approximate $f_{\bar{Y}_n}(s)$ it is necessary to know the whole cumulant generating function, not just the first four cumulants.

Comparison with Edgeworth expansion

To use the saddlepoint expansion to approximate $f_{\bar{Y}_n}(s)$ it is necessary to know the whole cumulant generating function, not just the first four cumulants.

Also necessary to solve the equation $K'_Y(\hat{\phi}) = s$ for each value of s.

The leading term in saddlepoint expansion is not the normal (or any other) density; in fact it will not usually integrate to 1, although it can be renormalised to do so.

The leading term in saddlepoint expansion is not the normal (or any other) density; in fact it will not usually integrate to 1, although it can be renormalised to do so.

Saddlepoint expansion is an asymptotic expansion in powers of n^{-1}, rather than $n^{-1/2}$ as in the Edgeworth expansion. The main correction for skewness has been absorbed by the leading term.

The leading term in saddlepoint expansion is not the normal (or any other) density; in fact it will not usually integrate to 1, although it can be renormalised to do so.

Saddlepoint expansion is an asymptotic expansion in powers of n^{-1}, rather than $n^{-1/2}$ as in the Edgeworth expansion. The main correction for skewness has been absorbed by the leading term.

It is not as easy to integrate the saddlepoint approximation to obtain an approximation to the distribution function of $\bar{Y}_n$.

Accuracy

Saddlepoint approximation is generally very accurate.

Accuracy

Saddlepoint approximation is generally very accurate.

In distributions that differ from the normal density in terms of asymmetry, such as the gamma distribution, the saddlepoint approximation is extremely accurate throughout the range of s.

Renormalisation

Use as the approximation to $f_{\bar{Y}_n}(s)$ a renormalised version:

$$f_{\bar{Y}_n}(s) \doteq c\{n/K_Y''(\hat{\phi})\}^{1/2} \exp\big[n\{K_Y(\hat{\phi}) - \hat{\phi}s\}\big]$$

where c is determined, usually numerically, so that the right-hand side integrates to 1.

If the $O(n^{-1})$ correction term is constant in s, renormalised approximation will be exact. For scalar random variables this happens only in the case of the normal, gamma and inverse Gaussian distributions.

If the $O(n^{-1})$ correction term is constant in s, renormalised approximation will be exact. For scalar random variables this happens only in the case of the normal, gamma and inverse Gaussian distributions.

In general, the n^{-1} correction term $\{3\hat{\rho}_4(\hat{\phi}) - 5\hat{\rho}_3^2(\hat{\phi})\}/24$ varies only slowly with s and the relative error in the renormalised approximation is $O(n^{-3/2})$.

Distribution function approximation

It is not easy to integrate the saddlepoint approximation to obtain an approximation to the distribution function of S_n.

Lugannani-Rice

The result is given by the Lugannani-Rice approximation:

$$F_{S_n}(s) = \Phi(r_s) + \phi(r_s)\Big(\frac{1}{r_s} - \frac{1}{v_s}\Big) + O(n^{-1}),$$

where

$$\begin{aligned} r_s &= \operatorname{sgn}(\hat{\phi})\sqrt{2n\{\hat{\phi}K'_Y(\hat{\phi}) - K_Y(\hat{\phi})\}} \\ v_s &= \hat{\phi}\sqrt{nK''_Y(\hat{\phi})}, \end{aligned}$$

and $\hat{\phi} \equiv \hat{\phi}(s)$ is the saddlepoint, satisfying $nK'_Y(\hat{\phi}) = s$.

An alternative approximation

The expansion can be expressed in the asymptotically equivalent form

$$F_{S_n}(s) = \Phi(r_s^*)\{1 + O(n^{-1})\},$$

with

$$r_s^* = r_s - \frac{1}{r_s} \log \frac{r_s}{v_s}.$$

Derivation of saddlepoint approximation

Usually derived by one of two methods.

Derivation of saddlepoint approximation

Usually derived by one of two methods.

First uses the inversion formula and contour integration, choosing the contour of integration to pass through the saddlepoint of the integrand 'on the line of steepest descent'.

A more statistical derivation

We associate with the density $f(y)$ for Y_1 a 'tilted' exponential family density $f(y;\lambda)$ defined by

$$f(y;\lambda) = \exp\{y\lambda - K_Y(\lambda)\}f(y)$$

where K_Y is the cumulant generating function of Y_1, under $f(y)$. Then the sum $S_n = Y_1 + \cdots + Y_n$ has associated density

$$f_{S_n}(s;\lambda) = \exp\{s\lambda - nK_Y(\lambda)\}f_{S_n}(s)$$

from which

$$f_{S_n}(s) = \exp\{nK_Y(\lambda) - s\lambda\}f_{S_n}(s;\lambda).$$

Now use the Edgeworth expansion to obtain an approximation to the density $f_{S_n}(s;\lambda)$, remembering that cumulants all must refer to cumulants computed under the tilted density $f(y;\lambda)$.

