STATISTICS IN MEDICINE, VOL. 15, 1333-1348 (1996)

A TECHNIQUE FOR MEASURING EPIDEMIOLOGICALLY USEFUL FEATURES OF BIRTHWEIGHT DISTRIBUTIONS

DAVID M. UMBACH

Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709, U.S.A.

AND

ALLEN J. WILCOX

Epidemiology Branch, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709, U.S.A.

SUMMARY

Birthweight distributions have been conceptualized as a predominant Gaussian distribution contaminated in the tails by an unspecified 'residual' distribution. Acknowledging this idea, we propose a technique for measuring certain features of birthweight distributions useful to epidemiologists: the mean and variance of the predominant distribution; the proportions of births in the low- and high-birthweight residual distributions, and the boundaries of support for these residual distributions. Our technique, based on an underlying multinomial sampling distribution, involves estimating parameters in a mixture model for the multinomial bin probabilities after having chosen the support of the residual distribution with a model selection criterion. A modest simulation study and experience with a few actual datasets indicate that use of a Bayesian information criterion (BIC) as model selection criterion is superior to use of Akaike's information criterion (AIC) in this application.

1. INTRODUCTION

Birthweight is strongly associated with infant survival. Since infant mortality is a rare event and birthweight data are plentiful (collected routinely as a vital statistic), birthweight is frequently analysed as a surrogate for infant well-being.

National registries of births represent a heterogenous collection of sexes, ethnic affinities, gestational ages and maternal parities. The underlying birthweight distribution is likely a complex mixture of many components. The general shape of birthweight distributions, however, is remarkably consistent from population to population. Conditional distributions of birthweight given different gestational ages, for example, have qualitatively similar shapes,[1] shapes that are also similar to that of the marginal birthweight distribution. Generally speaking, these distributions are mound-shaped with heavier tails than a normal distribution. Most often, excess births appear in the lower tail skewing the distribution, but they may occasionally also appear in the upper tail.

Seeing this general shape and focusing on extra births at low birthweights, some epidemiologists[2-4] have promoted a simplified conceptualization of birthweight distributions as consisting

CCC 0277-6715/96/131333-16 *Received July 1994* *Revised July 1995*

of two overlapping components: a predominant normal distribution for most (usually 95 per cent or more) of births and a 'residual' distribution that consists of excess low birthweights. While these two components may oversimplify the biology underlying the birthweight distribution, epidemiologists find the description useful. They typically summarize the two components by the mean and standard deviation of the predominant distribution and the fraction of live births in the residual distribution. One can use the parameters of the predominant distribution, which approximates the distribution of term births ($\geqslant 37$ week gestational age),[4] to standardize weight-specific perinatal mortality rates, and thus provide a convenient way to compare such rates among populations.[5,6] Knowing that low birthweight is associated with high risk of mortality, epidemiologists view the low-birthweight residual distribution as containing those births at particular risk for adverse outcomes.[5] Consequently, one can use the fraction of births in this residual distribution as a measure of risk for any population of births, even when information on infant survival is unavailable. The fraction of births in the residual distribution corresponds approximately to the fraction of preterm births (<37 week gestational age) weighing less than 2500 g.[4] Of course, summaries provided by the concepts of predominant and residual distributions are available even when information on gestational ages of births is lacking. Although the idea of a residual distribution originally focused on the lower tail, certain birthweight distributions may have excess births in the upper tail;[1] however, this upper-tail residual is not viewed as an index of risk.

The conceptualization of birthweight distributions as having predominant and residual components is a qualitative description, inherently imprecise, but its full use requires assignment of numerical values to features of that description. The term 'residual' derives from the notion of births remaining in the lower tail after having fit a predominant distribution to the data. In effect, how one determines the predominant distribution *defines* both the predominant and residual distributions. Kiely and Kleinman[6] determine the predominant distribution by fitting a truncated normal distribution to the birthweights that exceed a prespecified cut-off of 2500 g. The fraction of all birthweights in the residual distribution is the proportion of all births falling below the cut-off minus the proportion of births from the predominant distribution expected to fall below the cut-off. Wilcox and Russell[4] see the fixed cut-off, while convenient, as a potential distortion. Just as the location of the predominant distribution may change among populations, the cutpoint marking the upper bound of the residual distribution may also change. They determine the predominant distribution by fitting left-truncated normal distributions to a sequence of decreasing cut-offs, choosing one based on an index of model adequacy.

Despite the qualitative nature of the basic conceptualization, each of these methods provides a different but well-defined way to 'measure' birthweight distributions – to assign values to three features of birthweight distributions useful to epidemiologists namely the predominant mean, the predominant standard deviation, and the fraction of births in the lower-tail residual. Both existing methods, however, have shortcomings. Neither is designed to measure excess births in the upper tail of the birthweight distribution. Moreover, when the upper tail has excess births, both methods will present a distorted picture, tending to inflate the predominant standard deviation and to reduce the fraction of births in the lower-tail residual. Finally, both methods are inefficient because they determine the predominant distribution with data only from weight categories regarded as uncontaminated by the residual distribution.

Our purpose is to provide an alternative way to measure epidemiologically useful features of birthweight distributions that explicitly allows excess births in either (or both) of the tails and that uses information from every weight category to determine the predominant distribution. Although we derive our approach as an estimation procedure for a statistical model, we regard the measurement process as different in spirit from what statisticians usually mean by estimation. In particular, parameters are well-defined entities in a specific model. Our measurement technique is an attempt to quantify the qualitatively described ideas of predominant and residual birthweight distributions by choosing a model from a family of models that embody the qualitative description. In essence, we define the component distributions through the measurement procedure. While we use the terminology of estimation as a convenient manner of speaking, we view these 'estimates' as 'measurements'.

Since birthweight data usually appear grouped into weight classes, we exploit the multinomial nature of the data and model bin probabilities as a mixture of a normal distribution and an unspecified multinomial distribution on a subset of the bins. The two components of the mixture represent the predominant distribution and the residual distribution, respectively. Our technique examines a sequence of subsets of bins as support for the residual distribution and chooses the particular subset that optimizes a model selection criterion; then, given the chosen support set, we maximize the mixture likelihood to estimate parameters. So that the model chosen conforms qualitatively to the epidemiologists' ideal, we constrain the residual component to have support only in the tails. Our technique measures the parameters of a predominant distribution, the fractions of residual births in both the lower and upper tails, and the boundaries of support of the lower- and upper-tail residual components.

