Likelihood

N. Reid, University of Toronto

1 Introduction

In 1997 a study conducted at the University of Toronto concluded that the risk of a traffic accident increased by four-fold when the driver was using a cellular telephone (Redelmeier and Tibshirani, 1997a). The report also stated that such a large increase was very unlikely to be due to chance, or to unmeasured confounding variables, although the latter could not be definitively ruled out. Figure 1a shows the *likelihood function* for the important parameter in the investigators' model, the relative risk of an accident. Figure 1b shows the log of the likelihood function plotted against the log of the relative risk. The likelihood function was the basis for the inference reported.[1] As with most real life studies, there were a number of decisions related first to data collection, and then to modelling the observed data, that involved considerable creativity and a host of small but important decisions relating to details of constructing the appropriate likelihood function. A non-technical account of some of these is given in Redelmeier and Tibshirani (1997c), and a more statistically oriented version is given in Redelmeier and Tibshirani (1997b). In this vignette we are simply using the data to provide an illustration of the likelihood function.

Assume we are considering a parametric model $f(y;\theta)$, which is the probability density function with respect to a suitable measure for a random variable Y. The parameter is assumed to be k-dimensional and the data assumed to be n-dimensional, often representing a sequence of independent and identically distributed random variables: $Y = (Y_1, \dots Y_n)$. The *likelihood*

[1] The point estimate of relative risk from the likelihood function is actually 6.3, although 4.0 was the reported value. The maximum likelihood estimate was downweighted by an ad hoc method devised to accomodate some complexities in the study design.

function is defined to be a function of θ, proportional to the model density:

$$L(\theta) = L(\theta; y) = cf(y; \theta), \tag{1}$$

where c can depend on y but not on θ. Within the context of the given parametric model, the likelihood function measures the relative plausibility of various values of θ, for a given observed data point y. The notation for the likelihood function emphasizes that the parameter θ is the quantity that varies, and that the data value is considered fixed. The constant of proportionality in the definition is needed, for example, to accomodate one-to-one transformations of the random variable Y that do not involve θ, as these clearly should have no effect on our inference about θ. Another way to say the same thing is that the likelihood function is not calibrated in θ, or that only relative values $L(\theta_1)/L(\theta_2)$ are well determined.

The likelihood function was proposed by Fisher (1922) as a means of measuring the relative plausibility of various values of θ by comparing their likelihood ratios. When θ is one- or two-dimensional, the likelihood function can be plotted and provides a visual assessment of the set of likelihood ratios. Several authors, beginning with Fisher, have suggested that ranges of plausible values for θ can be directly determined from the likelihood function, first by determining the *maximum likelihood estimate* $\hat{\theta} = \hat{\theta}(y)$, the value of θ that maximizes $L(\theta; y)$, and then using as a guideline[2]

$L(\hat{\theta})/L(\theta) \in (1, 3)$	quite plausible
$L(\hat{\theta})/L(\theta) \in (3, 10)$	somewhat implausible
$L(\hat{\theta})/L(\theta) \in (10, \infty)$	highly implausible

General introductions to the definition of the likelihood function and its informal use in inference are given in Fisher (1956), Edwards (1972), Kalbfleisch (1985) and Azzalini (1996).

[2]The ranges suggested here are taken from Kass and Raftery (1995), attributed to Jeffreys, and other authors have suggested different cut-off points. For example Fisher (1956, p. 71) suggested using 2, 5 and 15, and Royall (1997) suggested 4, 8 and 32.

2 Likelihood function and inference

2.1 Bayesian inference

While the use of likelihood as a plausibility scale is sometimes of interest, probability statements are usually preferred in applications. The most direct way to obtain these is by combining the likelihood with a *prior probability* function for θ, to obtain a posterior probability function:

$$\pi(\theta|y) \propto \pi(\theta)L(\theta;y) \tag{2}$$

where the constant of proportionality is $\int \pi(\theta)L(\theta;y)d\theta$.