Now use the Edgeworth expansion to obtain an approximation to the density $f_{S_n}(s;\lambda)$, remembering that cumulants all must refer to cumulants computed under the tilted density $f(y;\lambda)$.

Since λ is arbitrary, it is chosen so that the Edgeworth expansion for the tilted density is evaluated at its mean, where the $n^{-1/2}$ term in the expansion is zero. This value $\hat{\lambda} \equiv \hat{\lambda}(s)$ is defined by $nK'_Y(\hat{\lambda}) = s$ and we obtain

$$f_{S_n}(s) \doteq \exp\{nK_Y(\hat{\lambda}) - \hat{\lambda}s\}\{2\pi nK''_Y(\hat{\lambda})\}^{-1/2}.$$

The factor $\{2\pi nK''_Y(\hat{\lambda})\}^{-1/2}$ comes from the normal density evaluated at its mean.

Exponential family case

Suppose $f(y)$ is itself in the exponential family,

$$f(y;\theta) = \exp\{y\theta - c(\theta) - h(y)\}.$$

Exponential family case

Suppose $f(y)$ is itself in the exponential family,

$$f(y;\theta) = \exp\{y\theta - c(\theta) - h(y)\}.$$

Then since $K_Y(t) = c(\theta + t) - c(\theta)$, it follows that
$\hat{\lambda} \equiv \hat{\lambda}(s) = \hat{\theta} - \theta$, where $\hat{\theta}$ is the MLE based on $s = y_1 + \cdots + y_n$.

The approximation is

$$\begin{aligned} f_{S_n}(s;\theta) &\doteq \exp\big[n\{c(\hat\theta)-c(\theta)\}-(\hat\theta-\theta)s\big] \\ &\times \{2\pi n c''(\hat\theta)\}^{-1/2}, \end{aligned}$$

which can be expressed as

$$c\exp\{l(\theta)-l(\hat\theta)\}|j(\hat\theta)|^{-1/2}$$

where $l(\theta)$ is the log-likelihood function based on $(y_1,\ldots,y_n)$, or s, and $j(\hat\theta)$ is the observed information.

Since $\hat{\theta} = \hat{\theta}(s)$ is a one-to-one function of s, with Jacobian $|j(\hat{\theta})|$, we can obtain an approximation to the density of $\hat{\theta}$

$$f_{\hat{\theta}}(\hat{\theta}; \theta) \doteq c \exp\{l(\theta) - l(\hat{\theta})\}|j(\hat{\theta})|^{1/2}.$$

Laplace approximation of integrals

To evaluate the integral

$$g_n = \int_a^b e^{-ng(y)} dy.$$

Laplace approximation of integrals

To evaluate the integral

$$g_n = \int_a^b e^{-ng(y)} dy.$$

The main contribution, for large n, will come from values of y near the minimum of $g(y)$, which may occur at a or b, or in the interior of the interval (a, b).

Assume that $g(y)$ is minimised at $\tilde{y} \in (a, b)$ and that $g'(\tilde{y}) = 0$, $g''(\tilde{y}) > 0$.

Assume that $g(y)$ is minimised at $\tilde{y} \in (a, b)$ and that $g'(\tilde{y}) = 0$, $g''(\tilde{y}) > 0$.

We can write

$$\begin{aligned} g_n &= \int_a^b e^{-n\{g(\tilde{y}) + \frac{1}{2}(\tilde{y}-y)^2 g''(\tilde{y}) + \cdots\}} dy \\ &\doteq e^{-ng(\tilde{y})} \int_a^b e^{-\frac{n}{2}(\tilde{y}-y)^2 g''(\tilde{y})} dy \\ &\doteq e^{-ng(\tilde{y})} \sqrt{\frac{2\pi}{ng''(\tilde{y})}} \int_{-\infty}^{\infty} \phi\left(y - \tilde{y}; \frac{1}{ng''(\tilde{y})}\right) dy \end{aligned}$$

where $\phi(y - \mu; \sigma^2)$ is the density of $N(\mu, \sigma^2)$.

Since ϕ integrates to one,

$$g_n \doteq e^{-ng(\tilde{y})} \sqrt{\frac{2\pi}{ng''(\tilde{y})}}.$$

A more detailed analysis gives

$$g_n = e^{-ng(\tilde{y})}\sqrt{\frac{2\pi}{ng''(\tilde{y})}}\left\{1+\frac{5\tilde{\rho}_3^2-3\tilde{\rho}_4}{24n}+O(n^{-2})\right\},$$

where

$$\begin{aligned}\tilde{\rho}_3 &= g^{(3)}(\tilde{y})/\{g''(\tilde{y})\}^{3/2},\\ \tilde{\rho}_4 &= g^{(4)}(\tilde{y})/\{g''(\tilde{y})\}^{2}.\end{aligned}$$

A similar analysis gives

$$\int_a^b h(y)e^{-ng(y)}dy = h(\tilde{y})e^{-ng(\tilde{y})}\sqrt{\frac{2\pi}{ng''(\tilde{y})}}\{1+O(n^{-1})\}.$$