The next section presents both the family of multinomial models that form the basis of the measurement technique and the procedure used to construct a sequence of support sets (models) for the residual distribution. It also describes how to select a particular model from the sequence. Section 3 contains methods for estimating parameters in these models and an approach to variance estimation. Section 4 elaborates on questions of model selection and variance estimation based on a small simulation study. Results from actual birthweight data appear in Section 5. The final section contains a discussion.

2. A FAMILY OF MULTINOMIAL MODELS

2.1. General Description of the Family

Because birthweight data are usually grouped into weight classes, we assume a multinomial model to describe the sampling distribution. Of course, if the original data are ungrouped or if the available grouping is regarded as too fine, one could impose some coarser grouping. Although several births to the same parents may appear in a set of data, we assume that obvious dependencies such as twins or other multiple births have been purged and will be treated separately; hence we take the observations as independent. A typical set of birthweight data might have 100 g bins with cutpoints ranging from 1000 g to 5700 g and might contain from several hundred to 100,000 or even more births.

Assume that we have $k + 1$ bins formed by k cutpoints where we regard the lowest and highest bins as unbounded to the left and right, respectively. We model bin probabilities as a mixture of two components: a predominant distribution that contributes to all bins and an unspecified residual distribution. We take the predominant distribution as normal but we could use any continuous univariate distribution. We view the residual distribution as 'contaminating' the predominant distribution and adopt corresponding terminology; in particular, contaminated bins intersect the support of the residual component and uncontaminated bins are disjoint from its support. Conceivably, every bin could be contaminated, but identifiability considerations restrict the degrees of freedom available for parameterizing the residual component. In fact, with $k + 1$ bins, estimation of the mean and standard deviation of the normal component along with the constraint that bin probabilities sum to unity leaves at most $k - 2$ degrees of freedom for the residual component. A straightforward way to honour this restriction is to model this component with a single parameter for each contaminated bin and to enforce the presence of at least three uncontaminated bins.

Let $\pi = (\pi_1, \pi_2, \pi_3, \dots, \pi_{k+1})$ be the vector of overall bin probabilities with $\Sigma\pi_i = 1$. Let $B = \{-\infty = b_0, b_1, b_2, b_3, \dots, b_k, b_{k+1} = \infty\}$ be the set of cutpoints augmented on either end to accommodate unbounded bins, $I = \{1, 2, 3, \dots, k+1\}$ be the set of bin indices, and $C \subset I$ be the set of indices of contaminated bins (C has at least three fewer elements than I). A mixture model for the overall bin probabilities π_i is

$$\pi_i = (1-\alpha)\Delta_i + \alpha\delta_i \text{ whenever } i \in I \qquad (1)$$

where α (the mixing proportion) is the probability of being from the residual component, δ_i is the conditional probability of being in bin i given membership in the residual component ($\delta_i \equiv 0$ whenever $i \notin C$), and Δ_i is the conditional probability of falling in bin i given membership in the predominant normal component with mean μ and standard deviation σ. Explicitly,

$$\Delta_i = \Phi\left(\frac{b_i - \mu}{\sigma}\right) - \Phi\left(\frac{b_{i-1} - \mu}{\sigma}\right) \qquad (2)$$

with $\Phi(\cdot)$ representing the standard normal cumulative distribution function. Of course, $\Sigma\delta_i = \Sigma\Delta_i = 1$. We employ a convenient reparameterization of (1):

$$\pi_i = \left(1 - \sum_{j \in C} \theta_j\right)\Delta_i + \theta_i \text{ whenever } i \in I \qquad (3)$$

where $\theta_i (= \alpha\delta_i)$ is the joint probability that a birthweight falls both in bin i and in the residual component ($\theta_i \equiv 0$ whenever $i \notin C$) and $\Sigma\theta_i (= \alpha)$ gives the probability of membership in the residual component.

A particular set of contaminated bins (or, equivalently, a particular set C) specifies what we term a model. Let $\omega = (\mu, \sigma, \boldsymbol{\theta}_C)$ where $\boldsymbol{\theta}_C$ is the vector of θ_i with indices in C. For a model with c contaminated bins, the parameter vector ω has $c + 2$ components because we have omitted those θ_i assumed to be zero. Although the fit the full parameter vector ω to use information in the contaminated bins in estimating the normal parameters, in fact interest is usually in a smaller parameter vector $(\mu, \sigma, \upsilon_L, \upsilon_H)$ where υ_L is the total residual among low birthweights and υ_H is the total residual among high birthweights. These two parameters are simply sums of θ_i over all contaminated bins in the lower and upper tails, respectively.

We could allocate contaminations to $0, 1, 2, \dots$, or $k - 2$ of the $k + 1$ available bins so the number of possible configurations of contaminated bins in the family grows rapidly as the number of bins increases. Thus, searching the entire family for models that adequately describe the data is completely impractical for the number of bins commonly encountered. More importantly, searching the entire family would completely ignore the conceptualization we seek to impose on the measurement process.

2.2. Search Strategy: A data-driven sequence of models within the Family

We confined searches to a data-dependent sequence of models designed to embody epidemiologists' concept of the structure of the distribution. In particular, we sought to have the predominant component in the 'centre' with a residual component in the 'tails' representing a small fraction (say, fewer than 10 per cent) of births. Our strategy is to fit the entire sequence of models and use the results to select a particular one. Among many conceivable rules for defining sequences, we focused on one.

The rule that we implemented begins with the model having no bins contaminated, the pure normal model, and proceeds to add one bin at a time to the set of contaminated bins until all but three bins contribute to the residual component. Thus, our earlier models are nested within later ones. At any stage, we choose the next bin to add based on how far that bin's observed count deviates from its expected count under the current model. Our rule adds the next bin by choosing from the two outermost uncontaminated bins the one with the larger Pearson residual (one could use other measures of discrepancy). (At some point, both residuals may be negative, but one can still make the choice.) Under this rule, a residual component can appear in the lower tail, upper tail, or both, and the uncontaminated bins within each tail are contiguous at every step. Generally, throughout most of this sequence, the fitted models correspond qualitatively to the epidemiologists' concept. Near the end of the sequence, however, fitted models can deviate by having residual components with most of the mass while the normal component has a small variance and essentially fills the last few uncontaminated bins.

