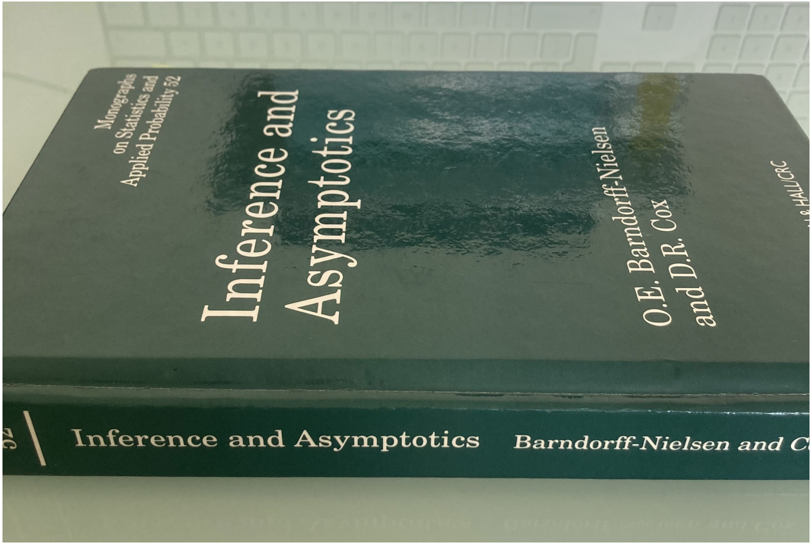
O. E. Barndorff-Nielsen's approximate conditional inference

Heather Battey
Department of Mathematics, Imperial College London

Aarhus, May 29, 2024


Monographs on Statistics and Applied Probability 52
Inference and Asymptotics
O.E. Barndorff-Nielsen and D.R. Cox
HALL/CRC
52
Inference and Asymptotics
Barndorff-Nielsen and

O. E. Barndorff-Nielsen's p^* formula

$$
\begin{aligned}
p^*(\hat{\theta};\theta|a^o) &= c|\hat{j}|^{1/2}e^{\bar{\ell}} \\
\bar{\ell} &= \ell(\theta;\hat{\theta},a^o) - \ell(\hat{\theta};\hat{\theta},a^o)
\end{aligned}
$$

Higher-order approximation by p^*

Under ordinary repeated sampling with sample size n

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

Approximation to the density function of the maximum likelihood estimator of θ, conditional on an ancillary statistic.

(Holds also for second-order approximate ancillaries).

O. E. Barndorff-Nielsen's p^* formula

$$
\begin{aligned}
p^*(\hat{\theta};\theta|a^o) &= c|\hat{j}|^{1/2}e^{\bar{\ell}} \\
\bar{\ell} &= \ell(\theta;\hat{\theta},a^o) - \ell(\hat{\theta};\hat{\theta},a^o)
\end{aligned}
$$

$\hat{\theta}$: arbitrary evaluation point.

θ: unknown parameter of the model (assumed correctly specified).

O. E. Barndorff-Nielsen's p^* formula

$$
\begin{aligned}
p^*(\hat{\theta};\theta|a^o) &= c|\hat{j}|^{1/2}e^{\bar{\ell}} \\
\bar{\ell} &= \ell(\theta;\hat{\theta},a^o) - \ell(\hat{\theta};\hat{\theta},a^o)
\end{aligned}
$$

$\hat{\theta}$: arbitrary evaluation point.

θ: unknown parameter of the model (assumed correctly specified).

ℓ: log-likelihood function.

O. E. Barndorff-Nielsen's p^* formula

$$
\begin{aligned}
p^*(\hat{\theta}; \theta | a^o) &= c|\hat{j}|^{1/2} e^{\bar{\ell}} \\
\bar{\ell} &= \ell(\theta; \hat{\theta}, a^o) - \ell(\hat{\theta}; \hat{\theta}, a^o)
\end{aligned}
$$

$\hat{\theta}$: arbitrary evaluation point.

θ: unknown parameter of the model (assumed correctly specified).

