Statistical Asymptotics

G. Alastair Young

Department of Mathematics
Imperial College London

APTS, April 2008

Motivation

Statistical inference requires approximation, using asymptotic theory, to densities or distribution functions.

Motivation

Statistical inference requires approximation, using asymptotic theory, to densities or distribution functions.

Exact answers are rarely available.

Motivation

Statistical inference requires approximation, using asymptotic theory, to densities or distribution functions.

Exact answers are rarely available.

Approximations based on results of probability theory.

Statistical asymptotics

Theory based on limit results combined in 'statistical asymptotics' with asymptotic techniques from analysis and development of asymptotic expansions.

Statistical asymptotics

Theory based on limit results combined in ‘statistical asymptotics’ with asymptotic techniques from analysis and development of asymptotic expansions.

Often a first-order approximation can be improved by incorporating higher-order terms in an asymptotic expansion.

Statistical asymptotics

Theory based on limit results combined in 'statistical asymptotics' with asymptotic techniques from analysis and development of asymptotic expansions.

Often a first-order approximation can be improved by incorporating higher-order terms in an asymptotic expansion.

Theory underlying approximation techniques is valid as some quantity, typically the sample size n [or more generally some 'amount of information'], goes to infinity, but the approximations obtained can be very accurate even for extremely small sample sizes.

No nuisance parameter case

Denote by l_r the rth component of $U(\theta)$, by l_{rs} the (r, s)th component of $\nabla_\theta \nabla_\theta^T l$. Let $[l_{rs}]^{-1} = [l^{rs}]$.

The maximum likelihood estimate for given observations y is, for regular problems, defined as the solution, assumed unique, of the likelihood equation

$$u(\hat{\theta}; y) = 0.$$

Test statistics

To test the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ.

Test statistics

To test the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ.

Three statistics that typically differ by $O_p(n^{-1/2})$ are:

Test statistics

To test the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ.

Three statistics that typically differ by $O_p(n^{-1/2})$ are:

(1) the likelihood ratio statistic

$$w(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\},$$

Test statistics

To test the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ.

Three statistics that typically differ by $O_p(n^{-1/2})$ are:

(1) the likelihood ratio statistic

$$w(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\},$$

(2) the score statistic

$$w_U(\theta_0) = U^T(\theta_0) i^{-1}(\theta_0) U(\theta_0),$$

Test statistics

To test the null hypothesis $H_0 : \theta = \theta_0$, where θ_0 is an arbitrary, specified, point in Ω_θ.

Three statistics that typically differ by $O_p(n^{-1/2})$ are:

(1) the likelihood ratio statistic

$$w(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\},$$

(2) the score statistic

$$w_U(\theta_0) = U^T(\theta_0) i^{-1}(\theta_0) U(\theta_0),$$

(3) the Wald statistic

$$w_p(\theta_0) = (\hat{\theta} - \theta_0)^T i(\theta_0)(\hat{\theta} - \theta_0).$$

Scalar case

For a scalar θ, (1) may be replaced by

$$r(\theta_0) = \operatorname{sgn}(\hat{\theta} - \theta_0)\sqrt{w(\theta_0)},$$

the signed root likelihood ratio statistic.

Scalar case

For a scalar θ, (1) may be replaced by

$$r(\theta_0) = \mathrm{sgn}(\hat{\theta} - \theta_0)\sqrt{w(\theta_0)},$$

the signed root likelihood ratio statistic.

Also (2) and (3) may be replaced by

$$r_U(\theta_0) = U(\theta_0)/\sqrt{i(\theta_0)}$$

and

$$r_p(\theta_0) = (\hat{\theta} - \theta_0)\sqrt{i(\theta_0)}$$

respectively.

Distributions

In a first-order asymptotic theory, the statistics (1)–(3) have, asymptotically, the chi-squared distribution with $d_\theta = \dim(\Omega_\theta)$ degrees of freedom. The 'signed root' versions have an $N(0, 1)$ distribution.

Distributions

In a first-order asymptotic theory, the statistics (1)–(3) have, asymptotically, the chi-squared distribution with $d_\theta = \dim(\Omega_\theta)$ degrees of freedom. The 'signed root' versions have an $N(0, 1)$ distribution.