This leads directly to Bayesian inferences of the sort: "Using the prior density $\pi(\theta)$, we conclude that values of θ greater than θ_U have posterior probability less than 0.05 and are hence inconsistent with the model and the prior". Jeffreys (1961) emphasized this use of the likelihood function, and investigated the possiblity of using 'flat' or 'noninformative' priors. He also suggested the plausibility range described above, this time in the context of Bayesian inference with a flat prior.

One difficulty in applying Bayesian inference is in the construction of a suitable prior, and there is recently renewed interest in the construction of 'noninformative' priors, which lead to posterior probability intervals that are in one way or another minimally affected by the prior density. One example of a noninformative prior is one for which the posterior probability limit θ_U described in the previous paragraph does in fact lead to an interval which, when considered as a confidence interval, has (at least approximately) coverage equal to its posterior probability content. If θ is a scalar parameter then the appropriate prior is Jeffreys' prior $\pi(\theta) \propto \{i(\theta)\}^{1/2}$, where $i(\theta)$ is the Fisher information in the model $f(y;\theta)$:

$$i(\theta) = E\left\{\frac{\partial L(\theta;Y)}{\partial\theta}\right\}^2 = \int \left\{\frac{\partial L(\theta;y)}{\partial\theta}\right\}^2 f(y;\theta)dy. \tag{3}$$

This result was derived by Welch and Peers (1963) in response to a question raised in Lindley (1958). Unfortunately there is not a satisfactory general prescription for such a *probability matching prior* when θ is multi-dimensional. Another type of noninformative prior, motivated rather differently, is the *reference prior* of Bernardo (Berger and Bernardo, 1992). Kass and Wasserman (1996) provide an excellent review of noninformative priors.

Another difficulty in applying Bayesian inference with multidimensional parameters, or in more complex situations, is the high dimensional integration needed either to evaluate the normalizing constant in (2), or to compute marginal posterior densities for particular parameters of interest from the multi-dimensional posterior. These difficulties have largely been solved by the introduction of a number of numerical methods, including importance sampling and Markov chain Monte Carlo methods. An introduction to Gibbs sampling is given in Casella and George (1992); see also the vignettes on Gibbs sampling and MCMC methods in this issue.

Bayesian inference respects the so-called *likelihood principle*, which states that inference from an experiment should be based only on the likelihood function for the observed data. Any inference which uses the sampling distribution of the likelihood function, as described in the next subsection, does not obey the likelihood principle. The discovery by Birnbaum (1962) that the principles of sufficiency and conditionality imply the likelihood principle led to considerable discussion in the sixties and seventies on various aspects of the foundations of inference. A good overview can be found in Berger and Wolpert (1984). There has more recently been less interest in these foundational issues.

2.2 Classical inference

Frequentist probability statements can be constructed from the likelihood function by considering the sampling distribution of the likelihood function and derived quantities. In fact this is practically necessary from a frequentist point of view, since the minimal sufficient statistic in any model is determined by the likelihood map $L(\theta; \cdot)$. This is why, for example, the Neyman-Pearson lemma concludes that the most powerful test depends on the likelihood ratio statistic.

The conventional derived quantities for a parametric likelihood function are the score function

$$\ell'(\theta) = \partial \log L(\theta)/\partial\theta, \tag{4}$$

the maximum likelihood estimate

$$\sup_\theta \ell(\theta) = \ell(\hat\theta), \tag{5}$$

and the observed Fisher information

$$j(\hat\theta) = -\left. \partial^2\ell(\theta)/\partial\theta^2 \right|_{\theta=\hat\theta} \tag{6}$$

where $\ell(\theta) = \log L(\theta)$, is the log-likelihood function.

In the case where $Y = (Y_1, \ldots, Y_n)$ is a sample of independent, identically distributed random variables, the log-likelihood function is a sum of n independent, identically distributed quantities, and under some conditions on the model a central limit theorem can be applied to the score function (4). More general sampling, such as the $(Y_1, \ldots, Y_n)$ independent, but not identically distributed, or weakly dependent, can be accomodated if the model satisfies enough regularity conditions to ensure a central limit theorem for a suitably standardized version of the score function. Under many types of sampling the score function is a martingale and the martingale central limit theorem can be applied. Thus for a wide class of models the following results can be derived:

$$\ell'(\theta)^T\{j(\hat{\theta})\}\ell'(\theta) \xrightarrow{d} \chi^2_p \tag{7}$$

$$(\hat{\theta} - \theta)^T\{j(\hat{\theta})\}^{-1}(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_p \tag{8}$$

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_p \tag{9}$$

where χ^2_p is the chi-squared distribution on p degrees of freedom, and p is the dimension of θ.