2.3. Selecting a Particular Model from the Sequence

The sequence provides a list of models from which we select one. We considered two well-known selection criteria based on penalized likelihoods: AIC, Akaike's information criterion;[7] and BIC, a Bayesian information criterion.[8] Both criteria have basically the same form: the maximum value of the likelihood function minus a penalty that depends on the number of fitted parameters in the model. For AIC, the penalty is simply the number of fitted parameters; for BIC, the penalty is the number of fitted parameters times half the natural logarithm of the total number of observations. Consequently, for any sample size larger than eight, BIC favours models with fewer parameters than AIC. Based on results from both asymptotic theory and small sample simulations, whenever the correct model is among the set of candidates, AIC applied to normal-theory regression or to certain time-series models tends to overfit – namely, to choose a model with more parameters than the correct model.[9,10] BIC chooses a more parsimonious model than AIC and may be preferred on that account. However, a trade-off between overfitting and underfitting always exists. Since we view our technique as a way to define operationally concepts described qualitatively, our concern is primarily with the repeatability of the measurement process rather than with selecting a 'correct' model. In Section 4, we report results from a limited simulation study that examines these two selection criteria with data that meet the structure that we assume for birthweight distributions.

To select a model is to decide which bins are contaminated. In particular, this choice determines two additional quantities of interest, b_L and b_H, the upper boundary of support for the low birthweight residual and the lower boundary of support for the high birthweight residual, respectively. We take as estimates of b_L and b_H, which are not explicitly part of ω, the corresponding bin boundaries for the selected model.

3. ESTIMATION OF MODEL PARAMETERS

3.1. Point Estimates for a Particular Model

For any particular model we use maximum likelihood to estimate parameters. In our models, however, maxima may occur at the boundary of the parameter space; individual θ_i parameters cannot be less than zero but are expected to be small, perhaps even equal to zero. To stay within the boundaries, we employed the EM algorithm[11] for maximizing the likelihood function.

Let $\mathbf{f} = (f_1, f_2, f_3, \ldots, f_{k+1})$ be the vector of observed bin counts. Denote the logarithm of the multinomial likelihood function for the observed data as $l(\omega | \mathbf{f}, B) \equiv l(\omega)$. Then, writing $\boldsymbol{\pi}(\omega)$ for $\boldsymbol{\pi}$ as given by (3) to emphasize the dependence of the bin probabilities on the model parameters

$$l(\omega) = \sum_{i=1}^{k+1} f_i \log \pi_i(\omega) + h(\mathbf{f}) \tag{4}$$

where $h(\mathbf{f})$ is constant in ω.

As a first step in maximizing $l(\omega)$ via the EM algorithm, one imagines having more information than is actually available and constructs a likelihood reflecting this additional information. In our case, suppose that, instead of counts in $k + 1$ bins that represent a mixture of predominant and residual components, we have augmented data, namely, a $2 \times (k + 1)$ array of bins where the first row contains counts from the predominant distribution and the second row contains counts from the residual distribution. Denote the corresponding bin counts by z_{ij} with z_1 and z_2 denoting the vectors of counts in the predominant and residual rows, respectively. Let p_{ij} be the corresponding bin probabilities with $\mathbf{p}_1(\omega)$ and $\mathbf{p}_2(\omega)$ denoting the vectors of bin probabilities associated with the predominant and residual rows, respectively. Of course, since the rows of the augmented table must add to the original table, $\mathbf{z}_1 + \mathbf{z}_2 = \mathbf{f}$, and $\mathbf{p}_1(\omega) + \mathbf{p}_2(\omega) = \boldsymbol{\pi}(\omega)$. Moreover, we chose the augmented data to reflect the two terms in equation (3) so that, using symbols defined previously,

$$p_{1j} = p_{1j}(\omega) = \left(1 - \sum_{i \in C} \theta_i\right) \Delta_j \text{ whenever } j \in I \tag{5}$$

and

$$p_{2j} = p_{2j}(\omega) = \theta_j \text{ whenever } j \in I. \tag{6}$$

Of course, the restriction that $\theta_j \equiv 0$ whenever bin j is uncontaminated is enforced (we can take those particular θ_j as known). Denote the logarithm of the multinomial likelihood function for the augmented data as $l^*(\omega | \mathbf{z}_1, \mathbf{z}_2, B)$ where (taking $0 \log 0 \equiv 0$)

$$l^*(\omega | \mathbf{z}_1, \mathbf{z}_2, B) = \sum_{i=1}^{2} \sum_{j=1}^{k+1} \mathbf{z}_{ij} \log p_{ij}(\omega) + h(\mathbf{z}_1, \mathbf{z}_2) \tag{7}$$

and $h(\mathbf{z}_1, \mathbf{z}_2)$ is constant in ω. Starting from $\hat{\omega}^{(0)}$, an initial estimate of the parameter vector, the EM algorithm generates a sequence $\{\hat{\omega}^{(s)}\}$ of parameter estimates by alternating between two steps, the E-step and the M-step.

In the problem at hand, the E-step computes $\mathbf{z}^{(s)} = E(\mathbf{Z} | \mathbf{f}, \hat{\omega}^{(s)})$ where $\hat{\omega}^{(s)}$ is the current estimate and $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$. That is, we allocate the observed counts to the $2 \times (k + 1)$ bins in the augmented data according to their expected values under the current parameter estimates. In particular,

$$z_{ij}^{(s)} = \frac{\hat{p}_{ij}^{(s)}}{\hat{\pi}_j^{(s)}} f_j \quad \text{for } j \in I,\ i \in \{1,2\}. \tag{8}$$

