Biometrika (2002), **89,** 4, *pp.* 933–938

Printed in Great Britain

Miscellanea

Saddlepoint approximations as smoothers

BY A. C. DAVISON

Institute of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland

anthony.davison@epfl.ch

AND SUOJIN WANG

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.

sjwang@stat.tamu.edu

SUMMARY

This note investigates the sense in which saddlepoint approximations act as smoothers of discrete distributions. The discrete problem is embedded in a continuous model that closely matches it on the discrete sample space, with saddlepoint approximation yielding an inference that is almost exact for the continuous model. The same applies to conditional distributions. An example is given and implications for inference are discussed.

Some key words: Conditional inference; Continuity correction; Exponential family; Hauck–Donner effect; Logistic regression; Saddlepoint approximation.

1. INTRODUCTION

Consider a logistic regression model in which the jth of n independent binary variables X_j has success probability $\{1+\exp(-v_j^{\mathrm{T}}\lambda - z_j\psi)\}^{-1}$, where v_j is a $p \times 1$ vector of covariates and λ is a vector of unknown nuisance parameters. It is natural to base inference for the scalar parameter ψ of interest on the conditional distribution of $U = \sum z_j X_j$ given $A = \sum v_j X_j$, which is free of λ. The main obstacle to doing so is enumeration of the conditional sample space, which sometimes has very few support points. This restricts the conditional significance levels for tests on ψ, resulting in a loss of power that has generated a protracted debate on the merits and demerits of conditional inference, particularly for the 2×2 table (Yates, 1984; Agresti, 1992). Moreover the inferences can be unstable, as a slight change to the observed value of A can have a large effect on the conditional significance levels available. This is irritating in practice and objectionable in principle: why should a well-defined and apparently sensible procedure behave thus? Saddlepoint approximation to the conditional distribution of U given A gives more stable inferences, suggesting that some form of smoothing is implicitly being performed. Indeed, Pierce & Peters (1999) argue that use of averaged exact conditional significance levels is not only appropriate in discrete models but is essentially performed by using procedures based on saddlepoint approximations but devised for continuous problems.

The purpose of this paper is to investigate the sense in which the saddlepoint approximation without adjustment for discreteness acts as a smoother of discrete probability models. Our conclusion is that the probability mass function for a discrete random variable T, possibly multivariate, is approximated by the density of a continuous random variable whose support is the interior of

the convex hull of the support of T; often this is infinite. The mass function and density differ at the support points of T, but with relative error $O(n^{-1})$, where n is the sample size. The same is true when saddlepoint approximation is applied directly to a conditional distribution, with inference based on a continuous approximation to the exact conditional density. Thus one may think of the discrete problem as being embedded in a continuous model that closely matches it on the discrete sample space, with saddlepoint approximation yielding an inference that is almost exact for the continuous model.

Conditional densities associated with the continuous model do not have a limited set of significance levels and they are not unstable with respect to changes in the conditioning event. Hence the continuous model does not share the drawbacks associated with conditioning in the discrete case. On the other hand continuity correction is needed for accurate approximation of the underlying discrete distribution.

Section 2 contains our argument, with discussion in § 3 and an example in § 4.

2. Saddlepoint approximation

Let $g_n(w)$ be the probability mass function of a random variable W_n taking values on a lattice with step c_n. Assume that W_n is standardised so that $c_n^{-1}g_n(w)$ has a nondegenerate limit density as $n\to\infty$ and $c_n\to 0$; note that the ratio $c_n^{-1}g_n(w)$ could be considered to be the derivative of the distribution function of W_n at w. The simplest example and the case we consider below is when W_n is the standardised average $n^{1/2}(\bar{X}-\mu)$ of a random sample of integer random variables $X_1,\ldots,X_n$ with mean μ. Then $E(W_n)=0$ and $c_n=n^{-1/2}$.

Denote the cumulant generating function of X_i by $K(t)$, assumed to exist in an open interval containing the origin. Then the probability mass function of W_n at a support value w has saddlepoint approximation (Daniels, 1954)

$$h_n(w)=\{2\pi nK''(t_w)\}^{-1/2}e^{nK(t_w)-n^{1/2}wt_w}, \tag{1}$$

where t_w is the unique solution to $K'(t)=n^{-1/2}w$. Typically this nonnegative function extends smoothly to all values of w inside the support of W_n, because (1) depends smoothly on the cumulant generating function $K(t)$ and its derivatives, all of which are very well behaved for models of practical interest.