Similar results are available for inference about component parameters: writing $\theta = (\psi, \lambda)$, and denoting by $\hat{\lambda}_\psi$ the *restricted maximum likelihood estimate* of λ for ψ fixed,

$$\sup_\lambda \ell(\psi, \lambda; y) = \ell(\psi, \hat{\lambda}_\psi; y) = \ell_p(\psi), \tag{10}$$

we have, for example,

$$2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\} \xrightarrow{d} \chi^2_q, \tag{11}$$

where q is the dimension of ψ. The function $\ell_p(\psi)$ defined in (10) is called the *profile log likelihood* function.

These limiting results are taken as the size of the sample, n, in an independent sampling context, increases, with the dimension of θ held fixed. More generally limit statements can be derived for the limit as the amount of Fisher information in Y increases.

The approximations suggested by these limiting results, such as

$$\hat{\theta} \mathrel{\dot\sim} N\{\theta, j(\hat{\theta})\}, \tag{12}$$

called *first order* approximations, are widely used in practice for inference about θ. The development of high speed computers throughout the last half of the 20th century has enabled accurate and fast computation of maximum likelihood estimators in a wide variety of models, and most statistical packages have general purpose routines for calculating derived likelihood quantities. This has meant in particular that development of alternative methods of point and interval estimation derived in the first half of the century are less important for applied work than they once were.

2.3 Likelihood as pivotal

A major development in statistical inference based on the likelihood function of the past twenty years is the discovery that the likelihood function can be used directly to provide an approximate sampling distribution for derived quantities that is more accurate than approximations like (12). The main result, usually called Barndorff-Nielsen's approximation, was initially developed in a series of papers in the August 1980 issue of *Biometrika*[3], all of which derived in one version or another that

$$f(\hat{\theta};\theta|a) \doteq c|j(\hat{\theta})|^{1/2}\exp\{\ell(\hat{\theta})-\ell(\theta)\}; \tag{13}$$

the right hand side is often called Barndorff-Nielsen's p^* approximation. This formula generalizes an exact result for location models due to Fisher (1934). The renormalizing constant c is equal to $(2\pi)^{-p/2}\{1+O(n^{-1})\}$. In some generality (13) is a *third order* approximation, meaning the ratio of the right hand side to the true sampling density of $\hat{\theta}$ (given a) is $1+O(n^{-3/2})$. In spite of its importance, a rigorous proof of (13) is not yet available, although Skovgaard (1990) gives a very careful and helpful derivation. It is necessary to condition on a statistic a in order that (13) be meaningful, because the likelihood function appearing on the right hand side depends on the data y, yet it is being used as the sampling distribution for $\hat{\theta}$. The role of a is to complete a one to one transformation from y to $(\hat{\theta}, a)$. In order for (13) to be useful for inference it is necessary for a to have a distribution either exactly or approximately free of θ; otherwise we have lost information about θ in reducing to the conditional model.

The importance of (13) for the theory of inference is that it shows that the distribution of the maximum likelihood estimator (and other derived

[3]Barndorff-Nielsen, Cox, Durbin, Hinkley (1980)

quantities) is obtained to a very high order of approximation directly from the likelihood function, as it is in a location model.