ℓ: log-likelihood function.

a^o: observed value of an (approximate) ancillary statistic.

O. E. Barndorff-Nielsen's p^* formula

$$
\begin{aligned}
p^*(\hat{\theta};\theta|a^o) &= c|\hat{j}|^{1/2}e^{\bar{\ell}} \\
\bar{\ell} &= \ell(\theta;\hat{\theta},a^o) - \ell(\hat{\theta};\hat{\theta},a^o)
\end{aligned}
$$

$\hat{\theta}$: arbitrary evaluation point.

θ: unknown parameter of the model (assumed correctly specified).

ℓ: log-likelihood function.

a^o: observed value of an (approximate) ancillary statistic.

$|\hat{j}|$: determinant of the observed information evaluated at $\hat{\theta}$.

A more explicit notation

$$f_{\hat{\theta}|A}(t|a^o;\theta)dt \simeq c(\theta,a^o)\left|j(t;y(t,a^o))\right|^{1/2}\exp\left\{\ell(\theta;y(t,a^o)) - \ell(t;y(t,a^o))\right\}dt$$

where t is an arbitrary evaluation point for the conditional density function of $\hat{\theta}$ and $y(t,a^o)$ is any value of $y=(y_1,\ldots,y_n)$ such that $A(y)=a^o$ and $\hat{\theta}(y)=t$.

Motivation for p^*

Exact conditional inference is compelling but:

- is only available in limited settings;
- even when available, typically takes great ingenuity.

p^* is intended to apply seamlessly to any problem (caveats).

$p^* \rightarrow$ **approximate conditional inference**

A "likelihood function" L_{MP} based on a version of p^* for nuisance parameters is called a modified (profile) likelihood function.

In most examples where exact conditional inference is available, inference based on L_{MP} coincides with exact inference to higher-order accuracy in n.

The problem of conditioning

Bardorff-Nielsen and Cox, 1994, p. 32

Consider a population of individuals and an event A of interest, for instance that an individual dies of heart disease before age 70. ...Now suppose that a series of new individuals is drawn randomly from the population under study and for each it is required to calculate the probability of event A .... If each probability is to be relevant to the individual in question, it must be conditional on observed relevant features, such as age, sex, smoking habits and blood pressure. ...
... Note, however, that, especially if we condition directly, we must limit the conditioning: otherwise we would reach the position where each individual is not only unique, but also uninformative about other individuals ....

Two types of conditioning

- Conditioning by model formulation: conditioning synonymous with specification of the model.
- Technical conditioning: abstract (model+data)-based partitioning of the sample space.

Two types of conditioning

- Conditioning by model formulation: conditioning synonymous with specification of the model.
- Technical conditioning: abstract (model+data)-based partitioning of the sample space.

Fisherian inferential separations specify where to limit the conditioning to ensure relevance while avoiding degeneracy.

Exact conditional inference

Notation

Model for random variable Y parametrised by θ and provisionally assumed true:

$$f_Y(y;\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta)$$

Arbitrary evaluation point $y = (y_1, \ldots, y_n)$.

Sufficiency reduction, e.g. $s(y) = \sum_i y_i$.

Observed outcome y^o.

Sufficient statistic $S = s(Y)$.

Observed value $s^o = s(y^o)$.

Sufficiency reduction

All information in Y relevant for inference on θ is encapsulated in $S = s(Y)$.

$$f_Y(y;\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i;\theta) = g(s(y);\theta)h(y)$$

Take S to be minimal sufficient, i.e. of lowest dimension.

Minimal sufficiency

Let d be the dimension of S. Let d_θ be the dimension of θ.

If $d > d_\theta$, then any estimator of θ must sacrifice information on θ by the definition of minimal sufficiency.

Starting point for p^*: determine a one-to-one transformation of the minimal sufficient statistic $S \cong (\hat{\theta}, A)$ where A is an ancillary statistic.

(If S is minimal sufficient, then so is $S' = (\hat{\theta}, A) \cong S$, so without loss of generality, take $S = (\hat{\theta}, A)$).