Confidence regions at level $1 - \alpha$ are formed approximately as, for example,

$$\{\theta : w(\theta) \leq \chi^2_{d_\theta,\alpha}\},$$

where $\chi^2_{d_\theta,\alpha}$ is the upper α point of the relevant chi-squared distribution.

Orders

Since $U(\theta_0)$ and $i(\theta_0)$ refer to the total vector Y of dimension n, then as $n \to \infty$:

$$
\begin{aligned}
U(\theta_0) &\equiv \sqrt{n}\bar{U}(\theta_0) = O_p(n^{1/2}), \\
i(\theta_0) &\equiv n\bar{i}(\theta_0) = O(n), \\
\hat{\theta} - \theta_0 &= O_p(n^{-1/2}),
\end{aligned}
$$

where $\bar{i}(\theta_0)$ is the average information per observation and $\bar{U}(\theta_0)$ is a normalised score function. If the observations are IID, $\bar{i}$ is the information for a single observation.

Estimation of information

As $n \to \infty$, we have in probability that, provided $i(\theta)$ is continuous at $\theta = \theta_0$,

$$\begin{aligned} j(\hat{\theta})/n &\to \bar{i}(\theta_0), \\ j(\theta_0)/n &\to \bar{i}(\theta_0). \end{aligned}$$

Estimation of information

As $n \to \infty$, we have in probability that, provided $i(\theta)$ is continuous at $\theta = \theta_0$,

$$\begin{aligned} j(\hat{\theta})/n &\to \bar{i}(\theta_0), \\ j(\theta_0)/n &\to \bar{i}(\theta_0). \end{aligned}$$

Therefore, in the definitions of the various statistics, $i(\theta_0)$ can be replaced by $i(\hat{\theta})$, $j(\hat{\theta})$, $j(\theta_0)$ etc. etc.

If $\theta = \theta_0$, the various modified statistics differ typically by $O_p(n^{-1/2})$, so that their asymptotic distributions are the same under H_0.

Distribution theory

A serious issue concerns the asymptotic existence, uniqueness and consistency of the maximum likelihood estimate. There are no very satisfactory general theorems on such questions.

Distribution theory

A serious issue concerns the asymptotic existence, uniqueness and consistency of the maximum likelihood estimate. There are no very satisfactory general theorems on such questions.

We assume that $\hat{\theta}$ is well defined and consistent.

In considerable generality U is asymptotically normal with zero mean and variance $i(\theta)$.

In considerable generality U is asymptotically normal with zero mean and variance $i(\theta)$.

Suppose that $U(\theta) = U(\theta; Y) = [l_r(\theta)]$ has been shown to be asymptotically $N_d(0, i(\theta))$,

$$\{n\bar{i}(\theta)\}^{-1/2} U(\theta) \xrightarrow{d} N_d\left(0, I_d\right),$$

I_d is identity matrix, '$1/2'$ indicates matrix square root.

An aside: summation convention

Whenever an index occurs both as a subscript and as a superscript in an expression, summation over possible values of that index is to be assumed.

Distribution of $\hat{\theta}$

Expand the score $l_r(\theta)$ in a Taylor series around θ, writing

$$\begin{aligned} l_r(\theta) &= U_r(\theta) = \sqrt{n}\bar{l}_r(\theta) = \sqrt{n}\bar{U}_r(\theta), \\ l_{rs}(\theta) &= n\bar{l}_{rs}(\theta) = -j_{rs}(\theta) = -n\bar{j}_{rs}(\theta), \\ \bar{\delta}^r &= \sqrt{n}(\hat{\theta}^r - \theta^r), l_{rst}(\theta) = n\bar{l}_{rst}(\theta), \\ i(\theta) &= n\bar{i}(\theta), \text{ etc.} \end{aligned}$$

Then, $l_r(\hat{\theta}) = 0$, so

$$\begin{aligned} \sqrt{n}\bar{l}_r(\theta) &+ n\bar{l}_{rs}(\theta)\bar{\delta}^s/\sqrt{n} \\ &+ \tfrac{1}{2}n\bar{l}_{rst}(\theta)\bar{\delta}^s\bar{\delta}^t/n + \cdots = 0. \end{aligned}$$

To a first-order approximation, ignoring the third term, we have

$$\begin{aligned}\bar{\delta}^r &= -\bar{l}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}) \\ &= \bar{j}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}).\end{aligned}$$

To a first-order approximation, ignoring the third term, we have

$$\begin{aligned}\bar{\delta}^r &= -\bar{l}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}) \\ &= \bar{j}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}).\end{aligned}$$

Now $j^{rs}/i^{rs} \xrightarrow{p} 1$, so

$$\bar{\delta}^r = \bar{i}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}),$$

a linear function of asymptotically normal variables of zero mean.