A result related to (13) and more directly useful for inference is the approximation of the cumulative distribution function for $\hat{\theta}$. In the case that θ is a scalar this is expressed as

$$\Pr(\hat{\Theta} \leq \hat{\theta}; \theta | a) = F(\hat{\theta}; \theta | a) \doteq \Phi(r) + \phi(r) \left(\frac{1}{r} - \frac{1}{q} \right) \tag{14}$$

where

$$r = \operatorname{sign}(q) \sqrt{[2\{\ell(\hat{\theta}) - \ell(\theta)\}]} \tag{15}$$

$$q = \{\ell_{;\hat{\theta}}(\hat{\theta}) - \ell_{;\hat{\theta}}(\theta)\}\{j(\hat{\theta})\}^{-1/2}, \tag{16}$$

where $\ell_{;\hat{\theta}}(\theta) = \partial \ell(\theta; \hat{\theta}, a)/\partial \hat{\theta}$. As with (13) this is an approximation with relative error $O(n^{-3/2})$. Two advantages of (14) over (13) are that it gives tail areas or p-values directly, and that it depends on a rather weakly, through a first derivative on the sample space. Approximation (14) shows that the first order approximation to the likelihood ratio statistic (the scalar parameter version of (9)), provides the leading term in an asymptotic expansion to its distribution, that the next term in the expansion is easily computed directly from the likelihood function, and that in frequentist based inference, the sample space derivative of the log likelihood function plays an essential role. This last result has the potential to clarify (and also narrow) the difference between frequentist and Bayesian inference. Approximation (14) is often called the Lugannani and Rice approximation, as a version for exponential families was first developed in Lugannani and Rice (1980). There are analogous versions of (14), (15) and (16) for inference about a scalar component of θ in the presence of a nuisance parameter: a partial review is given in Reid (1996), and more recent work is presented in Skovgaard (1996), Barndorff-Nielsen and Wood (1998) and Fraser, Reid and Wu (1999).

3 Partial likelihood and all that

3.1 Nuisance parameters

We defined at (10) the profile log likelihood function $\ell_p(\psi)$ which is often used in problems in which the parameter of the model, θ is partitioned into a parameter of interest ψ and a nuisance parameter λ. Typically λ is introducted into the model to make it more realistic. More generally we can define

$$\ell_p(\psi) = \sup_{\psi=\psi(\theta)} \ell(\theta). \tag{17}$$

The profile likelihood is not a real likelihood function, in that it is not proportional to the sampling distribution of an observable quantity. However, there are limiting results analogous to (7) through (9), such as (11), that continue to provide first order approximations. These approximations are expected to be poor if the dimension of the nuisance parameter λ is large relative to n, as it is known that the results break down if the dimension of θ increases with n. More intuitively, since no adjustment is made for errors of estimation of the nuisance parameter in (17), it is likely that the apparent precision of (17) is overstated.

Several methods have been suggested for constructing a likelihood function that is better suited to problems with nuisance parameters. In some models there may exist a conditional or marginal distribution that contains all the information about the parameter of interest, or is at least free of the nuisance parameter, and this density provides a true conditional or marginal likelihood. In fact Figure 1 is a plot of the conditional likelihood of a component of the minimal sufficient statistic for the model, this likelihood depending only on the relative risk of an accident and not on nuisance parameters describing the background risk. More precisely, in that model we have the factorization

$$f(y; \psi, \lambda) \propto f(s|t; \psi) f(t; \psi, \lambda) \tag{18}$$

and Figure 1 shows $L_c(\psi) \propto f(s|t;\psi)$. The justification for ignoring the term $f(t;\psi,\lambda)$ is not entirely clear, and not entirely agreed upon, although the claim is usually made that this component contains "little" information about ψ in the absence of knowledge of λ. A review of some of this work is given in Reid (1995).

In models where a conditional or marginal likelihood is not available, a natural alternative is a suitably defined approximate conditional or marginal likelihood, and approximation (13) has led to several suggestions for *modified profile likelihoods.* These typically have the form

$$\ell_m(\psi) = \ell_p(\psi) - \frac{1}{2}|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| + B(\psi) \tag{19}$$

for some choice of $B(\cdot)$ which is of the same asymptotic order as the second term in (18), typically $O_p(1)$. The original modified profile likelihood is due to Barndorff-Nielsen (1983); Cox and Reid (1987) suggested the use of (18) with $B(\psi) = 0$, and several other versions have been proposed. A brief overview is given in Mukerjee and Reid (1999); see also Severini (1998).