A further refinement of the method gives

$$\begin{aligned}
&\int_a^b e^{-n\{g(y)-\frac{1}{n}\log h(y)\}}\,dy \\
&= \int_a^b e^{-nq_n(y)}\,dy, \quad \text{say,} \\
&= e^{-ng(y^*)}h(y^*)\sqrt{\frac{2\pi}{nq_n''(y^*)}} \\
&\times \{1+(5\rho_3^{*2}-3\rho_4^*)/(24n)+O(n^{-2})\},
\end{aligned}$$

where

$$q_n'(y^*)=0, \rho_j^* = q_n^{(j)}(y^*)/\{q_n''(y^*)\}^{j/2}.$$

The p^* formula

Recall the convention that the minimal sufficient statistic can be re-expressed, by a one-to-one smooth transformation, as $(\hat{\theta}, a)$ where a is an exact or approximate ancillary.

The p^* formula

Recall the convention that the minimal sufficient statistic can be re-expressed, by a one-to-one smooth transformation, as $(\hat{\theta}, a)$ where a is an exact or approximate ancillary.

We can write the log-likelihood $l(\theta; y)$ as $l(\theta; \hat{\theta}, a)$.

The p^* formula

Recall the convention that the minimal sufficient statistic can be re-expressed, by a one-to-one smooth transformation, as $(\hat{\theta}, a)$ where a is an exact or approximate ancillary.

We can write the log-likelihood $l(\theta; y)$ as $l(\theta; \hat{\theta}, a)$.

Under a transformation model, the maximal invariant statistic serves as the ancillary.

The p^* formula

Recall the convention that the minimal sufficient statistic can be re-expressed, by a one-to-one smooth transformation, as $(\hat{\theta}, a)$ where a is an exact or approximate ancillary.

We can write the log-likelihood $l(\theta; y)$ as $l(\theta; \hat{\theta}, a)$.

Under a transformation model, the maximal invariant statistic serves as the ancillary.

Under (m, m) exponential models the MLE is minimal sufficient and no ancillary is called for.

Example: Location Model

Have $Y_1, \ldots, Y_n$ independent random variables with

$$Y_j = \theta + \epsilon_j, \;\; j = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent, each having the known density function $\exp\{g(\cdot)\}$.

Example: Location Model

Have $Y_1, \ldots, Y_n$ independent random variables with

$$Y_j = \theta + \epsilon_j, \quad j = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent, each having the known density function $\exp\{g(\cdot)\}$.

The log-likelihood is given by

$$l(\theta) = \sum g(y_j - \theta).$$

Let $a = (a_1, \ldots, a_n)$, where $a_j = y_j - \hat{\theta}$: a is ancillary.

Let $a = (a_1, \ldots, a_n)$, where $a_j = y_j - \hat{\theta}$: a is ancillary.

Write $Y_j = a_j + \hat{\theta}$, so that the log- likelihood is

$$l(\theta; \hat{\theta}, a) = \sum g(a_j + \hat{\theta} - \theta).$$

A Further Example

Let $Y_1, \ldots, Y_n$ be an independent sample from a full (m, m) exponential density

$$\exp\{y^T\theta - k(\theta) + D(y)\}.$$

A Further Example

Let $Y_1, \ldots, Y_n$ be an independent sample from a full (m, m) exponential density

$$\exp\{y^T\theta - k(\theta) + D(y)\}.$$

The log-likelihood is

$$l(\theta) = \sum y_j^T\theta - nk(\theta).$$

Since $\hat{\theta}$ satisfies the likelihood equation

$$\sum y_j - nk'(\hat{\theta}) = 0,$$

the log-likelihood may be written

$$l(\theta; \hat{\theta}) = nk'(\hat{\theta})^T \theta - nk(\theta).$$

Outside exponential families and transformation models it is usually necessary to work with approximate ancillaries.

Useful approximate ancillaries can often be constructed from signed log-likelihood ratios or from score statistics.

Outside exponential families and transformation models it is usually necessary to work with approximate ancillaries.

Useful approximate ancillaries can often be constructed from signed log-likelihood ratios or from score statistics.

Consider scalar θ.

The Efron-Hinkley ancillary

Let i and j be the expected and observed information and let $l_\theta = \frac{\partial l}{\partial \theta}$, $l_{\theta\theta} = \frac{\partial^2 l}{\partial \theta^2}$ etc.

Use the notation $\nu_{2,1} = E(l_{\theta\theta} l_\theta; \theta)$, $\nu_{2,2} = E(l_{\theta\theta} l_{\theta\theta}; \theta)$, $\nu_2 = E(l_{\theta\theta}; \theta)$.

Define

$$\gamma = i^{-1}(\nu_{2,2} - \nu_2^2 - i^{-1}\nu_{2,1}^2)^{1/2},$$

and use circumflex to denote evaluation at $\hat{\theta}$.

