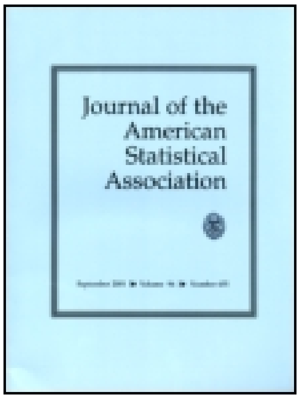

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Likelihood

N. Reid [a]

[a] Department of Statistics , University of Toronto , Ontario , Canada , M5S 3G3

Published online: 17 Feb 2012.

To cite this article: N. Reid (2000) Likelihood, Journal of the American Statistical Association, 95:452, 1335-1340

To link to this article: http://dx.doi.org/10.1080/01621459.2000.10474343

——— (1985), "Asymptotic Behavior of M-Estimators of p Regression Parameters When p^2/n is Large; II. Normal Approximation," *The Annals of Statistics*, 13, 1403–1417.

Portnoy, S., and Koenker, R. (1997), "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error and Absolute-Error Estimation," *Statistical Science*, 12, 279–300.

Portnoy, S., and Welsh, A. (1992), "Exactly What is Being Modelled by the Systematic Component in a Heteroscedastic Linear Regression," *Statistical Probability of Letters*, 13, 253–258.

Richardson, A. M., and Welsh, A. H. (1995), "Robust Restricted Maximum Likelihood in Mixed Linear Models," *Biometrics*, 51, 1429–1439.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J., and Hubert, M. (1999), "Regression Depth" (with discussion), *Journal of the American Statistical Association*, 94, 388–433.

Rousseeuw, P. J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Ruppert, D., and Carroll, R. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828–838.

Simpson, D. G. (1987), "Minimum Hellinger Distance Estimation for the Analysis of Count Data," *Journal of the American Statistical Association*, 82, 802–807.

Stigler, S. M. (1977), "Do Robust Estimators Work With Real Data?" (with discussion), *The Annals of Statistics*, 5, 1055–1077.

——— (1986), *The History of Statistics*, Cambridge, MA: Belknap Press.

Tukey, J. (1960), "A Survey of Sampling from Contaminated Distributions," in *Contributions to Probability and Statistics*, ed. I. Olkin, Stanford, CA: Stanford University Press.

Welsh, A. H., and Richardson, A. M. (1997), "Approaches to the Robust Estimation of Mixed Models. Robust Inference," in *Handbook of Statistics*, 15, Amsterdam: North-Holland.

Likelihood

N. Reid

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 1(a) shows the *likelihood function* for the important parameter in the investigators' model, the relative risk of an accident. Figure 1(b) 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. (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 a method devised to accommodate some complexities in the study design.) As with most real life studies, there were a number of decisions related first to data collection, and then to modeling 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 nontechnical account of some of these was given by Redelmeier and Tibshirani (1997c), and a more statistically oriented version was given by Redelmeier and Tibshirani (1997b). In this vignette I am simply using the data to provide an illustration of the likelihood function.

Assume that one is considering a parametric model $f(\mathbf{y};\boldsymbol{\theta})$, which is the probability density function with respect to a suitable measure for a random variable $\mathbf{Y}$. The parameter is assumed to be k-dimensional and the data are assumed to be n-dimensional, often representing a sequence of iid random variables: $\mathbf{Y} = (Y_1, \ldots Y_n)$. The *likelihood function* is defined to be a function of $\boldsymbol{\theta}$, proportional to the model density,

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

where c can depend on $\mathbf{y}$ but not on $\boldsymbol{\theta}$. Within the context of the given parametric model, the likelihood function measures the relative plausibility of various values of $\boldsymbol{\theta}$, for a given observed data point $\mathbf{y}$. The notation for the likelihood function emphasizes that the parameter $\boldsymbol{\theta}$ 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 accommodate one-to-one transformations of the random variable $\mathbf{Y}$ that do not involve θ, as these clearly should have no effect on our inference about $\boldsymbol{\theta}$. Another way to say the same thing is that the likelihood function is not calibrated in $\boldsymbol{\theta}$, or that only relative values $L(\boldsymbol{\theta}_1)/L(\boldsymbol{\theta}_2)$ are well determined.