The M-step finds $\hat{\omega}^{(s+1)}$ by maximizing the augmented data log-likelihood of (7) with respect to ω. Ignoring the parameter-free term and using (5) and (6), we can re-express (7) as

$$\begin{aligned} l^*(\omega | z_1, z_2, B) &= \sum_{j=1}^{k+1} z_{1j} \log(1 - \Sigma\theta_i)\Delta_j + \sum_{j=1}^{k+1} z_{2j} \log \theta_j \\ &= \sum_{j=1}^{k+1} z_{1j} \log \Delta_j + \left[\sum_{j=1}^{k+1} z_{2j} \log \theta_j + n_1 \log(1 - \Sigma\theta_i)\right] \end{aligned} \tag{9}$$

where $n_1 = \Sigma z_{1j}$, the number of births in the predominant component. From the second line of (9), one can see that this maximization decomposes into two operations: estimation of $\boldsymbol{\theta}_C$, and estimation of μ and σ (the determinants of Δ_j). Let $n = \Sigma f_i$, the total number of observed births. Maximizing the portion of (9) in square brackets yields

$$\hat{\theta}_j^{(s+1)} = \hat{p}_{2j}^{(s+1)} = \frac{z_{2j}^{(s)}}{n} \quad \text{for } j \in I. \tag{10}$$

Maximizing the term in (9) involving Δ_j for $\hat{\mu}$ and $\hat{\sigma}$ involves fitting a normal distribution to grouped data, accomplished easily using the method of scoring.[12]

Because the EM algorithm tends to converge much more slowly than the Newton–Raphson method, accurate specification of starting values is advantageous. For any fixed model, however, *a priori* specification of reasonable starting values for the θ_i parameters is difficult. One effective way to specify starting values is to base them on the maximum likelihood estimates of μ and σ under a truncated normal model fitted to the uncontaminated bins. The corresponding estimator of θ_i in a contaminated bin is

$$\hat{\theta}_i^* = \frac{f_i}{n} - \frac{f_U}{n} \frac{\hat{\Delta}_i^*}{\hat{\Delta}_U^*} \tag{11}$$

where superscript '*' denotes estimates based on a truncated normal distribution fitted to uncontaminated bins and the subscript U indicates summation over all uncontaminated bins. Again, $0 \leqslant \hat{\theta}_i^* \leqslant 1$ for $i \in C$ is *not* guaranteed; the truncated normal estimates of ω may be outside Ω. Whenever the truncated normal estimates are within Ω, they do coincide with the maximum likelihood estimates. (The estimates based on the truncated normal fit solve the score equations for the full likelihood. Whenever the solutions to these score equations are within Ω, they are the maximum likelihood estimates.) Estimation via the truncated normal scheme is much faster than via the EM algorithm, so that calculation of truncated normal estimates for every model in the sequence and omitting EM calculations for models whose truncated normal estimates honour the boundaries of Ω saves time. When the truncated normal estimates transgress the boundaries, they usually do so in multiple bins. In those cases, one needs other starting values. Since one fits models in sequence, a second effective way to specify starting values for a particular model is to use the final fitted values $\hat{\omega}$ from the previous model in the sequence with θ_i for the single newly contaminated bin estimated as $\max((f_i - \hat{f}_i)/n, \varepsilon)$ where $\hat{f}_i$ denotes the fitted value under the previous model in the sequence and ε is a small positive quantity. The idea here is that when $(f_i - \hat{f}_i)/n$ is positive, it should provide a good estimate of θ_i; however, if it is not positive, one should start θ_i within the parameter space.

Several features of these fitted models merit comment. First, one might think that each model must fit the data perfectly in contaminated bins; however, this is not necessarily so. The normal distribution that fits best throughout the entire data set may allow an observed count in a bin with an estimated zero contamination to be below that expected from the normal component. This phenomenon would seem to require strong adherence to the fitted normal model in the uncontaminated cells. Second, if no bins are empty, the fully parameterized model will provide an exact fit to the data unless the data configuration in the three uncontaminated bins precludes an exact fit by any normal distribution. For example, assuming bins of equal width, if one sees fewer births in the middle bin than in either of its neighbours, no normal distribution can provide an exact fit. On the other hand, if some bins have observed counts of zero, the fully parameterized model will not provide an exact fit to the data because the predominant normal component would never predict a zero probability for any bin. We elected to remove zeros by combining any empty bins with adjacent bins, arbitrarily joining bins with zero counts to the first non-zero bin 'outside' them, that is, away from the median of the sample (we joined zero-count bins outside all non-zero bins to be endmost non-zero bin).

3.2. Assessing Variability in Parameter Estimates

Because the data often represent a complete list of the birthweights in a certain geographic region within a prescribed time period, some may argue that the birthweights at hand are the entire population, not a sample, and estimates of variability are unnecessary. We believe that most investigators would wish to assess variability under the idea that the data at hand represent a sample from a hypothetical population. Two features complicate estimation of variability in the parameter estimates. First, some of the θ_i parameters may lie at the boundary of the parameter space – a situation where the asymptotic distribution of the MLE is not the usual multivariate normal distribution,[13] and variances derived from the information matrix may not be useful. One can construct confidence regions based on the likelihood function.[14] but these are difficult to implement for high dimensional parameters. Second, and more importantly, the model selection process introduces extra variability.

We recommend resampling methods to assess variability. If one resamples the data vector with replacement and recreates the entire model selection process for each bootstrap sample, one can generate a list of estimates of $(\mu, \sigma, \upsilon_L, \upsilon_H)$ as well as b_L and b_H, one from each bootstrap sample. Since each entry in this list could come from a different selected model, variability includes the contribution of the model selection process. Non-parametric bootstrapping also copes with estimates at the boundary. However, bootstrapping the entire selection process can be time consuming, especially as the number of bins grows.

4. RESULTS FOR SIMULATED DATA

To evaluate our method's ability to measure features of a known distribution that explicitly meets the assumptions for which the method was designed, we employed a modest simulation study to examine a single underlying population at seven sample sizes (1, 5, 25, 50, 75, 100, and 125 thousand births) and three different bin widths (100, 200 and 500 g). We programmed the simulations and the fitting procedure in GAUSS (Aptech Systems, Inc., Maple Valley, WA) using the random number generator supplied with the system. For each sample size and bin width, we generated 100 independent samples from the underlying distribution. For each sample, we estimated parameters using the entire model selection procedure and recorded estimates of the reduced parameter vector $(\mu, \sigma, \upsilon_L, \upsilon_H)$ as well as b_L and b_H for both the AIC- and BIC-selected models. We report in greater detail results from 100 g bins, pointing out the few differences that arise with the other bin widths.