The cumulative distribution function $G_n(w)$ of W_n has a Lugannani–Rice-type saddlepoint approximation (Daniels, 1987)

$$H_n(w)=\begin{cases}\Phi(r_{w_+})+\phi(r_{w_+})(r_{w_+}^{-1}-s_{w_+}^{-1}), & \text{if } w_+\neq 0,\\ \frac{1}{2}+\frac{1}{6}(2\pi n)^{-1/2}K^{(3)}(0)/\{K''(0)\}^{3/2}-\frac{1}{2}\{2\pi nK''(0)\}^{-1/2}, & \text{if } w_+=0,\end{cases} \tag{2}$$

where ϕ and Φ are the standard normal density and distribution functions and

$$w_+=w+c_n,\quad r_w=\mathrm{sgn}(t_w)[2n\{n^{-1/2}t_ww-K(t_w)\}]^{1/2},\quad s_w=(1-e^{-t_w})\{nK''(t_w)\}^{1/2}.$$

Approximations (1) and (2) have relative error $O(n^{-1})$ if w lies in the support of W_n.

If we pretend that the X_i are continuous, we can compute the continuous version of the saddlepoint approximation for the distribution function of W_n, that is

$$\tilde{H}_n(w)=\begin{cases}\Phi(r_w)+\phi(r_2)(r_w^{-1}-\tilde{s}_w^{-1}), & \text{if } w\neq 0,\\ \frac{1}{2}+\frac{1}{6}(2\pi n)^{-1/2}K^{(3)}(0)/\{K''(0)\}^{3/2}, & \text{if } w=0,\end{cases}$$

where $\tilde{s}_w=t_w\{nK''(t_w)\}^{1/2}$. In the continuous case w can take any value inside the convex hull of the support of W_n. In fact $\tilde{H}_n(w)$ differs from $G_n(w)$ by a relative error of $O(n^{-1/2})$ in a normal deviation region. We discuss below the use of $\tilde{H}_n(w)$ in discrete data problems and particularly for approximate conditional inference.

We write the distribution function of W_n as

$$G_n(w)=\sum_{i=a}^{b}g_n(w_i),$$

where $\{w_i\}$ contains the support of W_n in increasing order as i runs from a to ∞, a may be finite or $-\infty$, and $w_b = w$. By (1), $G_n(w)$ may be approximated by

$$\hat{G}_n(w) = \sum_{i=a}^{b} h_n(w_i) = c_n \sum_{i=a}^{b} f_n(w_i) \tag{3}$$

with relative error $O(n^{-1})$, where $f_n(w) = c_n^{-1} h_n(w)$ is a smooth function on the interior of the convex hull of the support of W_n with a nondegenerate normal density as its limit as $n \to \infty$.

We apply the Euler–Maclaurin formula (Barndorff-Nielsen & Cox, 1989, pp. 68–71) to the rightmost summation in (3), giving

$$\begin{aligned}
\hat{G}_n(w) &= c_n \left\{ c_n^{-1} \int_{w_a}^{w} f_n(u)\, du + \tfrac{1}{2} f_n(w) + O(c_n) \right\} \\
&= \int_{w_a}^{w+c_n/2} f_n(u)\, du - \int_{w}^{w+c_n/2} f_n(u)\, du + \tfrac{1}{2} c_n f_n(w) + O(c_n^2) \\
&= \int_{w_a}^{w_c} f_n(u)\, du + O(n^{-1}),
\end{aligned} \tag{4}$$

where $w_c = w + c_n/2$ and here we have taken $c_n = n^{-1/2}$. The first equality uses the fact that, as its limit is a normal density, $f_n(w_a)$ goes to zero in exponential order since w_a is either $-\infty$ or converges to $-\infty$ at rate $n^{1/2}$. Thus $f_n(w_a) = O(n^{-1/2})$. Equations (3) and (4) indicate that, apart from an $O(n^{-1})$ error, the discrete distribution function of W_n may be approximated at its support points by the smooth function $\hat{G}_n(w)$, which behaves like a distribution function. Moreover, the integral in (4) can be further approximated with a relative error of $O(n^{-1})$ by the r^* formula (Barndorff-Nielsen & Cox, 1994, p. 211)