The likelihood function was proposed by Fisher (1922) as a means of measuring the relative plausibility of various values of $\boldsymbol{\theta}$ by comparing their likelihood ratios. When $\boldsymbol{\theta}$ 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, suggested that ranges of plausible values for $\boldsymbol{\theta}$ can be directly determined from the likelihood function, first by determining the *maximum likelihood estimate,* $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$, the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta};\mathbf{y})$, and then using as a guideline

$$L(\hat{\boldsymbol{\theta}})/L(\boldsymbol{\theta}) \in (1,3), \quad \text{very plausible;}$$

$$L(\hat{\boldsymbol{\theta}})/L(\boldsymbol{\theta}) \in (3,10), \quad \text{somewhat implausible;}$$

and

$$L(\hat{\boldsymbol{\theta}})/L(\boldsymbol{\theta}) \in (10,\infty) \quad \text{highly implausible.}$$

N. Reid is Professor, Department of Statistics, University of Toronto, Ontario, Canada M5S 3G3 (E-mail: *reid@utstat.utoronto.ca*).

Journal of the American Statistical Association
December 2000, Vol. 95, No. 452, Vignettes

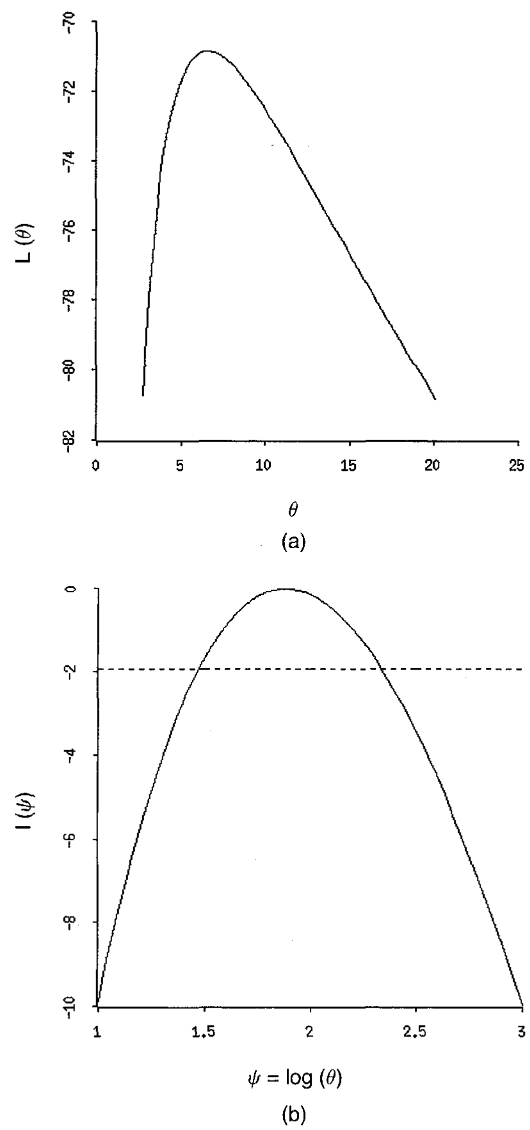

Figure 1. Likelihood and Log-Likelihood Function for the Relative Risk; Based on Redelmeier and Tibshirani (1997c).

The ranges suggested here are taken from Kass and Raftery (1995), attributed to Jeffreys. Other authors have suggested different cutoff points; for example, Fisher (1956, p. 71) suggested using 2, 5, and 15, and Royall (1997) suggested 4, 8, and 32. General introductions to the definition of the likelihood function and its informal use in inference were given by Fisher (1956), Edwards (1972), Kalbfleisch (1985), Azzalini (1996), and Royall (1997).

2. LIKELIHOOD FUNCTION AND INFERENCE

2.1 Bayesian Inference

Although 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 $\boldsymbol{\theta}$, to obtain a posterior probability function,

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

where the constant of proportionality is $\int \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}; \mathbf{y})\, \mathbf{d}\boldsymbol{\theta}$.

This leads directly to Bayesian inferences of this sort: Using the prior density $\pi(\theta)$, we conclude that values of θ greater than θ_U have posterior probability less than .05 and are hence inconsistent with the model and the prior. Jeffreys (1961) emphasized this use of the likelihood function, and investigated the possibility of using "flat" or "noninformative" priors. He also suggested the plausibility range described earlier, in the context of Bayesian inference with a flat prior.