We assumed that birthweights were recorded in 45 100 g bins plus two end bins of indefinite length, the lower one having upper bound 1100 g and the upper one having lower bound 5600 g. The predominant distribution representing 97 per cent of the population had $\mu = 3500$ and $\sigma = 435$, approximating what one might expected from an actual birthweight distribution. The residual distributions were more arbitrary: we assigned the low-birthweight contamination that contained 2 per cent of the population equally to the 16 lowermost bins ($b_L = 2600$ g) and the high-birthweight contamination that contained 1 per cent of the population equally to each of the 10 uppermost bins ($b_H = 4700$ g). These proportions and boundaries are reasonable for actual data, and, whereas strict uniformity seems unlikely, what simple shape might be more realistic is unclear. Although the high-birthweight contamination was smaller in total than the low-birthweight

Table I. Performance of the model selection procedure in identifying the correct set of contaminated bins when facing 100 g bins

Sample size	Number of samples out of 100							
	Selected set coincides with correct set		Selected set is a subset of the correct set		Selected set contains the correct set		Selected set and correct set are not nested*	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
1,000	1	0	95	100	1	0	3	0
5,000	6	0	84	100	6	0	4	0
25,000	50	0	30	100	18	0	2	0
50,000	70	15	10	85	19	0	1	0
75,000	77	36	1	64	22	0	0	0
100,000	77	69	0	31	23	0	0	0
125,000	79	79	1	21	20	0	0	0

* In one tail, the selected bins are a subset of the correct set; in the other tail, the correct set is a subset of those selected

contamination, its support began farther from the predominant mean; consequently, we regard the upper-tail contamination as easier to detect. For example, in the uppermost bin of lower-tail residual support, the odds of coming from the contaminating distribution are about 0·15 whereas the corresponding odds are 0·69 in the lowermost bin of upper-tail residual support.

For the other two bin widths, we used the same underlying distribution with different bin boundaries. For 200 g bins, we recorded the data in 22 200 g bins plus two end bins of indefinite length, the lower one having upper bound 1200 g and the upper one having lower bound 5600 g. In this configuration, the lower-tail residual was uniform in 8 bins ($b_L = 2600$ g); however, the upper-tail residual in 6 bins now has $b_H = 4600$ g and is no longer strictly uniform since the new bin boundaries straddle the former b_H (consequently, the first and last bins in the upper residual distribution have half the contamination of the other four). For 500 g bins, we recorded the data in eight 500 g bins plus two end bins of indefinite length, the lower one having upper bound 1500 g and the upper one having lower bound 5500 g. With this bin configuration, neither the upper- nor the lower-tail residual is uniform. In particular, a birth from the bin with upper bound 3000 g has odds of contamination of about 0·01 – making the correct bin boundary difficult to detect. Moreover, the full model with only 3 uncontaminated bins is the correct model for our 500 g bin configuration.

One key for understanding the performance of the model sequence and the model selection procedures is how well they can identify the boundaries of support of the low- and high-birthweight residual components, b_L and b_H, respectively. If a procedure tends to choose these boundaries correctly, that is, to choose the correct set of contaminated bins, it should estimate the parameter vector with little bias. If a procedure, however, tends to select models whose set of contaminated bins is a subset of the correct set, that is, to underfit, then the procedure will tend to underestimate the proportion of births in the residual components and the overestimate the variance of the predominant component. The opposite biases will accrue if a procedure tends to overfit. The simulations that used 100 g bins showed that both AIC and BIC improved their ability to identify correctly both b_L and b_H as sample size increased (Table I) with AIC choosing the correct model more often than BIC except at the largest sample size. The ability of AIC to identify correctly both b_L and b_H was virtually the same with 200 g bins; however, BIC performed

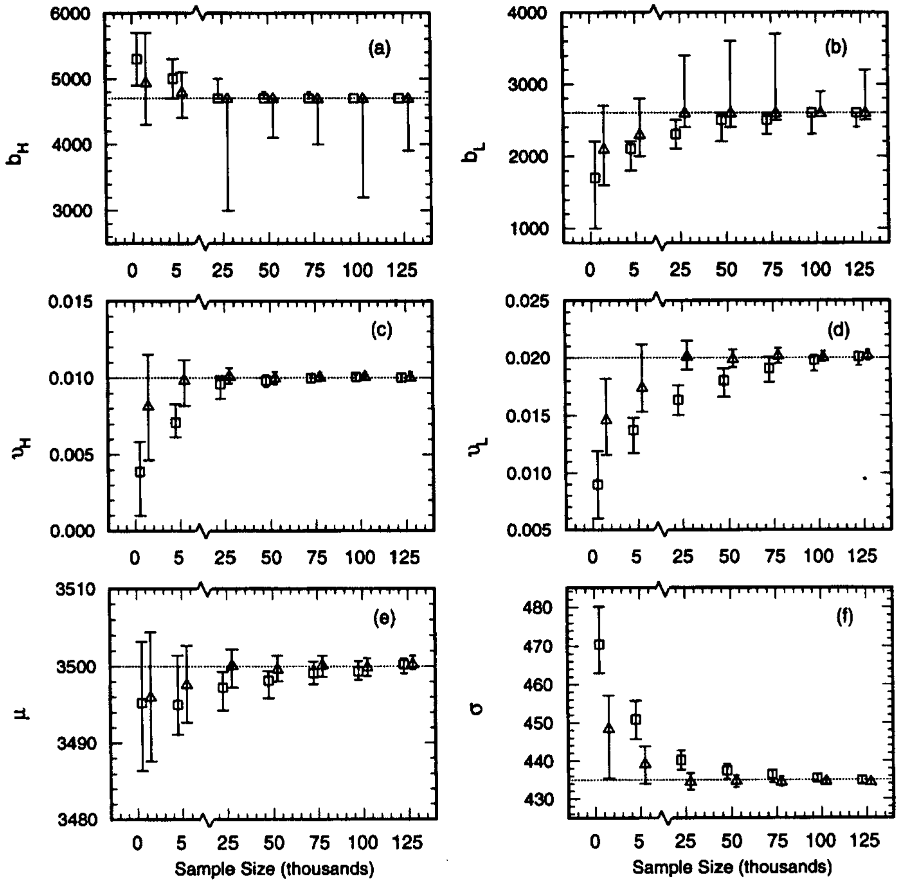

Figure 1. Parameter estimates based on simulated samples using 100 g bins from a known birthweight distribution plotted as a function of sample size. Plotted points, horizontally offset at each sample size so error bars do not overlap, are the medians of 100 estimates with error bars to the extremes in panels (*a*) and (*b*) and to the quartiles in the remaining panels. Triangles represent AIC-selected estimates; squares, BIC-selected estimates; horizontal dotted line marks the true value in the population. (*a*) lower bound on high-birthweight residual's support, b_H; (*b*) upper bound on low-birthweight residual's support, b_L; (*c*) proportion of population in high-birthweight residual, υ_H; (*d*) proportion of population in low-birthweight residual, υ_L; (*e*) mean of predominant normal distribution, μ; and (*f*) standard deviation of predominant normal distribution, σ