$$\tilde{G}_n(w_c) = \Phi\{r^*(w_c)\}, \tag{5}$$

where $r^*(w) = r_w + r_w^{-1} \log(\tilde{s}_w / r_w)$ with $\tilde{s}_w$ and r_w given after (2). Hence $\tilde{G}_n(w_c)$ approximates $G_n(w)$ with a relative error of $O(n^{-1})$ in a normal deviation region. In the large deviation situation with $w = O(n^{1/2})$, the error rate is not relative, because of (4), but the relative error is generally bounded; it is $O(1)$. This is readily seen in the application of the Euler–Maclaurin formula with $f_n(w)$ equal to the standard normal density function, for which $f_n'(w) = -w f_n(w)$, $f_n''(w) = (w^2 - 1) f_n(w)$ so that the absolute error is

$$n^{-1/2} O\{n^{-1/2} f_n'(w) + n^{-1} f_n''(w)\} = O\{n^{-1/2} f_n(w)\},$$

which is the same order as $G_n(w)$ when $w = O(n^{1/2}) \to -\infty$; a similar argument applies to the upper tail probability. Furthermore, for a moderately large deviation region with $w = O(n^p)$ for $0 < p < \frac{1}{2}$, the relative error is $O(n^{2p-1})$.

An alternative to the development above leading to (5) starts with the discrete version (2) of the Lugannani–Rice formula. In a normal deviation region with $w = O(1)$, $1 - e^{-t_w} = t_w - t_w^2/2 + O(t_w^3)$. Moreover, provided $w_+ \neq 0$,

$$\begin{aligned}
\tilde{G}_n(w_+) &= \Phi(r_{w_+}) + \phi(r_{w_+})(r_{w_+}^{-1} - \tilde{s}_{w_+}^{-1}) + O(n^{-1}) \\
&= H_n(w) + \phi(r_{w_+})(s_{w_+}^{-1} - \tilde{s}_{w_+}^{-1}) + O(n^{-1}) \\
&= H_n(w) + \tfrac{1}{2}\{nK''(t_w)\}^{-1/2} \phi(r_{w_+}) + O(n^{-1}).
\end{aligned} \tag{6}$$

On the other hand, using the facts that

$$w_+ = w_c + \tfrac{1}{2} n^{-1/2}, \quad dr_w/dw = \{K''(t_w)\}^{-1/2} + O(n^{-1/2}),$$

we have

$$\tilde{G}_n(w_+) = \Phi\{r^*(w_c)\} + \tfrac{1}{2}\{nK''(t_w)\}^{-1/2} \phi(r_{w_+}) + O(n^{-1}).$$

This and (6) give the desired result (5). Detailed derivations may be obtained from the authors.

3. Comments

Several comments spring to mind. First, we have shown that (5) approximates the distribution function $G_n(w)$ of the discrete W_n at each of its support points with error $O(n^{-1})$, but is continuous inside the convex hull of the support. In fact this involves two approximations, the first arising when the mass function $g_n(w)$ is replaced by the saddlepoint quantity (1) at the points $w = w_i$, and the second occurring when the integral of (1) over w is approximated by (5). For a related application of saddlepoint smoothing in bootstrap distributions, see Wang (1990).

Secondly, continuity correction is needed only if the original discrete model is regarded as the target to which approximation is sought. The usual use of $r^*(w)$ as a continuous approximation to a discrete problem is first-order correct, having error $O(n^{-1/2})$. If a continuity correction $\frac{1}{2}c_n$ is added to w, then $r^*(w_c)$ and (5) give a second-order correct approximation for the discrete distribution, with error $O(n^{-1})$. One could argue that it is more appropriate to regard the continuous distribution arising from integrating (1) as the baseline. One might assert that this continuous version of the problem is more nearly ideal, in the sense that there is no theoretical restriction on the achievable significance levels, although in practice only a subset of them can be observed because of the discreteness of W_n; this embedding plays a more central role in the conditional situations discussed below. This discussion reinforces the recommendation of Pierce & Peters (1999) that continuity correction not be used, but from the viewpoint that one should not attempt to reproduce results from a discrete model in which restricted significance levels depend heavily on a conditioning event.