3.2 Partial likelihood

In more complex models there is often a partition analogous to (18), say

$$L(\theta; y) = L_1(\psi; y) L_2(\psi, \lambda; y) \tag{20}$$

where it seems intuitively obvious that the second component cannot provide information about ψ in the absence of knowledge of λ. The most famous model for which this is the case is Cox's proportional hazards model for failure time data, where L_1 depends on the observed failure times, and L_2 depends on the failure process between observed failure times. Cox (1972) proposed basing inference about the parameters of interest on L_1, which he called a conditional likelihood, later changed to partial likelihood (Cox, 1975). Cox (1972) showed that a martingale central limit theorem could be applied to the score statistic computed from L_1, leading to asymptotic normality for derived quantities such as the partial maximum likelihood estimate.

There have since been a number of related models where a partial likelihood leads to an adequate first order approximation (Andersen et al., 1993; Murphy and van der Vaart, 1997). There is not yet a theory of higher order approximations in this setting. Likelihood partitions such as (20) are discussed in some generality in Cox (1999).

3.3 Pseudo-likelihood

One interpretation of partial likelihood is that the probability distribution of only part of the observed data is modelled, as this makes the problem tractable and with luck provides an adequate first order approximation. A similar construction was suggested for complex spatial models by Besag (1977), using the conditional distribution of the nearest neighbours of any given point, and using the product of these conditional distributions as a pseudo-likelihood function. A more direct approach to likelihood inference in spatial point processes is described in Geyer (1999).

3.4 Quasi-likelihood

The last thirty years have also seen the development of an approach to modelling that does not specify a full probability distribution for the data, but instead specifies the form of, for example, the mean and the variance of each observation. This is the point of view emphasized in the development of generalized linear models (McCullagh and Nelder, 1989), and is central to the theory of generalized estimating equations (Diggle, Liang and Zeger, 1994). A *quasi-likelihood* is a function that is compatible with the specified mean and variance relations. Although it may not exist, when it does it does it has in fairly wide generality the same asymptotic distribution theory as a likelihood function (McCullagh, 1983; Li and McCullagh, 1994).

3.5 Likelihood and nonparametric models

Suppose we have a model in which we assume $Y_1, \dots Y_n$ are independent and identically disributed from a completely unknown distribution function $F(\cdot)$. The natural estimate of $F(\cdot)$ is the empirical distribution function:

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i \leq y\}. \tag{21}$$

Although it is not immediately clear what the likelihood function or likelihood ratio is in a nonparametric setting, for a suitably defined likelihood $F_n(\cdot)$ is the maximum likelihood estimator of $F(\cdot)$. This is generalized to much more complex sampling, including censoring, in Andersen et al. (1993).

The empirical distribution function plays a central role in two inferential techniques closely connected to likelihood inference: the bootstrap and empirical likelihood. The nonparametric bootstrap uses samples from F_n for constructing an inference, usually by Monte Carlo resampling. The parametric bootstrap uses samples from $F(\cdot; \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator. There is a close connection between the parametric bootstrap and the asymptotic theory of Section 2.3, although the precise relationship is still elusive. A good review is given in DiCiccio and Efron (1996).

An alternative to the nonparametric bootstrap is the empirical likelihood function, which is a particular type of profile likelihood function for a parameter of interest, treating the distribution of the data otherwise as the nuisance 'parameter'. The empirical likelihood was developed by Owen (1988), and has been shown to have an asymptotic theory similar to that for parametric likelihoods.

Empirical likelihood and likelihoods related to the bootstrap are described in Efron and Tibshirani (1993).

4 Conclusion

From either a Bayesian or frequentist perspective, the likelihood function plays an essential role in inference. The maximum likelihood estimator, once regarded on an equal footing among competing point estimators, is now typically the basis for most inference and subsequent point estimation, although some refinement is needed in problems with large numbers of nuisance parameters. The likelihood ratio statistic is the basis for most tests of hypotheses and interval estimates. The emergence of the centrality of the likelihood function for inference, partly due to the large increase in computing power, is one of the central developments in the theory of statistics during the latter half of the twentieth century.

References

Andersen, P.K., Borgan, O. Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Azzalini, A. (1998). *Statistical Inference.* Chapman and Hall, London.

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

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* Chapman and Hall, London.