To a first-order approximation, ignoring the third term, we have

$$\begin{aligned}\bar{\delta}^r &= -\bar{l}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}) \\ &= \bar{j}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}).\end{aligned}$$

Now $j^{rs}/i^{rs} \xrightarrow{p} 1$, so

$$\bar{\delta}^r = \bar{i}^{rs}(\theta)\bar{l}_s(\theta) + O_p(n^{-1/2}),$$

a linear function of asymptotically normal variables of zero mean. It follows that $[\bar{\delta}^r]$ is asymptotically normal with zero mean and covariance matrix $[\bar{i}^{rs}]$. We have

$$\{n\bar{i}(\theta)\}^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N_d(0, I_d).$$

Other quantities

By direct expansion in θ around $\hat{\theta}$ we have, writing $\hat{j}_{rs} = j_{rs}(\hat{\theta})$,

$$w(\theta) = \hat{j}_{rs}(\hat{\theta} - \theta)^r(\hat{\theta} - \theta)^s + o_p(1)$$

or equivalently

$$w(\theta) = i^{rs} l_r l_s + o_p(1),$$

so $w(\theta) \xrightarrow{d} \chi^2_d$.

Other quantities

By direct expansion in θ around $\hat{\theta}$ we have, writing $\hat{j}_{rs} = j_{rs}(\hat{\theta})$,

$$w(\theta) = \hat{j}_{rs}(\hat{\theta} - \theta)^r(\hat{\theta} - \theta)^s + o_p(1)$$

or equivalently

$$w(\theta) = i^{rs} l_r l_s + o_p(1),$$

so $w(\theta) \xrightarrow{d} \chi^2_d$.

The asymptotic χ^2 distribution of the Wald and score statistics follows similarly.

Signed root statistic

When the dimension of θ is $d = 1$, we have that the signed root likelihood ratio statistic

$$r = \mathrm{sgn}(\hat{\theta} - \theta)\sqrt{w(\theta)}$$

satisfies

$$r = \hat{j}^{-1/2} U + o_p(1)$$

so that $r \xrightarrow{d} N(0, 1)$.

A Confidence Interval

For scalar θ, we have $i(\hat{\theta})^{1/2}(\hat{\theta} - \theta)$ asymptotically $N(0, 1)$, so an approximate $100(1 - \alpha)\%$ confidence interval for θ is

$$\hat{\theta} \mp i(\hat{\theta})^{-1/2}\Phi^{-1}(1 - \alpha/2).$$

Profile likelihood

Consider the multiparameter problem in which $\theta = (\theta^1, \dots, \theta^d) \in \Omega_\theta$, an open subset of $\mathbb{R}^d$.

Profile likelihood

Consider the multiparameter problem in which $\theta = (\theta^1, \dots, \theta^d) \in \Omega_\theta$, an open subset of $\mathbb{R}^d$.

Interest lies in inference for a subparameter or parameter function $\psi = \psi(\theta)$.

Profile likelihood

Consider the multiparameter problem in which $\theta = (\theta^1, \ldots, \theta^d) \in \Omega_\theta$, an open subset of $\mathbb{R}^d$.

Interest lies in inference for a subparameter or parameter function $\psi = \psi(\theta)$.

The profile likelihood $L_{\rm p}(\psi)$ for ψ is

$$L_{\rm p}(\psi) = \sup_{\theta : \psi(\theta) = \psi} L(\theta),$$

the supremum of $L(\theta)$ over all θ that are consistent with the given value of ψ.

Profile likelihood

Consider the multiparameter problem in which $\theta = (\theta^1, \ldots, \theta^d) \in \Omega_\theta$, an open subset of $\mathbb{R}^d$.

Interest lies in inference for a subparameter or parameter function $\psi = \psi(\theta)$.