One difficulty in applying Bayesian inference is in constructing a suitable prior, and interest has been renewed 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 that, 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's 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 by Lindley (1958). Unfortunately, there is no satisfactory general prescription for such a *probability matching prior* when $\boldsymbol{\theta}$ is multidimensional. Another type of noninformative prior, motivated rather differently, is the *reference prior* of Bernardo (Berger and Bernardo 1992). Kass and Wasserman (1996) provided 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 multidimensional posterior. These difficulties have largely been solved by the introduction of a number of numerical methods, including importance sampling and Markov chain Monte Carlo (MCMC) methods. An introduction to Gibbs sampling was given by 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 that uses the sampling distribution of the likelihood function, as described in the next section, 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 1960s and 1970s on various aspects of the foundations of inference. A good overview was provided by Berger and Wolpert (1984). More recently, there has 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 standpoint, because the likelihood map is sufficient, which in particular implies that the minimal sufficient statistic in any model is determined by the likelihood map $L(\boldsymbol{\theta};\cdot)$. 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

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

the maximum likelihood estimate

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

and the observed Fisher information

$$\mathbf{j}(\hat{\boldsymbol{\theta}}) = -\partial^2 l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^2|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{6}$$

where $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$, is the log-likelihood function.

In the case where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a sample of iid random variables, the log-likelihood function is a sum of n iid 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 $(Y_1, \ldots, Y_n)$ independent, but not identically distributed, or weakly dependent, can be accommodated 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:

$$\mathbf{l}'(\boldsymbol{\theta})^T\{\mathbf{j}(\hat{\boldsymbol{\theta}})\}\mathbf{l}'(\boldsymbol{\theta}) \xrightarrow{d} \chi^2_p, \tag{7}$$

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

and

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

where χ^2_p is the chi-squared distribution on p degrees of freedom and p is the dimension of $\boldsymbol{\theta}$.

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

$$\sup_{\boldsymbol{\lambda}} l(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{y}) = l(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\psi}; \mathbf{y}) = l_p(\boldsymbol{\psi}), \tag{10}$$

one has, for example,

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

where q is the dimension of $\boldsymbol{\psi}$. The function $l_p(\boldsymbol{\psi})$ 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 $\boldsymbol{\theta}$ held fixed. More generally, limit statements can be derived for the limit as the amount of Fisher information in $\mathbf{Y}$ increases.

The approximations suggested by these limiting results, such as

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

called first-order approximations, are widely used in practice for inference about $\boldsymbol{\theta}$. The development of high-speed computers throughout the last half of the twentieth 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 likelihood-based inference of the past 20 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 articles in the August 1980 issue of *Biometrika* (Barndorff-Nielsen; Cox; Durbin; Hinkley), all of which derived in one version or another that

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

The right side of (13) 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 side to the true sampling density of $\hat{\boldsymbol{\theta}}$ (given $\mathbf{a}$) is $1+O(n^{-3/2})$. Despite its importance, a rigorous proof of (13) is not yet available, although Skovgaard (1990) gave a very careful and helpful derivation. It is necessary to condition on a statistic $\mathbf{a}$ so that (13) is meaningful, because the likelihood function appearing on the right side depends on the data $\mathbf{y}$, yet it is being used as the sampling distribution for $\hat{\boldsymbol{\theta}}$. The role of $\mathbf{a}$ is to complete a one-to-one transformation from $\mathbf{y}$ to $(\hat{\boldsymbol{\theta}}, \mathbf{a})$. For (13) to be useful for inference, $\mathbf{a}$ must have a distribution either exactly or approximately free of $\boldsymbol{\theta}$; otherwise, we have lost information about $\boldsymbol{\theta}$ 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 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 where θ is a scalar, this is expressed as

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

where

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

and

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

where $l_{;\hat{\theta}}(\theta) = \partial l(\theta; \hat{\theta}, \mathbf{a})/\partial\hat{\theta}$. 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 $\mathbf{a}$ 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 by Lugannani and Rice (1980). There are analogous versions of (14), (15), and (16) for inference about a scalar component of $\boldsymbol{\theta}$ in the presence of a nuisance parameter; a partial review was given by Reid (1996), and more recent work was presented by Barndorff-Nielsen and Wood (1998), Fraser, Reid, and Wu (1999), and Skovgaard (1996). (See also the approximations vignette by R. Strawderman.)

