Published in final edited form as:

Int Stat Rev. 2010 April 1; 78(1): 102–116. doi:10.1111/j.1751-5823.2010.00107.x.

Profile Likelihood and Incomplete Data

Zhiwei Zhang

Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20892-7510, USA

Summary

According to the law of likelihood, statistical evidence is represented by likelihood functions and its strength measured by likelihood ratios. This point of view has led to a likelihood paradigm for interpreting statistical evidence, which carefully distinguishes evidence about a parameter from error probabilities and personal belief. Like other paradigms of statistics, the likelihood paradigm faces challenges when data are observed incompletely, due to non-response or censoring, for instance. Standard methods to generate likelihood functions in such circumstances generally require assumptions about the mechanism that governs the incomplete observation of data, assumptions that usually rely on external information and cannot be validated with the observed data. Without reliable external information, the use of untestable assumptions driven by convenience could potentially compromise the interpretability of the resulting likelihood as an objective representation of the observed evidence. This paper proposes a profile likelihood approach for representing and interpreting statistical evidence with incomplete data without imposing untestable assumptions. The proposed approach is based on partial identification and is illustrated with several statistical problems involving missing data or censored data. Numerical examples based on real data are presented to demonstrate the feasibility of the approach.

Keywords

Censoring; evidential analysis; law of likelihood; likelihood paradigm; missing data; partial identification

1 Introduction

Statisticians are routinely asked to interpret observed data as statistical evidence, even though there is no consensus on what constitutes statistical evidence and how to measure its strength. Hacking (1965) suggests that the likelihood function is the mathematical representation of statistical evidence and that the likelihood ratio (LR) measures the strength of statistical evidence for one hypothesis versus another. This point of view has led to a likelihood paradigm for interpreting statistical evidence that carefully distinguishes evidence about a parameter from error probabilities and personal belief (Royall, 1997; Blume, 2002). Royall (2000) analyzes the probability of observing misleading evidence under parametric models, and Blume (2008) provides a parallel analysis for sequential trials. Royall & Tsou (2003) and Blume et al. (2007) develop adjusted likelihood functions with certain robustness properties under model failure.

Like other paradigms of statistics, the likelihood paradigm faces challenges when data are observed incompletely, due to non-response or censoring, for instance. Such problems, discussed extensively in the frequentist estimation theory as well as in Bayesian statistics, do not appear to have been considered in the likelihood paradigm. This is not surprising given the relative sparsity of the likelihood paradigm literature, which is nonetheless growing rapidly. Some of the estimation methods for incomplete data are likelihood-based and would be easily adaptable to the likelihood paradigm. However, the existing methods, likelihood-based or not, generally require assumptions about the mechanism that governs the incomplete observation of data. In the case of missing data, it is common practice to assume missing-completely-at-random (MCAR) or missing-at-random (MAR) in the sense of Rubin (1976) or specify a model for the missing-data mechanism. With censored data, one might assume independent censoring, possibly conditional on a set of covariates, or specify a model for the censoring mechanism. Such assumptions usually rely on external information and cannot be validated with the observed data. Without reliable external information, the use of untestable assumptions driven by convenience could potentially compromise the interpretability of the resulting likelihood as an objective representation of the observed evidence.

Example (Proportion). To fix ideas, consider the problem of making inference about a proportion with missing data. Suppose that X is a Bernoulli variable, and that the parameter of interest is $\gamma = P(X = 1)$, a population proportion. If a random sample of X is completely observed, then the evidence about γ can be represented with a standard binomial likelihood. Suppose, however, that X is missing on some subjects, which can happen because of patient dropout in clinical trials, for instance. Let R be the observation indicator, so $R = 1$ if X is observed and 0 otherwise. If X is assumed MCAR in the sense that R and X are independent, then the likelihood for γ is proportional to a binomial likelihood based on the complete cases only. Alternatively, a selection model could be specified for the conditional distribution of R given X , and the resulting likelihood would involve both the complete and the incomplete cases (Little & Rubin, 2002). None of these assumptions can be tested with the observed data, and it is natural to ask exactly what the data suggest about γ without these untestable assumptions.

The goal of this work is to develop, for a class of incomplete data problems including the above example, a likelihood method to describe and quantify the observed evidence without imposing untestable assumptions. A notable consequence of eliminating such assumptions is that the parameter of interest may become partially identified. Manski (2003) gives an extensive discussion on the notion of partial identification, which has originated and become increasingly popular in the econometrics literature. Early research in this area has focused on characterization of identification regions, and recent research efforts have been directed toward confidence sets for identification regions as well as for the parameters of interest (e.g., Horowitz & Manski, 1998, 2000; Imbens & Manski, 2004; Beresteanu & Molinari, 2008; Zhang, 2009). The framework of Zhang (2009) for partial identification (PI) with incomplete data will be adopted in the present paper. The focus here, of course, is on developing a likelihood method for evidential analysis instead of constructing confidence sets.

The rest of the paper is organized as follows. Section 2 gives a brief overview of the likelihood paradigm. Section 3 describes the incomplete data problem in some generality and discusses the associated identification problem, using the notation and terminology of Zhang (2009). A profile likelihood approach is then proposed in Section 4. The proportion example introduced earlier will be used as an illustration throughout this discussion. The practical implications of the proposed approach are considered in Section 5, with an asymptotic discussion followed by a simulation study. The approach will be further

illustrated in Sections 6–8 with additional statistical problems involving missing data or censored data. In most cases the illustration includes a numerical example based on real data. Section 9 gives concluding remarks.