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

Berger, J.O. and Bernardo, J. (1992). On the development of reference priors (with discussion). *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds, *Bayesian Statistics IV*, Oxford University Press, 35–60.

Berger, J.O. and Wolpert, R. (1980). *The Likelihood Principle.* Institute of Mathematical Statistics, Hayward.

Besag, J. (1977). Efficiency of pseudo-likelihood estimation for simple Gaussian fields. *Biometrika* **64**, 616–618.

Birnbaum, A. (1962). On the foundations of statistical inference. *Amer. Statist. Assoc.* **57**, 269-306.

Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46**, 167–174.

Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D.R. (1980). Local ancillarity. *Biometrika* **67**, 279–286.

Cox, D.R. (1999). Some remarks on likelihood factorization. preprint.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with dicussion). *J.R. Statist. Soc.* B **49**, 1-39.

DiCiccio, T.J. and Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statist. Science* **11**, 189–228.

Diggle P.J., Liang K.Y., and Zeger S.L. (1994). *The Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Durbin, J. (1980). Approximations for densities of sufficient statistics. *Biometrika* **67**, 311–333.

Edwards, A.F. (1960). *Likelihood.* Oxford University Press, Oxford.

Efron, B. and Tibshirani, R.J. (1996). *An Introduction to the Bootstrap.* Chapman and Hall, London.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.* A **222**, 309–368.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc.* A **144**, 285–307.

Fisher, R.A. (1956). *Statsitical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.

Fraser, D.A.S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264

Geyer, C. (1999). Likelihood inference for spatial point processes. in *Stochastic Geometry: Likelihood and Computation*, O.E. Barndorff-Nielsen, W.S. Kendall, and M.N.M. van Lieshout, eds, Chapman and Hall/CRC, London, 79–140.

Hinkley, D.V. (1980). Likelihood as approximate pivotal. *Biometrika* **67**, 287–292.

Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press, Oxford.

Kalbfleisch, J.G. (1985). *Probability and Statistics, Vol. II.* 2nd edition. Springer-Verlag, New York.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795.

Kass, R.E and Wasserman, L. (1996). Formal rules for selecting prior distributions: a review and annotated bibliography. *J. Am. Statist. Assoc.* **91**, 1343–70.

Li, B. and McCullagh, P. (1994). Potential functions and conservative estimating functions. *Ann. Statist.* **22**, 340–356.

Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem. *J. R. Statist. Soc.* B **20**, 102–107.

Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.* **12**, 475–490.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* 2nd ed. Chapman and Hall, London.

Mukerjee, R. and Reid, N. (1999). On confidence intervals associated with the usual and adjusted likelihoods. *J. R. Statist. Soc.* B, to appear.

Murphy, S.A. and van der Vaart, A. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.* **25**, 1471–1509.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Redelmeier, D. and Tibshirani, R.J. (1997a). Association between cellular phones and car collisions. *New Engl. J. Med.* Feb. 12, 1997, 1–5.

Redelmeier, D. and Tibshirani, R.J. (1997b). Is using a cell phone like driving drunk? *Chance* **10**, 5–9.

Redelmeier, D. and Tibshirani, R.J. (1997c). Cellular telephones and motor vehicle collisions: somne variations on matched pairs analysis. *Canad. J. Statist.* **25**, 581–593.

Reid, N. (1995). The roles of conditioning in inference. *Statist. Science* **10**, 138–157.

Reid, N. (1996). Likelihood and higher-order approximations to tail areas: A review and annotated bibliography. *Canad. J. Statist.* **24**, 141-66.

Royall, R. (1997). *Statistical Evidence.* Chapman and Hall, London.

Severini, T.A. (1998). An approximation to the modified profile likelihood function. *Biometrika* **82**, 1–23.

Skovgaard, I.M. (1990). On the density of minimum contrast estimators. *Ann. Statist.* **18**, 779–789.

Skovgaard, I.M. (1996). An explicit large-deviation approximation to one parameter tests. *Bernoulli* **2**, 145-65.

Welch, and Peers, (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc.* B **25**, 318–329.

Figure 1: Likelihood and log-likelihood function for the relative risk; based on Redelmeier and Tibshirani (1997c)