Define

$$\gamma = i^{-1}(\nu_{2,2} - \nu_2^2 - i^{-1}\nu_{2,1}^2)^{1/2},$$

and use circumflex to denote evaluation at $\hat{\theta}$.

Then the Efron–Hinkley ancillary is defined by

$$a = (\hat{i}\hat{\gamma})^{-1}(\hat{j} - \hat{i}).$$

The formula

The conditional density function $f(\hat{\theta}; \theta \mid a)$ for the MLE $\hat{\theta}$ given an ancillary statistic a is, in wide generality, exactly or approximately equal to

$$p^*(\hat{\theta}; \theta \mid a) = c(\theta, a)|j(\hat{\theta})|^{1/2} \exp\{l(\theta) - l(\hat{\theta})\},$$

i.e.

$$f(\hat{\theta}; \theta \mid a) \doteq p^*(\hat{\theta}; \theta \mid a).$$

The formula

The conditional density function $f(\hat{\theta}; \theta \mid a)$ for the MLE $\hat{\theta}$ given an ancillary statistic a is, in wide generality, exactly or approximately equal to

$$p^*(\hat{\theta}; \theta \mid a) = c(\theta, a)|j(\hat{\theta})|^{1/2} \exp\{l(\theta) - l(\hat{\theta})\},$$

i.e.

$$f(\hat{\theta}; \theta \mid a) \doteq p^*(\hat{\theta}; \theta \mid a).$$

Here, $c(\theta, a)$ is a normalising constant, determined so that the integral of p^* with respect to $\hat{\theta}$, for fixed a, equals 1.

Discussion

The formula gives the exact conditional distribution of the MLE for a considerable range of models. It is the case for virtually all transformation models, for which $c(\theta, a)$ is independent of θ.

Discussion

The formula gives the exact conditional distribution of the MLE for a considerable range of models. It is the case for virtually all transformation models, for which $c(\theta, a)$ is independent of θ.

Among models for which formula is exact is the inverse Gaussian distribution (not a transformation model). Under many of these models the normalising constant c equals $(2\pi)^{-d/2}$ exactly, $d = \dim(\theta)$.

Discussion

The formula gives the exact conditional distribution of the MLE for a considerable range of models. It is the case for virtually all transformation models, for which $c(\theta, a)$ is independent of θ.

Among models for which formula is exact is the inverse Gaussian distribution (not a transformation model). Under many of these models the normalising constant c equals $(2\pi)^{-d/2}$ exactly, $d = \dim(\theta)$.

In general, $c = c(\theta, a) = (2\pi)^{-d/2}\bar{c}$, where $\bar{c} = 1 + O(n^{-1})$.

Outside the realm of exactness cases, formula is quite generally accurate to relative error of order $O(n^{-3/2})$:

$$f(\hat{\theta};\theta \mid a) = p^*(\hat{\theta};\theta \mid a)\big(1 + O(n^{-3/2})\big).$$

Outside the realm of exactness cases, formula is quite generally accurate to relative error of order $O(n^{-3/2})$:

$$f(\hat{\theta};\theta \mid a) = p^*(\hat{\theta};\theta \mid a)\big(1 + O(n^{-3/2})\big).$$

The p^* formula is equivalent to the saddlepoint approximation in exponential families, with θ the natural parameter.

Distribution function approximation

Integration of the p^* formula to obtain an approximation to the distribution function of the MLE is intricate. Suppose we wish to evaluate $\Pr(\hat{\theta} \leq t; \theta \mid a)$.

Notation

Write

$$r_t \equiv r_t(\theta) = \mathrm{sgn}(t-\theta)\sqrt{2(l(t;t,a) - l(\theta;t,a))},$$

and let

$$v_t \equiv v_t(\theta) = j(t;t,a)^{-1/2}\{l_{;\hat{\theta}}(t;t,a) - l_{;\hat{\theta}}(\theta;t,a)\},$$

in terms of the sample space derivative $l_{;\hat{\theta}}$ defined by

$$l_{;\hat{\theta}}(\theta;\hat{\theta},a) = \frac{\partial}{\partial\hat{\theta}} l(\theta;\hat{\theta},a),$$

and with j the observed information.

The formula

Then

$$P(\hat{\theta} \leq t; \theta \mid a) = \Phi\{r_t^*(\theta)\}\{1 + O(n^{-3/2})\},$$

where $r_t^*(\theta) = r_t(\theta) + r_t(\theta)^{-1} \log\{v_t(\theta)/r_t(\theta)\}$.

The formula

Then

$$P(\hat{\theta} \leq t; \theta \mid a) = \Phi\{r_t^*(\theta)\}\{1 + O(n^{-3/2})\},$$

where $r_t^*(\theta) = r_t(\theta) + r_t(\theta)^{-1} \log\{v_t(\theta)/r_t(\theta)\}$.