2 The Likelihood Paradigm

The likelihood paradigm is concerned with interpreting statistical evidence and answering questions like “what do the data say?” with regard to two competing hypotheses. Hypothesis tests and posterior probabilities are commonly used to answer such questions in practice. Royall (1997) points out that the Neyman–Pearson theory for testing hypotheses is really a decision theory designed to answer such questions as “what should I do?”. Serious logical inconsistencies can arise when hypothesis tests are used to represent statistical evidence (Royall, 1997, Chapter 2). Posterior probabilities, on the other hand, incorporate prior information as well as the data and may not provide an objective representation of the observed evidence. The Bayesian approach may be more appropriate for questions like “what should I believe?”. A proper concept of evidence is missing from the major statistical theories that dominate today's statistical practice.

The missing concept of evidence can be found in the law of likelihood (Hacking, 1965), which can be stated as follows. Suppose a random variable Y^* follows a distribution with density $f(\cdot; \gamma)$, which is known up to a parameter γ taking values in Γ . Then the likelihood for γ , based on the observation $Y^* = y^*$, is given by $L(\gamma) = f(y^*; \gamma)$. According to the law of likelihood, the data provide evidence supporting one hypothesis $H_1 : \gamma = \gamma_1$ over another hypothesis $H_2 : \gamma = \gamma_2$ if $L(\gamma_1) > L(\gamma_2)$, and the strength of that evidence is measured by the ratio $L(\gamma_1)/L(\gamma_2)$. Unlike conventional hypothesis tests, such a comparison does not require the hypotheses to be pre-specified. Royall (1997) and Blume (2002) discuss the implications of the law of likelihood and compare the resulting likelihood paradigm with other paradigms of statistics. Royall (1997) also suggests benchmarks for describing the strength of evidence. Specifically, a LR exceeding 8 (respectively, 32) is said to represent moderate (respectively, strong) evidence.

When γ is a vector, some of its components may be of greater interest than others. Suppose $\gamma = (\eta, \nu)$, where η is of primary interest and ν is a nuisance parameter. The law of likelihood does not prescribe a specific procedure to deal with nuisance parameters. Under special circumstances, one may be able to obtain a marginal, conditional or partial likelihood for η alone (Royall, 1997, Chapter 7). Another approach, based on the profile likelihood $\tilde{L}(\eta) = \sup_{\nu} L(\eta, \nu)$, is more generally applicable. Royall (2000) shows that $\tilde{L}$ has reasonable asymptotic properties, even though it does not correspond to a real probability density function.

While the likelihood paradigm requires a reasonable likelihood function to operate on, it does not necessarily require correct specification of a parametric model. There are different approaches to robust inference under the likelihood paradigm. Royall & Tsou (2003) start with a working parametric model for which the object of inference (i.e., the probability limit of the maximum likelihood estimator) remains of interest even when the model fails. They apply an exponential adjustment to the likelihood function based on the working model and show that the adjusted likelihood is robust in an asymptotic sense. Alternatively, one could focus on defining the parameter of interest via estimating functions without specifying the entire distribution of Y^* , because the relevant evidence about the parameter can then be represented with an empirical likelihood function, which has asymptotic properties parallel to those of the various parametric likelihood functions discussed earlier. As will be illustrated in Section 7, nonparametric likelihood functions can be very helpful in robust evidential analysis.

3 Incomplete Data and Partial Identification

Suppose now that Y^* is not completely observed; instead we observe $Y = h(Y^*)$, where h is a known function. Let P^* denote the set of all possible distributions of Y^* ; then the induced model for Y is $P = \{P^*h^{-1} : P^* \in P^*\}$. It is often convenient to parameterize P as $\{P_\theta : \theta \in \Theta\}$, where Θ may be finite- or infinite-dimensional. Suppose θ is identifiable in the usual sense that $\theta_1 = \theta_2$ whenever $P_{\theta_1} = P_{\theta_2}$. This implies that θ is determined by P^* , the distribution of Y^* , so we can write $\theta = S(P^*)$ for a known functional $S : P^* \rightarrow \Theta$. The parameter of interest will be written as $\gamma = T(P^*)$ for another (known) functional T . There is no need to consider a nuisance component of γ now because γ is not required to determine P^* .

Example (Proportion, Continued). The latent variable Y^* in this example is a collection of independent copies of (X, R) , and its distribution P^* is defined by the probabilities $p_{xr} = P^*(X = x, R = r)$, $x, r = 0, 1$. The parameter of interest is then

$$\gamma = P^*(X=1) = p_{10} + p_{11}.$$

The observable Y may be thought of as independent copies of (RX, R) , since X can be recovered from RX if and only if $R = 1$. Note that Y is essentially a trinomial variable and its distribution is defined by

$$\theta = (p_{01}, p_{11}, p_{\cdot 0}).$$

Here and in the sequel, dot $(\cdot)$ in place of a subscript denotes summation over that subscript; in particular, $p_{\cdot 0} = p_{00} + p_{10}$.

A pair of values (θ, γ) will be said to be *compatible* if they can possibly obtain together, that is, if there exists $P^* \in P^*$ such that $S(P^*) = \theta$ and $T(P^*) = \gamma$. If the value of θ is known, then γ must be one of those values compatible with θ . The collection of such values is called the *identification region* for γ given θ . Formally, this is written as

$$\Gamma|_{\theta} = TS^{-1}(\theta) = \{T(P^*) : S(P^*) = \theta\},$$

where the superscript -1 denotes inverse image. Conversely, if the value of γ is given, then θ must be one of those values compatible with γ . This latter set of values is called the *inverse identification region* for θ given γ and is denoted by

$$\Theta|_{\gamma} = ST^{-1}(\gamma) = \{S(P^*) : T(P^*) = \gamma\}.$$

The parameter γ is said to be *completely identified* if $\Gamma|_{\theta}$ is a singleton, *unidentified* if $\Gamma|_{\theta} = \Gamma$, and *partially identified* if $\Gamma|_{\theta}$ is a proper subset of Γ consisting of more than one element. This terminology refines the standard terminology by drawing a distinction between the partially identified case and the truly unidentified case.