3. PARTIAL LIKELIHOOD AND ALL THAT

3.1 Nuisance Parameters

I defined at (10) the profile log-likelihood function $l_p(\psi)$, which is often used in problems in which the parameter of the model $\boldsymbol{\theta}$ is partitioned into a parameter of interest ψ and a nuisance parameter $\boldsymbol{\lambda}$. Typically $\boldsymbol{\lambda}$ is introduced into the model to make it more realistic. More generally, one can define

$$l_p(\psi) = \sup_{\psi=\psi(\boldsymbol{\theta})} l(\boldsymbol{\theta}). \quad (17)$$

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)–(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 $\boldsymbol{\lambda}$ is large relative to n, as it is known that the results break down if the dimension of $\boldsymbol{\theta}$ increases with n. More intuitively, because no adjustment is made for errors of estimation of the nuisance parameter in (10) or (17), it is likely that the apparent precision of (10) or (17) is overstated. Several methods have been suggested for constructing a likelihood function better suited to problems with nuisance parameters. Some models may contain 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, that model has the factorization

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

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

$$l_m(\psi) = l_p(\psi) - \frac{1}{2}\,|\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\psi, \hat{\boldsymbol{\lambda}}_\psi)| + B(\psi) \quad (19)$$

for some choice of $B(\cdot)$ 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 using (18) with $B(\psi) = 0$, and several other versions have been proposed. Brief overviews were given by Mukerjee and Reid (1999) and Severini (1998).

3.2 Partial Likelihood

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

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

where it seems intuitively obvious that the second component cannot provide information about ψ in the absence of knowledge of $\boldsymbol{\lambda}$. 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) also 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 are many related models where a partial likelihood leads to an adequate first-order approximation (Andersen, Borgan, Gill, and Keiding 1993; Murphy and van der Vaart 1997). There is not yet a theory of higher-order approximations in this setting, however. Likelihood partitions, such as (20), were discussed in some generality by Cox (1999).

3.3 Pseudolikelihood

One interpretation of partial likelihood is that the probability distribution of only part of the observed data is modeled, 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 neighbors of any given point, and using the product of these conditional distributions as a pseudolikelihood function. A more direct approach to likelihood inference in spatial point processes was described by Geyer (1999).

3.4 Quasi-Likelihood

The last 30 years have also seen the development of an approach to modeling 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 viewpoint is 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 has in fairly wide generality the same asymptotic distribution theory as a likelihood function (Li and McCullagh 1994; McCullagh 1983).

3.5 Likelihood and Nonparametric Models

Suppose that we have a model in which we assume that $Y_1, \ldots Y_n$ are iid 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 was generalized to much more complex sampling, including censoring, by 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 was given by DiCiccio and Efron (1996).

An alternative to the nonparametric bootstrap is the empirical likelihood function, 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 and 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 were described by Efron and Tibshirani (1993).

4. CONCLUSION

Whether from a Bayesian or a 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*, New York: Springer-Verlag.

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

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

——— (1983), "On a Formula for the Distribution of the Maximum Likelihood Estimator," *Biometrika*, 70, 343–365.

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 *Bayesian Statistics IV*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 35–60.

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

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," *Journal of the American Statistical Association*, 57, 269–306.

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler, *American Statistician*, 46, 167–174.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

——— (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

——— (1980), "Local Ancillarity," *Biometrika*, 67, 279–286.

——— (1999), "Some Remarks on Likelihood Factorization," in *State of the Art in Probability and Statistics*, eds. M. C. M. de Gunst, C. A. J. Klaussen, A. W. van der Waart, Hayward: Institute of Mathematical Statistics.

Cox, D. R., and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 49, 1–39.

DiCiccio, T. J., and Efron, B. (1996), "Bootstrap Confidence Intervals" (with discussion), *Statistical Science*, 11, 189–228.

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

Durbin, J. (1980), "Approximations for Densities of Sufficient Statistics," *Biometrika*, 67, 311–333.

Edwards, A. F. (1972), *Likelihood*, Cambridge, U.K.: Cambridge University Press.