The random variable $r^*(\theta)$ corresponding to $r_t^*(\theta)$ [replace fixed t by random $\hat{\theta}$] is an adjusted form of the signed root likelihood ratio statistic, $N(0, 1)$ to (relative) error $O(n^{-3/2})$, conditional on ancillary a.

Conditional inference in exponential families

Suppose that $Y_1, \ldots, Y_n$ are independent, identically distributed from the exponential family density

$$f(y; \psi, \lambda) = \exp\{\psi\tau_1(y) + \lambda\tau_2(y) - d(\psi, \lambda) - Q(y)\},$$

where we will suppose for simplicity that the parameter of interest ψ and the nuisance parameter λ are both scalar.

Conditional inference in exponential families

Suppose that $Y_1, \ldots, Y_n$ are independent, identically distributed from the exponential family density

$$f(y; \psi, \lambda) = \exp\{\psi\tau_1(y) + \lambda\tau_2(y) - d(\psi, \lambda) - Q(y)\},$$

where we will suppose for simplicity that the parameter of interest ψ and the nuisance parameter λ are both scalar.

The natural statistics are $T = n^{-1}\sum\tau_1(y_i)$ and $S = n^{-1}\sum\tau_2(y_i)$. From the general properties of exponential families, the conditional distribution of T given $S = s$ depends only on ψ, so that inference about ψ may be derived from a conditional likelihood, given s.

The log-likelihood based on the full data $y_1, \ldots, y_n$ is

$$n\psi t + n\lambda s - nd(\psi, \lambda),$$

ignoring terms not involving ψ and λ, and a conditional log-likelihood function is the full log-likelihood minus the log-likelihood function based on the marginal distribution of S.

The log-likelihood based on the full data $y_1, \ldots, y_n$ is

$$n\psi t + n\lambda s - nd(\psi, \lambda),$$

ignoring terms not involving ψ and λ, and a conditional log-likelihood function is the full log-likelihood minus the log-likelihood function based on the marginal distribution of S.

We consider an approximation to the marginal distribution of S, based on a saddlepoint approximation to the density of S, evaluated at its observed value s.

The cumulant generating function of $\tau_2(Y_i)$ is given by

$$K(z) = d(\psi, \lambda + z) - d(\psi, \lambda).$$

The cumulant generating function of $\tau_2(Y_i)$ is given by

$$K(z) = d(\psi, \lambda + z) - d(\psi, \lambda).$$

The saddlepoint equation is therefore given by

$$d_\lambda(\psi, \lambda + \hat{z}) = s.$$

With s the observed value of S, the likelihood equation for the model with ψ held fixed is

$$ns - nd_\lambda(\psi, \hat{\lambda}_\psi) = 0,$$

so that $\lambda + \hat{z} = \hat{\lambda}_\psi$, where $\hat{\lambda}_\psi$ denotes the maximum likelihood estimator of λ for fixed ψ.

Applying the saddlepoint approximation, ignoring constants, we approximate the marginal likelihood function based on S as

$$|d_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2} \exp\{n[d(\psi, \hat{\lambda}_\psi) - d(\psi, \lambda)] - (\hat{\lambda}_\psi - \lambda)s\};$$

the resulting approximation to the conditional log-likelihood function is given by

$$n\psi t + n\hat{\lambda}_\psi^T s - nd(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\log|d_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|$$

$$\equiv l(\psi, \hat{\lambda}_\psi) + \frac{1}{2}\log|d_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|.$$

The score function

Given an ancillary a, the MLE $\hat{\theta}$ and the score vector $U = \nabla l$, with components l_r, will in general be in one-to-one correspondence for a region of values of $\hat{\theta}$ around the true parameter value θ, and this region will carry all the probability mass, except for an asymptotically negligible amount.

The score function

Given an ancillary a, the MLE $\hat{\theta}$ and the score vector $U = \nabla l$, with components l_r, will in general be in one-to-one correspondence for a region of values of $\hat{\theta}$ around the true parameter value θ, and this region will carry all the probability mass, except for an asymptotically negligible amount.

Obtain a density approximation for U by multiplying density approximation for $\hat{\theta}$ by Jacobian of (one-to-one) transformation.

Details

The Jacobian of the transformation from $\hat{\theta}$ to the vector of derivatives $l_r = l_r(\theta; \hat{\theta}, a)$ is the matrix $l_{;}$ of mixed second-order log model derivatives

$$l_{r;s} = l_{r;s}(\theta; \hat{\theta}, a) = \frac{\partial}{\partial \theta^r} \frac{\partial}{\partial \hat{\theta}^s} l(\theta; \hat{\theta}, a).$$

Then an approximation of high accuracy to the conditional distribution of the score vector given a is provided by

$$p(u;\theta \mid a) \doteq p^*(u;\theta \mid a),$$

where

$$p^*(u;\theta \mid a) = c(\theta, a)|\hat{j}|^{1/2}|l;|^{-1}e^{l-\hat{l}}.$$

Bartlett correction

The first-order approximation to the distribution of the likelihood ratio statistic $w(\psi)$ is

$$\Pr_\theta\{w(\psi) \leq \omega^\circ\} = P\{\chi^2_q \leq \omega^\circ\}\{1 + O(n^{-1})\},$$

where $q = d_\psi, \theta = (\psi, \lambda)$, say.