Example (Proportion, Continued). In this example, the identification region for γ given θ is easily seen to be

$$\Gamma|_{\theta} = [p_{11}, p_{11} + p_{\cdot 0}],$$

and the inverse identification region for θ given γ is

$$\Theta|_{\gamma} = \{\theta: p_{11} \leq \gamma \leq p_{11} + p_{\cdot 0}\}.$$

The identification of γ depends on $p_{\cdot 0}$, the probability that X is missing. Specifically, γ is completely identified if $p_{\cdot 0} = 0$ (X always observed), unidentified if $p_{\cdot 0} = 1$ (X always missing), and partially identified otherwise.

If completely identified, γ is essentially a function of θ and evidence about θ easily translates into evidence about γ . On the other hand, in the unidentified case, γ is completely arbitrary even if θ is known and there is little hope to make any meaningful inference about γ without additional information. This paper is primarily concerned with evidential interpretations about a partially identified γ .

4 A Profile Likelihood Approach

We shall take as our starting point a likelihood L for θ based on Y , which can be parametric, semiparametric or nonparametric, depending on the nature of θ . At this stage, there is no need to be specific about this likelihood, as long as it provides a natural representation of the observed evidence about θ . For motivation, let us consider the completely identified case where γ is a function of θ (but θ may not be a function of γ). In this case, each value of γ corresponds to a distinct θ -set given by $\Theta|_{\gamma}$, and it seems reasonable to represent the evidence about γ with the profile likelihood:

$$\tilde{L}(\gamma) = \sup_{\theta \in \Theta|_{\gamma}} L(\theta). \quad (1)$$

Without complete identification, the hypothesis that γ equals some fixed value γ_1 is no longer equivalent to the hypothesis $\theta \in \Theta|_{\gamma_1}$. The former implies the latter but not vice versa. Nonetheless, it seems natural to continue to use the profile likelihood $\tilde{L}$ given by (1) to represent the evidence about γ . As a consequence of partial identification, $\tilde{L}$ does not generally have a unique maximizer. If $\hat{\theta}$ is a maximum likelihood estimate of θ , then $\tilde{L}$ attains its maximum at every point in $\Gamma|\hat{\theta}$. On the other hand, the profile likelihood can take on lower values outside $\Gamma|\hat{\theta}$ so as to provide an informative representation of evidence, unless γ is unidentified (in which case $\tilde{L}$ is constant). The profile likelihood $\tilde{L}$ can be summarized numerically through support sets of the form $S_k = \{\gamma : \tilde{L}(\gamma) > \sup_{\Gamma} \tilde{L}(\Gamma)/k\}$, $k > 1$. This looks similar to the likelihood ratio confidence sets derived by Zhang (2009), but the interpretations are very different. It can be shown that S_k is the smallest subset of Γ supported by a factor of k over its own complement. Confidence sets, on the other hand, do not have a direct evidential interpretation (Royall, 1997).

Example (Proportion, Continued). Let (X_i, R_i) , $i = 1, \dots, n$, be independent copies of (X, R) . Recall that Y is essentially a trinomial variable with probability vector $\theta = (p_{01}, p_{11}, p_{\cdot 0})$. A standard likelihood for θ is

$$L(\theta) = p_{01}^{N_{01}} p_{11}^{N_{11}} p_{\cdot 0}^{N_{\cdot 0}},$$

where, $N_{01} = \sum_{i=1}^n R_i (1 - X_i)$, $N_{11} = \sum_{i=1}^n R_i X_i$ and $N_{\cdot 0} = \sum_{i=1}^n (1 - R_i)$. It follows that $\hat{\theta} = (\hat{p}_{01}, \hat{p}_{11}, \hat{p}_{\cdot 0}) = n^{-1}(N_{01}, N_{11}, N_{\cdot 0})$. To avoid trivialities, all components of $\hat{\theta}$ are assumed positive. Then the profile likelihood for γ is given by

$$\tilde{L}(\gamma) = \begin{cases} L\left(\frac{N_{01}(1-\gamma)}{N_{01}+N_0}, \gamma, \frac{N_{01}(1-\gamma)}{N_{01}+N_0}\right), & 0 < \gamma < \hat{p}_{11}; \\ L(\hat{\theta}), & \hat{p}_{11} \leq \gamma \leq \hat{p}_{11} + \hat{p}_{00}; \\ L\left(1 - \gamma, \frac{N_{11}\gamma}{N_{11}+N_0}, \frac{N_{01}\gamma}{N_{11}+N_0}\right), & \hat{p}_{11} + \hat{p}_{00} < \gamma < 1. \end{cases} \quad (2)$$

Without missing data, this reduces to the usual binomial likelihood.

The method can be illustrated numerically with real data from the Obstructed Coronary Bypass Graft Trial, discussed previously by Hollis (2002). In this trial, 110 patients with obstructed coronary bypass grafts were treated with a stent, and the outcome of interest was restenosis at 6 months (an undesirable event). However, 24 of these patients dropped out of the study with unknown 6-month restenosis status. Of the remaining 86 patients, 32 had restenosis at 6 months while 54 did not. Based on these data, Figure 1 displays the profile likelihood for the 6-month restenosis rate given by (2) as well as a binomial likelihood from a complete-case analysis (restricted to patients with observed outcomes). In this and the subsequent likelihood plots, each likelihood function is divided by its maximum value so the peak value is invariably 1. Obviously, the profile likelihood with no assumptions about the missing-data mechanism appears “flatter” than the complete-case likelihood assuming MCAR. This is a natural consequence of partial identification, as will be discussed in Section 5.