a bit better with 200 g bins at sample sizes of 25,000 births or more, correctly identifying both boundaries in about 10 more simulated samples at each of those sample sizes than with 100 g bins (results for 200 g bins not shown). With 500 g bins, neither procedure could identify the correct set of bins in more than 20 per cent of the simulated samples, but AIC did better than BIC (results for 500 g bins not shown). Both procedures became more accurate as sample sizes increased except with 500 g bins where neither performed well. At smaller sample sizes, whenever either selection procedure erred at any bin width, it tended to underfit. At larger sample sizes, BIC continued to underfit when erring whereas AIC tended to overfit except with 500 g bins where the configuration prohibited overfitting (see Table I for 100 g data). In our simulations, we never observed BIC to overfit at any sample size or bin width.

Focus first on estimation of b_H using 100 g bins (Figure 1(*a*)). Whenever the sample size exceeded 25,000 births, both AIC- and BIC-selected models appeared to provide a median-unbiased estimator of this boundary, that is, the median of the 100 simulated boundary estimates coincided with the true boundary. However, the range of the BIC estimates was strikingly smaller than that of the AIC estimates. Also, whereas the BIC-selected models rarely underestimated but could moderately overestimate the boundary, the AIC-selected models rarely overestimated but could severely underestimate the boundary. Whenever the sample size was 1000 or 5000 births, both selection procedures tended to overestimate b_H (BIC doing somewhat worse) but the ranges of the estimates were more nearly equal for both selection procedures. The performance of AIC and BIC with respect to estimation of b_L was, for the most part, a mirror image of that for b_H (Figure 1(*b*)). One notable difference was that the median of the BIC estimates of b_L did not coincide with the true value until the sample size reached 100,000 births. Thus, estimators of b_L and b_H based on the BIC-selected model, while more biased at smaller sample sizes, tended to be more stable than their AIC counterparts as the sample size grew. This same generalization held for the simulations using the 200 g bins (data not shown); in fact, 200 g bin width made little difference in the procedures' abilities to estimate b_L and b_H except that the BIC estimates based on the wider bins appeared somewhat less biased at smaller sample sizes than BIC estimates based on the narrower bins. At the 500 g bin width, medians of estimates of b_L from both AIC and BIC differed from the correct value at all sample sizes; however, estimates of b_H had medians close to the correct value for both AIC and BIC at sample sizes larger than 5000 births.

The general observation that bias decreased as sample size increased held for the other parameters at each bin width as well. The variability in the AIC estimates and the BIC estimates from 100 g bins in Figures 1(*c*)–(*f*) appears more nearly equal than it does in Figures (*a*) and (*b*), but this appearance is largely an artifact of showing interquartile range instead of overall range. When the number of births exceeded 5000 and bin width was 100 g. AIC-based estimates from a small proportion of samples were quite out of line with estimates from the others while BIC-based estimates did not show this tendency. For example, at 100,000 births, AIC estimates of υ_H ranged from 0·0092 to 0·58 (the second highest estimate was 0·0119) while BIC estimates ranged from 0·0092 to 0·0109. The collection of BIC estimates for each parameter showed no such extreme outliers at any number of births whereas AIC estimates showed one or more for several parameters when the number of births exceeded 5000 (although individual outliers were only rarely as egregious as the example). Such outliers arose when the AIC-selected model was at or near the end of the model sequence, and the chosen model no longer matched the conceptualization that we intended to capture.

The bias of AIC estimates at each sample size was similar whether we used 200 g or 100 g bins. In addition, with 200 g bins, the tendency for AIC occasionally to deliver extreme outliers, while still present, was somewhat weaker. BIC estimates based on 200 g bins were perhaps slightly less biased than their 100 g counterparts, and BIC showed no extreme outliers. With 500 g bins, both AIC and BIC underestimated the lower-tail residual fraction, and, to a smaller degree, the predominant mean – underestimation that persisted even with sample sizes of 125,000 births. The upper-tail residual fraction and the predominant standard deviation estimated from 500 g bins showed declining bias similar to that seen using other bin widths. The greatest benefit of the larger bin width (fewer bins in total) was the marked saving in computing time.

Based on these simulations, we might prefer AIC-selected estimates when the number of births is 5000 or fewer since AIC appears less biased and both methods seem to have comparable variability at these sample sizes (but see Section 5). For samples of more than 5000 births, we would prefer BIC-selected estimates which, while somewhat biased, are not subject to extreme variability. The 200 g bin width is preferred because of time savings and a slight reduction of bias