Separations within the minimal sufficient statistic

Likelihood function depends on the data only through S.

Realisable separation $S = (\hat{\theta}, A)$.

Notional idealised separation $S = (C(A), A)$.

Separates the information in S into components of dimensions d_θ and d_A without loss or redundancy.

Notional idealised separation

Notional idealised separation $S = (C(A), A)$.

Ancillary A; "maximal co-ancillary" $C(A)$

$$C(a^o) \stackrel{d}{=} S \mid \{A = a^o\}.$$

The observed value $a^o = a(y^o) = a(s^o)$ leaves $d_\theta = d - d_A$ degrees of freedom of variation of S consistent with the constraint $a(s) = a^o$.

Think of $C(a^o)$ as having a distribution on the d_θ-dimensional co-ancillary manifold:

$$\mathcal{C}(a^o) = \{s \in \mathbb{R}^d : a(s) = a^o\} \subset \mathbb{R}^d.$$

Ancillary statistic A

Ancillary A is defined through its properties w.r.t. θ.

Several property-based definitions have been put forward of varying stringency (e.g. B-N & Cox, 1994, p. 38).

Idealised situation: distribution of A does not depend on θ.

That does not mean that A is irrelevant for inference on θ (A is part of the minimal sufficient statistic).

It means that A, by itself, carries no info on the value of θ.

A vague but practically useful definition

Ancillary statistic: *A is ancillary for θ if, from observation of A alone, no information about the value of θ can in general be extracted.*

This appears to be the implicit definition used by Fisher.

Formalised constructions along these lines have been proposed e.g. Barndorff-Nielsen (1973). On *M*-ancillarity. *Biometrika*, 60, 447–455.

Relevance through conditioning

The conditioning event $\{A = a^o\}$ isolates hypothetical samples for which $s^o = (\hat{\theta}^o, a^o)$ is one realisation, retaining only the variability in S that is relevant for determining the horizontal position of the normed log-likelihood function, rather than its shape, the latter being fixed by $\{A = a^o\}$.

Hypothetical replication

Inferential statements about θ inevitably involve hypothetical replication.

Two samples of the same size can produce log-likelihood functions that differ appreciably in shape, and yet are maximized at the same point.

Example: linear regression. Relevant precision characterised by $X^{\mathrm{T}}X$, not $\mathbb{E}(X^{\mathrm{T}}X)$: $X^{\mathrm{T}}X$ is ancillary when X is considered random.

The ancillary A separates samples of the same size according to their information content.

An exact conditional analysis with nuisance parameters

2×2 **table in original and standardised form**

	0 failure	1 success	
0 control	$N_{0\|0}$	$N_{1\|0}$	$N_{\bullet\|0}$
1 treated	$N_{0\|1}$	$N_{1\|1}$	$N_{\bullet\|1}$
	$N_{0\|\bullet}$	$N_{1\|\bullet}$	N

	0 failure	1 success	
0 control	$\hat{p}_{0\|0}$	$\hat{p}_{1\|0}$	$\hat{p}_{\bullet\|0}$
1 treated	$\hat{p}_{0\|1}$	$\hat{p}_{1\|1}$	$\hat{p}_{\bullet\|1}$
	$\hat{p}_{0\|\bullet}$	$\hat{p}_{1\|\bullet}$	1

Degrees of freedom for 2×2 table

	0 failure	1 success	
0 control			
1 treated			
			1

If the row and column totals are ignored, there are three degrees of freedom for variation of the entries of the table: $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ belong to the unit simplex in $\mathbb{R}^4$.

Degrees of freedom for 2×2 table

	0 failure	1 success	
0 control			$\hat{p}_{\bullet\|0}$
1 treated			$\hat{p}_{\bullet\|1}$
			1

Knowledge of (one of the) row totals leaves 2 degrees of freedom for how the table can be filled in.

Degrees of freedom for 2×2 table

	0 failure	1 success	
0 control			$\hat{p}_{\bullet\|0}$
1 treated			$\hat{p}_{\bullet\|1}$
	$\hat{p}_{0\|\bullet}$	$\hat{p}_{1\|\bullet}$	1

Knowledge of row and column totals leaves 1 degree of freedom for how the table can be filled in.