5 Practical Implications

We now consider the practical implications of the proposed profile likelihood method, first asymptotically and then empirically. The following asymptotic discussion involves standard conditions and elementary arguments, which are omitted in order to avoid distractions. For a completely identified parameter, a reasonable likelihood is expected to be consistent in the sense that the true value γ_0 eventually dominates every other fixed value. This can be formalized as $L(\gamma)/L(\gamma_0) \xrightarrow{P} 0$ for any $\gamma \neq \gamma_0$. The same property would be unrealistic to expect in the partially identified case, because the observed data provide no information to distinguish among the values in $\Gamma|_{\theta_0}$, where θ_0 is the true value of θ . This limitation applies to any likelihood method that avoids making untestable identifying assumptions, and the profile likelihood $\tilde{L}$ is no exception. As a consequence, $\tilde{L}(\gamma_0)$ may fail to dominate $\tilde{L}(\gamma)$ if $\gamma \neq \gamma_0$ lies inside or on the boundary of $\Gamma|_{\theta_0}$. On the other hand, no value of γ can dominate γ_0 asymptotically under $\tilde{L}$. This is in contrast to likelihood methods based on untestable identifying assumptions, which are consistent when the assumptions hold but can be dangerously inconsistent, allowing γ_0 to be dominated by other values, when the identifying assumptions fail. In large samples, L will be largely proportional to the indicator function of $\Gamma|_{\theta_0}$, which represents all that could possibly be learned about γ_0 from the data without external information. In this sense, the profile likelihood method provides an objective representation of the observed evidence without additional assumptions. A sharper likelihood function can be obtained by invoking untestable identifying assumptions. While this is often reasonable to do, one should recognize that the resulting likelihood represents not only the observed evidence but also the identifying assumptions. The effect of the identifying assumptions is basically shrinking the identification region into a single value, which may or may not be equal (or close) to the true value, depending on the validity of the assumptions.

The following simulation study sheds some light on the finite-sample behaviour of the proposed approach in comparison with standard methods based on identifying assumptions.

The study is conducted in the setting of the proportion example, under a selection model given by

$$\text{logit} \{P^*(R=1|X=x)\} = \alpha + \beta x, \quad x=0, 1. \quad (3)$$

Note that β determines the (non-)ignorability of the missing-data mechanism, with $\beta = 0$ corresponding to MCAR. In this study, data are simulated with $n = 100$, $\gamma = P^*(X = 1) = 0.5$, $\beta \in \{0, 1, -1\}$, and α chosen such that $p_0 = P(R = 0) = 0.1$ or 0.3 . In each scenario (combination of parameter values), 1,000 replicate samples are generated. The proposed method will be compared with a more standard likelihood method that can be described as follows. Note that model (3) is not identifiable without additional assumptions or constraints. One way to constrain the model is to fix the value of β , which allows us to obtain a standard profile likelihood for γ as

$$\tilde{L}_\beta(\gamma) = \sup_\alpha \left\{ \left(\frac{(1-\gamma)e^\alpha}{1+e^\alpha} \right)^{N_{01}} \left(\frac{\gamma e^{\alpha+\beta}}{1+e^{\alpha+\beta}} \right)^{N_{11}} \left(\frac{1-\gamma}{1+e^\alpha} + \frac{\gamma}{1+e^{\alpha+\beta}} \right)^{N_{10}} \right\}. \quad (4)$$

Note that, for $\beta = 0$ (MCAR), the last term disappears and the profile likelihood reduces to a binomial likelihood based on the complete cases. The maximization in (4) can be carried out with an EM algorithm which naturally treats (N_{00}, N_{10}) as the missing data. Given the current value $\alpha^{(k)}$, the E-step amounts to computing

$$\begin{aligned} E\{N_{00}|N_{01}, N_{11}, N_{10}; \gamma, \alpha^{(k)}, \beta\} &= \frac{N_{10}(1-\gamma) \left(\frac{1+e^{\alpha^{(k)+\beta}}}{1+e^{\alpha^{(k)}}} \right)}{(1-\gamma) \left(\frac{1+e^{\alpha^{(k)+\beta}}}{1+e^{\alpha^{(k)}}} \right) + \gamma \left(\frac{1+e^{\alpha^{(k)}}}{1+e^{\alpha^{(k)+\beta}}} \right)} =: \widehat{N}_{00}^{(k)} \\ E\{N_{10}|N_{01}, N_{11}, N_{10}; \gamma, \alpha^{(k)}, \beta\} &= N_{10} - \widehat{N}_{00}^{(k)} =: \widehat{N}_{10}^{(k)}, \end{aligned}$$

and the M-step sets $\alpha^{(k+1)}$ to the maximizer of the complete-data log-likelihood

$$\widehat{N}_{00}^{(k)} \log \frac{1}{1+e^\alpha} + N_{01} \log \frac{e^\alpha}{1+e^\alpha} + \widehat{N}_{10}^{(k)} \log \frac{1}{1+e^{\alpha+\beta}} + N_{11} \log \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}},$$

which can be found using a Newton–Raphson algorithm. To address the uncertainty about β , L_β will be calculated for several different values of β (0, 1, -1). This will be referred to as a sensitivity analysis (SA) approach. The proposed method based on PI will be abbreviated as PI.

Suppose the true value $\gamma_0 = 0.5$ is to be compared with the alternative $\gamma_1 = 0.3$. Such a comparison can be made via LRs, and Table 1 gives the frequency distribution of the LRs for each method in each scenario. In Table 1, the LRs for γ_0 versus γ_1 are grouped into five intervals: 0–1/32 (strong evidence for γ_1), 1/32–1/8 (moderate evidence for γ_1), 1/8–8 (weak evidence), 8–32 (moderate evidence for γ_0), and 32+ (strong evidence for γ_0). It is easy to see that, under the SA approach, the mode of the likelihood (i.e., the maximum likelihood estimate) decreases with β_{guess} , the working value of β . Indeed, Table 1 shows that smaller values of β_{guess} (e.g., -1) are more likely to produce large LRs for γ_0 versus γ_1 , whereas larger values of β_{guess} (e.g., 1) are more prone to false claims of strong evidence for γ_1 . Setting β_{guess} equal to β_{true} , the true value of β , is not necessarily advantageous for comparing a given pair of hypotheses. The PI method is reliably more conservative than SA (for any β_{guess}) and less likely to indicate strong evidence in either direction. This may be desirable, especially when comparing against SA with $\beta_{\text{guess}} = 1$ in the last scenario.