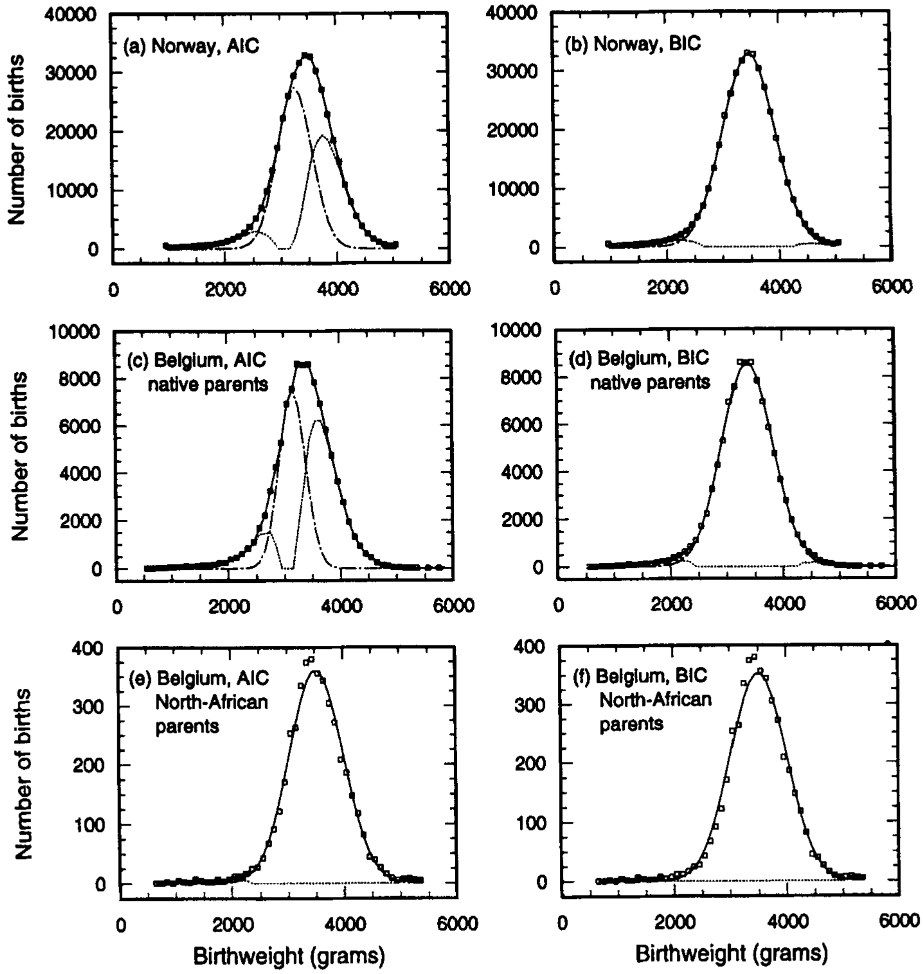

Figure 2. Actual birthweight data and frequency polygons for fitted models based on 100 g bins. Squares are observed birthweight data plotted at bin midpoints; alternating dash-dot line is fitted predominant normal distribution; pecked line is fitted residual distribution; solid line is fitted birthweight distribution = sum of predominant and residual. (*a*) AIC-selected model for Norwegian data; (*b*) BIC-selected model for Norwegian data; (*c*) AIC-selected model for the native-parent Belgian data, (*d*) BIC-selected model for the native-parent Belgian data; (*e*) AIC-selected model for the North-African-parent Belgian data; (*f*) BIC-selected model for the North-African-parent Belgian data

with BIC. The coarse grid of 500 g bins greatly speeds processing at the cost of some bias. We must, however, temper these conclusions since they are based on simulations from only a single underlying population.

5. RESULTS FOR ACTUAL BIRTHWEIGHT DATA

We applied our measuring technique to three sets of birthweight measurements, one from the Medical Birth Registry of Norway and two based on provisional data from the Belgian Ministry of Health. The data from Norway consisted of 394,386 singleton live and still births from 1967 to

Table II. Results from applying the proposed measurement technique to actual birthweight data in 100 g bins

Parameter	Parameter estimates (standard errors)*					
	Norway		Belgium: native-born parents		Belgium: North-African-born parents	
	AIC	BIC	AIC	BIC	AIC	BIC
μ	3252	3465	3146	3370	3513	3505
	(113)	(1·4)	(82)	(2·6)	(230)	(8·8)
σ	311	460	229	447	478	494
	(60)	(2·0)	(45)	(5·0)	(126)	(9·4)
υ_L	0·065	0·030	0·102	0·024	0·015	0·009
	(0·051)	(0·001)	(0·038)	(0·003)	(0·182)	(0·003)
υ_H	0·389	0·008	0·458	0·008	0·005	0·004
	(0·170)	(0·002)	(0·120)	(0·004)	(0·237)	(0·002)
b_L	2900	2600	2900	2400	2300	1800
	(198)	(14·1)	(115)	(79)	(454)	(200)
b_H	3200	4300	3200	4300	5000	5100
	(336)	(77·8)	(160)	(113)	(599)	(118)

* Standard errors are derived from 100 bootstrap samples repeating the entire model selection process

1984 whose gestational age exceeded 28 weeks and were summarized in 42 bins (lowest bin having upper bound 1000 g with 100 g spacing except for endmost bins). One Belgian data set consisted of 99,547 singleton live and still births to native-born parents in 1984 and was recorded in 51 bins (lowest upper bound 600 g with 100 g spacing except for endmost bins and penultimate bin of 300 g). The other Belgian data set consisted of 4435 singleton live and still births to North-African-born parents and was recorded in 48 bins (lowest upper bound 700 g with 100 g spacing except endmost bins). Figure 2 shows plots of observed birthweight distributions along with fitted frequency polygons for the predominant distribution, residual distribution, and their total for both AIC- and BIC-selected models with the corresponding parameter estimates in Table II.

A glance at Figures 2(*a*)–(*d*) reveals that the models selected by AIC and BIC differed sharply in the way they decomposed the Norwegian and native-parent Belgian data into predominant and residual components but the selected models are more similar for the North-African-parent Belgian data (Figure 2(*d*) and (*e*)). For the two larger data sets, the AIC-selected model was the final model in the sequence and fit the data perfectly, but the resulting description was not all consistent with the conceptualization that we sought to enforce. On the other hand, the BIC-selected models visually appeared to fit the data satisfactorily and reflected the desired conceptualization. For the smallest data set, both selection procedures provided similar acceptable descriptions of the birthweight distributions. Note that all three birthweight distributions exhibited residual components in the upper tail.

The variability in the parameter estimates from all three data sets was much larger among AIC-based as opposed to BIC-based estimates (Table II). In fact, the AIC-based estimates often had standard errors an order of magnitude or more larger than their BIC-based counterparts. This result reflects the same volatility in the AIC-based estimates evident in the simulations reported in Section 4. However, with the actual data sets, the extreme volatility of the AIC estimates extended to the smallest data set, a sample size where in the simulated data AIC and BIC seemed to have comparable variability.