Thirdly, our argument easily extends to conditional distributions, simply by taking $G_n(w)$ to be the conditional distribution of a discrete variable. If a saddlepoint approximation is available for the corresponding conditional density, then the orders of error in (1) and (2) continue to hold, and our argument goes through, with (5) a smooth approximation to the discrete conditional distribution. In many exponential family models, however, the resulting approximation arises from or is equivalent to a double saddlepoint approximation, and it is then the result of a smoothing on the original unconditional sample space. Double saddlepoint approximation to the conditional mass function of a discrete random variable W_n conditional on another discrete variable A_n uses the ratio $h_n(w, a)/h_n(a)$, where the joint and marginal density approximations $h_n(w, a)$ and $h_n(a)$ have formulae similar to (1) and arise from essentially the same argument. Now $h_n(w, a)$ differs from the exact joint mass function of (W_n, A_n) by $O(n^{-1})$ at its support points, but is defined for any (w, a) inside the convex hull of the support of (W_n, A_n). Division by $h_n(a)$ renormalises $h_n(w, a)$ to have integral $1 + O(n^{-1})$ over w for each fixed a within the support of A_n. Thus the smoothing takes place before conditioning in the sense that the joint mass function of (W_n, A_n) is replaced by a smooth density-like function defined on the interior of the convex hull of the support of (W_n, A_n). Rescaled slices of this smooth function provide density approximations for W_n conditional on A_n, and integrals of those rescaled slices provide distribution function approximations. Thus here smoothing takes place before conditioning.

By contrast, when the saddlepoint approximation is applied directly to a conditional distribution, so that conditioning takes place before smoothing, the argument in § 2 is in effect applied directly to the conditional distribution. In this case saddlepoint approximation need not produce tail probabilities that are stable across values of A_n, because the conditional distributions obtained for adjacent values of A_n are not necessarily constrained to be related.

4. Example

We illustrate our argument by a constructed example in which independent binary variables $X_1, \ldots, X_9$ depend on a covariate z taking values 0, 1, 2, 3, 5, 7, 11, 13 and 17 respectively through a logistic regression with linear predictor $\lambda + \psi z$. As mentioned in § 1, a standard argument leads us to consider the conditional distribution of $U = \sum z_j X_j$ given the sum $A = \sum X_j$ of the responses; here the nuisance parameter λ is scalar. Figure 1 shows these conditional distributions for $a = 3$ and $a = 4$ when $\psi = 0$, together with their saddlepoint approximations $\Phi\{r^*(u)\}$ from (5) evaluated

for continuous u. They differ from the distributions of the corresponding W_n only because U is a weighted sum of X_j and the horizontal axis has not been centred at zero and rescaled. The approximations appear excellent, but close inspection suggests that the approximation is not always improved by continuity correction; see the inset, which shows the right-most functions around $u = 30$.

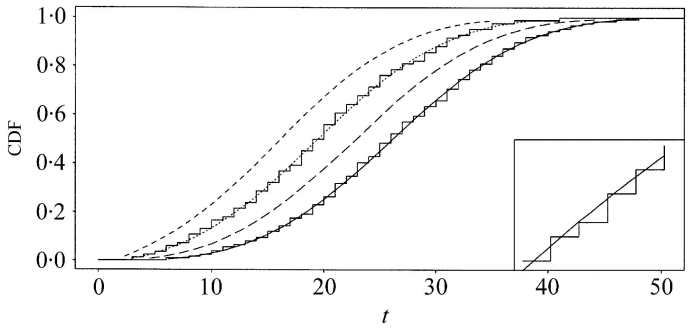

Fig. 1. Exact and appropriate conditional distributions, CDF's, of U given $A = a = 3$ and 4. The exact conditional cumulative distribution functions are step functions, and the approximations for $a = 3$ and $a = 4$ are respectively dotted and solid. Shown also are continuous distribution functions when $a = 2{\cdot}5$ (dashes) and $a = 3{\cdot}5$ (large dashes). The inset shows a close-up of the cumulative distribution function for $a = 4$ and its approximation around $u = 30$.