Increasing p_{j0} from 0.1 to 0.3 widens the identification interval and magnifies the difference between PI and SA. Overall, the simulation results suggest that the PI method operates like a systematic and exhaustive SA seeking consensus among all possible identifying assumptions (hence the conservative behaviour).

6 Comparing Two Proportions

The proposed profile likelihood approach will be further illustrated with other statistical problems in the rest of the paper. The general notation of Sections 3 and 4 will be used throughout, although its exact meaning will depend on the particular problem. In this section, we extend the proportion example slightly and consider the problem of comparing two proportions with missing data. For $j = 1, 2$, let $\{X_{ji} : i = 1, \dots, n_j\}$ denote a random sample from a Bernoulli distribution with success probability γ_j . The two samples are assumed independent, as in randomized clinical trials. As in the previous example, let R_{ji} be the observation indicator for X_{ji} , so $R_{ji} = 1$ if X_{ji} is observed and 0 otherwise. Thus Y^* consists of (X_{ji}, R_{ji}) , $i = 1, \dots, n_j$, $j = 1, 2$, and the distribution of Y^* is characterized by $p_{jxr} = P^*(X_{j1} = x, R_{j1} = r)$, $j = 1, 2$, $x = 0, 1$, $r = 0, 1$. Now, we can write $\gamma_j = p_{j1} = p_{j10} + p_{j11}$, $j = 1, 2$. The observable consists $(R_{ji}X_{ji}, R_{ji})$, $i = 1, \dots, n_j$, $j = 1, 2$, and the identifiable parameter can be written as $\theta = (\theta_1, \theta_2)$, where $\theta_j = (p_{j01}, p_{j11}, p_{j00})$, $j = 1, 2$. Between θ and γ (γ_1, γ_2), the compatibility relation can be expressed as $p_{j11} \leq \gamma_j \leq p_{j11} + p_{j00} = 1$, 2 .

A standard likelihood for θ based on Y is given by $L(\theta) = L_1(\theta_1)L_2(\theta_2)$, where

$$L_j(\theta_j) = p_{j01}^{N_{j01}} p_{j11}^{N_{j11}} p_{j00}^{N_{j00}} \text{ with } N_{j01} = \sum_{i=1}^{n_j} R_{ji}(1 - X_{ji}), N_{j11} = \sum_{i=1}^{n_j} R_{ji}X_{ji} \text{ and } N_{j00} = \sum_{i=1}^{n_j} (1 - R_{ji}), j = 1, 2.$$

Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ denote the maximum likelihood estimate of θ . For each γ_j , the profile likelihood L_j involves only $L_j(\theta_j)$ and takes the form (2) with an additional subscript j to each relevant quantity. Each L_j is maximized at the interval $[\underline{\gamma}_j, \bar{\gamma}_j]$, where $\underline{\gamma}_j = \hat{p}_{j11}$ and $\bar{\gamma}_j = \hat{p}_{j11} + \hat{p}_{j00}$, and its maximum value is just $L_j(\hat{\theta}_j)$. The profile likelihood for γ is simply $L(\gamma) = L_1(\gamma_1)L_2(\gamma_2)$, which is maximized at $[\underline{\gamma}_1, \bar{\gamma}_1] \times [\underline{\gamma}_2, \bar{\gamma}_2]$.

Let $\gamma_e = e(\gamma)$ denote an effect measure for comparing γ_1 with γ_2 . Common choices of γ_e include the difference $\gamma_2 - \gamma_1$, the ratio γ_2/γ_1 and the odds ratio $\gamma_2(1 - \gamma_1)/\{\gamma_1(1 - \gamma_2)\}$. The profile likelihood for γ_e can be computed as $\tilde{L}_e(\gamma_e) = \sup\{\tilde{L}(\gamma) : e(\gamma) = \gamma_e\}$. It is easy to verify that this is identical to the original definition of $L_e(\gamma_e)$ according to (1). Suppose $e(\gamma) = \gamma_2 - \gamma_1$; then $\tilde{L}_e(\gamma_e) = \sup_{\gamma_1} \tilde{L}(\gamma_1, \gamma_1 + \gamma_e)$ is maximized at $[\underline{\gamma}_2 - \bar{\gamma}_1, \bar{\gamma}_2 - \underline{\gamma}_1]$ and its maximum value is equal to $L(\hat{\theta})$. For $\gamma_e > \bar{\gamma}_2 - \underline{\gamma}_1$, a moment of thought reveals that the maximization of $\tilde{L}(\gamma_1, \gamma_1 + \gamma_e)$ over γ_1 can be restricted to the interval $[(\bar{\gamma}_2 - \gamma_e) \vee 0, (1 - \gamma_e) \wedge \underline{\gamma}_1]$, where $\vee$ denotes maximum and $\wedge$ denotes minimum. The maximizer can be found analytically as in Miettinen & Nurminen (1985) and Farrington & Manning (1990) or numerically using a bisection method (Zhang, 2006). At $\gamma_e < \underline{\gamma}_2 - \bar{\gamma}_1$, $\tilde{L}_e$ can be evaluated in a similar fashion by maximizing $\tilde{L}(\gamma_1, \gamma_1 + \gamma_e)$ over $\gamma_1 \in [(-\gamma_e), \vee \bar{\gamma}_1, (\gamma_2 - \gamma_e) \wedge 1]$.