We refit the same three data sets using 200 g and 500 g bins (estimates not shown). BIC-based estimates using the other bin widths were comparable to their counterparts using 100 g bins. However, the resampling variances for estimates from 200 g or 500 g bins were somewhat larger than those for the 100 g bins – which one might expect since adjustment of the boundaries of residual support in larger steps could well induce more variability than adjustment in steps of 100 g. The AIC-based estimates using the larger bins were generally closer to their BIC-based counterparts than to the AIC-based estimates using 100 g bins. Moreover, use of 200 g or 500 g bins reduced the variability of the AIC-based estimates compared to use of 100 g bins. We attribute this decrease in variability of the AIC-based estimates using larger bins to the fact that outliers, while still present, were generally less extreme than with 100 g bins. However, the variability of AIC-based estimates still exceeds that of BIC-based estimates in these three examples, except for 500 g bins.

6. DISCUSSION

Our method offers advantages over methods previously employed[4,6] for describing birthweight distributions. We allow the residual distribution to have support in the upper or lower tail or in both – a capability that appears to be important with actual data. In addition, our method uses information from all bins to estimate the predominant distribution instead of ignoring information in bins taken as contaminated. While designed explicitly for birthweight data, one could employ the technique in any setting where contamination of an hypothesized predominant distribution might distort results.

Based on our simulations and experience with actual birthweight data, we recommend use of the BIC-selected estimates. At the bin widths that we investigated, these estimates are more reproducible than their AIC counterparts. The BIC-based estimates that used 200 g bins appear slightly less biased at small sample sizes but slightly more variable than those using 100 g bins. Considering the great savings in computer time available with 200 g bins, we prefer that width for routine use. We would make 500 g bins a second choice, particularly if the parameters of the predominant distribution were the primary interest. The use of 500 g bins greatly increases the time savings but also increases the possibility of bias from missing contamination that narrower bins would detect.

Employing our technique to compare two birthweight samples with strikingly different sizes may be problematic because sample-size-induced biases in measurement may appear to be actual differences between the populations. One can, however, ameliorate this problem by applying the technique to repeated subsamples (with replacement) of the same size from each original sample. Comparing measurements from sets of subsamples of a common size should diminish biases in estimated differences between populations because the bias in each population's own measurement should have similar magnitude.

Several alternative approaches for describing birthweight distribution are available, but the most obvious seem to lack ease of interpretation comparable to ours. One alternative is to fit a flexible four-parameter distribution that allows for adjustable degrees of skewness and tail weight, but the additional parameters are not necessarily epidemiologically interpretable. Since one might expect birthweight data to reflect a complex mixture, a second alternative approach is to fit a finite mixture of an unspecified number of normal components. However, one may find it daunting to provide a reasonable interpretation for the resulting components, and determining how many components is analogous to our problem of determining how many contaminated bins.

Although we couch our presentation in terms of the maximum likelihood estimation of parameters in certain multinomial models, we view our technique as an instrument for measuring birthweight distributions in terms of quantifiable concepts already familiar and useful to epidemiologists. The issue is not simply whether these models fit the data but rather whether they meet an established notion of what constitutes an acceptable description of a birthweight distribution. Even when the model that underlies our estimation procedure does not hold, our technique would still measure operationally-defined features of the birthweight distribution. This idea provides justification for giving variance properties more importance than unbiasedness in recommending BIC-based estimates at all sample sizes.

The description of birthweight distributions on which our procedure is based is somewhat arbitrary in the sense that models that offer disparate descriptions should be capable of equally good fits. For example, a normal distribution with large variance might fit both tails while 'contamination' provides mass in the centre. Alternatively, non-normal specifications for the predominant component might produce comparable fits. Oja *et al.*,[15] addressing the problem of jointly modelling birthweight and gestational age, use the log-normal family to model birthweight given gestational age. Justification for the conceptualization that our procedure enforces must come from epidemiological research that evaluates whether such descriptions lead to improved understanding of neonatal health.

ACKNOWLEDGEMENTS

The authors would like to thank Professor Pierre Buekens and Professor Rolv Skjærven for their kind permission to use data from Belgium and Norway, respectively, Dr. Beth Gladen and Dr. Joseph Haseman for carefully reading the manuscript and for suggestions that improved the presentation, and two anonymous reviewers for posing questions that prompted further improvements.

REFERENCES

1. Wilcox, A. J. and Skjærven, R. 'Birthweight and perinatal mortality: the effect of gestational age', *American Journal of Public Health*, **82**, 378–382 (1992)
2. Adams, M. S., MacLean, C. J. and Niswander, J. D. 'Discrimination between deviant and ordinary low birth weight: American Indian infants', *Growth*, **32**, 153–159 (1968).
3. Ashford, J. R., Brimblecombe, F. S. W. and Fryer, J. G. 'Birthweight and perinatal mortality in England and Wales 1956–65', *in* McLachlan, G. (ed), *Problems and Progress in Medical Care* (3rd series), Oxford University Press, London, for Nuffield Provincial Hospitals Trust, 1968, pp. 1–30.
4. Wilcox, A. J. and Russell, I. T. 'Birthweight and perinatal mortality: I. On the frequency distribution of birthweight', *International Journal of Epidemiology*, **12**, 314–318 (1983).
5. Wilcox, A. J. and Russell, I. T. 'Birthweight and perinatal mortality: III. Towards a new method of analysis', *International Journal of Epidemiology*, **15**, 188–196 (1986).
6. Kiely, J. L. and Kleinman, J. C. 'Birth-weight-adjusted infant mortality in evaluations of perinatal care: towards a useful summary measure', *Statistics in Medicine*, **12**, 377–392 (1993).
7. Akaike, H. 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control*, **AC-19**, 716–723 (1974).
8. Schwarz, G. 'Estimating the dimension of a model', *The Annals of Statistics*, **6**, 461–464 (1978).
9. Hurvich, C. M. and Tsai, C.-L. 'Regression and time series model selection in small samples', *Biometrika*, **76**, 297–307 (1989).

10. Mills, J. A. and Prasad, K. 'A comparison of model selection criteria', *Econometric Reviews*, **11**, 201–234 (1992).

11. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B*, **39**, 1–38 (1977).

12. Rao, C. R. *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.

13. Self, S. G. and Liang, K.-Y. 'Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions', *Journal of the American Statistical Association*, **82**, 605–610 (1987).

14. Feng, Z. and McCulloch, C. E. 'Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space', *Statistics & Probability Letters*, **13**, 325–332 (1992).

15. Oja, H., Koiranen, M. and Rantakallio, P. 'Fitting mixture models to birth weight data: a case study', *Biometrics*, **47**, 883–897 (1991).