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

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society*, Ser. A, 222,

309–368.
——— (1934), "Two New Properties of Mathematical Likelihood," *Proceedings of the Royal Society*, Ser. A, 144, 285–307.
——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.
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*, eds. O. E. Barndorff-Nielsen, W. S. Kendall, and M. N. M. van Lieshout, London: Chapman and Hall/CRC, pp. 79–140.
Hinkley, D. V. (1980), "Likelihood as Approximate Pivotal," *Biometrika*, 67, 287–292.
Jeffreys, H. (1961), *Theory of Probability*, Oxford, U.K.: Oxford University Press.
Kalbfleisch, J. G. (1985), *Probability and Statistics*, Vol. II (2nd ed.), New York: Springer-Verlag.
Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.
Kass, R. E., and Wasserman, L. (1996), "Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography," *Journal of the American Statistical Association*, 91, 1343–1370.
Li, B., and McCullagh, P. (1994), "Potential Functions and Conservative Estimating Functions," *The Annals of Statistics*, 22, 340–356.
Lindley, D. V. (1958), "Fiducial Distributions and Bayes's Theorem," *Journal of the Royal Statistical Society*, Ser. B, 20, 102–107.
Lugannani, R., and Rice, S. (1980), "Saddlepoint Approximation for the Distribution of the Sum of Independent Random Variables," *Advances in Applications and Probabilities*, 12, 475–490.
McCullagh, P. (1983), "Quasi-Likelihood Functions," *Annals of Statistics*, 11, 59–67.
McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.
Mukerjee, R., and Reid, N. (1999), "On Confidence Intervals Associated With the Usual and Adjusted Likelihoods," *Journal of the Royal Statistical Society*, Ser. B, 61, 945–953.
Murphy, S. A., and van der Vaart, A. (1997), "Semiparametric Likelihood Ratio Inference," *The Annals of Statistics*, 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 England Journal of Medicine*, February 12, 1997, 1–5.
——— (1997b), "Is Using a Cell Phone Like Driving Drunk?," *Chance*, 10, 5–9.
——— (1997c), "Cellular Telephones and Motor Vehicle Collisions: Some Variations on Matched Pairs Analysis," *Canadian Journal of Statistics*, 25, 581–593.
Reid, N. (1995), "The Roles of Conditioning in Inference," *Statistics in Science*, 10, 138–157.
——— (1996), "Likelihood and Higher-Order Approximations to Tail Areas: A Review and Annotated Bibliography," *Canadian Journal of Statistics*, 24, 141–166.
Royall, R. (1997), *Statistical Evidence*, London: Chapman and Hall.
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," *The Annals of Statistics*, 18, 779–789.
——— (1996), "An Explicit Large-Deviation Approximation to One-Parameter Tests," *Bernoulli*, 2, 145–165.
Welch, B. L., and Peers, H. W. (1963), "On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods," *Journal of the Royal Statistical Society*, Ser. B, 25, 318–329.

Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts

James ROBINS and Larry WASSERMAN

1. INTRODUCTION

Statistics is intertwined with science and mathematics but is a subset of neither. The "foundations of statistics" is the set of concepts that makes statistics a distinct field. For example, arguments for and against conditioning on ancillaries are purely statistical in nature; mathematics and probability do not inform us of the virtues of conditioning, but only on how to do so rigorously. One might say that foundations is the study of the fundamental conceptual principles that underlie statistical methodology. Examples of foundational concepts include ancillarity, coherence, conditioning, decision theory, the likelihood principle, and the weak and strong repeated-sampling principles. A nice discussion of many of these topics was given by Cox and Hinkley (1974).

There is no universal agreement on which principles are "right" or which should take precedence over others. Indeed, the study of foundations includes much debate and controversy. An example, which we discuss in Section 2, is the likelihood principle, which asserts that two experiments that yield proportional likelihood functions should yield identical inferences. According to Birnbaum's theorem, the likelihood principle follows logically from two other principles: the conditionality principle and the sufficiency principle. To many statisticians, both conditionality and sufficiency seem compelling yet the likelihood principle does not. The mathematical content of Birnbaum's theorem is not in question. Rather, the question is whether conditionality and sufficiency should be elevated to the status of "principles" just because they seem compelling in simple examples. This is but one of many examples of the type of debate that pervades the study of foundations.

This vignette is a selective review of some of these key foundational concepts. We make no attempt to be complete

James Robins is Professor, Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115. Larry Wasserman is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.
This research was supported by National Institutes of Health grants R01-CA54852-01 and R01-A132475-07 and National Science Foundation grants DMS-9303557 and DMS-9357646. The authors thank David Cox, Phil Dawid, Sander Greenland, Erich Lehmann, and Isabella Verdinelli for many helpful suggestions.

Journal of the American Statistical Association
December 2000, Vol. 95, No. 452, Vignettes