As an example, the Obstructed Coronary Bypass Graft Trial is actually a randomized study comparing stent placement with balloon angioplasty. The angioplasty arm consisted of another 110 patients. In this arm, the 6-month restenosis status was confirmed for 37 patients, ruled out for 43, and unknown for the other 30. Figure 2 shows a contour plot of the joint profile likelihood for the 6-month restenosis rates in the two treatment groups. Figure 3 presents two profile likelihood functions for the difference between the two rates: the proposed one based on PI and a more standard one assuming MAR. The proposed profile likelihood is largely neutral about the superiority of stent placement to balloon angioplasty, with both regions (superiority and non-superiority) containing plausible values that maximize the likelihood. Therefore, any evidence for superiority that might be found in the MAR-based profile likelihood should be attributed to the MAR assumption.

7 Means of Bounded Variables

Suppose now that X follows an arbitrary distribution supported by the known interval $[a, b]$ and is not always observed, with R the observation indicator as usual. Then the latent variable Y^* consists of independent copies (X_i, R_i) , $i = 1, \dots, n$. The distribution of Y^* can be characterized by (p, F_0, F_1) , where $p = P(R = 1)$ and F_r denotes the conditional distribution of X given $R = r$, $r = 0, 1$. Let F denote the marginal distribution of X ; then $F = (1 - p)F_0 + pF_1$. Suppose the target parameter is $\gamma = EX = M(F)$, where M denotes the mean functional: $M(F) = \int x dF(x)$. This allows us to write $\gamma = (1 - p)M(F_0) + pM(F_1)$. As in previous examples, the observable can be conceptualized as $Y = \{(R_i X_i, R_i) : i = 1, \dots, n\}$, a transformation of Y^* . The distribution of Y identifies p and F_1 but not F_0 , so $\theta = (p, F_1)$. Obviously, θ and γ are compatible if and only if

$$pM(F_1) + (1 - p)a \leq \gamma \leq pM(F_1) + (1 - p)b, \quad (5)$$

and γ is partially identified if $p \in (0, 1)$ and $-\infty < a < b < \infty$. The boundedness of X is often reasonable to assume, although precise specification of the bounds may be difficult in practice.

A nonparametric likelihood for θ Based on Y can be defined as

$$L(\theta) = \prod_{i=1}^n (pF_1\{X_i\})^{R_i} (1 - p)^{1-R_i}, \quad (6)$$

where $F_1\{x\}$ denotes the point mass of F_1 at x . Maximizing this likelihood yields the maximum likelihood estimate $\hat{\theta} = (\hat{p}, \hat{F}_1)$, where $\hat{p}$ is the sample proportion of complete cases and $\hat{F}_1$ is the empirical distribution of the observed values of X . Maximizing the likelihood (6) subject to (5) leads to the profile likelihood for γ , which can also be written as

$$\begin{aligned} \tilde{l}(\gamma) &= \log \tilde{L}(\gamma) \\ &= n \log n + \max \left\{ \sum_{i=1}^n \log w_i : w_i > 0 \forall i, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i X_{ai} \leq \gamma \leq \sum_{i=1}^n w_i X_{bi} \right\}, \end{aligned} \quad (7)$$

where $X_{ai} = R_i X_i + (1 - R_i)a$ and $X_{bi} = R_i X_i + (1 - R_i)b$, $i = 1, \dots, n$ (Zhang, 2009, Appendix A.2). The constant $n \log n$ on the right-hand side of (7) serves to standardize the log-likelihood by setting its maximum value at 0. The maximum of $\tilde{l}$ is attained at $\gamma \in [\bar{X}_a, \bar{X}_b]$, where

$$\begin{aligned} \bar{X}_a &= n^{-1} \sum_{i=1}^n X_{ai} = \hat{p}M(\hat{F}_1) + (1 - \hat{p})a, \\ \bar{X}_b &= n^{-1} \sum_{i=1}^n X_{bi} = \hat{p}M(\hat{F}_1) + (1 - \hat{p})b. \end{aligned}$$

For $\gamma < \bar{X}_a$, the profile log-likelihood is easily seen to be

$$\tilde{l}(\gamma) = n \log n + \max \left\{ \sum_{i=1}^n \log w_i : w_i > 0 \forall i, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i X_{ai} = \gamma \right\},$$

that is, the empirical log-likelihood for the pseudo-observations X_{ai} (with missing X 's replaced by a). Following Owen (1988, section 3), we have

$$\tilde{l}(\gamma) = - \sum_{i=1}^n \log \{1 + t(\gamma)(X_{ai} - \gamma)\},$$

where $t(\gamma)$ satisfies

$$\sum_{i=1}^n \frac{X_{ai} - \gamma}{1 + t(\gamma)(X_{ai} - \gamma)} = 0. \quad (8)$$

For $\gamma \in (a, \bar{X}_a)$, $t(\gamma)$ must lie in $(-(\max_{1 \leq i \leq n} X_{ai} - \gamma)^{-1}, -(\min_{1 \leq i \leq n} X_{bi} - \gamma)^{-1})$, on which the left side of (8) is strictly decreasing in $t(\gamma)$, so $\tilde{l}(\gamma)$ can be evaluated using a bisection method or, more efficiently, a safeguarded zero finding algorithm (Owen, 1988). The computation for $\gamma \in (\bar{X}_b, b)$ is completely analogous. Obviously, $\tilde{l}$ reduces to the empirical log-likelihood for the X_i if they are completely observed.

The preceding discussion can be easily extended to the two-sample situation as in Section 6, with the parameter space for $\gamma = (\gamma_1, \gamma_2)$ replaced by $[a, b]^2$.