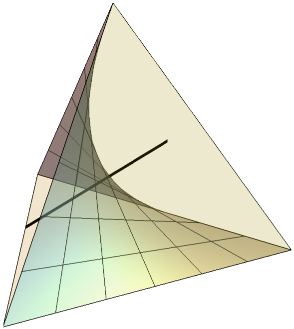
Conditioning in the 2×2 table

	0 failure	1 success	
0 control	$\hat{p}_{0\|0}$	$\hat{p}_{1\|0}$	$\hat{p}_{\bullet\|0}$
1 treated	$\hat{p}_{0\|1}$	$\hat{p}_{1\|1}$	$\hat{p}_{\bullet\|1}$
	$\hat{p}_{0\|\bullet}$	$\hat{p}_{1\|\bullet}$	1

Fisher argued that is it appropriate to condition on row and column totals in the analysis, these being ancillary.

After conditioning, the values of $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ have a distribution constrained to a one-dimensional subspace of the unit simplex.

Geometric exposition of Fisher's conditional analysis



Curved manifold (Feinberg & Gilbert, 1970): the set of true multinomial probabilities consistent with independence of the two binary variables.

Black line (co-ancillary manifold): constraint within the simplex (sample space for the standardised table) imposed by the marginal totals $\hat{p}_{1|\bullet} = 0.6$, $\hat{p}_{\bullet|1} = 0.4$.

Fisher's analysis: based on the distribution of $(\hat{p}_{0|0}, \hat{p}_{1|0}, \hat{p}_{0|1}, \hat{p}_{1|1})$ constrained to the line.

An example with many nuisance parameters (Cox, 1958)

Used by Barndorff-Nielsen (1983) to illustrate the behaviour of modified profile likelihood in an extreme example.

One individual from each of n pairs is randomised to treatment, the other is the untreated control. Pairwise table:

	0 failure	1 success	
0 control			1
1 treated			1
			2

The design fixes the row totals.

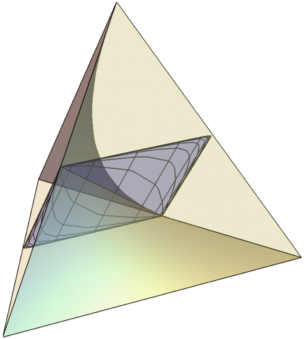
Logistic model for the probabilities

Binary outcomes on n matched pairs. For the ith pair the model is

$$\begin{aligned} p^{(i)}_{1|0} = \text{pr}(\text{success} \mid \text{control}) &= \frac{e^{\alpha_i}}{1+e^{\alpha_i}}, \quad p^{(i)}_{0|0} = 1 - p^{(i)}_{1|0} \\ p^{(i)}_{1|1} = \text{pr}(\text{success} \mid \text{treated}) &= \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}, \quad p^{(i)}_{0|1} = 1 - p^{(i)}_{1|1} \end{aligned}$$

The logistic model is intermediate between a general multinomial representation and one in two independent binomials.

Logistic parametrisation of matched pair problem



Flat plane: subspace compatible with row totals $(\frac{1}{2}, \frac{1}{2})$ from matched pair design.

Curved contours of plane contours of equal β in the logistic parametrisation $(\alpha, \beta) \mapsto e^{\alpha+\beta}/(1 + e^{\alpha+\beta}) = \mathrm{pr}(\text{success}|\text{treated})$.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

$$
\begin{array}{c|cc|}
 & \text{F} & \text{S} \\ \cline{2-3}
\text{C} & 1 & 0 \\
\text{T} & 1 & 0 \\ \cline{2-3}
\end{array}
\qquad
\begin{array}{c|cc|}
 & \text{F} & \text{S} \\ \cline{2-3}
\text{C} & 1 & 0 \\
\text{T} & 0 & 1 \\ \cline{2-3}
\end{array}
\qquad
\begin{array}{c|cc|}
 & \text{F} & \text{S} \\ \cline{2-3}
\text{C} & 0 & 1 \\
\text{T} & 1 & 0 \\ \cline{2-3}
\end{array}
\qquad
\begin{array}{c|cc|}
 & \text{F} & \text{S} \\ \cline{2-3}
\text{C} & 0 & 1 \\
\text{T} & 0 & 1 \\ \cline{2-3}
\end{array}
$$

Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables

Because there are pair-specific nuisance parameters, we start by considering n separate pairwise tables. Four possibilities:

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

Number of tables of each type: R^{00}, R^{01}, R^{10}, R^{11}.

Four possible pairwise tables

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

	F	S	
C			1
T			1
	2	0	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	0	2	

In the leftmost and rightmost tables (concordant pairs), conditioning on column totals leaves no degrees of freedom.

Four possible pairwise tables

	F	S
C	1	0
T	1	0

	F	S
C	1	0
T	0	1

	F	S
C	0	1
T	1	0

	F	S
C	0	1
T	0	1

	F	S	
C			1
T			1
	2	0	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	1	1	

	F	S	
C			1
T			1
	0	2	

In the leftmost and rightmost tables (concordant pairs), conditioning on column totals leaves no degrees of freedom.

In the two inner tables (discordant pairs) there remains one degree of freedom after conditioning.

Conditional analysis based on discordant pairs

Conditioning in the pairwise tables leads us to discard concordant pairs.

- R^{01} tables of type $\boxed{\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}}$ contribute $\boxed{\begin{matrix} R^{01} & 0 \\ 0 & R^{01} \end{matrix}}$
- R^{10} tables of type $\boxed{\begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}}$ contribute $\boxed{\begin{matrix} 0 & R^{10} \\ R^{10} & 0 \end{matrix}}$

Discordant pair table:

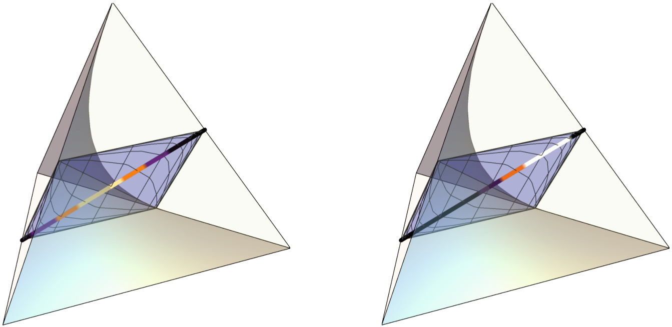
	F	S	
C	R^{01}	R^{10}	m
T	R^{10}	R^{01}	m
	m	m	

Conditional on row and column totals $m = R^{01} + R^{10}$

$$R^{01} \sim \text{Bin}(m, e^{\beta}/(1 + e^{\beta})).$$

Have eliminated all the nuisance parameters $\alpha_1, \ldots, \alpha_n$.

Binomial distribution on the co-ancillary manifold



Induced discrete distributions on the co-ancillary manifold $C(a^o)$ (straight line) corresponding to $\beta = 0$ (left) and $\beta = 2$ (right) from $m = 7$ discordant pairs.

Approximate conditional inference

Two broad situations

Data are realisations of $Y = (Y_1, \ldots, Y_n)$.

Likelihood $L(\theta; y^o) = L(\theta; s^o) = L(\theta; (t^o, a^o))$.

Joint density at arbitrary y:

$$\mathrm{pr}(Y \in [y, y + dy)) = f_Y(y; \theta) dy$$

Two situations

- No reduction from n to $d < n$:
  change variables $y \to (t, a)$.
- Reduction $y \mapsto s(y) \in \mathbb{R}^d$ $d < n$:
  change variables $s \to (t, a)$.

Setting 1: no reduction (1/2)

Typical when θ is a location parameter of a location family.