The profile likelihood $L_\mathrm{p}(\psi)$ for ψ is

$$L_\mathrm{p}(\psi) = \sup_{\theta : \psi(\theta) = \psi} L(\theta),$$

the supremum of $L(\theta)$ over all θ that are consistent with the given value of ψ.

The profile log-likelihood is $l_\mathrm{p} = \log L_\mathrm{p}$.

The usual case

Usually ψ is a component of a given partition $\theta = (\psi, \chi)$ of θ into sub-vectors ψ and χ of dimension $d_\psi = d - d_\chi$ and d_χ respectively.

Then

$$L_{\rm p}(\psi) = L(\psi, \hat{\chi}_\psi),$$

where $\hat{\chi}_\psi$ denotes the maximum likelihood estimate of χ for a given value of ψ.

Properties of profile likelihood

Properties of profile likelihood

The maximum profile likelihood estimate of ψ equals $\hat{\psi}$.

Properties of profile likelihood

The maximum profile likelihood estimate of ψ equals $\hat{\psi}$.

The profile log-likelihood ratio statistic $2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\}$ equals the log-likelihood ratio statistic for $H_0 : \psi = \psi_0$,

$$2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\} \equiv 2\{l(\hat{\psi}, \hat{\chi}) - l(\psi_0, \hat{\chi}_0)\} \equiv w(\psi_0),$$

where l is the log-likelihood and $\hat{\chi}_0 \equiv \hat{\chi}_{\psi_0}$.

Properties of profile likelihood

The maximum profile likelihood estimate of ψ equals $\hat{\psi}$.

The profile log-likelihood ratio statistic $2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\}$ equals the log-likelihood ratio statistic for $H_0 : \psi = \psi_0$,

$$2\{l_{\mathrm{p}}(\hat{\psi}) - l_{\mathrm{p}}(\psi_0)\} \equiv 2\{l(\hat{\psi}, \hat{\chi}) - l(\psi_0, \hat{\chi}_0)\} \equiv w(\psi_0),$$

where l is the log-likelihood and $\hat{\chi}_0 \equiv \hat{\chi}_{\psi_0}$.

The asymptotic null distribution of the profile log-likelihood ratio statistic is $\chi^2_{d_\psi}$.

Multiparameter problems: further statistics

To test $H_0 : \psi = \psi_0$, in the presence of nuisance parameter χ.

Multiparameter problems: further statistics

To test $H_0 : \psi = \psi_0$, in the presence of nuisance parameter χ.

Partition the maximum likelihood estimate, the score vector, the information matrix and its inverse:

$$
\begin{aligned}
U(\theta) &= \begin{pmatrix} U_\psi(\psi,\chi) \\ U_\chi(\psi,\chi) \end{pmatrix}, \\
i(\theta) &= \begin{bmatrix} i_{\psi\psi}(\psi,\chi) & i_{\psi\chi}(\psi,\chi) \\ i_{\chi\psi}(\psi,\chi) & i_{\chi\chi}(\psi,\chi) \end{bmatrix}, \\
i^{-1}(\theta) &= \begin{bmatrix} i^{\psi\psi}(\psi,\chi) & i^{\psi\chi}(\psi,\chi) \\ i^{\chi\psi}(\psi,\chi) & i^{\chi\chi}(\psi,\chi) \end{bmatrix}.
\end{aligned}
$$

Wald statistic

We have $\hat{\psi}$ asymptotically normally distributed with mean ψ_0 and covariance matrix $i^{\psi\psi}(\psi_0, \chi_0)$, which can be replaced by $i^{\psi\psi}(\psi_0, \hat{\chi}_0)$.

Wald statistic

We have $\hat{\psi}$ asymptotically normally distributed with mean ψ_0 and covariance matrix $i^{\psi\psi}(\psi_0, \chi_0)$, which can be replaced by $i^{\psi\psi}(\psi_0, \hat{\chi}_0)$.