In the case of IID sampling, it can be shown that

$$E_\theta w(\psi) = q\{1 + b(\theta)/n + O(n^{-2})\},$$

and so $E_\theta w'(\psi) = q\{1 + O(n^{-2})\}$, where $w' = w/\{1 + b(\theta)/n\}$.

In the case of IID sampling, it can be shown that

$$E_\theta w(\psi) = q\{1 + b(\theta)/n + O(n^{-2})\},$$

and so $E_\theta w'(\psi) = q\{1 + O(n^{-2})\}$, where $w' = w/\{1 + b(\theta)/n\}$.

The adjustment procedure of replacing w by w', is known as Bartlett correction.

Discussion

Bartlett correction yields remarkably good results under continuous models.

Division by $\{1 + b(\theta)/n\}$ adjusts not only the mean but simultaneously all the cumulants—and hence the whole distribution—of w towards those of χ^2_q. It can be shown that

$$P_\theta\{w'(\psi) \leq \omega^\circ\} = P\{\chi^2_q \leq \omega^\circ\}\{1 + O(n^{-2})\}.$$

Discussion

Bartlett correction yields remarkably good results under continuous models.

Division by $\{1 + b(\theta)/n\}$ adjusts not only the mean but simultaneously all the cumulants—and hence the whole distribution—of w towards those of χ^2_q. It can be shown that

$$P_\theta\{w'(\psi) \leq \omega^\circ\} = P\{\chi^2_q \leq \omega^\circ\}\{1 + O(n^{-2})\}.$$

In practice $b(\theta)$ will be replaced by $b(\psi, \hat{\lambda}_\psi)$. The above result still holds, even to $O(n^{-2})$.

The effect of the Bartlett correction is due to the special character of the likelihood ratio statistic, and the same device applied to, for instance, the score test does not have a similar effect.

The effect of the Bartlett correction is due to the special character of the likelihood ratio statistic, and the same device applied to, for instance, the score test does not have a similar effect.

Also, under discrete models this type of adjustment does not generally lead to an improved χ^2 approximation.

Modified profile likelihood

The profile likelihood $L_{\mathrm{p}}(\psi)$ for a parameter of interest ψ can largely be thought of as if it were a genuine likelihood.

Modified profile likelihood

The profile likelihood $L_{\mathrm{p}}(\psi)$ for a parameter of interest ψ can largely be thought of as if it were a genuine likelihood.

This amounts to behaving as if the nuisance parameter over which the maximisation has been carried out were known. Inference on ψ based on treating $L_{\mathrm{p}}(\psi)$ as a proper likelihood may therefore be grossly misleading if the data contain insufficient information about χ, or if there are many nuisance parameters.

Modified profile likelihood

The profile likelihood $L_{\mathrm{p}}(\psi)$ for a parameter of interest ψ can largely be thought of as if it were a genuine likelihood.

This amounts to behaving as if the nuisance parameter over which the maximisation has been carried out were known. Inference on ψ based on treating $L_{\mathrm{p}}(\psi)$ as a proper likelihood may therefore be grossly misleading if the data contain insufficient information about χ, or if there are many nuisance parameters.

Modify.

Definition

The modified profile likelihood $\tilde{L}_{\mathrm{p}}(\psi)$ for a parameter of interest ψ, with nuisance parameter χ, is defined by

$$\tilde{L}_{\mathrm{p}}(\psi) = M(\psi) L_{\mathrm{p}}(\psi),$$

where M is a modifying factor

$$M(\psi) = \left| \frac{\partial \hat{\chi}}{\partial \hat{\chi}_{\psi}} \right| |\hat{j}_{\psi}|^{-1/2}.$$

Here $\partial\hat{\chi}/\partial\hat{\chi}_\psi$ is the matrix of partial derivatives of $\hat{\chi}$ with respect to $\hat{\chi}_\psi$, where $\hat{\chi}$ is considered as a function of $(\hat{\psi}, \hat{\chi}_\psi, a)$ and $\hat{j}_\psi = j_{\chi\chi}(\psi, \hat{\chi}_\psi)$, the observed information on χ assuming ψ is known.

Comments

The modified profile likelihood $\tilde{L}_{\mathrm{p}}$ is, like L_{p}, parameterisation invariant.

Comments

The modified profile likelihood $\tilde{L}_{\mathrm{p}}$ is, like L_{p}, parameterisation invariant.