- t is of dimension d_θ
- a is of dimension $n - d_\theta$.

$$f_Y(y;\theta)dy = f_Y(y(t,a);\theta)|J_{y\to(t,a)}|dt_1\cdots dt_{d_\theta}da_1\cdots da_{n-d_\theta}$$

Condition on $A = a^o$ by dividing by the marginal density of A at a^o. Obtained by integrating out $t_1,\ldots t_{d_\theta}$.

Setting 1: no reduction (2/2)

Fisher (1934): take $T = \hat{\theta}$ and "configuration statistic" A.

$$f_{T|A}(t|a^o;\theta)dt = c(a)\frac{f_Y(y(t,a^o);\theta)}{f_Y(y(t,a^o);t)}dt$$

This is a special case of p^* in a more explicit notation.

OB-N notation:

$$\begin{aligned} p^*(\hat{\theta};\theta|a^o) &= c|\hat{j}|^{1/2}e^{\bar{\ell}} \\ \bar{\ell} &= \ell(\theta;\hat{\theta},a^o) - \ell(\hat{\theta};\hat{\theta},a^o) \end{aligned}$$

Setting 2: reduction by sufficiency (1/2)

Simplest example: canonical exponential family

$$f_Y(y;\theta)dy = \exp\left\{\sum_{i=1}^{n} s(y_i)^{\mathrm{T}}\theta - n\kappa(\theta)\right\}\prod_{i=1}^{n} f_0(y_i)dy_i$$

Minimal sufficient statistic $s(y) = \sum_i s(y_i)$. No ancillary statistic.

Density of sum approximated by inversion of characteristic function.

Duality between canonical parameter space $\Theta \subset \mathbb{R}^d$ and parameter space $\mathscr{S} \subset \mathbb{R}^d$ for $\mathbb{E}S$ (also sample space for S) is the bridge between S and $\hat{\theta}$ needed to specify the Jacobian.

Setting 2: reduction by sufficiency (2/2)

Curved exponential family. Single observation case:

$$f_Y(y;\psi)dy = \exp\left\{s(y)^{\mathrm{T}}\theta(\psi) - \kappa(\theta(\psi))\right\}\prod_{i=1}^{n} f_0(y_i)dy_i$$

Dimension of ψ is smaller than that of s ($d_\psi < d$).

Ancillary complement A of dimension $d_A = d - d_\psi$.

Varying $\psi \in \Psi$ defines a d_ψ-dimensional differentiable manifold Θ_Ψ in Θ. Duality between Θ and $\mathscr{S}$ means that to Θ_Ψ there corresponds a differentiable manifold $\mathscr{S}_\Psi$ in $\mathscr{S}$.

Duality is the bridge between S and $\hat{\theta}$ needed for Jacobian.

Duality

Barndorff-Nielsen (1978), *Information and Exponential Families*, Ch. 9.

Barndorff-Nielsen (1980). Conditionality resolutions. *Biometrika*, 67, 293–310.

Barndorff-Nielsen and Cox (1994), pp. 66–70. Esp. Fig. 2.1.

Validity of p^* in general models

This is more difficult to ascertain and was proved by:

Skovgaard (1990). On the density of minimum contrast estimators. *Biometrika*, 18, 779–789.

Construction of A

Volume 67, issue 2 of *Biometrika*, especially Barndorff-Nielsen (1980). Conditionality resolutions. *Biometrika*, 67, 293–310.

Skovgaard (1990). On the density of minimum contrast estimators. *Biometrika*, 18, 779–789 (especially pp. 787–788).

Barndorff-Nielsen and Cox (1994). *Inference and Asymptotics*, pp. 226–235.

Barndorff-Nielsen and Wood (1998). On large deviations and choice of ancillary for p^* and r^*. *Bernoulli*, 4, 35–63.

A version of p^* for nuisance parameters

The most interesting examples have $\theta = (\psi, \lambda)$, where λ is a nuisance parameter.

A version of p^* for nuisance parameters leads to modified profile likelihood and higher-order inference based on r^*.

Acknowledgements

Thanks are due to Nancy Reid for guiding me through some of the literature on p^* and r^*.

Many omissions.