8 Application to Censored Data

Besides missing data, the proposed method can also be used to deal with censored data. Let U be a failure time and V a censoring time with distributions F and G , respectively. These are not completely observed because of censoring. Instead we observe $X = U \wedge V$ together with a censoring indicator $\Delta = I(U \leq V)$. As usual, suppose the scientific interest lies in the distribution function $F(u)$ or, equivalently, the survival function $1 - F(u)$. (To simplify notation, the same symbols are used here to denote probability distributions and the associated distribution functions.) A common assumption for handling censored data is independent censoring, namely that U and V are independent, which cannot be validated with the observed data (Tsiatis, 1975). The available methods that allow dependent censoring typically rely on other untestable assumptions about the censoring mechanism. In contrast, the proposed method can be used to represent the observed evidence about $F(u)$ without imposing untestable assumptions.

This problem can be cast in the framework of Section 3 by taking $Y^* = (U_i, V_i) : i = 1, \dots, n$ and $Y = \{(X_i, \Delta_i) : i = 1, \dots, n\}$. Let $p = P(\Delta = 1)$ and let F_δ (respectively, G_δ) denote the conditional distribution of U (respectively, V) given $\Delta = \delta$, $\delta = 0, 1$. It follows that $F = (1 - p)F_0 + pF_1$, $G = (1 - p)G_0 + pG_1$ and $X \sim (1 - p)G_0 + pF_1$. The distribution of Y identifies $\theta(p, G_0, F_1)$, while the target parameter is

$$\gamma = F(u) = (1 - p)F_0(u) + pF_1(u)$$

for a given time u . The compatibility relation between θ and γ can be characterized as

$$pF_1(u) \leq \gamma \leq pF_1(u) + (1 - p)G_0(u) \quad (9)$$

(Peterson, 1976; Zhang, 2009). A nonparametric likelihood for θ based on Y is given by

$$L(\theta) = \prod_{i=1}^n [pF_1\{X_i\}]^{\Delta_i} [(1-p)G_0\{X_i\}]^{1-\Delta_i}, \quad (10)$$

which is maximized at $\hat{\theta} = (\hat{p}, \hat{G}_0, \hat{F}_1)$, where $\hat{p} = n^{-1} \sum_{i=1}^n \Delta_i$, $\hat{G}_0$ the empirical distribution of $\{X_i : \Delta_i = 0\}$, and $\hat{F}_1$ is the empirical distribution $\{X_i : \Delta_i = 1\}$. The corresponding profile likelihood for γ is just the maximum of (10) subject to (9), an equivalent expression for which

$$\begin{aligned} \tilde{l}(\gamma) &= \log \tilde{L}(\gamma) = n \log n \\ &+ \max \left\{ \sum_{i=1}^n \log w_i : w_i > 0 \forall i, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \Delta_i I(X_i \leq u) \leq \gamma \leq \sum_{i=1}^n w_i I(X_i \leq u) \right\}, \end{aligned}$$

where $I(\cdot)$ is the indicator function. Obviously, $\tilde{l}$ attains its maximum value, 0, at $\gamma \in [N_1/n, N/n]$, where

$$\begin{aligned} N_1 &= \sum_{i=1}^n \Delta_i I(X_i \leq u) = n \widehat{pF}_1(u), \\ N &= \sum_{i=1}^n I(X_i \leq u) = n \{ \widehat{pF}_1(u) + (1 - \widehat{p}) \widehat{G}_0(u) \}. \end{aligned}$$

For the other values of γ , it is easy to see that

$$\tilde{l}(\gamma) = \begin{cases} n \log n + N_1 \log \frac{\gamma}{N_1} + (n - N_1) \log \frac{1-\gamma}{n-N_1}, & 0 < \gamma < \frac{N_1}{n}; \\ n \log n + N \log \frac{\gamma}{N} + (n - N) \log \frac{1-\gamma}{n-N}, & \frac{N}{n} < \gamma < 1; \end{cases}$$

If no censoring has occurred by time u (i.e., $N_1 = N$), then $\tilde{l}$ is simply the binomial log-likelihood based on N successes out of n trials. This can again be extended to the two-sample situation along the lines of Section 6.

To illustrate this approach, let us consider a clinical study of patients with renal insufficiency, initially reported by Nahman et al. (1992). The outcome of interest in this study was the time (in months) from catheter placement to first exit-site infection, which could have been censored due to catheter failure. Figure 4 displays the profile likelihood for the probability of staying infection-free for 6 months, labelled as the 6M survival probability for brevity, in the subgroup of patients who had their catheters placed surgically.

9 Concluding Remarks

This paper presents a general profile likelihood method to represent statistical evidence with incomplete data. In contrast to standard methods for incomplete data, the proposed method does not require untestable assumptions that help identify the parameter, and provides an objective representation of evidence in the observed data alone. This is not to suggest that untestable identifying assumptions should not be made. Where appropriate, such assumptions can lead to sharpened inference and improved efficiency, as we have seen in Table 1 and Figures 1 and 3. One should be mindful, however, that such assumptions cannot be validated with the data and must be based on external information. Without total confidence in any set of identifying assumptions, it may be helpful to use the proposed method to understand “what the data say”.

As we have seen in the numerical examples, the proposed method typically yields likelihood functions that appear “flatter” than those resulting from additional assumptions. The difference becomes more apparent as the incomplete observation of data becomes more severe (with a larger proportion of missing data, e.g.). These observations should not be interpreted as an indication of inefficiency or even uselessness of the proposed method. Rather, they indicate just how little information there is in the incompletely observed data and how strong the additional assumptions really are. Imposing additional assumptions has the effect of mixing external information with the data. The resulting likelihood may appear more concentrated, but it obscures the two sources of information. The proposed method can help to clarify the contributions of different sources of information.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. The author would like to thank two anonymous referees whose comments have improved the manuscript greatly.