An alternative expression for the modifying factor M is

$$M(\psi) = |I_{\chi;\hat{\chi}}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a)|^{-1} \times |j_{\chi\chi}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a)|^{1/2}.$$

This follows from the likelihood equation for $\hat{\chi}_\psi$:

$$l_\chi(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a) = 0.$$

Differentiation with respect to $\hat{\chi}$ yields

$$l_{\chi\chi}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a)\frac{\partial \hat{\chi}_\psi}{\partial \hat{\chi}} + l_{\chi;\hat{\chi}}(\psi, \hat{\chi}_\psi; \hat{\psi}, \hat{\chi}, a) = 0.$$

Justification

Asymptotically, $\tilde{L}_{\mathrm{p}}$ and L_{p} are equivalent to first-order.

Justification

Asymptotically, $\tilde{L}_{\mathrm{p}}$ and L_{p} are equivalent to first-order.

The reason for using $\tilde{L}_{\mathrm{p}}$ rather than L_{p} is that the former arises as a higher-order approximation to a marginal likelihood for ψ when such a marginal likelihood function is available, and to a conditional likelihood for ψ when this is available.

Details

Suppose that the density $f(\hat{\psi}, \hat{\chi}; \psi, \chi \mid a)$ factorises, either as

$$f(\hat{\psi}, \hat{\chi}; \psi, \chi \mid a) = f(\hat{\psi}; \psi \mid a) f(\hat{\chi}; \psi, \chi \mid \hat{\psi}, a)$$

or as

$$f(\hat{\psi}, \hat{\chi}; \psi, \chi \mid a) = f(\hat{\chi}; \psi, \chi \mid a) f(\hat{\psi}; \psi \mid \hat{\chi}, a).$$

In the first case modified profile likelihood can be obtained as an approximation (using the p^*-formula) to the marginal likelihood for ψ based on $\hat{\psi}$ and conditional on a, i.e. to the likelihood for ψ determined by $f(\hat{\psi}; \psi \mid a)$.

In the first case modified profile likelihood can be obtained as an approximation (using the p^*-formula) to the marginal likelihood for ψ based on $\hat{\psi}$ and conditional on a, i.e. to the likelihood for ψ determined by $f(\hat{\psi}; \psi \mid a)$.

In the second case it is obtained as an approximation to the conditional likelihood for ψ given $\hat{\chi}$ and a.

Further comments

Note that if $\hat{\chi}_\psi$ does not depend on ψ,

$$\hat{\chi}_\psi = \hat{\chi},$$

then

$$\tilde{L}_{\mathrm{p}}(\psi) = |\hat{j}_\psi|^{-1/2} L_{\mathrm{p}}(\psi).$$

Further comments

Note that if $\hat{\chi}_\psi$ does not depend on ψ,

$$\hat{\chi}_\psi = \hat{\chi},$$

then

$$\tilde{L}_{\mathrm{p}}(\psi) = |\hat{j}_\psi|^{-1/2} L_{\mathrm{p}}(\psi).$$

In the case that ψ and χ are orthogonal, which is a weaker assumption, both hold to order $O(n^{-1})$.

Bayesian asymptotics

The key result is that the posterior distribution is asymptotically normal. Write

$$\pi_n(\theta \mid y) = f(y;\theta)\pi(\theta) / \int f(y;\theta)\pi(\theta)d\theta$$

for the posterior density. Denote by $\hat{\theta}$ the MLE.

Proof

For θ in a neighbourhood of $\hat{\theta}$ we have, by Taylor expansion,

$$\log\left\{\frac{f(y;\theta)}{f(y;\hat{\theta})}\right\} \doteq -\tfrac{1}{2}(\theta-\hat{\theta})^T j(\hat{\theta})(\theta-\hat{\theta}).$$

Proof

For θ in a neighbourhood of $\hat{\theta}$ we have, by Taylor expansion,

$$\log\left\{\frac{f(y;\theta)}{f(y;\hat{\theta})}\right\} \doteq -\tfrac{1}{2}(\theta-\hat{\theta})^T j(\hat{\theta})(\theta-\hat{\theta}).$$

Provided the likelihood dominates the prior, we can approximate $\pi(\theta)$ in a neighbourhood of $\hat{\theta}$ by $\pi(\hat{\theta})$.

Proof

For θ in a neighbourhood of $\hat{\theta}$ we have, by Taylor expansion,

$$\log\left\{\frac{f(y;\theta)}{f(y;\hat{\theta})}\right\} \doteq -\tfrac{1}{2}(\theta-\hat{\theta})^T j(\hat{\theta})(\theta-\hat{\theta}).$$

Provided the likelihood dominates the prior, we can approximate $\pi(\theta)$ in a neighbourhood of $\hat{\theta}$ by $\pi(\hat{\theta})$.

Then we have

$$f(y;\theta)\pi(\theta) \doteq f(y;\hat{\theta})\pi(\hat{\theta})\exp\{-\tfrac{1}{2}(\theta-\hat{\theta})^T j(\hat{\theta})(\theta-\hat{\theta})\}.$$

Then, to first order,

$$\pi_n(\theta \mid y) \sim N\big(\hat{\theta}, j^{-1}(\hat{\theta})\big).$$

Another approximation

When the likelihood does not dominate the prior, expand about the posterior mode $\hat{\theta}_\pi$, which maximises $f(y;\theta)\pi(\theta)$.