The two dashed smooth curves standing alone show the continuous saddlepoint approximation to the conditional distribution when $a = 2{\cdot}5$ and $a = 3{\cdot}5$; of course in this case the discrete distributions do not exist. The continuous curve is readily constructed, however, by creating a set of 'binary' responses summing to a and for which u has whatever value is desired; quantities such as $r^*(u)$ are then straightforwardly obtained using output from a routine such as `glm` in S-Plus (Davison, 1988; Brazzale, 1999). When $a = 3{\cdot}5$, for instance, we take

$$z = \delta(1, 1, 1, 0{\cdot}5, 0, 0, 0, 0, 0)^{\mathrm{T}} + (1 - \delta)(0, 0, 0, 0, 0, 0{\cdot}5, 1, 1, 1)^{\mathrm{T}},$$

and allow δ to vary in the range $(0, 1)$. Then $a = \sum x_j = 3{\cdot}5$, while any value of $U = \sum z_j x_j$ in the interval $(4{\cdot}5, 44{\cdot}5)$ can be produced by an appropriate choice of δ. Saddlepoint approximation then yields the 'conditional distribution' corresponding to this a. Such curves could be produced for any a in the interval $(0, 9)$, thereby giving a smooth joint density for (U, A). There is a heuristic analogy with quasilikelihood models: one can imagine the model with $a = 3{\cdot}5$ as corresponding to 90 notional observations in which 35 responses were positive, these observations however being so overdispersed that they correspond to a mere nine 'ordinary' binary responses.

A numerical problem arises with binary logistic response models, for the following reason. The statistic $r^*(u) = r_u + r_u^{-1} \log(\tilde{s}_u / r_u)$ depends on the signed root r_u of the likelihood ratio statistic and on $\tilde{s}_u$; both have the same sign. The loglikelihood for a logistic model has the property that, as u approaches the limits of its conditional range, $u_-(a)$ and $u_+(a)$, say, $|r_u|$ becomes large, but $|\tilde{s}_u|$ decreases to zero; see Hauck & Donner (1977). The result is that $|r^*(u)|$ decreases as u continuously approaches either limit, so $\Phi\{r^*(u)\}$ is not a strictly increasing function. Fortunately this seems not to affect applications. In the present case, for example, when $a = 4$ we have $u_-(a) = 6$ and $u_+(a) = 48$, and trouble erupts only when u is within about 0·2 of the limits of its range. Continuity-corrected approximations to the limiting probabilities would involve evaluating $\Phi\{r^*(u)\}$ for $u = 6{\cdot}5$ and $47{\cdot}5$, and hence would not fail; Fig. 1 suggests that these approximations would be adequate.

Acknowledgement

We thank a referee for useful comments. This work was supported by the Swiss National Science Foundation, the Texas Advanced Research Program, the U.S. National Cancer Institute and the Texas A&M University Center for Environmental and Rural Health.

References

Agresti, A. (1992). A survey of exact inference for contingency tables (with Discussion). *Statist. Sci.* **7**, 131–77.
Barndorff-Nielsen, O. E. & Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics.* London: Chapman and Hall.
Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics.* London: Chapman and Hall.
Brazzale, A. R. (1999). Approximate conditional inference in logistic and loglinear models. *J. Comp. Graph. Statist.* **8**, 653–61.
Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631–50.
Daniels, H. E. (1987). Tail probability approximations. *Int. Statist. Rev.* **54**, 37–48.
Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *J. R. Statist. Soc.* B **50**, 445–61.
Hauck, W. W. & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *J. Am. Statist. Assoc.* **72**, 851–3.
Pierce, D. A. & Peters, D. (1999). Improving on exact tests by approximate conditioning. *Biometrika* **86**, 265–77.
Wang, S. (1990). Saddlepoint approximations in resampling analysis. *Ann. Inst. Statist. Math.* **42**, 115–31.
Yates, F. (1984). Tests of significance for 2×2 contingency tables (with Discussion). *J. R. Statist. Soc.* A **147**, 426–63.

[*Received July* 2000. *Revised December* 2001]