So a version of the Wald test statistic for the nuisance parameter case is:

$$w_p(\psi_0) = (\hat{\psi} - \psi_0)^T [i^{\psi\psi}(\psi_0, \hat{\chi}_0)]^{-1} (\hat{\psi} - \psi_0).$$

Score statistic

A version of the score statistic for testing $H_0 : \psi = \psi_0$ is:

$$w_u(\psi_0) = U_\psi(\psi_0, \hat{\chi}_0)^T i^{\psi\psi}(\psi_0, \hat{\chi}_0) U_\psi(\psi_0, \hat{\chi}_0).$$

Score statistic

A version of the score statistic for testing $H_0 : \psi = \psi_0$ is:

$$w_u(\psi_0) = U_\psi(\psi_0, \hat{\chi}_0)^T i^{\psi\psi}(\psi_0, \hat{\chi}_0) U_\psi(\psi_0, \hat{\chi}_0).$$

This test has the advantage that MLE has to be obtained only under H_0, and is derived from the asymptotic normality of U.

Asymptotic distributions

Both $w_p(\psi_0)$ and $w_u(\psi_0)$ have asymptotically a chi-squared distribution with d_ψ degrees of freedom.

Effects of parameter orthogonality

Assume that it is possible to make the parameter of interest ψ and the nuisance parameter, now denoted by λ, orthogonal.

Effects of parameter orthogonality

Assume that it is possible to make the parameter of interest ψ and the nuisance parameter, now denoted by λ, orthogonal.

Any transformation from, say, (ψ, χ) to (ψ, λ) necessary to achieve this leaves the profile log-likelihood unchanged.

The matrices $i(\psi, \lambda)$ and $i^{-1}(\psi, \lambda)$ are block diagonal. Therefore, $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent.

The matrices $i(\psi, \lambda)$ and $i^{-1}(\psi, \lambda)$ are block diagonal. Therefore, $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent.

Also, $\hat{\lambda}_\psi$, the MLE of λ for specified ψ, varies only slowly in ψ in the neighbourhood of $\hat{\psi}$, and there is a corresponding slow variation of $\hat{\psi}_\lambda$ with λ: if $\psi - \hat{\psi} = O_p(n^{-1/2})$, then $\hat{\lambda}_\psi - \hat{\lambda} = O_p(n^{-1})$.

The matrices $i(\psi, \lambda)$ and $i^{-1}(\psi, \lambda)$ are block diagonal. Therefore, $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent.

Also, $\hat{\lambda}_\psi$, the MLE of λ for specified ψ, varies only slowly in ψ in the neighbourhood of $\hat{\psi}$, and there is a corresponding slow variation of $\hat{\psi}_\lambda$ with λ: if $\psi - \hat{\psi} = O_p(n^{-1/2})$, then $\hat{\lambda}_\psi - \hat{\lambda} = O_p(n^{-1})$.

For a nonorthogonal nuisance parameter χ, we would have $\hat{\chi}_\psi - \hat{\chi} = O_p(n^{-1/2})$.

Sketch Proof, Scalar Case

If $\psi - \hat{\psi} = O_p(n^{-1/2}), \chi - \hat{\chi} = O_p(n^{-1/2})$, we have

$$l(\psi, \chi) = l(\hat{\psi}, \hat{\chi}) - \frac{1}{2}\{\hat{j}_{\psi\psi}(\psi - \hat{\psi})^2 \\ + 2\hat{j}_{\psi\chi}(\psi - \hat{\psi})(\chi - \hat{\chi}) + \hat{j}_{\chi\chi}(\chi - \hat{\chi})^2\} + O_p(n^{-1/2}).$$

It then follows that

$$\begin{aligned}\hat{\chi}_\psi - \hat{\chi} &= \frac{-\hat{j}_{\psi\chi}}{\hat{j}_{\chi\chi}}(\psi - \hat{\psi}) + O_p(n^{-1}) \\ &= \frac{-i_{\psi\chi}}{i_{\chi\chi}}(\psi - \hat{\psi}) + O_p(n^{-1}).\end{aligned}$$

Then, because $\psi - \hat{\psi} = O_p(n^{-1/2})$, $\hat{\chi}_\psi - \hat{\chi} = O_p(n^{-1/2})$ unless $i_{\psi\chi} = 0$, the orthogonal case, when the difference is $O_p(n^{-1})$.

Further remarks

So far as asymptotic theory is concerned, we can have $\hat{\chi}_\psi = \hat{\chi}$ independently of ψ only if χ and ψ are orthogonal. In this special case we can write $l_{\mathrm{p}}(\psi) = l(\psi, \hat{\chi})$.