References

- Beresteanu A, Molinari F. Asymptotic properties for a class of partially identified models. *Econometrica* 2008;76:763–814.
- Blume JD. Likelihood methods for measuring statistical evidence. *Stat. Med* 2002;21:2563–2599. [PubMed: 12205699]
- Blume JD. How often likelihood ratios are misleading in sequential trials. *Commun. Stat. Theory Meth* 2008;37:1193–1206.
- Blume JD, Su L, Olveda RM, McGarvey ST. Statistical evidence for GLM regression parameters: A robust likelihood approach. *Stat. Med* 2007;26:2919–2936. [PubMed: 17211856]
- Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med* 1990;9:1447–1454. [PubMed: 2281232]
- Hacking, I. *Logic of Statistical Inference*. Cambridge University Press; New York: 1965.
- Hollis S. A graphical sensitivity analysis for clinical trials with nonignorable missing binary outcome. *Stat. Med* 2002;21:3823–3834. [PubMed: 12483769]
- Horowitz JL, Manski CF. Censoring of outcomes and regressors due to nonresponse: identification and estimation using weights and imputations. *J. Econometrics* 1998;84:37–58.
- Horowitz JL, Manski CF. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Amer. Statist. Assoc* 2000;95:77–81.
- Imbers GW, Manski CF. Confidence intervals for partially identified parameters. *Econometrica* 2004;72:1845–1857.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd ed.. Wiley; New York: 2002.
- Manski, CF. *Partial Identification of Probability Distributions*. Springer-Verlag; New York: 2003.
- Miettinen O, Nurminen M. Comparative analysis of two rates. *Stat. Med* 1985;4:213–226. [PubMed: 4023479]
- Nahman NS, Middendorf DF, Bay WH, McElligott R, Powell S, Anderson J. Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: clinical results in 78 patients. *J. Am. Soc. Nephrol* 1992;3:103–107. [PubMed: 1391701]
- Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 1988;75:237–249.
- Peterson AV. Bounds for a joint distribution function with fixed sub-distribution functions: applications to competing risks. *Proc. Natl. Acad. Sci. USA* 1976;73:11–13. [PubMed: 16592299]
- Royall, R. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall; Boca Raton, FL: 1997.

- Royall R. On the probability of observing misleading statistical evidence. *J. Amer. Statist. Assoc* 2000;95:760–768.
- Royall R, Tsou T-S. Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. *J. R. Stat. Soc. Ser. B Stat. Methodol* 2003;65:391–404.
- Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–592.
- Tsiatis AA. A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci. USA* 1975;72:20–22. [PubMed: 1054494]
- Zhang Z. Non-inferiority testing with a variable margin. *Biom. J* 2006;48:948–965. [PubMed: 17240654]
- Zhang Z. Likelihood-based confidence sets for partially identified parameters. *J. Statist. Plann. Inference* 2009;139:696–710.

Figure 1. Likelihood functions for the restenosis rate at 6 months after stent placement in the Obstructed Coronary Bypass Graft Trial: the proposed profile likelihood (solid line) vs. a standard binomial likelihood from a complete-case analysis (dashed line).

Figure 2. Contour plot of the joint profile likelihood for the 6-month restenosis rates in the two treatment groups in the Obstructed Coronary Bypass Graft Trial.

Figure 3. Profile likelihood functions based on PI (solid line) and the MAR assumption (dashed line) for the difference in 6-month restenosis rate (stent-angioplasty) in the Obstructed Coronary Bypass Graft Trial, with the dotted line separating regions of superiority and non-superiority.

Figure 4. Profile likelihood for the 6-month survival probability in patients with surgically placed catheters in the study of Nahman et al. (1992).

Empirical comparison of the proposed method based on PI with a SA approach in terms of the frequency distribution of LRs for comparing the true value $\gamma_0 = 0.5$ against $\gamma_1 = 0.3$ in the setting of the proportion example. Here $p_0 = P(R = 0)$ is the proportion of missing data, β_{true} is the true value of β , β_{guess} is the value used to implement SA, and $\Gamma|_{\theta_0}$ is the identification interval for γ . See Section 5 for details.

Table 1

p_0	β_{true}	$\Gamma _{\theta_0}$	Method	β_{guess}	Freq. Dist. (%) of LRs for $\gamma_0 = 0.5$ vs. $\gamma_1 = 0.3$							
					0-32-1	32-1-8-1	8-1-8	8-32	32+			
0.1	0	[0.45, 0.55]	SA	0	0.1	0.5	6.5	5.4	87.5			
				1	1.1	0.9	13.2	7.7	77.1			
				-1	0.0	0.1	2.7	2.9	94.3			
	1	[0.47, 0.57]	SA	0	0.1	0.0	1.9	3.2	94.8			
				1	0.2	0.3	5.3	4.9	89.3			
				-1	0.0	0.1	0.5	0.7	98.7			
0.3	0	[0.43, 0.53]	SA	0	0.7	1.1	15.0	8.8	74.4			
				1	2.6	2.3	22.0	13.1	60.0			
				-1	0.3	0.3	7.1	6.2	86.1			
	1	[0.35, 0.65]	SA	0	0.4	0.6	12.9	9.8	76.3			
				1	6.6	5.7	36.4	14.0	37.3			
				-1	0.0	0.0	1.1	1.4	97.5			
	1	[0.40, 0.70]	SA	0	0.0	0.1	1.4	1.2	97.3			
				1	0.5	0.5	10.6	8.3	80.1			
				-1	0.0	0.0	0.1	0.2	99.7			
	-1	[0.30, 0.60]	SA	0	6.7	6.7	38.3	15.0	33.3			
				1	33.7	13.7	39.8	6.3	6.5			
				-1	0.4	0.4	12.2	9.8	77.2			