Another approximation

When the likelihood does not dominate the prior, expand about the posterior mode $\hat{\theta}_\pi$, which maximises $f(y;\theta)\pi(\theta)$.

Then

$$\pi_n(\theta \mid y) \sim N\big(\hat{\theta}_\pi, j_\pi^{-1}(\hat{\theta}_\pi)\big),$$

where j_π is minus the matrix of second derivatives of $f(y;\theta)\pi(\theta)$.

A more accurate approximation

We have

$$\begin{aligned}\pi_n(\theta \mid y) &= f(y;\theta)\pi(\theta) / \int f(y;\theta)\pi(\theta)d\theta \\ &\doteq \frac{c\exp\{l(\theta;y)\}\pi(\theta)}{\exp\{l(\hat{\theta};y)\}|j(\hat{\theta})|^{-1/2}\pi(\hat{\theta})},\end{aligned}$$

by Laplace approximation of the denominator.

A more accurate approximation

We have

$$\begin{aligned}\pi_n(\theta \mid y) &= f(y;\theta)\pi(\theta) / \int f(y;\theta)\pi(\theta)d\theta \\ &\doteq \frac{c\exp\{l(\theta;y)\}\pi(\theta)}{\exp\{l(\hat{\theta};y)\}|j(\hat{\theta})|^{-1/2}\pi(\hat{\theta})},\end{aligned}$$

by Laplace approximation of the denominator.

We can rewrite as

$$\pi_n(\theta \mid y) \doteq c|j(\hat{\theta})|^{1/2}\exp\{l(\theta) - l(\hat{\theta})\} \times \{\pi(\theta)/\pi(\hat{\theta})\}.$$

Posterior expectations

To approximate to the posterior expectation of a function $g(\theta)$ of interest,

$$E\{g(\theta) \mid y\} = \frac{\int g(\theta) e^{n\bar{l}_n(\theta)} \pi(\theta) d\theta}{\int e^{n\bar{l}_n(\theta)} \pi(\theta) d\theta},$$

where $\bar{l}_n = n^{-1} \sum_{i=1}^{n} \log f(y_i; \theta)$ is the average log-likelihood function.

Posterior expectations

To approximate to the posterior expectation of a function $g(\theta)$ of interest,

$$E\{g(\theta) \mid y\} = \frac{\int g(\theta) e^{n\bar{l}_n(\theta)} \pi(\theta) d\theta}{\int e^{n\bar{l}_n(\theta)} \pi(\theta) d\theta},$$

where $\bar{l}_n = n^{-1} \sum_{i=1}^{n} \log f(y_i; \theta)$ is the average log-likelihood function.

Rewrite the integrals as

$$E\{g(\theta) \mid y\} = \frac{\int e^{n\{\bar{l}_n(\theta) + q(\theta)/n\}} d\theta}{\int e^{n\{\bar{l}_n(\theta) + p(\theta)/n\}} d\theta}$$

and use the modified version of the Laplace approximation.

Applying this to the numerator and denominator gives

$$\begin{aligned} E\{g(\theta) \mid y\} &\doteq \frac{e^{n\bar{l}_n(\theta^*)+q(\theta^*)}}{e^{n\bar{l}_n(\tilde{\theta})+p(\tilde{\theta})}} \\ &\times \frac{\{-n\bar{l}_n''(\tilde{\theta}) - p''(\tilde{\theta})\}^{1/2}}{\{-n\bar{l}_n''(\theta^*) - q''(\theta^*)\}^{1/2}} \frac{\{1+O(n^{-1})\}}{\{1+O(n^{-1})\}} \end{aligned}$$

where θ^* maximises $n\bar{l}_n(\theta) + \log g(\theta) + \log \pi(\theta)$ and $\tilde{\theta}$ maximises $n\bar{l}_n(\theta) + \log \pi(\theta)$.

Applying this to the numerator and denominator gives

$$\begin{aligned} E\{g(\theta) \mid y\} &\doteq \frac{e^{n\bar{l}_n(\theta^*)+q(\theta^*)}}{e^{n\bar{l}_n(\tilde{\theta})+p(\tilde{\theta})}} \\ &\times \frac{\{-n\bar{l}_n''(\tilde{\theta}) - p''(\tilde{\theta})\}^{1/2}}{\{-n\bar{l}_n''(\theta^*) - q''(\theta^*)\}^{1/2}} \frac{\{1+O(n^{-1})\}}{\{1+O(n^{-1})\}} \end{aligned}$$

where θ^* maximises $n\bar{l}_n(\theta) + \log g(\theta) + \log \pi(\theta)$ and $\tilde{\theta}$ maximises $n\bar{l}_n(\theta) + \log \pi(\theta)$.

Detailed analysis shows that the relative error is, in fact, $O(n^{-2})$. If the integrals are approximated in their unmodified form the result is not as accurate.