Further remarks

So far as asymptotic theory is concerned, we can have $\hat{\chi}_\psi = \hat{\chi}$ independently of ψ only if χ and ψ are orthogonal. In this special case we can write $l_\mathrm{p}(\psi) = l(\psi, \hat{\chi})$.

In the general orthogonal case, $l_\mathrm{p}(\psi) = l(\psi, \hat{\chi}) + o_p(1)$, so that a first-order theory could use $l_\mathrm{p}^*(\psi) = l(\psi, \hat{\chi})$ instead of $l_\mathrm{p}(\psi) = l(\psi, \hat{\chi}_\psi)$.

Distribution theory

The log-likelihood ratio statistic $w(\psi_0)$ can be written as

$$w(\psi_0) = 2\{l(\hat{\psi}, \hat{\chi}) - l(\psi_0, \chi)\} - 2\{l(\psi_0, \hat{\chi}_0) - l(\psi_0, \chi)\},$$

as the difference of two statistics for testing hypotheses without nuisance parameters.

Taylor expansion about (ψ_0, χ), where χ is the true value of the nuisance parameter, gives, to first-order (i.e. ignoring terms of order $o_p(1)$),

$$
\begin{aligned}
w(\psi_0) &= \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix}^T i(\psi_0, \chi) \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} \\
&\quad - (\hat{\chi}_0 - \chi)^T i_{\chi\chi}(\psi_0, \chi)(\hat{\chi}_0 - \chi).
\end{aligned}
$$

The linearised form of the MLE equations is

$$\begin{bmatrix} i_{\psi\psi} & i_{\psi\chi} \\ i_{\chi\psi} & i_{\chi\chi} \end{bmatrix} \begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} = \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix},$$

so

$$\begin{bmatrix} \hat{\psi} - \psi_0 \\ \hat{\chi} - \chi \end{bmatrix} = \begin{bmatrix} i^{\psi\psi} & i^{\psi\chi} \\ i^{\chi\psi} & i^{\chi\chi} \end{bmatrix} \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix}.$$

Also $\hat{\chi}_0 - \chi = i_{\chi\chi}^{-1} U_\chi$, to first-order. Then, to first-order,

$$w(\psi_0) = [U_\psi^T \, U_\chi^T] \begin{bmatrix} i^{\psi\psi} & i^{\psi\chi} \\ i^{\chi\psi} & i^{\chi\chi} \end{bmatrix} \begin{bmatrix} U_\psi \\ U_\chi \end{bmatrix} - U_\chi^T i_{\chi\chi}^{-1} U_\chi.$$

Then,

$$w(\psi_0) \sim Q_U - Q_{U_\chi} = Q_{U_\psi . U_\chi},$$

a difference of two quadratic forms, and is thus asymptotically $\chi^2_{d_\psi}$, by properties of multivariate normal distribution.

The Wald statistic $w_p(\psi_0)$ is based directly on a quadratic form of $\hat{\psi} - \psi_0$, and so can be seen immediately to be asymptotically $\chi^2_{d_\psi}$, from properties of the multivariate normal distribution.

The Wald statistic $w_p(\psi_0)$ is based directly on a quadratic form of $\hat{\psi} - \psi_0$, and so can be seen immediately to be asymptotically $\chi^2_{d_\psi}$, from properties of the multivariate normal distribution.

Note that to first-order we have

$$w_p(\psi_0) = [i^{\psi\psi} U_\psi + i^{\psi\chi} U_\chi]^T (i^{\psi\psi})^{-1} [i^{\psi\psi} U_\psi + i^{\psi\chi} U_\chi].$$

We can express the statistic $w_U(\psi_0)$ in terms of the score vector U. To first-order we have

$$w_U(\psi_0) = (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi)^T i^{\psi\psi} (U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi).$$

This follows since, to first-order,

$$\begin{aligned} U_\psi(\psi_0, \hat{\chi}_0) &= U_\psi + \frac{\partial U_\psi}{\partial \chi}(\hat{\chi}_0 - \chi) \\ &= U_\psi - i_{\psi\chi} i_{\chi\chi}^{-1} U_\chi. \end{aligned}$$

The equivalence of the three statistics, and therefore the asymptotic distribution of $w_U(\psi_0)$, follows on showing that the three first order quantities are identical.